Advancing Image and Video Recognition with Less Supervision

Anna Kukleva

A dissertation submitted towards the degree
Doctor of Engineering Science (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

Saarbrücken, 2024

Date of Colloquium:	01.08.2024
Dean of the Faculty:	Prof. Dr. Roland Speicher
Chair of the Committee:	Prof. Dr. Vera Demberg
Advisor:	Prof. Dr. Bernt Schiele
Reviewers:	Prof. Dr. Hilde Kuehne
	Prof. Dr. Dima Damen
	Prof. Dr. Kate Saenko
Academic Assistant:	Dr. Xinting Hu

Abstract

Deep learning is increasingly relevant in our daily lives, as it simplifies tedious tasks and enhances quality of life across various domains such as entertainment, learning, automatic assistance, and autonomous driving. However, the demand for more data to train models for emerging tasks is increasing dramatically. Deep learning models heavily depend on the quality and quantity of data, necessitating high-quality labeled datasets. Yet, each task requires different types of annotations for training and evaluation, posing challenges in obtaining comprehensive supervision. The acquisition of annotations is not only resource-intensive in terms of time and cost but also introduces biases, such as granularity in classification, where distinctions like specific breeds versus generic categories may arise. Furthermore, the dynamic nature of the world causes the challenge that previously annotated data becomes potentially irrelevant, and new categories and rare occurrences continually emerge, making it impossible to label every aspect of the world.

Therefore, this thesis aims to explore various supervision scenarios to mitigate the need for full supervision and reduce data acquisition costs. Specifically, we investigate learning without labels, referred to as self-supervised and unsupervised methods, to better understand video and image representations. To learn from data without labels, we leverage injected priors such as motion speed, direction, action order in videos, or semantic information granularity to obtain powerful data representations. Further, we study scenarios involving reduced supervision levels. To reduce annotation costs, first, we propose to omit precise annotations for one modality in multimodal learning, namely in text-video and image-video settings, and transfer available knowledge to large copora of video data. Second, we study semi-supervised learning scenarios, where only a subset of annotated data alongside unlabeled data is available, and propose to revisit regularization constraints and improve generalization to unlabeled data. Additionally, we address scenarios where parts of available data is inherently limited due to privacy and security reasons or naturally rare events, which not only restrict annotations but also limit the overall data volume. For these scenarios, we propose methods that carefully balance between previously obtained knowledge and incoming limited data by introducing a calibration method or combining a space reservation technique with orthogonality constraints. Finally, we explore multimodal and unimodal open-world scenarios where the model is asked to generalize beyond the given set of object or action classes. Specifically, we propose a new challenging setting on multimodal egocentric videos and propose an adaptation method for vision-language models to generalize on egocentric domain. Moreover, we study unimodal image recognition in an open-set setting and propose to disentangle open-set detection and image classification tasks that effectively improve generalization in different settings.

In summary, this thesis investigates challenges arising when full supervision for training models is not available. We develop methods to understand learning dynamics and the role of biases in data, while also proposing novel setups to advance training with less supervision.

Zusammenfassung

Deep Learning wird zunehmend relevant in unserem täglichen Leben, da es mühsame Aufgaben vereinfacht und die Lebensqualität in verschiedenen Bereichen wie Unterhaltung, Lernen, automatische Unterstützung und autonomes Fahren verbessert. Die Nachfrage nach mehr Daten zur Schulung von Modellen für aufkommende Aufgaben steigt jedoch dramatisch an. Deep Learning Modelle sind stark abhängig von der Qualität und Quantität der Daten, was hochwertige gelabelte Datensätze erfordert. Doch jede Aufgabe erfordert unterschiedliche Arten von Annotationen für Training und Evaluation, was Herausforderungen bei der Beschaffung darstellt. Die Beschaffung von Annotationen ist nicht nur ressourcenintensiv in Bezug auf Zeit und Kosten, sondern führt auch zu Verzerrung, wie z.B. Granularität in der Klassifizierung, wo Unterscheidungen wie spezifische Tierrassen gegenüber generischen Kategorien entstehen können. Darüber hinaus führt die dynamische Natur der Welt dazu, dass zuvor annotierte Daten potenziell irrelevant werden und neue Kategorien und seltene Ereignisse kontinuierlich auftauchen, was es unmöglich macht, jeden Aspekt der Welt zu kennzeichnen.

Daher zielt diese Arbeit darauf ab, verschiedene Supervisionszenarien zu erkunden, um den Bedarf an vollständiger supervison zu reduzieren und die Kosten für die Datenerfassung zu senken. Speziell untersuchen wir das Lernen ohne Lables, das als self-supervised und unsupervised bezeichnet wird, um Video- und Bildrepräsentationen besser zu verstehen. Um aus Daten ohne Labels zu lernen, nutzen wir injizierte Priors wie Bewegungsgeschwindigkeit, -richtung, Handlungsreihenfolge in Videos oder semantische Informationsgranularität, um leistungsstarke Datenrepräsentationen zu erhalten. Weiterhin untersuchen wir Szenarien mit reduzierter Supervision. Um die Kosten für Annotationen zu reduzieren, schlagen wir zunächst vor, präzise Annotationen für eine Modalität im multimodalen Lernen zu unterlassen, nämlich in Text-Video- und Bild-Video-Szenarien, und vorhandenes Wissen auf große Korpora von Videodaten zu übertragen. Zweitens untersuchen wir Semi-Supervised Lernszenarien, bei denen nur eine Teilmenge annotierter Daten neben unannotierten Daten verfügbar ist, und schlagen vor, Regularisierungsbeschränkungen zu überdenken und die Verallgemeinerung auf unannotierten Daten zu verbessern. Zusätzlich behandeln wir Szenarien, in denen Teile der verfügbaren Daten aufgrund von Datenschutz- und Sicherheitsgründen oder natürlich seltenen Ereignissen von Natur aus begrenzt sind, was nicht nur die Annotationen einschränkt, sondern auch das gesamte Datenvolumen begrenzt. Für diese Szenarien schlagen wir Methoden vor, die sorgfältig zwischen zuvor erhaltenem Wissen und eintreffenden begrenzten Daten abwägen, indem wir eine Kalibrierungsmethode einführen oder eine Raumreservierungstechnik mit Orthogonalitätsbeschränkungen kombinieren. Schließlich untersuchen wir multimodale und unimodale Szenarien in einer offenen Welt, in denen das Modell gebeten wird, über den gegebenen Satz von Objekt- oder Aktionsklassen hinaus zu generalisieren. Speziell schlagen wir eine neues herausforderndes Szenario für multimodale egozentrische Videos vor und schlagen eine Anpassungsmethode für Vision-Sprach-Modelle vor, um in der egozentrischen Domäne zu generalisieren. Darüber hinaus untersuchen wir die unimodale Bilderkennung in einem Open-Set Szenario und schlagen vor, Open-Set-Erkennung und Bildklassifizierungsaufgaben zu entflechten, die die Generalisierung in verschiedenen Einstellungen effektiv verbessern.

Zusammenfassend untersucht diese Arbeit die Herausforderungen, die entstehen, wenn eine vollständige Überwachung für das Training von Modellen nicht verfügbar ist. Wir entwickeln Methoden, um das Lernverhalten und die Rolle von Verzerrungen in Daten zu

verstehen, während wir gleichzeitig neuartige Setups vorschlagen, um das Training mit weniger Supervision voranzutreiben.

Acknowledgements

I would like to express my deepest gratitude to all those who supported me throughout my PhD journey.

First and foremost, I would like to thank my advisor, Bernt Schiele, for his invaluable guidance, patience, and support throughout my research. His vision and experience guided my projects, and I greatly appreciate his openness to students' needs and truly enjoyed his sense of humor. I am deeply grateful to Bernt for creating an atmosphere in the group that encourages curiosity, deep research discussions, and the pursuit of new initiatives. His constant availability and support made a significant difference in my PhD experience.

I am also extremely grateful to Hilde Kuehne. Her support extended far beyond my PhD and has been a constant throughout my research journey. As the mentor of my first research project when I was a master student at Bonn University, Hilde taught me how to take my first steps in research. Our continued collaboration throughout my PhD on multiple projects has been invaluable. Her insights, expertise, and belief in my work, combined with her unique positive outlook on any problem, have motivated me to persevere. I hope that some of her positivity has taken root in me as well.

I am also thankful for the mentorship of Christian Rupprecht. His insights, unique perspectives, and innovative ideas were crucial in guiding one of my projects to completion. It was a valuable experience to work with him, and I am grateful for the opportunity.

Special thanks to my collaborators with whom I had the pleasure of working: Nina, Fan, Moritz, Wei, Noor, Enea, Philipp, and Dengxin. I learned so much from each of you, and I believe our close collaborations have significantly contributed to the breadth of my understanding and to the researcher I am today.

I was fortunate to do an internship at Meta Zurich during the final year of my PhD. I would like to thank Fadime Sener for being an excellent manager, and Edo, Burgra, Eric, and Shugao for creating an incredible team atmosphere. Not only did we have fruitful discussions, but I also gained valuable insights into the workings of the industry.

I would like to particularly thank all my colleagues at D2 at MPI: Nina, Julian, David, Moritz, Sid, Anurag, Bharat, Fan, Garvita, Yaoyao, Mattia, Sukrut, Xinting, Zhi, Ada, Haoran, Jiahao, Jonas, Max, Mo, Sweta, Jan, Jovita, Julia, Steffen, Margret, Gerard, Vladimir, Aymen, Verica, and all other current and former colleagues. The in-depth discussions, push-up and coffee breaks, retreats, seminars, and reading groups made the research environment enjoyable and supportive.

A special thanks to Connie Balzert, an amazing secretary and person. Her support has been invaluable, ensuring that everyone in our department can focus on research. I have truly enjoyed our three-language conversations, and her role in the life of our department cannot be overstated.

To my friends, who have been a source of strength and happiness, thank you for your understanding and for keeping me grounded during this challenging process. I am grateful that many of my colleagues have become close friends. Outside MPI, I would like to specifically thank Sebastian, Fadime, Masha, Vova, Stas, and Yann.

Last but not least, I want to thank my love and husband, Nik, and my family for their incredible support and for always being there for me during this journey.

Contents

Introduction 1

Contents

Deep learning is increasingly relevant in our daily lives. Recent advances in training large-scale language models (LLMs) [BMR+20, TAB+23, TLI+23] demonstrate how research-based tools can be integrated continuously in our daily life from verifying grammar errors in our emails to organizing our daily routine. Development in all different directions such as entertainment [BBC+19], education [MRRB+23], automatic assistance [CGR+24], autonomous driving [MSWL23] and many others [MRA+23, DMCD19], targets to improve our lives and avoid tedious and time-consuming tasks. So far, the advances in deep learning are based on the availability of data and respective annotations for one or another task. However, collecting annotations is a time-consuming and labour-intensive process [KSH12, TKM+24]. It can depend on the skills, and experience of the annotators and on the size of the project itself. Annotations can introduce bias, which can lead to unfair or discriminatory outcomes [BCZ+16, BG18]. This is because the annotations are often created by humans, who are not immune to biases themselves. Whenever we deal with private data [HWVDI+19], creating annotations involves revealing personal information about individuals. This is especially true for sensitive data, such as medical records or financial information. The world is constantly changing, and the data used in the first place may quickly become outdated [DLAM+21]. This can make it difficult for models to keep up with the real world and provide accurate predictions. It is impossible to label all of the data in the world, as our perception of the world and different tasks are not discriminatory and depends on individual experiences [RZ11]. Finally, ensuring the quality of annotations is critical for performance. However, it is hard to guarantee high quality of the data and annotations due to human mistakes [NBMS21], and low-quality annotations can lead to models that make poor predictions. All these challenges indicate that if we want to move to more data and more different tasks, we need to find ways to reduce the annotation and data collection costs.

Transitioning from accurately curated datasets and controlled environments to real-world applications introduces unexplored challenges in the field of computer vision. This shift necessitates a deeper exploration of techniques to effectively utilize available data while reducing the reliance on annotations for the new downstream tasks. The central question in deep learning is how to learn robust representations [LBH15], which serve as the bridge between raw input data and the desired outcomes for downstream tasks. Learned representations are pivotal for

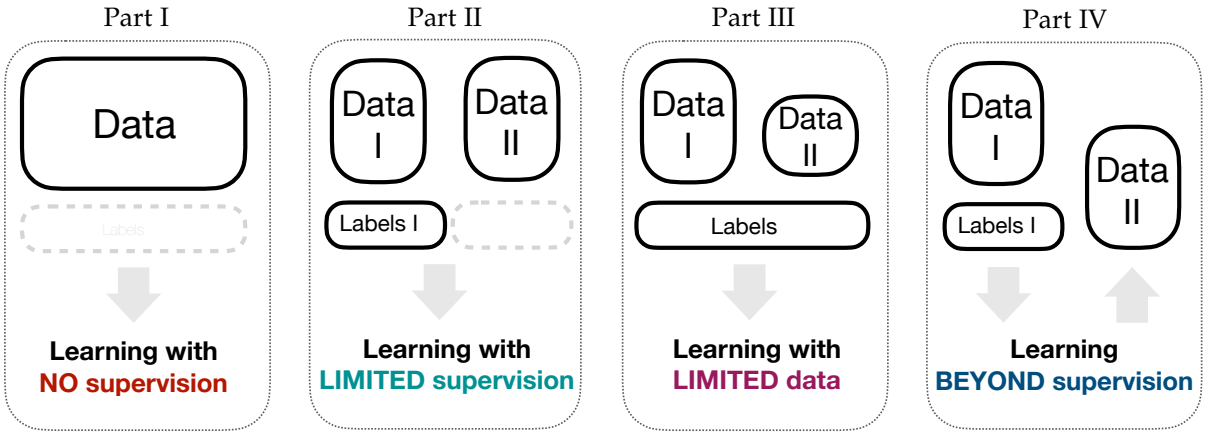


Figure 1.1: **Levels of supervision tackled in this thesis.** First, in Part I, we discuss learning representations relying on human annotations of the data. Then, in Part II, we explore various scenarios where only partial annotations are accessible during training, alongside of unlabeled data. In Part III, we explore scenarios where only a limited number of labeled samples are available during training. Lastly, in Part IV, we discuss methods that tailored for unconstrained environments and classification tasks that extend beyond the classes seen during training.

accurately addressing various tasks. However, as the process of collecting data is expensive, there is an increasing interest in methods that can autonomously learn representations or adapt them with varying degrees of supervision, e.g. leveraging learning from noisy web data at scale [MZA+19], exploring self-supervised learning strategies [ODM+23], and facilitating knowledge transfer across different modalities [RKH+21]. Despite the remarkable progress achieved with foundation models [RKH+21, SHG+22, GENL+23] — trained on extensive datasets — specialized tasks still necessitate task-specific data for effective model training and evaluation. Therefore, there is a substantial need to explore methods to reduce the reliance on annotated data while ensuring the continued advancement of computer vision applications in real-world scenarios.

This thesis focuses on exploring various supervision scenarios to mitigate the need for full supervision and reduce data acquisition costs. In Figure 1.1, we depict a high-level overview of the problems address in this thesis. In Part I, we focus on self-supervised and unsupervised learning methods to learn powerful representations without any labels that generalize well across different downstream tasks. To learn these representations, we explore and inject different perceptional biases that reflect structure of the data or the downstream task for both images and videos. In Part II, we gradually increase availability of supervision and explore different multi- and uni-modal settings with limited supervision. We explore how to reduce annotation cost to transfer image and video representations to the downstream tasks relying only on language modality, how to annotate unlabeled videos relying on images from the web, or how to exploit large language models to improve the quality of the large-scale video dataset collected from the web. On the other side, we explore uni-modal settings by limiting available annotations for the available classes while relying additionally on the unlabeled data. To improve the performance we revisit regularization techniques that allow for better utilization of the unlabeled data. In Part III, we study challenges when acquiring data incrementally, with new information arriving in continuous portions and with only few labeled samples, so called few-shot class incremental learning scenario. Typically, the model first pretrained on the large dataset to ensure meaningful initial representations. However, to improve the performance

across all classification tasks with abundant data and with very limited available data, it is necessary to explore regularization, influence of pretraining and space reservation that restricts representations for over represented classes and allow to balance between different classes. In Part IV, we further explore the generalization of pretraining in open world scenarios. When the model relies only on visual input data, the open world scenario often targets filtering all unknown classes during evaluation and predicting with high certainty classification of the known classes. Whereas for jointly trained vision-language models the open world scenario is reformulated to become the ability to classify classes that have not been seen during training. These models rely on the knowledge from the large-scale web vision-text pretraining, therefore posses general knowledge acquired from the open world. However, these models do not generalize well on specific downstream tasks. Therefore, we explore adaptation methods that allows to keep the generalization knowledge of these models while transferring them to specific downstream tasks.

For the rest of this chapter, we discuss each topic and explain our contributions. Then, we provide an outline of the thesis with relevant publications.

1.1 Learning Without Supervision

The learning without supervision paradigm comprises two interconnected sub-fields: self-supervised learning and unsupervised learning. Both approaches involve extracting powerful representations from raw data without relying on labeled information[SSSK21]. However, they differ in their overarching objectives. Self-supervised learning aims to derive robust representations from extensive amounts of unlabeled data, often referred to as the pretraining stage. Models trained in this phase serve as initialization and can be readily adapted to various downstream tasks. In contrast, unsupervised learning directly targets specific downstream tasks, such as clustering or semantic segmentation. Consequently, unsupervised learning frameworks involve learning targeted representations and subsequently grouping input features to accomplish the desired task [MGL+18]. In light of the availability of large-scale datasets sourced from the web, there has been a growing interest in leveraging self-supervised methods to learn robust representations. These generalized representations can then be applied to unsupervised tasks, thereby enhancing their performance.

To leverage unlabeled data effectively, it is essential to devise tasks that can be tackled without direct human supervision. These tasks often draw upon human experiences or deliberately introduced biases, such as controllable input transformations. Existing self-supervised methods can be broadly categorized into two groups: discriminative and generative methods [DGE15, OLB+23]. Discriminative methods focus on learning feature representations that facilitate solving pretext tasks. Their objective functions resemble those of supervised methods, assessing how well the model performs on pretext tasks like jigsaw puzzles [NF16] or instance discrimination [MM20, HFW+20]. On the other hand, generative models aim to learn the distribution over the pixel space, utilizing various input corruption techniques such as masking [WFX+22] or noise injection [LMS17, ZXL+20] with the following reconstruction the original input. Despite the diverse range of pretext tasks, instance discrimination has emerged as one of the most effective self-supervised approaches [GSA+20, CKNH20, ZMKG23]. In this thesis, we heavily rely on contrastive learning [CXH21, MZA+19], a type of instance discrimination method where the model is invariant to different transformations of the input.

Despite the successes of self-supervised methods, these approaches remain primarily empirically driven and necessitate further theoretical analysis and comprehension. For instance, understanding the influence of implicit uniform priors during training on downstream perfor-

mance [ABD+22], understanding the semantic differences between various methods [BIS+23], and exploring why methodologies like BYOL [GSA+20] and SimSiam [CH21] do not collapse [TCG21] are crucial directions. A more profound grasp of the learning dynamics can not only help to optimize existing algorithms and enhance their performance but also reveal more efficacious solutions. Furthermore, formulating effective and optimal self-supervised tasks remains an ongoing challenge [BIS+23]. On one hand, determining the most suitable pretext task for capturing comprehensive information from vast datasets is a substantial task. By scaling both models and data and experimenting with diverse augmentations, improvements in generalization performance across various tasks can be achieved [GDG+17, WDH+23]. On the other hand, identifying prior information to inject for specific tasks to facilitate unsupervised learning is equally crucial [VGVGVG21]. Some tasks may require additional adaptation, particularly if the domain significantly deviates from the web pretraining domain [MCN+23, JTC+22]. Addressing this limitation spurs further exploration in unsupervised learning, aiming to directly comprehend the underlying data structure for downstream tasks. The exploration of self-supervised and unsupervised tasks and their impact on diverse downstream tasks is central for deploying these models effectively in real-world scenarios.

1.1.1 Contributions

In this section, we provide a summary of our contributions in this area.

In Chapter 3, we explore the task of temporal detection and temporal segmentation of actions in untrimmed instructional videos in an unsupervised way. First, we propose to learn a continuous temporal embedding of frame-based features to obtain task-specific video feature representations. The embedding exploits the fact that some actions need to be performed in a certain order and we use a network to learn an embedding of frame-based features with respect to their relative time in the video. Further, to solve the downstream task in an unsupervised way, we propose a decoding of the videos into coherent action segments based on an ordered clustering of the embedded frame-wise video features. We observe that our embedding especially benefit from the particular structure of the instructional videos and brings advantage on the task of unsupervised temporal segmentation compared to more general representations.

In Chapter 4, we explore unintentional actions and propose a multi-stage framework that exploits inherent biases such as motion speed, motion direction, and order of activities to recognize such actions. We formulate the temporal transformations specifically for unintentional actions to bias the model to capture sudden changes in the behaviour of the video and then pretrain the model in a self-supervised way to predict these transformations. Moreover, our multi-stage approach models the temporal information on both the level of individual frames and full clips. We observe that our task-specific representations significantly improve the performance across different tasks for unintentional actions.

In Chapter 5, we analyse the behaviour of one of the most popular variants of self-supervised methods, i.e. contrastive methods, on long-tail data. In particular, we investigate the role of the temperature parameter τ in the contrastive loss and find that a large τ emphasises group-wise discrimination, whereas a small τ leads to a higher degree of instance discrimination. While τ has thus far been treated exclusively as a constant hyperparameter, in this work, we propose to employ a dynamic τ and show that a simple cosine schedule can yield significant improvements in the learnt representations.

1.2 Learning with Limited Supervision

To effectively train powerful and robust models that demonstrate generalization across diverse tasks, reliance on large-scale datasets is necessary [ZKHB22, DDM+23]. However, the process of collecting and annotating such datasets is labor-intensive and costly. In response to this challenge, current practical interests lie in strategies that mitigate the need for extensive labeled data. This includes leveraging weak annotations across multiple modalities [MKL+24, ZLH+21], or relying on a limited number of labeled samples supplemented by a large volume of unlabeled data during training [YSKX22]. Traditional large-scale fully labeled datasets often cover a narrow range of concepts due to limitations in expansion [ima]. Relaxing the requirements for annotation leads to a broader coverage of concepts within the dataset [SVB+21], which proves especially advantageous in domains where data availability is inherently sparse [GWB+22a] or resource limitations hinder the acquisition of sufficient data for downstream tasks [KSM+21, SKL+21]. Therefore, it is crucial to explore approaches that demand less supervision and can effectively learn from limited supervision.

When only a few labeled samples are available, constructing a successful learning system poses a significant challenge. In the past decade, efforts to enhance the reliability of such systems have led to the proposal of semi-supervised settings [CSZ09, ZG22]. In this framework, a small amount of labeled training data is accompanied by vast amounts of unlabeled data, with the goal of jointly learning from both types of data. The primary challenge of semi-supervised learning lies in effectively leveraging the limited labeled data and abundant unlabeled data to enhance generalization performance. One of the most common techniques employed is pseudo-labeling, where unlabeled data is assigned labels for training [SBL+20]. However, pseudo-labeling inherits the trade-off problem between quantity and quality, as it relies on hard thresholding, which can hinder learning [CTF+23]. Another prominent technique is consistency regularization, aimed at ensuring consistent predictions for similar data samples [SJT16, LA17]. Various works have focused on improving pseudo-labeling and consistency regularization from different angles to effectively leverage limited labeled data and transfer knowledge to unlabeled data [XSY+21, ZWH+21, BRS+22]. Nevertheless, the central question of how to leverage different types of supervision for a single modality remains challenging. Thus, we further explore regularization methods to enhance generalization.

Learning from large-scale multimodal data has attracted increasing attention in recent years, as the collection of noisy web data allows for significant scaling of datasets [ZLH+21, SVB+21]. A pioneering work in successful text-image large-scale pretraining is CLIP (Contrastive Language-Image Pre-training)[RKH+21], where the authors investigate the behavior of image classifiers trained with natural language at scale. This and similar works[IWW+21, WYY+22, JYX+21] demonstrate remarkable zero-shot performance and cross-modality retrieval across a wide range of tasks. To construct the large-scale dataset, the authors of CLIP curated a new dataset comprising 400 million image-text pairs to cover a broad set of visual concepts. The query set encompasses 500,000 instances for searching text-image pairs on the internet, based on the frequency of words occurring in Wikipedia. Following the success of raw alt-text image pair pretraining, numerous web datasets have emerged to enable large-scale pretraining. However, due to the inherent noise in the data, filtering becomes necessary to avoid suboptimal performance [LLXH22, LLSH23a, ARZV21, SVB+21, IWW+21]. This necessity is particularly pronounced for video web data, as the corresponding text often originates from automatic speech recognition systems (ASR). Supervision derived from ASR-video pairs [MZA+19] can be highly incoherent, incorporating unrelated speech or suffering from temporal misalignment between spoken words and video content. Consequently, the collection and annotation of video

data remain significantly more challenging than that of static images.

To bridge the gap between the success of text-image pretraining and text-video pretraining, efforts have been directed towards aligning automatic speech recognition (ASR) with video frames [HXZ22] and toward collecting cleaner and larger video datasets [XHZ+22, WHL+23, CLW+24]. Methods that leverage both image and video datasets [BNVZ21, NSS+22] have proven to be efficient in making the collection and labeling of video datasets more streamlined. These methods utilize less noisy alt-text image pairs, readily accessible from the web, as a labeling-free data source for video datasets or downstream video tasks. The enormous successes of Large Language Models (LLMs)[BMR+20, Ope23, TLI+23] have generated increasing interest in utilizing LLMs across various downstream tasks. One prominent direction is the enhancement of existing datasets or the creation of new and improved datasets for multimodal video training[ZMKG23, WHL+23]. LLMs facilitate the creation of automatic pipelines that improve alignment between modalities, enhance action descriptions, augment existing language data, and summarize the content of video frames.

Despite the success of integrating general knowledge from LLMs and utilizing the collection of video data from various sources, several inherent limitations remain. These limitations arise from the fundamental differences between textual and visual representations, presenting significant challenges for multimodal understanding and processing. For instance, the distinction between discrete and symbolic text representations and the continuous and perceptual nature of visual representations creates difficulties in accurately aligning textual and visual cues. Furthermore, textual and visual data sources differ significantly in terms of availability, noise levels, and reliability. Addressing these and other challenges requires the development of innovative approaches in multimodal representation learning, fusion techniques, and model interpretability, thereby laying the groundwork for more robust and reliable multimodal systems.

1.2.1 Contributions

In this section, we provide a summary of our contributions in this area.

In Chapter 6, we study semi-supervised learning scenario, where only a subset of annotated data alongside unlabeled data is available, and propose to revisit consistency regularization idea. We find that enforcing invariance by decreasing distances between features from differently augmented images leads to improved performance. Moreover, encouraging equivariance instead, by increasing the feature distance, further improve the performance. Our FeatDistLoss imposes consistency and equivariance on the classifier and feature level and shows improved performance over the previous methods by a significant margin.

To address the challenge of transferring large-scale pretraining to the downstream tasks that requires labeled video-language pairs, in Chapter 7, we propose a new setting, text-video retrieval with uncurated & unpaired data, that during training utilizes only text queries together with uncurated web videos without any paired text-video data. To explore the new setting, we propose an approach, In-Style, that learns the style of the text queries and transfers it to uncurated web videos. We evaluate our model on retrieval performance over multiple datasets and demonstrate the advantages of our style transfer framework and close the gap towards fully supervised setting without the need of full supervision.

In Chapter 8, we exploit labeling-free web images for adapting to unlabeled target videos. This poses two major challenges: spatial domain shift between web images and video frames; and the modality gap between image and video data. To address these challenges, we propose a cycle-consistency based approach for unsupervised image-to-video domain adaptation. We

leverage the joint spatial information in images and videos on the one hand and, on the other hand, train an independent spatio-temporal model to bridge the modality gap. By alternating between spatial and spatio-temporal learning with knowledge transfer between the two, we show that our approach achieves significant performance improvements over previous methods.

As large-scale annotation-free web video training data remains sub-optimal for training text-video models, in Chapter 9, we propose to leverage the capability of large language models (LLMs) to obtain fine-grained video descriptions aligned with videos. Specifically, we prompt an LLM to create plausible video descriptions based on ASR narrations of the video for a large-scale instructional video dataset. To this end, we introduce a prompting method that is able to take into account a longer text of subtitles and align the captions to the video temporally by generating timestampes for each produced caption. In this way, we obtain human-style video captions without human supervision on a large scale and show significant performance improvements over many different benchmark datasets.

1.3 Learning with Limited Data

The human brain is remarkably proficient at generalizing new information based on learned knowledge from only a few samples for a given task. However, when only a limited number of samples are available, and it is difficult to collect additional data—such as in medical domains or scenarios involving private information—adaptation to new environments on-the-fly becomes crucial. Few-shot incremental learning [THC+20] exemplifies this setting, aiming to teach models to understand the world continuously, with new information arriving in small batches over time. Naturally, the main challenges in this context include catastrophic forgetting [GH10, KPR+17], where previously learned concepts are at risk of being overwritten by the latest updates, and overfitting [CLK+19, SSZ17], as the model may memorize scarce input data at the expense of its generalization ability. In essence, the neural network must strike a balance: maintaining learned knowledge while also adapting to new tasks. Moreover, due to the limited number of samples, standard empirical risk minimization algorithms may prove unreliable and unable to accurately approximate the optimal expected risk, leading to model overfitting [Biso6]. These challenges, coupled with the increasing relevance to real-world scenarios, have received significant attention in recent years.

More formally, the few-shot incremental learning for image classification setting involves two separate sets of classes: base classes and novel classes. Base classes represent the data used for pretraining, where each class has an abundant number of samples to learn robust representations. All base classes are available at once. However, novel classes are scarce and are typically represented by only 1 or 5 training samples per class [TLL+24]. Moreover, the classes arrive in a sequential manner. This unbalanced distribution, not available all at once, leads to the challenges mentioned above. One of the strongest baselines involves first learning powerful representations based on the base classes. Subsequently, the learned model is utilized to represent the incoming classes by a simple average of the representations in the embedding space [SSZ17]. This approach ensures a balanced classifier across all classes and excludes training on the novel classes, thus preventing overfitting. Notably, the effectiveness of this method heavily relies on the pretraining method. Research has shown [TWK+20] that stronger representations can lead to better separation between the base and novel classes.

Recent methods further explore various types of regularization techniques to calibrate the degree of learning for base and novel classes [CRF+21, her, MSR21, THC+20]. One of the most prominent recent approaches is the space reservation technique [ZYM+22, YYL+23, her],

which is based on the neural collapse paradigm [Kot22]. Contrary to the intuition that similar classes should be closer to each other in the embedding space, neural collapse suggests that for classification problems, the optimal structure of the corresponding classifier is a simplex equiangular tight frame. This structure places vertices on a hypersphere, making them linearly separable and positioned at the maximum possible distance from each other. Based on this insight, recent methods explore different approaches to implicitly reserve space through augmentations [PZW+22] or explicitly through fixed prototypes spread uniformly while preserving maximum distance from each other [YYL+23]. Despite recent progress, these methods encounter real-world challenges and demonstrate miscalibration and remain overconfident either on the base classes or on the novel classes.

1.3.1 Contributions

In this section, we provide a summary of our contributions in this area.

In Chapter 10, we propose a three-stage framework that allows to explicitly and effectively address catastrophic and calibration challenges that arise due to the incremental few-shot learning paradigm. While the first phase learns base classes with many samples, the second phase learns a calibrated classifier for novel classes from few samples based on base-normalized cross entropy while also preventing catastrophic forgetting. Our base-normalized cross-entropy amplifies the softmax output of novel classes to overcome the bias towards the base classes. In the final phase, we address the problem of calibration of the overall model across base and novel classes. Our framework achieves significant improvement over the previous methods for generalizaed and incremental few-shot learning.

In Chapter 11, we propose the OrCo framework built on two core principles: features' orthogonality in the representation space, and contrastive learning. In particular, we improve the generalization of the embedding space by employing a combination of supervised and self-supervised contrastive losses during the pretraining phase. Additionally, we introduce OrCo loss to address challgens arising from data limitations during incremental sessions. Through feature space perturbations and orthogonality between classes, the OrCo loss maximizes margins and reserves space for the following incremental data which allows accommodation of the incoming classes. Our experimental results showcase improved performance over the previous methods especially on the novel classes.

1.4 Learning Beyond Supervision

In the vast landscape of machine learning, open-world classification presents a unique set of challenges and opportunities that distinguish it from traditional closed-world classification tasks [JDC+20]. While closed-world classification assumes a fixed and exhaustive set of classes [KTS+14], open-world classification tackles the dynamic nature of data, where new classes can emerge over time, and existing classes may evolve or become irrelevant.

This dynamic nature of open-world classification poses a fundamental question: how can machine learning models effectively classify data instances when faced with an ever-expanding and uncertain set of possibilities? This question underscores the need for novel approaches that can adapt to new information, identify previously unseen classes to handle instances that do not fit within known categories [XLSA18, BB15] or classify new classes relying on external source of general knowledge [ZYLL22a, ZYLL22b].

Single-modality methods often need to determine with high certainty whether the incoming

input instance belongs to the previously seen distribution [GZJ+20, LWOL20, OP19, MSMD21], as seen in tasks such as anomaly detection [RVG+19] and novelty detection [ZGZ+20]. The interplay between the classification task for closed-set classes and the detection of outliers presents an important challenge as a multitask learning objective for image classification. Simply detecting unseen classes is commonly referred to as open-set learning [JDC+20], while learning in an open-world context indicates learning beyond the given set of classes. Previously, methods often relied on clustering approaches to identify distinct novel classes for further classification [SXL18, ZZL+21a, HRE+20]. However, with the advancements in vision-language models (VLMs) pretrained on large-scale datasets and accumulating more general knowledge about the world, recent methods utilize these models for classification beyond the given set of classes [ELRS22, ZLZ+23, WLYL23, ZYLL22a]. These methods adapt the pretrained VLMs to the given domain or downstream task. Similar to few-shot class incremental learning, one of the main challenges is catastrophic forgetting. Recent methods aim to achieve reliable performance on the closed-set classes while maintaining the generalization performance of the pretrained model and preventing overfitting to the closed-set classes. The common approach is to introduce a small set of new parameters for optimization that do not drastically change the model but allow for adaptation to the current domain [JHZ+22, LYF+23, HSW+22]. These methods leverage unseen information from the new domain, such as different input styles like cartoon style or egocentric videos, by adapting to the given domain based on the closed-set data while simultaneously transferring adapted general knowledge to the domain. To bridge the gap between fully supervised training on closed-set datasets and open-world generalization, integrating open-world learning techniques with traditional classification methods will create robust and adaptive machine learning systems, capable of handling dynamic and evolving datasets effectively.

1.4.1 Contributions

In this section, we provide a summary of our contributions in this area.

In Chapter 12, we study the challenging and realistic open-set semi-supervised learning setting, where the goal is to both correctly classify inliers and to detect outliers. We find that inlier classification performance can be largely improved by incorporating high-confidence pseudo-labeled data, regardless of whether they are inliers or outliers. By using non-linear transformations we propose to separate the features used for inlier classification and outlier detection in the mutlitask learning, preventing adverse effects between them. In experiments, we show that our framework greatly improves both inlier classification and outlier detection performance compared to existing methods.

In Chapter 13, we explore the open world setting by adapting vision-language models to the unexplored domain of egocentric videos. To address this problem, we propose a simple yet effective cross-modal adaptation framework. Using a video adapter, our pipeline learns to align frozen text embeddings to each egocentric video directly in the shared embedding space. Our novel adapter architecture is specifically tailored to egocentric videos, focusing attention on the hand region. We evaluate our approach in the open-world scenarios on the Epic-Kitchens, Ego4D, and EGTEA datasets for fine-grained cross-dataset action generalization, demonstrating the effectiveness of our method.

1.5 Outline

In this section, we provide a summary of the thesis by briefly outlining each chapter and establishing connections between them. We also acknowledge any relevant publications and collaborations with other researchers.

Chapter 2, Related Work. This chapter surveys related work which tackle the challenges of learning with less supervision with a particular focus on the four directions of the thesis: learning without supervision, learning with limited supervision, learning with limited data, and learning in open world. We discuss how these works relate to the methods and contributions presented in this thesis. Discussions of related work specific to the following chapters are provided within each chapter.

Part I, Learning Without Supervision

Chapter 3, Unsupervised Learning of Action Classes. In this chapter, we tackle the problem of unsupervised temporal action segmentation and classification for instructional videos. We discuss how to learn task-specific representation space based on the fact that some actions need to be performed in a certain order with the further decoding into action segments.

The content of this chapter corresponds to the CVPR 2019 publication with the title "*Unsupervised learning of action classes with continuous temporal embedding*" [KKSG19]. Anna Kukleva and Hilde Kuehne are the co-first authors of this work, under the supervision of Juergen Gall. Anna Kukleva contributed the main ideas for the temporal embedding, conducted all experiments and made the figures for the paper. Hilde Kuehne acted as Anna Kukleva's supervisor during the project and played a significant role in the paper's writing. Fadime Sener contributed to implementation of the baselines.

Chapter 4, Self-Supervised Training for Unintentional Actions. In this chapter, we explore unintentional actions and inherent biases that they present such as motion speed, motion direction, and order of activities. To recognize such actions on both level of individual frames and full clips, we propose mutli-stage framework with a transformer-based architecture in combination iwth conditional random fields to capture local and global temporal relations between intentional and unintentional actions.

The content of this chapter corresponds to the ECCV Workshop 2022 publication with the title "*Leveraging Self-Supervised Training for Unintentional Action Recognition*" [DKS22]. Anna Kukleva and Enea Duka are the co-first authors of this work, under the supervision of Bernt Schiele. This paper is based on the master thesis of Enea Duka. Anna Kukleva acted as Enea Duka's supervisor during the project and played a significant role in the paper's writing.

Chapter 5, Temperature Schedules for Self-Supervised Contrastive Methods. In this chapter, we develop a framework for understanding the underlying dynamics of the self-supervised contrastive learning loss through the lens of average distance maximisation with respect to the temperature parameter. While this temperature parameter has thus far been treated exclusively as a constant hyperparameter, in this work, we propose to employ a dynamic temperature parameter and show that a simple cosine schedule can yield significant improvements in the learnt representations on long-tail distributions.

The content of this chapter corresponds to the ICLR 2023 publication with the title "*Temperature Schedules for Self-Supervised Contrastive Methods on Long-Tail Data*" [KBS+23]. Anna Kukleva and Moritz Böhle are the co-first authors of this work, under the supervision of Bernt Schiele, Hilde Kuehne and Christian Rupprecht. Anna Kukleva and Moritz Böhle contributed in equal parts to the design of the temperature scheduling approach, the analytical framework, as well as the writing. The implementation of the experiments and the publication of the code for this project were fully handled by Anna Kukleva.

Part II, Learning with Limited Supervision

Chapter 6, Consistency Regularization for Semi-Supervised Learning. In this chapter, we tackle the semi-supervised learning setting, and propose to revisit the consistency regularization idea. We find that enforcing either invariance or equivariance improves the performance. Surprisingly, by increasing the feature distance between the samples our framework improves the performance by a significant margin compared to previous work.

The content of this chapter corresponds to the GCPR 2021 publication with the title "*Revisiting Consistency Regularization for Semi-Supervised Learning*" [FKS21]. The extended version of this work was invited to the special issue of the International Journal of Computer Vision (IJCV) and published at IJCV 2023 with the same title [FKDS23a]. Yue Fan was the lead author of this paper under the supervision of Bernt Schiele. The journal extension also involved supervision of Dengxin Dai. Anna Kukleva was involved in the weekly and scientific discussions, contributed to writing of the paper and creating the figures.

Chapter 7, In-Style: Bridging Text and Uncurated Videos. In this chapter, we propose a new setting, text-video retrieval with uncurated & unpaired data. This setting allows to avoid costly video annotations for the downstream tasks while relying only on targeted text data, uncurated videos and pretrained image-language models. We propose to learn and transfer the style of the text queries to the uncurated large-scale web video dataset to leverage our setting and close the gap between zero-shot evaluation and fully supervised fine-tuning.

The content of this chapter corresponds to the ICCV 2023 publication with the title "*In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval*" [SKSK23]. Anna Kukleva and Nina Shvetsova are the co-first authors of this work, under the supervision of Bernt Schiele and Hilde Kuehne. Anna Kukleva and Nina Shvetsova jointly led the project as co-first authors, sharing responsibilities across all stages of its development with equal contribution.

Chapter 8, Unsupervised Domain Adaptation to Learn from Image to Video. In this chapter, we explore how to annotate unlabeled videos by adapting video and image domains with web images as a source. We propose a cycle-based approach for the unsupervised image-to-video domain adaptation to tackle spatial domain shift between web images and video frames and the modality gap between image and video data.

The content of this chapter corresponds to the ECCV 2022 publication with the title "*CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video*" [LKS+22]. Wei Lin was the lead author of this paper under the supervision of Hilde Kuehne and Horst Bischof. Anna Kukleva was involved in the weekly and scientific discussions, contributed to writing of the paper and creating the figures.

Chapter 9, Prompting LLMs to Transform Video Annotations at Scale. In this chapter, we improve large-scale web video dataset by leveraging capabilities of large language models (LLMs). We prompt an LLM to create plausible video descriptions with timestamps based on ASR narrations of the video for a large-scale instructional video dataset. In this way, we obtain human-style video captions without human supervision on a large scale.

The content of this chapter corresponds to the paper with the title "*HowToCaption: Prompting LLMs to Transform Video Annotations at Scale*" [SKH+24]. This paper is ECCV 2024 publication. Anna Kukleva and Nina Shvetsova are the co-first authors of this work, under the supervision of Christian Rupprecht, Bernt Schiele and Hilde Kuehne. Nina Shvetsova proposed the idea and was leading the project. Anna Kukleva was involved equally with Nina Shvetsova in the project in the first months and then contributed to writing of the paper. Anna had to change focus due to the started internship. Xudong Hong participated in discussions related to the NLP aspect and gathered statistics of the generated captions.

Part III, Learning with Limited Data

Chapter 10, Few-Shot Learning by Explicit Learning and Calibration. In this chapter, we propose a three-stage framework that allows to explicitly and effectively address catastrophic and calibration challenges that arise due to the incremental few-shot learning paradigm. While the first phase learns base classes with many samples, the second phase learns a calibrated classifier for novel classes from few samples. To calibrate the training, we propose base-normalized cross-entropy to amplify the softmax output of novel classes and overcome the bias towards the base classes. In the final phase, we calibrate both novel and base classes.

The content of this chapter corresponds to the ICCV 2021 publication with the title "*Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration without Forgetting*" [KKS21b]. Anna Kukleva is the first author of this work, under the supervision of Hilde Kuehne and Bernt Schiele.

Chapter 11, Orthogonality and Contrast for Few-Shot Incremental Learning. In this chapter, we propose the OrCo framework built on two core principles: features' orthogonality in the representation space, and contrastive learning. Through feature space perturbations and orthogonality between classes, our framework maximizes margins and reserves space for the following incremental data. Our experimental results showcase performance improvements across benchmark datasets.

The content of this chapter corresponds to the CVPR 2024 publication with the title "*OrCo: Towards Better Generalization via Orthogonality and Contrast foro Few-Shot Class-Incremental Learning*" [AKS24]. This paper is accepted as a highlight at the main conference and as a spotlight at the 3^{rd} Workshop on Learning with Limited Labelled Data for Image and Video Understanding. Anna Kukleva and Noor Ahmed are the co-first authors of this work, under the supervision of Bernt Schiele. This paper is based on the master thesis of Noor Ahmed. Anna Kukleva acted as Noor Ahmed's supervisor during the project and played a significant role in the paper's writing.

Part IV, Learning Beyond Supervision

Chapter 12, Boosting Performance of Open-Set Semi-Supervised Learning. In this chapter, we tackle the realistic open-set semi-supervised learning setting. We explore the balance

between inlier classification and outlier detection tasks and find that the classification performance can be largely improved by incorporating high-confidence pseudo-labeled data, regardless of whether they are inliers and outliers. Moreover, we propose to separate the features used for classification and detection in the multitask learning to prevent adverse effects between the tasks.

The content of this chapter corresponds to the ECCV 2022 publication with the title "*SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning*" [FKDS23b]. Yue Fan was the lead author of this paper under the supervision of Dengxin Dai and Bernt Schiele. Anna Kukleva was involved in the weekly and scientific discussions, contributed to writing of the paper and creating the figures.

Chapter 13, X-MIC: Egocentric Action Generalization. In this chapter, we tackle the problem of adaptation of vision-language models to the egocentric video domain. We propose a simple yet effective cross-modal adaptation framework, which we call X-MIC. Using a video adapter, our pipeline learns to align frozen text embeddings to each egocentric video directly in the shared embedding space.

The content of this chapter corresponds to the CVPR 2024 publication with the title "*X-MIC: Cross-Modal Instance Conditioning for Egocentric Action Generalization*" [KSR+24]. This paper is accepted as a spotlight at the 3^{rd} Workshop on Learning with Limited Labelled Data for Image and Video Understanding. Anna Kukleva is the first author of this work, under the supervision of Fadime Sener, Bernt Schiele, and Shugao Ma. Edoardo Remelli, Bugra Tekin and Eric Sauser were involved in the weekly discussions. Edoardo Remelli and Bugra Tekin were involved in the writing of the paper. The work was done during the internship at Meta in 2023.

1.6 Publications

The content of this thesis has previously appeared in the following publications, ordered as outlined above:

- [KKSG19] **Anna Kukleva***, Hilde Kuehne*, Fadime Sener, and Jurgen Gall. (*equal contribution) "Unsupervised learning of action classes with continuous temporal embedding", CVPR 2019
- [DKS22] Enea Duka*, **Anna Kukleva***, and Bernt Schiele. (*equal contribution) "Leveraging Self-Supervised Training for Unintentional Action Recognition", ECCV Workshop 2022
- [KBS+23] **Anna Kukleva***, Moritz Boehle*, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. (*equal contribution) "Temperature Schedules for Self-Supervised Contrastive Methods on Long-Tail Data", ICLR 2023
- [FKS21] Yue Fan, **Anna Kukleva**, and Bernt Schiele. "Revisiting Consistency Regularization for Semi-Supervised Learning", GCPR 2021
- [FKDS23a] Yue Fan, **Anna Kukleva**, Dengxin Dai, and Bernt Schiele. "Revisiting Consistency Regularization for Semi-Supervised Learning", IJCV 2023
- [SKSK23] Nina Shvetsova*, **Anna Kukleva***, Bernt Schiele, and Hilde Kuehne. (*equal contribution) "In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval", ICCV 2023

- [LKS+22] Wei Lin, **Anna Kukleva**, Kunyang Sun, Horst Possegger, Hilde Kuehne, and Horst Bischof. "CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video", ECCV 2022
- [SKH+24] Nina Shvetsova*, **Anna Kukleva***, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. (*equal contribution) "HowToCaption: Prompting LLMs to Transform Video Annotations at Scale", ECCV 2024
- [KKS21b] **Anna Kukleva**, Hilde Kuehne, and Bernt Schiele. "Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration without Forgetting", ICCV 2021
- [AKS24] Noor Ahmed*, **Anna Kukleva***, and Bernt Schiele. (*equal contribution) "OrCo: Towards Better Generalization via Orthogonality and Contrast foro Few-Shot Class-Incremental Learning", CVPR 2024
- [FKDS23b] Yue Fan, **Anna Kukleva**, Dengxin Dai, and Bernt Schiele. "SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning", ICCV 2023
- [KSR+24] **Anna Kukleva**, Fadime Sener, Edoardo Remelli, Bugra Tekin, Eric Sauser, Bernt Schiele, and Shugao Ma. "X-MIC: Cross-Modal Instance Conditioning for Egocentric Action Generalization", CVPR 2024

Further contributions were made to the following works not discussed in this thesis:

- [SPK+23] Nina Shvetsova, Felix Petersen, **Anna Kukleva**, Bernt Schiele, and Hilde Kuehne. "Learning by Sorting: Self-supervised Learning with Group Ordering Constraints", ICCV 2023
- [LKP+23] Wei Lin, **Anna Kukleva**, Horst Possegger, Hilde Kuehne, and Horst Bischof. "TAEC: Unsupervised Action Segmentation with Temporal-Aware Embedding and Clustering", CEUR Workship 2023
- [FDKS22] Yue Fan, Dengxin Dai, **Anna Kukleva**, and Bernt Schiele. "COSSL: Co-learning of Representation and Classifier for Imbalanced Semi-Supervised Learning", CVPR 2022
- [VSK+21] Rosaura G VidalMata, Walter J Scheirer, **Anna Kukleva**, David Cox, and Hilde Kuehne. "Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences", WACV 2021

Related Work 2

Contents

In this chapter, we review literature on topics related to learning with less supervision. First, in Section 2.1, we discuss approaches for learning without any supervision signal, relying only on perceptional biases to obtain effective image and video representations. Following that, in Section 2.2, we discuss methods that depend on partial supervision or weak alignment between vision and language modalities. Moving onto Section 2.3, we discuss methods that require only a few labeled training samples during training and adapt to the novel tasks continuously. Finally, in Section 2.4, we cover related work on open-set and open-world classification. Subsequent chapters will also discuss related work, each tailored to the specific topic of the respective chapter.

2.1 Learning Without Supervision

The field of self-supervised representation learning (SSL) from visual data is a quickly evolving field. It involves learning effective representations from unlabeled data, leveraging inherent properties of images and videos and solving tasks based on these properties without relying on human annotations. Generally, self-supervised approaches fall into two categories: discriminative and generative methods [OLB+23]. Initially, we provide an overview of image-based methods followed by a discussion of video-specific techniques. Note that while we discuss them separately, image-based methods are also applicable to videos.

2.1.1 Discriminative Self-Supervised Learning

Puzzle Solvers. One of the earliest methods for self-supervised learning involves solving predefined puzzles, relying on injected transformations of the input to tackle the pretext

task [NF16, GSK18, NALV18, YPB16]. Predicting the type of geometric transformation applied to the input serves as an example of such tasks. Specifically, the model is required, for instance, to predict the direction of the rotation that was applied [GSK18]. Another popular task is solving jigsaw puzzles [NF16, PCJ21], where the objective is to accurately predict the relative locations of a certain number of patches obtained from the input image.

Contrastive Instance Discrimination. Contrastive learning methods [WXSL18] learn representations by forming positive pairs of images through augmentations and a loss formulation that maximizes their similarity while simultaneously minimizing the similarity to other samples. These methods have demonstrated remarkable representation quality and downstream performance by leveraging the InfoNCE loss [OLV18]. MoCo [HFW+20, CFGH20] establishes a dictionary look-up to incorporate a large set of negatives during training. In contrast, SimCLR [CKNH20, CKS+20] further enhances performance by integrating projection heads and employing strong augmentations. These frameworks serve as foundational methods upon which many subsequent approaches are built to enhance efficiency [FP22, ZZL+21b] or to investigate the role of components such as negatives [WWW+21, YHH+22, ZZP+22, ITAC18, KSP+20, RCSJ20, KAG22] or analyze learning dynamics [BZK+17, FV18, LFV20, LAV21]. Moreover, contrastive learning has found applications beyond SSL pre-training, including multi-modal learning [SCR+22], domain generalization [YBZ+22], semantic segmentation [VGVGVG21], 3D point cloud understanding [ADD+22], and 3D face generation [DYC+20].

Clustering. Clustering approaches rely on self-labeling of available data to generate supervision signals. However, when applying these methods to train deep learning models with large datasets, several challenges need to be addressed, including clustering large-scale data, issues of over-clustering and under-clustering, defining the number of clusters, and avoiding trivial solutions [RPY+22]. Deep Cluster [CBJD18] stands as one of the pioneering clustering-based self-supervised deep learning methods that have yielded remarkable results. These challenges have been tackled through the enforcement of uniform clusters via offline k-Means clustering. SeLA [ARV20] mitigates the collapse issue by employing the Sinkhorn-Knopp algorithm, while Online Deep Cluster [ZXL+20] utilizes online training to address them. The SWAV method [CMM+20] combines online clustering with a contrastive learning objective, proving to be a highly effective and stable approach. Additionally, CVLC [CMWP23] introduces a cross-view contrastive learning method that learns view-invariant representations and generates clustering results by contrasting cluster assignments across multiple views.

Non-Contrastive Instance Discrimination. Non-contrastive methods share a similar objective to contrastive methods, aiming to align two different transformations of the same input from student and teacher networks. However, they diverge by excluding negatives from the learning process. Instead, these methods employ various strategies to prevent collapsing, where the model predicts a constant value for any input. BYOL [GSA+20] addresses collapsing by proposing the use of the exponential moving average (EMA) for one of the backbones and employing different projection heads. Conversely, SimSiam [CH21] demonstrated that while EMA offered minimal improvement, the asymmetry of the projection head was crucial for performance. DINO [CTM+21] applies centering of the student outputs and utilizes discretization via softmax during training. The subsequent iteration, DINOv2 [ODM+23], merges DINO with a generative approach of masked image modeling in the embedding space, leveraging a larger pretraining dataset. Other methods in this category rely on correlation analysis, such as VICReg [BPL22], BarlowTwins [ZJM+21], and TWIST [WKZ+23].

2.1.2 Generative Self-Supervised Learning

Many methods in this category rely on information restoration, where a portion of the image data is deliberately distorted, and the network is tasked with recovering this information. For instance, in the task of image colorization, the input is grayscale, and the network is tasked with reconstructing the original RGB values [LMS17, ZIE16].

Presently, the most efficient methods belong to the family of masked image modeling, where a portion of the image is either removed or masked out, and the model is tasked with restoring the missing information [CRC+20, BDPW21]. This approach finds extensive application in the NLP community, where the objective is to restore masked words within sentences [DCLT19]. BEiT [BDPW21] stands as one of the pioneering frameworks that successfully employed this paradigm for training, predicting discrete visual tokens generated from visual patches of images by leveraging a discrete variational autoencoder. Subsequent versions, BEiT-v2 [PDB+] and BEiT-v3 [WBD+23], enhanced results by incorporating the CLIP tokenizer alongside a patch aggregation strategy. Another successful adaptation of the masked objective was accomplished by MAE [HCX+22], utilizing an asymmetric autoencoder framework that directly learns to reconstruct image patches. Notably, the model efficiently processes only unmasked patches (approximately 25%) in the encoder, resulting in significant computational savings. Recently, this strategy has been effectively applied across multiple modalities.

2.1.3 Video Representation Learning

The previous subsections have focused on methods tailored for image-based self-supervision. However, videos offer an additional temporal dimension that can be leveraged to formulate tasks for self-supervision. Early methods [MZH16, LHSY17a, FBGG17] concentrate on frame-level video augmentations, such as frame shuffling and odd frame insertion within an existing frame sequence. These methods train models to discriminate between these augmentations [MZH16]. Xu et al.[XXZ+19] modify this approach by augmenting videos at the clip level rather than the frame level, shuffling video clips and predicting their correct order, thereby leveraging spatial-temporal information. Han et al.[HXZ19] frame the problem as future clip prediction. Recent works [QMG+21, DGRS21] employ contrastive learning between different video augmentations. Other directions in self-supervised learning include utilizing temporal co-occurrence as a learning signal [IZKA15, WG15]. Wei et al.[WLZF18] exploit the arrow of time by detecting the direction of video playback. Another popular approach for video representation learning is learning temporal or spatial-temporal cycle-consistency as a proxy task[WJE19, JOE20]. The most recent methods adapt the concept of masked image modeling to videos from image-based self-supervised pretraining.

2.1.4 Connection to our work

In Chapter 3, we explore unsupervised temporal segmentation of instructional videos. We propose a framework that is based on the learned representations. To learn task-specific representations, we are the first to propose supervision based on the inherent order of actions within tasks. For instance, supervision based on the fact that some actions need to be performed in a certain order, for instance a subaction "take cup" will usually occur in the beginning of the activity "making coffee". We show that these learned representations bring an advantage for the task of unsupervised temporal segmentation compared to more general

representations [RTMFF15, CZ17, WS13, FGO+15].

In Chapter 4, we explore unintentional actions and propose to learn an embedding based on inherent biases as motion speed, motion direction, and temporal order of motion clues. We formulate the temporal transformations specifically to reflect the structure of unintentional activities to capture sudden changes in the video and pretrain the model in self-supervised way by recognizing the formulated transformations. Despite availability of variety of temporal self-supervised learning methods [IZKA15, WG15, WLZF18], we show that this amplification improves the unintentional action recognition by large margin with respect to previous work and the baselines.

In Chapter 5, we study contrastive learning methods and propose the framework to understand better the dynamic on the long-tail data with respect to temperature hyperparameter. This parameter controls the impact of the negatives during the training on semantic information learned by the model. In contrast to explicitly choosing a specific subset of negatives [WWW+21, YHH+22, ZZP+22, ITAC18], we discuss the Info-NCE loss [OLV18] through the lens of an average distance perspective with respect to all negatives and show that the temperature parameter can be used to implicitly control the effective number of negatives. Moreover, we focus on understanding the structure and learning dynamics of the objective function such as in [SPA+19, TWSM20, CLL21]. In this work, we rely on the previous findings, expand them to long-tailed data distributions and complement them with an understanding of the emergence of semantic structure.

2.2 Learning with Limited Supervision

Introducing weak supervision, such as partially annotated data points or weak alignment between different modalities, can significantly enhance performance on downstream tasks. In this section, we discuss topics such as semi-supervised learning and multi-modal learning, which enable the introduction of weak or partial annotations in the data.

2.2.1 Semi-Supervised Learning

Semi-supervised learning is a broad field aiming to leverage both labeled and unlabeled data, and self-supervised learning, as discussed in the previous section, has proven beneficial in this context. In [HFW+20, CKNH20, CKS+20, REH+20], self-supervised pre-training is employed to initialize models for subsequent fine-tuning. However, as these methods typically involve multiple training phases with numerous hyperparameters, such pretrainining is mainly avoided by the community. Alternatively, consistency regularization emerges as a powerful method for semi-supervised learning [RBH+15, SJT16, BAP14]. The idea is that the model should yield consistent predictions for perturbed versions of the same input. Various approaches have been explored to generate such perturbations. For example, [TV17] utilizes an exponential moving average of the trained model to produce additional input; [SJT16, LA17] employ random max-pooling and dropout [SHK+14]; [XDH+20, BCC+20, SBL+20, KMHK20] employ advanced data augmentation techniques; [BCG+19, VLK+19, BCC+20] utilize MixUp regularization [ZCDLP18], which encourages convex behavior "between" examples; [GWL21] enforce label consistency using alpha-divergence.

Another spectrum of popular approaches is pseudo-labeling [Scu65, Nes83, Lee13], where the model is trained with artificial labels. [AOA+20] trained the model with “soft” pseudo-labels from network predictions; [PXDL20] proposed a meta learning method that deploys a

teacher model to adjust the pseudo-label alongside the training of the student;[SBL+20, Lee13] learn from "hard" pseudo-labels and only retain a pseudo-label if the largest class probability is above a predefined threshold; [ZWH+21] further refines the thresholding mechanism by adaptively adjusting thresholds for different classes according to the learning effect of each class. Furthermore, there are many excellent works around generative models [KMRW14, Ode16, DGF16] and graph-based methods [LZL+18, LWHL19, BDLR06, Joa03]. As noise injection plays a crucial role in consistency regularization [XDH+20], the advanced data augmentation, especially combined with weak data augmentation, introduces stronger noise to unlabeled data and brings substantial improvements [BCC+20, SBL+20]. [SBL+20] proposes to integrate pseudo-labeling into the pipeline by computing pseudo-labels from weakly augmented images, and then uses the cross-entropy loss between the pseudo-labels and strongly augmented images. Besides the classifier level consistency, our model also introduces consistency on the feature level, which explicitly regularizes representation learning and shows improved generalization performance. PRG [DZQ+23] addresses challenges where labeled and unlabeled data distributions do not align, resulting in biased pseudo-labeling. The authors propose class-level guidance based on a Markov random walk modeled on a dynamic graph over class distributions. FreeMatch [WCH+23] analyzes the relationship between pseudo-labeling threshold and model learning status, proposing to adjust the confidence threshold adaptively to enhance performance, particularly in cases with extremely rare labeled data.

2.2.2 Learning from Multimodal Data

Image-Language Pretraining. There is a growing interest in multi-modal representations [RKH+21, LDF+20, SZC+20, LCC+20, LJS+20, SCR+22], which require aligned pairs across multiple modalities. However, acquiring human-annotated paired data is costly; hence, noisy web data [RKH+21, MZA+19] enables the significant scaling of such datasets. Many methods effectively utilize web image-text pairs [RKH+21, JYX+21, YWV+22] for large-scale pretraining. For contrastive-based vision-language representation learning methods, dual-encoder architectures are a common choice, featuring two parallel branches for each modality, contrasted against each other to learn a joint embedding space [RKH+21, SZC+20, LCC+20, LJS+20]. Recently, BLIP [LLXH22], BLIP2 [LLSH23a], and CoCa [YWV+22] have proposed a unified multi-task contrastive-generative framework, combining contrastive and captioning objectives. These methods leverage curated image-text and uncurated web image datasets, with BLIP additionally employing iterative generation and filtering of synthetic captions.

Video-Language Pretraining. Manual annotation of video datasets is even more time-consuming since it involves video trimming and localization of action boundaries. Currently, manually annotated video captioning datasets, e.g., MSR-VTT [XMYR16], and YouCook2 [ZXC18], are limited in size. Therefore, different methods of mining video with weak supervision from the Internet were considered. Datasets such as YouTube-8M [AEHKL+16] and IG-Kinetics-65M [GTM19] provided multiple class labels based on query click signals and metadata [AEHKL+16] or hashtags [GTM19]. However, short class labels are suboptimal supervision compared to textual descriptions [DJ21]. Therefore, [BNVZ21] considered scrapping videos with associated alt-text from the web, similarly to the image-based Conceptual Captions dataset [SDGS18], obtaining the WebVid2M dataset [BNVZ21] that contains 2.5M videos-text pairs and later WebVid10M with 10M pairs. [SLS+20] proposed to use meta information, such as titles, video descriptions, and tags from YouTube, as a textual annotation and created the WTS-70M dataset. And [NSS+22] proposed to transfer image captions from an image-text dataset to videos by searching videos with similar frames to the image and, therefore, collected

the VideoCC3M dataset. However, most videos in the WebVid2M and WebVid10M datasets do not have audio, which is an essential part of video analysis, and captions in the VideoCC3M dataset are derived from images and, therefore, tend to describe more static scenes rather than actions. At the same time, the title and tags of WTS-70M provide only high-level video descriptions.

As an alternative to this, [MZA+19] proposed the HowTo100M dataset, where instructional videos are naturally accompanied by dense textual supervision in the form of subtitles obtained from ASR (Automatic Speech Recognition) systems. The HowTo100M dataset with 137M clips sourced from 1.2M YouTube videos was proven to be effective for pre-training video-audio-language representations [RBH+21, CRD+21, SCR+22]. The followed-up YT-Temporal-180M [ZLH+21], HD-VILA-100M [XHZ+22] and InternVid [WHL+23] datasets are created by using the same idea, but expand the HowTo100M with more videos, higher diversity, and higher video resolution. However, the problem of misalignment and noisiness of ASR supervision in instructional videos, such as in the HowTo100M dataset, were already addressed in multiple works. MIL-NCE loss [MAS+20] and soft max-margin ranking loss [ABARB21] were proposed to adapt contrastive loss to misalignment in text-video pairs. [ZLH+21] proposed to use LLM to add punctuation and capitalization to ASR subtitles and remove mistranscription errors. [HXZ22] proposed to train temporal alignment networks to filter out subtitles that are not alignable to the video and determine alignment for the others. [LPB+22] goes beyond just removing mistranscription errors and ASR re-alignment, and proposed to match the sentences from ASR subtitles to a large base of descriptions of the steps from WikiHow dataset [KW18] (distant supervision).

Large-Language Models in Vision-Language Tasks. In recent years, there has been a remarkable success of LLMs in many language-related tasks [DCLT19, RWC+19, RSR+20]. Latest large language models such as GPT-3.5 [NXP+22], Alpaca [TGZ+23] or Vicuna [CLL+23] have demonstrated excellent zero-shot capabilities on common sense inference [CB23]. This success has prompted research into integrating common-sense knowledge into vision-language tasks to enhance their performance. In this regard, some methods [SMV+19, SZC+20, LBPL19, TB19] initialize the language part of vision-language models from pre-trained LLM. Another line of work [CLTB21, LLSH23b, ZMKG23] uses LLM as a decoder to enable vision-to-language generation. For example, the MiniGPT-4 [ZCS+23] model enhances a frozen Vicuna model by aligning visual encoder tokens with Vicuna's input token space, enabling visual reasoning capabilities, e.g., image question answering or image captioning. In this regard, some works [LRC+23, ZMKG23] adapted visually conditioned LLM for visual captioning and created captioning pseudo labels for large-scale video data that later used for vision-language tasks. However, these methods require human-annotated datasets to train a captioning model, while our method does not require any label data and aims to transform free available annotation (ASR subtitles) into textual descriptions.

2.2.3 Connection to our work

In Chapter 6, we follow the trend of [ZOKB19, BCC+20] to incorporate an auxiliary self-supervised loss alongside training optimizing a rotation prediction loss [GSK18]. Different from previous work, our work explores equivariant representations in the sense that differently augmented versions of the same image are represented by different points in the feature space despite the same semantic label. As we show that information like object location or orientation is more predictable from our model when features are pushed apart from each other.

In Chapter 7, in contrast to previous work [GVM+22, LXX+22], we propose to exclude

pre-annotated text-video paired data from the training and, relying on text descriptions only, generate text-video pairs leveraging uncurated web videos while transferring the style of original captions. Moreover, we adopt pre-trained image-language models for uncurated & unpaired text-video retrieval by transferring the caption style directly on uncurated videos without any aligned data during training.

In Chapter 8, we exploit web images to annotate video data and propose to transfer knowledge in the form of pseudo labels. Various works have shown how web images and videos can be used as labeling-free data source to improve the performance of action recognition [DZX+20, GSDG16, GYY+16, MBZ+17]. In our work, we specifically address the domain shift on the spatial level and transfer the knowledge to the spatio-temporal level, without using any annotations on target data. Compared to webly-supervised learning, image-to-video domain adaptation approaches actively address the domain shift between web images and target videos either by spatial alignment between web images and video frames [LWZK17, SSSN15, ZHT+16], or through class-agnostic domain-invariant feature learning for images and videos [LOW+20, YWCD19, YWSD18]. In contrast to these, we propose to transfer knowledge in the form of pseudo labels, without enforcing spatial information from the DA stage onto the spatio-temporal model. Moreover, we alternately conduct spatial alignment and spatio-temporal learning with knowledge transfer to each other. With category knowledge transferred to the spatial model, we perform class-aware domain alignment that induces domain-invariance within category-level.

In Chapter 9, we explore large-scale video-language pretraining. While ASR supervision can provide a scalable way to create a large video dataset with dense annotation [ZLH+21, XHZ+22, MZA+19], the quality of subtitles is still not on par with human-annotated captions. In our work, we propose to use LLM to create high-quality video captions given ASR subtitles, which allows us to create a detailed video description dataset at scale where each description is specific for every video and has proper sentence structure.

2.3 Learning with Limited Data

In many real-world scenarios, acquiring extensive datasets can be impractical or even impossible. For example, drug discovery task which tries to predict whether a new molecule has toxic effect [ATRPP17]. This limitation has led to the emergence of a few-shot learning methods. These algorithms are designed to tackle the scenarios where only a small number of examples are available for each class or task. Moreover, few-shot learning assumes that there are certain classes with sufficient labeled data, referred to as base classes. This data is typically available prior to tasks with limited data and is utilized for pretraining of the model. Therefore, the task of few-shot learning is how to learn such a model from these base classes that can generalize to novel classes with only a few labeled samples (referred to as support set) available [Xia20].

2.3.1 Few-Shot Learning

Few-shot learning relies on the transfer learning paradigm, where the fundamental operation involves pretraining the model on an extensive dataset and subsequently fine-tuning it on the limited support set. Many methods [WGHH18, SSZ17, QBL18, FAL17] employ a meta-learning framework instead of straightforward training on the base classes [Liu23]. Meta-learning framework entail training the model on multiple few-shot tasks, also known as episodes, generated from the base classes containing sufficient labeled data. Subsequently, the

performance of the model is evaluated on test episodes derived from novel classes.

Data-based Methods. To mitigate data scarcity, various methods employ data augmentation for few-shot learning to enhance model robustness [WGHH18]. Leveraging prior knowledge, these methods introduce different types of invariance during model training. To prevent overfitting, [BKVM15] propose simulating noise by erasing random pixels in images. Similarly, [DT17, YHO+19] suggest using random filling operations to introduce various types of noise. On the other hand, approaches like [ZCG+18, HG17, LZLF20] aim to learn domain-specific augmentations based on the training base classes and transfer them to novel classes. For instance, [LCL+20] transfer modes of variation from base to novel classes, while [WGHH18] propose hallucinating additional data samples for the novel classes by training an end-to-end model on the base classes in a meta-learning framework.

Metric-based Methods. Metric-based methods [VBL+16, SSZ17, QBL18, YHZS20, HCM+19] aim to learn a metric space suitable for further comparison between class representations without additional training. These methods utilize a pretrained model and compute similarity or distance between samples. For example, by calculating distances in the embedding space between a test sample and all training samples, the test sample can be assigned to the category of the nearest training sample. Prototypical Networks [SSZ17] exemplify this approach, learning a metric space in a meta-learning framework where the distance to class prototypes determines classification. The imprinting weights method [QBL18] demonstrates improved performance by utilizing class means as strong initializations for an evolving classifier. However, simple feature averaging is susceptible to noise; thus, approaches like [YJ06, WSL+20] explore methods to increase distances between prototypes to mitigate noise. Additionally, [LCL+20] propose leveraging positive and negative margins to reduce overfitting and enhance generalization. Furthermore, relational networks [SYZ+18] enable direct calculation of similarity between samples by the network itself, exploring alternative methods of distance and similarity calculation.

Optimization-based Methods. Optimization-based methods aim to adjust the optimization algorithm so that the model can converge within a small number of optimization steps and a limited number of input samples without overfitting. MAML [FAL17] is a method based on the meta-learning framework that effectively initializes the model for further training of the few-shot support set. TAML [JQ19] is an entropy-based approach that meta-learns an unbiased initial model with the largest uncertainty over the output labels by preventing it from over-performing in classification tasks. IMAML [RFKL19] is an implicit MAML algorithm, which depends only on the solution to the inner-level optimization and can efficiently handle vanishing gradients and memory constraints problems. Ravi et al.[RL16] propose an LSTM-based meta-learner to learn the exact optimization to train another network classifier in a few-shot manner. LEO[RRS+18] decouples training and representation by learning data-dependent latent model parameters and performing gradient-based optimization in a low-dimensional space. In [LMRS19], the authors target meta-learning with efficient linear classifiers relying on implicit differentiation of the optimization and the low-rank nature of the classifier in the few-shot setting. Note that MAML, one the most popular and effective optimization-based approaches, has been adopted in various tasks such as continual learning [JW19, GYP20], hate speech detection [ALT+23], and detecting security intrusions [LWY+23].

2.3.2 Few-Shot Class-Incremental Learning

Real-world applications often encounter various challenges when acquiring data incrementally, with new information arriving in continuous portions. This scenario is commonly referred to as Class Incremental Learning [RKSL17a, WCW+19, HVD15, LH17a, CMJG+18, HPL+19,

BP19, ZZW+21, ZXG+20, SCL+18, LMH+18, CDAT18]. Within CIL, the foremost challenge lies in preventing catastrophic forgetting [MC89, GMX+13, KPR+17], where previously learned concepts are susceptible to being overwritten by the latest updates. However, in a Few-Shot Class Incremental Learning (FSCIL) scenario [THC+20, ZSL+21, ZLX+23, KHSM22, ZFK+21, CRF+21, ZWY+22, YYL+23, her, MSR21, her], characterized by the introduction of new information with only a few labeled samples, two additional challenges emerge, reflected in the few-shot learning scenario: overfitting and intransigence [SSZ17, CLK+19]. Overfitting arises as the model may memorize scarce input data and lose its generalization ability. On the other hand, intransigence involves maintaining a delicate balance, preserving knowledge from abundant existing classes while remaining adaptive enough to learn new tasks from a highly limited dataset.

Data-based Methods. Data manipulation is a direct strategy to mitigate the catastrophic forgetting issue. A common approach is data replay, which involves showing the model data from previous tasks to prevent forgetting [CRF+21, DHT+21, LGC+22]. For instance, Cheraghian et al.[CRF+21] propose semantic-aware knowledge distillation by storing a small number of samples for previous classes. Liu et al.[LGC+22] suggest learning a generative model to produce additional data close to the decision boundary. Shankarampeta et al.[SY21] propose generating data in a feature space instead of the pixel space, training the generative model with the MAML method[FAL17]. FACT [ZWY+22] imposes constraints on real samples during training to encourage more compact representations in the embedding space while preserving space for virtual categories. Similarly, ALICE [PZW+22] generates virtual classes by merging two distinct classes from the base session and augmenting the data using standard geometric transformations.

Metric-based Methods. Similarly to few-shot learning, metric-based methods for few-shot class incremental learning aim to learn effective representations so that models can identify and utilize underlying patterns helpful for generalization to unseen classes through simple comparison between samples. Recently, metric learning has also been adopted in few-shot class incremental learning to improve representation learning. Mazumder et al.[MSR21] propose incorporating self-supervised learning to enhance the generalization capability of the backbone, while Peng et al.[PZW+22] suggest using the angular penalty loss, originally used in face recognition, to obtain well-clustered features. CLOM [ZZLL22] proposes a framework to address the issue where large margins can lead to good discrimination between base classes but hinder performance on novel classes. The authors design different loss functions for shallow and deep layers of the encoder network and integrate class relationships. Zhao et al.[ZFK+21] first train the feature extractor using metric learning loss and regularization loss, then decouple features based on their frequency as they found that low-frequency information may contribute more to preserving old knowledge. C-FSCIL[her] utilizes a meta-learning framework to train a feature extractor and rewritable memory. The C-FSCIL framework aligns class prototypes quasi-orthogonally to negate interference between classes, while NC-FSCIL [YYL+23] aligns class features with classifier prototypes, formed as a simplex equiangular tight frame, using dot-regression loss.

Optimization-based Methods. Similarly to few-shot learning, optimization-based few-shot class incremental methods leverage meta-learning frameworks to address overfitting and catastrophic forgetting. By creating pseudo-incremental tasks sampled from the base classes, these methods effectively train the backbone and make adjustments for upcoming new classes. C-FSCIL [her] demonstrates that meta-learning can facilitate the learning of effective representations that can be simply averaged to achieve improved performance. Zheng et al.[ZZ21] regulate the distribution of learned classes within the meta-learning framework by propos-

ing a class structure regularizer that prevents interference between new and old prototypes. CEC[ZSL+21] trains a graph attention mechanism to regularize the relations between class prototypes using a meta-learning framework.

2.3.3 Connection to our work

In Chapter 10, we introduce a three-stage optimization-based framework designed to address generalized few-shot learning and few-shot class incremental learning, effectively tackling catastrophic forgetting and calibration challenges. While some approaches focus on bias removal techniques [WCW+19, BP19, BP20], employing methods such as training additional hyperparameters [WCW+19], utilizing dual memory [BP19], or applying post-processing [BP20], our framework leverages the joint space to overcome data deficiencies and construct a stronger classifier for novel classes without introducing extra parameters. Furthermore, unlike standard optimization-based methods that rely on meta-learning frameworks [THC+20, AW20], we approach the problem from the perspective of classic parametric classification models [HZRS16, KSH12], which have proven effective in few-shot learning [TWK+20] and class incremental learning [PTD20].

In Chapter 11, we propose the OrCo metric-based framework built on two core principles: features' orthogonality in the representation space, and contrastive learning. Our method is closely related to metric-based methods [her, YYL+23]. Our approach utilizes contrastive learning with data-agnostic pseudo-targets and margin maximization through perturbations in the embedding space, thereby enhancing generalization in incremental sessions. Furthermore, we enhance the generalization of the embedding space by incorporating a combination of supervised and self-supervised contrastive losses during the pretraining phase.

2.4 Learning Beyond Supervision

Traditional methods often assume a closed-world scenario, where models are trained and tested on fixed, predefined classes or tasks. However, in many real-world applications, the environment is dynamic and constantly changing, presenting new classes, concepts, or data distributions over time. Learning in an open world addresses this challenge by enabling models to be robust with respect to the open world by identifying unknown yet aspects or continually adapt and learn from incoming information, without requiring retraining on all possible classes.

Traditional methods typically operate in a closed-world scenario, where models are trained and tested on fixed, predefined classes or tasks. However, in various real-world applications, the environment is dynamic and ever-changing, introducing new classes, concepts, or data distributions over time. Open-world learning tackles this challenge by enabling models to be robust in the open world. This involves identifying unknown aspects or continually adapting and learning from incoming information without the need for retraining on all possible data.

2.4.1 Open-Set Learning

In an open-set setting, only a limited number of classes is available during the model training phase, and the system must be capable of detecting unseen classes during evaluation [MC21, SRSB13]. Methods in this context must not only identify and classify instances belonging to the seen classes during training but also reject instances from unknown classes during testing.

By handling unknown classes during evaluation, such systems can operate robustly in an open-world scenario, where unmodeled situations can be encountered. Open-set classification is closely related to both standard classification and out-of-distribution detection [YZLL21, LLLS17, PAD18, YGX22]. Out-of-distribution detection involves identifying rare items and typically acts as a binary classifier, while open-set settings also entail classifying seen classes alongside detecting unseen items. In contrast, standard classification involves assigning input items to a fixed set of classes during evaluation, whereas open-set settings require a rejection mechanism to prevent misclassification of unseen classes as one of the seen classes. Several closed-set approaches have been adapted to include rejection mechanisms for open-set settings [BW08, FHW16, GEY17].

The scope of open-set work can be broadly divided into two categories: discriminative [PP19, CQS+20] and generative [PMJ+20] approaches. Discriminative methods rely on input data and corresponding labels for classification. For instance, one pioneering deep learning-based method for open-set recognition is OpenMax [BB16], which addresses overconfident classification scores for unseen test inputs by using logit scores instead of softmax scores. Another approach, introduced by Dhamija et al. [DGB18], involves incorporating background samples crawled from the internet during classifier training. The method enforces small feature magnitudes for these background samples, while seen samples for classification are enforced to have larger magnitudes with a defined margin. Vaze et al. [VHVZ22] explore the correlation between closed-set accuracy and open-set performance with different backbone architectures, finding a strong correlation across various loss functions and architectures, leading to a surprisingly simple yet effective baseline. Additionally, Wang et al. [WXY+22] propose a novel metric called OpenAUC for evaluating open-set and closed-set performance in a coupled manner. Optimization of OpenAUC risk has been shown to be effective in improving performance.

Generative approaches, on the other hand, leverage additional data generated by Generative Adversarial Networks (GANs) or Autoencoders (AEs) to identify outliers. G-OpenMax [GDCG17] extends the OpenMax approach by incorporating GANs to generate unseen items. This is achieved by training a conditional GAN and interpolating in the latent space. Similarly, OSRCI [NOF+18] follows the concept of G-OpenMax and generates counter-factual examples, which lie on the classification boundaries between seen and unseen classes. C2AE [OP19] and CROSR [YSK+19] argue that abnormal test-time samples should not be reconstructed as well as normal ones. They propose using AEs as the meta-recognition function, where the AE's encoder serves as the classifier. OpenGAN [KR21] suggests generating fake samples that mimic the open-set distribution by carefully selecting a GAN discriminator on real outlier data.

2.4.2 Open-World Learning

In an open-set setting, machine learning models are trained to recognize known classes while also being equipped to reject instances from unknown classes. However, in an open-world setting, models must not only recognize known classes but also adapt to dynamically changing environments by continuously learning and identifying entirely new classes or concepts that were not present during training [JDC+20]. The crucial distinction lies in the adaptability demanded in the open-world scenario, which extends beyond merely detecting unknown instances in the open-set setting.

Open-world computer vision focuses on algorithms that can learn and adapt to new and unseen classes of objects or situations. Some key approaches include novelty discovery [CKSP15, HLK17, HVZ19, QZCH21, WZW+19] and zero-shot learning [MNXA21, FWD+19, LLWJ18, AC19, XGF16, RKP18]. Novelty discovery aims to identify data points that deviate significantly

from existing patterns, potentially representing new classes [HVZ19, FSL+21]. UNO [FSL+21] utilizes a multi-view self-labeling strategy to generate pseudo-labels that can be treated homogeneously with ground-truth labels. Han et al.[HVZ19] propose to extend and improve Deep Embedded Clustering [XGF16] by introducing a representation bottleneck, temporal ensembling, consistency, and identifying the number of classes in the unlabeled data. On the other hand, zero-shot learning enables models to classify new classes without any labeled training data for those classes, relying on semantic relationships between classes [MNXA21, CLSZ20]. CompCos [MNXA21] solves compositional zero-shot learning based on cosine similarity between logits and a projection of learned primitive embeddings with an integrated feasibility estimation function. Meanwhile, Chen et al. [CLSZ20] propose a boundary-based out-of-distribution classifier to address generalized zero-shot learning, which classifies the unseen and seen domains using only seen samples for training. By leveraging the class centers and boundaries, the unseen classes can be separated from seen samples. These methods offer different strengths for handling open-world data: novelty detection helps identify potential new classes, while zero-shot learning can leverage existing knowledge to classify unseen classes, albeit often requiring strong semantic representations.

Combining vision data with other modalities such as languange can enhance open-world recognition capabilities. By aligning information from multiple sources, multimodal learning can create more robust and informative representations, allowing zero-shot recognition of novel concepts [RKH+21, LLXH22]. Many recent open world methods effectively align open-set and closed-set distributions by utilizing the alignment learned with visual-language models (VLMs) [ZYLL22b, RKM+23a]. Inspired by the NLP community [ZFC21, SRLI+20, JXAN20, GFC20, LARC21], prompt learning, where frozen general text models are adapted to the specific tasks based on the closed-set data and show prominent generalization performance in an open world scenarios, attracts more attention in recent years in computer vision. CoOp [ZYLL22b] learns appendable vectors in the text token space, and CoCoOp [ZYLL22a] introduces input-conditioned prompt learning, boosting performance but with high computational costs. MaPLe [KRM+23] leverages shared deep prompts for text and visual encoders, while PromptSRC [RKM+23a] suggests regularizing constraints for frozen encoders. Chen et al. [CGT+22] found that prompting boosts model transferability in tasks with fewer number of visual tokens, like image classification, but has limited impact in tasks with more tokens, such as video understanding. An alternative research direction explores adapting vision-language models with feature adapters [HGJ+19]. Clip-adapter [GGZ+23] learns new features through an additional bottleneck layer and blends them with original pre-trained features in a residual style showing better performance than prompt-based methods in few-shot generalization. These studies demonstrate the potential of combining open world techniques with multimodal learning for robust and adaptable recognition in open-world scenarios.

2.4.3 Connection to our work

In Chapter 12, we study the challenging and realistic open-set semi-supervised learning setting, where the goal is to both correctly classify inliers and to detect outliers. We adopt a simple confidence-based pseudo-labeling [SBL+20] for classifier training, which is an effective way of leveraging unlabeled data to improve the model performance. Compared to standard semi-supervised learning, SSB has an additional outlier detector, which enables the model to reject samples that do not belong to any of the inlier classes. Moreover, existing methods seek to alleviate the effect of out-of-distribution data by filtering them out in different ways so that the classification model is trained with inliers only [CZLG20, SKS21, HYG22]. In

contrast, we show that if the representations of the inlier classifier and the outlier detector are well-separated, OOD data turns out to be a powerful source to improve the inlier classification without degrading the detection performance. So, instead of filtering OOD data, we use a simple confidence-based pseudo-labeling to incorporate them into the training.

In Chapter 13, we explore open-world setting by adapting vision-language model to unexplored domain of egocentric videos. Prompt learning [ZYLL22b, ZYLL22a] and adapters [HSW+22, GGZ+23] are the two prominent directions for adaptation of the frozen models to the unseen image domains. Our approach falls under the adapter category. Unlike previous visual adapters, we introduce cross-modal instance-conditioned adapters specifically designed for open world generalization. Moreover, recent advancements in prompt learning extend to third-person videos [JHZ+22, LGZ+22, was23, CSMY24]. Our method builds on existing work while introducing an adapter architecture specifically tailored to egocentric domain, resulting in superior performance.

I

Learning Without Supervision

In the first part of this thesis, our attention is directed towards learning without supervision, encompassing both self-supervised and unsupervised methods. Our primary focus lies in learning powerful and robust representations for the downstream tasks for image and video recognition.

In Chapter 3, we explore unsupervised temporal action segmentation of instructional videos. We learn the representation space based on the fact that some actions need to be performed in a certain order with the further decoding into action segments.

In Chapter 4, we explore unintentional actions and inherent biases that they present such as motion speed, motion direction, and order of activities. We propose a framework to capture local and global temporal relations between intentional and unintentional actions.

In Chapter 5, we develop a framework for understanding the underlying dynamics of a popular self-supervised contrastive learning loss through the lens of average distance maximisation with respect to the temperature parameter. We study the effect of this objective function to the long-tail disitribution of the data.

3 Unsupervised Learning of Action Classes with Continuous Temporal Embedding

Contents

The intricate interplay between learning representations and deriving task-specific solutions mirrors the fundamental principles of unsupervised learning, where the system autonomously identifies underlying patterns and structures from raw data, facilitating a deeper understanding of the domain without explicit guidance or labeled examples. The task of temporally detecting and segmenting activities in untrimmed videos is a challenging task in the area of action recognition. One problem in this context arises from the need to define and label action boundaries to create annotations for training which is very time and cost intensive. To address this issue, we propose an unsupervised approach for learning action classes from untrimmed video sequences. To this end, we use a continuous temporal embedding of framewise features to benefit from the sequential nature of activities. Based on the latent space created by the embedding, we identify clusters of temporal segments across all videos that correspond to semantically meaningful action classes. The approach is evaluated on three challenging datasets, namely the Breakfast dataset, YouTube Instructions, and the 50Salads dataset. While previous works assumed that the videos contain the same high level activity, we furthermore show that the proposed approach can also be applied to a more general setting where the content of the videos is unknown.

This chapter is based on [KKSG19]. As the co-first author Anna Kukleva contributed main ideas for the temporal embedding, conducted all experiments, made the figures for the paper and contributed to writing. This work has been cited more than 100 times and used as a foundation for developing methods [WCL+22, LT21, VSK+21].

3.1 Introduction

The task of action recognition has seen tremendous success over the last years. So far, high-performing approaches require full supervision for training. But acquiring frame-level annotations of actions in untrimmed videos is very expensive and impractical for very large datasets. Recent works, therefore, explore alternative ways of training action recognition approaches without having full frame annotations at training time. Most of those concepts, which are referred to as weakly supervised learning, rely on ordered action sequences which are given for each video in the training set.

Acquiring ordered action lists, however, can also be very time consuming and it assumes that it is already known what actions are present in the videos before starting the annotation process. For some applications like indexing large video datasets or human behavior analysis in neuroscience or medicine, it is often unclear what action should be annotated. It is therefore important to discover actions in large video datasets before deciding which actions are relevant or not. Recent works [SY18, ABA+16] therefore proposed the task of unsupervised learning of actions in long, untrimmed video sequences. Here, only the videos themselves are used and the goal is to identify clusters of temporal segments across all videos that correspond to semantic meaningful action classes.

In this work we propose a new method for unsupervised learning of actions from long video sequences, which is based on the following contributions. The first contribution is the learning of a continuous temporal embedding of frame-based features. The embedding exploits the fact that some actions need to be performed in a certain order and we use a network to learn an embedding of frame-based features with respect to their relative time in the video. As the second contribution, we propose a decoding of the videos into coherent action segments based on an ordered clustering of the embedded frame-wise video features. To this end, we first compute the order of the clusters with respect to their timestamp. Then a Viterbi decoding approach is used such as in [RKIG18, KZN17, RKG17, LLK07] which maintains an estimate of the most likely activity state given the predefined order.

We evaluate our approach on the Breakfast [KAS14] and YouTube Instructions datasets [ABA+16], following the evaluation protocols used in [SY18, ABA+16]. We also conduct experiments on the 50Salads dataset [SM13] where the videos are longer and contain more action classes. Our approach outperforms the state-of-the-art in unsupervised learning of action classes from untrimmed videos by a large margin. The evaluation protocol used in previous works, however, divides the datasets into distinct clusters of videos using the ground-truth activity label of each video, i.e., unsupervised learning and evaluation are performed only on videos, which contain the same high level activity. This simplifies the problem since in this case most of the actions occur in all videos.

As a third contribution, we therefore propose an extension of our approach that allows to go beyond the scenario of processing only videos from known activity classes, i.e., we discover semantic action classes from all videos of each dataset at once, in a completely unsupervised way without any knowledge of the related activity. To this end, we learn a continuous temporal embedding for all videos and use the embedding to build a representation for each untrimmed video. After clustering the videos, we identify consistent video segments for all videos within a cluster. In our experiments, we show that the proposed approach not only outperforms the state-of-the-art using the simplified protocol, but it is also capable to learn actions in a completely unsupervised way. Code is available on-line.[1]

[1] https://github.com/annusha/unsup_temp_embed

The contributions of this work are as follows:

- A framework to learn of a continuous temporal embedding of frame-based features and the following decoding of the videos into coherent action segments based on an ordered clustering of the embedded frame-wise video features. The embedding exploits the fact that some actions need to be performed in a certain order and we use a network to learn an embedding of frame-based features with respect to their relative time in the video;

- Extended unsupervised setting for action segmentation to go beyond the scenario of processing only videos from known activity classes by discovering semantic action classes from all videos at once without any knowledge of the related activity;

- An extensive study to evaluate the performance of the proposed framework on three datasets, outperforming previous state-of-the-art methods. Furthermore, we perform a thorough analysis to evaluate the importance of each component.

3.2 Related Work

In this section, we discuss prior work on the temporal action segmentation of instructional videos. We will not revisit the topic of self-supervised learning on video data as previously discussed in Chapter 2.

Temporal action segmentation assumes precise boundaries for the start and end of each activity within the video. Acquiring such dense annotations is costly, resulting in the exploration of various methods to relax the necessity for dense annotations. Fully supervised training methods rely on dense annotations for these videos [FG19, LFV+17, MFD19]. In semi-supervised settings, dense annotations are provided for only a small subset of videos, significantly reducing annotation costs [SRY22, DY22]. To further diminish annotation expenses, only a subset of frames can be annotated, known as time-stamp annotations [LT21, RSTY22]. Another form of weak supervision comes from video transcripts [HFFN16, KRG17, RKG17, DX17, RKIG18], which provide the sequence of actions but are not temporally aligned with the videos, or from video tags [WXLVG17, RKG18].

Efforts have also been made towards unsupervised learning of action classes [SY18, SMS+21, DWZW22, LKP+23, WCL+22]. One of the pioneering works addressing the challenge of human motion segmentation without training data was proposed by Guerra-Filho and Aloimonos [GFA07]. They suggest a basic segmentation followed by clustering based on sensory-motor data. Utilizing these representations, they propose the application of a parallel synchronous grammar system to learn atomic action representations akin to words in language. Another contribution in this domain is by Fox et al. [FHSJ14], where a Bayesian nonparametric approach aids in jointly modeling multiple related time series without additional supervision, with application to motion capture data. In more recent work, Sener et al. [SY18] propose an iterative approach for learning action classes in an unsupervised way. Their method alternates between discriminative learning of the appearance of sub-activities from visual features and generative modeling of the temporal structure of sub-activities using a Generalized Mallows Model. Lin et al. [LKP+23] propose within-video and cross-video clustering to achieve better alignment between segments and the videos. Du et al. [DWZW22] propose detecting boundaries of actions by estimating similarities across smoothed frames, followed by a non-maximum suppression algorithm to identify candidate boundaries, and subsequent clustering for refinement.

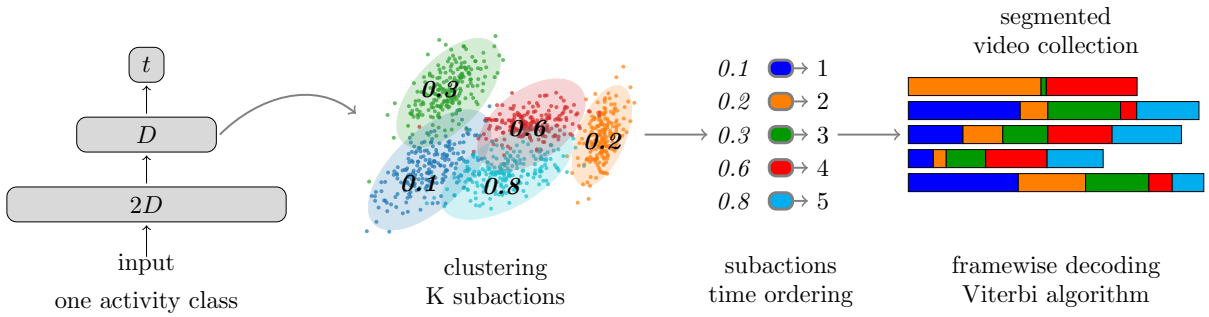


Figure 3.1: **Overview of the proposed pipeline.** We first compute the embedding of features with respect to their relative time stamp. The resulting embedded features are then clustered and the mean temporal appearance of each cluster is computed and the ordering of clusters is computed. Each video is then decoded with respect to this ordering given the overall proximity of each frame to each cluster.

3.3 Method

We begin with an overview in Section 3.3.1, followed by a step-by-step introduction of our pipeline, which consists of learning continuous temporal embedding in Section 3.3.2, clustering and ordering in Section 3.3.3, and frame labeling in Section 3.3.4. In Section 3.3.5, we discuss an extension of our method to the fully unsupervised case, and in Section 3.3.6, we discuss how to model the background class.

3.3.1 Overview

As input we are given a set $\{\boldsymbol{X_m}\}_{m=1}^{M}$ of M videos and each video $\boldsymbol{X_m} = \{x_{mn}\}_{n=1}^{N_m}$ is represented by N_m framewise features. The task is then to estimate the subaction label $l_{mn} \in \{1, \ldots, K\}$ for each video frame x_{mn}. Following the protocol of [ABA+16, SY18], we define the number of possible subactions K separately for each activity as the maximum number of possible subactions as they occur in the ground-truth.

Figure 3.1 provides an overview of our approach for unsupervised learning of actions from long video sequences. First, we learn an embedding of all features with respect to their relative time stamp as described in Section 3.3.2. The resulting embedded features are then clustered and the mean temporal occurrence of each cluster is computed. This step, as well as the temporal ordering of the clusters is described in Section 3.3.3. Each video is then decoded with respect to this ordering given the overall proximity of each frame to each cluster as described in Section 3.3.4.

We also present an extension to a more general protocol, where the videos have a higher diversity. Instead of assuming as in [ABA+16, SY18] that the videos contain the same high-level activity, we discuss the completely unsupervised case in Section 3.3.5. We finally introduce a background model to address background segments in Section 3.3.6.

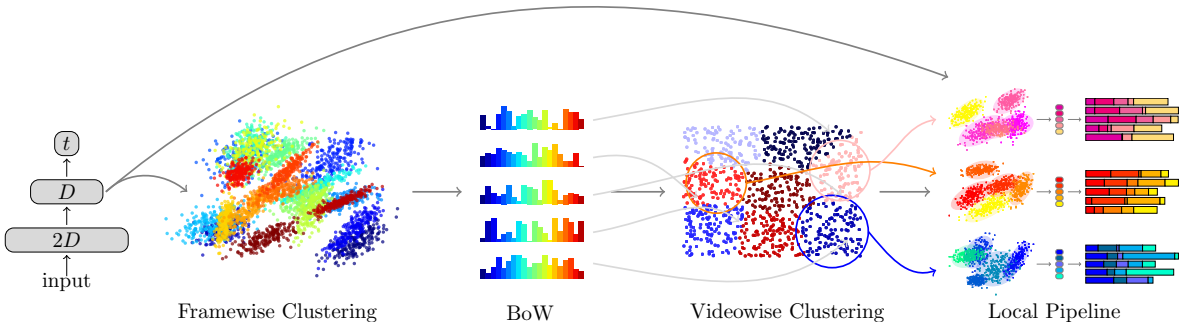


Figure 3.2: **Proposed pipeline for unsupervised learning with unknown activity classes.** We first compute an embedding with respect to the whole dataset at once. In a second step, features are clustered in the embedding space to build a bag-of-words representation for each video. We then cluster all videowise vectors into K' clusters and apply the previously described method for each video set.

3.3.2 Continuous Temporal Embedding

The idea of learning a continuous temporal embedding relies on the assumption that similar subactions tend to appear at a similar temporal range within a complex activity. For instance a subaction like "take cup" will usually occur in the beginning of the activity "making coffee". After that people probably pour coffee into the mug and finally stir coffee. Thus many subactions that are executed to conduct a specific activity are softly bound to their temporal position within the video.

To capture the combination of visual appearance and temporal consistency, we model a continuous latent space by capturing simultaneously relative time dependencies and the visual representation of the frames. For the embedding, we train a network architecture which optimizes the embedding of all framewise features of an activity with respect to their relative time $t(x_{mn}) = \frac{n}{N_m}$. As shown in Figure 3.1, we take an MLP with two hidden layers with dimensionality $2D$ and D, respectively, and logistic activation functions. As loss, we use the mean squared error between the predicted time stamp and the true time stamp $t(x_{mn})$ of the feature. The embedding is then given by the second hidden layer.

Note that this embedding does not use any subaction label associations as in [SY18, ABA+16], thus the network needs to be trained only once instead of retraining the model at each iteration. For the rest of the chapter, x_{mn} denotes the embedded D-dimensional features.

3.3.3 Clustering and Ordering

After the embedding, the features of all videos are clustered into K clusters by k-Means. Since in Section 3.3.4 we need the probability $p(x_{mn}|k)$, i.e., the probability that the embedded feature x_{mn} belongs to cluster k, we estimate a D-dimensional Gaussian distribution for each cluster:

$$p(x_{mn}|k) = \mathcal{N}(x_{mn}; \mu_k, \Sigma_k). \tag{3.1}$$

Note that this clustering does not define any specific ordering. To order clusters with respect to their temporal occurrence, we compute the mean over time stamps of all frames belonging

to each cluster

$$X(k) = \{x_{mn} | p(x_{mn}|k) \geq p(x_{mn}|k'), \forall k' \neq k\},$$
$$t(k) = \frac{1}{|X(k)|} \sum_{x_{mn} \in X(k)} t(x_{mn}). \tag{3.2}$$

The clusters are then ordered with respect to the time stamp so that $\{k_1, .., k_K\}$ is the set of ordered cluster labels subject to $0 \leq t(k_1) \leq .. \leq t(k_K) \leq 1$. The resulting ordering is then used for the decoding of each video.

3.3.4 Frame Labeling

We finally temporally segment each video $\boldsymbol{X_m}$ separately, i.e., we assign each frame x_{mn} to one of the ordered clusters $l_{mn} \in \{k_1, \ldots, k_K\}$. We first calculate the probability of each frame that it belongs to cluster k as defined by (3.1). Based on the cluster probabilities for the given video, we want to maximize the probability of the sequence following the order of the clusters $k_1 \rightarrow .. \rightarrow k_K$ to get consistent assignments for each frame of the video:

$$\begin{aligned}\hat{l}_1^{N_m} &= _{l_1,..,l_{N_m}} p\left(x_1^{N_m} | l_1^{N_m}\right) \\ &= _{l_1,..,l_{N_m}} \prod_{n=1}^{N_m} p\left(x_{mn}|l_n\right) \cdot p\left(l_n|l_{n-1}\right),\end{aligned}$$

where $p(x_{mn}|l_n = k)$ is the probability that x_{mn} belongs to the cluster k, and $p(l_n|l_{n-1})$ are the transition probabilities of moving from the label l_{n-1} at frame $n-1$ to the next label l_n at frame n,

$$p(l_n|l_{n-1}) = \mathbb{1}_{0 \leq l_n - l_{n-1} \leq 1}. \tag{3.3}$$

This means that we allow either a transition to the next cluster in the ordered cluster list or we keep the cluster assignment of the previous frame. Note that (3.3) can be solved efficiently using a Viterbi algorithm.

3.3.5 Unknown Activity Classes

So far we discussed the case of applying unsupervised learning to a set of videos that all belong to the same activity. When moving to a larger set of videos without any knowledge of the activity class, the assumption of sharing the same subactions within the collection cannot be applied anymore. As it is illustrated in Figure 3.2, we therefore cluster the videos first into more consistent video subsets.

Similar to the previous setting, we learn a D-dimensional embedding of the features but the embedding is not restricted to a subset of the training data, but it is computed for the whole dataset at once. Afterward, the embedded features are clustered in this space to build a video representation based on bag-of-words using quantization with a soft assignment. In this way, we obtain a single bag-of-words feature vector per video sequence. Using this representation, we cluster the videos into K' video sets. For each video set, we then separately infer clusters for subactions and assign them to each video frame as in Figure 3.1. However, we do not learn an embedding for each video set but use the embedding learned on the entire dataset for each video set. The impact of K and K' will be evaluated in the experimental section.

3.3.6 Background Model

As subactions are not always executed continuously and without interruption, we also address the problem of modeling a background class. In order to decide if a frame should be assigned to one of the K clusters or the background, we introduce a parameter τ which defines the percentage of features that should be assigned to the background. To this end, we keep only $1-\tau$ percent of the points within each cluster that are closest to the cluster center and add the other features to the background class. For the labeling described in c Section 3.3.4, we remove all frames that have been already assigned to the background before estimating $l_{mn} \in \{k_1, \ldots, k_K\}$ (3.3), i.e., the background frames are first labeled and the remaining frames are then assigned to the ordered clusters $\{k_1, \ldots, k_K\}$.

3.4 Experiments

3.4.1 Dataset

We evaluate the proposed approach on three challenging datasets: Breakfast [KAS14], YouTube Instructional [ABA+16], and 50Salads [SM13].

The Breakfast dataset is a large-scale dataset that comprises ten different complex activities of performing common kitchen activities with approximately eight subactions per activity class. The duration of the videos varies significantly, e.g. *coffee* has an average duration of 30 seconds while cooking *pancake* takes roughly 5 minutes. Also in regards to the subactivity ordering, there are considerable variations. For evaluation, we use reduced Fisher Vector features as proposed by [KGS16] and used in [SY18] and we follow the protocol of [SY18], if not mentioned otherwise.

The YouTube Instructions dataset contains 150 videos from YouTube with an average length of about two minutes per video. There are five primary tasks: *making coffee, changing car tire, cpr, jumping car, potting a plant*. The main difference with respect to the Breakfast dataset is the presence of a background class. The fraction of background within different tasks varies from 46% to 83%. We use the original precomputed features provided by [ABA+16] and used by [SY18].

The 50Salads dataset contains 4.5 hours of different people performing a single complex activity, making mixed salad. Compared to the other datasets, the videos are much longer with an average video length of 10k frames. We perform evaluation on two different action granularity levels proposed by the authors: mid-level with 17 subaction classes and eval-level with 9 subaction classes.

3.4.2 Evaluation Metrics

Since the output of the model consists of temporal subaction bounds without any particular correspondences to ground-truth labels, we need a one-to-one mapping between $\{k_1, .., k_K\}$ and the K ground-truth labels to evaluate and compare the method. Following [SY18] and [ABA+16], we use the Hungarian algorithm to get a one-to-one matching and report accuracy as the mean over frames (MoF) for the Breakfast and 50Salads datasets. Note that especially MoF is not always suitable for imbalanced datasets. We therefore also report the Jaccard index as intersection over union (IoU) as an additional measurement. For the YouTube Instruction dataset, we also report the F1-score since it is used in previous works. Precision and recall are

Comp. of temporal embedding strategies	
ImageNet [KSH12] + proposed	21.2%
I3D [CZ17] + proposed	25.1%
dense trajectories [WS13] + proposed	31.6%
video vector [RTMFF15] + proposed	30.1%
video darwin [FGO+15] + proposed	36.6%
ours	41.8%

Table 3.1: **Evaluation of the influence of the temporal embedding.** Results are reported as MoF accuracy on the Breakfast dataset.

computed by evaluating if the time interval of a segmentation falls inside the corresponding ground-truth interval. To check if a segmentation matches a time interval, we randomly draw 15 frames of the segments. The detection is considered correct if at least half of the frames match the respective class, and incorrect otherwise. Precision and recall are computed for all videos and F1 score is computed as the harmonic mean of precision and recall.

3.4.3 Continuous Temporal Embedding

In the following, we first evaluate our approach for the case of know activity classes to compare with [SY18] and [ABA+16] and consider the case of completely unsupervised learning in Section 3.4.7. First, we analyze the impact of the proposed temporal embedding by comparing the proposed method to other embedding strategies as well as to different feature types without embedding on the Breakfast dataset. As features we consider AlexNet fc6 features, pre-trained on ImageNet as used in [RTMFF15], I3D features [CZ17] based on the RGB and flow pipeline, and pre-computed dense trajectories [WS13]. We further compare with previous works with a focus on learning the temporal embedding [RTMFF15, FGO+15]. We trained these models following the settings of each paper and construct the latent space, which is used to substitute ours.

As can be seen in Table 3.1, the results with the continuous temporal embedding are clearly outperforming all the above-mentioned approaches with and without temporal embedding. We also used OPN [LHSY17b] to learn an embedding, which is then used in our approach. However, we observed that for long videos nearly all frames where assigned to a single cluster. When we exclude the long videos with degenerated results, the MoF was lower compared to our approach.

3.4.4 Mallow vs. Viterbi

We compare our approach, which uses Viterbi decoding, with the Mallow model decoding that has been proposed in [SY18]. The authors propose a rankloss embedding over all video frames from the same activity with respect to a pseudo ground-truth subaction annotation. The embedded frames of the whole activity set are then clustered and the likelihood for each frame and for each cluster is computed. For the decoding, the authors build a histogram of features with respect to their clusters with a hard assignment and set the length of each action with respect to the overall amount of features per bin. After that, they apply a Mallow model to sample different orderings for each video with respect to the sampled distribution. The

Mallow vs. Viterbi	
	Acc. (MoF)
Mallow+multi only	29.5%
Mallow-Viterbi	34.8%
Viterbi only	41.8%

Table 3.2: **Comparison of the Mallow model and Viterbi decoding.** Results are reported as MoF accuracy on the Breakfast dataset.

Comparison with rankloss and Mallow model		
	Rankloss emb.	Temp. emb.
Mallow model (MoF)	34.6%	29.5%
Viterbi dec. (MoF)	27.1%	41.8%

Table 3.3: **Influence of temporal embedding and Viterbi decoding together.** Comparison of proposed embedding and Viterbi decoding with respect to the previously proposed Mallow model [SY18]. Results are reported as MoF accuracy on the Breakfast dataset.

resulting model is a combination of Mallow model sampling and action length estimation based on the frame distribution.

For the first experiment, we evaluated the impact of the different decoding strategies with respect to the proposed embedding. In Table 3.2 we compare the results of decoding with the Mallow model only, Viterbi only, and a combination of Mallow-Viterbi decoding. For the combination, we first sample the Mallow ordering as described by [SY18] leading to an alternative ordering. We then apply a Viterbi decoding to the new as well as to the original ordering and choose the sequence with the higher probability. It shows that the original combination of Mallow model and multinomial distribution sampling performs worst on the temporal embedding. Also, the combination of Viterbi and Mallow model can not outperform the Viterbi decoding alone. To have a closer look, we visualize the observation probabilities as well as the resulting decoding path over time for two videos in Figure 3.3. It shows that the decoding, which is always given the full sequence of subactions, is able to marginalize subactions that do not occur in the video by just assigning only very few frames to those ones and the majority of frames to the clusters that occur in the video. This means that the effect of marginalization allows to discard subactions that do not occur. Overall, it turns out that this strategy of marginalization usually performs better than re-ordering the resulting subaction sequence as done by the Mallow model. To further compare the proposed setup to [SY18], we additionally compare the impact of different decoding strategies, Mallow model and Viterbi, with respect to the two embeddings, rankloss [SY18] and continuous temporal embedding, in Section 3.4.4. It shows that the rankloss embedding works well in combination with the multinomial Mallow model, but fails when combined with Viterbi decoding because of the missing temporal prior in this case, whereas the Mallow model is not able to decode sequences in the continuous temporal embedding space. This shows the necessity of a suitable combination of both, the embedding and the decoding strategy.

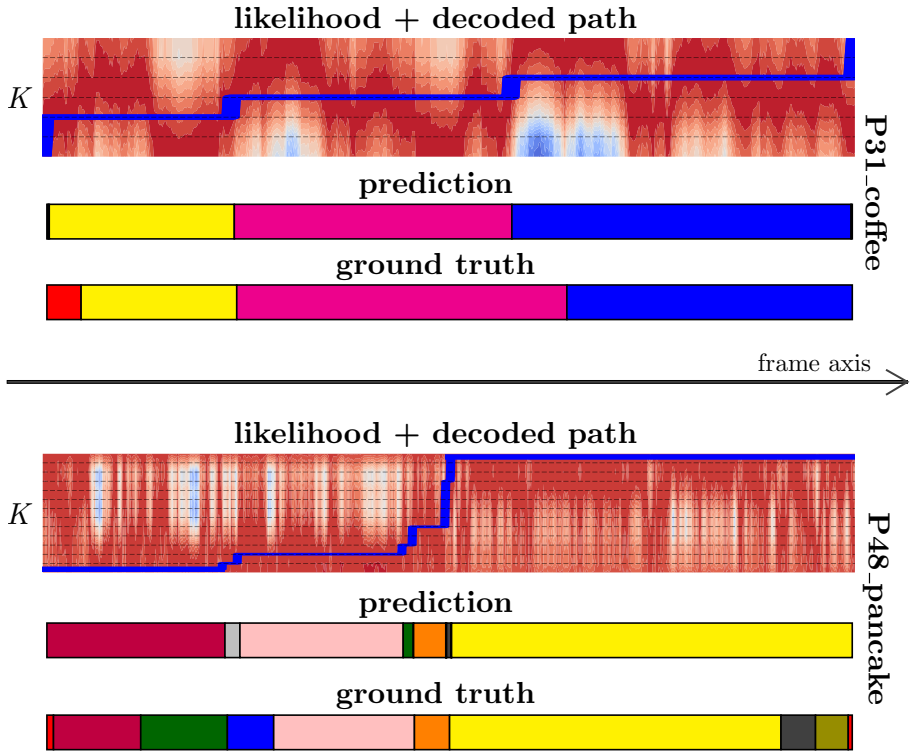


Figure 3.3: **Visualization of Viterbi decoding.** Comparison of Viterbi decoded paths with respective predicted and ground-truth segmentation for two videos. The observation probabilities with red indicating high and blue indicating low probabilities of belonging to subactions. It shows that the decoding assigns most frames to occurring subactions while marginalizing actions that do not occur in the sequence by assigning only a few frames.

3.4.5 Background Model

Finally, we assess the impact of the proposed background model for the given setting. For this evaluation, we choose the YouTube Instructions dataset. Note that for this dataset, two different evaluation protocols have been proposed so far. [ABA+16] evaluates results on the YTI dataset usually without any background frames, which means that during evaluation, only frames with a class label are considered and all background frames are ignored. Note that in this case it is not penalized if estimated subactions become very long and cover the background. Including background for a dataset with a high background portion, however, leads to the problem that a high MoF accuracy is achieved by labeling most frames as background. We therefore introduce for this evaluation the Jaccard index as intersection over union (IoU) as additional measurement, which is also common in comparable weak learning scenarios [RKG17]. For the following evaluation, we vary the ratio of desired background frames as described in Section 3.3.6 from 75% to 99% and show the results in Figure 3.4. As can be seen, a smaller background ratio leads to better results when computing MoF without background, whereas a higher background ratio leads to better results when the background is considered in the evaluation. When we compare it to the IoU with and without background, it shows that the IoU without background suffers from the same problems as the MoF in this case, but that the IoU with background gives a good measure considering the trade-off between background and class

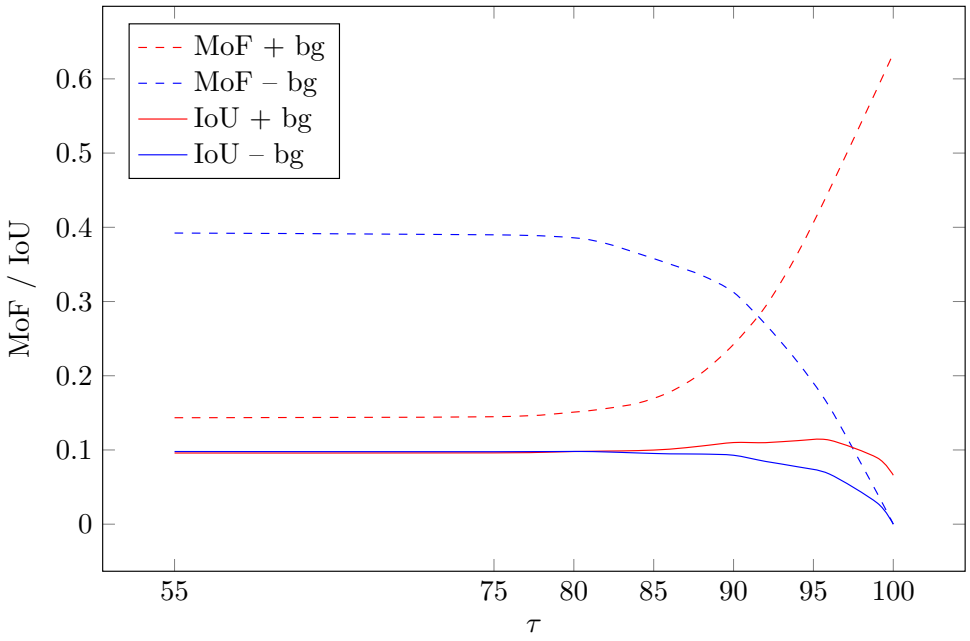


Figure 3.4: **Visualization of background influence on the performance.** Evaluation of different accuracy measurements with respect to the amount of sampled background on the YouTube Instructions dataset.

labels. For τ of 75%, our approach achieves 9.6% and 9.8% IoU with and without background, respectively, and 14.5% and 39.0% MoF with and without background, respectively.

3.4.6 Comparison to State-of-the-art

We further compare the proposed approach to current state-of-the-art approaches, considering unsupervised learning setups as well as weakly and fully supervised approaches for both datasets. However, even though evaluation metrics are directly comparable to weakly and fully supervised approaches, one needs to consider that the results of the unsupervised learning are reported with respect to an optimal assignment of clusters to ground-truth classes and therefore report the best possible scenario for this task.

We compare our approach to recent works on the Breakfast dataset in Section 3.4.6. As already discussed in Section 3.4.4, our approach outperforms the current state-of-the-art for unsupervised learning on this dataset by 7.2%. But it also shows that the resulting segmentation is comparable to the results gained by the best weakly supervised system so far [RKIG18] and outperforms all other recent works in this field. In the case of YouTube Instructions, we compare to the approaches of [ABA+16] and [SY18] for the case of unsupervised learning only. Note that we follow their protocol and report the accuracy of our system without considering background frames. Here, our approach again outperforms both recent methods with respect to Mof as well as F1-score. A qualitative example of the segmentation on both datasets is given in Figure 3.5. Although we cannot compare with other unsupervised methods on the 50Salads dataset, we compare our approach with the state-of-the-art for weakly and fully supervised learning in Section 3.4.7. Each video in this dataset has a different order of subactions and includes many repetitions of the subactions. This makes unsupervised learning very difficult

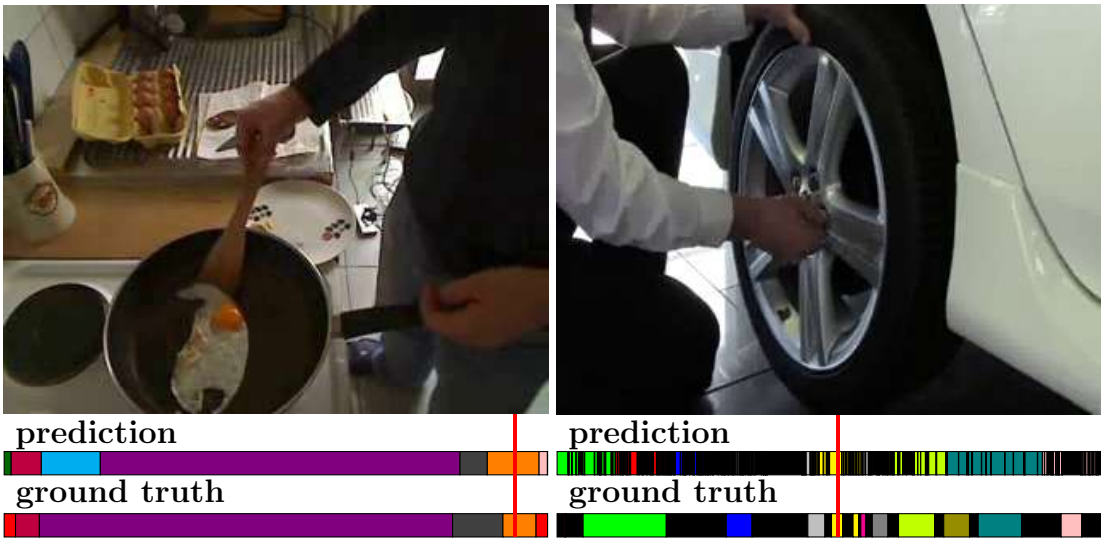


Figure 3.5: **Qualitative results.** Segmentation results on the Breakfast and the YouTube Instructions dataset.

Breakfast dataset		
	Fully supervised	
		MoF
HOGHOF+HTK		28.8%
TCFPN		52.0%
HTK+DTF w. CA		56.3%
RNN+HMM		60.6%
	Weakly supervised	
		MoF
ECTC		27.7%
GMM+CNN		28.2%
RNN-FC		33.3%
TCFPN		38.4%
NN-Vit.		43.0%
	Unsupervised	
	F1-score	MoF
Mallow	–	34.6%
Ours	26.4%	41.8%

Table 3.4: **Sota on Breakfast dataset.** Comparison of proposed method to other state-of-the-art approaches for fully, weakly and unsupervised learning on the Breakfast dataset.

YouTube Instructions		
	Unsupervised	
	F1-score	MoF
Frank-Wolfe	24.4%	–
Mallow	27.0%	27.8%
Ours	28.3%	39.0%

Table 3.5: **Sota on YouTube Instructions dataset.** Comparison of the proposed method to other state-of-the-art approaches for unsupervised learning on the YouTube Instructions dataset. We report results for a background ratio τ of 75%. Results of F1-score and MoF are reported without background frames as in [ABA+16, SY18].

50Salads		
Supervision	Granularity level	MoF
Fully supervised	eval	88.5%
Unsupervised (*Ours*)	eval	35.5%
Fully supervised	mid	67.5%
Weakly supervised	mid	49.4%
Unsupervised (*Ours*)	mid	30.2%

Table 3.6: **Sota on 50Salads dataset.** Comparison of proposed method to other state-of-the-art approaches for fully, weakly and unsupervised learning on the 50Salads dataset.

compared to weakly or fully supervised learning. Nevertheless, 30.2% and 35.5% MoF accuracy are still competitive results for an unsupervised method.

3.4.7 Unknown Activity Classes

Finally, we assess the performance of our approach with respect to a complete unsupervised setting as described in Section 3.3.5. Thus, no activity classes are given and all videos are processed together. For the evaluation, we again perform matching by the Hungarian method and match all subactions independent of their video cluster to all possible action labels. In the following, we conduct all experiments on the Breakfast dataset and report MoF accuracy unless stated otherwise. We assume in case of Breakfast $K' = 10$ activity clusters with $K = 5$ subactions per cluster, we then match 50 different subaction clusters to 48 ground-truth subaction classes, whereas the frames of the leftover clusters are set to background. For the evaluation of the activity clusters, we perform the Hungarian matching on activity level as described earlier.
Activity-level clustering. We first evaluate the correctness of the resulting activity clusters with respect to the proposed bag-of-words clustering. We therefore evaluate the accuracy of the completely unsupervised pipeline with and without bag-of-words clustering, as well as the case of hard and soft assignment. As can be seen in Section 3.4.7, omitting the quantization step significantly reduces the overall accuracy of the video-based clustering.
Influence of additional embedding. We also evaluate the impact of learning only one embedding for the entire dataset as in Figure 3.2 or learning additional embeddings for each video set. The results in Section 3.4.7 show that a single embedding learned on the entire dataset

Accuracy of activity clustering	
	mean over videos
no BoW	19.3%
BoW hard ass.	29.8%
BoW soft ass.	31.8%

Table 3.7: **Influence of clustering.** Evaluation of activity based clustering on Breakfast with $K' = 10$ activity clusters.

Multiple embeddings	
	MoF
full w add. cluster emb.	16.4%
full w/o add. cluster emb.	18.3%

Table 3.8: **Influence of embedding.** Evaluation of the impact of learning additional embeddings for each video cluster on the Breakfast dataset.

achieves 18.3% MoF accuracy. If we learn additional embeddings for each of the K' video clusters, the accuracy even slightly drops. For completeness, we also compare our approach to a very simple baseline, which uses k-Means clustering with 50 clusters using the video features without any embedding. This baseline achieves only 6.1% MoF accuracy. This shows that a single embedding learned on the entire dataset performs best. **Influence of cluster size.** For all previous evaluations, we approximated the cluster sizes based on the ground-truth number of classes. We therefore evaluate how the overall ratio of activity and subaction clusters influences the overall performance. To this end, we fix the overall number of final subaction clusters to 50 to allow mapping to the 48 ground-truth subaction classes and vary the ratio of activity (K') to subaction (K) clusters. Section 5.4.3 shows the influence of the various cluster sizes. It shows that omitting the activity clustering ($K' = 1$), leads to significantly worse results. Depending on the measure, good results are achieved for $K' = 5$ and $K' = 10$. **Unsupervised learning on YouTube Instructions.** Finally, we evaluate the accuracy for the completely unsupervised learning setting on the YouTube Instructions dataset in Section 3.4.7. We use $K = 9$ and $K' = 5$ and follow the protocol described in Section 3.4.5, i.e., we report the accuracy with respect

Influence of cluster size			
K' / K	mean over videos	MoF	IoU
1 / 50	10.9%	10.7%	4.0%
2 / 25	19.9%	15.3%	5.6%
3 / 16	25.6%	16.2%	6.1%
5 / 10	30.6%	18.8%	7.1%
10 / 5	31.8%	18.3%	13.2%

Table 3.9: **Influence of number of clusters.** Evaluation of the number of activity clusters (K') with respect to the number of subaction clusters (K) on the Breakfast dataset. The second column (mean over videos) reports the accuracy of the activity clusters (K') as in Section 3.4.7.

Influence of background ratio τ				
	MoF		IoU	
τ	wo bg.	w bg.	wo bg.	w bg.
60	19.8%	8.0%	4.9%	4.9%
70	19.6%	9.0%	4.9%	4.8%
75	19.4%	10.1%	4.8%	4.8%
80	18.9%	12.0%	4.8%	4.9%
90	15.6%	22.7%	4.3%	4.7%
99	2.5%	58.6%	1.5%	2.7%

Table 3.10: **Influence of the background.** Evaluation of τ reported as MoF and IoU without and with background on the YouTube Instructions dataset.

to the parameter τ as MoF and IoU with and without background frames. As we already observed in Figure 3.4, IoU with background frames is the only reliable measure in this case since the other measures are optimized by declaring all or none of the frames as background. Overall we observe a good trade-off between background and class labels for $\tau = 75\%$.

3.5 Conclusion

In this chapter, we propose a new method for unsupervised learning of actions in sequential video data. Given the idea that actions are not performed in an arbitrary order and thus bound to their temporal location in a sequence, we propose to learn a continuous temporal embedding to enforce clusters at similar temporal stages. These representations are tailored to the procedural videos where activities follow a specific order. We combine the temporal embedding with a frame-to-cluster assignment based on Viterbi decoding which outperforms all other approaches in the field. Additionally, we introduce the task of unsupervised learning without any given activity classes, which is not addressed by any other method in the field so far. We show that the proposed approach also works on this less restricted, but more realistic task. In the following Chapter 4 and Chapter 5, we will elaborate on the methods utilized to acquire different representations, each tailored to optimize distinct downstream tasks.

4 Leveraging Self-Supervised Training for Unintentional Action Recognition

Contents

This chapter addresses the challenge of learning effective representations for unintentional actions—instances where the intended activity is interrupted by accidents, like spilling a cup of coffee while carrying it to the table. Such rare occurrences are difficult to define precisely and highly dependent on the temporal context of the action. In this work, we explore such actions and seek to identify the points in videos where the actions transition from intentional to unintentional. We propose a multi-stage framework that exploits inherent biases such as motion speed, motion direction, and order to recognize unintentional actions. To enhance representations via self-supervised training for the task of unintentional action recognition we propose temporal transformations, called **T**emporal **T**ransformations of **I**nherent **B**iases of **U**nintentional **A**ctions (T^2IBUA). The multi-stage approach models the temporal information on both the level of individual frames and full clips. These enhanced representations show strong performance for unintentional action recognition tasks. We provide an extensive ablation study of our framework and report results that significantly improve over the state-of-the-art.

This chapter is based on [DKS22]. Anna Kukleva, as the co-first author, contributed to the conceptual development of this project and the writing of the paper. This paper is based on the master thesis of Enea Duka. Anna Kukleva acted as Enea Duka's supervisor during the project and played a significant role in the paper's writing.

4.1 Introduction

Video action understanding has witnessed great progress over the past several years in the fields of action detection, recognition, segmentation, caption generation, tracking and many others [DDM+21], [YXS+20], [KKSG19], [GZPB20], [WZWL21]. However, these methods im-

Figure 4.1: **Transition from intentional to unintentional action.** We can notice an abrupt change of motion in the first example, when the snowboarder starts to fall down the stairs, and a sudden change in speed in the second example when the skater starts rolling on the ground. In both cases, there is a strict order from intentional to unintentional action.

plicitly rely on the continuity of the underlying intentional action, such as one starts and then continues the same or related activity for some time. In this chapter, we study unintentional actions, for example, when one falls down by accident during jogging. This is a challenging task due to the difficulty in defining precisely the intentionality of an action while largely depending on the temporal context of the action. Moreover, annotating such videos is both costly and difficult with human accuracy being 88% in localising the transition between intentional and unintentional actions [ECV20]. The potential of this research covers assistance robotics, health care devices, or public video surveillance, where detecting such actions can be critical. In this work, we propose to look at unintentional actions from the perspective of the discontinuity of the intentionality and exploit inherent biases in this type of videos.

We design our framework to explore videos where a transition from intentional to unintentional action occurs. We are the first to explicitly leverage the inherent biases of unintentional videos in the form of motion cues and high-level action ordering within the same video, specifically the transition from unintentional to intentional and vice-versa.

To this end, we propose a three-stage framework where, starting from a pre-trained representation, the first two stages leverage self-supervised training to address unintentional biases on frame and video clip levels and further enhance these representations. We regularly observe abrupt motion changes in videos of unintentional actions when an unwitting event happens. Therefore, we formulate **T**emporal **T**ransformations to exploit **I**nherent **B**iases in videos that resemble **U**nintentional **A**ction motion changes (T^2IBUA). Using T^2IBUA we learn intra and inter-clip information in a multi-stage manner. In the first stage, we capture local information from the neighbouring frames by predicting the discrete label of the applied T^2IBUA on the frame level. Then, in the second stage, we train the model to integrate information from the entire video by predicting T^2IBUA on the clip level. See Fig. 4.4 for different T^2IBUA levels. Inspired by the growing popularity of transformer models in

vision [FXM+21], [SMV+19], [GCDZ19], [NBZA21], we utilize a multi-level transformer architecture that serves as a temporal encoder in our framework to model long-term relations. In the last stage of supervised downstream learning, we particularly benefit from the high-level ordering by deploying conditional random fields (CRF) to explicitly enforce smooth transitions within the same video from intentional to unintentional actions. By modelling these explicit global temporal relations, we enhance our results on various downstream tasks such as unintentional action classification, localisation and anticipation.

The contributions of this work are as follows:

- A framework that includes three learning stages for unintentional action recognition: by leveraging inherent biases in the first and in the second stages with self-supervised feature enhancement, the third stage employs supervised learning for unintentional action recognition on the top of these amplified representations;
- A multi-stage temporal encoder as a transformer-based architecture in combination with conditional random fields to capture local and global temporal relations between intentional and unintentional actions;
- We show state-of-the-art results for classification, detection, and anticipation tasks and perform various ablation studies that evaluate the effectiveness of each of the components of the proposed framework.

4.2 Related Work

In this section, we discuss prior work on the unintentional action recognition. We will not revisit the topic of self-supervised learning on video data as previously discussed in Chapter 2.

The topic of unintentional action recognition is recently introduced by Epstein et al. [ECV20]. The authors collect the dataset Oops! and propose a framework to study unintentional action recognition based on three directions such as classification, localization, and anticipation. The framework consists of two consecutive parts to learn the representations in a self-supervised way and then adapt the model to unintentional actions. Han et al. [HXZ20] propose to predict the future steps of the video based on the feature space on the given history frames. Epstein et al. [EV20] utilize a 3D CNN in combination with Transformer blocks to enhance representation learning and recognize discontinuities in the videos. Zatsarynna et al. [ZFG22] rely on the global context to learn good representations by training the model with a contrastive loss and order prediction loss. In contrast to [ECV20], we leverage self-supervised training to tailor pre-trained representations to unintentional actions and model temporal information in a multi-stage fashion through the multiple learning stages.

4.3 Method

In this and the next sections, we present our framework to study unintentional actions (UA) in videos. First, we provide an overview of our approach in Sec. 4.3.1. In Sec. 4.3.2 we detail T^{2}IBUA for self-supervised training, and then in Sec. 4.4 we describe the learning stages for our framework.

Notation: Let $X \in \mathcal{R}^{T \times W \times H \times 3}$ be an RGB video, where T, W and H are the number of frames, width and height respectively. We denote a clip sampled from this video as $x \in \mathcal{R}^{t \times W \times H \times 3}$

where t is the number of frames in the clip and $t \leq T$. As f we further denote individual frames from the video or clip. We denote T^{2}IBUA as a function $\mathcal{T}(\cdot)$, the spatial encoder as $\mathfrak{S}(\cdot)$ and the temporal encoder as $\mathfrak{T}(\cdot)$, and a linear classification layer as $MLP_i(\cdot)$, where i is a learning stage indicator.

4.3.1 Framework Overview

Unintentional actions (UAs) are highly diverse and additionally happen rarely in daily life, making it difficult to collect representative and large-scale datasets. To overcome this issue, we propose a three-stage learning framework for unintentional action recognition. Starting from the observation that unintentional actions are often related to changes in motion such as speed or direction, we aim to leverage such inherent biases of unintentional actions in our framework. In particular, simulating these inherent biases with a set of temporal video transformations, the first two stages of our framework greatly enhance pre-trained representation in a self-supervised fashion. More specifically, the first stage uses these transformations to learn intra-clip information (frame2clip) whereas the second stage uses the same transformations to address inter-clip information (frame2clip2video). The third stage then refines this representation via fine-tuning on the downstream unintentional action tasks using labelled data. We additionally enforce smooth transitions from intentional to unintentional action during the third stage by employing a conditional random field. While simulating the above-mentioned inherent motion biases cannot possibly cover the entire diversity of unintentional actions, it results in a powerful representation leveraged by the third stage of our approach and achieves a new state of the art for unintentional action recognition tasks.

For the first and the second stages, we generate labels based on the selected transformations. We extract frame features with a fixed pretrained spatial encoder, in particular using a pretrained ViT [DBK$^+$20] model. Furthermore, for fair comparison to previous work we also explore random initialization without any additional pre-training employing a Resnet3D architecture. Then we pass the frame features to the temporal encoder. The architecture of the temporal encoder changes from stage to stage that we discuss in detail in Sec. 4.4. For the first and the second stages, we use cross entropy loss with self-generated labels. The overview of the framework is shown in Fig. 4.2.

4.3.2 Temporal Transformations of Inherent Biases of Unintentional Actions

The potential diversity of unintentional actions encompasses all possible types of activities, since it is human nature to have failures during executing intentional actions. We aim to grasp the motion inherent biases of unintentional actions and propose several temporal transformations grouped in two categories: motion speed and motion direction transformations. Changes in the motion speed can correspond to unintentional actions when, for example, the person stumbles and tries to keep balance by moving faster. Whereas, changes in the motion direction can occur by falling down or unexpectedly sliding backwards. While these are just two examples, in practice similar observations hold for a wide variety of unintentional actions. We ground the set of transformations on the above intuition that connect temporal motion and inherent biases of unintentionality. In this work, we consider each video as an ordered set of units, which can be either frames or clips. We formulate the framework in a multi-stage manner and, thus, apply these transformations in the first stage to frames and in the second stage to clips. In Fig. 4.4 we show the difference between frame-level and clip-level T^{2}IBUA using the shuffle

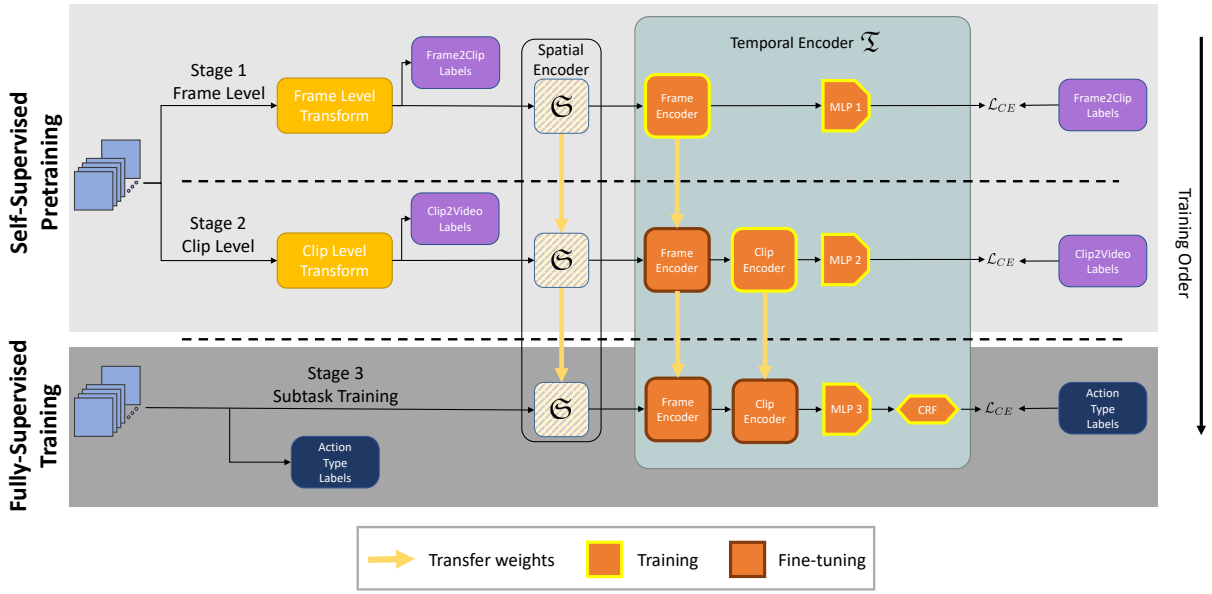


Figure 4.2: **Framework overview.** In the first and the second stages, we use self-supervised feature enhancement by predicting T^{2}IBUA . In the **first stage**, we enhance the representations so that they encode short-term dependencies based on neighbouring frames by training the frame encoder $\mathfrak{T}_{frame}$ and MLP_1. During the **second stage**, we further fine-tune representations so that they encode long-term dependencies using inter clip information by fine-tuning the frame encoder $\mathfrak{T}_{frame}$ and training the clip encoder $\mathfrak{T}_{clip}$ together with MLP_2. During the **stage three**, we train in a fully-supervised way for downstream tasks by fine-tuning the frame $\mathfrak{T}_{frame}$ and clip $\mathfrak{T}_{clip}$ encoders, while we train MLP_3 and the CRF parameters.

transformation.

We define T^{2}IBUA as follows:

- *Speed-up*: We synthetically vary the number of units of the video by uniformly subsampling them with the ratios $\{1/2, 1/4, 1/8\}$
- *Random point speed-up*: We sample a random index $ri \in \{1 \cdots t\}$ to indicate the unit from which we are synthetically speeding up the video. Specifically, we start subsampling of the video units after ri unit with ratio $1/\rho$:

$$[1, 2, \cdots, t-1, t] \rightarrow [1, 2, \cdots, ri, ri+\rho, \cdots, t-\rho, t].$$

  where t is the length of the video.
- *Double flip*: We mirror the sequence of the video units and concatenate them to the original counterpart:

$$[1, 2, \cdots, t-1, t] \rightarrow [1, 2, \cdots, t-1, t, t-1, \cdots, 2, 1].$$

- *Shuffle*: The video units are sampled in a random, non-repeating and non-sorted manner.
- *Warp*: The video units are sampled randomly and sorted increasingly by their original index.

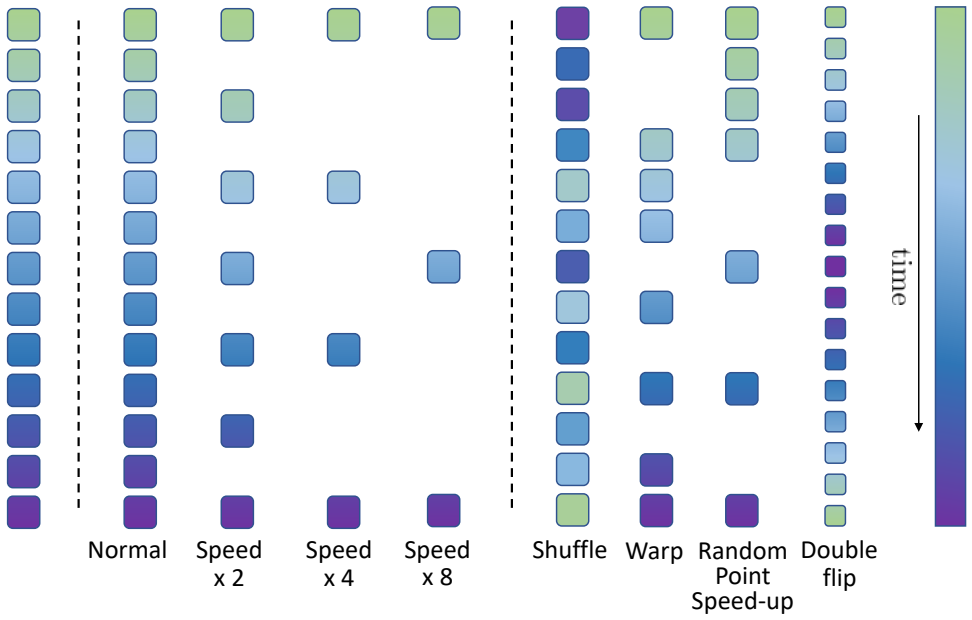


Figure 4.3: **T²IBUA.** We group T²IBUA in the **speed** group, where the speed of the video changes uniformly, and the **direction** group, where the speed changes non-uniformly, or we permute the units of the video.

All transformations are depicted in Fig. 4.3. We group T²IBUA in the motion speed group where we include the *Speed-up* variations and the motion direction group where we include the rest of the proposed T²IBUA.

T²IBUA prediction: For the feature enhancement, we associate a discrete label to each T²IBUA and train our framework in a self-supervised way to predict the label of the T²IBUA that we apply either on frame or clip level. Each input sequence is transformed into a set of 6 sequences. The correspondences between the self-generated labels and the 6 transformed sequences are as follows: 1 - Initial sequence without T²IBUA; {2, 3, 4} - *Speed-up* with sampled ratio $\in \{1/2, 1/4, 1/8\}$; 5 - *Random point speed-up*; 6 - *Double flip*; 7 - *Shuffle*; 8 - *Warp*. Specifically, we predict the transformation applied to each of the 6 resulting sequences correspondingly. Therefore, the model learns to distinguish various simulated inherent biases of unintentional actions on the same sequence.

4.4 Multi-Stage Learning for Unintentional Action Recognition

In this section, we first introduce the Transformer block, a building element for the different modules of our temporal encoder in Sec. 4.4.1. Then we discuss in details each learning stage of our framework in Sec. 4.4.2, 4.4.3 and 4.4.4.

4.4.1 Transformer block

In this work, to process long sequences, we utilize a transformer architecture. We integrate intra and inter clip information and capture information for unintentional action recognition

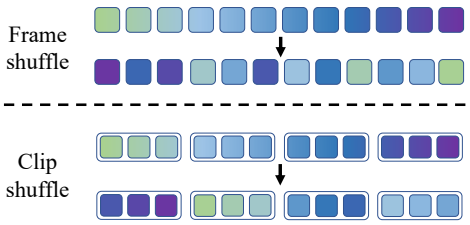


Figure 4.4: **Shuffle transformation. Top**: Frame level. Each frame moves randomly to any position. **Bottom**: Clip level. We group frames into clips. Frames cannot change the position within a clip, while clips are shuffled randomly.

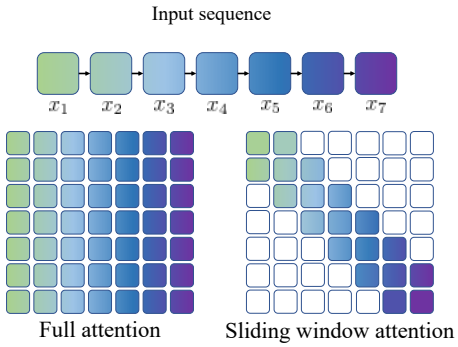


Figure 4.5: **Attention computation. Left**: Full attention. Required memory and processing time scale quadratically. **Right**: Sliding window attention mechanism. Required memory and processing time scale linearly with the input sequence length.

from different granularity levels, frame and clip levels. To aggregate information over the whole video, we need to process a long sequence of consecutive frames or clips from the video. Further, we refer to the elements of frame or clip sequences as units. In the datasets, all videos are usually of different lengths, and the vanilla transformer model [DBK+20] scales quadratically with the sequence length in the required memory and processing time. The Longformer [BPC20] addresses this problem by attending only to a fixed-sized window for each unit. The difference in the attention computation is shown in Fig. 4.5. We apply a window of size w so that each unit attends to $\frac{1}{2}w$ units from the past and $\frac{1}{2}w$ from the future. All building blocks for the temporal encoder in our work follow the structure of Longformer attention blocks. The transformer blocks can update the input sequence based on the temporal context or aggregate the temporal information from the entire sequence into one representation vector. For the former, for a given input sequence of units $U = \{u_1, \cdots, u_K\}$ we update the sequence and output $U' = \{u'_1, \cdots, u'_K\}$. For the latter, we expand the given input sequence with an additional classification unit cls that serves as aggregation unit across the whole sequence [DBK+20], [NBZA21]. Therefore, we map the expanded input sequence $U \cup \{cls\}$ to cls' that aggregates the sequence into one representation. For more details we refer to the original paper [BPC20].

4.4.2 [Stage 1] Frame2Clip (F2C) learning

In the first stage of our framework, we enhance the pre-trained features such they encode intra clip information based on neighboring frames. We operate within short clips and, thus, capture local motion information. First, we sample a short clip x of the length n from a long input video X, where $n < |X|$ and $|X|$ is the length of the input video. The clip x consists of sequential frames $x = \{f_1, \cdots, f_n\}$ from the same video X. We apply T^{2}IBUA $\mathcal{T}$ to this sequence of frames of the clip x as we show schematically in the shuffle transformation on the top of Fig. 4.4. As a label for the transformed clip $\mathcal{T}(x)$ we use the index that corresponds to the respective T^{2}IBUA, e.g. the shuffle transformation has index 7. Then, from the spatial encoder $\mathfrak{S}$ we obtain the spatial frame features for the sequence. Thereafter, to impose the intra connections between the frames of the same sequence, we use the temporal frame encoder $\mathfrak{T}_{frame}$ that is composed of the transformer blocks. The output of this encoder is one representation vector

for the clip, hence we expand the sequence of frame features with an additional unit cls_1 to aggregate the temporal information from the clip x into one representation $\tilde{x}$. Finally, we predict the T^{2}IBUA index with a classification layer MLP_1. We compute cross-entropy loss and optimize parameters of the temporal frame encoder $\mathfrak{T}_{frame}$ and a classification layer MLP_1. Formally, we apply the following pipeline in the first stage:

$$\hat{y}_{clip} = MLP_1(\mathfrak{T}_{frame}(\mathfrak{S}(\mathcal{T}(f_1, \cdots, f_n)), cls_1)), \tag{4.1}$$

where $\hat{y}_{clip}$ is the predicted label of the transformation applied to the clip x. Note that at this stage of the training, we process all sampled clips independently of each other. Our F2C learning implies intra-clip encoding, therefore, it can be also substituted with the more common spatio-temporal networks such as ResNet3D that learn representations of short clips.

4.4.3 [Stage 2] Frame2Clip2Video (F2C2V) learning

In the second stage of our framework, we further enhance the representations based on the clip level transformations. At this stage, we integrate inter clip information into the video representation to model long-term dependencies. The input video X of length $|X|$ we split into overlapping clips $z = \{x_1, x_2, \cdots, x_N\}$ that we sample with stride k from the video, where $N = \frac{|X|-n}{k} + 1$. Each clip consists of n frames $x_i = \{f_1^i, \cdots, f_n^i\}$. During F2C2V learning stage we apply T^{2}IBUA $\mathcal{T}$ to the sequence of clips of the whole video X as we show schematically on the bottom of Fig. 4.4. Specifically, the order of consecutive frames within each clip remains fixed, whereas the sequence of the clips is transformed $\mathcal{T}(z)$. As in the first stage, we first pass frames of the clips through the spatial encoder $\mathfrak{S}$ to obtain frame-level features, then the sequence of frames for each clip $x_i = \{f_1^i, \cdots, f_n^i\}$ is aggregated into one clip representation $\tilde{x}_i$ by the frame encoder $\mathfrak{T}_{frame}$. Note that the frame sequence during this stage follows the original order. Then we aggregate the transformed sequence of clip representations $\mathcal{T}(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N)$ into video representation vector $\tilde{X}$ with the temporal clip encoder $\mathfrak{T}_{clip}$. We predict T^{2}IBUA that was applied to the sequence of clips with the classification layer MLP_2. Formally, the second stage is as follows:

$$\tilde{x}_i = \mathfrak{T}_{frame}(\mathfrak{S}(f_1^i, \cdots, f_n^i), cls_1), \ \forall i \in \{1, \ldots, N\}; \tag{4.2}$$

$$\hat{y}_{video} = MLP_2(\mathfrak{T}_{clip}(\mathcal{T}(\tilde{x}_1, \cdots, \tilde{x}_N), cls_2)), \tag{4.3}$$

where $\hat{y}_{video}$ is the predicted T^{2}IBUA label applied to a sequence of clips from video X. To aggregate all the clips into one video vector we similarly use an additional classification token cls_2 as in the previous stage. At this stage, we optimize the parameters of the temporal clip encoder $\mathfrak{T}_{clip}$ and the classification layer MLP_2, the parameters of the temporal frame encoder $\mathfrak{T}_{frame}$ we transfer from the previous stage and fine-tune. Note that at each stage we utilize a new classification layer and discard the one from the previous stage.

4.4.4 [Stage 3] Downstream Transfer to Unintentional Action Tasks

In the last stage, we extend our framework to supervised unintentional action recognition tasks. For these tasks, we predict the unintentionality for each short clip rather than for the entire video. This stage is completely supervised with clip level labels, therefore we do not apply T^{2}IBUA. The input video X is divided into temporally ordered overlapping clips $z = \{x_1, x_2, \cdots, x_N\}$. We follow the pipeline of the second stage and extract clip level representation vectors $\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N\}$ with the temporal frame encoder $\mathfrak{T}_{frame}$. Then we use the temporal clip encoder $\mathfrak{T}_{clip}$ without the classification aggregation unit cls_2 given that for the

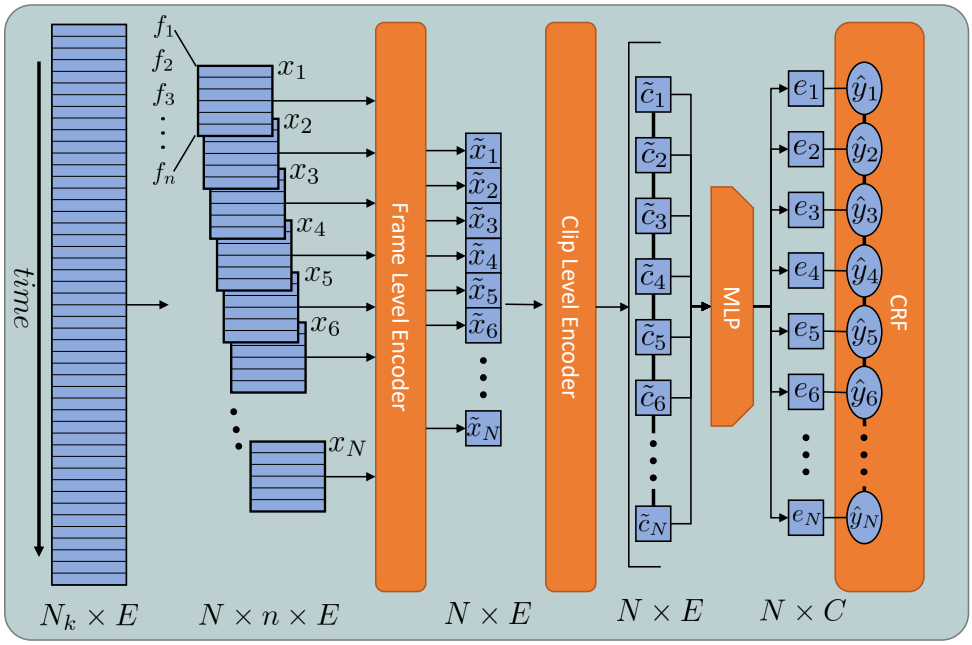


Figure 4.6: **Multi-stage temporal encoder.** The $\mathfrak{T}_{frame}$ encoder aggregates the frames of each input clip x_1 into single-vector clip representation $\tilde{x}_i$. $\mathfrak{T}_{clip}$ updates $\tilde{x}_i$ with inter-clip information resulting into $\tilde{c}_i$. *MLP* turns $\tilde{c}_1$ into emission scores e_i for the CRF layer, which predicts clip-level labels.

supervised tasks we use clip-level classification labels. The output of $\mathfrak{T}_{clip}$ then is the sequence of clip representations instead of one video representation vector. By using the temporal clip encoder $\mathfrak{T}_{clip}$ each clip representation vector is updated with the inter clip information from the whole video X. Finally, we pass the clip vectors through the classification layer MLP_3 to obtain the emission scores that we use in the CRF layer $\mathfrak{T}_{CRF}$. The overall downstream transfer stage is as follows:

$$\tilde{x}_i = \mathfrak{T}_{frame}(\mathfrak{S}(f_{i,1}, \cdots, f_{i,n}), cls_1), \ \forall i \in \{1, \ldots, N\}; \tag{4.4}$$

$$\{\hat{y}_1, \cdots, \hat{y}_N\} = \mathfrak{T}_{CRF}(MLP_3(\mathfrak{T}_{clip}(\tilde{x}_1, \cdots, \tilde{x}_N)), T), \tag{4.5}$$

where $\{\hat{y}_1, \cdots, \hat{y}_N\}$ predicted clip level labels and T is the transition matrix of the CRF layer.

The CRF layer aims to explicitly model the high-level temporal structure of the unintentional videos between the clips as shown in Fig. 4.1. For videos containing unintentional actions the high-level order imposes a smooth transition from normal to transition to unintentional clips, therefore we can incorporate this prior information on the order of the clip labels. The CRF layer facilitates dependencies between the current clip and the previously predicted labels and enforces neighbouring clips to have the same label. This layer is parametrized by the transition matrix $T \in \mathcal{R}^{C \times C}$, where $T_{i,j}$ is the score of transitioning from label i to label j and C stands for the number of classes. We denote by θ the parameters of the temporal encoder $\mathfrak{T}$ and MLP_3 together that we optimize during training, by $L \in \mathcal{R}^N$ the vector of clip labels for the input sequence of clips z and by $E \in \mathcal{R}^{N \times C}$ an emission matrix that consists of logits for each unit of the input sequence generated by MLP_3 layer. Then, we can calculate the scores for a label sequence $\hat{L}$ given the input sequence $z = \{x_1, \cdots, x_N\}$, trainable weights θ and the transition matrix T as

$$s(\hat{L}, z, \theta, T) = \sum_{t=1}^{n} \left(T_{L_{t-1}, Lt} + E_{t, L_t}\right). \tag{4.6}$$

Let L_{all} be the set of all possible vectors of labels for the input clip sequence. Then the loss

function is the negative log-likelihood:

$$\mathcal{L}(\theta, T) = -log \frac{exp(s(L_{gt}, z, \theta, T))}{\sum_{L \in L_{all}} exp(s(L, z, \theta, T))}, \tag{4.7}$$

where L_{gt} the ground truth labels for each clip of the input sequence z. The total loss is the average over all training sequences.

During the inference, we compute the optimal labels L_{opt} for the input sequence of clips z_{test} as

$$L_{opt} = _{L \in L_{all}} s(L, z_{test}, \theta^{\star}, T^{\star}), \tag{4.8}$$

where $\theta^{\star}$ and $T^{\star}$ are the optimized model parameters. Note that Eq. 4.8 can be solved efficiently with the Viterbi algorithm.

4.5 Experiments

In this section, we present experimental findings that validate our framework. First, we introduce the Oops! dataset, present different backbone models used in the framework and discuss the implementation details of our method. We further compare our framework to the state-of-the-art on the three subtasks in Sec. 4.5.1 and then we present extensive ablation experiments to validate the impact of each of the components in Sec. 12.4.2.

Dataset: Oops! [ECV20] is a collection of 20,338 amateur fail videos from YouTube. The dataset is split into three sets: 7,368 labelled videos for supervised training, 6,739 labelled videos for validation, and 6,231 unlabelled videos for pretraining. Each label in the first two sets consists of the timestamp where the action transitions from intentional to unintentional. Following the clip sampling procedure described in Sec. 4.4 we get $18,069$ intentional clips, $4,137$ transitional clips and $19,679$ unintentional clips.

Spatial Encoder: To disentangle the influence of different components of the framework, to provide a fair comparison to previous methods and to follow the trend of performant architectures, we use two backbones for the main experiments. Similarly to prior work [ECV20], [HXZ20], we employ ResNet3D (R3D) [HKS18]. This backbone is spatio-temporal and, therefore, substitutes both spatial $\mathfrak{S}$ and frame level $\mathfrak{T}_{frame}$ encoder as we discuss in Sec. 4.4.2, we refer to it as F2C level for R3D backbone. In this setup, we learn the representations from scratch instead of enhancing pre-trained representations to fairly compare to previous works. To decouple spatial and temporal dimensions we use ViT model [DBK$^+$20] pretrained on ImageNet-21K[RBBNZM21] (IN-21K). In this setup, we leverage the pre-trained image representations and further enhance them for unintentional actions. We note that we freeze ViT model, while R3D model we train from scratch with a random initialization. Each encoder consists of three stacked transformer blocks, each block constitutes of 16 parallel heads.

Implementation details: We sample clips of 16 frames from the input video with stride $k = 4$. We train for 100 epochs for the first and second stage and for 50 epochs for the third stage with the AdamW [LH17b] optimizer with weight decay of $1e-4$. The starting learning rate for all stages is $1e-4$ and is decreased to $1e-6$ using a cosine decay policy. During the third stage, the loss function is weighted as $\omega_i = {}^{max(\eta_1, \eta_2, \cdots, \eta_c)}/_{\eta_i}$, where η_1 is the number of labels in class i and c is the number of classes. We recalculate the weights during the anticipation task as class sample distribution changes over time.

Method	Backbone	Init.	Pretrain Dataset	Cls. Acc.	Loc. $\tau_L : 0.25$	Loc. $\tau_L : 1$	Ant. $\tau_A : 1$
K700 Supervision [ECV20]	R(18)3D	K700	-	64.0	46.7	75.9	59.7
Epstein et al. [ECV20]	R(18)3D	-	Oops!	61.6	36.6	65.3	56.7
Han et al. [HXZ20]	(2+3D)R18	-	(K400+Oops!)	64.4	-	-	-
Ours (F2C)	R(18)3D	-	Oops!	65.3	37.7	67.8	66.7
Ours (F2C2V)	R(18)3D	-	Oops!	**74.0**	**39.4**	**69.5**	**76.1**
Ours (F2C)	ViT*	IN-21K	Oops!	65.5	41.4	72.2	69.2
Ours (F2C2V)	ViT*	IN-21K	Oops!	**76.9**	**42.8**	**72.8**	**78.1**

Table 4.1: **Comparison of our approach to state-of-the-art for UA classification, localisation and anticipation.** F2C denotes the first (frame2clip) learning stage, F2C2V denotes the second (frame2clip2video) learning stage. τ_L indicates a time window in seconds for ground truth assignment for the localization task. τ_A indicates the time step in the future that we predict. Init. indicates fully supervised initialization of the backbone if applies. * indicates that the backbone is frozen during all stages.

4.5.1 Comparison to state-of-the-art

In this section we compare the performance of our framework on the three downstream tasks as classification, localization and anticipation. In Tab. 4.1 we provide a comparison to the previous state-of-the-art method across all three tasks. The first row in the table corresponds to supervised pretraining on the Kinetics 700 [SCN+20]. The benchmark that we follow for the fair comparison across all the tasks is defined in Oops! dataset [ECV20].

Unintentional Action Classification. We first compare our framework to the recent state-of-the-art work on unintentional action classification. We divide Tab. 4.1 into blocks with comparable backbones and pre-training methods. We can observe a significant improvement over the previous state-of-the-art method by approximately 10 points with R3D backbone and training from scratch. In contrast to Han et al. [HXZ20] we pretrain our model only on the Oops! dataset without additional external data. With our temporal encoder on the frame level, we gain minor improvement by 0.9 points indicating that T^2IBUA helps to improve even on the F2C level, while our two stage temporal encoder allows us to achieve substantial increase in the performance by 9.6 points. We notice further increase in performance when we use pre-trained frozen ViT as our backbone. For this setup, we enhance the pre-trained features and gain an improvement of 1.1 points with our frame level encoder and 12.5 points with our two stage temporal encoder. These results indicate the importance of connections between intra and inter clip information in the video

Unintentional Action Localisation. For this downstream task, we localize the transition point from intentional to unintentional action in the video. We directly validate the network trained on unintentional action classification task and detect the transition point from intentional action to unintentional. The transitional clip for this task we define as the clip with the maximum output score of being transitional. In Tab. 4.1 the localisation column shows the performance of our framework for two temporal localisation thresholds. By using R3D as our backbone, we improve by 2.8 and 4.2 points for the $\tau_L = 0.25$ and $\tau_L = 1$ respectively compared to Epstein et al. [ECV20].We imporove by 6.2 and 7.5 points for each threshold respectively when using ViT as our backbone. Additionally, we can observe the same trend as for the classification task,

T²IBUA	CRF	ViT		R3D	
		F2C	F2C2V	F2C	F2C2V
–	–	60.9	69.6	59.3	65.6
✓	–	65.5	74.3	63.8	70.4
✓	✓	65.0	**76.9**	65.3	**74.0**

Table 4.2: **Influence of T²IBUA and CRF layer.** We evaluate classification performance for two representations, pretrained ViT representation and R3D.

specifically, the clip level model outperforms the frame level model. Note that for this task we reuse the model that is trained for the classification task where the number of normal and unintentional clips is notably higher than the number of transitional clips. The localisation task requires the model to be very specific on the boundaries when the unintentionality starts. We suppose that this influences the smaller improvement for the localization task than for the other tasks.

Unintentional Action Anticipation. Further, we validate our framework on anticipation of unintentional actions. During the supervised training, we train our model to predict the label of the future clip. To directly compare to the previous work, we anticipate an action 1.5 seconds into the future as in previous work. Our model achieves new state-of-the-art results with a considerable improvement by 19.4 and 21.4 points for the R3D and ViT backbones respectively. Note that the performance of our framework is better for the anticipation task than for classification. As mentioned, the dataset includes more unintentional clips than normal clips, and in combination that we are able to predict unintentional clips more accurately, it leads to the difference in the performance between the downstream tasks.

4.5.2 Ablation study

We perform ablation studies on different components and backbones of our framework to assess their influence on the overall performance. We use the unintentional action classification task for all the evaluations in this section.

Feature enhancement stages. In this section we analyse the influence of the self-supervised training to enhance the representations with T²IBUA on the UA classification task. In Table 4.2 the first and the second rows show the performance with and without T²IBUA respectively. Specifically, for the first row, we skip the self-supervised procedure for the pretraining of the parameters and directly optimize from scratch the respective temporal encoders with the supervised task. For the second row, we include the self-supervised feature enhancement corresponding to the respective learning stages. We can observe that for each temporal encoder (F2C and F2C2V) we have a significant increase with T²IBUA representation enhancing (learning). Considering only intra clip information (F2C) we improve by 4.6 and 4.5 points for the ViT and R3D backbones respectively while using additionally inter clip information we improve by 4.7 and 4.8 points. These results confirm the importance of the self-supervised feature enhancement for UA recognition performance.

Learning Stages for Temporal Encoder. In this section, we evaluate the impact of multi-stage learning of our temporal encoder. In Table 4.2 we show in the F2C column the performance across different settings for the corresponding encoder, see the detailed structure in Figure 4.6 for ViT backbone, whereas R3D backbone comprises this stage. In the second column with

T^2IBUA	F2C	F2C2V
–	60.9	73.2
Speed	62.5	73.8
Direction	63.3	75.1
All	**65.5**	**76.9**

Table 4.3: **Influence of T^2IBUA groups.** We evaluate UA classification task for different stages of learning of the temporal encoder.

the F2C2V encoder, we observe a significant improvement by about $5-10$ points consistently across all the settings. Furthermore, we evaluate the performance of the spatial features from the fixed pretrained ViT model [DBK+20] that we use as input features to our temporal encoder. We obtain 58.4 points without T^2IBUA and without CRF. It supports the importance of the temporal encoder that is able to capture successfully inter and intra clip information.

T^2IBUA groups. We assess the influence of the T^2IBUA groups on the overall performance. Table 4.3 shows that both groups improve the performance for both encoder levels. We notice that the direction T^2IBUA have a more significant impact than the speed T^2IBUA. Whereas, in contrast to the separate groups, the combination of the speed and the direction transformations leads to greatly enhanced representations that effectively capture the inherent biases of UA. We additionally study an influence of each transformation separately that shows random point speed-up to be the most important, while the combination of all T^2IBUA is still predominant.

4.5.3 Qualitative results

In this section, we show qualitative results for our framework. Figures 4.7, 4.8, 4.9, 4.10, 4.11, 4.12 display these results. In each case, the first plot in the figure shows the confidence of our framework before using CRF. The x axis represents the time in seconds, while the y axis represents the confidence for that prediction. We mark the ground truth transition point t_{gt} of the video with the vertical line at $x = t_{gt}$. The second plot in each figure shows the predictions of the framework when we include CRF. The x axis shows the time in seconds, while the y axis shows the discrete clip label. The third plot in each figure, shows the shows the confidence of model in [ECV20] and has the same layout as the first plot. At the top of the figure, we show frames for clips close to the transition point in the video. The color of the frame border indicates the ground truth label for it. We take the frame in the middle of the clip each prediction is related to.

In Figures 4.7, 4.8, we show samples for which our network performs the best. First, we notice in the continuous plot that the prediction is correct for each clip. In addition, we can see that the confidence is high for all clips and there is a clean transition point and ordering between the different types of clips. These results translate to the discrete scatter plot, where we notice that all the predictions are correct. In contrast to our results, the results from [ECV20] are clearly more noisy and less correct.

In Figures 4.9, 4.10 we show samples for which the predictions are less accurate. We observe that the order of the clip types is preserved. However, the exact transition from one clip type to the next is less certain, specifically, the confidence of the predictions on the boarders between the different types is lower than the maximal possible score as in the previous case. The same is not true for results from [ECV20] where not only the prediction confidence is suboptimal but also the clip order is not preserved. In the scatter plots of both figures, we notice the

improvement on the performance due to the CRF layer.

Finally, in Figures 4.11, 4.12 we show samples on which our framework performs the worst. In this case, we observe that there are noisy predictions which violate the order of the clip types as well as the clean transitions between the clip types. We notice similar results for these samples when using the model from [ECV20]. CRF layer improves the quality of the predictions in these cases the most. It makes the prediction smoother and reduces the noise, but the transition point localisation remains poor.

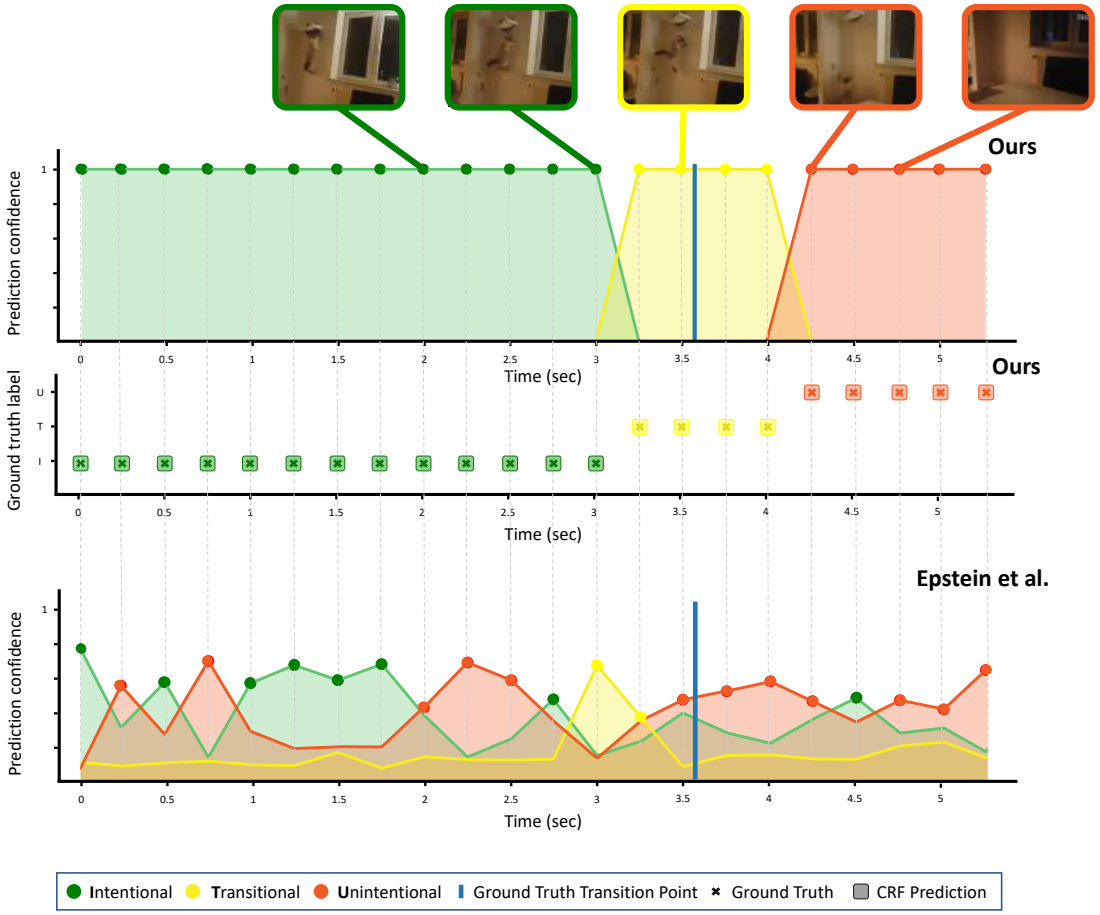


Figure 4.7: A cat falling while trying to hold on to an object on the wall. We notice that when we do not use CRF, all the predictions are correct and with high confidence. After adding CRF, the predictions remain all correct.

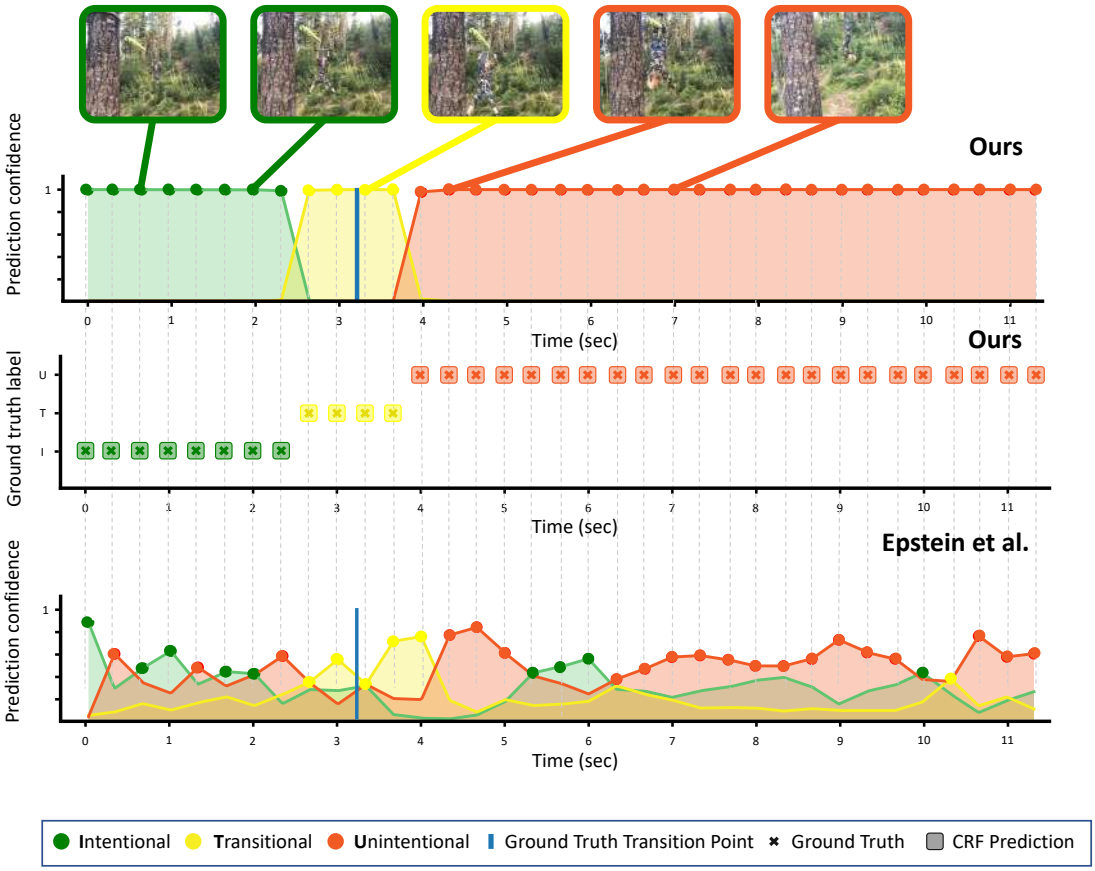


Figure 4.8: A person abruptly ending the zipline ride and turning upside-down at the end of it. We notice that when we do not use CRF, all the predictions are correct and with high confidence. After adding CRF, the predictions remain all correct.

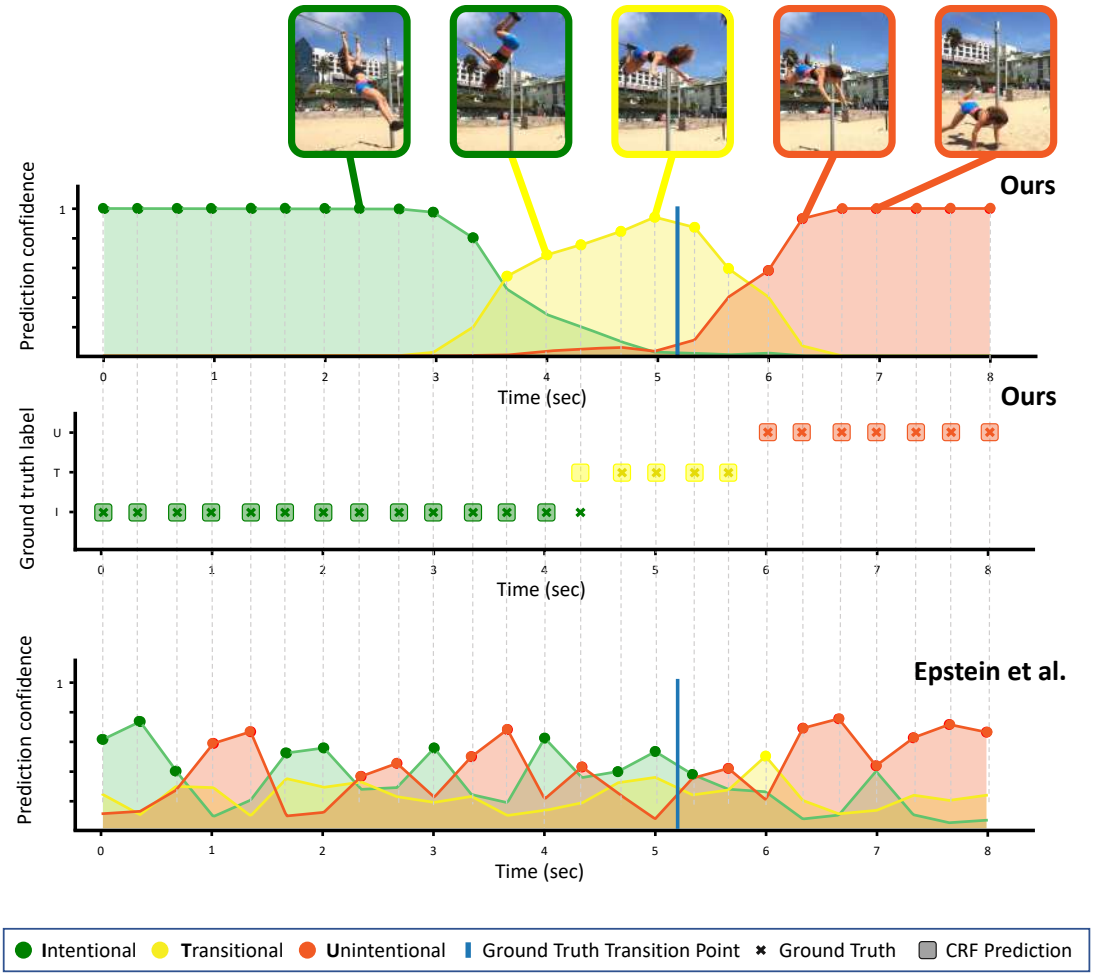


Figure 4.9: A person trying to perform tricks on a pull-up bar and falling down in the sand. We notice that when we do not use CRF, most of the predictions are correct and not all of them have a high confidence. Adding CRF in this case leads to more predictions being correct.

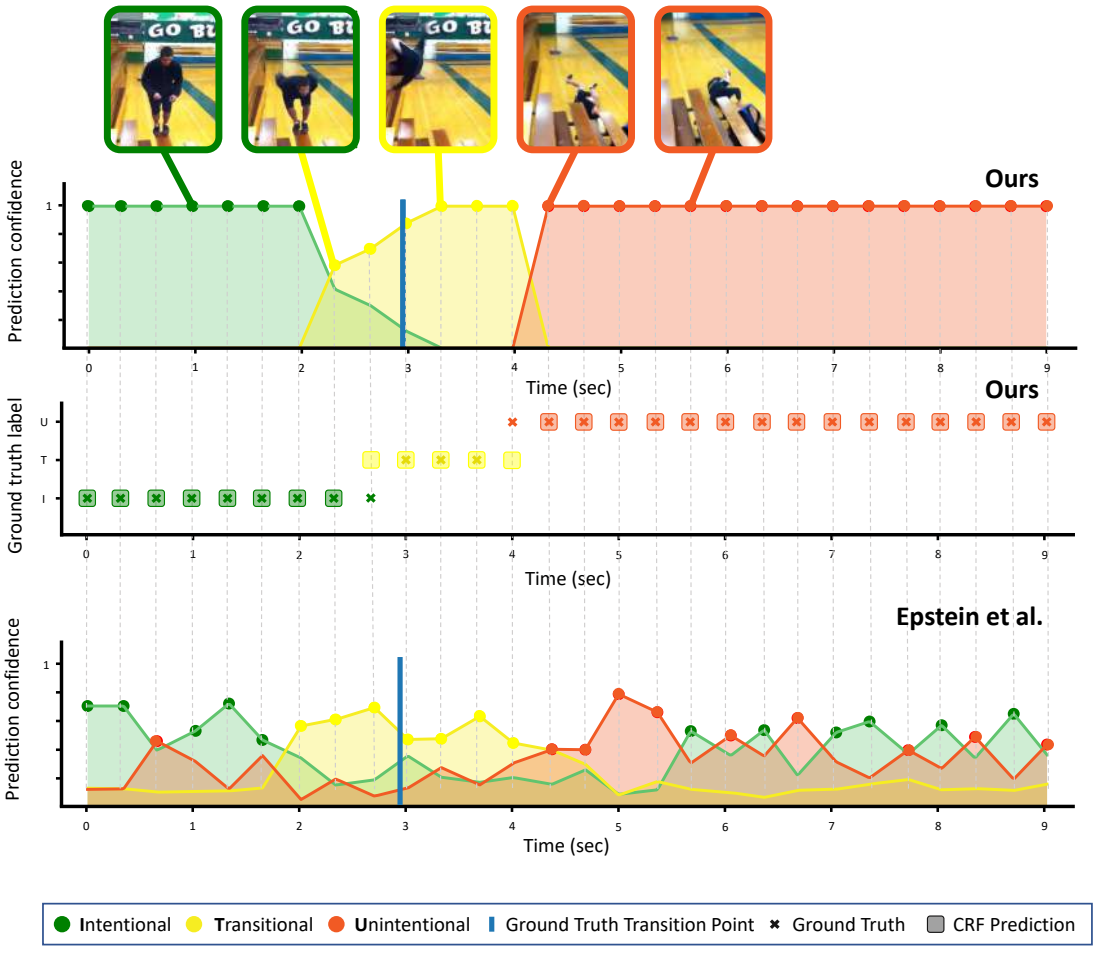


Figure 4.10: A person trying to perform a backflip from the top of a bench and falling on the floor. We notice that when we do not use CRF, most of the predictions are correct and not all of them have a high confidence. Adding CRF in this case leads to marginally better performance.

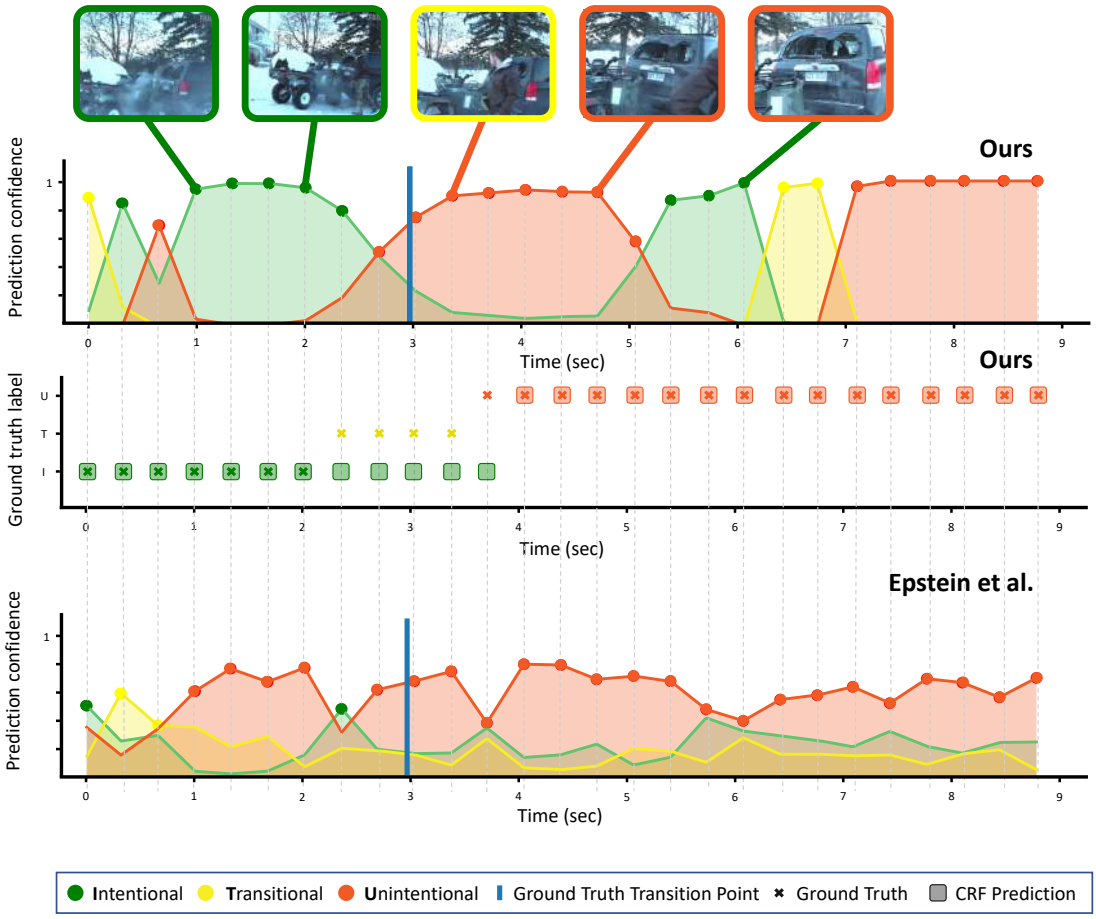


Figure 4.11: An ATV has hit the back of a car and the camera films the aftermath. We notice that when we do not use CRF, there are few predictions that are correct, and they are noisy overall. In addition, it is hard to correctly locate the transition point from intentional to unintentional action. Adding CRF makes the predictions overall smoother and less noisy. However, it is still hard to locate the transition point.

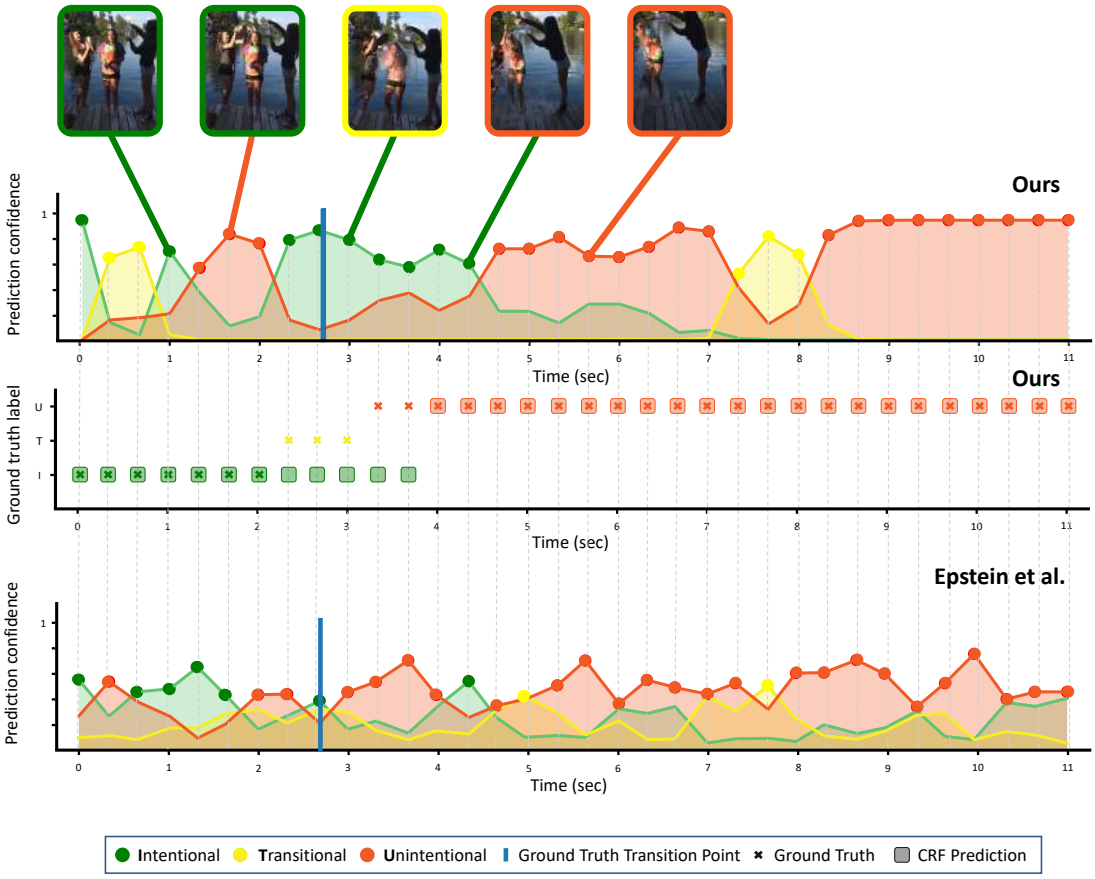


Figure 4.12: The person on the left falling in the lake while trying to throw a bucket of water on the person in the middle. We notice that when we do not use CRF, there are few predictions that are correct, and they are noisy overall. In addition, it is hard to correctly locate the transition point from intentional to unintentional action. Adding CRF makes the predictions overall smoother and less noisy. However, it is still hard to locate the transition point.

4.6 Conclusion

In Chapter 3, we similarly utilize the inherent biases found in videos, where the sequence of actions is predefined by the tasks. In this chapter, we explore the characteristics of unintentional actions. We propose a multi-stage framework to exploit inherent biases that exist in videos of unintentional actions for learning targeted representations. First, we simulate these biases with the temporal transformations that we employ for self-supervised training to enhance pre-trained representations. We formulate the representation enhancement task in a multi-stage way to integrate inter and intra clip video information. This leads to powerful representations that significantly enhance the performance on the downstream UA tasks. Finally, we employ a CRF to explicitly model global dependencies. We evaluate our model on the three unintentional action recognition tasks, classification, localisation, and anticipation, and achieve state-of-the-art performance across all of them. Transitioning to the next chapter, Chapter 5, we introduce a framework designed to clarify the underlying learning dynamics of self-supervised learning methods, with broad applicability to image and video recognition frameworks.

5 Temperature Schedules for Self-Supervised Contrastive Methods on Long-Tail Data

Contents

In this chapter, we transition from examining representations designed for specific types of videos to analyzing the behavior of one of the most popular variants of SSL: contrastive methods. We focus particularly on their performance on long-tail data, given that natural data often follows long-tail distributions. In particular, we investigate the role of the temperature parameter τ in the contrastive loss, by analysing the loss through the lens of average distance maximisation, and find that a large τ emphasises group-wise discrimination, whereas a small τ leads to a higher degree of instance discrimination. While τ has thus far been treated exclusively as a *constant* hyperparameter, in this work, we propose to employ a *dynamic* τ and show that a simple cosine schedule can yield significant improvements in the learnt representations. Such a schedule results in a constant 'task switching' between an emphasis on instance discrimination and group-wise discrimination and thereby ensures that the model learns both group-wise features, as well as instance-specific details. Since frequent classes benefit from the former, while infrequent classes require the latter, we find this method to consistently improve separation between the classes in long-tail data without any additional computational cost.

This chapter is based on [KBS+23]. Anna Kukleva, as the co-first author, contributed equally to the design of the temperature scheduling, the analytical framework, as well as the writing. the conceptual development of this project and the writing of the paper. The implementation of the framework was fully handled by Anna Kukleva.

5.1 Introduction

Deep Neural Networks have shown remarkable capabilities at learning representations of their inputs that are useful for a variety of tasks. Especially since the advent of recent self-supervised learning (SSL) techniques, rapid progress towards learning universally useful representations has been made.

Currently, however, SSL on images is mainly carried out on benchmark datasets that have

been constructed and curated for supervised learning (e.g. ImageNet [DDS+09], CIFAR [KH09], etc.). Although the labels of curated datasets are not *explicitly* used in SSL, the *structure* of the data still follows the predefined set of classes. In particular, the class-balanced nature of curated datasets could result in a learning signal for unsupervised methods. As such, these methods are often not evaluated in the settings they were designed for, i.e. learning from truly unlabelled data. Moreover, some methods (e.g. [ARV20, CMM+20]) even explicitly enforce a uniform prior over the embedding or label space, which cannot be expected to hold for uncurated datasets.

In particular, uncurated, real-world data tends to follow long-tail distributions [Ree01], in this chapter, we analyse SSL methods on long-tailed data. Specifically, we analyse the behaviour of contrastive learning (CL) methods, which are among the most popular learning paradigms for SSL.

In CL, the models are trained such that embeddings of different samples are repelled, while embeddings of different 'views' (i.e. augmentations) of the same sample are attracted. The strength of those attractive and repelling forces between samples is controlled by a temperature parameter τ, which has been shown to play a crucial role in learning good representations [CFGH20, CKNH20]. To the best of our knowledge, τ has thus far almost exclusively been treated as a *constant* hyper-parameter.

In contrast, we employ a *dynamic* τ during training and show that this has a strong effect on the learned embedding space for long-tail distributions. In particular, by introducing a simple schedule for τ we consistently improve the representation quality across a wide range of settings. Crucially, these gains are obtained without additional costs and only require oscillating τ with a cosine schedule.

This mechanism is grounded in our novel understanding of the effect of temperature on the contrastive loss. In particular, we analyse the contrastive loss from an average distance maximisation perspective, which gives intuitive insights as to why a large temperature emphasises *group-wise discrimination*, whereas a small temperature leads to a higher degree of *instance discrimination* and more uniform distributions over the embedding space. Varying τ during training ensures that the model learns both group-wise and instance-specific features, resulting in better separation between head and tail classes.

The contributions of this work are as follows:

- We carry out an extensive analysis of the effect of τ on imbalanced data;
- We analyse the contrastive loss from an average distance perspective to understand the emergence of semantic structure;
- We propose a simple yet effective temperature schedule that improves the performance across different settings;
- We show that the proposed τ scheduling is robust and consistently improves the performance for different contrastive learning methods and hyperparameter choices.

5.2 Related Work

In this section, we discuss more in details prior work on the role of negatives in contrastive learning, the imbalanced self-supervised learning and the analysis of contrastive learning. We will not revisit the topic of contrastive self-supervised learning as previously discussed in Chapter 2.

Negatives. The importance of negatives for contrastive learning is remarkable and noticed in many prior works [WWW+21, YHH+22, ZZP+22, ITAC18, KSP+20, RCSJ20, KAG22, BIS+23, LYZ+23]. [YHH+22] proposes decoupled learning by removing the positive term from the denominator, [RCSJ20] develops an unsupervised hard-negative sampling technique, [WWW+21] proposes to employ a triplet loss, [LYZ+23] proposes deep graph clustering methods with a comprehensive similarity measure criterion and a general dymanic sample weighing strategy, and [ZZP+22, KAG22] propose to improve negative mining with the help of different temperatures for positive and negative samples that can be defined as input-independent or input-dependent functions, respectively. In contrast to explicitly choosing a specific subset of negatives, we discuss the Info-NCE loss [OLV18] through the lens of an average distance perspective with respect to all negatives and show that the temperature parameter can be used to implicitly control the effective number of negatives.

Imbalanced Self-Supervised Learning. Learning on imbalanced data instead of curated balanced datasets is an important application since natural data commonly follows long-tailed distributions [Ree01, LMZ+19, WRH17]. In recent work, [KLX+20], [YX20], [LHGM21], [ZTC+22], [GS22] discover that self-supervised learning generally allows to learn a more robust embedding space than a supervised counterpart. [THvdO21] explores the down-stream performance of contrastive learning on standard benchmarks based on large-scale uncurated pre-training and propose a multi-stage distillation framework to overcome the shift in the distribution of image classes. [JCMW21, ZYW+22] propose to address the data imbalance by identifying and then emphasising tail samples during training in an unsupervised manner. For this, [JCMW21] compares the outputs of the trained model before and after pruning, assuming that tail samples are more easily 'forgotten' by the pruned model and can thus be identified. [ZYW+22] uses the loss value for each input to identify tail samples and then use stronger augmentations for those. Instead of modifying the architecture or the training data of the underlying frameworks, we show that a simple approach—i.e. oscillating the temperature of the Info-NCE loss [OLV18] to alternate between instance and group discrimination—can achieve similar performance improvements at a low cost.

Analysis of Contrastive Learning (CL). Given the success of CL in representation learning, it is essential to understand its properties. While some work analyses the interpretability of embedding spaces [BZK+17, FV18, LFV20, LAV21], here the focus lies on understanding the structure and learning dynamics of the objective function such as in [SPA+19, TWSM20, CLL21]. E.g., [CLL21] study the role of the projection head, the impact of multi-object images, and a feature suppression phenomenon. [WL21b] analyses the feature learning process to understand the role of augmentations in CL. [RSY+21] finds that an emphasis on instance discrimination can improve representation of some features at the cost of suppressing otherwise well-learned features. [WI20, WL21a] analyse the uniformity of the representations learned with CL. In particular, [WL21a] focuses on the impact of individual negatives and describe a uniformity-tolerance dilemma when choosing the temperature parameter. In this chapter, we rely on the previous findings, expand them to long-tailed data distributions and complement them with an understanding of the emergence of semantic structure.

5.3 Method

In the following, we describe our approach and analysis of contrastive learning on long-tailed data. For this, we will first review the core principles of contrastive learning for the case of uniform data (Section 5.3.1). In Section 5.3.2, we then place a particular focus on the

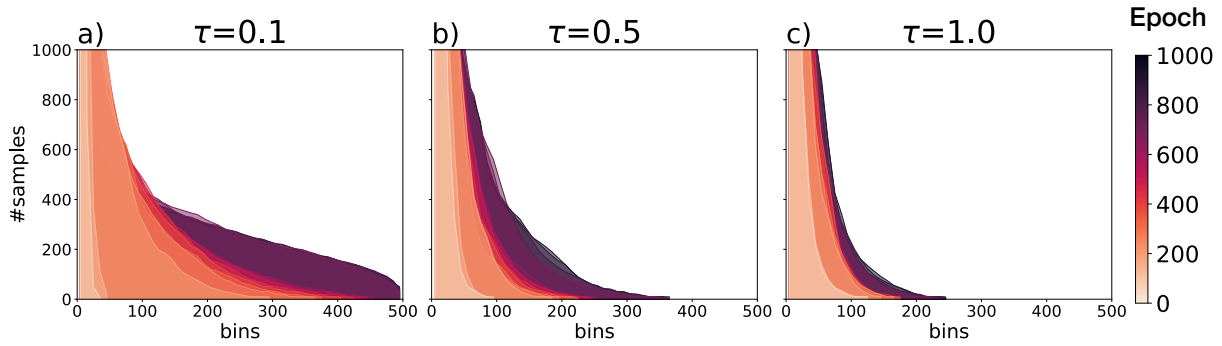


Figure 5.1: **Coverage of the embedding space during training.** To measure coverage we uniformly sample 500 bins on the unit sphere. Each training sample is assigned to the closest bin and we plot a histogram of the assignments. X-axis: bins. Y-axis: number of training samples in a bin. Colors denotes epochs: light is the 1st epoch of training, dark is the last. For small τ (a) the representations are more uniformly distributed (cf. Section 13.3).

temperature parameter τ in the contrastive loss and its impact on the learnt representations. Based on our analysis, in Section 5.3.3 we discuss how the choice of τ might negatively affect the learnt representation of rare classes in the case of long-tailed distributions. Following this, we describe a simple proof-of-concept based on additional coarse supervision to test our hypothesis. We then further develop temperature schedules (TS) that yield significant gains with respect to the separability of the learnt representations in Section 13.4.

5.3.1 Contrastive Learning

The Info-NCE loss is a popular objective for contrastive learning (CL) and has lead to impressive results for learning useful representations from unlabelled data [OLV18, WXSL18, HFW+20, CKNH20]. Given a set of inputs $\{x_1, \ldots, x_N\}$, and the cosine similarities s_{ij} between learnt representations $u_i = f(\mathcal{A}(x_i))$ and $v_j = g(\mathcal{A}(x_j))$ of the inputs, the loss is defined by:

$$\mathcal{L}_c = \sum_{i=1}^{N} -\log \frac{\exp(s_{ii}/\tau)}{\exp(s_{ii}/\tau) + \sum_{j \neq i} \exp(s_{ij}/\tau)}. \tag{5.1}$$

Here, $\mathcal{A}(\cdot)$ applies a random augmentation to its input and f and g are deep neural networks. For a given x_i, we will refer to u_i as the *anchor* and to v_j as a *positive* sample if $i = j$ and as a *negative* if $i \neq j$. Last, τ denotes the *temperature* of the Info-NCE loss and has been found to crucially impact the learnt representations of the model [WI20, WL21a, RSY+21].

Uniformity. Specifically, a small τ has been tied to more uniformly distributed representations, see Figure 5.1. For example, [WL21a] show that the loss is 'hardness-aware', i.e. negative samples closest to the anchor receive the highest gradient. In particular, for a given anchor, the gradient with respect to the negative sample v_j is scaled by its relative contribution to the denominator in Equation (5.1):

$$\frac{\partial \mathcal{L}_c}{\partial v_j} = \frac{\partial \mathcal{L}_c}{\partial s_{ij}} \times \frac{\partial s_{ij}}{\partial v_j} = \frac{1}{\tau} \times [\text{softmax}_k(s_{ik}/\tau)]_j \times \frac{\partial s_{ij}}{\partial v_j} \quad . \tag{5.2}$$

As a result, for sufficiently small τ, the model minimises the cosine similarity to the nearest negatives in the embedding space, as softmax approaches an indicator function that selects

the largest gradient. The optimum of this objective, in turn, is to distribute the embeddings as uniformly as possible over the sphere, as this reduces the average similarity between nearest neighbours, see also Figures 5.1 and 5.3.

Semantic structure. In contrast, a large τ has been observed to induce more semantic structure in the representation space. However, while the effect of small τ has an intuitive explanation, the phenomenon that larger τ induce semantic structure is much more poorly understood and has mostly been described empirically [WL21a, RSY+21]. Specifically, note that for any given positive sample, all negatives are repelled from the anchor, with close-by samples receiving exponentially higher gradients. Nonetheless, for large τ, tightly packed semantic clusters emerge. However, if close-by negatives are heavily repelled, how can this be? Should the loss not be dominated by the hard-negative samples and thus break the semantic structure?

To better understand both phenomena, we propose to view the contrastive loss through the lens of *average distance* maximisation, which we describe in the following section.

5.3.2 Contrastive learning as average distance maximisation

As discussed in the previous section, the parameter τ plays a crucial role in shaping the learning dynamics of contrastive learning. To understand this role better, in this section, we present a novel viewpoint on the mechanics of the contrastive loss that explain the observed model behaviour. In particular, and in contrast to [WL21a] who focused on the impact of *individual* negatives, for this we discuss the *cumulative* impact that all negative samples have on the loss.

To do so, we express the summands $\mathcal{L}_c^i$ of the loss in terms of distances d_{ij} instead of similarities s_{ij}:

$$0 \;\leq\; d_{ij} \;=\; \frac{1 - s_{ij}}{\tau} \;\leq\; \frac{2}{\tau} \quad \text{and} \quad c_{ii} = \exp(d_{ii}). \tag{5.3}$$

This allows us to rewrite the loss $\mathcal{L}_c^i$ as

$$\mathcal{L}_{\mathrm{c}}^i = -\log\left(\frac{\exp\left(-d_{ii}\right)}{\exp\left(-d_{ii}\right) + \sum_{j\neq i}\exp\left(-d_{ij}\right)}\right) = \log\left(1 + c_{ii}\sum_{j\neq i}\exp\left(-d_{ij}\right)\right). \tag{5.4}$$

As the effect c_{ii} of the positive sample for a given anchor is the same for all negatives, in the following we place a particular focus on the negatives and their relative influence on the loss in Equation (5.4); for a discussion of the influence of positive samples, please see Section 5.4.4.

To understand the impact of the temperature τ, first note that the loss monotonically increases with the sum $S_i = \sum_{j\neq i}\exp(-d_{ij})$ of exponential distances in Equation (5.4). As log is a continuous, monotonic function, we base the following discussion on the impact of τ on the sum S_i.

For small τ, the nearest neighbours of the anchor point dominate S_i, as differences in similarity are amplified. As a result, the contrastive objective maximises the average distance to nearest neighbours, leading to a uniform distribution over the hypersphere, see Figure 5.3. Since individual negatives dominate the loss, this argument is consistent with existing interpretations, e.g. [WL21a], as described in the previous section.

For large τ, (e.g. $\tau \geq 1$), on the other hand, the contributions to the loss from a given negative are on the same order of magnitude for a wide range of cosine similarities. Hence, the constrastive objective can be thought of as maximising the average distance over a wider range of neighbours. Interestingly, since distant negatives will typically outnumber close

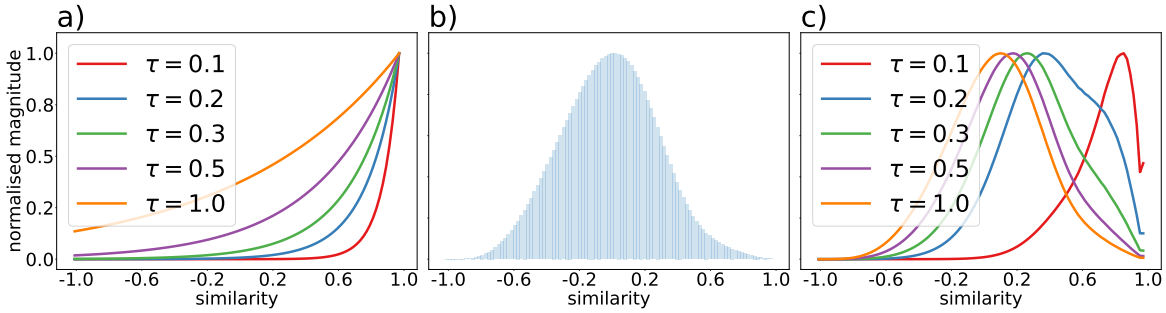


Figure 5.2: **Loss contribution by similarity.** X-axis: cosine similarity between anchor and negative. All curves are normalised such that their max y-value is 1. **a)**: influence of an individual negative sample to the loss depending on its similarity to anchor for different τ; **b)**: average histogram of distribution of negatives over the hypersphere with respect to their similarity to the anchor; **c)**: cumulative impact that negative samples have on the loss. The *cumulative* contribution of negatives shifts left, towards less similar samples, in contrast to individual contributions of negatives. As $\tau \to \infty$, the cumulative distribution coincides with the histogram b).

negatives, the strongest *cumulative* contribution to the contrastive loss will come from more distant samples, despite the fact that *individually* the strongest contributions will come from the closest samples. To visualise this, in Figure 5.2a, we plot the contributions of *individual* samples depending on their distance, as well as the distribution of similarities s_{ij} to negatives over the entire dataset in Figure 5.2b. Since the number of negatives at larger distances (e.g. $s_{ij} \approx 0.1$) significantly outnumber close negatives ($s_{ij} > 0.9$), the peak of the cumulative contributions[1] shifts towards lower similarities for larger τ, as can be seen in Figure 5.2c; in fact, for $\tau \to \infty$, the distribution of cumulative contributions approaches the distribution of negatives.

Hence, the model can significantly decrease the loss by increasing the distance to relatively 'easy negatives' for much longer during training, i.e. to samples that are easily distinguishable from the anchor by simple patterns. Instead of learning 'hard' features that allow for better *instance discrimination* between hard negatives, the model will be biased to learn easy patterns that allow for *group-wise discrimination* and thereby increase the margin between clusters of samples. Note that since the clusters as a whole mutually repel each other, the model is optimised to find a trade-off between the expanding forces between hard negatives (i.e. within a cluster) and the compressing forces that arise due to the margin maximisation between easy negatives (i.e. between clusters).

Importantly, such a bias towards easy features can prevent the models from learning hard features—i.e. by focusing on *group-wise discrimination*, the model becomes agnostic to instance-specific features that would allow for a better *instance discrimination* (cf. [RSY+21]). In the following, we discuss how this might negatively impact rare classes in long-tailed distributions.

5.3.3 Temperature schedules for contrastive learning on long-tail data

As discussed in Section 5.1, naturally occurring data typically exhibit long-tail distributions, with some classes occurring much more frequently than others; across the dataset, *head* classes appear frequently, whereas *tail* classes contain fewest number of samples. Since self-supervised

[1] To obtain the cumulative contributions, we group the negatives into 100 non-overlapping bins of size 0.02 depending on their distance to the anchor and report the sum of contributions of a given bin.

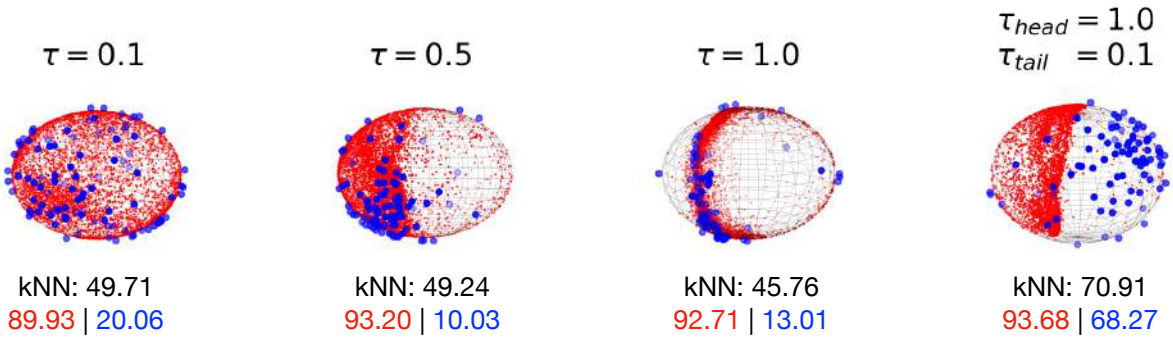


Figure 5.3: **Representations of a head and a tail class.** Visualisation of the influence of τ on representations of two semantically close classes (trained with all 10 classes). Red: single head class and blue: single tail class from CIFAR10-LT. Small τ=0.1 promotes uniformity, while large τ=1.0 creates dense clusters. With $\tau_{\{head/tail\}}$ we refer to coarse supervision described in Section 5.3.3 which separates tail from head classes. In black / red / blue, we respectively show the average kNN accuracy over all classes / the head class / the tail class.

learning methods are designed to learn representations from unlabelled data, it is important to investigate their performance on imbalanced datasets.

Claim: Tail classes benefit from instance discrimination. As discussed in Section 5.3.2, sufficiently large τ are required for semantic groups to emerge during contrastive learning as this emphasises group-wise discrimination. However, as shown by [RSY+21], this can come at the cost of encoding instance-specific features and thus hurt the models' instance discrimination capabilities.

We hypothesise that this disproportionately affects tail classes, as tail classes consist of only relatively few instances to begin with. Their representations should thus *remain distinguishable* from most of their neighbours and not be grouped with other instances, which are likely of a different class. In contrast, since head classes are represented by many samples, grouping those will be advantageous.

To test this hypothesis, we propose to explicitly train head and tail classes with different τ, to emphasise group discrimination for the former while ensuring instance discrimination for the latter.

Experiment: Controlling τ with coarse supervision. We experiment on CIFAR10-LT (a long-tail variant of CIFAR10 - see Section 5.4.1) in which we select a different τ depending on whether the anchor u_i is from a head or a tail class, i.e. of the 5 *most* or *least* common classes. We chose a relatively large τ (τ_{head}=1.0) for the 5 head classes to emphasise group-wise discrimination and a relatively small τ (τ_{tail}=0.1) for the 5 tail classes to encourage the model to learn instance-discriminating features.

As can be seen in Figure 5.3, this simple manipulation of the contrastive loss indeed provides a significant benefit with respect to the semantic structure of the embedding space, despite only weakly supervising the learning by adjusting τ according to a coarse (frequent/infrequent) measure of class frequency.

In particular, in Figure 5.3, we show the projections of a single head class and a single tail class onto the three leading PCA dimensions and the corresponding kNN accuracies. We would like to highlight the following results. First, without any supervision, we indeed find that the head class consistently performs better for larger values of τ (e.g. 1.0), whereas the tail class consistently benefits from smaller values for τ (e.g. 0.1). Second, when training the model

according to the coarse τ supervision as described above, we are not only able to maintain the benefits of large τ values for the head class, but significantly outperform all constant τ versions for the tail class, which improves the overall model performance on all classes; detailed results for all classes are provided in the appendix.

Temperature Schedules (TS) without supervision. Such supervision with respect to the class frequency is, of course, generally not available when training on unlabelled data and these experiments are only designed to test the above claim and provide an intuition about the learning dynamics on long-tail data. However, we would like to point out that the supervision in these experiments is very coarse and only separates the unlabelled data into *frequent* and *infrequent* classes. Nonetheless, while the results are encouraging, they are, of course, based on additional, albeit coarse, labels. Therefore, in what follows, we present an unsupervised method that yields similar benefits.

In detail, we propose to modify τ according to a cosine schedule, such that it alternates between an upper (τ_+) and a lower (τ_-) bound at a fixed period length T:

$$\tau_{\cos}(t) = (\tau_+ - \tau_-) \times (1 + \cos(2\pi\, t/T))/2 + \tau_- \; ; \tag{5.5}$$

here, t denotes training epochs. This method is motivated by the observation that τ controls the trade-off between learning easily separable features and learning instance-specific features.

Arguably, however, the models should learn both types of features: i.e. the representation space should be structured according to easily separable features that (optimally) represent semantically meaningful group-wise patterns, whilst still allowing for instance discrimination within those groups.

Therefore, we propose to *alternate* between both objectives as in Equation (5.5), to ensure that throughout training the model learns to encode instance-specific patterns, whilst also structuring the representation space along semantically meaningful features. Note that while we find a cosine schedule to work best and to be robust with respect to the choice for T (Section 13.4.4), we also evaluate alternatives. Even randomly sampling τ from the interval $[\tau_-, \tau_+]$ improves the model performance. This indicates that the *task switching* between group-wise discrimination (large τ) and instance discrimination (small τ) is indeed the driving factor behind the performance improvements we observe.

5.4 Experiments

In this section, we validate our hypothesis that simple manipulations of the temperature parameter in Equation (5.1) lead to better performance for long-tailed data. First, we introduce our experimental setup in Section 5.4.1, then in Section 5.4.2 we discuss the results across three imbalanced datasets and, finally, we analyse different design choices of the framework through extensive ablation studies in Section 13.4.4.

5.4.1 Implementation Details

Datasets. We consider long-tailed (LT) versions of the following three popular datasets for the experiments: CIFAR10-LT, CIFAR100-LT, and ImageNet100-LT. For most of the experiments, we follow the setting from SDCLR [JCMW21]. In case of **CIFAR10-LT/CIFAR100-LT**, the original datasets [KH09] consist of 60000 32x32 images sampled uniformly from 10 and 100 semantic classes, respectively, where 50000 images correspond to the training set and 10000 to a test set. Long-tail versions of the datasets are introduced by [CJL+19] and consist of a subset

of the original datasets with an exponential decay in the number of images per class. The imbalance ratio controls the uniformity of the dataset and is calculated as the ratio of the sizes of the biggest and the smallest classes. By default, we use an imbalance ratio 100 if not stated otherwise. Experiments in Table 5.1, Table 5.3 are the average over three runs with different permutations of classes. **ImageNet100-LT** is a subset of the original ImageNet-100 [TKI20] consisting of 100 classes for a total of 12.21k 256x256 images. The number of images per class varies from 1280 to 25.

Training. We use an SGD optimizer for all experiments with a weight decay of 1e-4. As for the learning rate, we utilize linear warm-up for 10 epochs that is followed by a cosine annealing schedule starting from 0.5. We train for 2000 epochs for CIFAR10-LT and CIFAR100-LT and 800 epochs for ImageNet100-LT. For CIFAR10-LT and CIFAR100-LT we use a ResNet18 [HZRS16] backbone. For ImageNet100-LT we use a ResNet50 [HZRS16] backbone. For both the MoCo [HFW+20] and the SimCLR [CKNH20] experiments, we follow [JCMW21] and use the following augmentations: resized crop, color jitters, grey scale and horizontal flip. MoCo details: we use a dictionary of size 10000, a projection dimensionality of 128 and a projection head with one linear layer. SimCLR details: we train with a batch size of 512 and a projection head that has two layers with an output size of 128. For evaluation, we discard the projection head and apply l2-normalisation. Regarding the proposed temperature schedules (TS), we use a period length of $T=400$ with $\tau_+=1.0$ and $\tau_-=0.1$ if not stated otherwise.

Evaluation We use k nearest neighbours (kNN) and linear classifiers to assess the learned features. For kNN, we compute $l2$-normalised distances between LT samples from the train set and the class-balanced test set. For each test image, we assign it to the majority class among the top-k closest train images. We report accuracy for kNN with $k=1$ (kNN@1) and with $k=10$ (kNN@10). Compared to fine-tuning or linear probing, kNN directly evaluates the learned embedding since it relies on the learned metric and local structure of the space. We also evaluate the linear separability and generalisation of the space with a linear classifier that we train on the top of frozen backbone. For this, we consider two setups: balanced few-shot linear probing (FS LP) and long-tailed linear probing (LT LP). For FS LP, the few-shot train set is a direct subset of the original long-tailed train set with the shot number equal to the minimum class size in the original LT train set. For LT LP, we use the original LT training set.

5.4.2 Effectiveness of Temperature Schedules

Contrastive learning with TS. In Table 5.1 we present the efficacy of temperature schedules (TS) for two well-known contrastive learning frameworks MoCo [HFW+20] and SimCLR [CKNH20]. We find that both frameworks benefit from varying the temperature and we observe consistent improvements over all evaluation metrics for CIFAR10-LT and CIFAR100-LT, i.e. the local structure of the embedding space (kNN) and the global structure (linear probe) are both improved. Moreover, we show in Table 5.3 that our finding also transfers to ImageNet100-LT. Furthermore, in Table 5.2 we evaluate the performance of the proposed method on the CIFAR10 and CIFAR100 datasets with different imbalance ratios. An imbalance ratio of 50 (imb50) reflects less pronounced imbalance, and imb150 corresponds to the datasets with only 30 (CIFAR10) and 3 (CIFAR100) samples for the smallest class. Varying τ during training improves the performance for different long-tailed data; for a discussion on the dependence of the improvement on the imbalance ratio, please see the appendix.

TS vs SDCLR. Further, we compare our method with SDCLR [JCMW21]. In SDCLR, SimCLR is modified s.t. the embeddings of the online model are contrasted with those of a pruned

method	CIFAR10-LT				CIFAR100-LT			
	kNN@1	kNN@10	FS LP	LT LP	kNN@1	kNN@10	FS LP	LT LP
MoCo	63.54	64.56	69.31	65.11	28.69	28.75	26.86	30.41
MoCo + TS	**64.99**	**65.01**	**72.87**	**66.86**	**30.31**	**29.75**	**28.97**	**32.05**
SimCLR	59.84	60.19	68.29	61.86	28.81	28.12	25.70	31.20
SimCLR + TS	**63.09**	**62.91**	**71.86**	**65.03**	**31.06**	**30.06**	**28.89**	**33.28**

Table 5.1: **Effect of temperature scheduling.** Comparison of MoCo vs MoCo+TS and SimCLR vs SimCLR+TS on CIFAR10-LT and CIFAR100-LT with kNN, few-shot and long-tail linear probe (FS LP and LT LP).

method	CIFAR-10-LT				CIFAR-100-LT			
	imb 50		imb 150		imb 50		imb 150	
	kNN@10	FS LP	kNN@10	FS LP	kNN@10	FS LP	kNN@10	FS LP
MoCo	69.12	74.16	59.13	65.76	32.22	33.53	25.36	22.73
MoCo + TS	**71.49**	**76.37**	**60.83**	**68.59**	**33.24**	**35.03**	**26.75**	**22.78**

Table 5.2: **Effect of imbalance ratio.** MoCo vs MoCo+TS on CIFAR10-LT and CIFAR100-LT for imbalance ratio 50 (imb50) and 150 (imb150). Evaluation metrics: kNN classifier and few-shot linear probe (FS LP).

method	CIFAR-10-LT			CIFAR-100-LT			ImageNet-100-LT		
	kNN@10	FS LP	LS LP	kNN@10	FS LP	LT LP	kNN@10	FS LP	LT LP
SimCLR	60.19	68.29	61.68	28.12	25.70	31.20	38.00	42.64	44.82
SDCLR	60.74	71.03	64.99	29.22	27.28	**34.23**	37.36	42.74	46.40
SimCLR + TS	**62.91**	**71.86**	**65.03**	**30.06**	**28.89**	33.28	**38.86**	**45.18**	**47.26**

Table 5.3: **Comparison with SDCLR.** SimCLR vs SDCLR vs SimCLR+TS on CIFAR10-LT, CIFAR100-LT, and ImageNet100-LT. Evaluation: kNN classifier, few-shot (FS LP) and long-tail linear probe (LT LP).

version of the same model, which is updated after every epoch. Since the pruning is done by simply masking the pruned weights of the original model, SDCLR requires twice as much memory compared to the original SimCLR and extra computational time to prune the model every epoch. In contrast, our method does not require any changes in the architecture or training. In Table 5.3 we show that this simple approach improves not only over the original SimCLR, but also over SDCLR in most metrics.

5.4.3 Ablations

In this section, we evaluate how the hyperparameters in Equation (5.5) can influence the model behaviour.

Cosine Boundaries. First, we vary the lower τ_- and upper τ_+ bounds of τ for the cosine schedule. In Table 5.4 we assess the performance of MoCo+TS with different τ_- and τ_+ on CIFAR10 with FS LP. We observe a clear trend that with a wider range of τ values the performance increases. We attribute this to the ability of the model to learn better 'hard' features with low τ and improve semantic structure for high τ. Note that 0.07 is the value for

$\tau_- \setminus \tau_+$	0.2	0.3	0.4	0.5	1.0
0.07	69.46	68.86	71.29	71.83	**73.26**
0.1	68.17	70.34	71.25	72.31	72.87
0.2	68.89	69.37	70.12	69.65	71.42

Table 5.4: **Influence of cosine boundaries.** Best performance with the largest difference between τ_- and τ_+. CIFAR10 with MoCo+TS, evaluating few-shot linear probes (FS LP).

τ in many current contrastive learning methods.

TS	FS LP
fixed	68.89
step	70.18
rand	70.26
oscil	71.50
cos	**72.31**

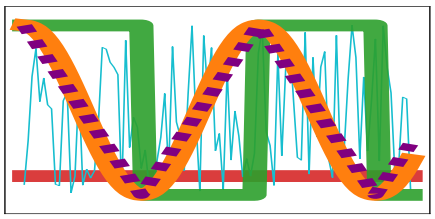


Table 5.5: **Alternative Schedules**. Constant, step function, and random sampling. All functions are bounded by 0.1 and 0.5.

Cosine Period. Further, we investigate if the length of the period T in Equation (5.5) impacts the performance of the model. In Table 5.6, we show that modifying the temperature τ based on the cosine schedule is beneficial during training independently of the period T. The performance varies insignificantly depending on T and consistently improves over standard fixed τ=0.2, whereas the best performance we achieve with T=400. Even though the performance is stable with respect to the length of the period, it changes within one period as we show in Figure 5.4. Here, we average the accuracy of one last full period over different models trained with different T and find that the models reach the best performance around $0.7\,T$. Based on this observation, we recommend to stop training after $(n - 0.3)\,T$ epochs, where n is the number of full periods.

Alternatives to Cosine Schedule. Additionally, we test different methods of varying the temperature parameter τ and report the results in Table 5.5: we examine a linearly oscillating (oscil) function, a step function, and random sampling. For the linear oscillations, we follow the same schedule as for the cosine version, as shown on the right of Table 5.5. For the step function, we change τ from a low (0.1) to a high (0.5) value and back every 200 epochs. For random, we uniformly sample values for τ from the range [0.1, 0.5]. In Table 5.5 we observe that both those methods for varying the τ value also improve the performance over the fixed temperature, while with the cosine schedule the model achieves the best performance. These results indicate that it is indeed the *task switching* between group-wise and instance-wise discrimination during training which is the driving factor for the observed improvements for unsupervised long-tail representation learning. We assume the reason why slow oscillation of the temperature performs better than fast (i.e. random) temperature changes is grounded in learning dynamics and the slow evolution of the embedding space during training.

5.4.4 Influence of the positive samples on contrastive learning

In Section 5.3.2, we particularly focused on the impact of the *negative samples* on the learning dynamics under the contrastive objective, as they likely are the driving factor with respect

T	T / #epochs	FS LP
no	fixed τ	68.89
200	0.1	71.86
400	0.2	72.87
1000	0.5	72.47
2000	1.0	72.22
4000	2.0	72.10

Table 5.6: **Influence of the period length T.** Few-shot linear probe accuracy (FS LP) of MoCo+TS on CIFAR10-LT.

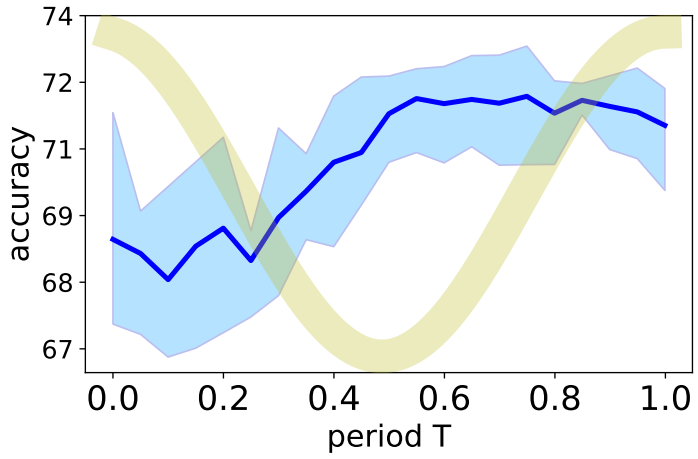


Figure 5.4: **Dependence on relative time of one period.** Blue: Average FS LP of last period of the models trained with $T = 200, 400, 1000, 2000$. Light blue: variance. Yellow: Relative cosine value over relative time. CIFAR10-LT trained with MoCo+TS.

to the semantic structure. In fact, we find that the positive samples should have an inverse relation with the temperature τ and thus cannot explain the observed learning dynamics, as we discuss in the following.

To understand the impact of the *positive samples*, first note their role in the loss (same as Equation (5.4)):

$$\mathcal{L}_{\mathrm{c}}^{i} = \log\left(1 + c_{ii} S_i\right) . \tag{5.6}$$

In particular, c_{ii} scales the entire sum $S_i = \sum_{j \neq i} \exp(-d_{ij})$. As such, encoding two augmentations of the same instance at a large distance is much more 'costly' for the model than encoding two different samples close to each other, as each and every summand S_i is amplified by the corresponding c_{ii}. As a result, the model will be biased to 'err on the safe side' and become invariant to the augmentations, which has been one of the main motivations for introducing augmentations in contrastive learning in the first place, cf. [TSP+20, CKNH20, CMM+20].

Consequently, the positive samples, of course, also influence the forming of clusters in the embedding space as they induce invariance with respect to augmentations. Note, however, that this does not contradict our analysis regarding the impact of negative samples, but rather corroborates it.

In particular, c_{ii} biases the model to become invariant to the applied augmentations for all values of τ; in fact, for small τ, this invariance is even emphasised as c_{ii} increases for small τ

and the influence of the negatives is diminished. Hence, if the augmentations were the main factor in inducing semantic structure in the embedding space, τ should have the opposite effect of the one we and many others [WL21a, ZZP+22, ZWBG21] observe.

Thus, instead of inducing semantic structure on their own, we believe the positive samples to rather play a critical role in influencing which features the model can rely on for grouping samples in the embedding space; for a detailed discussion of this phenomenon, see also [CLL21].

5.5 Conclusion

In this chapter, we discover the surprising effectiveness of temperature schedules for self-supervised contrastive representation learning on imbalanced datasets. In particular, we find that a simple cosine schedule for τ consistently improves two state-of-the-art contrastive methods over several datasets and different imbalance ratios, without introducing any additional cost. Importantly, our approach is based on a novel perspective on the contrastive loss, in which the average distance maximisation aspect is emphasised. This perspective sheds light on which samples dominate the contrastive loss and explains why large values for τ can lead to the emergence of tight clusters in the embedding space, despite the fact that individual instance always repel each other. Specifically, we find that while a large τ is thus necessary to induce semantic structure, the concomitant focus on group-wise discrimination biases the model to encode easily separable features rather than instance-specific details. However, in long-tailed distributions, this can be particularly harmful to the most infrequent classes, as those require a higher degree of instance discrimination to remain distinguishable from the prevalent semantic categories. The proposed cosine schedule for τ overcomes this tension, by alternating between an emphasis on instance discrimination (small τ) and group-wise discrimination (large τ). As a result of this constant 'task switching', the model is trained to both structure the embedding space according to semantically meaningful features, whilst also encoding instance-specific details such that rare classes remain distinguishable from dominant ones.

In this part, Part I, of the thesis, we explore targeted representations tailored for specific downstream tasks. In particular, in Chapter 3, we delve into temporal segmentation of procedural videos, while in Chapter 4, we focus on recognizing unintentional actions. We demonstrate how to utilize on task-specific inherent biases and also examine how specific hyperparameters within the general self-supervised framework can influence learning dynamics. Up to this point, our focus has been on leveraging data without explicit supervision, relying instead on prior task-related information. In the next part, Part II, we will discuss methods that leverage some form of supervision.

II

Learning with Limited Supervision

While the previous part of the thesis focuses on learning efficient representations without labels, this part considers the learning scenario with limited supervision. This scenario includes weak annotations across multiple modalities or relying on limited number of samples along with unlabeled data.

In Chapter 6, we investigate a semi-supervised learning scenario in which only a subset of annotated data is available alongside unlabeled data. Upon revisiting the popular consistency regularization concept, we observe that not only invariance but also equivariance can enhance performance.

In Chapter 7, we introduce a challenging setup utilizing uncurated & unpaired data to mitigate annotation costs for text-video retrieval. Additionally, we present the In-Style method, enabling the transfer of text query styles to large-scale uncurated web video data to allow low-cost efficient pretraining.

In Chapter 8, we further exploit web data by utilizing label-free web images as a source for adapting to unlabeled target videos. To efficiently transfer knowledge from images to videos, we decouple the spatial and temporal modalities in a cycle-based approach, alternating training between them.

In Chapter 9, we present a pipeline for automatically annotating large-scale web videos, leveraging the capabilities of large language models, thereby enabling zero-cost annotations at scale. To achieve this, we introduce a prompting method capable of accommodating long subtitles and explicitly aligning captions with the videos.

6 Revisiting Consistency Regularization for Semi-Supervised Learning

Contents

To mitigate the requirement for annotating each sample in the training set, we begin this part by delving into the semi-supervised setting. This scenario entails having only a limited number of labels available alongside a vast collection of unlabeled samples. Consistency regularization is one of the most widely-used techniques for semi-supervised learning (SSL). Generally, the aim is to train a model that is invariant to various data augmentations. In this chapter, we revisit this idea and find that enforcing invariance by decreasing distances between features from differently augmented images leads to improved performance. However, encouraging equivariance instead, by increasing the feature distance, further improves performance. To this end, we propose an improved consistency regularization framework by a simple yet effective technique, FeatDistLoss, that imposes consistency and equivariance on the classifier and the feature level, respectively. Experimental results show that our model defines a new state of the art across a variety of standard semi-supervised learning benchmarks as well as imbalanced semi-supervised learning benchmarks. Particularly, we outperform previous work by a significant margin in low data regimes and at large imbalance ratios. Extensive experiments are conducted to analyze the method, and the code will be published.

This chapter is based on [FKS21, FKDS23a]. As a second author, Anna Kukleva contributed to the project through detailed discussions, writing of the paper and creating the figures.

6.1 Introduction

Deep learning requires large-scale and annotated datasets to reach state-of-the-art performance [RDS+15, LMB+14]. As labels are not always available or expensive to acquire a wide range of semi-supervised learning (SSL) methods have been proposed to leverage unlabeled data [TV17, LA17, MMKI18, VLK+19, BCG+19, SBL+20, XDH+20, BCC+20, AOA+20, Lee13, PXDL20,

FOS20, BHB19, CKS+20].

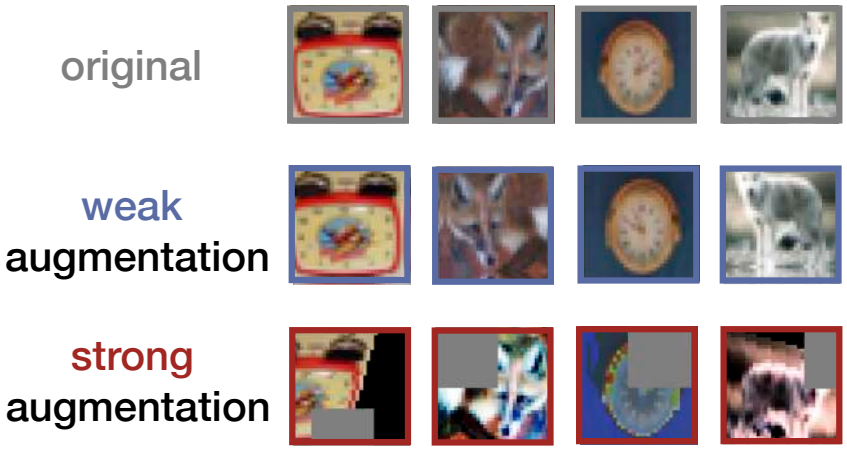


Figure 6.1: Examples of strongly and weakly augmented images from CIFAR-100 (please refer to Section 6.3.3 for details of strong and weak augmentation). The visually large difference between them indicates that it can be more beneficial if they are treated differently.

Consistency regularization [BAP14, LA17, SJT16] is one of the most widely-used SSL methods. Recent work [SBL+20, XDH+20, KMHK20] achieves strong performance by utilizing unlabeled data in a way that model predictions should be invariant to input perturbations. However, when using advanced and strong data augmentation schemes, we question if the model should be invariant to such strong perturbations. In Figure 6.1 we illustrate that strong data augmentation leads to perceptually highly diverse images. Thus, we argue that improving equivariance on such strongly augmented images can provide even better performance rather than making the model invariant to all kinds of augmentations. Moreover, existing works apply consistency regularization either at the feature level or at the classifier level. We find empirically that it is more beneficial to introduce consistency on both levels. To this end, we propose a simple yet effective technique, Feature Distance Loss (FeatDistLoss), to improve data-augmentation-based consistency regularization.

We formulate our FeatDistLoss as to explicitly encourage invariance or equivariance between features from different augmentations while enforcing the same semantic class label. Figure 6.2 shows the intuition behind the idea. Specifically, encouragement of equivariance for the same image but different augmentations (increase distance between stars and circles of the same color) pushes representations apart from each other, thus, covering more space for the class. Imposing invariance, on the contrary, makes the representations of the same semantic class more compact. In this work, we empirically find that increasing equivariance to differently augmented versions of the same image can lead to better performance especially when rather few labels are available per class (see section 6.4.3).

This chapter introduces the method *CR-Match* which combines FeatDistLoss with other strong techniques defining a new state-of-the-art across a wide range of settings of standard SSL benchmarks, including CIFAR-10, CIFAR-100, SVHN, STL-10, and Mini-Imagenet. More specifically, our contribution is fourfold. (1) We improve data-augmentation-based consistency regularization by a simple yet effective technique for SSL called *FeatDistLoss* which regularizes the distance between feature representations from differently augmented images of the same class as well as the classifier simultaneously. (2) We show that while encouraging invariance results in good performance, encouraging equivariance to differently augmented versions of the same image consistently results in even better generalization performance. (3) We provide

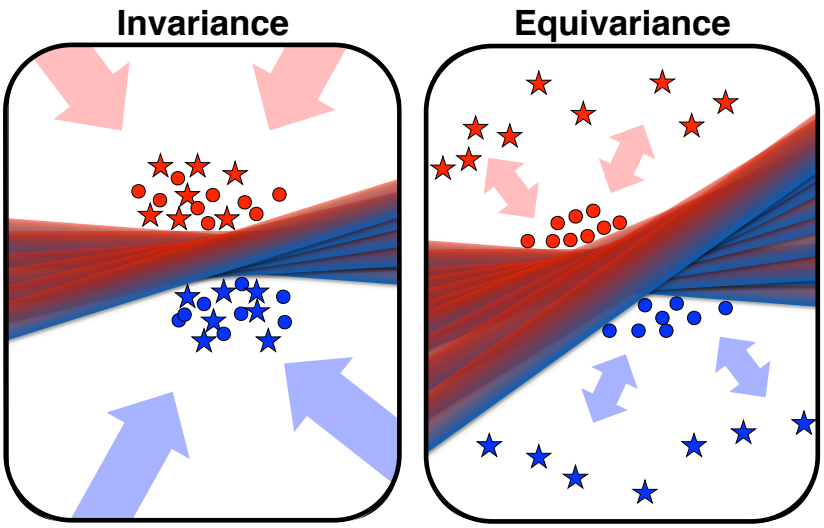


Figure 6.2: Binary classification task. Stars are features of strongly augmented images and circles are of weakly augmented images (please refer to Section 6.3.3 for details of strong and weak augmentation). While encouraging invariance by decreasing distance between features from differently augmented images gives good performance (left), encouraging equivariant representations by increasing the distance regularizes the feature space more, leading to even better generalization performance.

comprehensive ablation studies on different distance functions and different augmentations with respect to the proposed FeatDistLoss. (4) In combination with other strong techniques, we achieve *new state-of-the-art results* on most standard semi-supervised learning benchmarks as well as imbalanced semi-supervised learning benchmarks. In particular, our method outperforms previous methods by a significant margin in low data regimes and at large imbalance ratios.

A preliminary version of this work has been published in [FKS21]. In this work, we extend [FKS21] in three aspects: (1) We extend the existing standard SSL settings by providing evaluations on wider range of the datasets and showing the benefit of the proposed technique on top of various SSL methods. In particular, combining with the recently published method FlexMatch [ZWH$^+$21], we can push the state-of-the-art even further under the standard settings. Moreover, we evaluate our method on ImageNet to verify that the method scales to larger datasets as well. (2) We evaluate our methods under a more realistic and challenging setting: imbalanced SSL, where the training data is not only partially annotated but also exhibits long-tailed class distribution. We achieve new state-of-the-art results on multiple imbalanced SSL benchmarks across a wide range of settings. (3) To give more in-depth insight into our method, we provide pseudo-code and more analysis of the method, especially the robustness against important hyper-parameters.

6.2 Related Work

In this section, we discuss prior work on the equivariant representations. We will not revisit the topic of semi-supervised learning as previously discussed in Chapter 2.

Equivariant representations are recently explored by capsule networks [SFH17, HSF18]. They replaced max-pooling layers with convolutional strides and dynamic routing to preserve

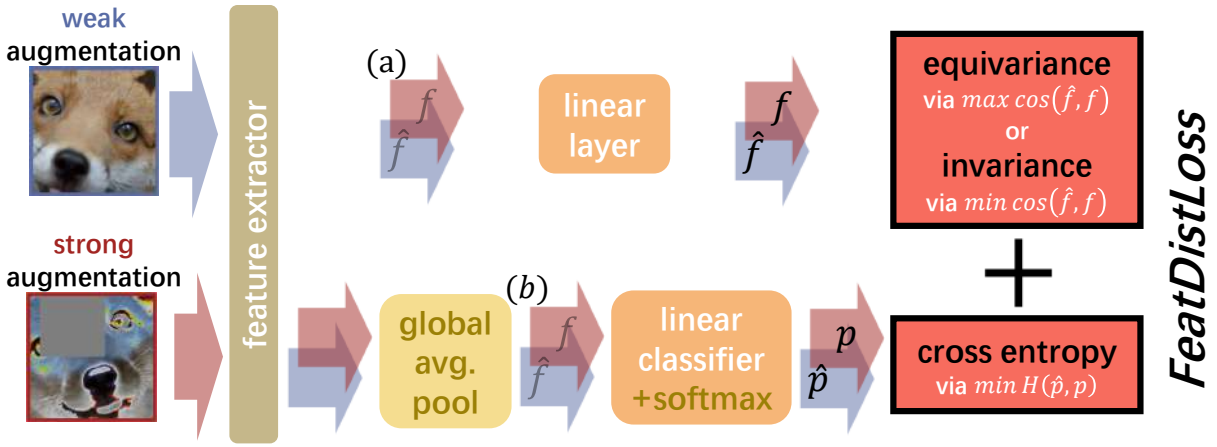


Figure 6.3: The proposed FeatDistLoss utilizes unlabeled images in two ways: On the classifier level, different versions of the same image should generate the same class label, whereas on the feature level, representations are encouraged to become either more equivariant (pushing away) or invariant (pulling together). f and $\hat{f}$ denote strong and weak features; p and $\hat{p}$ are predicted class distributions from strong and weak features; a) and b) denote features before and after the global average pooling layer. Our final model takes features from a) and encourages equivariance to differently augmented versions of the same image.

more information about the input, allowing for preservation of part-whole relationships in the data. It has been shown, that the input can be reconstructed from the output capsule vectors. Another stream of work on group equivariant networks [CW16a, WC19, CW16b] explores various equivariant architectures that produce transform in a predictable linear manner under transformations of the input. Different from previous work, our work explores equivariant representations in the sense that differently augmented versions of the same image are represented by different points in the feature space despite the same semantic label.

6.3 Method

Consistency regularization is highly-successful and widely-adopted technique in SSL [BAP14, LA17, SJT16, SBL+20, XDH+20, KMHK20]. In this work, we aim to leverage and improve it by even further regularizing the feature space. To this end, we present a simple yet effective technique FeatDistLoss to explicitly regularize representation learning and classifier learning at the same time. We describe our SSL method, called CR-Match, which shows improved performance across many different settings, especially in scenarios with few labels. In this section, we first describe our technique FeatDistLoss and then present CR-Match that combines FeatDistLoss with other regularization techniques inspired from the literature.

6.3.1 Feature Distance Loss

Background: The idea of consistency regularization [BAP14, LA17, SJT16] is to encourage the model predictions to be invariant to input perturbations. Given a batch of n unlabeled images

$\mathbf{u}_i, i \in (1, ..., n)$, consistency regularization can be formulated as the following loss function:

$$\frac{1}{n}\sum_{i=1}^{n} \|f(\mathcal{A}(\mathbf{u}_i)) - f(\alpha(\mathbf{u}_i))\|_2^2 \tag{6.1}$$

where f is an encoder network that maps an input image to a d-dimensional feature space; $\mathcal{A}$ and α are two stochastic functions which are, in our case, strong and weak augmentations, respectively (details in Section 6.3.3). By minimizing the L_2 distance between perturbed images, the representation is therefore encouraged to become more invariant with respect to different augmentations, which helps generalization. The intuition behind this is that a good model should be robust to data augmentations of the images.

FeatDistLoss: As shown in Figure 6.3, we extend the above consistency regularization idea by introducing consistency on the classifier level and invariance or equivariance on the feature level. FeatDistLoss thus allows to apply different types of control for these levels. In particular, when encouraging to reduce the feature distance, it becomes similar to classic consistency regularization, and encourages invariance between differently augmented images. As argued above, making the model predictions invariant to input perturbations gives good generalization performance. Instead, in this work we find it is more beneficial to treat images from different augmentations differently because some distorted images are largely different from their original images as demonstrated visually in Figure 6.1. Therefore, the final model (CR-Match) uses FeatDistLoss to increase the distance between image features from augmentations of different intensities while at the same time enforcing the same semantic label for them. Note that in Section 6.4.3, we conduct an ablation study on the choice of distance function, where we denote CR-Match as CR-Equiv, and the model that encourages invariance as CR-Inv.

The final objective for the FeatDistLoss consists of two terms: $\mathcal{L}_{Dist}$ (on the feature level), that explicitly regularizes feature distances between embeddings, and a standard cross-entropy loss $\mathcal{L}_{PseudoLabel}$ (on the classifier level) based on pseudo-labeling.

With $\mathcal{L}_{Dist}$ we either decrease or increase the feature distance between weakly and strongly augmented versions of the same image in a low-dimensional space projected from the original feature space to overcome the curse of dimensionality [Bel66]. Let $d(\cdot, \cdot)$ be a distance metric and z be a linear layer that maps the high-dimensional feature into a low-dimensional space. Given an unlabeled image $\mathbf{u}_i$, we first extract features with strong and weak augmentations by $f(\mathcal{A}(\mathbf{u}_i))$ and $f(\alpha(\mathbf{u}_i))$ as shown in Figure 6.3 (a), and then FeatDistLoss is computed as:

$$\mathcal{L}_{Dist}(\mathbf{u}_i) = d(z(f(\mathcal{A}(\mathbf{u}_i))), z(f(\alpha(\mathbf{u}_i)))) \tag{6.2}$$

Different choices of performing $\mathcal{L}_{Dist}$ are studied in Section 6.4.3, where we find empirically that applying $\mathcal{L}_{Dist}$ at (a) using cosine distance in Figure 6.3 gives the best performance. The use of the projection head z does not only reduce the computation burden as the original feature space is high-dimensional, but also brings additional performance improvements as shown in [CKNH20, CKS+20].

At the same time, images from strong and weak augmentations should have the same class label because they are essentially generated from the same original image. Inspired by [SBL+20], given an unlabeled image $\mathbf{u}_i$, a pseudo-label distribution is first generated from the weakly augmented image by $\hat{\mathbf{p}}_i = g(f(\alpha(\mathbf{u}_i)))$, and then a cross-entropy loss is computed between the pseudo-label and the prediction for the corresponding strongly augmented version as:

$$\mathcal{L}_{PseudoLabel}(\mathbf{u}_i) = \ell_{CE}(\hat{\mathbf{p}}_i, g(f(\mathcal{A}(\mathbf{u}_i)))) \tag{6.3}$$

where ℓ_{CE} is the cross-entropy, g is a linear classifier that maps a feature representation to a class distribution, and $\mathcal{A}(\mathbf{u}_i)$ denotes the operator for strong augmentations.

Putting it all together, FeatDistLoss processes a batch of unlabeled data $\mathbf{u}_i, i \in (1, ..., B_u)$ with the following loss:

$$\mathcal{L}_U = \frac{1}{B_u} \sum_{i=1}^{B_u} \mathbb{1}\{c_i > \tau\}(\mathcal{L}_{Dist}(\mathbf{u}_i) + \mathcal{L}_{PseudoLabel}(\mathbf{u}_i)) \tag{6.4}$$

where $c_i = max\ \hat{\mathbf{p}}_i$ is the confidence score, and $\mathbb{1}\{\cdot\}$ is the indicator function which outputs 1 when the confidence score is above a threshold. This confidence thresholding mechanism ensures that the loss is only computed for unlabeled images for which the model generates a high-confidence prediction. Therefore, it controls the trade-off between the quality and the quantity of contributing unlabeled samples. As is shown in Section 6.4.2, a higher threshold τ is normally preferred because it alleviates the instability early in the training by eliminating less confident unlabeled samples. As training progresses, the model produces more confident predictions and more samples will contribute to the final loss, which also provides a natural curriculum to balance labeled and unlabeled losses [SBL+20]. Moreover, the thresholding mechanism is applied for both the feature level consistency and the classifier level consistency so that the two losses are well-synchronized.

As mentioned before, depending on the function d, FeatDistLoss can decrease the distance between features from different data augmentation schemes (when d is a distance function, thus pulling the representations together), or increase it (when d is a similarity function, thus pushing the representations apart). As shown in Table 6.6, we find that both cases results in an improved performance. However, increasing the distance between weakly and strongly augmented examples consistently results in better generalization performance. We conjecture that the reason lies in the fact that FeatDistLoss by increasing the feature distance explores equivariance properties (differently augmented versions of the same image having distinct features but the same label) of the representations. It encourages the model to have more distinct weakly and strongly augmented images while still imposing the same label, which leads to both more expressive representation and more powerful classifier. As we will show in Section 6.4.3, information like object location or orientation is more predictable from models trained with FeatDistLoss that pushes the representations apart. Additional ablation studies of other design choices such as the distance function and the linear projection z are also provided in Section 6.4.3.

6.3.2 Overall CR-Match

Now we describe our SSL method called CR-Match leveraging the above FeatDistLoss. Pseudo-code for processing a batch of labeled and unlabeled examples is shown in algorithm 1.

Given a batch of labeled images with their labels as $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i) : i \in (1, ..., B_s)\}$ and a batch of unlabeled images as $\mathcal{U} = \{\mathbf{u}_i : i \in (1, ..., B_u)\}$. [1] CR-Match minimizes the following learning objective:

$$\mathcal{L}_S(\mathcal{X}) + \lambda_u \mathcal{L}_U(\mathcal{U}) + \lambda_r \mathcal{L}_{Rot}(\mathcal{U}) \tag{6.5}$$

where $\mathcal{L}_S$ is the supervised cross-entropy loss for labeled images with weak data augmentation regularization; $\mathcal{L}_U$ is our novel feature distance loss for unlabeled images which explicitly regularizes the distance between weakly and strongly augmented images in the feature space; and $\mathcal{L}_{Rot}$ is a self-supervised loss for unlabeled images and stands for rotation prediction from [GSK18] to provide an additional supervisory and regularizing signal.

[1] In practice, unlabeled data includes all labeled data without labels.

Algorithm 1

Require: Labeled batch $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i) : i \in (1, \ldots, B_s)\}$, unlabeled batch $\mathcal{U} = \{\mathbf{u}_i : i \in (1, \ldots, B_u)\}$, confidence threshold τ, FeatDistLoss weight λ_u, rotation prediction loss weight λ_r, classifier g, distance metric d, FeatDistLoss head z, rotation prediction head h.

▷ *Cross-entropy loss for labeled data*
$\mathcal{L}_S = \frac{1}{B_s} \sum_{i=1}^{B_s} \ell_{CE}(\mathbf{p}_i, g(\alpha(\mathbf{x}_i)))$
for $i = 1$ B_u **do**
    ▷ *Extract representation from weak data augmentation*
    $\mathbf{u}_i^w = f(\alpha(\mathbf{u}_i))$
    ▷ *Extract representation from strong data augmentation*
    $\mathbf{u}_i^s = f(\mathcal{A}(\mathbf{u}_i))$
    ▷ *Compute confidence score from the weakly augmented image*
    $c_i = max\ g(\mathbf{u}_i^w)$
end for
▷ *Cross-entropy loss with pseudo-label for unlabeled data*
$\mathcal{L}_{Pseudo} = \frac{1}{B_u} \sum_{i=1}^{B_u} \mathbb{1}\{c_i > \tau\}\ \ell_{CE}(g(\mathbf{u}_i^w), \mathbf{u}_i^s)$
▷ *Increase the feature distance for unlabeled data*
$\mathcal{L}_{Dist} = \frac{1}{B_u} \sum_{i=1}^{B_u} \mathbb{1}\{c_i > \tau\}\ -d(z(\mathbf{u}_i^w), z(\mathbf{u}_i^s))$
▷ rotation prediction loss
$\mathcal{L}_{Rot} = \frac{1}{4B_u} \sum_{i=1}^{B_u} \sum_{r \in \mathbb{R}} \ell_{CE}(r, h(R(\mathbf{u}_i^w, r)))$ **return** $\mathcal{L}_S + \lambda_u(\mathcal{L}_{Pseudo} + \mathcal{L}_{Dist}) + \lambda_r \mathcal{L}_{Rot}$

Fully supervised loss for labeled data: We use cross-entropy loss with weak data augmentation regularization for labeled data:

$$\mathcal{L}_S = \frac{1}{B_s} \sum_{i=1}^{B_s} \ell_{CE}(\mathbf{p}_i, g(f(\alpha(\mathbf{x}_i)))) \tag{6.6}$$

where ℓ_{CE} is the cross-entropy loss, $\alpha(\mathbf{x}_i)$ is the extracted feature from a weakly augmented image $\mathbf{x}_i$, g is the same linear classifier as in equation 6.2, and $\mathbf{p}_i$ is the corresponding label for $\mathbf{x}_i$.

Self-supervised loss for unlabeled data: Rotation prediction [GSK18] (RotNet) is one of the most successful self-supervised learning methods, and has been shown to be complementary to SSL methods [ZOKB19, BCG+19, REH+20]. Here, we create four rotated images by 0°, 90°, 180°, and 270° for each unlabeled image $\mathbf{u}_i$ for $i \in (1, ..., \mu B)$. Then, classification loss is applied to train the model predicting the rotation as a four-class classification task:

$$\mathcal{L}_{Rot} = \frac{1}{4B_u} \sum_{i=1}^{B_u} \sum_{r \in \mathbb{R}} \ell_{CE}(r, h(\alpha(R(\mathbf{u}_i, r)))) \tag{6.7}$$

where $\mathbb{R}$ is $\{0°, 90°, 180°, 270°\}$ and r refers to one of the four rotations, h denotes a three-layer MLP with its hidden dimension the same as the input dimension. Using a predictor head is shown to be beneficial for such an auxiliary loss [CKNH20, CKS+20]. Note that rotation prediction, though commonly used, might also have adverse effects. For example, numbers six and nine in most print fonts are centrosymmetric, rotating one upside down gives the other.

6.3.3 Implementation Details

Data augmentation: As mentioned above, CR-Match adopts two types of data augmentations: weak augmentation and strong augmentation from [SBL+20]. Specifically, the weak augmentation α corresponds to a standard random cropping and random mirroring with probability 0.5, and the strong augmentation $\mathcal{A}$ is a combination of RandAugment [CZSL20] and CutOut [DT17]. At each training step, we uniformly sample two operations for the strong augmentation from a collection of transformations and apply them with a randomly sampled magnitude from a predefined range. The complete table of transformation operations for the strong augmentation is provided in the supplementary material.

Other implementation details: For our results in Section 6.4 and Section 6.5, we minimize the cosine similarity in FeatDistLoss, and use a fully-connected layer for the projection layer z, which maps the feature from the original un-flattened 8192-dimension space into a 128-dimension space, the same dimension as the feature dimension for classification. The dimension of the original feature space and the patch size are fixed and depend on the architecture, which is chosen following the previous conventions [OOR+18, BCG+19, BCC+20, SBL+20]. In our case, $8192 = 8 \times 8 \times 128$, where the patch size is 8×8, and there are 128 feature maps. The predictor head h in rotation prediction loss consists of two fully-connected layers and a ReLU as non-linearity. We use the same $\lambda_u = \lambda_r = 1$ in all experiments since CR-Match shows good robustness within a range of loss weights in our preliminary experiments. We train our model for 512 epochs on CIFAR-10, CIFAR-100, and SVHN. On STL-10 and Mini-ImageNet, we train the model for 300 epochs. Other hyper-parameters are from [SBL+20] for the compatibility. Specifically, the confidence thresholds τ for pseudo-label selection is 0.95. We use SGD with momentum 0.9 and cosine learning rate schedule from [SBL+20] starting from 0.03, batch size B_s is 64 for labeled data, and B_u is $7 \times B_s$. The final performance is reported using an exponential moving average of model parameters as recommended by [TV17]. As a common practice, we repeat each experiment with five different data splits and report the mean and the standard deviation of the error rate.

6.4 Experiments

Following protocols from previous work [BCG+19, SBL+20], we conduct experiments on several commonly used SSL image classification benchmarks to test the efficacy of CR-Match. We show our main results in Section 6.4.1, where we achieve state-of-the-art error rates across all settings on SVHN [NWC+11], CIFAR-10 [KH09], CIFAR-100 [KH09], STL-10 [CNL11], and mini-ImageNet [RL16]. In our ablation study in Section 6.4.2 we analyze the effect of FeatDistLoss and RotNet across different settings. Finally, in Section 6.4.3 we extensively analyse various design choices for our FeatDistLoss.

6.4.1 Main Results

In the following, each dataset subsection includes two paragraphs. The first provides technical details and the second discusses experimental results.

CIFAR-10, CIFAR-100, and SVHN. We follow prior work [SBL+20] and use 4, 25, and 100 labels per class on CIFAR-100 and SVHN without extra data. For CIFAR-10, we experiment with settings of 4, 25, and 400 labels per class. We create labeled data by random sampling, and the remaining images are regarded as unlabeled by discarding their labels. Following [BCG+19,

	CIFAR-10			CIFAR-100		
Per class labels	4 labels	25 labels	400 labels	4 labels	25 labels	100 labels
Mean Teacher [TV17]	-	32.32±2.30*	9.19±0.19*	-	53.91±0.57*	35.83±0.24*
MixMatch [BCG+19]	47.54±11.50*	11.08±0.87	6.24±0.06	67.61±1.32*	39.94±0.37*	25.88±0.30
UDA [XDH+20]	29.05±5.93*	5.43±0.96	4.32±0.08*	59.28±0.88*	33.13±0.22*	24.50±0.25*
ReMixMatch [BCC+20]	19.10±9.64*	6.27±0.34	5.14±0.04	44.28±2.06*	27.43±0.31*	23.03±0.56*
FixMatch (RA) [SBL+20]	13.81±3.37	*5.07*±0.65	*4.26*±0.05	48.85±1.75	28.29±0.11	22.60±0.12
FixMatch (CTA) [SBL+20]	11.39±3.35	*5.07*±0.33	4.31±0.15	49.95±3.01	28.64±0.24	23.18±0.11
FeatMatch [KMHK20]	-	6.00±0.41	4.64±0.11	-	-	-
FlexMatch [ZWH+21]	**5.19**±0.05	5.33±0.12	4.47±0.09	45.91±1.76	28.11±0.20	23.04±0.28
CR-Match	10.70±2.91	**5.05**±0.12	**3.96**±0.16	*39.45*±1.69	*25.43*±0.14	**20.40**±0.08
CR-Match§	*5.52*±0.32	5.21±0.06	*4.26*±0.19	**35.72**±0.50	**24.61**±0.37	*20.91*±0.24

Table 6.1: Error rates on CIFAR-10, and CIFAR-100. A Wide ResNet-28-2 [ZK16] is used for CIFAR-10 and a Wide ResNet-28-8 with 135 filters per layer [BCG+19] is used for CIFAR-100. We use the same code base as [SBL+20] (i.e., same network architecture and training protocol) to make the results directly comparable. The best number is in bold and the second best number is in italic. *Numbers are generated by [SBL+20]. CR-Match§ refers to CR-Match combined with CPL [ZWH+21] from FlexMatch.

	STL-10	SVHN		
Per class labels	100 labels	4 labels	25 labels	100 labels
Mean Teacher [TV17]	21.34±2.39*	-	3.57±0.11*	3.42±0.07*
MixMatch [BCG+19]	10.18±1.46	42.55±14.53*	3.78±0.26	3.27±0.31
UDA [XDH+20]	7.66±0.56*	52.63±20.51*	2.72±0.40	*2.23*±0.07
ReMixMatch [BCC+20]	6.18±1.24	*3.34*±0.20*	3.10±0.50	2.83±0.30
FixMatch (RA) [SBL+20]	7.98±1.50	3.96±2.17	*2.48*±0.38	2.28±0.11
FixMatch (CTA) [SBL+20]	*5.17*±0.63	7.65±7.65	2.64±0.64	2.36±0.19
FeatMatch [KMHK20]	-	-	3.34±0.19†	3.10±0.06†
FlexMatch [ZWH+21]	6.15±0.25	20.81±5.26	17.32±2.07	12.90±2.68
CR-Match	**4.89**±0.17	**2.79**±0.93	**2.35**±0.29	**2.08**±0.07

Table 6.2: Error rates on STL-10 and SVHN. A Wide ResNet-28-2 and a Wide ResNet-37-2 [ZK16] is used for SVHN and STL-10, repectively. The same code base is adopted as [SBL+20] to make the results directly comparable. Notations follow Table 6.1.

	Mini-ImageNet	
Per class labels	40 labels	100 labels
Mean Teacher [TV17]	72.51±0.22	57.55±1.11
Label Propagation [ITAC19]	70.29±0.81	57.58±1.47
PLCB [AOA+20]	56.49±0.51	46.08±0.11
FeatMatch [KMHK20]	*39.05*±0.06	*34.79*±0.22
CR-Match	**34.87**±0.99	**32.58**±1.60

Table 6.3: Error rates on Mini-ImageNet with 40 labels and 100 labels per class. All methods are evaluated on the same ResNet-18 architecture. *Numbers are generated by [SBL+20]. †Numbers are produced without CutOut. The best number is in bold and the second best number is in italic. Notations follow Table 6.1.

SBL+20, BCC+20], we use a Wide ResNet-28-2 [ZK16] with 1.5M parameters on CIFAR-10 and SVHN, and a Wide ResNet-28-8 with 135 filters per layer (26M parameters) on CIFAR-100.

As shown in Table 6.1, Table 6.2, and Table 6.3 our method improves over previous methods across all settings, and defines a new state-of-the-art. Most importantly, we improve error rates in low data regimes by a large margin (e.g., with 4 labeled examples per class on CIFAR-100, we outperform FlexMatch and the second best method by 10.19% and 8.56% in absolute value respectively). Prior works [SBL+20, BCG+19, BCC+20] have reported results using a larger network architecture on CIFAR-100 to obtain better performance. On the contrary, we additionally evaluate our method on the small network used in CIFAR-10 and find that our method is more than 17 times ($17 \approx 26/1.5$) parameter-efficient than FixMatch. We reach 46.05% error rate on CIFAR-100 with 4 labels per class using the small model, which is still slightly better than the result of FixMatch using a larger model.

STL-10. STL-10 contains 5,000 labeled images of size 96-by-96 from 10 classes and 100,000 unlabeled images. The dataset pre-defines ten folds of 1,000 labeled examples from the training data, and we evaluate our method on five of these ten folds as in [SBL+20, BCC+20]. Following [BCG+19], we use the same Wide ResNet-37-2 model (comprising 5.9M parameters), and report error rates in Table 6.2.

Our method achieves state-of-the-art performance with 4.89% error rate. Note that FixMatch with error rate 5.17% used the more advanced CTAugment [BCC+20], which learns augmentation policies alongside model training. When evaluated with the same data augmentation (RandAugment) as we use in CR-Match, our result surpasses FixMatch by 3.09% (3.09%=7.98%-4.89%), which indicates that CR-Match itself induces a strong regularization effect.

Mini-ImageNet. We follow [ITAC19, AOA+20, KMHK20] to construct the mini-ImageNet training set. Specifically, 50,000 training examples and 10,000 test examples are randomly selected for a predefined list of 100 classes [RL16] from ILSVRC [DDS+09]. Following [KMHK20], we use a ResNet-18 network [HZRS16] as our model and experiment with settings of 40 labels per class and 100 labels per class.

As shown in Table 6.3, our method consistently improves over previous methods and achieves a new state-of-the-art in both the 40-label and 100-label settings. Especially in the 40-label case, CR-Match achieves an error rate of 34.87% which is 4.18% higher than the second best result. Note that our method is 2 times more data efficient than the second best method FeatMatch [KMHK20] (FeatMatch, using 100 labels per class, reaches a similar error rate as our method with 40 labeled examples per class).

ImageNet. To verify the effectiveness of our method on large scale datasets, we conduct experiments on ImageNet-1k. Following [ZWH+21], we take ∼10% (100,000) training images as the labeled set and construct unlabeled set using the rest of the images. The validation setting remains the same. We train a ResNet-50 [HZRS16] with the same hyper-parameters from [ZWH+21]. Note that FixMatch and FlexMatch use different protocols on ImageNet, and we follow the setup from FlexMatch therefore the numbers are directly comparable.

Table 6.4 shows the error rate comparison after running 2^{20} iterations. Our method outperforms the previous state-of-the-art by 1.04% absolute top-5 error rate, which demonstrates the efficacy of the proposed method at large scale dataset.

6.4.2 Ablation Study

In this section, we analyze how FeatDistLoss and RotNet influence the performance across different settings, particularly when there are few labeled samples. We conduct experiments

Method	Top-1	Top-5
FixMatch [SBL+20]	43.66*	21.80*
FlexMatch [ZWH+21]	41.85*	19.48*
CR-Match§	**40.69**	**18.44**

Table 6.4: Error rates on ImageNet after 2^{20} iterations. CR-Match§ refers to CR-Match combined with CPL [ZWH+21] from FlexMatch. *Numbers are from [ZWH+21].

RotNet	FeatDistLoss	MiniImageNet@40	CIFAR10@4	CIFAR100@4	SVHN@4
		35.13	11.86	46.22	2.42
	✓	34.14	**10.33**	43.48	2.34
✓		34.64	11.27	41.48	2.21
✓	✓	**33.82**	10.92	**39.22**	**2.09**

Table 6.5: Ablation studies across different settings. Error rates are reported for a single split.

on a single split on CIFAR-10, CIFAR-100, and SVHN with 4 labeled examples per class, and on MiniImageNet with 40 labels per class. Specifically, we remove the $\mathcal{L}_{Dist}$ from equation 6.4 and train the model again using the same training scheme for each setting. We do not ablate $\mathcal{L}_{Pseudo}$ and $\mathcal{L}_S$ due to the fact that removing one of them leads to a divergence of training.

We report final test error rates in Table 6.5. We see that both RotNet and FeatDistLoss contribute to the final performance while their proportions can be different depending on the setting and dataset. For MiniImageNet, CIFAR-100 and SVHN, the combination of both outperforms the individual losses. For CIFAR-10, FeatDistLoss even outperforms the combination of both. This suggests that RotNet and FeatDistLoss are both important components for CR-Match to achieve the state-of-the-art performance. Note that RotNet can be replaced by other types of self-supervision as well. We opt RotNet due to its superior performance in our initial experiments. On CIFAR-100 with 4 labels per class, CRMatch with SimCLR achieves an error rate of 42.50% compared to that of 39.22% from CRMatch with RotNet.

Figure 6.4 shows a more detailed analysis of the training process on CIFAR-100 with 4 labels per class for CR-Match and CR-Match without FeatDistLoss. The confidence threshold in CR-Match filters out unconfident predictions during training. Therefore, at each training step only images with confidence scores above the threshold contribute to the loss. We observe that CR-Match improves pseudo-labels for the unlabeled data, as it achieves a lower error rate of all unlabeled images as well as contributing unlabeled images during the training while maintaining the percentage of contributing images. The increasing of the pseudo-label error rate in Figure 6.4 middle is due to the increasing of the percentage of contributing pseudo-labels and the prediction confidence. At the beginning of the training, the contributing pseudo-labels are mostly correct as only a small number of samples are highly confident and, thus, selected. However, during the course of the training, the overall prediction confidence increases, resulting in more unlabeled data being used, which introduces more errors in pseudo-labels.

Effect of different confidence thresholds. For the main results in Section 6.4.1, we use a confidence threshold of 0.95 following [SBL+20]. We now study the model robustness against different confidence thresholds. Experiments are conducted on a single split with 4 labeled examples from CIFAR-100 on a Wide ResNet-28-2. Figure 6.5 shows the error rate of CR-Match when using a confidence threshold from 0.90 to 0.99. In general, the thresholding mechanism

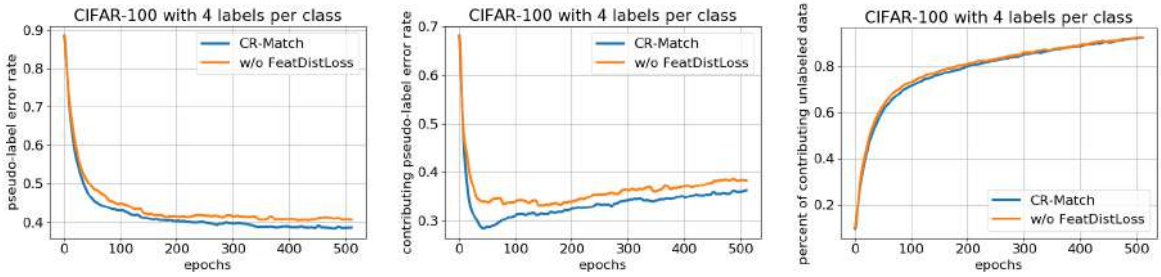


Figure 6.4: Ablation study of our best model on CIFAR-100 with 4 labels per class. **Left:** CR-Match has a lower pseudo-label error rate. **Middle:** If only the confident predictions are taken into account, CR-Match outperforms the other with a even larger margin in terms of pseudo-label error rate. **Right:** In spite of a better pseudo-label error rate on contributing unlabeled images, the percentage of contributing unlabeled images is maintained the same for CR-Match.

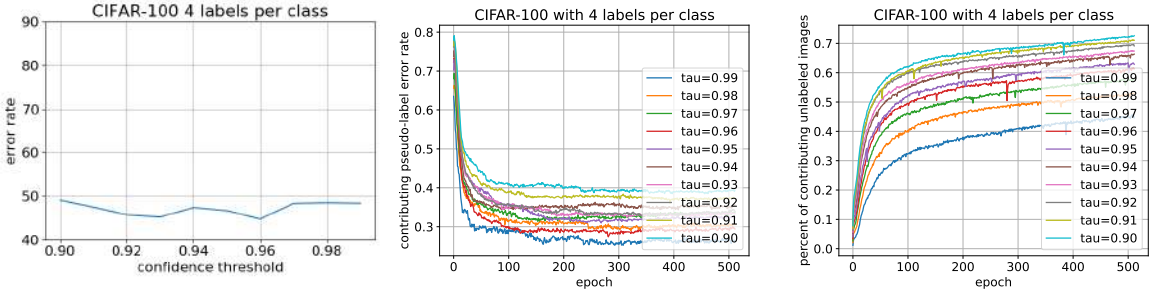


Figure 6.5: **Left:** Effect of different confidence thresholds on error rate. We run experiments on a single split of CIFAR-100 with 4 labels per class. The model is a Wide-ResNet-28-2. Our model shows good robustness against small changes in the confidence threshold. **Middle:** Effect of different confidence thresholds on pseudo-label error rate during the training. **Right:** Effect of different confidence thresholds on the number of unlabeled training samples.

provides the model a relatively smooth transition between learning from labeled data and learning from unlabeled data. A low percentage of the contributing unlabeled data at the beginning of the training can alleviate the potential error introduced by the low-quality pseudo-labels. This suggests that the quality of pseudo-labels is more important than the quantity for reaching a high accuracy at the early stage. As the model learns from the labeled data, the error rate of the pseudo-label decreases, and the model becomes more confident about its predictions. Then, the number of unlabeled data that contribute to the final loss gradually increases, which allows the model to continue learning from unlabeled data. Figure 6.5 left also implies that our model is quite robust against small changes in the confidence threshold.

6.4.3 Influence of Feature Distance Loss

In this section, we analyze different design choices for FeatDistLoss to provide additional insights of how it helps generalization. We focus on a single split with 4 labeled examples from CIFAR-100 and report results for a Wide ResNet-28-2 [ZK16]. For fair comparison, the same 4 random labeled examples for each class are used across all experiments in this section.

Different distance metrics for FeatDistLoss. Here we discuss the effect of different metric

functions d for FeatDistLoss. Specifically, we compare two groups of functions in Table 6.6: metrics that increase the distance between features, including cosine similarity, negative JS divergence, and L2 similarity (i.e. normalized negative L2 distance); metrics that decrease the distance between features, including cosine distance, JS divergence, and L2 distance. We find that both increasing and decreasing distance between features of different augmentations give reasonable performance. However, increasing the distance always performs better than the counterpart (e.g., cosine similarity is better than cosine distance). We conjecture that decreasing the feature distance corresponds to an increase of the invariance to data augmentation and leads to ignorance of information like rotation or translation of the object. In contrast, increasing the feature distance while still imposing the same label makes the representation equivariant to these augmentations, resulting in more descriptive and expressive representation with respect to augmentation. Moreover, a classifier has to cover a broader space in the feature space to recognize rather dissimilar images from the same class, which leads to improved generalization. In summary, we found that both increasing and decreasing feature distance improve over the model which only applies consistency on the classifier level, whereas increasing distances shows better performance by making representations more equivariant.

Metric		Error rate
Impose equivariance	cosine similarity	**45.52**
	L_2 similarity	46.22
	negative JS div.	46.46
Impose invariance	cosine distance	46.98
	L_2 distance	48.74
	JS divergence	47.48
CR-Match w/o FeatDistLoss		48.89

Table 6.6: Effect of different distance functions for FeatDistLoss. The same split on CIFAR-100 with 4 labels per class and a Wide ResNet-28-2 is used for all experiments. Metrics that pull features together performs worse than those that push features apart. The error rate of CR-Match without FeatDistLoss is shown at the bottom.

Transformations	Feature extractor	
	CR-Equiv	CR-Inv
Translation	33.22±0.28	36.80±0.30
Scaling	11.09±0.66	14.87±0.40
Rotation	15.05±0.33	21.92±0.32
ColorJittering	31.04±0.50	35.99±0.27

Table 6.7: Error rates of binary classification (whether a specific augmentation is applied) on the features from CR-Equiv (increasing the cosine distance) and CR-Inv (decreasing the cosine distance). We evaluate translation, scaling, rotation, and color jittering. Lower error rate indicates more equivariant features. Results are averaged over 10 runs.

Invariance and equivariance. Here we provide an additional analysis to demonstrate that increasing the feature distance provides equivariant features while the other provides invariant features. Based on the intuition that specific transformations of the input image should be

more predictable from equivariant representations, we quantify the equivariance by how accurate a linear classifier can distinguish between features from augmented and original images. Specifically, we compare two models from Table 6.6: the model trained with cosine similarity denoted as *CR-Equiv* and the model trained with cosine distance denoted as *CR-Inv*. We train a linear SVM to predict whether a certain transformation is applied for the input image. 1000 test images from CIFAR-100 are used for training and the rest (9000) for validation. The binary classifier is trained by an SGD optimizer with an initial learning rate of 0.001 for 50 epochs, and the feature extractor is fixed during training. We evaluate translation, scaling, rotation, and color jittering in Table 6.7. All augmentations are from the standard PyTorch library. The SVM has a better error rate across all augmentations when trained on CR-Equiv features, which means information like object location or orientation is more predictable from CR-Equiv features, suggesting that CR-Equiv produces more equivariant features than CR-Inv. Furthermore, if the SVM is trained to classify strongly and weakly augmented image features, CR-Equiv achieves a 0.27% test error while CR-Inv is 46.18%.

Regularization on the classifier level. As we described in Section 6.3, FeatDistLoss contains two levels of regularization: On the feature level, representations are encouraged to become more equivariant. On the classifier level, the same class label is imposed on different versions of the same image via pseudo-labeling. Here we provide more insights into the regularization on the classifier level in Table 6.8. Specifically, we conduct experiments on replacing or complementing the CE loss with Jensen-Shannon divergence. First, we can see that removing the classifier loss and using only the equivariant loss on the feature level leads to a significant drop on performance (from 45.52% to 91.53%). This is because $\mathcal{L}_{Dist}$ alone will just make the model aware of the difference between augmentations but does not help the classifier to distinguish between classes of unlabeled data, making the classifier unable to benefit from the usage of unlabeled data. Thus, the performance is on par with the model trained on labeled data only (91.28% error rate) Second, complementing the cross-entropy loss on the classifier level with Jensen-Shannon divergence, improves the performance (45.01%) while replacing it leads to inferior performance (76.83%).

Classifier level	Feature level	Error rate
None	None	91.28
None	Equiv.	91.53
CE	Equiv.	**45.52**
JSD	Equiv.	76.83
CE + JSD	Equiv.	**45.01**

Table 6.8: Effect of different regularization techniques on the classifier level. CE denotes cross-entropy loss. JSD denotes Jensen-Shannon divergence. Equiv. denotes the equivariance version of $\mathcal{L}_{Dist}$. Note that the chance level is 99%. None + None represents the model trained with labeled data only. The same split on CIFAR-100 with 4 labels per class and a Wide ResNet-28-2 is used for all experiments.

Different data augmentations for FeatDistLoss. In our main results in Section 6.4.1, FeatDistLoss is computed between features generated by weak augmentation and strong augmentation. Here we investigate the impact of FeatDistLoss with respect to different types of data augmentations. Specifically, we evaluate the error rate of CR-Inv and CR-Equiv under three augmentation strategies: weak-weak pair indicates that FeatDistLoss uses two weakly augmented images, weak-strong pair indicates that FeatDistLoss uses a weak augmentation and

Error rate	CR-Inv	CR-Equiv
Weak-Weak	48.88	48.51
Weak-Strong	**46.98**	**45.52**
Strong-Strong	48.57	48.05

Table 6.9: Effect of combinations of weak and strong augmentation in FeatDistLoss on a Wide ResNet-28-2 for CR-Inv and CR-Equiv.

a strong augmentation, and strong-strong pair indicates that FeatDistLoss uses two strongly augmented images.

As shown in Table 6.9, using either CR-Inv or CR-Equiv using weak-strong pairs conistently outperforms the other augmentation settings (weak-weak and strong-strong). Additionally, CR-Equiv consistently achieves better generalization performance across all three settings. In particular, in the case advocated in this chapter, namely using weak-strong pairs, CR-Equiv outperforms CR-Inv by 1.46%. Even in the other two settings, CR-Equiv leads to improved performance even though only by a small margin. This suggests that, on the one hand, that it is important to use different types of augmentations for our FeatDistLoss. And on the other hand, maximizing distances between images that are inherently different while still imposing the same class label makes the model more robust against changes in the feature space and thus gives better generalization performance.

Linear projection and confidence threshold in FeatDistLoss. As mentioned in Section 13.3, we apply $\mathcal{L}_{Dist}$ at (a) in Figure 6.3 with a linear layer mapping the feature from the encoder to a low-dimensional space before computing the loss, to alleviate the curse of dimensionality. Also, the loss only takes effect when the model's prediction has a confidence score above a predefined threshold τ. Here we study the effect of other design choices in Table 6.10. While features after the global average pooling (i.e. (b)) gives a better result than the ones directly from the feature extractor, (b) performs worse than (a) when additional projection heads are added. Thus, we use features from the feature extractor in CR-Match. The error rate increases

Features taken from Fig. 6.3 at	feature	feature + linear	feature + MLP
(a)	48.37	**45.52**	47.52
(b)	47.37	46.10	47.15

Table 6.10: Effect of the projection head z, and the place to apply $\mathcal{L}_{Dist}$. (a) denotes un-flattened features taken from the feature extractor directly. (b) denotes features after the global average pooling. MLP has 2 FC layers and a ReLU. Removing the linear projection head harms the test error, and a non-linear projection head does not improve the performance further.

from 45.52% to 48.37% and 47.52% when removing the linear layer and replacing the linear layer by a MLP (two fully-connected layers and a ReLU activation function), respectively. This suggests that a lower dimensional space serves better for comparing distances, but a non-linear mapping does not give further improvement. Moreover, when we apply FeatDistLoss for all pairs of input images by removing the confidence threshold, the test error increases from 45.52% to 46.94%, which suggests that regularization should be only performed on features that are actually used to update the model parameters, and ignoring those that are also ignored by the model.

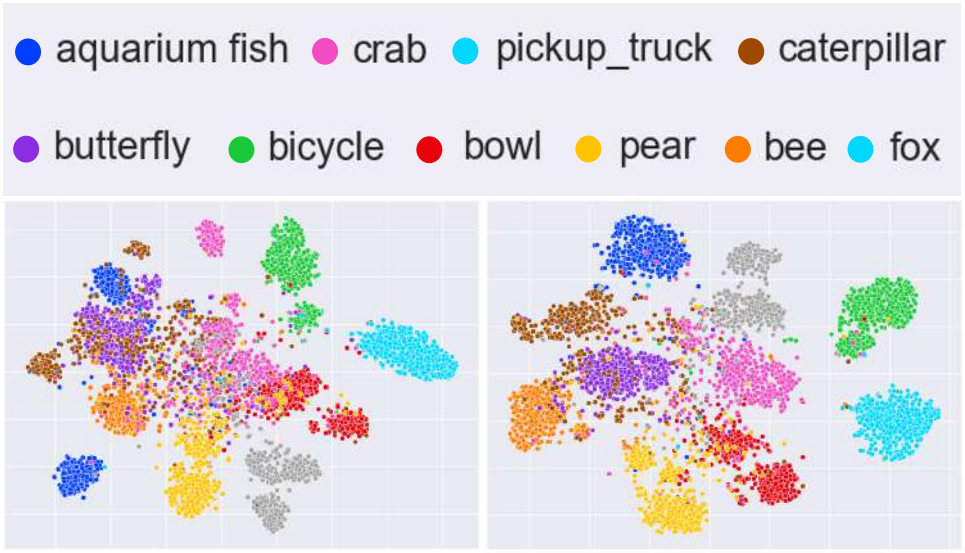


Figure 6.6: We plot t-SNE of input image features extracted by a CR-Match model trained without FeatDistLoss (left) and a CR-Match model with it (right). The better separation from CR-Match suggests that FeatDistLoss improves decision boundaries.

FeatDistLoss improves decision boundaries. As suggested by Figure 6.2, models trained with FeatDistLoss tend to have improved decision boundaries. Here we take two models from section 6.4.2, CR-Match (39.22% error rate) and CR-Match without FeatDistLoss (41.48% error rate), and plot t-SNE plots of features extracted from unlabeled images. As shown in Figure 6.6, CR-Match with FeatDistLoss produces better separation between classes. For example, CR-Match forms two clearer clusters for caterpillar and butterfly, while CR-Match without FeatDistLoss mostly mixes them up. Another example is that the overlap between crab, bowl, and pear is much less for CR-Match compared to CR-Match without FeatDistLoss. Moreover, the improved decision boundaries also lead to better per-class error rate. The standard deviation of per-class error rates for CR-Match is 4.34% lower than that from CR-Match without FeatDistLoss (30.83% v.s. 26.49%).

Additional analysis on FeatDistLoss. To further verify the importance of FeatDistLoss, we show in Figure 6.7 the contribution of FeatDistLoss compared to other losses. The model is CR-Equiv. trained on CIFAR-100 with 4 labels per class. We can see that during the training, the two components of FeatDistLoss, $\mathcal{L}_{Dist}$ and $\mathcal{L}_{PseudoLabel}$, account for a large portion of the overall loss, thus, the gradient. Note that $\mathcal{L}_{Dist}$ is the negative cosine distance, thus, ranging from 1 to -1.

6.5 Experiments on Imbalanced SSL

In this section, we go beyond the standard setting and evaluate the efficacy of our method under imbalanced SSL settings where both labeled and unlabeled data follow class imbalanced distributions. We first present the problem setup of imbalanced SSL. Then, we introduce the construction of the datasets before showing the final evaluation results.

Problem setup and notations. For a K-class classification problem, there is a labeled set $\mathcal{X} = \{(\mathbf{x}_n, y_n) : n \in (1, ..., N)\}$ and an unlabeled set $\mathcal{U} = \{\mathbf{u}_m : m \in (1, ..., M)\}$, where $\mathbf{x}_n, \mathbf{u}_m \in \mathbb{R}^d$ are training examples and $y_n \in \{1, ..., K\}$ are class labels for labeled examples. N_k

and M_k denote the numbers of labeled and unlabeled examples in class k, respectively, i.e., $\sum_{k=1}^{K} N_k = N$ and $\sum_{k=1}^{K} M_k = M$. Without loss of generality, we assume the classes are sorted by the number of training samples in descending order, i.e., $N_1 \geq N_2 \geq ... \geq N_k$. The goal is to train a classifier $f : \mathbb{R}^d \to \{1, ..., K\}$ on $\mathcal{X} \cup \mathcal{U}$ that generalizes well on a class-balanced test set.

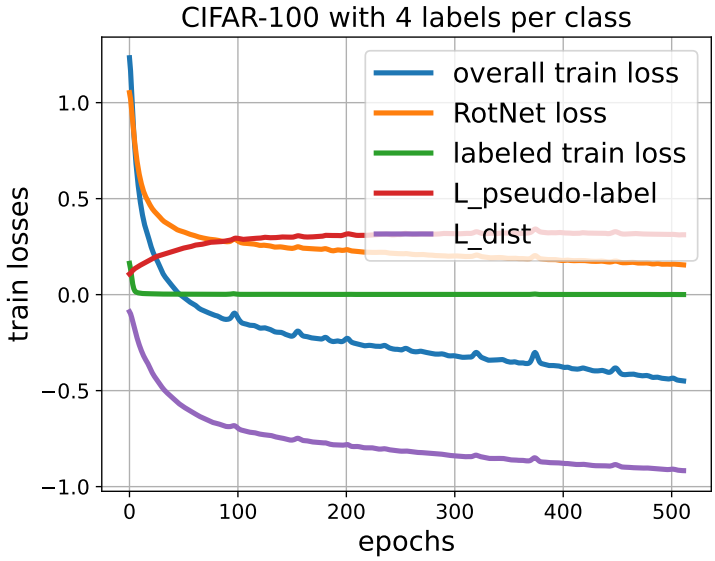


Figure 6.7: The amount of the contribution of the regularization term in the loss. The model is CR-Equiv. trained on CIFAR-100 with 4 labels per class.

Datasets. We consider three common datasets in the field to evaluate the efficacy of CRMatch for imbalanced SSL: CIFAR10-LT [KH09], CIFAR100-LT [KH09], and Semi-Aves [SM21].

For CIFAR-10-LT and CIFAR100-LT, we follow the convention [KHP+20, WSM+21] and randomly select some training images for each class determined by a pre-defined imbalance ratio γ as the labeled and the unlabeled set. Specifically, we set $N_k = N_1 \cdot \gamma^{-\frac{k-1}{K-1}}$ for labeled data and $M_k = M_1 \cdot \gamma^{-\frac{k-1}{K-1}}$ for unlabeled data. We use $N_1 = 1500$; $M_1 = 3000$ for CIFAR-10 and $N_1 = 150$; $M_1 = 300$ for CIFAR-100, respectively. Following [KHP+20, WSM+21], we report results with imbalance ratio $\gamma = 50$, 100 and 150 for CIFAR10-LT and $\gamma = 20$, 50 and 100 for CIFAR100-LT. Therefore, the number of labeled samples for the least class is 10 and 1 for CIFAR-10 with $\gamma = 150$ and CIFAR-100 with $\gamma = 100$, respectively.

Semi-Aves is a subset of bird species from the Aves kingdom of the iNaturalist 2018 dataset. There are 200 in-class and 800 out-of-class categories. The dataset consists of a labeled set L_{in} with 3,959 labeled images, an in-class unlabeled set U_{in} with 26,640 images, an out-of-class unlabeled set U_{out} with 122,208 images, a validation set L_{val} of 2,000 images, and 8,000 test images. The training data in L_{in}, U_{in}, and U_{out} has imbalanced distributions, specifically L_{in} has 5 to 43 images and U_{in} has 16 to 229 images per class. The validation data and test data have a uniform distribution with 40 and 10 images per class, repectively. In our experiments, we use L_{in} or $L_{in} \cup L_{val}$ as the labeled set and U_{in} as the unlabeled set. We do not use unlabeled images from U_{out} since out-of-class unlabeled images are found empirically harmful to the final performance [OOR+18] and making good use of out-of-class unlabeled images is out of the scope of this chapter. More details on the class distribution can be found in [SM21].

Implementation details. Due to the performance superiority of $\mathcal{L}_{U-equiv}$ over $\mathcal{L}_{U-inv}$, we use CR-Equiv throughout this section. For all experiments in this section, we use the same

Class index		1	2	3	4	5	6	7	8	9	10	Avg.
Recall	FixMatch + CReST+	**98.6**	99.3	85.8	77.4	84.4	63.4	77.2	55.5	37.4	34.6	71.3
	CR-Match + CReST+	**98.6**	**99.6**	**88.8**	**82.5**	**86.7**	**67.8**	**78.7**	**57.0**	**42.7**	**40.6**	**74.3**
Precision	FixMatch + CReST+	53.3	61.4	71.4	57.9	77.0	82.4	93.0	**97.1**	97.6	**98.0**	78.9
	CR-Match + CReST+	**56.1**	**62.9**	**75.1**	**63.7**	**78.6**	**83.3**	**94.5**	**97.1**	**98.1**	97.1	**80.7**

Table 6.11: Class-wise precision and recall (%) on the balanced test set of CIFAR-10-LT. Models are trained with imbalance ratio $\gamma = 150$.

hyper-parameters and design choices from the CIFAR experiments in Section 6.4.1. We deploy FixMatch [SBL+20] as the base SSL method due to its superiority under the standard SSL settings. A Wide ResNet-28-2 [ZK16] is used as the backbone as recommended by [OOR+18]. We base our implementation on the public codebases of each methods. Therefore, method-specific hyper-parameters follow the same as in their original papers [KHP+20, WSM+21]. For example, all experiments on CIFAR-LT are trained with batch size 64 using Adam optimizer [KB15] with a constant learning rate of 0.002 without any decay. We train the models for 500 epochs, each of which has 500 steps, resulting in a total number of 2.5×10^5 training iterations. On Semi-Aves, we follow the hyper-parameters from [OKK22]. For example, the models are trained for 90 epochs with a batch size of 256, and the optimizer is SGD with a learning rate of 0.04. For all experiments, we report the average test accuracy of the last 20 epochs following [OOR+18].

Results on CIFAR-10 and CIFAR-100. Table 6.12 and Table 6.13 compare our method with various SSL algorithms and long-tailed recognition algorithms on CIFAR-10-LT and CIFAR-100-LT with various imbalance ratios γ. Adding our method shows improved performance in most of settings. Our method combining with CoSSL [FDKS22] achieves the best or comparable performance across all settings. In particular, CRMatch + CoSSL outperforms others at large imbalance ratios (82.29% v.s. the second best 81.28% on CIFAR-10 at imbalance ratio $\gamma = 150$), which indicates the superiority of our method in handling severe dataset imbalance.

To analyze how the improvement is obtained, we compare the class-wise precision and recall of CReST+ and CReST+ with our method in Table 6.11. Both models are trained with imbalance ratio $\gamma = 150$ on CIFAR-10-LT using the same data split. The class indices are sorted according to the number of samples in descending order, i.e., class 1 has the largest number of data. For CReST+, the head classes tend to have higher precision but lower recall while the tail classes have lower precision but higher recall. By adding our method, the recall on the tail classes can be significantly improved without sacrificing much precision, which leads to the overall better performance. Similarly, the precision of the head classes is improved while the recall remains at the same level.

Results on Semi-Aves. As Semi-Aves is naturally imbalanced ($\gamma \approx 9$ and 4 for L_{train} and $L_{in} \cup L_{val}$, respectively), we compare CRMatch with other methods using different numbers of labeled data. We report the raw performance of backbone algorithms as well as the performance with CoSSL [FDKS22] considering its superior performance on CIFAR-10-LT and CIFAR-100-LT. From Table 6.14, we can see that CRMatch outperforms other backbone algorithms by a large margin in both settings. While CoSSL leads to improvement in all methods, CRMatch still achieves the best performance, which demonstrates the effectiveness of our method in realistic settings.

	CIFAR-10-LT		
	γ=50	γ=100	γ=150
vanilla	65.2±0.05*	58.8±0.13*	55.6±0.43*
Long-tailed recognition methods			
Re-sampling [Jap00]	64.3±0.48*	55.8±0.47*	52.2±0.05*
LDAM-DRW [CWG+19]	68.9±0.07*	62.8±0.17*	57.9±0.20*
cRT [KXR+20]	67.8±0.13*	63.2±0.45*	59.3±0.10*
SSL methods			
FixMatch [SBL+20]	81.58±0.34	74.74±1.35	70.04±0.77
ReMixMatch [BCC+20]	82.79±0.17	76.81±0.23	72.53±1.16
FlexMatch [ZWH+21]	81.89±0.25	74.94±0.96	70.09±0.42
CR-Match	82.87±0.04	76.54±0.87	72.14±0.76
FixMatch + DARP [KHP+20]	82.46±0.30	76.51±0.50	71.88±1.02
ReMixMatch + DARP [KHP+20]	82.88±0.23	76.77±0.29	72.90±0.95
FlexMatch + DARP [KHP+20]	81.93±0.22	74.84±0.66	70.46±0.58
CR-Match + DARP	83.22±0.27	77.32±0.29	73.44±0.06
FixMatch + CReST+ [WSM+21]	82.25±0.08	76.31±0.23	71.70±0.83
ReMixMatch + CReST+ [WSM+21]	83.71±0.17	79.13±0.19	75.17±0.31
FlexMatch + CReST+ [WSM+21]	82.75±0.25	77.23±0.35	72.21±0.11
CR-Match + CReST+	84.11±0.32	78.55±0.55	74.21±0.11
FixMatch + CoSSL [FDKS22]	86.63±0.24	83.10±0.48	80.15±0.59
ReMixMatch + CoSSL [FDKS22]	*87.55±0.06*	*84.15±0.65*	*81.28±0.95*
FlexMatch + CoSSL [FDKS22]	86.30±0.30	81.61±0.74	78.80±0.73
CR-Match + CoSSL [FDKS22]	**88.11±0.17**	**84.80±0.54**	**82.29±0.33**

Table 6.12: Classification accuracy (%) on CIFAR-10-LT using a Wide ResNet-28-2 under the uniform test distribution of three different class-imbalance ratios γ. The numbers are averaged over 5 different folds. We use the same code base as [KHP+20] for fair comparison following [OOR+18]. Numbers with * are taken from the original papers. The best number is in bold and the second best number is in italic.

	CIFAR-100-LT		
	γ=20	γ=50	γ=100
FixMatch [SBL+20]	49.58±0.90	42.10±0.38	37.46±0.48
ReMixMatch [BCC+20]	51.46±0.51	44.37±0.62	39.29±0.59
FlexMatch [ZWH+21]	51.00±0.75	42.86±0.42	37.20±0.51
CR-Match	52.03±0.42	44.37±0.57	39.32±0.31
FixMatch + DARP [KHP+20]	50.89±0.86	43.12±0.61	38.19±0.47
ReMixMatch + DARP [KHP+20]	51.95±0.40	45.24±0.46	39.50±0.58
FlexMatch + DARP [KHP+20]	50.78±0.71	42.81±0.36	36.99±0.66
CR-Match + DARP	49.33±0.32	44.13±0.38	39.18±0.80
FixMatch + CReST+ [WSM+21]	51.87±0.11	45.25±0.06	40.41±0.35
ReMixMatch + CReST+ [WSM+21]	51.22±0.38	45.91±0.33	41.24±0.79
FlexMatch + CReST+ [WSM+21]	51.16±0.63	43.12±0.57	38.09±0.58
CR-Match + CReST+	53.77±0.36	46.44±0.58	40.94±0.43
FixMatch + CoSSL [FDKS22]	53.99±0.87	47.78±0.53	42.87±0.61
ReMixMatch + CoSSL [FDKS22]	**55.92±0.69**	**49.10±0.59**	44.10±0.68
FlexMatch + CoSSL [FDKS22]	53.46±0.79	46.83±0.80	41.42±0.58
CR-Match + CoSSL [FDKS22]	*55.34±0.43*	*48.83±0.87*	**44.21±0.61**

Table 6.13: Classification accuracy (%) on CIFAR-100-LT under the uniform test distribution of three different class-imbalance ratios γ. The numbers are averaged over 5 different folds. We reproduce all numbers using the same codebase from [KHP+20] for a fair comparison. The best number is in bold and the second best number is in italic.

	Semi-Aves	
	$\mathcal{X} = L_{in} \cup L_{val}$	$\mathcal{X} = L_{in}$
FixMatch [SBL+20]	53.15	42.46
ReMixMatch [BCC+20]	51.28	40.10
FlexMatch [ZWH+21]	52.78	43.50
CRMatch	*54.53*	44.42
FixMatch + CoSSL [FDKS22]	54.15	*44.58*
ReMixMatch + CoSSL [FDKS22]	54.13	43.97
FlexMatch + CoSSL [FDKS22]	53.98	44.09
CRMatch + CoSSL [FDKS22]	**54.90**	**45.81**

Table 6.14: Classification accuracy (%) on Semi-Aves under the uniform test distribution. L_{train} and $L_{in} \cup L_{val}$ have imbalance ratio $\gamma \approx 9$ and $\gamma \approx 4$, respectively. The best number is in bold and the second best number is in italic.

6.6 Conclusion

The idea of consistency regularization gives rise to many successful works for SSL [BAP14, LA17, SJT16, SBL+20, XDH+20, KMHK20]. While making the model invariant against input perturbations induced by data augmentation results in improved performance, the scheme tends to be suboptimal when augmentations of different intensities are used. In this chapter, we propose a simple yet effective improvement, called FeatDistLoss. It introduces consistency regularization on both the classifier level, where the same class label is imposed for versions of the same image, and the feature level, where distances between features from augmentations of different intensities is increased. By encouraging the representation to distinguish between weakly and strongly augmented images, FeatDistLoss encourages more equivariant representations, leading to improved classification boundaries, and a more robust model. Through extensive experiments we show the superiority of our training framework, and define a new state-of-the-art on both standard and imbalanced semi-supervised learning benchmarks. Particularly, our method outperforms previous methods in low data regimes by significant margins, e.g., on CIFAR-100 with 4 annotated examples per class, our error rate (39.45%) is 4.83% better than the second best (44.28%). In future work, we are interested in integrating more prior knowledge and stronger regularization into SSL to further push the performance in low data regimes. In the following Chapter 7, Chapter 8, and Chapter 9, we will explore how to reduce annotation cost for by leveraging mutiple modalities of the input data.

7 In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval

Contents

In this chapter, we investigate multimodal video-language learning, leveraging large pre-trained foundational models for multimodal alignment to lower video annotation costs for text-video retrieval tasks. Large-scale noisy web image-text datasets have shown effectiveness in learning robust vision-language models, thus motivating our approach. However, to transfer them to the task of video retrieval, models still need to be fine-tuned on hand-curated paired text-video data to adapt to the diverse styles of video descriptions. To address this problem without the need for hand-annotated pairs, we propose a new setting, text-video retrieval with uncurated & unpaired data, that uses only text queries together with uncurated web videos during training without any paired text-video data. To this end, we propose our approach, In-Style, that learns the style of the text queries and transfers it to uncurated web videos. Moreover, to improve generalization, we show that one model can be trained with multiple text styles. To this end, we introduce a multi-style contrastive training procedure, that improves the generalizability over several datasets simultaneously. We evaluate our model on retrieval performance over multiple datasets to demonstrate the advantages of our style transfer framework on the new task of uncurated & unpaired text-video retrieval and improve state-of-the-art performance on zero-shot text-video retrieval.

This chapter is based on [SKSK23]. As the co-first author Anna Kukleva led the project jointly with Nina Shvetsova, sharing responsibilities across all stages of its development with equal contribution.

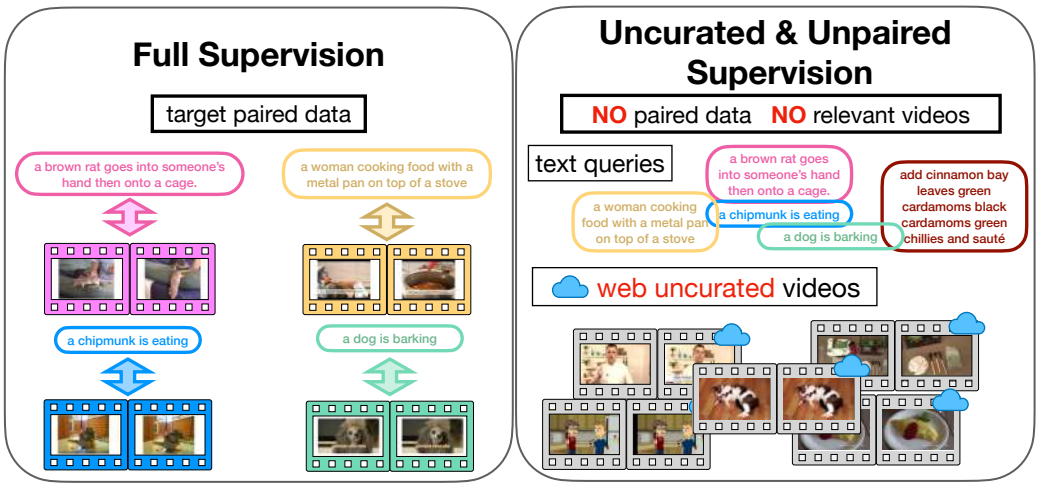


Figure 7.1: **Training data for supervised and uncurated & unpaired settings for text-video retrieval. Left**: standard supervised text-video retrieval, aligned and paired data is given for each target setting of the same distribution as target test set; **Right**: our uncurated & unpaired text-video retrieval setting. No paired data is available during training, only text queries, whereas to support training, we use uncurated web videos.

7.1 Introduction

Vision-language retrieval refers to the task of retrieving an image or a video from a large data pool given a textual description of the content. Especially the field of text-image retrieval has seen remarkable progress, mainly spurred by the combination of image and text models trained on large-scale web collections [RKH+21, LLXH22] of image-text pairs. While advances in video retrieval also rely on pre-trained image-language models, which serve for better task transfer, most systems still require a fine-tuning on downstream data. This requires hand-annotated text-video pairs, namely a trimmed segment of a larger video that is precisely described by the corresponding text pair, for the training and testing of each target downstream dataset. Collecting such aligned pairs of text and videos can be time and cost intensive, and particularly gathering videos that comply with national regulations and copyright can be a challenge. Also, in case of relying on free web content, some videos can become unavailable over time while the respective curated annotations stay available for download but do not have matching videos.

To address this problem, we propose a new setup, *text-video retrieval with uncurated & unpaired data*, assuming the availability of text queries only and without related videos during training (Figure 7.1). The setting is motivated by the fact that it can be considered easier to collect or generate text data, e.g. by producing topic-specific text queries, rather than providing a video to match a specific context. To allow the training of a text-video retrieval system based on the given text, we assume to have access to an uncurated video collection as the only source of available videos.

As different domains and datasets contain diverse styles of textual descriptions of videos, we propose a novel method, *In-Style*, to transfer the caption style of given text queries to uncurated web videos, which can be from a deviating distribution compared to the given text queries. To transfer the style of the text queries, we leverage large image-language models [LLXH22, RKH+21] by creating pseudo pairs that correspond to the given text queries and videos from the uncurated collection by matching them in the shared embedding space [RKH+21]. Thus,

we identify a subset of videos that have more similarity to the text queries than the rest of the videos. We then adopt an image-to-text captioning model (captioner) to mimic the style of our text queries by training with these pseudo pairs. The stylized captioner is now capable of producing relevant video descriptions in the desired style; therefore, we re-annotate the web videos with the captioner to obtain aligned paired data; we call them generated pairs. Finally, we show that generated pairs help to adapt models pre-trained on large-scale web data [LLXH22, SCR+22] to the desired single or multiple styles of given text queries.

We evaluate our model on text-video retrieval over 5 benchmark datasets. Specifically, we demonstrate the advantages of the In-Style method on the new task of uncurated & unpaired text-video retrieval with image-language [LLXH22] and video-language [SCR+22] pre-trained backbones. We show the generalization of the proposed approach by training a single model for multiple datasets at once leading to an improved state-of-the-art zero-shot text-video retrieval performance.

We summarize our contributions in the following: (i) we introduce a new task of *text-video retrieval with uncurated & unpaired data* where during training, only text queries are available, whereas for standard text-video retrieval task, paired text-video data is used; (ii) we propose a novel method, In-Style, to transfer the style of text queries in an unsupervised way, showing that style is an important component for language-based retrieval tasks; therefore, we repurpose large pre-trained image-language models to generate pseudo-captions of the same style for uncurated web videos; (iii) we demonstrate the advantages of our In-Style method for the new task over 5 different datasets with individual models for each dataset as well as one generalized model and we achieve state-of-the-art performance on zero-shot text-video retrieval.

7.2 Related Work

In this section, we discuss prior work on the text-video retrieval topic. We will not revisit the topic of vision-language large-scale pretraining as previously discussed in Chapter 2.

Text-video retrieval methods usually focus on learning modules that are able to capture relations between features from text and video modalities [YKK18, LANZ19, GSAS20, CZJW20, CBL+21, DKKP21, WZY21]. Currently, many approaches leverage pre-training on large-scale video-text [BNVZ21, MZA+19, MAS+20] or image-text [LLZ+21, LLXH22] datasets with a further adaptation of the backbone to individually downstream datasets. In this context, ClipBERT [LLZ+21] proposed sparse sampling instead of using dense full-length videos that allow lightweight training. However, foundation models [BHA+21] such as CLIP [RKH+21], combining the success of transformer architectures [DBK+20] using a contrastive objective [OLV18] and being trained on large collections of text-image pairs from the web, providing a strong zero-shot [LJZ+22, PQOBTM21] baseline on downstream tasks that outperforms many previous methods. Therefore, more recent approaches focus on adapting text-image CLIP pre-trained models for text-video retrieval [GVM+22, BCJ+22, FXXC21, GLC+21, LXX+22]. X-pool [GVM+22] introduces cross-modal attention to reason between text and frames of a video, TS2-Net [LXX+22] proposes dynamic adjustments over temporal and spatial token dimensions, which allows fine-tuning spatial model on video data without architecture changes. Another way to leverage foundation models is to enhance training data [WLF+23, ZMKG23]. Cap4Video [WLF+23] generates auxiliary captions for available curated training videos by using ZeroCap [TSSW22] that optimizes GPT-2 [RWC+19] text generation using a CLIP-based loss [RKH+21]. LaViLa [ZMKG23] proposes to generate additional narrations for a dense coverage of long videos from the Ego4D dataset [C+, GWB+22a] by fine-tuning a pre-trained large language model [RWC+19] on existing annotated text-video paired data. In contrast, we

propose to exclude pre-annotated text-video paired data from the training and, relying on text descriptions only, generate text-video pairs leveraging uncurated web videos while transferring the style of original captions.

7.3 Method

In this section, we introduce the proposed uncurated & unpaired text-video retrieval training setup. Typically, models for text-video retrieval are trained on *paired* text-video data. Given a set of pairs of captions t_i and corresponding videos v_i: $\{(t_i, v_i)\} \in D$, where D is a data distribution, the goal is to learn a similarity function $s(t_i, v_j)$ that calculates the similarity between the caption t_i and the video v_j. The training can be done from scratch, but typically pre-trained image-language [RKH+21, LLXH22] or video-language models [LJS+20, MZA+19] are fine-tuned on the target paired text-video data and then evaluated on the test set from the same distribution D [LXX+22, FXXC21]. If the evaluation is performed on multiple datasets, the model is usually fine-tuned for each dataset individually.

In contrast, we propose *a text-video retrieval with uncurated & unpaired data*, where only target text queries are available during training without any videos. More precisely, given a set of text descriptions $\{t_i\}$ from data distribution D, we aim to learn useful information about the similarity $s(t_i, v_j)$ in D relying only on the clean textual descriptions. We further assume that a large set of freely accessible web videos $V' = \{v'_j\} \in D'$ without any paired text is available to support the training (such as videos of the HowTo100M dataset [MZA+19]). We note that the data distribution D' in the support video dataset can deviate from the distribution D.

Finally, to avoid to train different models individually for each target dataset, we further consider learning a *generalized* model that maintains the performance of individual models over a set of K datasets of different caption styles and coming from different data distributions $D_1, ..., D_K$.

7.4 In-Style Method

To address the task of uncurated & unpaired text-video retrieval, we aim to transfer the style of the text queries (the only available curated information) to an uncurated web video dataset. To this end, we rely on web-scale pre-trained image-language models as a supervisory signal and leverage them as a matching module and pre-trained captioning model that we adapt throughout the training process. The steps of the proposed In-Style method are shown in Figure 7.2. The first step is *Pseudo Matching*, described in Section 7.4.1, which matches the given text queries to the most relevant videos from the set of all uncurated web videos. The following *Style Transfer* step (Section 7.4.2) adapts the pre-trained captioning model (captioner) to the target text style by training it on the previously obtained pseudo pairs. The captioner is then used to generate new style-adapted captions for all available web videos, which are then filtered to avoid too noisy pairs; we refer to the resulting filtered web videos with style-adapted video descriptions as generated pairs. Finally, we adapt a pre-trained vision-language model for the task of text-video retrieval on the generated pairs (Section 7.4.3). Moreover, in Section 7.4.3, we propose the training of a generalized model on multiple styles of text queries at the same time and introduce a new contrastive objective, In-Style, that improves training on more than one text style at once.

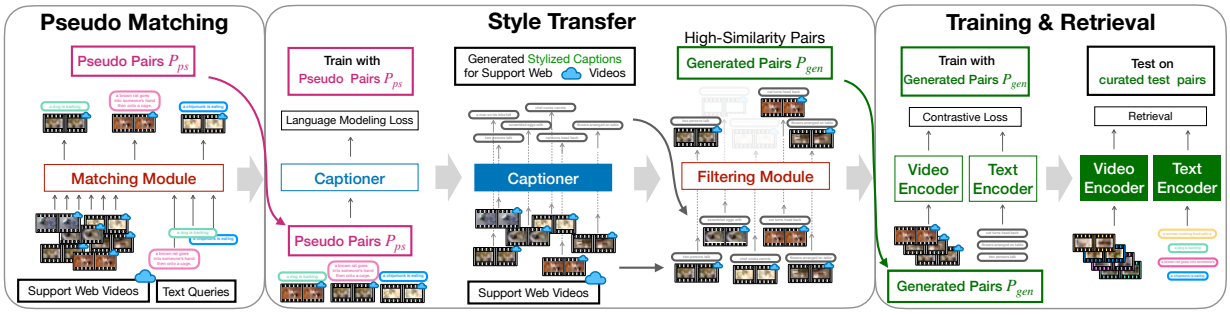


Figure 7.2: **The proposed In-Style method.** First, in the pseudo matching step, pseudo pairs P_{ps}, which consist of text queries and the most related web videos from the support set, are created. Style Transfer: captioner is tuned with the obtained pseudo pairs P_{ps} to adapt it to the style of the given text queries. Next, stylized new captions are generated for all videos in the support set and then filtered to avoid noisy captions; the resulting set of generated pairs P_{gen} contains web videos and aligned captions of the desired style. To complete the retrieval task, we adapt the dual video-text encoder model with the generated pairs P_{gen} and evaluate on curated paired text-video test sets.

7.4.1 Pseudo Matching

First, we obtain pseudo video-text pairs, with each pair containing one of the available text queries and the most relevant uncurated video from the web collection. For pseudo matching, we leverage image-language models such as CLIP [RKH+21] or BLIP [LLXH22] that excel in zero-shot retrieval performance [LJZ+22]. Such models usually follow a dual-encoder architecture: encoders f_t and f_v projects text t and image x into a common multimodal embedding space. The similarity of text and image is computed as a cosine similarity in this common space: $sim(t,x) = \frac{f_t(t)^\top f_v(x)}{\|f_t(t)\| \|.f_v(x)\|}$. We use this metric to match the available text queries to the closest video.

Since available videos can vary in overall duration (for example, five or more minutes) and cover a lot of different actions, we divide all videos into non-overlapping clips of s-seconds. We denote $V' = \{v'_j\}$ as a set of all such video clips. Then we calculate a multimodal representation for each video clip v'_j as an average representation of m uniformly sampled frames of a video. Using precomputed embeddings, we connect every caption t_i with a video v' with maximum similarity from available set of videos V', such as:

$$v_i^p = \arg\max_{v'_j \in V'} \text{sim}(t_i, v'_j). \tag{7.1}$$

To increase the diversity of matched videos, we don't allow multiple captions to match the same video clip; therefore, when video clip v_i^p is matched, we exclude it from V'. Thus, we obtain a set of pseudo text-video pairs $P_{ps} = \{(t_i, v_i^p)\}$. In Section 7.5, we show that this step allows us to introduce a weak supervision that may not find the exact match but provides a basis for further style transfer.

7.4.2 Style Transfer

We aim to transfer style of the given text queries to other, unrelated web videos by generating new captions with the desired style. Inspired by the ability of language models conditioned on visual input [LLXH22] to generate plausible descriptions for diverse visual inputs, we propose to adapt the pre-trained image captioner g using the obtained set of noisy pseudo text-video pairs P_{ps}. By doing this, we adapt the captioner to both, the style of the captions as well as the style of the web videos. This allows us to generate new stylized captions P_{gen} for the full support set of videos V' using this captioner.

Captioner. More specifically, we follow the BLIP [LLXH22] captioner architecture, which we extend for video captioning by conditioning the model not only on a single image, but on a number of video frames. To this end, we apply the image encoder on each frame individually and inject a joint set of visual tokens into the text decoder model, which produces text in an autoregressive manner. To train the captioner g on the pseudo text-video pairs P_{ps}, we utilize the common language model loss that optimizes cross-entropy loss between ground truth and predicted probabilities of the next token given a correct set of previous tokens in the sentence. Following BLIP, we also use label smoothing with parameter 0.1 while calculating cross-entropy.

Stylization of Captions. For each video $v'_i \in V'$, we generate a caption $t^g_i = g(v'_i)$ with a captioner g trained on pseudo pairs by using a nucleus sampling [HBD+20]. Nucleus sampling was shown to generate more diverse and detailed captions than a beam search [VCS+15, LLXH22].

Filtering. As the captioner g is adapted on pseudo pairs and shifts the model closer to a vocabulary of given text queries $\mathcal{D}$, some of the generated captions t^g_i might be noisy and not descriptive for the web videos. Therefore, we further filter the generated pairs based on a similarity score $s(t^g_i, v'_i)$ utilizing the large pre-trained image-language dual encoders the same way as it was used for creating pseudo text-video pairs (Section 7.4.1). Leaving only pairs with similarity higher a threshold $s(t^g_j, v'_j) > th$, we obtain a paired set of web videos and stylized related captions $P_{gen} = \{(t^g_j, v'_j)\}$. In Section 7.5.7, we show that even a noisy set of pseudo pairs is enough to adapt a captioner for generating captions in a desired text style and that stylized captions combined with the following filtering provide a strong learning signal to boost the performance of retrieval in target distribution D.

7.4.3 Training and Retrieval

Single-Style Training. To allow for text-video retrieval based on the stylized captions and the paired video data, we train a dual-encoder architecture [LLXH22] on the set of generated pairs P_{gen} with the contrastive loss [OLV18]. We show that P_{gen} provides better supervision than P_{ps} or even a combination P_{gen} + P_{ps}. Practically, we consider several pre-trained models: the image-text model BLIP [LLXH22], which we adapted for video as described in Section 7.4.1, as well as video-text model EAO [SCR+22], which is pre-trained on the HowTo100M dataset with ASR-video pairs, which serve as noisy supervision. Following previous works, we use symmetric contrastive loss, which brings together text t^g_i and video v_i from a text-video pair $(t^g_i, v_i) \in P_{gen}$ (a positive pair) in shared video-text embedding space, and contrasting them on

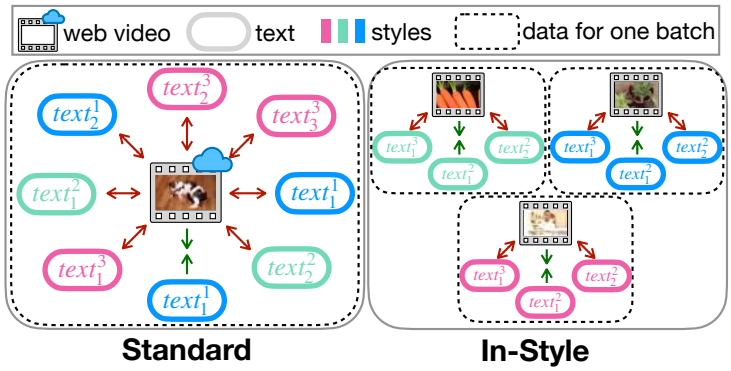


Figure 7.3: **Multi-dataset training. Left**: Standard contrastive training with multiple datasets. **Right**: Ours In-Style training procedure. Each batch consists of text queries that belong only to the same style. Note that we use only web videos from the support set; therefore, all videos are from the same distribution.

video and text from different pairs (negatives), that are pushed apart:

$$L = -\frac{1}{2B}\sum_{i=1}^{B}(\log\frac{\exp(\frac{s(t_i^g,v_i)}{\tau})}{\sum_{j=1}^{B}\exp(\frac{s(t_i^g,v_j)}{\tau})} + \log\frac{\exp(\frac{s(v_i,t_i^g)}{\tau})}{\sum_{i=1}^{B}\exp(\frac{s(v_i,t_j^g)}{\tau})}), \tag{7.2}$$

where τ denotes a temperature parameter, and B is a number of pairs.

For the fine-tuning of the the BLIP model, we follow the original setup and utilize the extension of contrastive training with a momentum encoder and a queue that keeps more negatives, as well as soft labels. For the fine-tuning of the EAO model, we follow the respective setup without a momentum encoder or soft labels.

Multi-Style Training. Finally, we consider training a generalized model on multiple sources of text queries coming from different data distributions $D_1, ..., D_K$. Let's denote $P^1_{gen}, ..., P^N_{gen}$ set of generated pairs for the captions from $D_1, ..., D_N$ respectively. Here, different sources can have various styles that might highlight different aspects of videos in their captions (Table 7.3). As an example, captions in the YouCook2 dataset [ZXC18] are more "action"-oriented, e.g. "combine macaroni sauce and cheese" or "stir in crushed tomatos", while captions of the LSMDC dataset [RRS] are third-person descriptions, e.g. "Someone gazes at the beautiful animal" or "Someone chews the sweet'. In standard training [LLXH22], all different styles with their matching videos would be present in contrastive loss together, which can lead to a mixture of different visual topics and text styles, which are easy to separate and which might include only few hard negatives per sample. To avoid this possibly noisy setting, we propose to modify the training procedure and to select video-caption pairs with captions from the same data source for contrastive loss. Formally, during training, we iterate over generated sets of pair $P^1_{gen}, ..., P^N_{gen}$ sampling a minibatch $\{(t_i^g, v_i)\}_{i=1}^B$ from a single set $P_{gen} \in \{P^1_{gen}, ..., P^N_{gen}\}$ and calculating loss $L(\{(t_i^g, v_i)\}_{i=1}^B)$ performing one optimization step with a minibatch (Figure 7.3). We note that for BLIP training we keep separate queues for each set P_{gen}.

We show in Section 7.5.4 that this setting can be beneficial for learning a generalized model. Our intuition is that text queries with the same style provide stronger negatives for the model, allowing the model to concentrate on the content of the captions rather than a style.

Pre-trained Model	Method	Supervision	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Mean			
			R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
BLIP [LLXH22]	Zero-shot	none	34.1	60.2	70.6	3	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	2	14.5	29.3	36.4	32.5	24.3	44.5	53.4	22.5
	In-Style (ours)	only text	**36.2**	**61.8**	**71.9**	**3**	**8.6**	**21.6**	**30.0**	**37**	**32.1**	**61.9**	**71.2**	**3**	**44.8**	**72.5**	**81.2**	**2**	**16.1**	**33.6**	**39.7**	**25**	**27.6**	**50.3**	**58.8**	**14**
	GT fine-tuning	T-V pairs	42.9	69.7	78.9	2	12.6	32.0	43.6	15	40.2	70.6	79.3	2	48.1	76.6	85.0	2	23.8	41.1	50.9	10	33.5	58.0	67.5	6.2
EAO [SCR+22]	Zero-shot	none	9.9	24.0	32.6	28	19.8	42.9	55.1	8	6.6	19.0	26.8	42	18.0	40.4	52.3	9	3.6	8.5	13.0	177	11.6	27.0	36.0	52.8
	In-Style (ours)	only text	**16.4**	**35.8**	**48.9**	**10**	**20.3**	**46.4**	**58.8**	**7**	**13.2**	**31.6**	**44**	**15**	**23.4**	**50**	**62.4**	**5**	**4.9**	**12.3**	**16.7**	**94**	**15.64**	**35.22**	**46.16**	**26.2**
	GT fine-tuning	T-V pairs	22.8	47.8	60.3	6	26.7	55.9	68.6	4	19.2	43.1	54.4	8	25.1	53.6	65.7	5	8.9	21.2	29.4	40	20.5	44.3	55.7	12.6

Table 7.1: **Text-video retrieval with style transfer.** Comparison between upper bound, where the retrieval model trained with ground truth aligned text-video pairs (T-V pairs), zero-shot respective models (no style transfer or tuning) and our In-Style method, where we follow our new setting of *uncurated & unpaired text-video retrieval* for style transfer based only on input text queries.

Training Dataset	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Mean			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
MSR-VTT	*36.2*	*61.8*	*71.9*	3	7.6	18.8	25.9	62	29.0	54.5	65.4	4	43.3	70.7	79.9	2	15.2	28.5	35.3	31	26.3	46.9	55.7	20.4
YouCook2	31.5	55.5	64.4	4	*8.6*	*21.6*	*30.0*	**37**	25.1	53.9	65.2	4	41.1	67.3	76.8	2	14.2	28.8	36.9	30	24.1	45.4	54.7	15.4
Didemo	34.0	58.5	68.9	3	6.8	17.2	24.5	69	*32.1*	***61.9***	***71.2***	**3**	43.7	71.6	80.5	2	16.6	30.5	38.4	28	26.6	47.9	56.7	21
MSVD	36.0	59.4	69.5	3	6.4	16.4	23.6	70	27.0	54.9	65.0	4	***44.8***	*72.5*	*81.2*	**2**	14.5	27.4	34.8	32	25.7	46.1	54.8	22.2
LSMDC	33.9	60.3	69.9	3	7.1	18.1	25.6	68	31.7	59.9	69.1	3	44.6	71.7	80.0	2	*16.1*	**33.6**	39.7	**25**	26.6	48.7	56.8	20.2
Target dataset (mean over diagonal)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*27.5*	***50.3***	*58.8*	***14***
All five datasets – standard training	36.4	62.1	71.8	3	**8.7**	21.4	29.4	44	31.4	62.5	71.2	3	44.7	72.9	81.5	2	16.3	31.9	39.5	25	27.5	50.2	58.7	15.4
All five datasets – In-Style (ours)	**36.7**	**61.9**	**72.3**	**3**	8.5	**21.8**	**30.4**	38.5	**32.6**	61.8	**71.2**	**3**	44.7	**73.1**	**82.0**	**2**	**16.6**	32.2	**39.8**	26	**27.8**	50.2	**59.1**	14.5

Table 7.2: **Generalization performance of different models over all datasets.** Mean denotes an average of R1, R5, R10, MR over 5 datasets, correspondingly. **Top**: the proposed In-Style method with the input text queries only from one *respective* training dataset. **Bottom:** training with 5 different text query styles. Comparison between standard multi-dataset training and proposed In-Style procedure.

7.5 Experiments

We evaluate the proposed uncurated & unpaired text-video retrieval approach on five popular benchmark datasets: MSR-VTT [XMYR16], YouCook2 [ZXC18], MSVD [CD11], LSMDC [RRS], and DiDeMo [AHWS+17]. All datasets cover different styles of captions and videos, which includes YouTube and Flickr videos on various topics and video clips from movies. As a source of support videos, we use the large-scale web dataset HowTo100M [MZA+19]. We additionally test our model with text queries from the VATEX dataset [WWC+19] as well as with third-party text queries (not video captions), specifically with the recipe steps from Food.com dataset [MLNM19] and task descriptions from WikiHow dataset [KW18] datasets.

7.5.1 Dataset Details

MSR-VTT [XMYR16] contains in total 10k videos on various topics and 200K captions. More precisely, every 20 captions describe the same video in different words. We use split 9K+1K [GSAS20] in evaluation, resulting in 180K captions for training and 1K text-video for testing.

YouCook2 [ZXC18] is a dataset of 14K cooking instructional video clips, where each clip is annotated with a short cooking recipe step. Following[MZA+19, SCR+22], we use a 10K+3.5K training-testing split, leveraging 10K captions for training.

MSVD [CD11] contains 2K video snippets, where each is associated with approximately 40 sentences. The standard split consists of 1200 videos for training, 100 for validation, and 670 for testing. The training set contains 48K captions.

LSMDC [RRS] is a collection of 202 movies sliced into 118K movie-clips with one description

Dataset	Examples
MSR-VTT (∼43 symbols in a text)	1) The peoples are sharing their view on this car of different models 2) Someone is showing the ingredients for a dish they are going to make 3) A man is playing an instrument
YouCook2 (∼39 symbols in a text)	1) Combine macaroni sauce and cheese 2) Grate and cube potatoes 3) Stir in crushed tomatos
DiDeMo (∼147 symbols in a text)	1) A dog runs down a hill and stop behind a shrub. Dog sniffs and chews at patch of grass on rock. the dog approaches, then begins to sniff the cluster of plants first time hand is seen petting dog. 2) Only big screen is visible the camera first pans to the large screen. The view shifts from the basketball court to the fans in the seats across the stadium. 3) A bus stops. The bus stops at the end of the driveway. A kid is coming out of a school bus.
MSVD (∼31 symbols in a text)	1) The cats are fighting 2) The lady sliced a vegetable 3) A man is eating a pizza
LSMDC (∼46 symbols in a text)	1) SOMEONE goes to the kitchen, wets a towel, comes back to the bed, kneels it, places the towel on SOMEONE's brow. 2) He slaps SOMEONE again. 3) SOMEONE moves off through the crowd.

Table 7.3: Three random examples of text descriptions in different datasets. With the dataset name, we also report the median length of a text in the dataset.

Method	Image-Text Datasets	Video-Text Datasets	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC			
			R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
HowTo100M [MZA+19]	-	HowTo100M	7.5	21.2	29.6	38	6.1	17.3	24.8	46	-	-	-	-	-	-	-	-	-	-	-	-
SupportSet [PHA+21]	-	HowTo100M	8.7	23.0	31.1	31	-	-	-	-	-	-	-	-	8.9	26.0	37.9	18	-	-	-	-
VATT [AYQ+21]		HowTo100M+AS	-	-	29.7	49	-	-	45.5	13	-	-	-	-	-	-	-	-	-	-	-	-
EAO§ [SCR+22]	-	HowTo100M	9.9	24.0	32.6	28	**19.8**	**42.9**	**55.1**	**8**	6.6	19.0	26.8	42	18.0	40.4	52.3	9	3.6	8.5	13.0	177
Nagrani et al. [NSS+22]	-	VideoCC3M	19.4	39.5	50.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Frozen in Time [BNVZ21]	CC+COCO	WebVid-2M	24.7	46.9	57.2	7	-	-	-	-	21.1	46.0	56.2	7	-	-	-	-	-	-	-	-
CLIP-straight [PQOBTM21]	WIT	-	31.2	53.7	64.2	4	-	-	-	-	-	-	-	-	37.0	64.1	73.8	**2**	11.3	22.7	29.2	56.5
CLIP4CLIP [LJZ+22]	WIT	HowTo100M	32.0	57.0	66.9	4	-	-	-	-	-	-	-	-	38.5	66.9	76.8	**2**	15.1	28.5	36.4	28
Nagrani et al. [NSS+22]	WIT	VideoCC3M	33.7	57.9	67.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLIP‖ [LLXH22]	CC+COCO+3more*	-	33.3	57.3	67.5	3.5	5.8	15.0	21.9	76	24.6	50.4	59.7	5.3	37.0	63.3	72.6	3	15.2	28.2	35.9	35
In-Style (ours) (CLIP)	WIT	HowTo100M†+VATEX‡	35.0	59.6	70.4	**3**	5.1	14.0	20.3	103	26.6	50.5	62.6	5	38.6	66.3	77.9	3	16.0	**31.6**	38.5	**26.5**
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+VATEX‡	**36.0**	**61.9**	**71.5**	**3**	6.8	16.7	24.5	63	**29.4**	**59.2**	**68.6**	**3.5**	**44.9**	**72.7**	**81.1**	**2**	16.4	30.1	38.7	28
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+WikiHow	34.2	59.6	69.0	**3**	7.3	19.2	27.1	46	29.7	56.2	67.4	4	42.8	70.2	79.1	**2**	**17.0**	30.8	**39.6**	27
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+Food.com	32.8	54.9	65.8	4	7.2	19.8	27.9	47	25.7	52.8	63.1	5	39.5	64.9	74.9	**2**	14.5	28.9	37.2	30.5
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+Target‡	36.2	61.8	71.9	3	8.6	21.6	30.0	37	32.1	61.9	71.2	3	44.8	72.5	81.2	2	16.1	33.6	39.7	25
In-Style (ours) (EAO)	-	HowTo100M+Target†	16.4	35.8	48.9	10	20.3	46.4	58.8	7	13.2	31.6	44.0	15	23.4	50.0	62.4	5	4.9	12.3	16.7	94

Table 7.4: **Zero-shot comparison with other methods. Top**: zero-shot retrieval with methods pre-trained on video-language or/and images-language web or/and curated datasets which exclude target datasets during training. For our In-Style method, the VATEX dataset is used as a source of text queries. **Bottom**: uncurated & unpaired text-video retrieval with text queries from the respective target datasets for comparison purposes. Note that this setting is not zero-shot. † denotes that only videos were used (without paired text) and ‡ – only text (without videos). §For EAO, performance with S3D backbone is reported. ‖For BLIP, the performance of dual encoder architecture is reported (not image-grounded text encoder). *CC [CSDS21]+COCO [LMB+14]+VG [KZG+17]+SBU [OKB11] +LAION [SVB+21]. AS denotes AudioSet [GEF+17].

Training Data	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Average			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
— (zero-shot)	34.1	60.2	70.6	3	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	2	14.5	29.3	36.4	32.5	24.3	44.5	53.3	22.5
Pseudo pairs P_{ps}	35.0	61.4	70.9	3	7.5	19.6	28.9	43	**33.1**	59.8	71.2	**3**	44.3	72.4	81.0	2	16.8	32.7	**40.4**	**25**	27.3	49.2	58.4	15.2
Generated pairs P_{gen}	**36.2**	**61.8**	**71.9**	**3**	8.6	21.6	**30.0**	**37**	32.1	**61.9**	**71.2**	**3**	**44.8**	**72.5**	**81.2**	**2**	16.1	**33.6**	39.7	**25**	**27.6**	**50.3**	**58.8**	**14.0**
Combined P_{ps} + P_{gen}	36.0	61.3	71.5	3	**8.9**	**21.8**	29.8	**37**	32.6	61.8	70.2	3	44.4	72.2	80.8	2	**17.1**	32.4	**40.4**	26	27.8	49.9	58.5	14.2

Table 7.5: **Different types of training pairs for text-video retrieval step.** We evaluate text-video retrieval with pseudo pairs P_{ps} only, with generated pairs P_{gen} only, and the combination of both P_{ps} + P_{gen}.

Training data	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Average			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
— (zero-shot)	34.1	60.2	70.6	**3**	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	**2**	14.5	29.3	36.4	32	24.3	44.5	53.3	22.5
P_{gen} with zero-shot captioner	**36.3**	61.6	71.8	**3**	7.1	18.4	25.6	65	28.7	56.3	65.0	4	43.8	71.2	80.1	**2**	16.0	29.2	37.7	30	26.3	47.3	56.1	20.8
In-Style P_{gen} (non-target)	36.0	61.9	71.5	**3**	6.8	16.7	24.5	63	29.4	59.2	68.6	3.5	44.9	**72.7**	81.1	**2**	**16.4**	30.1	38.7	28	26.7	48.1	56.9	19.9
In-Style P_{gen} (target)	36.2	**61.8**	**71.9**	**3**	**8.6**	**21.6**	**30.0**	37	**32.1**	**61.9**	**71.2**	**3**	**44.8**	72.5	**81.2**	**2**	16.1	**33.6**	**39.7**	**25**	**27.6**	**50.3**	**58.8**	**14**

Table 7.6: **Source of generated pairs P_{gen} for text-video retrieval.** Comparison between zero-shot BLIP (no adaption of retrieval model), zero-shot BLIP captioner, and adapted BLIP captioner with our In-Style method with either text queries from VATEX (non-target) or text queries from the target datasets.

per clip with about 100K for training, while 7408 and 1000 text-video paired samples are used for validation and testing, respectively.

DiDeMo [AHWS+17] is a fine-grained text-video dataset. 10K Flickr videos are paired with multiple detailed sentences (40K sentences in total). During training, we use the single sentences (33K captions), whereas for evaluation on the test set, we follow [BNVZ21] and concatenate all the descriptions for video into one paragraph, acting as a video-paragraph retrieval task (we do not use ground truth time-stamp annotations).

VATEX [WWC+19] dataset contains 35K video clips with multiple annotated captions for a video, covering 600 different human activities. The training set contains 260K captions.

Food.com [MLNM19] is a text dataset that contains more than 230K recipe texts with over 2.2M recipe steps crawled from websites. We use recipe steps as text queries in our training.

WikiHow [KW18] is a large-scale text dataset using the online WikiHow knowledge base. The dataset contains more than 230K articles covering a variety of topics/tasks and descriptions of steps to solve these tasks. We use only headline steps as text quires, which gives us 1.7M captions.

HowTo100M [MZA+19] is a dataset of instructional videos that cover a large variety of topics. The dataset consists of more than 1M videos that were collected by querying on YouTube 23,000 different "how to" tasks. In our default setup, we use 8-second non-overlapping clips from a 100K random subset of the dataset (no more than 15 clips per video) as a support video dataset, resulting in ∼1.4M video clips.

7.5.2 Implementation Details

Model. We leverage the pre-trained dual-encoder CLIP (ViT-B/32) model [RKH+21] in the matching module and the filtering module. Captioner weights are initialized with BLIP (ViT-B/16) captioner [LLXH22] which is pre-trained on five different image-text datasets, including LAION [SVB+21] with 129M images. For retrieval, we consider two architectures: dual encoder image-text initialized with BLIP (ViT-B/16), and dual encoder video-text architecture initialized from EAO [SCR+22] pre-trained on HowTo100M with noisy ASR narrations. We follow [SCR+22] and use a model with a S3D [XSH+18] feature extractor and weights that were pre-trained with a video-text-audio triplet, but only utilize the video-text encoder and report all results without audio.

Training. For training, we uniformly sample $m = 8$ frames per video with a resolution of 224×224, augmented with RandAugment [CZSL20]. For the captioner training and BLIP-architecture retrieval model, we use AdamW optimizer [LH19] with a weight decay of 0.05 and a batch size of 128, and a learning rate $1.0e-05$ for captioner and $1.0e-06$ for retrieval. Following [SCR+22], for the EAO model, we used Adam optimizer [KB15] without weight decay.

Evaluation. For testing, we use $m = 64$ frames for the fine-grained DiDeMo dataset, and $m = 12$ for all others, following [LJZ+22]. For text-video retrieval, we report standard recall metrics for R1, R5, R10, and the median rank (MR).

7.5.3 Text Query Style

We consider text style as a set of attributes and properties of the text shared across a text corpus. Such properties might be the usage of stop words, sentence construction, sentiment, text length, etc. To highlight those differences, we show three text examples from the different datasets in Table 7.3. For example, the YouCook2 test queries always start with an action verb, while in other datasets, the subject+verb+object structure is mostly used. While in the MSR-VTT dataset, frequent words are third person nouns like "man", "woman", "person", "people", the DiDeMo uses more words about camera position like "camera", "left", "right", "screen", "view", and the LSMDC mostly describes a subject as "someone". While the MSR-VTT and the MSVD datasets might look similar, Table 7.3 shows that sentences in the MSR-VTT are 1.5 times longer than in the MSVD on average. We consider such properties as style properties of the text.

7.5.4 Uncurated & Unpaired Text-Video Retrieval

Single Dataset Training. First, we demonstrate the efficiency of the proposed style transfer method in uncurated & unpaired text-video retrieval on five different downstream datasets in Table 7.1. We present results for the image-text pre-trained BLIP [LLXH22] model as well as for the video-text pre-trained EAO [SCR+22] model. We consider three evaluation scenarios: 1) zero-shot performance; 2) the performance of our style transfer method in the text-video retrieval task with uncurated & unpaired data where only text queries are available during training; 3) training with the ground truth aligned text-video pairs, which can be considered as an upper bound for our task. It shows that the proposed In-Style method significantly outperforms zero-shot performance even without using any aligned training samples from the target distribution. This supports the hypothesis that the style of the text queries is an important component of text-video retrieval. Moreover, we observe that the gap between training with ground truth aligned pairs and the style transfer can be remarkably small, especially on the MSVD dataset, indicating the benefits with respect to a potential annotation cost reduction in the proposed setup.

Multi-Dataset Training. Second, we evaluate the proposed multi-dataset training procedure with the In-Style method in Table 7.2. Here, a minibatch is compiled from a single text source as shown in Figure 7.3. This is favorable compared to the standard training, where data points in a minibatch are randomly sampled from all data sources together. It shows that the proposed procedure leads to improved retrieval performance compared to individually trained models and better generalization across all datasets compared to the standard multi-dataset training. We attribute the performance increase compared to standard multi-dataset training to the fact that considering the captions of only the same style in contrastive loss provides a model with a

cleaner learning signal with stronger text negative counterparts. As an example, "add sliced cucumber" in YouCook2 style would be a stronger negative in comparison to a correct "add sliced tomato" query than a "a person in a video puts sliced cucumber in a salad" in MSR-VTT style.

7.5.5 Comparison with SOTA

We further compare the proposed method with zero-shot retrieval baselines in Table 13.1. We report the performance of BLIP and CLIP backbones trained with text queries from the VATEX dataset, thus text queries do not follow distribution of any of the test datasets. The closest counterpart to our model is Nagrani et al. method [NSS+22], which utilizes the pre-trained image-text CLIP backbone, which is further trained with the VideoCC3M dataset [NSS+22] – a video-text dataset collected by automatic transferring image captions from text-image CC3M dataset [SDGS18]. The conceptual difference between [NSS+22] and our method is that [NSS+22] proposes to transfer *image* captions from the image-caption dataset by pairing images to videos, while the proposed In-Style method adapts the model to the *video* captions. While noting that a direct comparison to different state-of-the-art methods is limited due to different pre-training datasets, it can be observed that the proposed In-Style method achieves the best results on four out of five datasets, underperforming only in YouCook2, which might benefit from HowTo100M pretraining. We additionally validate the statement that text queries can be used without any corresponding videos by using texts from WikiHow [KW18] and Food.com [MLNM19] datasets that contain descriptions of different actions/steps to solve tasks or cook meals. In Table 13.1, we show that style transfer from both datasets especially benefits YouCook2 retrieval performance that we attribute to the similarity in text styles. However, style transfer from the WikiHow, which is more diverse and covers a larger variety of topics, also improves the performance over the baselines on the DiDemo, MSVD, and LSMDC datasets.

7.5.6 Efficiency of Style Transfer

Training Pairs. In Table 7.5, we compare the performance of the models trained either with pseudo pairs P_{ps} or with generated pairs P_{gen}, or with a combination of them $P_{ps} + P_{gen}$. All setups boost the performance of text-video retrieval by a large margin compared to zero-shot text-video retrieval. The generated pairs P_{gen} achieve a better performance than pseudo pairs P_{ps} on all datasets except LSMDC, whereas a combination of $P_{ps} + P_{gen}$ does not improve performance on average. We note that the number of pairs in P_{gen} is significantly larger than in P_{ps} (Table 7.8) for all datasets except LSMDC (a dataset of movies, which might contain a larger domain shift to YouTube videos compared to other datasets). We assume that in this case P_{gen} contains better-aligned pairs since each generated text description is conditioned on the corresponding video, while in P_{ps} a fixed set of descriptions is matched (see examples in Figure 7.4) explaining the performance drop with $P_{ps} + P_{gen}$.

Style Transfer. In Table 7.6, we consider how much the text style transfer in the generated pairs P_{gen} influences the retrieval performance. For this, we considered three sets of P_{gen} for the training retrieval model: 1) P_{gen} generated with zero-shot BLIP captioner; 2) In-Style P_{gen} generated with captioner trained on P_{ps} with text queries from a different non-target dataset (we used the VATEX dataset); 3) In-Style P_{gen} with a captioner trained on P_{ps} with text queries from the target dataset. We observe that training the model with generated text-video pairs (from uncurated web videos from the HowTo100M dataset) by a zero-shot image-pretrained

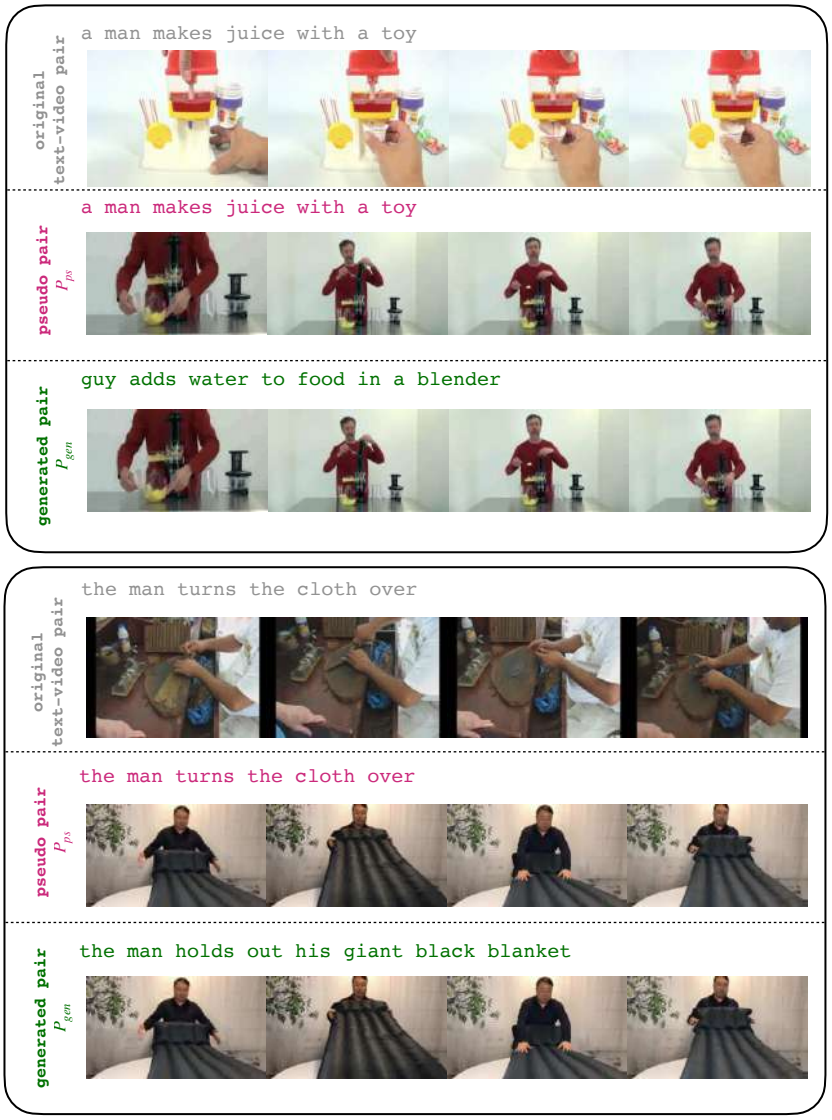


Figure 7.4: **Qualitative evaluation of P_{ps} and P_{gen} on the MSR-VTT (left) and DiDeMo (right) datasets.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen}).

Filt. Thr.	R1	R5	R10	MR
0.26	43.9	71.8	80.8	**2**
0.27	44.2	72.2	80.9	**2**
0.28	44.8	**72.5**	**81.2**	**2**
0.29	45.0	72.3	80.9	**2**
0.30	**45.1**	72	80.8	**2**

Table 7.7: Filtering threshold

Dataset	#Pseudo Pairs	#Generated Pairs
MSR-VTT	180k	495k
YouCook	10k	168k
Didemo	33k	280k
MSVD	48k	379k
LSMDC	101k	144k

Table 7.8: Number of P_{ps} and P_{gen}

Matching	R1	R5	R10	MR
Random	39.1	66.3	75.8	**2**
BLIP	44.1	71.4	80.0	**2**
CLIP	**44.8**	**72.5**	**81.2**	**2**

Table 7.9: Matching method

captioner already improves the performance in all video retrieval datasets. We attribute this to the content and style adaptation of the image-language model to the specific appearances in the videos. However, such models tend to generate "static" descriptions that do not involve actions. Thus, text queries from non-target video datasets, namely the VATEX dataset, improve the retrieval performance further. Yet, we notice that YouCook2 does not benefit from the VATEX text queries as from the zero-shot generated captions. Finally, using training text queries from the target dataset excels on the considered benchmarks.

7.5.7 Ablation Study

Matching Method. To obtain generated pairs, we train the captioner with pseudo pairs that were created by a matching module. In Table 7.9, we consider two options for the matching module: image-text pre-trained dual encoders from BLIP [LLXH22] and CLIP [RKH+21], as well as the "Random" option where text queries are simply matched with the random videos. We report the text-video retrieval performance of our final model using the given option of the matching module. We observe that matching module based on CLIP leads to better performance. We attribute that to the robustness of CLIP to the noisy web data as it was trained on large-scale web image-text pairs, whereas BLIP utilizes additional filtering to reduce the noise in the training.

Filtering Threshold. In Table 7.7, we consider the effect of filtering on the quality of the

Training pairs	B@4	ROUGE	CIDEr
– (zero-shot)	0.305	0.519	0.610
Pseudo pairs	0.559	0.628	1.059
GT pairs	0.659	0.680	1.296

Table 7.10: Captioning performance

generated pairs P_{gen}. We find threshold $th = 0.28$ works the best, indicating that filtering is an important step for our style transfer framework.

Captioning Perforformance Finally, we evaluate the captioning performance of the captioner trained with pseudo pairs P_{ps} with the standard NLP metrics BLEU@4, ROUGE and CIDEr. Table 7.10 demonstrates that the captioner trained with pseudo pairs almost doubles the zero-shot captioner performance, significantly reducing the gap to the training with ground truth supervision.

7.6 Conclusion

In this chapter, we address a new task of *text-video retrieval with uncurated & unpaired data*, where during training only text queries are available. Motivated by the fact that different domains imply diverse styles of video descriptions, we introduce the In-Style method that preserves the style of the given input queries and transfers it to the support set of unrelated web videos, creating aligned text-video pairs with the style of the input. Utilization of obtained text-video pairs as supervision leads to a significant performance boost in text-video retrieval. Moreover, we show the performance generalization of a single model that we train with multiple styles simultaneously, proposing a training procedure for multi-dataset training. We evaluate the proposed model over multiple datasets and show the advantages of the In-Style method on the task of uncurated & unpaired text-video retrieval and achieve new state-of-the-art results for zero-shot text-video retrieval. In Chapter 8, we will discuss a method for aligning image and video modalities and utilizing webly annotated image data to enhance action recognition. Meanwhile, in Chapter 9, we will further explore contributions related to large-scale video-language pretraining, particularly focusing on enhancing web video datasets for pretraining.

8 CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video

Contents

To further investigate different modalities, we examine image-to-video adaptation, aiming to exploit label-free web images as a source for adapting to unlabeled target videos. This poses two major challenges: (1) spatial domain shift between web images and video frames; (2) modality gap between image and video data. To address these challenges, we propose Cycle Domain Adaptation (CycDA), a cycle-based approach for unsupervised image-to-video domain adaptation. We leverage the joint spatial information in images and videos on the one hand and, on the other hand, train an independent spatio-temporal model to bridge the modality gap. We alternate between the spatial and spatio-temporal learning with knowledge transfer between the two in each cycle. We evaluate our approach on benchmark datasets for image-to-video as well as for mixed-source domain adaptation achieving state-of-the-art results and demonstrating the benefits of our cyclic adaptation.

This chapter is based on [LKS+22]. As a second author, Anna Kukleva contributed to the project in the scientific discussions, writing of the paper and creating the figures.

8.1 Introduction

The task of action recognition has seen tremendous success in recent years with top-performing approaches typically requiring large-scale labeled video datasets [Fei20, WSS21, YXS+20], which can be impractical in terms of both data collection and annotation effort. In the meanwhile, webly-supervised learning has been explored to leverage the large amount of easily accessible web data as a labeling-free data source for video recognition [GSN17, GHZ+18, LWZK17, WXLVG17, YSL+18, ZLL+17].

In this work, we address the problem of image-to-video adaptation with webly-labeled images as the source domain and unlabeled videos as the target domain to allow for action classification without video annotation. This setting provides two major challenges: (1) the spatial domain shift between web images and video frames, based on difference in image

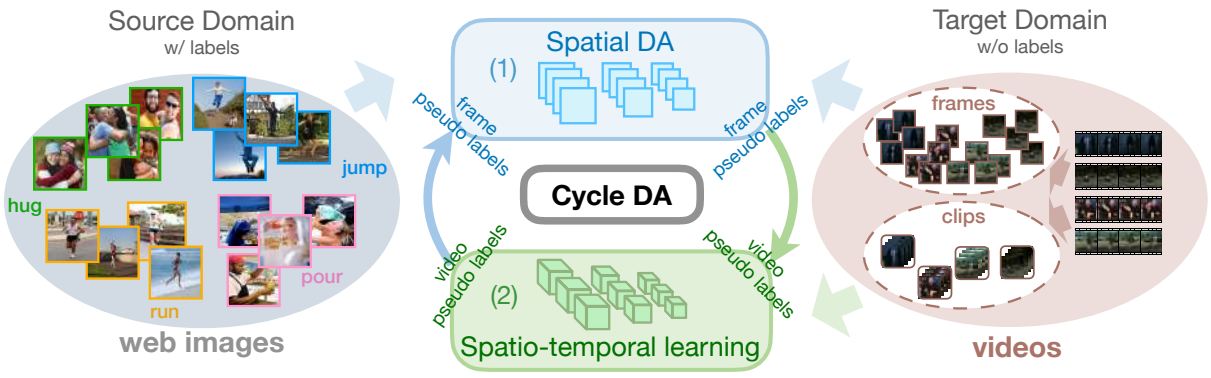


Figure 8.1: **Cycle Domain Adaptation (CycDA) pipeline**. We address image-to-video adaption by training a spatial model and a spatio-temporal model alternately, passing pseudo labels to supervise each other in a cycle. The two alternating steps are: (1) domain alignment on the spatial model with pseudo labels from the spatio-temporal model, and (2) training the spatio-temporal model with updated pseudo labels from the spatial model.

styles, camera perspectives and semantic drifts; (2) the modality gap between spatial images and spatio-temporal videos. Specifically, this modality gap restrains that merely spatial knowledge can be transferred from source to target domain. Previous works on action recognition with web supervision either learn from web data directly [GSDG16, GYY+16] or perform joint training by combining the web source with annotated target data [DZX+20, MBZ+17]. To specifically address the domain shift between web images and target videos, some approaches perform class-agnostic domain-invariant feature learning either within [KS20] or across modalities [YWSD18, YWCD19, LOW+20], in the absence of ensuring domain-invariance on the category-level.

In this context, we propose Cycle Domain Adaptation (CycDA), i.e. alternating knowledge transfer between a spatial model and a spatio-temporal model. Compared to other works, we address the two challenges at hand, domain-alignment and closing the modality gap in separate stages, cycling between both of them. An overview of the CycDA is given in Fig. 8.1. With the category knowledge from the spatio-temporal model, we achieve enhanced category-level domain invariance on the spatial model. With updated knowledge transferred from the spatial model, we attain better spatio-temporal learning. In this manner, we can better tackle each challenge for the corresponding model, with the updated knowledge transferred from the other.

More specifically, we propose a four stage framework to address the domain shift between images and videos on different levels. In stage 1, we enforce *class-agnostic* domain alignment on the spatial model between images and video frames. In stage 2, we use supervision from the spatial model to learn a spatio-temporal video model, bridging the gap between the two modalities. Stage 3 then focuses on *class-aware* domain alignment on the spatial model, given pseudo labels computed by the video model trained on stage 2. In stage 4, we update the video model with the improved pseudo labels from the spatial model of stage 3.

We first evaluate our approach on several challenging settings for web image based action recognition, where a single cycle already outperforms baselines and state-of-the-arts. Second, we show how CycDA can be flexibly applied for mixed-source image&video-to-video DA settings, leading to a performance competitive to the state-of-the-art requiring only 5% of the provided source videos.

We summarize our contributions as follows: (1) We propose to address web image-to-

video domain adaptation by decoupling the domain-alignment and spatio-temporal learning to bridge the modality gap. (2) We propose cyclic alternation between spatial and spatio-temporal learning to improve spatial and spatio-temporal models respectively. (3) We provide an extensive evaluation with different benchmark tasks that shows state-of-the-art results on unsupervised image-to-video domain adaptation and a competitive performance for the mixed-source image&video-to-video setting.

8.2 Related Work

In this section, we discuss prior work on the image-to-video domain adaptation and video-tovideo domain adaptation. We will not revisit the topic of multimodal pretraining as previously discussed in Chapter 2.

Image-to-video Domain Adaptation. Compared to webly-supervised learning, image-to-video DA approaches actively address the domain shift between web images and target videos either by spatial alignment between web images and video frames [LWZK17, SSSN15, ZHT+16], or through class-agnostic domain-invariant feature learning for images and videos [LOW+20, YWCD19, YWSD18]. Li et al. [LWZK17] use a spatial attention map for cross-domain knowledge transfer from web images to videos. In this case, the DA is addressed on the spatial level without transfer to the temporal level. Liu et al. [LOW+20] perform domain-invariant representation learning for images, video keyframes and videos, and fuse features of different modalities. Furthermore, hierarchical GAN [YWSD18], symmetric GAN [YWCD19] and spatio-temporal causal graph [CWHL21] are proposed to learn the mapping between image features and video features. Closest to our work is probably the work of Kae et al. [KS20] which also employs a spatial and a spatio-temporal model for two stages of class-agnostic domain alignment, proposing to copy the weights from the spatial to the spatio-temporal model. In contrast to these, we propose to transfer knowledge in the form of pseudo labels, without enforcing spatial information from the DA stage onto the spatio-temporal model.

Video-to-video DA. Compared to image-to-video adaption, video-to-video adaptation methods adapt annotated source videos to unlabeled target videos [CKA+19, SSP+21, KTZ+21, CSSH20, CSCH20, MD20, LHW+20, PCAN20, JNDV18], focusing mainly on the problem of domain alignment. Chen et al. [CKA+19] align the features spatially on the frame-level and temporally on the scale-level. Others propose feature alignment via self-attention [CSSH20], cross-domain co-attention [PCAN20], or across two-stream modalities of RGB and optical flow [KTZ+21, MD20]. Sahoo et al. [SSP+21] propose temporal contrastive learning and background mixing for domain-invariance. In this work, we focus on image-to-video adaptation and use only single stream of RGB. We further show that we are able to extend the pipeline to the mixed-source case, where we achieve competitive performance compared to video-to-video adaptation methods while requiring only a small amount of source videos.

8.3 Method

We propose four stages to tackle image-to-video adaptation, which we summarize in Sec. 8.3.1. Afterwards, we detail each stage and motivate the cycling of stages in Sec. 8.3.2. Our CycDA can be flexibly extended (Sec. 8.3.3) for mixed-source video adaptation, where a limited amount of annotated source videos are available.

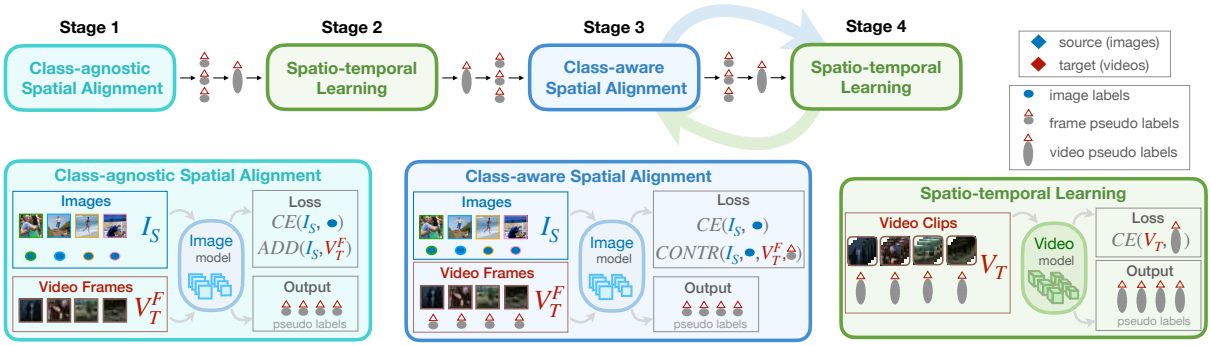


Figure 8.2: **Our CycDA framework.** Our method alternates between spatial alignment (stage 1 and 3) and spatio-temporal learning (stage 2 and 4). See text for details.

8.3.1 System Overview

The task of unsupervised image-to-video DA is to learn a video classifier given labeled source images and unlabeled target videos. In order to close the domain gap across these two different modalities, we employ (1) a spatial (image) model to train on source web images and frames sampled from target videos, and (2) a spatio-temporal (video) model to train on target video clips. We propose a training pipeline that alternately adapts the two models by passing pseudo labels to supervise each other in a cycle. This facilitates the knowledge transfer between both models, where pseudo labels efficiently guide the model through the corresponding task, i.e. semantic alignment (image model) or spatio-temporal learning (video model).

As shown in Fig. 8.2, our CycDA pipeline consists of four training stages. The initial pseudo labels from *class-agnostic* domain alignment (Stage 1) are improved via video spatio-temporal learning (Stage 2). In Stage 3, these pseudo labels on target data, together with ground truth labels on source, enable semantic alignment through the class-aware domain alignment. This again refines the pseudo labels, which further provide enhanced supervision for spatio-temporal learning in stage 4. In this manner, one iteration of CycDA facilitates the alternating knowledge transfer between the image and video models, whereby pseudo labels are improved on each stage and further provide strengthened supervision for better learning on the next stage.

Notations: First, we denote the feature extractor as $E(\cdot;\theta_E)$, the classifier as $C(\cdot;\theta_C)$, and the domain discriminator as $D(\cdot;\theta_D)$. Then, we have the image model $\phi^I = \{E^I(\cdot;\theta_E^I),\ C^I(\cdot;\theta_C^I),\ D^I(\cdot;\theta_D^I)\}$ and the video model $\phi^V = \{E^V(\cdot;\theta_E^V),\ C^V(\cdot;\theta_C^V)\}$. We use the superscripts I, V and F to denote modalities of image, video and video frame, correspondingly. S and T stand for *source* and *target* domains respectively. The labeled source image domain is denoted as $I_S = \{(i_j, l(i_j))|_{j=1}^{N_S^I}\}$, where $l(\cdot)$ is the ground truth label of the corresponding image. The unlabeled target video domain is $V_T = \{v_j|_{j=1}^{N_T^V}\}$ and each video v_j has M_j frames, the set of frames of unlabeled target videos $V_T^F = \{\{v_{j,m}^F|_{m=1}^{M_j}\}|_{j=1}^{N_T^V}\}$.

8.3.2 Stages

Stage 1 - Class-agnostic Spatial Alignment. In the first stage, we learn the class-agnostic domain alignment between source web images and frames sampled from unlabeled target

videos. Thus, we reduce the domain gap between the appearance of the web images and target videos even if the classes could be incorrectly aligned during this stage. We train the image model ϕ^I with a supervised **c**ross **e**ntropy loss $\mathcal{L}_{CE}(I_S)$ and an **a**dversarial **d**omain **d**iscrimination loss $\mathcal{L}_{ADD}(I_S, V_T^F)$ on source images and target frames. With the classification loss on source images given as

$$\mathcal{L}_{CE}(I_S) = \sum_{(i_j, l(i_j)) \in I_S} -l(i_j) \cdot \log(C^I(E(i_j; \theta_E^I); \theta_C^I)) \tag{8.1}$$

and the binary cross entropy loss for domain discrimination given as

$$\mathcal{L}_{ADD}(I_S, V_T^F) = \sum_{i_j, v_{j',m}^F} \log D^I(E^I(i_j; \theta_E^I); \theta_D^I) + \log(1 - D^I(E^I(v_{j',m}^F; \theta_E^I); \theta_D^I)), \tag{8.2}$$

the overall objective is $\min_{\theta_E^I, \theta_C^I} \mathcal{L}_{CE}(I_S) + \beta \max_{\theta_E^I} \min_{\theta_D^I} \mathcal{L}_{ADD}(I_S, V_T^F)$, where β is the trade-off weight between the two losses. We train the domain discriminator D^I to distinguish between extracted features from different domains, while the feature extractor is trained adversely based on the domain discrimination task. In this case, the feature extractor learns to yield domain invariant features that the domain discriminator is unable to differentiate. The adversarial training is performed with a gradient reversal layer (GRL) [GL15, GUA+16] which reverses the sign of gradients of the domain discrimination loss on the feature extractor during back-propagation. The domain alignment on this stage is class-agnostic as there is not yet any pseudo label for category knowledge in the target domain. The alignment is performed globally at the domain level.

Stage 2 & Stage 4 - Spatio-Temporal Learning. In this stage, we use the trained image model ϕ^I from the previous stage to generate pseudo labels for the target videos. Then we perform supervised spatio-temporal learning with the pseudo labeled target data.

Specifically, we first use ϕ^I to predict the pseudo label $\hat{l}(\cdot)$ for each frame of the target videos. We employ a spatio-temporal model that trains on target videos capturing both spatial and temporal information in the target domain only. To select new pseudo label candidates, we temporally aggregate frame-level predictions into a video-level prediction. We discard predictions with confidence lower than a threshold δ_p and perform a majority voting among the remaining predictions to define the video label. From all videos, we only keep those that have at least one frame with a minimum confidence. We set the confidence threshold δ_p such that $p \times 100\%$ of videos remain after the thresholding.

We denote the target video set after thresholding as $\tilde{V}_T$. For stage 2, the supervised task on pseudo labeled target videos is to $\min_{\theta_E^V, \theta_C^V} \hat{\mathcal{L}}_{CE}(\tilde{V}_T)$, with

$$\hat{\mathcal{L}}_{CE}(\tilde{V}_T) = \sum_{v_j \in \tilde{V}_T} -\hat{l}(v_j) \cdot \log(C^V(E^V(V_j; \theta_E^V); \theta^V)). \tag{8.3}$$

In stage 4, we repeat the process as described above and re-train the video model on the target data with the updated pseudo labels from the third stage.

Stage 3 - Class-aware Spatial Alignment. The adversarial learning for domain discrimination in the first stage aligns features from different domains globally, but not within each category (c.f. Fig. 8.3 (b) and (c)). In this case, target samples in a category A can be incorrectly aligned with source samples in a different category B. This would lead to inferior classification performance of the target classifier. To evade this misalignment, we perform class-aware domain alignment in the third stage between the source web images and the target video

frames. Since the source data consists exclusively of images, we apply alignment on the spatial model between images and frames. Furthermore, as the target data is unlabeled, in order to align features across both domains within each category, we generate pseudo labels by the model ϕ^V from the second stage to provide category knowledge. Specifically, we use the video model to generate video-level labels that we disseminate into frame-level labels. To align images and video frames we use cross-domain contrastive learning by maximizing the similarity between samples across domains of the same class and minimizing the similarity between samples from different classes. We use $z = E^I(i;\theta_E^I)$ to denote the feature computed by the feature extractor on image i. The set of source image features is $Z_S^I = \{E^I(i;\theta_E^I)|i \in I_S\}$ and the set of target frame features is $Z_T^F = \{E^I(v^F;\theta_E^I)|v^F \in V_T^F\}$. During training, for each pseudo labeled target sample $z_j^F \in Z_T^F$, we randomly choose two samples from the source domain: a positive sample of the same label and a negative sample of a different label, i.e. $z_{j\,+}^I, z_{j\,-}^I \in I_S$. The contrastive loss is formulated as

$$\mathcal{L}_{CONTR}(I_S, V_T^F) = -\sum_{z_j^F \in Z_T^F} \log \frac{h(z_j^F, z_{j\,+}^I)}{h(z_j^F, z_{j\,+}^I) + h(z_j^F, z_{j\,-}^I)}. \tag{8.4}$$

Following [CKNH20], we set $h(u,v) = \exp(\text{sim}(u,v)/\tau)$, where we use the cosine similarity $\text{sim}(u,v) = u^\top v/(\|u\|\|v\|)$ and τ is the temperature parameter. Thus, the objective of stage 3 on the image model is $\min_{\theta_E^I, \theta_C^I} \mathcal{L}_{CE}(I_S) + \mathcal{L}_{CONTR}(I_S, V_T^F)$.

In the third stage, an alternative of exploiting pseudo labels from the video model from the second stage is to self-train the video model on the target data, as self-training is a common practice in DA [LWL21, ZDJZ20, ZYKW18, ZYL+19]. However, the category-level domain alignment with supervision from the source domain further regularizes the learning of class distribution in the target domain. This results in a significantly improved target classifier, as we show in Sec. 8.4.5.

Cycling of the Stages. The pseudo labels from the video model are exploited for class-aware domain alignment on the image model (stage 3) and the updated pseudo labels from the image model can supervise the training of the video model (stage 4). In this manner, stage 3 and stage 4 can be performed iteratively. We show in the evaluation (Table 8.1 and Fig. 8.5) the impact of this cyclic learning setup and how several iterations of CycDA can further improve the performance of the target classifier.

8.3.3 Mixed-source Video Adaptation

Image-to-video DA applies to the case in which the source domain consists only of web images. However, other possible settings presume limited amount of annotated videos with the domain shift to the unlabeled target videos. We refer to this case as mixed-source video adaptation. CycDA can be adjusted for this setting as follows. We denote the labeled source video domain as $V_S = \{(v_j, l(v_j))|_{j=1}^{N_S^V}\}$. For the class-agnostic (stage 1) and class-aware domain alignment (stage 3) stages we replace the source image domain $\{I_S\}$ by the mixed-source domain data $\{I_S, F_S\}$ which consists of web images and frames sampled from source videos. The supervised classification, adversarial domain discrimination and cross-domain contrastive learning are adapted accordingly. For the spatio-temporal learning of the video model ϕ^V (stage 2 and 4) we include additional supervised classification w.r.t. the ground truth labels for the source videos, therefore the overall loss is $\mathcal{L}_{CE}(V_S) + \hat{\mathcal{L}}_{CE}(\tilde{V}_T)$. In this case, the annotated source videos are utilized to regularize domain alignment on the image model, and provide further supervision

for learning the classification task on the video model. In Sec. 8.4.4, we demonstrate that in the context of mixed-source video adaptation, even a limited amount of source videos is sufficient to achieve results competitive to video-to-video adaptation approaches that employ the entire source video dataset.

8.4 Experiments

8.4.1 Datasets

To evaluate our CycDA framework for image-to-video adaptation, we conduct experiments on 3 real-world image-video action recognition benchmark settings. Videos are from two large-scale action recognition datasets, UCF101 [SZS12] and HMDB51 [KJG+11]. Web images are from the EADs (Extensive Action Dataset) [CWHL21], Stanford40 [YJK+11] and the BU101 dataset [MBZ+17].

The three image-to-video adaptation tasks are: (1) Stanford40 → UCF101: the UCF101 action dataset contains 13320 videos collected from YouTube with 101 action classes. The Stanford40 dataset contains 9532 images collected from Google, Bing and Flickr, comprised of 40 action classes. Following [CWHL21, YWCD19, YWSD18], we select the 12 common action classes between the two datasets for image-to-video action recognition. (2) EADs→HMDB51: HMDB51 has 6766 videos with 51 action classes collected from online videos and movie clips. The EADs dataset consists of Stanford40 and the HII dataset [TZIC16]. It has 11504 images from 50 action classes. There are 13 shared action classes between the two datasets. (3) BU101→UCF101: BU101 consists of 23.8K web action images from 101 classes that completely correspond to classes on UCF101. We use data of all classes for evaluation on large-scale image-to-video adaptation.

UCF101 and HMDB51 both have three splits of training and test sets. Following [LWZK17, YWSD18, YWCD19], we report the average performance over all three splits.

The UCF-HMDB dataset [CKA+19] is a benchmark for video-to-video DA. It consists of the 12 common classes between UCF101 and HMDB51. On this dataset, we perform two types of evaluations: (1) *frame-to-video adaptation*: we use only a single frame from each source video to adapt to target videos; and (2) mixed-source video adaptation: we use source and target videos of UCF-HMDB, and extend the source domain with web images from BU101.

8.4.2 Implementation Details

For the image model, we use a ResNet-18 [HZRS16] pretrained on ImageNet [DDS+09]. We freeze the first 6 sequential blocks and train with a learning rate of 0.001 to perform the domain alignment between web images and frames sampled from target videos. To avoid redundancy in the video frames, we uniformly divide a video into 5 segments. In each training epoch, we randomly sample one frame from each segment. As trade-off weight for domain discrimination, we follow the common practices in [GL15, GUA+16, CLB+20] to gradually increase β from 0 to 1. The temperature parameter τ is set to 0.05.

For the video model, we employ I3D Inception v1 [CZ17] pretrained on the Kinetics dataset [KCS+17], which is common practice, e.g. [CWHL21, CSSH20, KTZ+21, MD20, SSP+21]. We train the RGB stream only. To validate the efficacy of our CycDA pipeline, we use the I3D backbone with a shallow classifier of 2 FC layers, without any temporal aggregation module (e.g. GCN in [SSP+21] or self-attention module in [CSSH20]). We extract a clip of 64 frames

Method	Backbone	E→H	S→U	B→U
source only	ResNet18	37.2	76.8	54.8
DANN [GUA+16]*	ResNet18	39.6	80.3	55.3
UnAtt [LWZK17]	ResNet101	-	-	66.4
HiGAN [YWSD18]	ResNet50, C3D	44.6	95.4	-
SymGAN [YWCD19]	ResNet50, C3D	55.0	97.7	-
CycDA (1 iteration)	ResNet50, C3D	56.6	98.0	-
DANN [GUA+16]+I3D*	ResNet18, I3D	53.8	97.9	68.3
HPDA [CWHL21]*	ResNet50, I3D	38.2	40.0	-
CycDA (1 iteration)	ResNet18, I3D	60.5	99.2	69.8
CycDA (2 iterations)	ResNet18, I3D	60.3	**99.3**	72.1
CycDA (3 iterations)	ResNet18, I3D	**62.0**	99.1	**72.6**
supervised target	ResNet18, I3D	83.2	99.3	93.1

Table 8.1: Results on E→H (13 classes), S→U (12 classes) and B→U (101 classes), averaged over 3 splits. ResNet, C3D and I3D are pretrained on ImageNet[HZRS16], Sports-1M[KTS+14] and Kinetics400[KCS+17]. * denotes our evaluation.

from each video. Following [SSP+21], we use a learning rate of 0.001 for the backbone and 0.01 on other components. We keep $p = 70\%$ and 80% of videos in stage 2 and stage 4.

8.4.3 Image-to-video DA

We compare the proposed approach to other image-to-video adaptation methods on the three described benchmark settings as shown in Table 8.1. As CycDA enables the iterative knowledge transfer between the image model and the video model, we can repeat stage 3 and stage 4 multiple times. We therefore report the performance for the first three iterations. We add the lower bound (source only) and the upper bound (ground truth supervised target) for reference.

We compare against several approaches: DANN [GUA+16] is classical adversarial domain discrimination on the image-level. UnAtt [LWZK17] applies a spatial attention map on video frames. HiGAN [YWSD18] and SymGAN [YWCD19] employ GANs for feature alignment (on backbone of ResNet50 and C3D) and define the current state-of-the-art on E→H and S→U. We also evaluate CycDA with the same backbones for fair comparison. DANN [GUA+16]+I3D is a strong baseline that trains the I3D model with pseudo labels from an adapted image model. HPDA [CWHL21] is a recent partial DA approach and for a fair comparison, we re-run its official implementation in a closed-set DA setting. Our CycDA outperforms all other approaches already after the first iteration. Except for the saturation on S→U, running CycDA for more iterations leads to a further performance boost on all evaluation settings.

We further explore the potential of CycDA on UCF-HMDB, which is a benchmark for video-to-video adaptation. For a strict comparison, we select data from the same source video dataset used in the video-to-video adaptation methods, without using any auxiliary web data for training. However, instead of directly using the source videos, we perform *frame-to-video* adaptation where we use only one frame from each source video to adapt to target videos. Here we sample the middle frame from each video and report the results in Table 8.2 (case B). We see that even when using only one frame per source video, on U→H, CycDA (83.3%) can already outperform TA3N [CKA+19] (81.4%) and SAVA [CSSH20] (82.2%) which use all source

	DA setting	Method	Video backbone	Source data		U→H	H→U
				web image	videos (U or H) in %		
A:	video-to-video	AdaBN [LWS+18]	ResNet101	-	100%	75.5	77.4
		MCD [SWUH18]	ResNet101	-	100%	74.4	79.3
		TA^{3}N [CKA+19]	ResNet101	-	100%	78.3	81.8
		ABG [LHW+20]	ResNet101	-	100%	79.1	85.1
		TCoN [PCAN20]	ResNet101	-	100%	87.2	89.1
		DANN [GUA+16]	I3D	-	100%	80.7	88.0
		TA^{3}N [CKA+19]	I3D	-	100%	81.4	90.5
		SAVA [CSSH20]	I3D	-	100%	82.2	91.2
		MM-SADA [MD20]	I3D	-	100%	84.2	91.1
		CrossModal [KTZ+21]	I3D	-	100%	84.7	92.8
		CoMix [SSP+21]	I3D	-	100%	86.7	93.9
B:	frame-to-video	CycDA	I3D	-	one frame	83.3	80.4
C:	mixed-source to video	CycDA	I3D	BU*	0%	77.8	88.6
				BU*	5%	82.2	93.1
				BU*	10%	82.5	93.5
				BU*	50%	84.2	95.2
				BU*	100%	88.1	98.0
	supervised target		I3D	-	-	94.4	97.0

Table 8.2: Results of Cycle Adaption on UCF-HMDB in comparison to video-to-video adaptation (case A) approaches. For frame-to-video adaptation (case B), we use only one frame from each source video to adapt to target videos. For mixed-source video adaptation (case C), we combine BU101 web images and source videos as the source data. *We sample 50 web images per class from 12 classes in BU101.

videos for video-to-video adaptation. This demonstrates the strength of CycDA to exploit the large informativity in single images such that they could potentially replace videos as the source data. On H→U, our source domain contains only 840 frames from the 840 videos in the HMDB training set on UCF-HMDB, which leads to an inferior performance. We show that this can be easily addressed by adding auxiliary web data in Sec. 8.4.4.

8.4.4 Mixed-source image&video-to-video DA

For mixed-source adaptation, we assume to have both web images and some amount of source videos in the source domain. To evaluate this case, we use the source and target videos on UCF-HMDB, and extend the source domain with web images of the 12 corresponding action classes in BU101. We notice that using all web data from BU101 leads to performance saturation on the target video set of UCF-HMDB. Therefore, to validate the efficacy of CycDA, we only sample 50 web images per class as auxiliary training data. We vary the amount of source videos in the mixed-source domain and report the results in Table 8.2 (case C). First, by training with only sampled web images (without any source videos), we achieve baseline results of 77.8% (BU→H) and 88.6% (BU→U). By adding only 5% of videos to the mixed-source domain, we already achieve performance comparable to the video-to-video adaptation methods, i.e. 82.0% (BU+U→H) and 93.1% (BU+H→U). Furthermore, increasing the amount of source videos from 5% to 50% leads to another improvement of 2%. As web images are more informative than sampled video frames, using web images as auxiliary training data can thus significantly reduce the amount of videos required. Finally, with sampled web images and all source videos,

	Experiment	stage 1	stage 2	stage 3	stage 4	Acc
A:	source only	$\mathcal{L}_{CE}(I_S)$	-	-	-	39.0
B:	source only + video model	$\mathcal{L}_{CE}(I_S)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	-	50.5
C:	class-agnostic DA + video model	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	-	52.3
D:	case C + vid. self-train×1	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	55.4
E:	case C + vid. self-train×2	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	56.4
F:	CycDA	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{CONTR}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	**60.8**

Table 8.3: Stage-wise ablation study of the CycDA training pipeline on EADs → HMDB51 split 1. A: source only on image model. B: source only training on image model and video model training. C: class-agnostic DA (stage 1) and video model training (stage 2). D: case C + one stage of self-training the videos model. E: case C + two stages of self-training the video model. F: CycDA with stage 1∼4. Category-level pseudo labels of case C and F are compared in Fig. 8.4.

we outperform all video-to-video adaptation methods, even exceeding the supervised target model for BU+H→U by 1%. Considering that we only use 50 web images per class, this further demonstrates that CycDA can exploit both, the information in web images and knowledge from the source data with domain shift, for a potentially improved learning.

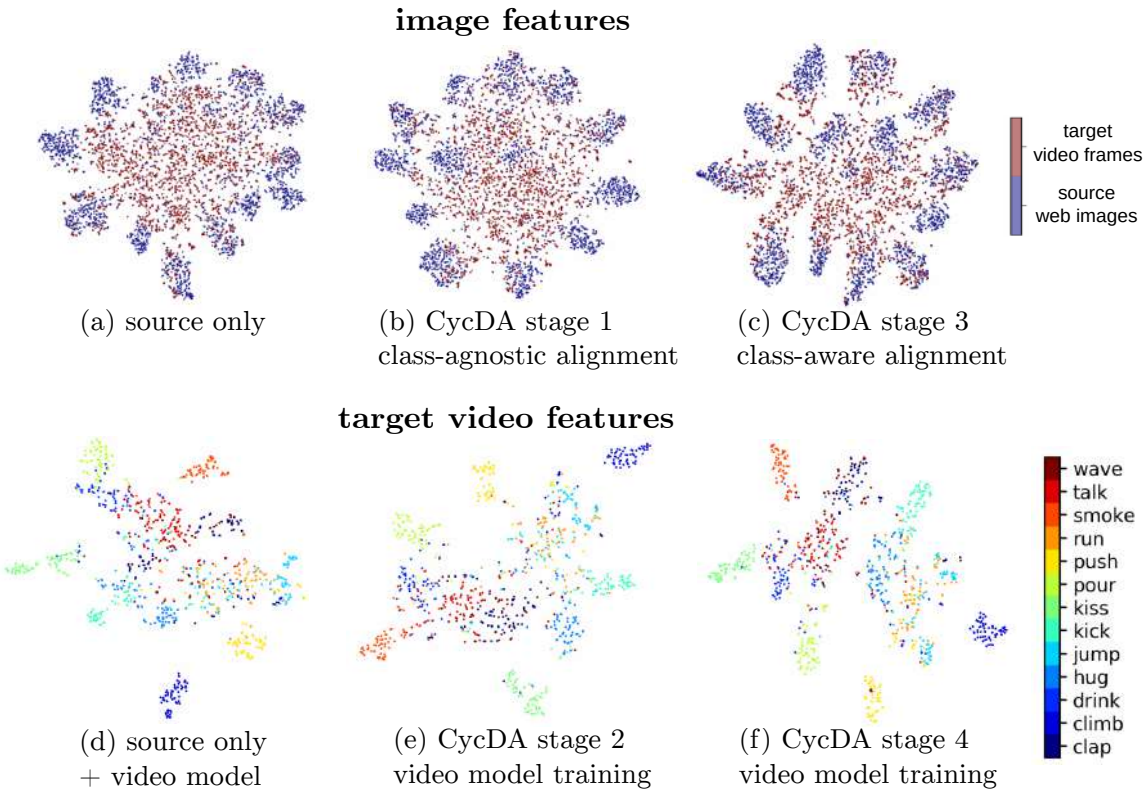
8.4.5 Ablation study

We perform several ablation studies to validate the proposed CycDA pipeline. We conduct these experiments on the setting of EADs → HMDB51 (split 1).

Stage-wise ablation study. We first validate the efficacy of the CycDA pipeline by switching the stages with alternate counterparts. We report the quantitative results of six ablation settings in Table 8.3. Case A (source only), training the image model on web images only and predicting on target test set, demonstrates the lower bound of 39%. Case B (source only + video model), training with pseudo labeled target videos, is a vanilla baseline. It shows that training the video model with supervision from the image model already significantly improves performance by 11.5%. In case C, with class-agnostic domain alignment on the image model, the performance of B is improved by 1.8%. In case F, after completing the cycle with class-aware domain alignment in stage 3 and training the video model in stage 4 with the updated pseudo labels, we achieve the best performance with 60.8%.

Additionally, we conduct ablation experiments using pseudo labels from case C to self-train the video model. Although self-training the video model for 1 (case D) or 2 (case E) stages exhibits performance improvement compared to case C, it is still clearly outperformed by CycDA. This indicates that the elaborate step of knowledge transfer from the video model to the image model and class-aware domain alignment are critical for a good performance.

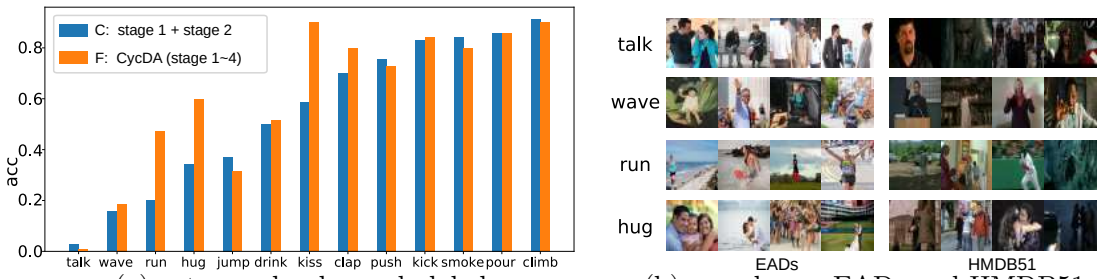
We further illustrate the t-SNE [VdMH08] feature visualizations of ablation cases in Fig. 8.3. By observing image features of the source only case (Fig. 8.3(a)), we see that web images (blue) gather in category clusters after supervised training, while target video frames (red) are far less discriminative and highly misaligned with source due to large domain shift. The class-agnostic domain alignment (Fig. 8.3(b)) in stage 1 results in a slightly better global alignment of source and target features. With category knowledge from the video model, the class-aware domain alignment (Fig. 8.3(c)) demonstrates distinctly better category-level association between source and target. By passing pseudo labels from the image model of Fig. 8.3(a)(b)(c) to supervise the



(a) source only (b) CycDA stage 1 class-agnostic alignment (c) CycDA stage 3 class-aware alignment

(d) source only + video model (e) CycDA stage 2 video model training (f) CycDA stage 4 video model training

Figure 8.3: t-SNE visualizations of image features (for both source and target) and target video features (colored w.r.t. ground truth). We plot the results of source only, source only and video model training, and results of each stage in CycDA.



(a) category-level pseudo labels (b) samples on EADs and HMDB51

Figure 8.4: (a) Category-level pseudo label analysis on target. We compare the accuracy of category-level pseudo labels on target videos for the ablation case C and F from Table 8.3. Here we plot the accuracy of pseudo labels of the 13 common classes on EADs → HMDB51. (b) Samples of four categories on EADs and HMDB51.

spatio-temporal training, we get the corresponding video models whose features are plotted in Fig. 8.3(d)(e)(f). The vanilla baseline of source only + video model (Fig. 8.3(d)) leads to highly undiscriminative features on the difficult classes, which is slightly improved by class-agnostic alignment (Fig. 8.3(e)). Class-aware domain alignment (Fig. 8.3(f)) results in more pronounced category clusters with larger inter-class distances.

Category-level pseudo label analysis on target. In Fig. 8.4(a), we plot the accuracy of category-wise pseudo labels on target videos for the ablation cases C and F from Table 8.3. On case C (blue) which consists of stage 1 (class-agnostic spatial domain alignment) and stage

Experiment		stage 3	stage 4	Acc
stage 1∼2		-	-	52.3
stage 1∼4	A:	$\mathcal{L}_{CE}(V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	55.6
	B:	$\mathcal{L}_{CE}(I_S, V_T^F)$		56.4
	C:	$\mathcal{L}_{CE}(I_S, V_T^F)$, $\mathcal{L}_{ADD}(I_S, V_T^F)$		58.0
	D:	$\mathcal{L}_{CE}(I_S)$, $\mathcal{L}_{CONTR}(I_S, V_T^F)$		**60.8**

Table 8.4: Comparison of DA strategies in stage 3 on EADs → HMDB51 split 1. A: supervised classification on pseudo labeled target frames. B: supervised classification on source and pseudo labeled target. C: B + adversarial domain discrimination. D: supervised classification on source + cross-domain contrastive learning.

2 (spatio-temporal learning), image-to-video action recognition has varying performance on different action classes. More specifically, it yields better results on appearance-based actions with distinct background (e.g. *climb*), or on actions with discriminative pose or gesture (e.g. *smoke*, *pour*). On the contrary, there is inferior performance on fine-grained actions with subtle movements (e.g. *talk*) or actions that are semantically highly generalized with large inter-class variation (e.g. *run*, *wave*), c.f. in Fig. 8.4(b).

By comparing the pseudo label accuracy of cases C and F, we see that the complete CycDA with stage 3 and stage 4 contributes to a significant performance boost on the difficult classes (e.g. *run, hug, kiss*) while keeping the performance on the easy classes. This can be attributed to the class-aware domain alignment that improves cross-domain association on the difficult classes while keeping the alignment on easy ones. Note that better class-aware domain alignment also leads to more support samples for difficult classes, which could result in slight performance drop on the easier classes.

Domain adaptation strategies in stage 3. In stage 3, we transfer knowledge from the video model to the image model by passing pseudo labels from the video model to provide category information in domain alignment. We compare different DA strategies that use pseudo labels from the video model in Table 8.4. Intuitively, performing the supervised task on both source and target (case B) outperforms the case of pseudo labeled target only (case A). Adding the adversarial domain discrimination (case C) leads to a further boost. The cross-domain contrastive learning (case D) has the best performance among the 4 cases. In comparison to the case of only stage 1 and 2, a complete cycle with 4 stages leads to performance improvement of at least 3.3% for all cases. This indicates the benefits of transferring the knowledge from the video model onto the image model. The proposed CycDA generalizes well on different strategies using pseudo labels from the previous stage.

Temporal aggregation of frame-level pseudo labels. Before video model training, we temporally aggregate frame-level pseudo labels. Here we compare three strategies of temporal aggregation in Table 8.5. Frame-level thresholding followed by temporal averaging (case C) outperforms the other two cases which perform temporal averaging before video-level thresholding. Frame-level thresholding filters out frames of low confidence scores and effectively increases the accuracy of video pseudo labels.

Number of iterations. Finally, we illustrate the performance after several iterations on EADs→HMDB51 in Fig. 8.5. It shows that within the first five iterations, performing CycDA iteratively results in a slight increase of performance. Further, within all the 9 iterations, CycDA delivers relatively stable performance, with the result fluctuating between 60.0% and 62.2% without dropping below the ablated alternatives shown in Table 8.3.

	Aggregation			Acc
A:	avg.	+	class-balanced thresh.	50.3
B:	avg.	+	thresh.	55.9
C:	thresh.	+	avg.	**60.8**

Table 8.5: Temporal aggregation of frame-level pseudo labels into video-level pseudo labels on EADs → HMDB51 split 1.

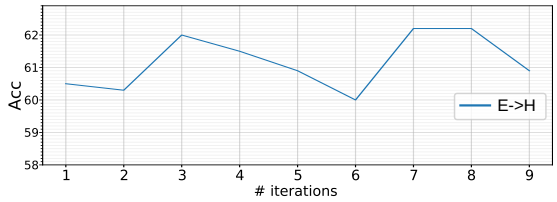


Figure 8.5: Performance of multiple iterations of CycDA (stage 3 and 4) on EADs → HMDB51. Average on 3 splits.

8.5 Conclusion

In this chapter, we address the image-to-video adaptation problem. In our CycDA framework, we propose to alternately perform spatial domain alignment to address the domain shift between images and target frames, and spatio-temporal learning to bridge the modality gap. Our evaluations across several benchmarks and datasets demonstrate that CycDA exploits the large informativity in the source images for enhanced performance on the target video classifier. In the next chapter, Chapter 9, we further extend the possibilities of transferring knowledge between modalities for low-cost video annotations. We will show how to leverage large language models for large-scale automated generation of video descriptions.

9 HowTo-Caption: Prompting LLMs to Transform Video Annotations at Scale

Contents

In this chapter, we explore how to further decrease annotation costs for large-scale video datasets and investigate the capabilities of large language models applied to instructional videos and their subtitles. Instructional videos are a common source for learning text-video or even multimodal representations by leveraging subtitles extracted with automatic speech recognition systems (ASR) from the audio signal in the videos. However, in contrast to human-annotated captions, both speech and subtitles naturally differ from the visual content of the videos and thus provide only noisy supervision. As a result, large-scale annotation-free web video training data remains sub-optimal for training text-video models. In this work, we propose to leverage the capabilities of large language models (LLMs) to obtain high-quality video descriptions aligned with videos at scale. Specifically, we prompt an LLM to create plausible video captions based on ASR subtitles of the instructional video. To this end, we introduce a prompting method that is able to take into account a longer text of subtitles, allowing us to capture the contextual information beyond one single sentence. We further prompt the LLM to generate timestamps for each produced caption based on the timestamps of the subtitles and finally align the generated captions to the video temporally. In this way, we obtain human-style video captions at scale without human supervision. We apply our method to the subtitles of the HowTo100M dataset, creating a new large-scale dataset, HowToCaption. Our evaluation shows that the resulting captions not only significantly improve the performance over many different benchmark datasets for zero-shot text-video retrieval but also lead to a disentangling of textual narration from the audio, boosting the performance in text-video-audio tasks.

This chapter is based on [SKH+24]. As the co-first author Anna Kukleva led the project jointly with Nina Shvetsova, sharing responsibilities in the first months of the project with equal contribution. Anna had to change focus due to the started internship and then she contributed to writing of the paper.

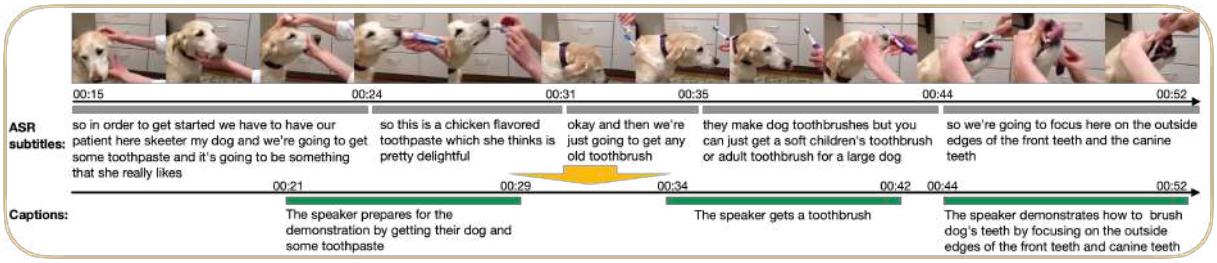


Figure 9.1: **ASR subtitles deviate from human-written captions.** The subtitles contain a lot of filler phrases, e.g., "we're going to", and extra information, e.g., "they make dog toothbrushes". We propose to generate human-style video captions based on the ASR subtitles and their timestamps that we further temporally realign with the video.

9.1 Introduction

Textual descriptions of visual information allow for navigating large amounts of visual data. Recently, image-text cross-modal learning has achieved remarkable performance in many downstream tasks by pre-training on large-scale web datasets consisting of text-image pairs [RKH+21, SVB+21, JYX+21]. To collect video data on a similar scale, media platforms such as YouTube can be used as a great source of freely available videos [AEHKL+16, ZXC18, SLS+20, MZA+19, ZLH+21, XHZ+22]. Most of these videos include some narrations, e.g., in instructional videos [MZA+19], people explain and show how to accomplish one or another task. To transform spoken language from the videos into subtitles, current automatic speech recognition (ASR) systems [RKX+23] can be used, providing aligned text-video annotated pairs for free. This automatic supervisory signal can easily scale to large video datasets. However, such video web data poses additional challenges [HXZ22, MZA+19]: (1) spoken and visual information in the video can deviate from each other, e.g., when speakers provide information beyond what is visible or when spoken instructions do not temporally align with the actions shown, (2) speech contains filler words and phrases, such as "I'm going to", and can be incomplete and sometimes contains grammatical errors, and (3) ASR transcripts usually do not have punctuation and may contain errors. Therefore, ASR subtitles provide only weak, noisy supervision for videos.

To address this problem, we propose a new framework, HowToCaption, that leverages large language models (LLMs) [CLL+23] to generate human-style captions on a large scale for web-video instructional datasets based on corresponding ASR subtitles (Figure 9.1). By carefully designing prompts, we show that the LLM can effectively map long, noisy subtitles into concise and descriptive human-style video captions. Moreover, we obtain an initial temporal alignment of the generated captions to the video based on ASR timestamps by tasking the LLM to predict timestamps for each caption. For additional quality improvement, we apply alignment and filtering within short temporal windows with respect to the generated timestamp. This approach can generate aligned text-video pairs on a large scale without any human intervention.

Beyond providing better annotation, the new captions provide the advantage that they are no longer a direct output of the speech signal, thus effectively decoupling audio and text. Current methods usually avoid using audio [MAS+20, HXZ22], as the ASR subtitle is directly derived from speech, thus leading to the problem that any text-to-audio+video retrieval would mainly retrieve the closest speech signal while disregarding the video. Being able to generate

captions that deviate from the speech thus allows to extend retrieval to audio+video without the need for fine-tuned regularization.

To verify the effectiveness of the proposed HowToCaption method, we generate new captions for the large-scale HowTo100M dataset [MZA+19], obtaining a new HowToCaption dataset. We evaluate the quality of the improved narrations on various challenging zero-shot downstream tasks over four different datasets, namely YouCook2 [ZXC18], MSR-VTT [XMYR16], MSVD [CD11], and LSMDC [RRS]. It shows that the generated captions not only provide a better training signal but also allow for a decoupling of speech and caption annotation, allowing a retrieval based on audio, vision, and subtitles at scale. We release a new HowToCaption dataset with high-quality textual descriptions to show the potential of generated captions for web text-video pairs. We also make code publicly available. We summarize the contributions of the chapter as follows:

- We propose a HowToCaption method to efficiently convert noisy ASR subtitles of instructional videos into accurate video captions, which leverages recent advances in LLMs and generates high-quality video captions at scale without any human supervision.

- We create a new HowToCaption dataset with high-quality human-style textual descriptions with our proposed HowToCaption method.

- Utilizing the HowToCaption dataset for training text-video models allows us to significantly improve the performance over many benchmark datasets for text-to-video retrieval. Moreover, since new textual annotation allows us to disentangle audio and language modalities in instructional videos, where ASR subtitles were highly correlated to audio, we show a boost in text-video+audio retrieval performance.

9.2 Method

9.2.1 Problem Statement

Given a dataset of N untrimmed long-term instructional videos V_n with corresponding noisy ASR (automatic speech recognition) subtitles S_n, our goal is to create "human-like" video captions C_n (with $1 \leq n \leq N$). Note that our task does not assume access to any paired training data $((V_n, S_n), C_n)$. The goal is to create the video captions C_n in a *zero-shot* setting given only videos and subtitles (V_n, S_n). More formally, for each given video V_n, we also have a set of subtitles of spoken text in the video, $S_n = \{s_{n,j}, t^s_{n,j}, t^e_{n,j}\}_{j \leq |S_n|}$ with their start t^s and end timestamps t^e recognized by ASR-systems. For each video V_n, our goal is to generate dense captions $C_n = \{c_{n,i}, \tau^s_{n,i}, \tau^e_{n,i}\}_{i \leq |C_n|}$ and their timestamps, where each caption $c_{n,i}$ describes a segment of the video, that starts at $\tau^s_{n,i}$ and ends at $\tau^e_{n,i}$.

The generated captions aim to serve for vision-language or vision-language-{other modalities (e.g., audio)} tasks, providing language supervision in the form of "human-written-like" captions rather than scrambled noisy ASR subtitles. That enables the potential of collecting large-scale datasets with long-term videos and their dense textual descriptions for free, without human supervision.

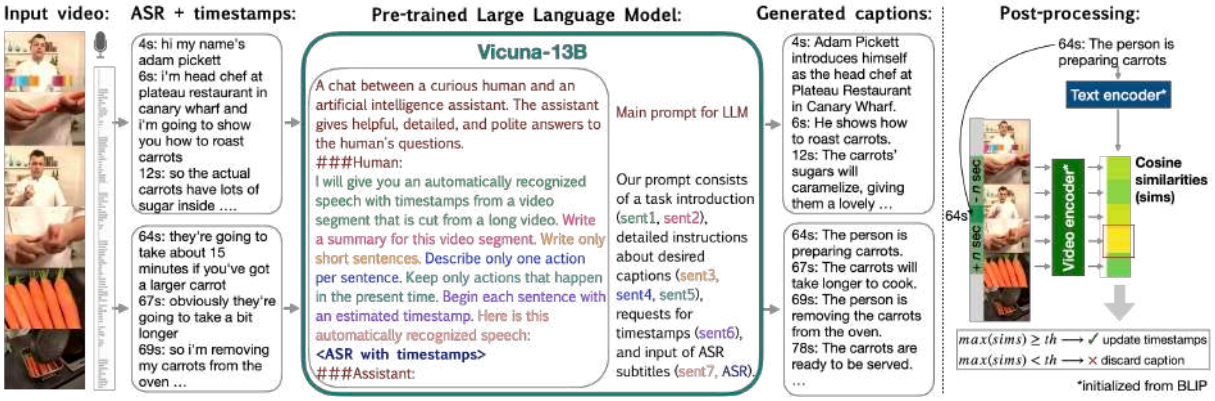


Figure 9.2: **Schematic visualization the proposed HowToCaption method**. Obtained from the automatic speech recognition system (ASR), subtitles are divided into blocks that contain longer contextual information. A large pre-trained language model is then used to generate plausible video captions based on ASR subtitles, along with timestamps for each caption. These generated captions and timestamps are further post-processed to enhance their alignment to the video and filter out captions with low similarity to the corresponding video by leveraging a pre-trained text-video model.

9.2.2 Video-Language Retrieval Model

Before we will describe our method for generating the HowToCaption dataset, we will briefly recap the video-language retrieval models (V-L model), as it is one of the main use cases for this dataset. Moreover, we also use a V-L model to improve the temporal alignment in the dataset.

We base our video-language retrieval model (V-L model) on the pre-trained BLIP image-language dual encoder model [LLXH22]. We maintain the architecture of the text encoder $f(c) \in \mathbb{R}^d$ but, following CLIP4CLIP [LJZ+22], adapt the image encoder $g(I)$ to a video encoder by averaging image embeddings obtained from uniformly sampled frames of the video: $g(V_n) = \sum_{I \in V_n} g(I) \in \mathbb{R}^d$. Dual encoder models typically learn a cross-modal embedding space [RKH+21, LJZ+22] via training with the symmetric InfoNCE loss [OLV18]. The training is based on a similarity metric (often cosine distance) between embeddings $\rho_{n,i,m} = \text{sim}\left(f(c_{n,i}), g(V_m)\right)$ scaled by a temperature parameter ν, resulting in the following loss function:

$$L = -\frac{1}{2|B|} \sum_{(n,i)\in B} \left(\log \frac{\exp(\rho_{n,i,n}/\nu)}{\sum_{(m,j)\in B} \exp(\rho_{n,i,m}/\nu)} + \right. \\ \left. + \log \frac{\exp(\rho_{n,i,n}/\nu)}{\sum_{(m,j)\in B} \exp(\rho_{m,j,n}/\nu)} \right) \tag{9.1}$$

where B is a batch of training sample indices (n, i).

9.2.3 HowToCaption Method

To generate captions for the instructional videos, we propose to leverage recent large language models that demonstrate great zero-shot performance in many different tasks formulated with

ASR: move them around to help direct the path
Caption: Matt Swanson gives a tip to use buckets to direct the path of the ball

ASR: she might need it for later so this is stage one of hiding the bone ...
Caption: Dog wants to hang out near dirt or other dogs with bones to acquire more bones

ASR: so it's not going to really show
Caption: Making a bow with two colors

ASR: but this yeah and it just stays or it won't get off it's busy here
Caption: Make sure the bottle stays together

Figure 9.3: **Examples video-captions pairs from our HowToCaption dataset.** ASR subtitles with only noisy supervision for the video are converted from spoken to written-language-style captions.

natural language. Namely, we prompt the LLM to read the ASR subtitles of the video and create a plausible video description based on this. Since one subtitle only covers a small part of the video and lacks a global context, we propose to aggregate multiple subtitles together with their timestamp information. Then, we task the LLM to create detailed descriptions based on the entire input and estimate timestamps for each generated sentence.

The overview of our approach is shown in Figure 9.2. For each video, we first slice a given sequence of subtitles into blocks that contain long context information about the video. Then, the ASR subtitles of each block are summarised into a video caption using the LLM that we prompt with our task description. The LLM also predicts timestamps for each sentence in the video caption, which we further refine in our post-processing step based on similarities of a caption sentence to video clips in the neighboring area of predicted timestamps. FIX

9.2.3.1 *LLM Prompting*

For our language prompt (shown in Figure 13.2), we leverage the same "main" prompt for the LLM, as in the Vicuna-13B model [CLL+23]: "A chat between a curious human and..." that defines the requirement for LLM to give a helpful answer to our questions. Then, we describe our request, what data we need to process, and how it should be processed: "I will give you an automatically recognized speech...". We found structuring the prompt in the way that the task description given at the beginning of the prompt and the long ASR input S_n at the end is beneficial. Then, we give detailed instructions about how to process ASR subtitles. We found that instructions such as "Write only short sentences" or "Describe only one action per sentence" are beneficial, as they encourage the creation of concise captions that better match the video content. The instruction "Keep only actions that happen in the present time" is intended to filter out unrelated chats, advice, or comments from the captions; we observed that it also resulted in performance enhancements. Lastly, we request the model to predict a timestamp for each generated caption and, finally, input timestamps + ASR subtitles that need to be processed. The LLM response follows the start timestamp + caption format given in the prompt and, therefore, can be automatically parsed with a simple script into a set of captions and timestamps $C_n = \{c_{n,i}, \tau^s_{n,i}, \tau^e_{n,i}\}_{i \leq |C_n|}$, where we assign $\tau^e_{n,i} = \tau^s_{n,i} + \Delta_{sec}$, where Δ_{sec} is a constant video clip length parameter (number of seconds). Please see Section 9.3.3 and the supplement for a detailed evaluation of these choices.

9.2.3.2 *Post-processing: Alignment & Filtering*

ASR subtitles suffer from bad temporal alignment [HXZ22, MZA+19]. Although the LLM prompted to produce video captions can filter some noise in the ASR subtitles, some generated captions are still misaligned with the video. Therefore, inspired by the TAN method [HXZ22] that automatically predicts the alignability of subtitles and matching timestamps, we further improve our obtained captions with an alignment & filtering post-processing step (Figure 9.2).

To this end, we utilize the video-language encoder model (f, g). Given a generated caption $c_{n,i}$ and its start and end timestamps $(\tau^s_{n,i}, \tau^e_{n,i})$ that corresponds to a part of the video clip $V_n^{[\tau^s_{n,i}, \tau^e_{n,i}]}$, we use the V-L model to compute alignment similarity scores $\rho_{n,i}(\delta) = \text{sim}\left(f(c_{n,i}), g(V_n^{[\tau^s_{n,i}+\delta, \tau^e_{n,i}+\delta]})\right)$ between the caption and video clips with time offsets $\delta \in \mathbb{Z}, |\delta| \leq T$ around predicted timestamps. Then we *align* the caption to the video clip by finding the best offset around the timestamp $\delta^*_{n,i} = \underset{\delta \in \{-T,\dots,T\}}{\arg\max}\ \rho_{n,i}(\delta)$ and *filter* out pairs if $\rho_{n,i}(\delta^*_{n,i}) < \kappa$, where κ is a similarity score threshold.

To further improve the alignment of captions, we perform multiple rounds of alignment & filtering. In practice, we found that the improvement after two rounds is marginal. For subsequent rounds, we finetune (c.f. Section 9.2.2) the V-L model on the aligned & filtered video-captions pairs $\{(v_i, c_i)\}$, resulting in new alignment scores $\rho'_{n,i}(\delta)$. Since fine-tuning V-L models often leads to forgetting, we employ two modifications in the fine-tuning and second alignment processes. First, during fine-tuning, we add regularization $L_{\text{align}} = \alpha \frac{1}{2|B|} \sum_{(n,i)\in B} (\text{sim}(f(c_{n,i}), f^*(c_{n,i})) + \text{sim}(g(V_n), g^*(V_n))$ where f^* and g^* denote frozen text and video encoders, α is a regularization weight, and $(n, i) \in B$ represents the samples batch B. This regularization prevents the model from forgetting [HPL+19]. Then, during alignment & filtering, we use the average of the similarities of the finetuned and original model. We show an impact of these modifications in the supplement.

9.2.4 HowToCaption Dataset

We apply the proposed HowToCaption approach to 1.2M long-term instructional videos and ASR subtitles of the HowTo100M dataset and obtain the HowToCaption dataset. By prompting the Vicuna-13B model, we obtain $\sim$ 70M initial captions. After alignment & filtering (details in Section 9.3.2) we obtain 25M high-quality video-caption pairs. We show examples from our HowToCaption dataset in Figure 9.3. We note that generated captions follow different text styles, e.g., the first and the second examples contain a long description of an object and its actions, the third describes the process, and the last one is instruction. The average length of the generated captions is 9 words. We provide additional examples, statistics, and analysis of the HowToCaption dataset in the supplement.

9.3 Experiments

To evaluate the proposed HowToCaption dataset for large-scale pre-training of vision-language models, we train a T-V model as described in Section 9.2.2 on the HowToCaption dataset and assess its zero-shot video-text retrieval performance on four widely recognized and diverse video-text benchmarks: YouCook2 [ZXC18], MSR-VTT [XMYR16], MSVD [CD11], and LSMDC [RRS]. While the YouCook2 dataset consists of instructional cooking videos and might be considered as an in-domain benchmark for the HowToCaption dataset, the other datasets

Prompt	YouCook2 R10↑	YouCook2 MR↓	MSR-VTT R10↑	MSR-VTT MR↓	MSVD R10↑	MSVD MR↓	LSMDC R10↑	LSMDC MR↓	**Average** R10↑	**Average** MR↓
x1: Here is an automatically recognized speech from a video: <ASR with timestamps>. x2: Write a synopsis for this video. x3: Begin each sentence with an estimated timestamp.	37.5	22.5	71.0	3	80.5	2	37.3	30	56.6	14.4
<x1> <x2> x4: Write only short sentences. <x3>	39.3	20.5	71.4	3	81.0	2	36.5	32.5	57.1	14.5
<x1> <x2> <x4> x5: Describe only one action per sentence. <x3>	39.8	20	71.0	3	80.9	2	37.2	30.5	57.2	13.9
<x1> <x2> <x4> <x5> x6: Keep only actions that happen in the present time. <x3>	39.5	19.5	71.6	3	81.2	2	37.9	29	57.6	13.4
<x1> x2′: Write *a summary* for this video. <x4> <x5> <x6> <x3>	40.4	19	71.4	3	81.4	2	37.1	30	57.6	13.5
x1′: Here is an automatically recognized speech from *a video segment that is cut from a long video*: <ASR with timestamps> x2″: Write a summary for *this video segment*. <x4> <x5> <x6> <x3>	40.0	20	72.0	3	81.1	2	37.8	29	57.7	13.5
I will give you an automatically recognized speech with timestamps from a video segment that is cut from a long video. <x2″> <x4> <x5> <x6> <x3> *Hereis this automatically recognized speech: <ASR with timestamps>*	40.6	19	72.0	3	81.6	2	37.7	30	58.0	13.5

Table 9.1: **Ablation of LLM prompts.** We step by step construct a prompt for an LLM that concisely and in detail describes the caption generation task. To emphasize our incremental adjustments, we label the sentences as x*n* (where *n* is an index). Each prompt consists of sentences that were already used in previous prompt versions (e.g., <x1>, <x2>) and new sentences introduced in the current prompt (e.g., x4: Write only ...). With each prompt, we obtain 2M video-text pairs from 100k HowTo100M videos that we later use for T-V model training (lower-resource setup). Downstream zero-shot text-video retrieval performance is reported.

Method	YouCook2 R1↑	YouCook2 R5↑	YouCook2 R10↑	YouCook2 MR↓	MSR-VTT R1↑	MSR-VTT R5↑	MSR-VTT R10↑	MSR-VTT MR↓	MSVD R1↑	MSVD R5↑	MSVD R10↑	MSVD MR↓	LSMDC R1↑	LSMDC R5↑	LSMDC R10↑	LSMDC MR↓	Average R1↑	Average R5↑	Average R10↑	Average MR↓
No context: single ASR subtitle	11.1	27.9	38.4	21	37.7	**62.4**	**72.6**	**3**	43.3	71.7	80.2	**2**	16.5	30.4	**38.4**	**30**	27.1	48.1	57.4	14
almond Long context: multiple ASR+timestamps	**12.1**	**30.0**	**40.6**	**19**	**37.9**	61.6	72	**3**	**43.9**	**72.7**	**81.6**	**2**	**16.8**	**31.4**	37.7	**30**	**27.7**	**48.9**	**58.0**	**13.5**

Table 9.2: **Effect of a longer context.** For the "no context" option, we predict captions from individual ASR subtitles. With our "long context" option, we input multiple ASR subtitles with timestamps and the model generated captions based on longer context. This ablation is done in lower-resource setup.

encompass a broader range of topics and video types, including non-instructional YouTube videos and movies. To evaluate the properties of HowToCaption dataset in comparison with other large-scale pre-training datasets, we also train our T-V model on HowTo100M [MZA+19], on HowTo100M with step labels [LPB+22], HTM-AA [HXZ22], VideoCC3M [NSS+22], and WebVid2M [BNVZ21] datasets and compare zero-shot text-video retrieval performance.

9.3.1 Datasets and Metrics

Pre-training Datasets. HowTo100M is a dataset of 1.2M instructional videos with ASR subtitles collected by querying YouTube with 23k different "how to" tasks from WikiHow articles. We consider three versions of annotations of this dataset: *Sentencified HowTo100M*, with pre-processed ASR subtitles by structuring them into full sentences by [HXZ22]; *HowTo100M with Distant Supervision*, where ASR subtitles were linked to WikiHow [KW18] step descriptions via distant supervision by [LPB+22]; and *HTM-AA* [HXZ22], an auto-aligned (AA) version of HowTo100M, where subtitle timestamps were adjusted to improve alignment to videos, discarding non-alignable subtitles. **WebVid2M** [BNVZ21] is a large open-domain dataset of 2.5M of short videos scrapped from the internet with their alt-text. **VideoCC3M** [NSS+22] is a dataset of 10M video-text pairs collected by transferring captions from image-text CC3M

Caption Post-processing	YouCook2		MSR-VTT		MSVD		LSMDC		**Average**	
	R10↑	MR↓	R10↑	MR↓	R10↑	MR↓	R10↑	MR↓	R10↑	MR↓
Lower bound: original ASR as supervision	39.3	20	61.7	5	77.1	2	31.5	56	52.4	20.8
No post-processing	40.2	18	65.9	4	79.8	2	34.4	40	55.1	16.0
Filtering (using BLIP)	42.5	16	71.2	3	81.7	2	37.4	30	58.2	12.8
Alignment & filtering (using BLIP)	42.4	17	71.7	3	82.2	2	38.5	29.5	58.7	12.9
Alignment & filtering (with ours)	44.1	15	73.3	3	82.1	2	38.6	29	59.5	12.3

Table 9.3: **Effect of alignment & filtering.** With each post-processing variant, we obtain 25M video-text pairs that we later use for T-V model training. Downstream zero-shot text-video retrieval performance is reported.

dataset [CSDS21] to videos with similar visual content.

Downstream Datasets. YouCook2 [ZXC18] is a dataset of instructional cooking videos, where each video clip is annotated with a recipe step. We used 3.5k test set for evaluation. **MSR-VTT** [XMYR16] contains 10k YouTube videos on various topics and human descriptions. Following prior work [BNVZ21, NSS+22], we use the 1k test set for evaluation. **MSVD** [CD11] is a dataset of video snippets with their textual summary. The evaluation set consists of 670 videos with 40 captions corresponding to each video. We follow standard practice [BNVZ21, LJZ+22] and count each caption-video pair towards the metrics. **LSMDC** [RRS] is a collection of movies sliced into video clips with human-written descriptions. The test set consists of 1k video-caption pairs.

Metrics. To evaluate zero-shot text-video retrieval, we used standard Recall@K metrics where $K \in 1, 5, 10$ (R1, R5, R10) and Median Rank (MedR).

9.3.2 Implementation Details

As an LLM, we utilize Vicuna-13B [CLL+23], which is LLAMA [TLI+23] model fine-tuned to follow natural language instructions. We additionally experiment with the MiniGPT-4 model [ZCS+23] to generate captions from subtitles grounded on visual content. To create the HowToCaption dataset, we leverage subtitles with timestamps released by [HXZ22] (Sentencified HowTo100M), where officially released subtitles [MZA+19] for HowTo100M videos were post-processed by structuring them into full sentences. For our T-V model (described in Section 9.2.2), we use ViT-B/16 visual encoder and $\text{BERT}_{\text{base}}$ textual encoder that are initialized with $\text{BLIP}_{\text{CapFilt-L}}$ pre-trained weights. We uniformly sample 4 frames from a video clip during training and 12 frames during evaluation. For HowToCaption method, we use $T = 10$ seconds offset for alignment and adaptive threshold κ to leave 25M most similar pairs after filtering. Following [CRD+21] that found that 8-sec clips are optimal for training on HowTo100M, we set $\Delta_{\text{sec}} = 8$.

9.3.3 Ablation Studies

Prompt Engineering. Since prompting LLM with subtitles from 1.2M videos is resource-intensive, we perform the prompt engineering ablations in a lower-resource setup, where we use a 100k subset of HowTo100M ($\sim$ 10% of all videos) to create dense captions with the LLM and use the threshold κ to obtain the 2M most confident video-caption pairs. Here, we train the T-V model for 150k iterations and then evaluate zero-shot on downstream tasks. In Table 13.8, we begin with a basic prompt for an LLM, gradually refining it to generate captions more suitable for vision-language tasks. It is essential to recognize that the impact of various prompts

Video-Text Training Data	YouCook2				MSR-VTT				MSVD				LSMDC				**Average**			
	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓
- (zero-shot, with BLIP initialization)	6.1	16.2	23.6	69	34.3	59.8	70.6	3	38.5	65.0	74.0	2	14.7	29.5	36.5	31	23.4	42.6	51.2	26.3
HowTo100M with ASRs	12.2	29.1	39.3	20	30.8	52.6	61.7	5	39.2	68.3	77.1	2	12.9	24.7	31.5	56	23.8	43.7	52.4	20.8
HowTo100M with distant supervision	8.3	21.5	30.3	34	28.6	54.0	66.3	5	38.5	68.6	79.4	2	12.1	24.7	32.4	42.5	21.9	42.2	52.1	20.9
HTM-AA	13.4	32.2	43.5	15	29.8	54.1	64.3	4	38.7	68.6	78.7	2	11.9	23.9	30.5	46	23.5	44.7	54.3	16.8
HowToCaption (ours)	13.4	33.1	44.1	15	37.6	62.0	73.3	3	44.5	73.3	82.1	2	17.3	31.7	38.6	29	28.2	50.0	59.5	12.3
VideoCC3M	5.3	15.1	21.7	84	33.9	57.9	67.1	4	39.6	66.7	76.8	2	14.8	29.4	35.8	33	23.4	42.3	50.4	30.8
WebVid2M	7.3	20.7	29.0	46	38.5	61.7	71.9	3	44.5	73.4	82.1	2	17.8	31.2	39.8	25	27.0	46.8	55.7	19.0

Table 9.4: **Zero-shot text-to-video retrieval performance of model trained on different video-text datasets.** For each dataset, we train our T-V model and report downstream zero-shot text-video retrieval performance.

Method	Vision Encoder	Image-Text Data	Video-Text Data	YouCook2				MSR-VTT				MSVD				LSMDC			
				R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓
Nagrani et al.[NSS+22]	ViT-B + fusion. b.	-	VideoCC3M	-	-	-	-	18.9	37.5	47.1	-	-	-	-	-	-	-	-	
Frozen-in-Time [BNVZ21]	ViT-B/16 + temp.	CC+COCO	WebVid-2M	-	-	-	-	24.7	46.9	57.2	7								
CLIP-straight [PQOBTM21]	ViT-B/32	WIT	-	-	-	-	-	31.2	53.7	64.2	4	37.0	64.1	73.8	2	11.3	22.7	29.2	56.5
CLIP4CLIP [LJZ+22]	ViT-B/32	WIT	HTM100M	-	-	-	-	32.0	57.0	66.9	4	38.5	66.9	76.8	2	15.1	28.5	36.4	28
VideoCoCa [YZW+22]	∼ViT-B/18 + temp.	JFT-3B	VideoCC3M	16.5	-	-	-	31.2	-	-	-	-	-	-	-	-	-	-	-
BLIP§ [LLXH22]	ViT-B/16	5 datasets‡	-	6.1	16.2	23.6	69	34.3	59.8	70.6	3	38.5	65.0	74.0	2	14.7	29.5	36.5	30.5
almond **Ours**	ViT-B/16	5 datasets‡	HTM-Captions	13.4	33.1	44.1	15	37.6	62	73.3	3	44.5	73.3	82.1	2	17.3	31.7	38.6	29

Table 9.5: **Comparison in zero-shot text-to-video retrieval with dual-encoder baseline methods.** "+ fusion b." denotes a usage of a fusion bottleneck., "+ temp" denotes of usage of temporal attention. ‡CC [CSDS21]+COCO [LMB+14]+VG [KZG+17]+SBU [OKB11] +LAION [SVB+21]. §For BLIP, the performance of dual encoder architecture is reported.

on performance can vary across datasets, as certain prompts may yield captions better aligned with specific downstream tasks. Notably, incorporating key phrases such as "Write only short sentences" or "Describe only one action per sentence" leads to performance improvements on 3 out of 4 datasets. Additionally, the use of the phrase "Keep only actions that happen in the present time" also resulted in performance enhancements. Furthermore, structuring the task description at the beginning and presenting the data to be processed at the end (the final modification), also boosts performance. We also examine the impact of leveraging a longer context for caption prediction. In Table 9.2, we compare caption generation with "no context", where captions are predicted from individual ASR subtitles. In this option, timestamps of the input ASR are used as timestamps of the prediction caption. With our "long context" option, we input multiple ASR subtitles with their timestamps, and the model predicts both captions and timestamps based on longer context. We found that using a longer context is beneficial, resulting in an average improvement 0.6 p.p. in R10, and particularly advantageous for the YouCook2 and MSVD datasets.

Alignment & Filtering. Further, we assess the impact of the proposed alignment & filtering procedure on the quality of captions of the acquired dataset in Table 9.3. We examine the performance of the T-V model when trained on differently post-processed versions of the dataset. Remarkably, we discover that the obtained video-caption pairs, even without any post-processing, significantly outperform the original ASR-based supervision. Subsequently, by employing the alignment and filtering procedure to leave only 25M pairs based on video-caption similarities derived from BLIP pre-trained weights, we achieve a notable performance enhancement of 3.6 p.p. in R10. Furthermore, alignment & filtering with our proposed fine-tuning without forgetting yields an additional 0.8 p.p. boost in R10 performance.

Method	Vision Enc	YouCook2				MSR-VTT			
		R1↑	R5↑	R10↑	MR↓	R1↑	R5↑	R10↑	MR↓
MIL-NCE‡	S3D	15.1	38.0	51.2	10	9.9	24.0	32.4	29.5
TAN‡	S3D	20.1	45.5	59.5	7.0	-	-	-	-
MMT	Transformer	-	-	-	-	-	14.4	-	66
AVLNet	R152+RX101	19.9	36.1	44.3	16	8.3	19.2	27.4	47
MCN	R152+RX101	18.1	35.5	45.2	-	10.5	25.2	33.8	-
EAO	S3D	24.6	48.3	60.4	6	9.3	22.9	31.2	35
Ours	S3D	25.5	51.1	63.6	5	13.2	30.3	41.5	17

Table 9.6: **Zero-shot text-video+audio retrieval.** MIL-NCE [MAS+20], TAN [HXZ22], MMT [GSAS20], AVLNet [RBH+21], MCN [CRD+21], EAO [SCR+22]. ‡ denote text-video only retrieval models. R152+RX101 denotes ResNet-152+ResNeXt101.

9.3.4 Comparison with State-of-the-art

Comparision With Other Web Datasets. In Table 9.4, we assess the pre-training effectiveness of our proposed HowToCaption dataset compared to other web video-language datasets. Specifically, we evaluate different textual annotations of HowTo100M videos: sentencified ASR subtitles [HXZ22], task steps from distant supervision [LPB+22], and auto-aligned ASR subtitles [HXZ22]. Additionally, we conduct evaluations on WebVid2M [BNVZ21] and VideoCC3M [NSS+22] datasets. Our findings indicate that the model pre-trained on our HowToCaption dataset significantly outperforms models pre-trained on other versions of HowTo100M annotations, with an average improvement of 5.2 p.p. in R10. This improvement is most pronounced for the MSR-VTT, MSVD, and LSMDC datasets, which feature full-sentence captions. Interestingly, for the YouCook2 dataset with captions in the form of step descriptions like "cut tomato", HTM-AA already exhibits a high baseline performance, but our HowToCaption dataset still provides a performance boost. We also observe that the VideoCC3M dataset does not improve the initial BLIP performance on any datasets except for the MSVD. We attribute it to the fact that the VideoCC3M dataset adopts captions from the CC3M dataset [CSDS21] and transfers them to videos, potentially not introducing significantly new knowledge for the BLIP-initialised model since BLIP was pre-trained on multiple datasets including CC3M. On the other hand, WebVid2M demonstrated performance improvements across all datasets, but our HowToCaption dataset notably outperforms WebVid2M on YouCook2 and MSR-VTT, only underperforming on LSMDC.

Comparison with SOTA in Zero-shot Text-Video Retrieval. In Table 9.5, we also conduct a comparison with zero-shot dual-encoder retrieval baselines. It is important to acknowledge that comparing state-of-the-art methods can be challenging due to variations in backbone capacity, training objectives, and other factors. Therefore, we focus on a comparison with dual-encoder models in the zero-shot settings. Nevertheless, it is worth highlighting that our approach consistently outperforms the baseline methods in zero-shot text-video retrieval across all datasets.

Text-Video+Audio Retrieval. It is known that instructional video datasets, e.g., HowTo100M or HD-VILA, suffer from a high correlation of audio modality to a textual description, therefore hindering building a text-video+audio retrieval system where the video is extended with audio. The usage of ASR narrations as supervisory textual description leads retrieval models to primarily perform speech recognition on the audio, hindering true language-audio connections. Therefore, training text-video+audio systems on these datasets usually requires additional

regularization, such as shifting audio timestamps or assigning lower weights to the audio loss [SCR+22]. Our HowToCaption dataset resolves this issue by providing richer textual descriptions, allowing us to train a text-video+audio retrieval system without regularization. To evaluate this, we train a multimodal Everything-At-Once (EAO) [SCR+22] model that learns to fuse any combinations of text, video, and audio modalities on our proposed HowToCaption dataset without any additional tricks and evaluate zero-shot text-video+audio retrieval performance. Table 9.6 shows the proposed model significantly outperforms all baselines and the directly comparable EAO model.

9.3.5 Qualitative Examples

HowToCaption Dataset. We present an extension of Fig. 3 from the main paper with video-text examples of our HowToCaption dataset and corresponding ASR subtitles in Figure 9.4. We see that our HowToCaption method effectively transforms noisy ASR subtitles into proper captions, leveraging the complete ASR context for caption generation. We demonstrate additional video-text examples of our HowToCaption dataset in Figure 9.5 and Figure 9.6. In Figure 9.6, we also showcase instances of failure cases. One such case involves a failure where the LLM was unable to generate a caption and instead copied the input ASR subtitles: "DP Move Safe lets operators get out of the classroom..." However, in this example, the ASR subtitles contain a third-person video description with a subjet+verb+object sentence structure that justifies the coping input description without modification. Other failure cases include video-caption pairs, where the caption corresponds to the video only partially, e.g., "Cover it with lid" action is not visible on the video while "until the seviayan is cooked" is visible.

LLM Caption Generation. In Table 9.7, we showcase captions generated by the Vicuna-13B model, presenting both the input ASR subtitles and their corresponding generated captions for comparison. We observe the LLM is able to transform scrambled ASR subtitles into "human-written-like" descriptions. However, we also note that sometimes LLM fails to produce descriptions. We present some failure cases in Table 9.8, which include 1) direct input repetition: instances where the LLM duplicates ASR input without modification; 2) ineffective reformulation: the LLM attempts to convert ASR content into descriptions using ineffective structures like "A person says..."; 3) failure to follow the requested structure: instances where the LLM output doesn't follow "a timestamp: a sentence" structure for output, e.g., using "Summary: " to write a video description without timestamps.

9.3.6 Limitations

Our method relies on pre-trained large foundational models, including the large language model Vicuna [CLL+23] and the vision-language model BLIP [LLXH22]. Consequently, our HowToCaption method and the proposed HowToCaption dataset may inherit limitations present in these models. Notably, large language models have several shortcomings, such as biases from their training data, which can lead to the generation of potentially misleading content and the propagation of societal biases [**?**]. Additionally, since our text-video model is initialized from the BLIP model that was pre-trained on curated and filtered datasets, it might be less robust to noisy low-quality videos in our alignment & filtering step. In our robustness analysis, we found that the model is capable of filtering noisy input, but 1-2% of noisy data is still passing the filter. Finally, since our dataset is sourced from the HowTo100M [MZA+19] dataset, it follows the same data distribution, focusing solely on "how-to" topics.

Caption (118s-126s): Matt Swanson gives a tip to use buckets to direct the path of the ball
ASR: 7s: hi i'm matt swanson
9s: with matt swanson's we'll go is gonna help me change the direction of your ball play
16s: if you're struggling with a slice or a hook and it's moving too much in that direction
21s: i'm gonna give you a little tip that we use
25s: that's very easy that you can do when you're out to help
84s: change that
91s: you're slicing now
109s: too much
116s: so use these buckets when you're out at the range
120s: move them around to help direct the path
123s: make sure the clubface is closing if you're trying to get rid of the slice opening
128s: if you're trying to hit a fade use these tips and you'll get better

Caption (87s-95s): Dog wants to hang out near dirt or other dogs with bones to acquire more bones
ASR: **87s: so this is stage one of hiding the bone**
90s: burying the bone
90s: there's so much more involved
92s: let's watch how she behaves
94s: when she comes back from burying the bone she becomes a bit annoying
97s: she wants to hang out next to dirt or somebody else who has a bone so that she might acquire yet another bone
103s: now what i have found in my experience with her and remember every dog is different
107s: nobody does things the same way but she lets those bones ferment for about two days before she goes back and finally retrieves the bone
115s: and then she sits down and enjoys it
116s: and then something makes it really special
...

Caption (187s-195s): Making a bow with two colors
ASR: 185s: so if you don't get your your ribbon twisted there's no up or down side to it
191s: so it's not going to really show
193s: now once i get my three loops on my lighter color i'm going to make a little loop and this is basically just to hide my wire
202s: that i'm going to use
205s: the wire i use either a teen or a 20 gauge wire and what i'm going to do
210s: this little loop is going to be my hide for the wire
215s: so i just slide that through like that and pull real tight and twist
227s: that will keep your loops good and snug once you get done
237s: just kind of work your your loops around
243s: and now you have a bow that's made with two colors
246s: these are great for easter baskets you can use those for mother's day
251s: we have an assortment of colors so we even have some for the holidays for fall
257s: it's the type that you can use any time of the year and it makes great bows

Caption (68s-76s): Make sure the bottle stays together
ASR: 0s: hi there the amateur scientist here
4s: and today in this video i'm going to show you how to make a very cheap and from recycled materials
12s: a mosquitoes trap
14s: now because we all know that mosquitoes can bite and make you itch and they are very uncomfortable
20s: so today in this video i'm going to show you how to make a mosquito trap from a plastic bottle and on their amulet or zoom
28s: so let's get into the video
30s: first of all we're going to need a plastic bottle and i and some very cheap one
45s: you could also use these for sugar
61s: so first grab the bottle and and put the upper part into the lower part
69s: but this yeah and it just stays or it won't get off
73s: it's busy here
74s: good deeper cheap wine
79s: hello this with it like this or just a little bit more like that
...

Figure 9.4: **Extended example of video-captions pairs from our HowToCaption dataset (an extension of Fig. 3 of the main paper)**. The ASR subtitles within the corresponding video clip are bolded. We note that some details in the generated captions are derived from a long ASR context.

Caption: Video segment starts with a shot of David's face, which is described as funny
ASR: and the bottom is actually has holes in it because it gets so incredibly hot so you cannot submerge it in water so we ask you to just rinse it out real quick look at david's face he is so funny

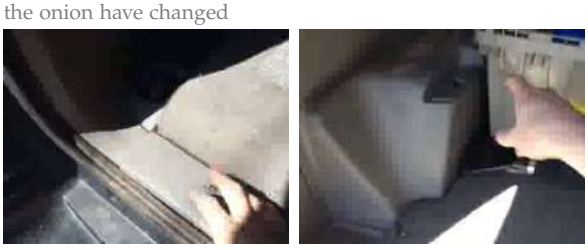
Caption: Brutus is encouraged to swallow his medication
ASR: if i put it in a piece of food he'll chew it up and spit it out he knows oh baby i've never seen him do this though

Caption: Adds two cans of red kidney beans to the chili
ASR: you could also use a vegetable broth all right so we're mixing this well

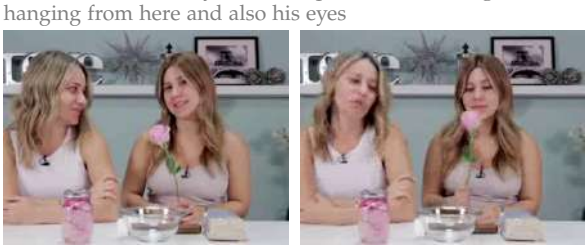
Caption: She explains the charm tool's little piece of metal acts as a spacer to hold the charms open
ASR: has this little piece of patootie metal right here that acts as a spacer to hold the charms open so that gap is visible in the back so it's easier to slip the charms on and off

Caption: Adding chopped onions and green chillies to the pan
ASR:once the oil is hot enough we will add our onions and green chillies we need to cook the onions for some time maybe like 2 to 3 minutes until you start noticing that the colors of the onion have changed

Caption: Paints the top part
ASR: i also notice how the blue continues onto the front of him just like right there so be careful with that next you take the white color and you would paint the webbing that he's hanging from here and also his eyes



Caption: Shows where wire runs along inside of vehicle
ASR: then ran alongside the gasket right here and runs down here and then this we took off and then ran the wiring in through here put this back down



Caption: You are making a rose petal exfoliating face scrub
ASR: you guys one of our favorite diys ever had to do with rose petals so we thought let's make another one

Figure 9.5: **Examples video-captions pairs from our HowToCaption dataset.** Since ASR subtitles' timestamps do not always correspond to the timestamps of video clips from the HowToCaption dataset, we show ASR subtitles that intersect with video clip boundaries.


Caption: Soft bed in the car
ASR: but honestly when people see a lone prius in the parking lot no one thinks hey i wonder if someone's sleeping in there because come on it's a prius what fits in my car i have a soft bed i have a closet blackout curtains a desk kitchen table and chair a pantry a bike a laundry basket travel kit for emergencies


Caption: Asks viewers to choose favorite pet
ASR: i think i like them all for different reasons so that's hard so they can all three be my favorite can't they and look at her little friend


Caption: DP Move Safe lets operators get out of the classroom and out of their truck faster where they learn how to perform every task and do it safely (failure)
ASR: dp move safe lets operators get out of the classroom and out of their truck faster where they learn how to perform every task and do it safely


Caption: Cover it with a lid for 15 minutes until the seviayan is cooked (failure)
ASR: today, we will prepare a sweet recipe.. ..called 'sevaiyan' (vermicelli) so come on, let's see how to make sweet sevaiyan.


Caption: Walk upstairs to show light in the ceiling
ASR: i placed them over here because it's a little bit lighter on this side of the stair case then the other side there's only one light in the ceiling here so i'm gonna walk upstairs and i'm gonna let you see it from the top of the stairs one more time


Caption: The speaker adds white paint to the brush to keep the color bright
ASR: so i lay it on with the flat of the brush which deposits it a little heavier it holds up a little better and notice i keep adding white as ...


Caption: Outdoor Edge has instructional gated processing DVDs available on their website (failure)
ASR: this is one of the big issues with large diameter sausage products remember processing your own wild game animal can be fun easy and very rewarding if you have the tools and the knowledge to do the job you're watching outdoor edges


Caption: The group harvests fresh berries from the farm (failure)
ASR: the kids and i are here at a local blue red patches to manage to harvest some fresh berries at pcc are fresh and frozen organic ...

Figure 9.6: **Examples video-captions pairs from our HowToCaption dataset.** Failure cases are marked as (failure). Since ASR subtitles' timestamps do not always correspond to the timestamps of video clips from the HowToCaption dataset, we show ASR subtitles that intersect with video clip boundaries.

Table 9.7: **Examples of caption generation with the Vicuna-13B.** Input ASR subtitles and output generated captions by the Vicuna-13B (without any post-processing). For brevity, lengthy examples are shortened with the symbol "...".

Input ASR subtitles	Generated captions
0s: hi guys it is bill with septic flow 4s: i am here at a brand new construction that i'm actually working on 8s: here 9s: on the back is the septic field 10s: we've already had bulldozing contractor actually cover up the the septic system 15s: but i'm going to show you why 17s: you want to use septic flow to pre perk the system 22s: septic flow is a product that actually helps get rid of sodium and phosphate buildup and we're actually going to just dump it right here in this two - inch 29s: it goes right out there 29s: we're going to run some water behind it for new construction 33s: the reason you want to do that is because we are actually pre perking the system getting it ready to take those phosphates and sodium buildup 41s: that happens through all different kinds of things 43s: you know whether you're using 44s: i've had people ask me can you use organic detergent 47s: will that prevent sodium buildup 50s: absolutely not 50s: soap by nature of the saponification process that it goes through it's just part of it	0s: Bill is at a new construction site. 4s: He is showing the septic field. 8s: The bulldozing contractor has covered up the septic system. 10s: Bill explains why they use septic flow. 17s: Septic flow helps get rid of sodium and phosphate buildup. 22s: Bill demonstrates how they will dump septic flow in a two-inch area. 29s: They will run water behind it for new construction. 33s: Bill explains that they are pre-perking the system. 41s: The system needs to be prepared to take away phosphates and sodium buildup. 44s: People ask if using organic detergents can prevent sodium buildup. 50s: The answer is no, soap is part of the saponification process and will cause buildup.
104s: the great thing about the smart chicken 107s: with it being air chilled 108s: it's going to pick up a lot of these flavors when we put the the ginger the garlic 111s: and the sage in here 112s: it's just going to combine well with the meat especially while it's cooking on the grill 116s: all those flavors are just going to pop once the sandwich is complete and once it's all done 121s: now that we have it all mixed i'm just going to separate that out into two patties 129s: um you don't have to pat it up too much 131s: it takes a lot to get it on the grill so just kind of separate it out into two balls 136s: even 136s: and that's our chicken 138s: it's ready for the grill ...	104s: The speaker prepares the chicken by air-chilling it and adding ginger, garlic, and sage. 108s: The chicken will pick up the flavors while cooking on the grill. 111s: The speaker mixes the chicken and its seasoning. 116s: The flavors will pop once the sandwich is complete. 121s: The speaker separates the chicken mixture into two patties. 129s: The patties are not patted too much, as they will be grilled soon. 131s: The chicken is ready for the grill. 138s: The grill should be hot, around 400 degrees, for grilling the chicken. 141s: The chicken will flatten itself out and get a nice sear on the bottom when grilled. 149s: The speaker recommends not touching the chicken for about four minutes so the bottom can set up. 157s: The chicken will fall apart if touched too soon.
9s: hello home bruce and welcome to my channel 15s: so i am down to my last bottle of strawberry wine 17s: i've got some in a glass so i thought it would be a fantastic idea to make some more 23s: so it is a bit early in the season to be getting strawberries 27s: it has been rather cold so i am actually using frozen strawberries so i hope you enjoyed the video and i'm gonna have a drink 42s: so for this recipe i am actually going to be using a beauty bucket 46s: in my case it is an up cycled ten liter mayonnaise dough which you would have seen in the sakae video 52s: now if you don't have a brewing bucket of some description which has got a wide mouth so it's less likely to puke ... black	9s: Hello and welcome to the channel. 15s: The host is down to their last bottle of strawberry wine. 17s: They have some wine in a glass and decide to make more. 23s: The host mentions that it's early in the season for strawberries. 27s: The host is using frozen strawberries in the recipe. 42s: The host is using a beauty bucket for the recipe. 46s: The bucket is an upcycled mayo container. 52s: The host mentions that a wide-mouthed bucket is important to avoid puking. 59s: The host mentions that puking is a common issue in fruit brewing. 62s: The host mentions a strawberry brewing bucket that they have. ...

Table 9.8: **Illustrative failures in caption generation with the Vicuna-13B.** Input ASR subtitles and output generated captions where the Vicuna-13B failed to generate video descriptions based on the subtitles. Failures include 1) direct input repetition; 2) ineffective reformulation of the ASR subtitles into descriptions using structures like "A person says .. "; 3) failure to follow the requested structure "a timestamp: a sentence". For brevity, lengthy examples are shortened with the symbol "...".

Input ASR subtitles	**LLM output: generated captions**
2s: i got my barbecue shoes on 9s: go down to the house the sausage 14s: i got my barbecue shoes 16s: welcome to the barbecue pit boys 18s: hi 18s: and today we're gonna do some country meatballs and gravy on the grill 22s: all right 23s: it's real easy to get started here 25s: we got some large granulated salt we got some black peppercorn there we have some granulated onion here 31s: we got some red pepper flakes and here we got some granulated garlic 36s: now instead of using those spices you could just say use your favorite blend of seasonings here ...	2s: i got my barbecue shoes on 9s: go down to the house the sausage 14s: i got my barbecue shoes 16s: welcome to the barbecue pit boys 18s: hi 18s: and today we're gonna do some country meatballs and gravy on the grill 22s: all right 23s: it's real easy to get started here 25s: we got some large granulated salt we got some black peppercorn there we have some granulated onion here 31s: we got some red pepper flakes and here we got some granulated garlic 36s: now instead of using those spices you could just say use your favorite blend of seasonings here ...
426s: you need three out of five 432s: so first jack jack 435s: ready 436s: oh good man 437s: oh no 438s: come on 448s: everything 449s: here we go 449s: okay actually okay 450s: and then for - 451s: since you're a little older you guys stand behind the line here 456s: so let's have you stand behind jack jack's face 462s: so stand up 463s: okay here we go back there ...	426s: A person announces that three out of five is needed. 432s: Jack Jack is mentioned. 435s: Someone says "ready." 436s: A person says "oh good man." 437s: A person says "oh no." 438s: Someone says "come on." 448s: Everything is mentioned. 449s: The current video segment starts. 450s: The game is continued. 451s: A person tells the players to stand behind the line. 456s: Jack Jack stands behind his face. 462s: The game continues. 463s: Someone tells Jack Jack to stand up. 465s: The game continues behind Jack Jack's face. 471s: Three is mentioned. ...
3s: so we got to the campground 7s: we want to use the oven 10s: we need to light the manual pilot 11s: we only have to light it once for the whole weekend if we're not moving around 15s: so all we need to do is make sure the propane works the oven works or the stovetop works 22s: now we just need to use the oven 24s: so now we're down at the ovens level 26s: we're going to light the pilot light so all we need to do is have a stick lighter 35s: turn the knob to pilot push and hold light 41s: the pilot the pilots burning one two three i let go didn't turn the knob the pilots staying lit 49s: the pilot is good for all weekend 51s: when you go to cook your pizza set it to the temperature you want pizzas done turn it back to pilot and the pilot will stay lit all weekend 63s: when you're ready to leave then we turn it to off that turns off the pilot 69s: we don't want that pilot on while we're driving all the time and constantly using propane 75s: that's why they give us the option 80s: pilot or off is off black	3s: Campground 7s: Oven 10s: Manual pilot light 11s: Light pilot once for the weekend 15s: Check if propane and oven work 22s: Use the oven 24s: At oven's level 26s: Light pilot 35s: Turn knob to pilot, push and hold 41s: Pilot lit 49s: Good for all weekend 51s: Set temperature, turn to pilot, and leave 63s: Turn off when ready to leave 69s: Off while driving 75s: Option to choose pilot or off 80s: Off is off. Summary: A group checks if the campground's oven works and lights the manual pilot. They only need to light it once for the weekend as long as they're not moving around. They use the oven and make sure the propane works before they leave. The option to choose pilot or off is given. The pilot should be off while driving.

9.4 Conclusion

Freely available web videos serve as a rich source of multimodal text-video data. Nevertheless, training on such data presents challenges, primarily due to weak supervision offered by video subtitles for text-visual learning. In this chapter, we address this problem by leveraging the capabilities of large-language models (LLMs). We propose a novel approach, HowToCaption, that involves prompting an LLM to create detailed video captions based on ASR subtitles. Simultaneously, we temporally align the generated captions to the videos by predicting timestamps with the LLM, further followed by the alignment & filtering step, which additionally ensures synchronization with the video content. To validate the efficacy of the proposed HowToCaption method, we curate a new large-scale HowToCaption dataset, featuring high-quality human-style textual video descriptions derived from the videos and ASR subtitles of the HowTo100M dataset. Our HowToCaption dataset helps to improve performance across multiple text-video retrieval benchmarks and also separates textual subtitles from the audio modality, enhancing text-to-video-audio tasks. This work demonstrates the ability of LLMs for creating annotation-free, large-scale text-video datasets.

In Part II, various strategies are discussed to minimize annotation costs for visual tasks. This includes semi-supervised learning in Chapter 6, as well as multimodal learning explored in Chapter 7, Chapter 8, and Chapter 9. We investigate image-to-video alignment and language-to-video alignment, leveraging readily available supervision, and demonstrate performance comparable to fully supervised models.

III

Learning with Limited Data

While the previous parts focused learning representations with abundant amounts of data, either without any labels or with some restrictions on label availability, this part considers scenarios where only a limited number of training samples are available.

In Chapter 10, we investigate generalized and incremental few-shot learning settings, focusing on addressing challenges such as catastrophic forgetting and calibration. Our base-normalized cross-entropy approach amplifies the softmax output of novel classes, thereby enhancing learning while preserving previous knowledge based on limited samples. Finally, we employ data replay to tackle the calibration challenge that is widely used in a standard incremental learning methods.

In Chapter 11, we address bias resulting from the disproportionate availability of data for pretraining and subsequent tasks. With an abundance of data samples from base classes, models often exhibit overconfidence in these classes, leading to confusion with novel classes. To mitigate overconfident predictions, we employ space reservation techniques, enhance pretraining method, and utilize margin enhancement techniques.

10 Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration Without Forgetting

Contents

In this chapter, we explore scenarios where only a limited number of samples are available for training, and obtaining additional data in the form of in-domain unlabeled data is challenging. These scenarios, such as generalized and incremental few-shot learning, face three major challenges: learning novel classes from only few samples per class, preventing catastrophic forgetting of base classes, and classifier calibration across novel and base classes. In this work we propose a three-stage framework that allows to explicitly and effectively address these challenges. While the first phase learns base classes with many samples, the second phase learns a calibrated classifier for novel classes from few samples while also preventing catastrophic forgetting. In the final phase, calibration is achieved across all classes. We evaluate the proposed framework on four challenging benchmark datasets for image and video few-shot classification and obtain state-of-the-art results for both generalized and incremental few shot learning.

This chapter is based on [KKS21b]. As the first author, Anna Kukleva conducted all experiments and was the main writer.

10.1 Introduction

In this chapter we are interested in two practically important learning scenarios, namely generalized few-shot learning (GFSL) [GK18, RLFZ19, QBL18, SSS+20] and incremental few-shot learning (IFSL) [THC+20, CL21]. In both scenarios it is possible to learn a performant classifier for a set of base classes for which many training samples exist. However, for the novel classes, only few training samples are available such that a *novel class learning* is challenging. Additionally, in generalized few-shot learning and in incremental learning it is important to prevent *catastrophic forgetting* of the base classes during novel class learning. Last, but not least, *classifier calibration* across classes has to be addressed, due to the imbalance in the amount of training samples. While previous work focuses on addressing a subset of these challenges [GK18, RLFZ19, KDGM+19, THC+20, CL21], in this chapter we aim to address all three.

To this end, we propose a three phase framework to explicitly address these challenges. The

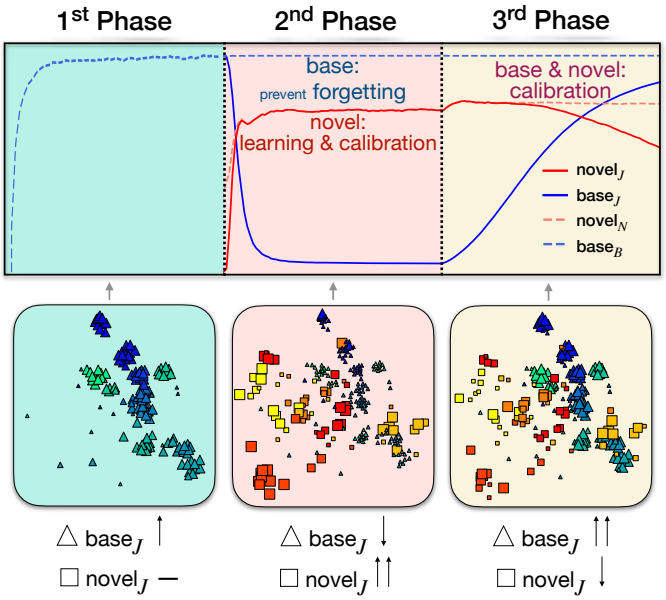


Figure 10.1: Overview of the performance of our framework during different phases. J indicates performance in the joint space, B and N denote performance in the base and novel spaces respectively. During the 1^{st} phase we train the model on the base classes. During the 2^{nd} phase we try to achieve a high performance on novel classes in the joint space and prevent forgetting of base classes. In the 3^{rd} phase we calibrate the two classifiers and achieve balanced performance between base and novel classes. T-SNE plots show the performance of the test samples at each phase. The symbol size shows the confidence of the model for the sample.

first phase is devoted to general representation learning as in previous work [GK18, RLFZ19, WGHH18]. Here, we utilize a large base dataset for pretraining and obtain high performance for base classification. In the second phase we concentrate on learning novel classes. In contrast to the prior work, we pay special attention to training a calibrated classifier for the novel classes while simultaneously preventing catastrophic forgetting for the base classes. More specifically we propose base-normalized cross entropy that amplifies the softmax output of novel classes to overcome the bias towards the base classes, and simultaneously enforce the model to preserve previous knowledge via explicit weight constraints. In the third phase we address the problem of calibrating the overall model across base and novel classes. In Fig. 10.1 we show how the model develops during all three phases by plotting the test accuracy of base and novel classes in the separate and joint spaces. The contributions of this work are as follows:

1. A framework to explicitly address the problems of generalized few-shot-learning by balancing between learning novel classes, forgetting base classes and calibration across them in three phases;

2. Base-normalized cross-entropy to overcome the bias learned by the model on the base classes in combination with weight constraints to mitigate the forgetting problem in the second phase;

3. An extensive study to evaluate the proposed framework on images and videos showing state-of-the-art results for generalized and incremental few shot learning.

10.2 Method

In the following we introduce the general setting and the motivation of our method based on four separate performance measures introduced below. We then discuss the second phase training as the most crucial part to achieve strong performance for both novel and base classes, followed by our third phase training. Finally, we discuss how to generalize our method to incremental few shot learning.

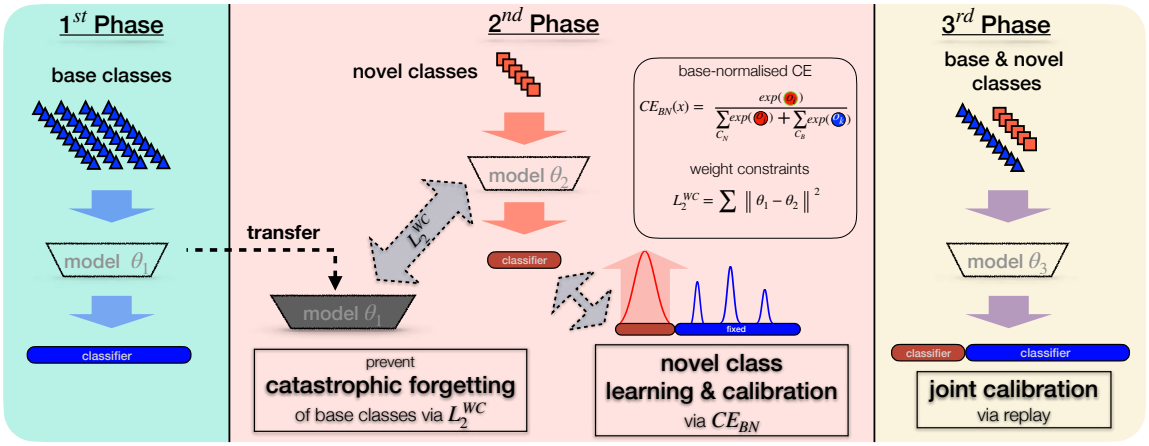


Figure 10.2: Overview of our framework. To achieve balanced performance on base and novel classes we deal with three problems *learning novel classes*, *catastrophic forgetting*, and *calibration* that we address in different phases of our framework. In the 1^{st} phase we pretrain model on base classes with abundant data. During the 2^{nd} phase we employ L_2^{WC} weight constraints to preserve knowledge and base-normalized cross entropy (CE_{BN}) to calibrate learning of novel classes in the joint space with base classes. In the 3^{rd} phase we calibrate the performance with the balanced replay of novel and base samples.

Setting: In both generalized and incremental few-shot learning we have a set of base classes C_B with many training samples. Additionally, we have one or several sets of novel classes C_N with only few training samples. In generalized few-shot learning we have just one set of novel classes, while in incremental few-shot learning we have a sequence of such sets. In the following, to keep the notation simple, we discuss our approach based on a single set of novel classes, whereas in incremental learning the approach is applied to the sequence of such sets of novel classes. Note that incremental few-shot learning in our work is the same as the few-shot class incremental learning in [CL21, THC+20].

Performance measures and approach: In few-shot learning we are interested to achieve best performance for both base and novel classes simultaneously. Therefore, in order to monitor performance for both sets of classes, we are considering four different measures (see Fig. 10.3) First, we denote $B_{/B}$ the classification performance of *base* samples in the space of only *base* classes C_B, and $N_{/N}$ the classification performance of *novel* samples in the space of only *novel* classes C_N. More importantly, we are interested in the performance in the *joint* (J) space where both base and novel classes are accounted for simultaneously $C_B \cup C_N$. For this we consider the performance of base and novel samples separately in the *joint* space, that is $N_{/J}$ and $B_{/J}$. We prefer these two measures rather than using only the joint performance in joint space due to the imbalance of the number of classes [HPL+19, BP19, WCW+19] with $|C_B| \gg |C_N|$ (e.g. 64 base vs. 5 novel).

These measures are directly related to the three challenges mentioned above: novel class

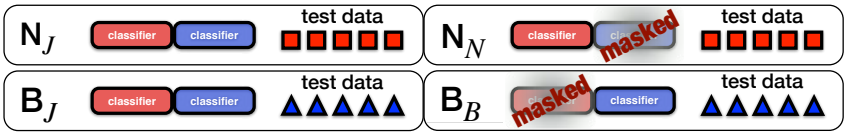


Figure 10.3: Classification layers for different evaluation protocols.

learning is measured by $N_{/N}$ and $N_{/J}$, catastrophic forgetting by $B_{/B}$ and $B_{/J}$, while calibration is related to $B_{/J}$ and $N_{/J}$. While ideally we would like to address all the measures simultaneously, we found this to be difficult in practice. Instead, during the first phase of our framework, we optimize for $B_{/B}$. In the second phase, during novel class learning, we are aiming for a calibrated classifier for the novel classes and thus optimize for both $N_{/N}$ and $N_{/J}$, instead of only $N_{/N}$ as in standard few-shot learning. Simultaneously, we aim to prevent catastrophic forgetting by an additional weight regularization that keeps $B_{/B}$ high (see Fig. 10.1). In the third and last phase we aim to calibrate across novel and base classes and thus optimize for both $B_{/J}$ and $N_{/J}$.

Model parameters: We denote the backbone parameters as θ_1, θ_2, and θ_3 for the first, the second, and the third phase respectively. As classifier we use a linear classification layer without bias that we train on the top of the backbone. Practically, during the second phase we introduce a classification layer for novel classes. To evaluate performance in the joint space we concatenate the output of the two classifiers before the normalization. o_i denotes the output logit of the model for the classification into class i. We train θ_1 on a large dataset of base classes to obtain a good representation. For the second and the third phase we initialize θ_{phase} with the parameters $\theta_{phase-1}$ and fine-tune on the corresponding to phase set.

10.2.1 Second Phase - Novel Class Training

Base-Normalized Cross Entropy (CE_{BN}) Recently, Tian et al. [TWK+20] showed that in few-shot learning competitive classification on novel classes can be achieved given good representations using standard cross entropy without meta-learning and prototypes. We follow this idea and train θ_2 using a pretrained model θ_1 from the first phase and fine-tune it with a new classification layer for novel classes. The standard way to fine-tune the model on the training set that includes C_N classes is

$$CE(x) = \sum_{C_N} y_i ln\Big(\frac{exp(o_i)}{\sum_{C_N} exp(o_j)}\Big), \tag{10.1}$$

where o_i is the logit of the corresponding class $i \in C_N$, and y_i equals to one if x belongs to class i, otherwise to zero.

One problem here is that even if the model is capable of learning information about the new classes C_N well, there is no guarantee that this performance is replicated in the joint space $N_{/J}$. By training two disjoint classifiers we learn classification weights that satisfy the classification problem either on novel classes $N_{/N}$ or on base classes $B_{/B}$, but so far the model does not learn any correlation between base classification weights and novel classification weights.

To this end, we propose to provide the model with information about the base class distribution in the joint space using readily available information. Specifically, for each novel training sample we compute logits not only for the novel C_N classes, but also for the base C_B classes (note, in the second phase the base classes classifier is kept fixed). We use these logits to

compute classification scores with the softmax function, thus the normalization of each score includes the base class logits that initially prevail in the sum, as follows:

$$CE_{BN}(x) = \sum_{C_N} y_i ln\Big(\frac{exp(o_i)}{\sum\limits_{C_N} exp(o_j) + \sum\limits_{C_B} exp(o_k)}\Big). \tag{10.2}$$

With this normalization, the novel model learns output probabilities for the novel classes directly in the joint space, and specifically increasing magnitude of novel class logits with respect to base class logits. This allows to have a good classification accuracy for novel classes $N_{/N}$ in the second phase and at the same time helps to match this accuracy in the joint space $N_{/J}$. Note that we do not use any base class training samples during this learning phase and we keep the weights of the base classifier fix.

Knowledge Preservation After the first phase the model performs well for the base classes and we aim to keep this capability. In FSL [DCRS20, RRBV19] and IL [LH17a, RKSL17b] multiple works show that adaptive representations can be beneficial for learning novel classes, specifically to fine-tune the parameters of the representation. In IL, the typical way to preserve knowledge from base classes [CMJG+18, LLSL19, WCW+19, DCO+20, LH17a, RKSL17b] is knowledge distillation (KD) [HVD15] that is applied in the form of KL-divergence between logits of base classes from adapted and old models.

As an alternative to keep the network to remember about the previous knowledge, we propose to utilise explicit weight constraints (WC) of the model with respect to the old model from the first phase. We formulate it in form of a L_2 penalization over adaptive parameters of the representation [LH17a, KPR+17, EP04]:

$$L_2^{WC} = \sum \|\theta_1 - \theta_2\|^2, \tag{10.3}$$

where θ_2 denotes adaptive parameters of the backbone excluding classification parameters during the second phase and θ_1 are the parameters of the model after base pretraining. The above constraint forces the model to keep the representation learned on base samples, but still allows the model to adjust the weights of the representation to better fit novel classes while not diverging a lot from the old model. The overall loss for the second phase is thus:

$$L = CE_{BN} + \lambda L_2^{WC}, \tag{10.4}$$

where λ controls the strength of knowledge preservation.

10.2.2 Third Phase - Joint Calibration

Balanced Replay The first and the second phase account for the performance on the base classes ($B_{/B}$) and for the novel class learning in both spaces ($N_{/N}$ and $N_{/J}$). For the third phase, due to the difference of number in training samples for base and novel classes and preservation of $B_{/B}$ during the second phase, the model is able to obtain good performance in the joint space as well. Empirically we found that during the second phase $B_{/J}$ performance can drop drastically, but due to keeping the base class performance $B_{/B}$ we can achieve good $B_{/J}$ performance in the third phase.

To achieve a balanced performance in the joint space of base and novel classes we apply the replay technique that is common in incremental learning [RKSL17b, LSL+20, RKH+20, LLSL19]. Specifically, we randomly draw only once base training samples, one per class, and join these samples with the novel training data. Moreover, in our case we require the least possible

memory [RKSL17b] to store exemplars of base classes, an essential component for incremental learning.

We continue training the model on the balanced dataset in the joint space. Due to the initial strong bias towards base classes from the first phase ($B_{/B}$), the model can improve its performance for base classes in the joint space ($B_{/J}$) quickly while at the same time overwriting the novel class performance at least partially ($N_{/J}$).

10.2.3 From Generalized to Incremental Learning

The main difference between GFSL [RLFZ19, SSS+20] and IFSL [THC+20, CL21] is the number of few-shot tasks. So far we considered the case of GFSL and it can be regarded as the first two tasks in terms of incremental setting: the training of the base classes refers to task one, the training of novel classes to task two in the incremental setup. As the current framework addresses the joint generalized problem in three phases, base classes, novel classes, and joint classes, we can easily extend the architecture to more tasks by repeating the novel class training. Specifically, for each new few-shot task we apply the second phase to learn a good joint classification for the current classes.

To evaluate the performance of the current joint space we finalize the training with the last phase of recuperation. To this end, we keep exemplars from base and novel classes from different tasks and perform training with the base-normalized cross-entropy loss and L_2^{WC} weights constraints as before. So, each time when we need to evaluate the joint performance on all classes, we apply the third phase.

For example, to report accuracy after five tasks, we learn representation parameters from base classes in the first task, as next stage we then apply the second phase sequentially for the second, third, fourth, and fifth tasks respectively, each time enlarging the classifier by the number of new classes in the task. After the fifth task we apply the third phase, balanced replay, where for each few-shot task we use all available data, and one sample per class for the data from the first task. During test the performance of the model is evaluated on a set that contains all previously seen classes.

10.3 Experiments

This section validates our proposed LCwoF-framework. First, we compare our method to the previous state-of-the-art work on both GFSL and IFSL in Section 10.3.1. Then we analyze each phase and the components separately to show the importance and connections of each to the improved performance in Section 10.3.2.

Datasets: *mini-ImageNet* [VBL+16] is a 100-class subset of ImageNet [DDS+09]. For FSL we follow [RLFZ19] and use a subsets 64-12-24 classes that corresponds to base-val-novel classes, and for IFSL we follow [CL21, THC+20] with a subset of 60 and 40 classes for base training and incremental few-shot testing with 5 classes per each novel set. *tiered-ImageNet* [RTR+18] is a larger subset of Imagenet [DDS+09] with categorical splits for for base, validation, and novel classes. Here, each high-level category (e.g. dog that includes different breeds) belongs only to one of the splits. *mini-Kinetics* [XKD+20] is a 100-class subset of the Kinetics video classification dataset [KCS+17]. We use the splits from [XKD+20]. *UCF101* [SZS12] is a video dataset with 101 classes in total. We follow [KDGM+19] with a splitting of 50-51 for base and novel classes for FSL on videos. We additionally introduce and evaluate more challenging division of the dataset.

method	mini-ImageNet 5w1s						mini-ImageNet 5w5s					
	$N_{/N}$ ($5_{/5}$)	$B_{/B}$ ($64_{/64}$)	$N_{/J}$ ($5_{/69}$)	$B_{/J}$ ($64_{/69}$)	$hm_{/J}$	$am_{/J}$	$N_{/N}$ ($5_{/5}$)	$B_{/B}$ ($64_{/64}$)	$N_{/J}$ ($5_{/69}$)	$B_{/J}$ ($64_{/69}$)	$hm_{/J}$	$am_{/J}$
CONV4												
PN [SSZ17]°	53.88	54.02	0.02	54.02	0.04	27.02	70.84	60.42	2.99	60.41	5.70	31.70
DFSL [GK18]°	55.80	69.93	40.30	58.54	47.74	49.42	72.24	70.24	58.26	59.89	59.06	59.07
RGFSL [SSS+20]	55.08	65.14	39.86	54.65	46.10	47.25	72.32	67.79	56.32	59.30	57.71	57.81
LCwoF (ours) *lim*	58.32	72.75	47.16	55.07	**50.81**	**51.12**	73.63	71.82	62.23	59.94	**61.06**	**61.09**
ResNet												
PN [SSZ17]⋆	-	-	-	-	-	42.73	-	-	-	-	-	57.05
IW [QBL18](i)	47.17	61.78	31.25	47.72	37.77	39.49	67.56	69.07	46.96	58.92	52.26	52.94
DFSL [GK18] (c)	56.83	70.15	41.32	58.04	48.27	49.68	72.82	70.03	59.27	58.68	58.97	58.98
AAN [RLFZ19](c)	56.14	77.58	45.61	63.92	53.24	54.76	69.72	77.58	60.82	64.14	62.43	62.48
AAN [RLFZ19](orig)	-	-	-	-	-	54.95	-	-	-	-	-	63.04
LCwoF (ours) *lim*	60.78	79.89	53.78	62.89	**57.39**	**57.84**	77.65	79.96	68.58	64.53	**66.49**	**66.55**
LCwoF (ours) *unlim*	61.15	80.10	53.33	62.99	**57.75**	**58.16**	77.88	80.09	67.17	66.59	**66.88**	**66.88**

Table 10.1: Comparison to state-of-the-art on mini-ImageNet 5w1s (left) and 5w5s (right) with backbones CONV4 and ResNet. *lim* denotes limited access to base train samples during the third phase, for *unlim* we do not apply such restrictions. ° indicates results copied from RGFSL [SSS+20], ⋆ indicates results from AAN [RLFZ19], (c) denotes that we run available code on the corresponding data, (i) states for our re-implementation of the respective method, (orig) indicates original numbers from the respective paper.

method	tiered-ImageNet 5w1s			tiered-ImageNet 5w5s		
	$N_{/J}$	$B_{/J}$	$hm_{/J}$	$N_{/J}$	$B_{/J}$	$hm_{/J}$
IW [QBL18](i)	44.95	62.53	52.30	71.85	56.11	63.01
DFSL [GK18](c)	47.32	36.10	40.96	67.94	39.08	49.61
AAN [RLFZ19](c)	54.39	55.85	55.11	57.76	64.13	64.93
LCwoF *lim*	57.13	60.39	**58.71**	69.05	63.44	**66.12**
LCwoF *unlim*	59.79	60.86	**58.75**	70.20	63.01	**66.41**

Table 10.2: Comparison to state-of-the-art on tiered-ImageNet 5w1s (left) and 5w5s (right) with ResNet backbone. $N_{/J}$ equal to $5_{/205}$, $B_{/J}$ to $200_{/205}$. *lim*, *unlim*, (i) and (c) see in Table 10.1.

Implementation details For the FSL experiments on mini-ImageNet and tiered-ImageNet we employ the same ResNet12 with DropBlock [GLL18] as backbone and pretrain it on base classes for 500 epochs with SGD optimizer with momentum with the learning rate (lr) of 1e-3 that is decayed by 0.1 at 75, 150, and 300 epochs. For the second and the third phase we use lr of 1e-2 and 1e-3 respectively for the classification layers and decayed by 0.1 lr for the backbone parameters. For IFSL experiments we use ResNet18 and follow the same pretraining steps as above. We use different architecture choices for GFSL and IFSL to remain comparable to previous works after the base pretraining. For the second phase we always train the model for 150 epochs, while for the third phase we use validation set to choose the number of epochs for each dataset. For videos we preextract features with C3D model [TBF+15] pretrained on large-scale Sports-1M [KTS+14] dataset. We apply average pooling over temporal domain to obtain one feature vector per video. As a backbone we use 2-layer MLP. We also clip gradients at value 100 for the experiments.

Evaluation We evaluate the proposed framework primarily with respect to the harmonic mean ($hm_{/J}$) [SES+19, SSS+20, KDGM+19] that is computed between base and novel performance in the joint space. Additionally, we report performance of base and novel classes in their

	mini-Kinetics 5w1s			mini-Kinetics 5w5s		
method	$N_{/J}$	$B_{/J}$	$hm_{/J}$	$N_{/J}$	$B_{/J}$	$hm_{/J}$
IW [QBL18](i)	45.56	48.56	47.01	56.92	49.17	52.76
DFSL [GK18](c)	50.81	44.51	47.45	70.29	46.31	55.83
GFSV [XKD$^+$20]	13.70	88.70	23.73	22.30	88.70	35.64
ANN [RLFZ19](c)	46.13	35.96	40.41	56.99	43.21	49.15
LCwoF *lim*	47.51	50.84	**49.12**	63.65	54.55	**58.75**
LCwoF *unlim*	46.26	51.94	**48.93**	65.40	52.70	**58.37**

Table 10.3: Comparison to state-of-the-art on mini-Kinetics 5w1s (left) and 5w5s (right) with MLP backbone on pre-extracted features. $N_{/J}$ equal to $5_{/69}$, $B_{/J}$ to $64_{/69}$. *lim*, *unlim*, (i) and (c) see in Table 10.1.

respective subspaces ($B_{/B}$ and $N_{/N}$), in the joint space ($B_{/J}$ and $N_{/J}$), and the arithmetic mean over the joint space ($am_{/J}$) as in [RLFZ19]. 5w1s and 5w5s denote 5 novel classes with 1 and 5 training samples per classes respectively. For the state-of-the-art comparison, we average over 600 episodes [SSS$^+$20, LLX$^+$19, DCRS20], for all other experiments over 100 episodes. *unlim* denotes access to the entire base training set, whereas for *lim* setup we use small subset.

10.3.1 Comparison to state-of-the-art

Generalized Few-Shot Learning We compare our performance on image and video benchmarks: mini-ImageNet, tiered-ImageNet, Kinetics and UCF101 in Tables 10.1, 10.2, 10.3, and 10.4 respectively. For mini-ImageNet and tiered-ImageNet, we train the respective backbones from scratch on the base classes. For Kinetics and UCF101, we preextract video features as in [KDGM$^+$19, XKD$^+$20] and then train a shallow MLP model on the base classes.

On mini-ImageNet in Table 10.1 we provide an evaluation in separate and joint space on 5w1s and 5w5s setups. For both backbones, conv4 [SSS$^+$20] and ResNet, we achieve significant improvements over state-of-the-art results in terms of $hm_{/J}$. Here, we can observe that previous methods drop in performance on both novel ($N_{/N}$) and base ($B_{/B}$) classes, whereas we address the problem by explicitly balancing between forgetting, learning, and calibration and achieve better performance.

On tiered-ImageNet, Table 10.2, we can observe a similar pattern and achieve strong improvements. Here, even with more base classes, we are able to calibrate the performance between novel and base classes.

We further evaluate the performance of the proposed idea on two video datasets. Our results on Kinetics, shown in Table 10.3, and UCF101, shown in Table 10.4, show that the proposed framework is able to perform well on the pre-extracted features. Results on UCF101 we present on two different splits for training and testing. In Table 10.4 the first two lines correspond to splits provided by [KDGM$^+$19] thus can be directly compared. The second part of the table shows the evaluation for the setup with the original UCF train/test split as defined in [SZS12].

Note that on both image datasets we obtain significant gains in performance while applying *unlim* sampling strategy, while on the Kinetics and UCF101 there is a slight decrease in comparison to *lim*. We speculate that it happens due to the fixed feature preextraction whereas on images we train models on raw images. Across all the datasets, setting and architectures we consistently achieve significantly better performance than previous work.

Incremental Few-Shot Learning We compare our method with the current few-shot class

method	UCF101 50w1s		
	$N_{/J}$	$B_{/J}$	$hm_{/J}$
ProtoG [KDGM+19]	52.30	75.30	61.72
LCwoF (ours) *lim*	54.41	91.41	**68.22**
IW [QBL18](i)	45.22	76.15	56.74
LCwoF (ours) *lim*	50.78	70.72	**59.11**
LCwoF (ours) *unlim*	49.12	69.98	**57.72**

Table 10.4: Comparison to state-of-the-art on UCF101 50w1s with MLP backbone on pre-extracted features. $N_{/J}$ equal to $50_{/101}$, $B_{/J}$ to $51_{/101}$. *lim*, *unlim* and (i) see in Table 10.1.

	mini-ImageNet 5w5s							
method	2	3	4	5	6	7	8	9
$hm_{/J}$	+5	+10	+15	+20	+25	+30	+35	+40
FT$^{\diamond}$	7.23	7.39	4.87	2.40	2.06	1.84	1.57	1.40
Joint$^{\diamond}$	8.92	17.02	21.86	20.54	22.92	22.85	24.41	24.95
iCaRL [RKSL17b]$^{\diamond}$	8.45	13.86	14.92	13.00	14.06	12.74	12.16	11.71
UCIR [HPL+19]$^{\diamond}$	9.62	14.14	15.58	13.19	13.63	13.11	12.76	11.96
PN [SSZ17]$^{\diamond}$	9.76	14.72	16.78	19.09	20.06	19.37	18.98	18.90
ILVQ [XSZ12]$^{\diamond}$	9.66	16.08	17.78	20.05	20.35	19.64	19.06	18.89
SDC [YTL+20]$^{\diamond}$	20.51	18.79	17.36	20.47	19.21	18.27	20.79	21.77
IW [QBL18]$^{\diamond}$	25.32	20.45	22.62	25.48	22.54	20.66	21.27	22.27
IDLVQ [CL21]	21.69	20.44	21.98	25.19	22.99	20.82	21.56	22.65
LCwoF *lim*	**41.24**	**38.96**	**39.08**	**38.67**	**36.75**	**35.47**	**34.71**	**35.02**

Table 10.5: IFSL. Comparison to state-of-the-art on mini-ImageNet. Number of base classes for the first task is 60. Each next task increases joint space by 5 novel classes with 5 training samples per class. $^{\diamond}$ indicates results copied from IDLVQ [CL21]. $hm_{/J}$ between base and novel classes in the joint space for each task.

incremental methods in Tables 10.5and 10.6. As in the previous experiments we use the hm accuracy that we compute between base (first set) and novel classes. We provide more detail on the performance of each task in Tables 10.7, 10.8, and 10.9, specifically showing performance of base and novel classes separately, as well as the standard accuracy of all classes in the joint space. Our method notably outperforms other methods in the field due to the fact that we address directly the balance between the performance on base and novel classes. We show that we obtain higher novel accuracy in the joint space for every task.

10.3.2 Ablation Studies

Here we investigate the influence of several components of our framework and the impact of various hyperparameters.

Calibration and forgetting during the second phase In this subsection we analyse the influence of the base-normalized cross entropy as a technique to address the calibration problem as well as the influence of knowledge preservation to address the forgetting issue. In Fig. 10.4 we show the behaviour of the training process during the second and the third phase with and without base-normalized cross entropy and knowledge preservation as explicit weight constraints L_2^{WC}. Removing both elements, as shown in the top left sub-figure, results in a drastic drop in $B_{/B}$

mini-ImageNet 5w5s				
$B_{/J}$ (60)	2	5	7	9
Joint$^{\diamond}$	63.30	62.18	61.86	61.89
SDC [YTL^{+}20]$^{\diamond}$	63.58	60.29	59.05	59.87
IW [QBL18]$^{\diamond}$	63.52	61.17	60.63	59.64
IDLVQ [CL21]	63.77	61.22	60.97	60.44
LCwoF *lim*	57.33	51.38	47.60	47.73
$N_{/J}$	2	5	7	9
#cl	5	20	30	40
Joint$^{\diamond}$	4.80	12.30	14.01	15.62
SDC [YTL^{+}20]$^{\diamond}$	12.23	12.33	10.81	13.30
IW [QBL18]$^{\diamond}$	15.81	16.09	12.45	13.69
IDLVQ [CL21]	13.07	15.86	12.55	13.94
LCwoF *lim*	**32.20**	**31.12**	**28.27**	**27.65**

Table 10.6: IFSL. Comparison to state-of-the-art on mini-ImageNet. Number of base classes for the first task is 60. Each next task increases joint space by 5 novel classes with 5 training samples per class. $^{\diamond}$ indicates results copied from IDLVQ [CL21]. Top: base classification accuracy in the joint space; Bottom: novel classification accuracy in the joint space.

mini-ImageNet									
method	1	2	3	4	5	6	7	8	9
$hm_{/J}$	60	+5	+10	+15	+20	+25	+30	+35	+40
FT$^{\diamond}$	-	7.23	7.39	4.87	2.40	2.06	1.84	1.57	1.40
Joint$^{\diamond}$	-	8.92	17.02	21.86	20.54	22.92	22.85	24.41	24.95
iCaRL [RKSL17b]$^{\diamond}$	-	8.45	13.86	14.92	13.00	14.06	12.74	12.16	11.71
UCIR [HPL^{+}19]$^{\diamond}$	-	9.62	14.14	15.58	13.19	13.63	13.11	12.76	11.96
PN [SSZ17]$^{\diamond}$	-	9.76	14.72	16.78	19.09	20.06	19.37	18.98	18.90
ILVQ [XSZ12]$^{\diamond}$	-	9.66	16.08	17.78	20.05	20.35	19.64	19.06	18.89
SDC [YTL^{+}20]$^{\diamond}$	-	20.51	18.79	17.36	20.47	19.21	18.27	20.79	21.77
IW [QBL18]$^{\diamond}$	-	25.32	20.45	22.62	25.48	22.54	20.66	21.27	22.27
IDLVQ [CL21]	-	21.69	20.44	21.98	25.19	22.99	20.82	21.56	22.65
LCwoF (base biased)	-	25.56	30.59	27.29	28.08	29.91	27.97	30.30	32.73
LCwoF (balanced hm)	-	**41.24**	**38.96**	**39.08**	**38.67**	**36.75**	**35.47**	**34.71**	**35.02**

Table 10.7: IFSL. Comparison to state-of-the-art on mini-ImageNet based on harmonic mean metric between base and novel classes. $^{\diamond}$ indicates results copied from IDLVQ [CL21].

and $B_{/J}$ that prohibits the model to quickly recover during the replay (third) phase since all previous knowledge is lost. The left bottom sub-figure shows the impact of CE_{BN}. With CE_{BN} the model easily achieves good performance on $N_{/J}$ in the joint space that matches $N_{/N}$, but here both $B_{/B}$ and $B_{/J}$ drop during the second phase training that again prevents recuperation. Compared to that, the right top sub-figure includes base knowledge preservation that keeps both $B_{/B}$ and $B_{/J}$ relatively high, and further facilitates complete recovery for the base classes. But without CE_{BN} novel learning in the joint space suffers during both the second and the third phase. The right bottom sub-figure shows performance if both, CE_{BN} and knowledge preservation, are used. The model is able to keep a certain performance of $B_{/B}$ during the second phase while achieving high accuracy on the novel classes ($N_{/N}$ and $N_{/J}$). During the third phase we calibrate the model by the replay technique and achieve good balance between the two disjoint sets of classes.

mini-ImageNet									
$B_{/J}$ (60)	1	2	3	4	5	6	7	8	9
FT$^{\diamond}$	64.25	32.28	20.87	6.95	3.17	3.16	1.92	1.53	1.46
Joint$^{\diamond}$	64.25	63.30	62.83	62.16	**62.18**	**62.68**	**61.86**	**61.87**	**61.89**
iCaRL [RKSL17b]$^{\diamond}$	64.25	51.66	48.97	45.62	37.39	30.86	28.68	26.83	24.47
UCIR [HPL^{+}19]$^{\diamond}$	64.25	52.87	50.16	44.78	37.48	28.75	25.58	22.97	21.57
PN [SSZ17]$^{\diamond}$	64.25	59.27	58.88	58.69	58.22	57.63	57.03	56.80	56.47
ILVQ [XSZ12]$^{\diamond}$	64.25	60.24	59.62	59.02	58.61	57.71	57.16	56.83	56.49
SDC [YTL^{+}20]$^{\diamond}$	64.62	63.58	62.78	61.12	60.29	59.37	59.05	59.97	59.87
IW [QBL18]$^{\diamond}$	64.71	63.52	62.96	62.13	61.17	61.27	60.63	59.86	59.64
IDLVQ [CL21]	**64.77**	**63.77**	**63.22**	**62.44**	61.22	61.47	60.97	60.66	60.44
LCwoF (base biased)	64.45	63.53	62.07	61.55	60.85	59.26	58.25	57.23	55.98
LCwoF (balanced hm)	64.45	57.33	53.31	52.87	51.38	48.25	47.60	47.51	47.73
$N_{/J}$	1	2	3	4	5	6	7	8	9
#*cl*	-	5	10	15	20	25	30	35	40
FT$^{\diamond}$	-	4.07	4.49	3.75	1.93	1.53	1.77	1.61	1.36
Joint$^{\diamond}$	-	4.80	9.84	13.26	12.30	14.03	14.01	15.21	15.62
iCaRL [RKSL17b]$^{\diamond}$	-	4.60	8.09	8.92	7.87	9.10	8.19	7.86	7.70
UCIR [HPL^{+}19]$^{\diamond}$	-	5.29	8.23	9.43	8.00	8.93	8.81	8.83	8.27
PN [SSZ17]$^{\diamond}$	-	5.32	8.41	9.79	11.42	12.14	11.67	11.39	11.35
ILVQ [XSZ12]$^{\diamond}$	-	5.25	9.29	10.47	12.09	12.35	11.86	11.45	11.34
SDC [YTL^{+}20]$^{\diamond}$	-	12.23	11.05	10.12	12.33	11.46	10.81	12.58	13.30
IW [QBL18]$^{\diamond}$	-	15.81	12.21	13.83	16.09	13.81	12.45	12.93	13.69
IDLVQ [CL21]	-	13.07	12.19	13.34	15.86	14.14	12.55	13.11	13.94
LCwoF (base biased)	-	16.00	20.30	17.53	18.25	20.00	18.40	20.60	23.12
LCwoF (balanced hm)	-	**32.20**	**30.70**	**31.00**	**31.12**	**29.68**	**28.27**	**27.34**	**27.65**

Table 10.8: IFSL. Comparison to state-of-the-art on mini-ImageNet. Top: performance of the base samples in the joint space after each task. Bottom: performance of the novel samples in the joint space after each novel task. $^{\diamond}$ indicates results copied from IDLVQ [CL21].

mini-ImageNet both$^{\uplus}$									
	1	2	3	4	5	6	7	8	9
$J_{/J}$	60	+5	+10	+15	+20	+25	+30	+35	+40
FT$^{\diamond}$	64.25	30.11	18.53	6.31	2.86	2.86	1.87	1.56	1.42
Joint$^{\diamond}$	64.25	58.80	55.26	52.38	49.71	48.37	45.91	44.68	43.38
iCaRL [RKSL17b]$^{\diamond}$	64.25	48.04	43.13	38.28	30.01	24.46	21.85	19.84	17.76
UCIR [HPL^{+}19]$^{\diamond}$	64.25	49.21	44.17	37.71	30.11	22.92	19.99	17.96	16.25
PN [SSZ17]$^{\diamond}$	64.25	55.12	51.67	48.91	46.52	44.25	41.91	40.07	38.42
ILVQ [XSZ12]$^{\diamond}$	64.25	56.01	52.43	49.31	46.98	44.37	42.06	40.11	38.43
SDC [YTL^{+}20]$^{\diamond}$	64.62	59.63	55.39	50.92	48.30	45.28	42.97	42.51	41.24
IW [QBL18]$^{\diamond}$	64.71	59.85	55.71	52.47	49.90	47.31	44.57	42.57	41.26
IDLVQ [CL21]	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84
TOPIC [THC^{+}20]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42
LCwoF (base biased)	64.45	**59.88**	**56.10**	**52.75**	**50.20**	**47.71**	**44.97**	**43.74**	**42.84**
LCwoF (balanced hm)	64.45	55.40	50.08	48.49	46.28	42.78	41.16	40.08	39.70

Table 10.9: IFSL. Comparison to state-of-the-art on mini-ImageNet based on joint performance of base and novel samples in the joint space. $^{\diamond}$ indicates results copied from IDLVQ [CL21].

Knowledge Preservation One important factor of the method is that we aim to retain the knowledge that the model obtained in the first phase, specifically the $B_{/B}$ performance, during the second phase. In this section we evaluate two different methods to achieve this objective,

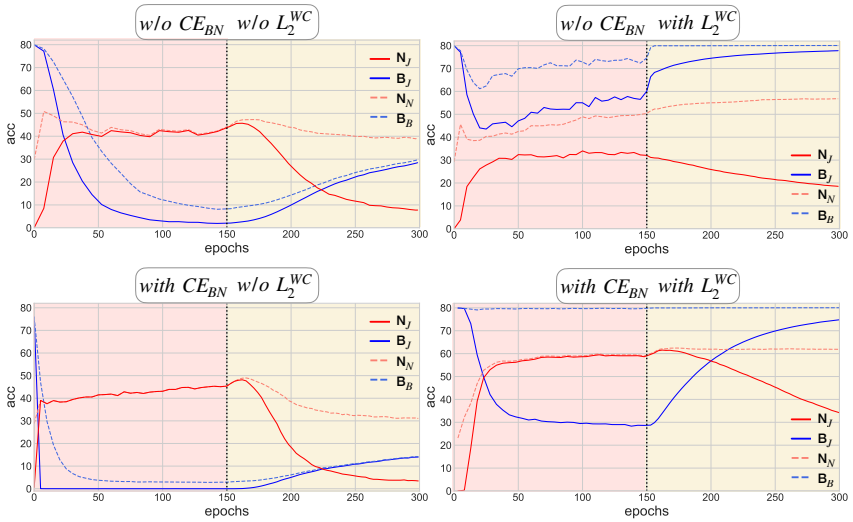


Figure 10.4: Influence of the proposed CE_{BN} and L_2^{WC} on the training during the 2^{nd} and the 3^{rd} phase. Red lines correspond to novel classes, blue to base classes. Solid lines represent performance in the joint space, dashed lines in separate corresponding subspaces. *Top left*: base performance drops as $B_{/B}$ and $B_{/J}$ and is not able to recover during the 3^{rd} phase; *bottom left*: both $N_{/N}$ and $N_{/J}$ achieve high performance, base performance still is not able to recover; *top right*: L_2^{WC} helps to recuperate base performance in the joint space while novel performance drops; *bottom right*: by addressing both $N_{/J}$ with CE_{BN} and $B_{/B}$ with L_2^{WC} we allow the model recover during the 3^{rd} phase on $B_{/J}$ while not losing drastically on novel classes.

method	5w1s $N_{/J}$ ($5/_{69}$)	5w1s $B_{/J}$ ($64/_{69}$)	5w1s $hm_{/J}$	5w5s $N_{/J}$ ($5/_{69}$)	5w5s $B_{/J}$ ($64/_{69}$)	5w5s $hm_{/J}$
L_2^{WC}	53.28	63.24	**57.83**	68.61	64.73	66.61
KD	50.28	57.72	53.74	70.29	64.67	**67.36**
KD^+	45.87	67.20	54.52	68.01	65.88	66.93
$L_2^{WC} + KD$	53.43	60.51	56.75	68.45	63.47	65.86

Table 10.10: Comparison of knowledge preservation techniques, such as L_2^{WC} as explicit weights constraints, knowledge distillation (KD) for old classes during the 2^{nd} phase, KD^+ includes KD on the additional 1000 unlabeled images during the second phase, and combination $L_2^{WC} + KD$. Results are computers on mini-ImageNet.

comparing the proposed explicit weight constraints L_2^{WC} with knowledge distillation that is formulated via KL-divergence. Knowledge distillation is a common technique to preserve knowledge in incremental learning [CMJG$^+$18, DCO$^+$20, LLSL19, LH17a, RKSL17b, WCW$^+$19], where abundant training data is available for new classes.

In Table 10.10 we evaluate the performance on 100 episodes on mini-ImageNet for the 5w1s and 5w5s settings. Comparing L_2^{WC} and KD knowledge preservation techniques, we find that the latter marginally outperforms the other if more data is available, as in the 5w5s setting, whereas plain weight constraints are more efficient in the lowest data regime with 1 training sample per class (5w1s). Additionally, we evaluate KD by including additional unlabeled 1000 images from the validation set during the second phase for KD loss computations, denoted as KD^+ in Table 10.10. We find that it helps to improve in the 5w1s setting, but still stays lower than L_2^{WC}. Applying both techniques at the same time does not give an improvement.

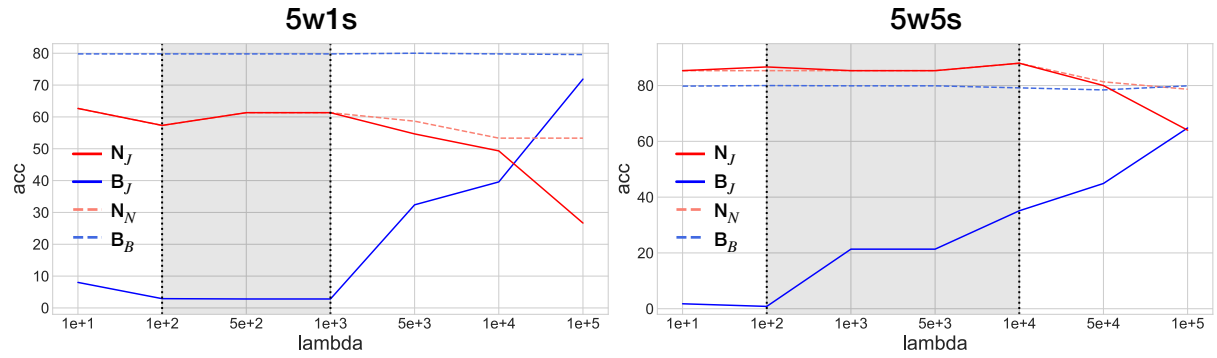


Figure 10.5: Evaluation of the performance after the 2^{nd} phase of the framework on the base and novel classes that depends on importance weight λ of the L_2^{WC} term. Filled areas correspond appropriate λ that allows us to achieve the desirable performance (before $N_{/J}$ drops). Results on mini-ImageNet.

	5w1s		
λ	$N_{/J}$	$B_{/J}$	$hm_{/J}$
1e+4	50.89	50.68	50.78
5e+4	53.06	58.56	55.67
1e+5	52.65	58.70	55.51
LCwoF	53.28	63.24	**57.83**

Table 10.11: Best possible performance of the model that can be achieved during the second phase (without the third phase). λ stands for the importance weight of the L_2^{WC} term. Higher λ, higher knowledge preservation, higher accuracy $B_{/J}$. We apply GCE and L_2^{WC} during the second phase with various λ. Results on mini-ImageNet.

Impact of λ In this section we study the influence of λ for the L_2^{WC} loss and how to preserve more knowledge and drop less on the base classes $B_{/J}$, but also how the model behaviour changes if we enforce it to preserve even more. In Fig. 10.5 we plot the accuracy after training in the second phase, before the replay phase for different λ. In Fig. 10.5 we fill these areas with gray color, that allow to reach our two objectives for the second phase, i.e., achieving good performance in the joint space on novel classes $N_{/J}$ and keeping good accuracy in the base space $B_{/B}$. Specifically higher λ helps to further keep $B_{/B}$ performance while novel learning $N_{/N}$ and $N_{/J}$ starts decreasing with higher values. It shows the start of the decrease depends on the number of available training samples: the more training data, the more we can keep from base classes by choosing a higher λ.

Early Stopping: Second Phase We observe that during the second phase, usually $B_{/J}$ starts dropping even when $N_{/J}$ has already achieved some reasonable accuracy, as can e.g. be seen in the bottom right subplot of Figure 10.4. We, therefore, also report the best performance that can be achieved with different λ during the second phase in Table 10.11. Note that λ influences the contribution of the knowledge preservation part L_2^{WC}. Thus, $B_{/J}$ will drop faster if a lower λ is chosen. By adjusting λ, we find that the proposed technique can also be helpful during the second phase. It shows that in this case, balanced model performance can be reached with higher λ and that we can achieve high performance even without the third replay phase.

Impact of batch normalization We use batch norm layers in the model that capture statistic from the base classes during the first phase. At the second phase our data is highly limited

	$N_{/J}$	$B_{/J}$	$hm_{/J}$
no	54.39	59.13	56.66
yes	53.28	63.24	**57.83**

Table 10.12: Influence of freezing batch norm during the 2^{nd} and the 3^{rd} phases, results on mini-ImageNet 5w1s.

		$N_{/J}$	$B_{/J}$	$hm_{/J}$
1	cos.norm.	47.91	59.69	53.15
2	bias	52.96	62.05	57.14
3	no bias	53.28	63.24	**57.83**

Table 10.13: Comparison between different linear layers: cos is cosine normalization of the weights and the embedding space, bias stands for linear layer with the bias term, no bias stands for linear layer without the bias term. Results on mini-ImageNet 5w1s.

thus we fix batch norm during further training. Table 10.12 shows that the performance drops more than 1 point when the model tries to accumulate new statistic from 1 training sample per class and to adapt parameters respectively.

Impact of normalization One of the common strategies to unify magnitudes of base and novel classifiers is to use cosine normalization of the embedding and weights [HPL+19, RLFZ19, SSS+20, MGB21, GK19, RKSL17b]. We experiment with such a setup (Table 10.13, lines 1 & 3) for our framework and find a decline in performance for both, $N_{/J}$ and $B_{/J}$. Note that the performance of the model after the first phase on $B_{/B}$ is the same as without cosine normalization. But if we attempt to match $N_{/J}$ and $N_{/N}$ during the second phase, we find that constrained magnitudes of logits due to normalization restrict the performance and do not allow to achieve our second phase objectives.

Analysis of classification layer As default, we conduct all our experiments with a linear classification layer without bias term (Table 10.13, lines 2 & 3). We therefore assess the performance of the model with and without bias. We find that during the first phase, it is the same, but that during the second and the third phase it is beneficial to use the latter giving a boost of about 0.7 in the performance $hm_{/J}$.

10.4 Conclusion

This chapter addresses major challenges in generalized few-shot and incremental few-shot learning with our three-phase framework. First, we learn a powerful representation by training a model on base classes. In the second phase, concerned with novel class learning, we employ base-normalized cross entropy that calibrates novel class classifiers to overcome the bias towards base classes. Additionally, during that phase we preserve knowledge about base classes via weight constraints. In the third phase, to achieve calibrated classifiers across both base and novel classes, we employ balanced replay. We show that each phase of the framework allows to study and address the essential problems of the task explicitly. We evaluate our proposed framework on four benchmark image and video datasets and achieve state-of-the-art performance across all settings. This work can be seen as a first step towards more explicitly addressing calibration, learning and knowledge preservation jointly to further improve deep learning for imbalanced settings beyond the ones addressed in this chapter. In the next Chapter 11, we tackle the domain of incremental few-shot learning from the perspective of

metric learning. We demonstrate the effective utilization of abundant pretraining data and limited data for novel tasks by introducing techniques such as space reservation and margin enhancement.

11 OrCo: Towards Better Generalization via Orthogonality and Contrast for Few-Shot Class-Incremental Learning

Contents

To tackle the challenges of a continuously expanding learning space with limited data, this chapter explores the effects of pretraining, space reservation, and margin enhancement techniques for Few-Shot Class Incremental Learning (FSCIL). FSCIL methods inherently face the challenge of catastrophic forgetting as data arrives incrementally, making models susceptible to overwriting previously acquired knowledge. Moreover, given the scarcity of labeled samples available at any given time, models may be prone to overfitting and find it challenging to strike a balance between extensive pretraining and the limited incremental data. To address these challenges, we propose the OrCo framework built on two core principles: features' orthogonality in the representation space, and contrastive learning. In particular, we improve the generalization of the embedding space by employing a combination of supervised and self-supervised contrastive losses during the pretraining phase. Additionally, we introduce the OrCo loss to address challenges arising from data limitations during incremental sessions. Through feature space perturbations and orthogonality between classes, the OrCo loss maximizes margins and reserves space for the following incremental data. This, in turn, ensures the accommodation of incoming classes in the feature space without compromising previously acquired knowledge. Our experimental results showcase state-of-the-art performance across three benchmark datasets, including mini-ImageNet, CIFAR100, and CUB datasets. The code will be made publicly available.

This chapter is based on [AKS24]. This paper accepted as a highlight at CVPR 2024. Anna Kukleva, as the co-first author, contributed to the conceptual development of this project and the writing of the paper. This paper is based on the master thesis of Noor Ahmed. Anna Kukleva acted as Noor Ahmed's supervisor during the project and played a significant role in the paper's writing.

11.1 Introduction

Real-world applications frequently encounter various challenges when acquiring data incrementally, with new information arriving in continuous portions. This scenario is commonly re-

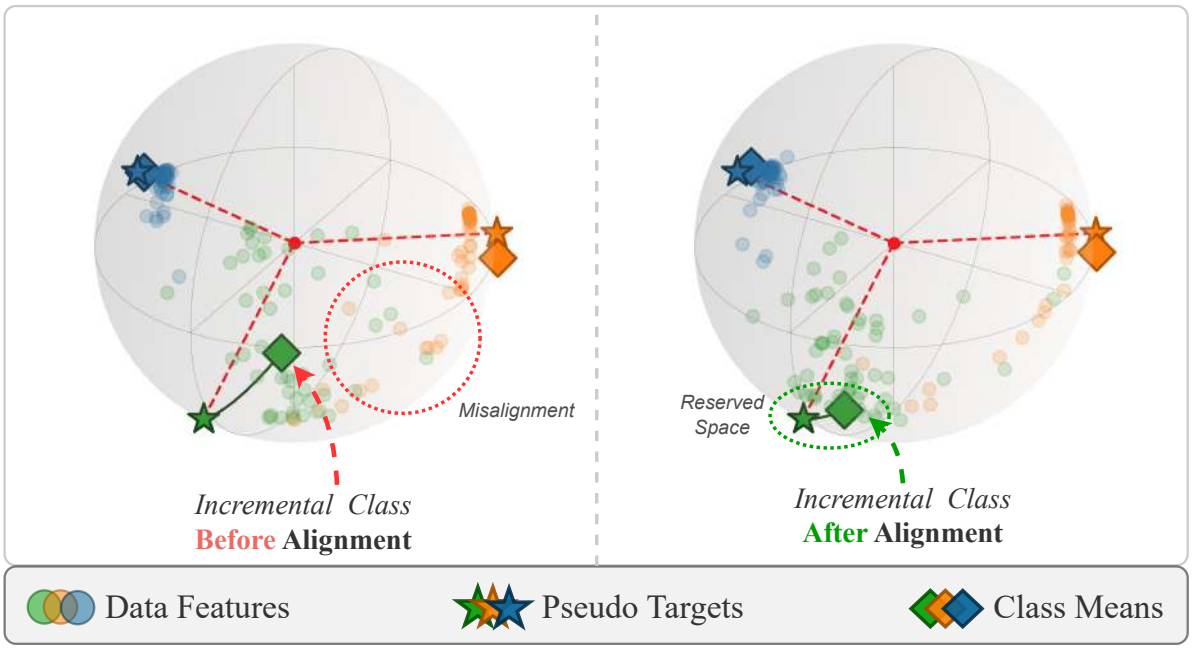


Figure 11.1: **PCA analysis on feature space before and after alignment.** Left: Before aligning incremental classes to orthogonal pseudo-targets. Right: After aligning incremental classes to assigned targets using **OrCo** loss. Our loss effectively reduces misalignment. Additionally, it enhances generalization for incoming classes by explicitly reserving space.

ferred to as Class Incremental Learning (CIL) [RKSL17a, WCW+19, HVD15, LH17a, CMJG+18, HPL+19, BP19, ZZW+21, ZXG+20, SCL+18, LMH+18, CDAT18]. Within CIL, the foremost challenge lies in preventing catastrophic forgetting [MC89, GMX+13, KPR+17], where previously learned concepts are susceptible to being overwritten by the latest updates. However, in a Few-Shot Class Incremental Learning (FSCIL) scenario [THC+20, ZSL+21, KKS21a, ZLX+23, KHSM22, ZFK+21, CRF+21, ZWY+22, YYL+23, her, MSR21, her], characterized by the introduction of new information with only a few labeled samples, two additional challenges emerge: overfitting and intransigence [SSZ17, CLK+19]. Overfitting arises as the model may memorize scarce input data and lose its generalization ability. On the other hand, intransigence involves maintaining a delicate balance, preserving knowledge from abundant existing classes while remaining adaptive enough to learn new tasks from a highly limited dataset. Advances in dealing with catastrophic forgetting, overfitting and intransigence are important steps toward improving the practical value of these methods.

Catastrophic forgetting is commonly tackled in CIL methods [RKSL17a, HPL+19, KK18], which assume ample labeled training data. However, standard CIL methods struggle in scenarios with limited labeled data, such as FSCIL [THC+20]. To address the three challenges posed by FSCIL, recent approaches [ZLX+23, ZSL+21] focus primarily on regularizing the feature space during incremental sessions, mitigating the risk of overfitting. These methods rely on a frozen backbone pretrained with standard cross-entropy on a substantial amount of data from the base session. However, we argue that achieving high performance on the pretraining dataset may not necessarily result in optimal generalization in subsequent incremental sessions with limited data. Therefore, in the first phase, we propose enhancing feature space generalization through contrastive learning, leveraging data from the base session.

In this work, we introduce the OrCo framework, a novel approach built on two fundamental

pillars: features' mutual orthogonality on the representation hypersphere and contrastive learning. During the first phase, we leverage supervised [KTW+20] and self-supervised contrastive learning [GH10, CKNH20, OLV18] for pretraining the model. The interplay between these two learning paradigms enables the model to capture various types of semantic information that is particularly beneficial for the novel classes with limited data [CFN+22, ICP+21, ADC+22], implicitly addressing the *intransigence* issue. After the pretraining, we generate and fix mutually orthogonal random vectors, further referred to as pseudo-targets. In the second phase, we aim to align the fixed pretrained backbone to the pseudo-targets using abundant base data. The learning objective during this phase is our OrCo loss, which consists of three integral components: perturbed supervised contrastive loss (PSCL), loss term that enforces orthogonality of features in the embedding space, and standard cross-entropy loss. Notably, our PSCL leverages generated pseudo-targets to maximize margins between the classes and to preserve space for incremental data, enhancing orthogonality through contrastive learning (see figure 11.1). The third phase of our framework, which we apply in each subsequent incremental session, similarly aims to align the model with the pseudo-targets, but using only few-shot data from the incremental sessions. During the third phase, our PSCL addresses limited data challenges, mitigating the *overfitting* problem to the current incremental session and *catastrophic forgetting* of the previous sessions through margin maximization.

We summarize the contributions of this work as follows:

- We introduce the novel OrCo framework designed to tackle FSCIL, that is built on orthogonality and contrastive learning principles throughout both pretraining and incremental sessions.
- Our perturbed supervised contrastive loss introduces perturbations of orthogonal, data-independent vectors in the representation space. This approach induces increased margins between classes, enhancing generalization.
- We showcase robust performance on three datasets, outperforming previous state-of-the-art methods. Furthermore, we perform a thorough analysis to evaluate the importance of each component.

11.2 Method

We begin with necessary preliminaries in section 11.2.1, followed by the description of our OrCo framework in section 11.2.2 and OrCo loss in section 11.2.3.

11.2.1 Preliminaries

FSCIL Setting. FSCIL consists of multiple incremental sessions. An initial 0-th session is often reserved to learn a generalisable representation on an abundant base dataset. This is followed by multiple few-shot incremental sessions with limited data. To formalise, an M-Session N-Way and K-Shot FSCIL task consists of $D_{seq} = \{D^0, D^1, ..., D^M\}$. These are all the datasets written in sequence where $D^i = \{(x_i, y_i)\}_{i=1}^{|D^i|}$ is the dataset for the i-th session. The 0-th session dataset D^0, also referred to as base dataset, consists of C^0 classes, each with a large number of samples. The training set for each following few-shot incremental session ($i > 0$) has N classes. Each of these classes has K samples, typically ranging from 1 to 5 samples per class. Taking the i-th session as an example, the model's performance is assessed on validation sets from the current

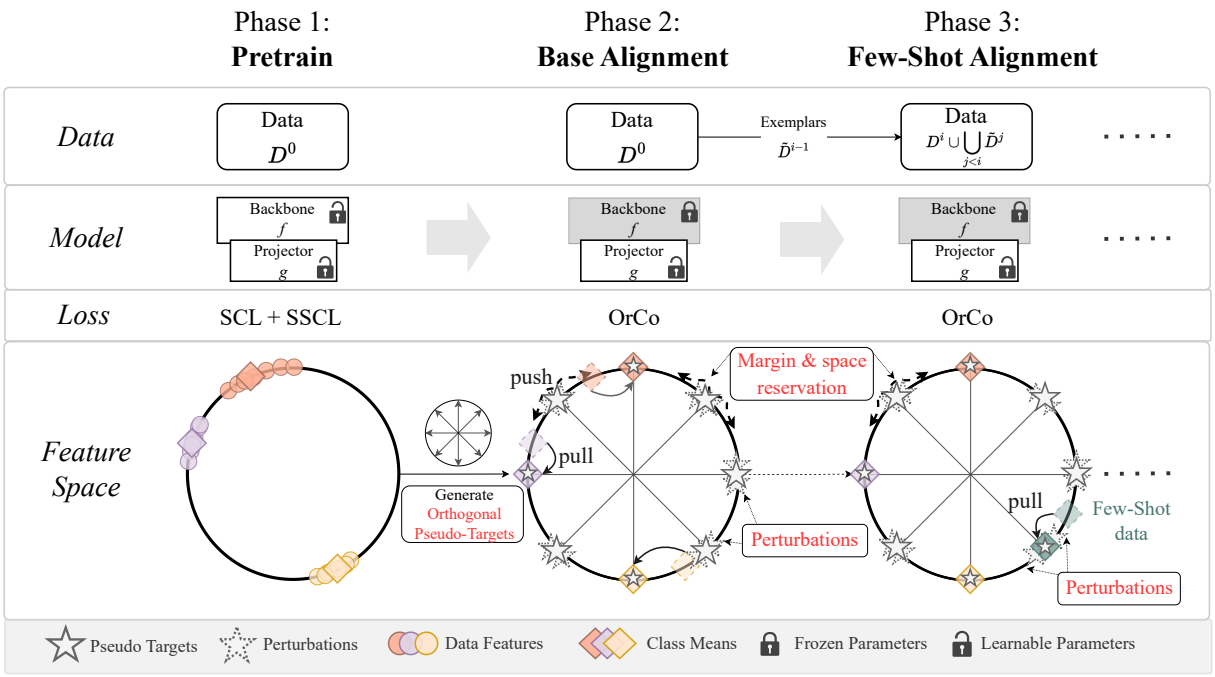


Figure 11.2: **Overview of OrCo framework.** Our OrCo framework is a three-phase approach for FSCIL. Phase 1 (Pretrain): We pretrain both backbone and projection head with SCL and SSCL on base dataset D^0. Before the next phase, we generate mutually orthogonal pseudo-targets. Phase 2 (Base Alignment): We aim to align the base dataset D^0 to the pseudo-targets through our OrCo loss. This involves pulling class means towards the nearest pseudo-targets and pushing forces based on perturbations around unassigned pseudo-targets (grey stars without assigned colored class means) to increase the margin and preserve space for incoming classes. Phase 3 (Few-Shot Alignment): Phase 3, employed in each subsequent incremental session, is similar to Phase 2 and assigns pseudo-targets to incremental class means with further alignment with our OrCo loss.

(i-th) and all previously encountered datasets ($< i$). The entire FSCIL task comprises a total of C classes. In our OrCo framework, we use base dataset D^0 for pretraining the model during phase 1 and for base alignment during phase 2.

Target Generation. We employ a Target Generation loss, similarly as in [LCY+22], to generate mutually orthogonal vectors across the representation hypersphere with a dimensionality of d. First, we define a set of random vectors $T = \{t_i\}$ where $\{t_i\} \in \mathbb{R}^d$. The optimization of the following loss with respect to these random vectors maximizes the angle between any pair of vectors $t_i, t_j \in T$, thereby ensuring their mutual orthogonality:

$$\mathcal{L}_{TG}(T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \log \sum_{j=1}^{|T|} e^{t_i \cdot t_j / \tau_o} \tag{11.1}$$

where τ_o is the temperature parameter. These optimized vectors, which we further refer to as pseudo-targets, remain fixed throughout our training process.

Contrastive Loss. The objective of contrastive representation learning is to create an embedding space where similar sample pairs are in close proximity, while dissimilar pairs are distant. In this work, we adopt the InfoNCE loss [KTW+20, OLV18] as our contrastive objective.

With positive set P_i and negative set N_i defined for each data sample z_i, called anchor, this loss aims to bring any $z_j \in P_i$ closer to its anchor z_i and push any $z_k \in N_i$ further away from the anchor z_i:

$$\mathcal{L}_{CL}(i;\theta) = \frac{-1}{|P_i|} \sum_{z_j \in P_i} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum\limits_{z_k \in N_i} \exp(z_i \cdot z_k/\tau)}, \tag{11.2}$$

where τ is the temperature parameter. In the classical self-supervised contrastive learning (SSCL) scenario, where labels for individual instances are unavailable, the positive set comprises augmentations of the anchor, and all other instances are treated as part of the negative set [OLV18]. In contrast, for the supervised contrastive loss (SCL), the positive set includes all instances from the same class as the anchor, while the negative set encompasses instances from all other classes [KTW+20]. To enhance clarity, we denote the supervised contrastive loss as $\mathcal{L}_{SCL}$ and self-supervised contrastive loss as $\mathcal{L}_{SSCL}$.

Supervised Contrastive Loss. Given a set of sample-label pairs $(x_i, y_i) \in Z^{SCL}$, we define the positive set P_i^{SCL} for x_i as the collection of pairs (x_j, y_j) where j varies over all instances such that $y_j = y_i$. Correspondingly, the negative set N_i^{SCL} is defined as $Z^{SCL} \setminus P_i^{SCL}$. Then, supervised contrastive loss (SCL) is defined as:

$$\mathcal{L}_{SCL}(i;\theta) = \frac{-1}{|P_i^{SCL}|} \sum_{x_j \in P_i^{SCL}} \log \frac{\exp(x_i \cdot x_j/\tau)}{\sum\limits_{x_k \in N_i^{SCL}} \exp(x_i \cdot x_k/\tau)}.$$

Self-Supervised Contrastive Loss. Given a set of samples $x_i \in Z^{SSCL}$, we define a positive for x_i as $A(x_i)$ where $A(\cdot)$ is a random transformation. Then, self-supervised contrastive loss (SSCL) loss is defined as:

$$\mathcal{L}_{SSCL}(i;\theta) = -1 \cdot \log \frac{\exp(x_i \cdot A(x_i)/\tau)}{\sum\limits_{x_k \in Z^{SSCL}, i \neq k} \exp(x_i \cdot x_k/\tau)}.$$

Orthogonality Loss In this section, we further expand on the orthogonality loss in our framework. We employ the orthogonality loss as an implicit geometric constraint on the set O predicated on the current batch. O contains the following: mean features for all classes within batch μ_j, assigned targets not represented within batch and all unassigned targets T_u^i.

Formally, let us assume the session i with data D^i and classes C^i. In order to define a set O we compute some preliminaries. Firstly, for every training batch B we compute the within-batch mean for all data features. This is computed as:

$$\mu_j = \frac{1}{|C^j|} \sum_{k=0}^{|C^j|} z_k, \forall j \in C_B^i \tag{11.3}$$

where $C_B^i \in C^i$ refers to all classes appearing in this particular batch. The combined set of all means can then be termed M_B. For the classes that did not appear in this batch we define as $\neg C_B^i = C^i \setminus C_B^i$. Subsequently we define a mapping function from seen class labels to the assigned pseudo target.

$$h : C^i \to T^i \tag{11.4}$$

We incorporate the remaining real data by adding the following set of assigned pseudo targets as $\neg T_B^i = h(\neg C_B^i)$. For completeness, we combine the above with the unassigned targets T_u^i leading to the following definition of O:

$$O = \{M_B \cup \neg T_B^i \cup T_u^i \mid B\} \tag{11.5}$$

Finally, the orthogonality loss takes the form:

$$\mathcal{L}_{ORTH}(O) = \frac{1}{|O|} \sum_{i=1}^{|O|} \log \sum_{j=1}^{|O|} e^{o_i \cdot o_j / \tau_o}, o_i, o_j \in O \tag{11.6}$$

In essence, the orthogonality loss introduces a subtle geometric constraint between real class features and the pseudo-targets. Additionally the batch-wise construction helps regularise the loss function.

Although the improvement from the Orthogonality loss are not prominent like in PSCL, it remains measurable and consistent across settings (e.g. in table 11.2 line 1 vs line 2, line 3 vs line 4), and thus contributes to our results. Additionally, we would like to highlight that each loss term in equation 11.10 incorporates orthogonality constraints either implicitly or explicitly. E.g. the orthogonality enforced on the pseudo-targets implies an implicit orthogonality among the features as they are pushed to these targets.

We consider SCL, SSCL and the orthogonality losses as the cornerstone guiding our work due to their discriminative nature, robustness, and extendability.

11.2.2 OrCo Framework

Overview. Our OrCo framework (see figure 11.2) for FSCIL begins with a pretraining of the model in the first phase, focusing on learning representations, which are transferable to the new tasks. To achieve this, we leverage both supervised and self-supervised contrastive losses [CFN$^+$22, ICP$^+$21]. Before the second phase, we generate a set of mutually orthogonal vectors, which we term as pseudo-targets. In the subsequent second phase, referred to as base alignment in figure 11.2, we allocate pseudo-targets to class means and ensure alignment through our OrCo loss, using abundant base data D^0. The third phase, implemented in each subsequent incremental session, similarly focuses on assigning incoming but few-shot data to unassigned pseudo-targets, followed by alignment through our OrCo loss. Our OrCo loss comprises three key components: cross-entropy, orthogonality loss, and our novel perturbed supervised contrastive loss (PSCL). The cross-entropy loss aligns incremental data with assigned fixed orthogonal pseudo-targets, the orthogonality loss enforces a geometric constraint on the entire feature space to mimic the pseudo-targets distribution, and our PSCL enhances crucial robustness for FSCIL tasks through margin maximization and space reservation, leveraging mutual orthogonality of pseudo-targets.

Phase 1: Pretrain. In the first pretraining phase, we learn an encoder that accumulates knowledge and generates distinctive features. Using a combination of supervised contrastive loss (SCL) and self-supervised contrastive loss (SSCL), we enhance feature separation within classes, improving model transferability to incremental sessions [CFN$^+$22, ICP$^+$21]. To this end, we train the model encoder f and MLP projection head g using base data D^0, mapping input images to $\mathcal{R}^d$ feature space. The pretraining loss is then defined as:

$$\mathcal{L}_{pretrain}(D^0; f, g) = (1 - \alpha) * \mathcal{L}_{SCL} + \alpha * \mathcal{L}_{SSCL}, \tag{11.7}$$

where α controls the contribution of each contrastive loss.

Pseudo-targets. During the first phase, we do not employ any explicit class vectors that can be used for linear classification. Therefore, we generate data-independent mutually orthogonal

pseudo-targets $T = \{t_j\} \in \mathcal{R}^d$ on the hypersphere by optimizing loss shown in equation 11.1, where $|T| >= C$. Further, these pseudo-targets are fixed and assigned to classes, which, in turn, maximize margins between the classes and improve generalization.

Phase 2: Base Alignment. In addition to the pretraining phase, we introduce the second phase based on the base dataset D^0. This phase initiates alignment between the projection head g and the set of generated pseudo-targets T. More specifically, we create class means by averaging features with the same labels. Then, we employ a one-to-one matching approach, utilizing the Hungarian algorithm [Kuh55], to assign class means with the most fitting set of pseudo-targets T^0, where $|T^0| = |C^0|$. Note that each class $y_j \in C^0$ is then associated with the respective pseudo-target $t_j^0 \in T^0$ and we denote the remaining unassigned pseudo-targets as $T_u^0 = T \setminus T^0$. Further, we use pseudo-targets T^0 as base class representations for classification. Despite the optimal initial assignment, we enhance the alignment of the projection head g and the respective pseudo-targets T^0 through the optimization of our OrCo loss. Further insights into the specifics and motivation behind our OrCo loss, we elaborate on in section 11.2.3.

Phase 3: Few-Shot Alignment. In the third phase of our framework, applied in each subsequent incremental session, our goal remains to align incoming data with the pseudo-targets. The following sessions introduce few shot incremental data D^i for the i-th session. Building on previous methods [CMJG+18, WCW+19, KKS21a, CRF+21], we maintain some exemplars from previously seen classes, constituting a joint set $D_{joint}^i = \{D^i \cup \{\bigcup_{j=0}^{i-1} \tilde{D}^j\}\}$, where $\tilde{D}^j$ denotes saved exemplars from earlier sessions. Keeping random exemplars from previous sessions serves to mitigate both overfitting and catastrophic forgetting issues. Similarly to the second phase, we assign pseudo-targets to incremental class means. To be specific, we determine the optimal assignment between T_u^{i-1} and the current class means, resulting in the optimal assignment set of pseudo-targets T^i. Respectively, the unassigned set of pseudo-targets becomes $T_u^i = T_u^{i-1} \setminus T^i$. During incremental session i, we optimize our OrCo loss given data D_{joint}^i, pseudo-targets $T = \{T^i\}_0^i \cup T_u^i$ and the assignment between $T = \{T^i\}_0^i$ and the respective classes.

11.2.3 OrCo Loss

During the second and the third phases, we optimize the parameters of the projection head g with our OrCo loss. This loss comprises three integral components: our novel perturbed supervised contrastive loss (PSCL), cross-entropy loss (CE) and orthogonality loss (ORTH), see figure 11.3. The aim of optimizing the OrCo loss is to align classes with their assigned pseudo-targets, simultaneously maximizing the margins between classes. This, in turn, enhances overall generalization performance.

To maximize the margins in the representation space, we introduce uniform perturbations of the pseudo-targets T, resulting in perturbed pseudo-targets $\tilde{T} = \{\tilde{t}_j\}_0^{|T|}$ defined as

$$\tilde{t}_j = t_j + \mathcal{U}(-\lambda, \lambda), \tag{11.8}$$

where $\mathcal{U}$ stands for uniform distribution and λ defines sampling boundaries. To utilize the introduced perturbations, we redefine positive P_j and negative N_j sets for the contrastive loss for the anchor z_j in equation 11.7. Note that during incremental session i the positive set P_j^i in the standard SCL contains all $z_k \in D_{joint}^i$ such that the label y_k is equal to anchor label y_j, e.g. in figure 11.3, all yellow circles belong to the positive set for yellow anchor z_j. And the negative set consists of remaining samples $N_j^i = D_{joint}^i \setminus P_j^i$, in figure 11.3, the negative set is composed of all other colors.

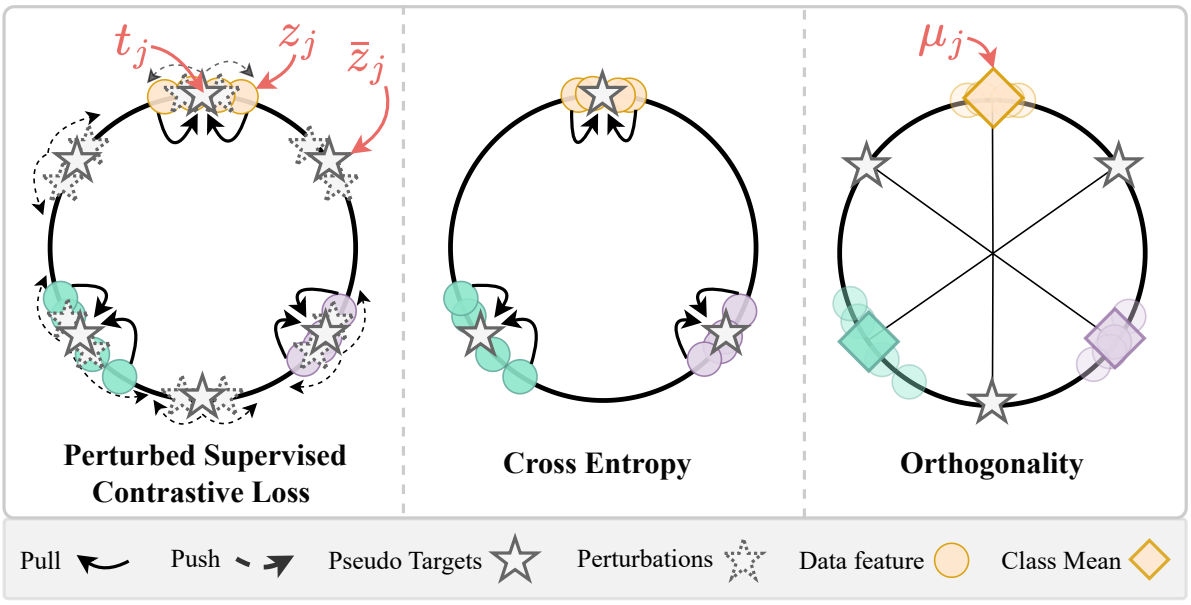


Figure 11.3: **OrCo loss** consists of three components: our proposed perturbed supervised contrastive loss (PSCL), cross-entropy loss (CE), and orthogonality loss (ORTH). z_j denotes the real data anchor point for a contrastive loss, $\bar{z}_j$ denotes the proposed unassigned pseudo-target anchor, and t_j denotes an additional positive sample for yellow class in the form of an assigned pseudo-target.

To adapt standard SCL to PSCL, we expand the definition of the positive set. The anchor z_j, with its assigned pseudo-target $t_j \in T$, becomes an additional positive pair, see figure 11.3. Furthermore, considering the previously defined pseudo-target perturbations, we incorporate them into the positive set, resulting in $\tilde{P}_j^i = P_j^i \bigcup t_j \bigcup \tilde{t}_j$. In figure 11.3, the positive set for yellow anchor z_j contains all yellow circles and additionally the pseudo-target t_j with its perturbations. This extension of the positive set introduces additional pushing forces for incremental classes and, therefore, enables the maximization of margins between classes. We show that this approach proves to be especially advantageous in scenarios with limited samples, as it mimics augmentations in the feature space.

On the other hand, we expand the anchor definition. In standard SCL, each anchor z_j belongs to the set of real training data D_{joint}^i, e.g. in figure 11.3, anchors for standard SCL are only circles. However, we propose to use anchors from both real data and unassigned pseudo-targets (circles and unassigned stars in figure 11.3), specifically $z_j \in D_{joint}^i$ and $\bar{z}_j \in T_u^i$. The positive set for the anchor $\bar{z}_j \in T_u^i$ (unassigned pseudo-target) contains only corresponding perturbed pseudo-targets $\tilde{P}_j^i = \{\tilde{t}_j\}$ (dashed stars around $\bar{z}_j$ in figure 11.3), while the negative set $\tilde{N}_j^i = \{D_{joint}^i \bigcup \tilde{T}_u^i\} \setminus \{\tilde{t}_j\}$ includes all real data and other perturbed pseudo-targets. This approach ensures that each unassigned pseudo-target pushes all other classes away, thereby promoting space preservation for the following incremental sessions.

To complement PSCL, we employ cross-entropy loss for sample z_j that pulls class features to their assigned targets during the few-shot incremental session i:

$$\mathcal{L}_{CE}(z_j) = -\sum_{c \in \{C^I\}_1^i} y_c \log \frac{\exp(z_j t_c^T)}{\sum_{k \in \{C^I\}_1^i} \exp(z_k t_c^T)}. \tag{11.9}$$

Finally, the orthogonality loss as defined in equation 11.1 applies a geometric constraint to class features such as to mimic the ideal target orthogonality. Our OrCo loss is a combination of the three losses introduced above:

$$\mathcal{L}_{OrCo} = \mathcal{L}_{PSCL} + \mathcal{L}_{CE} + \mathcal{L}_{ORTH}. \tag{11.10}$$

To test our representation space, we assign a label based on the nearest assigned pseudo-target.

11.3 Experiments

In section 11.3.1, dataset and evaluation protocol are introduced. Then we compare with state-of-the-art methods on 3 popular benchmarks in section 11.3.2. Finally, we validate the effectiveness of each of the components in section 11.3.3.

11.3.1 Datasets and Evaluation

We conduct evaluation of our OrCo framework on three FSCIL benchmark datasets: mini-ImageNet [RDS+15], CIFAR100 [KH09] and CUB200 [WBW+11]. In the setting formalised by [THC+20] mini-ImageNet and CIFAR100 are organized into 60 base classes and 40 incremental classes structured in a 5-way, 5-shot FSCIL scenario for a total of 8 sessions. CUB200, a dataset for fine-grained bird species classification, contains equal number of base and incremental classes for a total of 200 classes. The dataset is organised as a 10-way, 5-shot FSCIL task and presents a rigorous challenge.

Performance measure Commonly used FSCIL datasets all have a quantity bias towards the base classes. mini-ImageNet and CIFAR100 have both 60% of the data in the base classes and CUB200 with 50%. Consequently, standard accuracy measures like Top-1 accuracy will be skewed in favour of the base-classes. For instance, a method which has a base accuracy $A_{base} = 100\%$ on CUB200 and performs weakly on the first incremental session $A^1_{inc} = 10\%$ would produce a Top-1 average accuracy $A^1_{cls} = 91.82\%$. At first glance, this accuracy may not entirely represent inherent biases in a method, though such measures are commonly used to benchmark performance. To tackle this, harmonic mean has risen as a robust evaluation measure in FSCIL [KKS21a, ZWY+22, ZLX+23]. In the given scenario, the harmonic mean would penalise the method aggressively resulting in a metric score of $A^1_{hm} = 18.18\%$, accurately indicating bias. More concretely, we compute harmonic mean by combining base class accuracy and incremental session accuracy: $A^j_{hm} = (2 \times A_{base} \times A^j_{inc})/(A_{base} + A^j_{inc})$. In addition to this, we propose average harmonic mean (aHM) which is simply averaging the harmonic mean scores from all sessions for a consolidated view.

Implementation Details Our model is optimised using LARS [YGG17] for the pretraining phase and SGD with momentum for phase 2 and 3. For CUB200 dataset, we skip the pretraining following [YYL+23, ZSL+21, THC+20] and initialize the model with ImageNet pretrained weights. For the second and third phase, we finetune only the projection head. For the PSCL loss we choose a perturbation magnitude $\lambda =$ 1e-2. We train the projection head for 10 epochs during the second phase and 100 epochs for the third phase. Cosine scheduling is employed with a maximum learning rate set to 0.1. Augmentations include, random crop, random horizontal flip, random grayscale and a random application of color jitter.

Method	Base Acc	Session-wise Harmonic Mean (%) ↑								aHM	ΔaHM
		1	2	3	4	5	6	7	8		
IW [QBL18]	83.10	49.49	45.09	45.98	46.30	44.67	42.48	43.26	45.65	45.36	**+12.76**
FACT [ZWY+22]	75.78	27.20	27.84	27.94	25.17	22.46	20.54	20.88	21.25	24.16	**+33.96**
CEC [ZSL+21]	72.17	31.91	31.84	30.98	30.74	28.14	26.78	26.96	27.42	29.35	**+28.78**
C-FSCIL [her]	76.60	9.74	20.53	28.68	31.91	34.85	35.05	37.72	37.92	29.55	**+28.57**
LIMIT [ZYM+22]	73.27	40.34	33.58	31.81	31.74	29.32	29.11	29.57	30.28	31.97	**+26.15**
LCwoF [KKS21a]	64.45	41.24	38.96	39.08	38.67	36.75	35.47	34.71	35.02	37.49	**+20.63**
BiDist [ZLX+23]	74.67	42.42	43.86	43.87	40.34	38.97	38.01	36.85	38.47	40.35	**+17.77**
NC-FSCIL [YYL+23]	**84.37**	62.34	61.04	55.93	53.13	49.68	47.08	46.22	45.57	52.62	**+5.50**
OrCo	83.30	**68.71**	**63.87**	**60.94**	**57.98**	**55.27**	**52.41**	**52.68**	**53.12**	**58.12**	

Table 11.1: **Sota comparison on mini-ImageNet.** aHM denotes the average of the harmonic mean across all sessions. IW [QBL18] is evaluated based on the model learning in our pretrain phase. A detailed breakdown of the results into base and individual incremental sessions is in the supplement.

PSCL	CE	ORTH	**Session-wise Harmonic Mean (%) ↑**								aHM
			1	2	3	4	5	6	7	8	
	✓		65.46	56.29	44.12	36.96	26.90	21.11	18.90	16.19	35.74
	✓	✓	65.30	56.21	43.96	37.30	28.31	22.01	19.66	16.64	36.17
✓			50.70	45.42	42.68	39.84	38.71	37.94	36.26	35.87	40.93
✓		✓	52.34	47.24	43.79	41.62	41.15	39.68	38.69	37.34	42.73
✓	✓		68.04	63.94	60.22	**58.00**	**55.44**	51.51	51.88	52.74	57.72
✓	✓	✓	**68.71**	**63.87**	**60.94**	57.98	55.27	**52.41**	**52.68**	**53.12**	**58.12**

Table 11.2: **Influence of UniCon loss components.** PSCL denotes perturbed supervised contrastive loss, CE denotes cross-entropy, ORTH denotes orthogonality loss. See figure 11.3 for visualization of each component. Ablation study on mini-ImageNet.

Sampling	aHM
Rand	57.23
Orth	**58.12**

Table 11.3: **Importance of explicit orthogonality loss for pseudo-target generation.** Rand denotes random sampling from normal distribution.

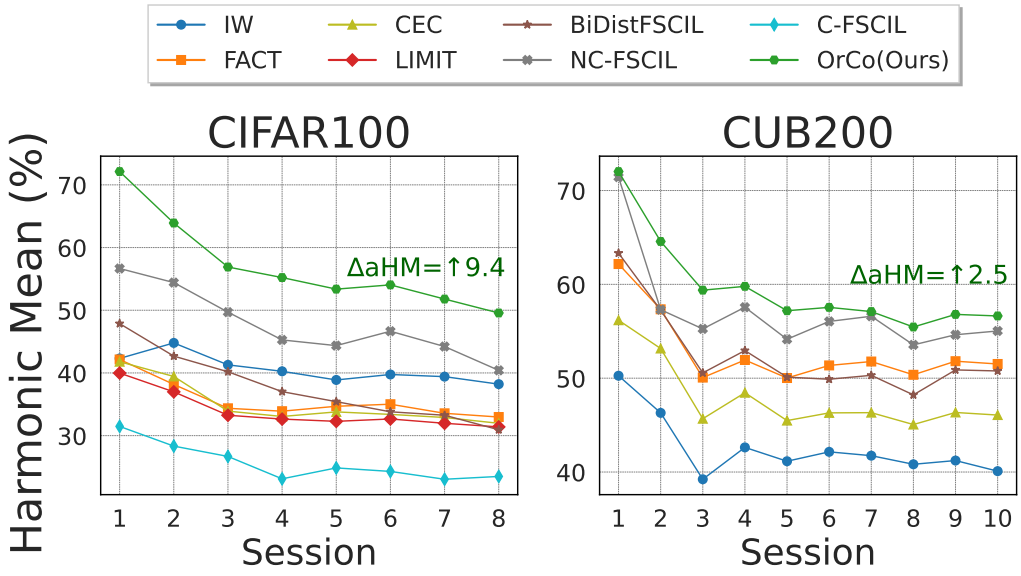


Figure 11.4: **Sota comparisons on CIFAR100 and CUB200 datasets.** Performance curves, that measure harmonic mean, of our method comparing to recent sota methods. Left: CIFAR100. Right: CUB200. ΔaHM denotes the average harmonic mean improvement over the runner-up method.

11.3.2 Comparison to state-of-the-art

In this section, we conduct a comparative analysis of our proposed OrCo with recent state-of-the-art approaches. Table 11.1 presents the results obtained on the mini-ImageNet dataset, while figure 11.4 illustrates the evaluation results on the CUB200 and CIFAR100 datasets. Our method demonstrates superior performance across all three datasets, surpassing previous state-of-the-art methods by a significant margin, particularly achieving improvements of 9.4% and 5.5% on CIFAR100 and mini-ImageNet, respectively. Notably, the effectiveness of OrCo is consistently evident across all incremental sessions.

In addition to reporting results for the standard FSCIL methods, we also present the performance of the Imprinted Weights method (IW) [QBL18] based on our model pretrained during Phase 1. The robust performance of this method indicates the efficacy of our pretraining strategy in facilitating effective transferability to downstream tasks, such as incremental few-shot learning sessions, thereby addressing the intransigence problem.

11.3.3 Analysis

To validate the effectiveness of each component of our framework, in this section, we show an analysis based on the mini-ImageNet dataset.

OrCo loss. We assess the efficacy of the components comprising our OrCo loss in table 11.2. The OrCo loss consists of three integral components illustrated in figure 11.3: the cross-entropy loss (CE), the orthogonality loss (Orth), and the perturbed supervised contrastive loss (PSCL). We observe that CE struggles to generalize on underrepresented incremental classes. On the contrary, PSCL enhances the robust SCL approach with pseudo target perturbations and provides better class separation. PSCL, on its own, shows steady generalization, with only a 14.83% drop in harmonic mean. CE, however, while starting strong, ultimately becomes biased

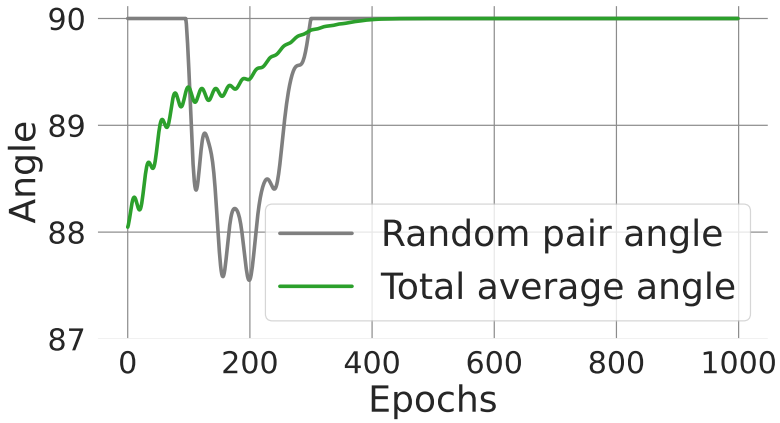


Figure 11.5: **Measurement of angle during orthogonality optimization.** The green curve corresponds to the evolution of the average angle between all pairs during the optimization. The gray curve shows measurements of random pairs at each epoch.

towards base classes, leading to a significant 49.26% drop in harmonic mean. By integrating the dynamic yet fundamentally discerning features of CE with the stability offered by PSCL, a significant enhancement in harmonic mean is achieved. Moreover, the orthogonality loss (ORTH) plays a crucial role in consistently improving performance (+0.4%). Its integration into our loss formulation further underscores the significance of orthogonality.

Influence of mutually orthogonal pseudo-targets. We note that independent randomly sampled vectors from a Gaussian distribution $\mathcal{N}(0,1)$ are theoretically orthogonal on the surface of a unit sphere. However, in practice, we observe only a near-orthogonal behavior, as illustrated in our training curve for orthogonal pseudo-target generation in figure 11.5. To examine the impact of a perfectly orthogonal target space, we assess our method using both randomly sampled targets from a Gaussian distribution and our generated orthogonal targets, as presented in table 11.3. We observe a 0.89% improvement in performance when explicit orthogonality constraints are applied. This finding suggests that an aligned and orthogonal feature space is more effective in addressing data imbalances between base and incremental sessions. Consequently, we incorporate orthogonality as a fundamental principle in our framework, recognizing its significant role in enhancing the overall effectiveness of our model.

Pseudo-targets perturbations. OrCo relies on perturbations of fixed pseudo-targets to introduce a margin between previously encountered and incoming classes. We compare OrCo against a variant where the contrastive loss does not receive any pseudo-targets' perturbations (w/o). In contrast to this, our perturbation schemes with sampling λ, as in equation 11.8, from Gaussian ($\mathcal{N}$) and uniform ($\mathcal{U}$) distributions consistently enhance the final session harmonic mean (HM_8) by over 30%, as shown in table 11.4.

For a detailed evaluation of our method, we employ false positive and cosine similarity analyses. By measuring the false positive rate within only incremental classes (FP_{inc}), we observe improved separation between the few-shot classes with the perturbed objective.

Subsequently, we calculate the average inter-class cosine similarity (Sim_{cls}) for all base and few-shot incremental classes, providing an indication of the spread of each class on the unit sphere. A lower value suggests more compact representations. Notably, we observe values at least 10 times lower for the training with perturbations. Lastly, we assess the availability of space around unassigned pseudo-targets ($Sim_{cls \to target}$) by computing the average similarity of all

Perturbation	$FP_{inc} \downarrow$	$Sim_{cls} \downarrow$	$Sim_{cls \to target} \downarrow$	$HM_8 \uparrow$
w/o	66.5	0.105	0.013	20.14
$\mathcal{N}$	54.6	**0.002**	**0.006**	50.23
$\mathcal{U}$	**52.5**	0.011	**0.006**	**53.12**

Table 11.4: **Influence of perturbations in PSCL.** Comparison of our PSCL loss with or without perturbations of pseudo-targets. $\mathcal{N}, \mathcal{U}$ denotes Gaussian and Uniform distributions, respectively, from which λ, as in equation 11.8, sampled during training. FP_{inc} refers to the False Positive rate among all incremental classes. Sim_{cls} computes the average pairwise cosine similarity between all class pairs. $Sim_{cls \to target}$ indicates the pairwise cosine similarity between classes and unassigned target pairs over all sessions. HM_8 refers to the 8-th and final session harmonic mean.

Pretrain Strategy	Phase 1:Accuracy	aHM
CE	85.70	55.20
SCL	85.18	57.38
SCL + SSCL (Ours)	**85.95**	**58.12**

Table 11.5: **Influence of pretraining.** aHM denotes average harmonic mean. CE is cross-entropy, SCL is supervised countrastive loss, and SSCL is self-supervised contrastive loss.

data features with respect to all unassigned targets. A higher average similarity corresponds to smaller margins between the features and the unassigned pseudo-targets. Table 11.4 illustrates that perturbations indeed increase the margin around unassigned pseudo-targets.

Influence of pretraining. To evaluate our pretraining strategy, we compare it against cross entropy (CE) and standard supervised contrastive loss (SCL) [KTW+20]. As shown in table 11.5, the addition of self-supervised contrastive loss (SCL+SSCL) to the pretraining session significantly enhances generalization on unseen data, showcasing improved transfer capabilities, which aligns with previous findings [CFN+22, ICP+21]. Additionally, we present the accuracy on the validation set for the base classes D^0 immediately after the pretrain phase for each strategy.

While all strategies exhibit close to 85% accuracy on the base validation set, our approach yields a 0.74% higher average harmonic mean compared to *SCL* and a notable 2.92% improvement over *CE*. The significance of this lies in the fact that our frozen backbone network, maintained during incremental sessions, is capable of producing strong and unique features even for unseen classes.

Frozen parameters. Table 11.6 illustrates how OrCo effectively addresses catastrophic forgetting by adopting a strategy of freezing the backbone and training only the projection head. The observed overall performance decay, along with a 2.7% greater loss of base accuracy across all sessions, demonstrates favorable outcomes for decoupling the learning process after Phase 1.

Fine-tuned params	Performance Decay ↓	Base Decay ↓
f, g	28.94	20.52
g	**26.99**	**17.83**

Table 11.6: **Influence of frozen paramters.** Analysing of catastrophic forgetting when 1) fine-tuning the entire model ($f + g$) and 2) fine-tuning only the projection head (g).

11.4 Conclusion

This chapter introduce the OrCo method to boost the performance of FSCIL by addressing its inherent challenges: catastrophic forgetting, overfitting, and intransigence. The OrCo framework is a novel approach that tackles these issues by leveraging features' mutual orthogonality on the representation hypersphere and contrastive learning. By combining supervised and self-supervised contrastive learning during pretraining, the model captures diverse semantic information crucial for novel classes with limited data, implicitly addressing the intransigence challenge. Employing the proposed OrCo loss during subsequent incremental sessions ensures alignment with the generated fixed pseudo-targets, maximizing margins between classes and preserving space for incremental data. This comprehensive approach not only enhances feature space generalization but also mitigates overfitting and catastrophic forgetting, marking steps toward improving the practical value of incremental learning methods in real-world applications.

Our analysis demonstrates that our method surpasses previously developed methods. However, addressing the bias resulting from the disproportionate availability of data for pretraining and subsequent tasks remains a challenging task that calls for further exploration in future research.

IV

Learning Beyond Supervision

While traditional methods often assume a closed-world scenario, in many real-world applications, the environment is dynamic and constantly changing, presenting new classes, concepts, or data distributions over time. Therefore, in this section, our emphasis shifts towards learning beyond the boundaries of the closed-set data provided during training.

In Chapter 12, we explore a realistic open-set semi-supervised scenario where the objective is twofold: accurately classify closed-set classes and detect outliers that fall outside these classes. Our analysis reveals that classification and detection pose conflicting challenges, hindering the learning of each task. Consequently, we propose decoupling these tasks using linear layers, enabling us to leverage unlabeled data without compromising performance.

In Chapter 13, we further explore the open-world scenario with the assistance of pretrained vision-language models. Specifically, we tackle the challenges of egocentric action recognition and emphasize the difficulty in generalizing due to the lack of pretraining web data for this domain. Therefore, we propose a simple yet effective adaptation method for foundational models to transition to new domains.

12

SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning

Contents

In this chapter, we discuss open-set setting that refers to the setting when recognition of unseen classes is required at the time of evaluation in combination with semi-supervised learning that effectively leverages unlabeled data to improve model generalization. Moreover, we tackle the problem when unlabeled data contain outliers from the novel categories that do not appear in the labeled set that is more challenging scenario than standard open-set setting. this chapter, we explore the open-set setting, where recognizing unseen classes during evaluation is necessary, together with semi-supervised learning methods that effectively utilize unlabeled data to enhance model generalization. Furthermore, we address the challenge of outliers from novel categories present in the unlabeled data, which presents a more complex scenario than the standard open-set setting. However, SSL models often underperform in such open-set scenarios due to closed-set classifiers. In this chapter, we study the challenging and realistic open-set SSL setting, where the goal is to both correctly classify inliers and to detect outliers. Intuitively, the inlier classifier should be trained on inlier data only. However, we find that inlier classification performance can be largely improved by incorporating high-confidence pseudo-labeled data, regardless of whether they are inliers or outliers. Also, we propose to utilize non-linear transformations to separate the features used for inlier classification and outlier detection in the multi-task learning framework, preventing adverse effects between them. Additionally, we introduce pseudo-negative mining, which further boosts outlier detection performance. The three ingredients lead to what we call **S**imple but **S**trong **B**aseline (SSB) for open-set SSL. In experiments, SSB greatly improves both inlier classification and outlier detection performance, outperforming existing methods by a large margin.

This chapter is based on [FKDS23b]. As a second author, Anna Kukleva contributed to the project in the scientific discussions, writing of the paper and creating the figures.

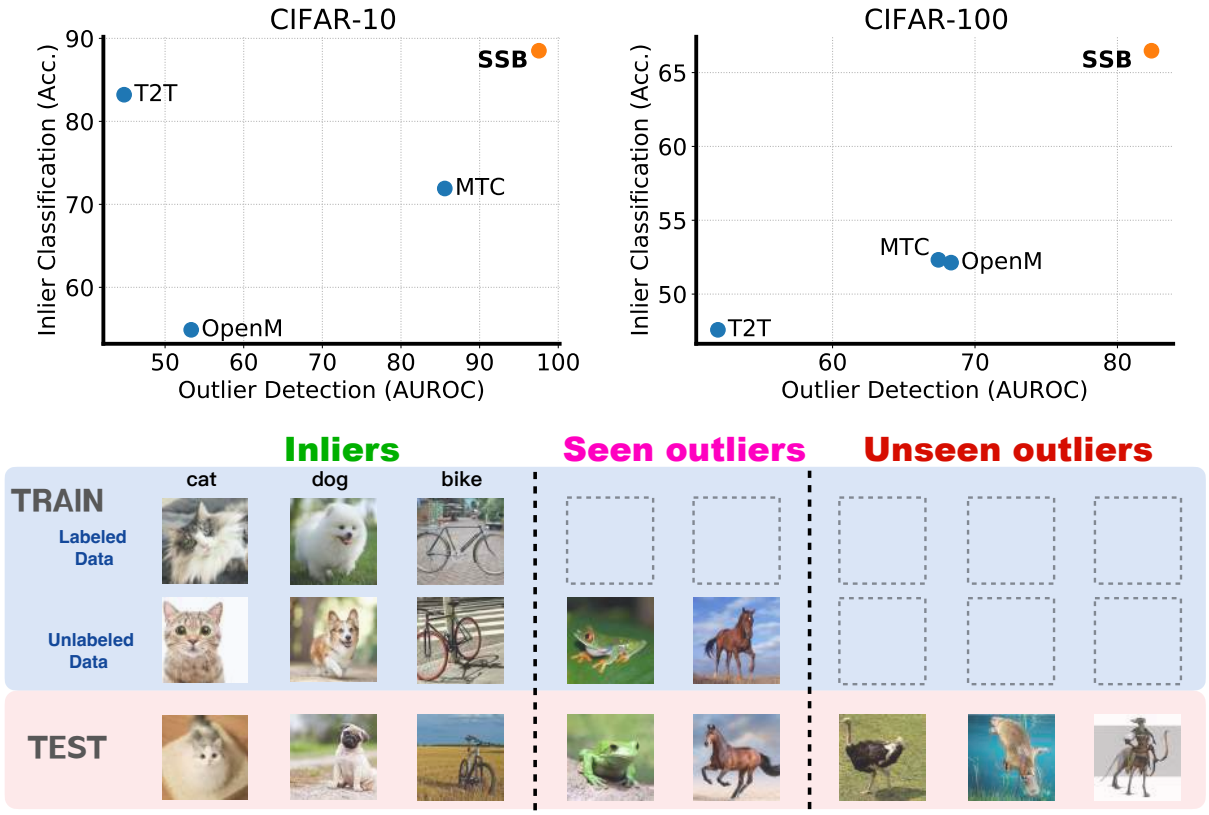


Figure 12.1: Open-set semi-supervised learning considers a realistic and challenging setting, where unlabeled data contains samples from novel classes (**seen outliers**) that do not appear in the labeled data. At test time, the model should correctly classify **inliers**, while identifying outliers seen during the training and, most importantly, **unseen outliers** that do not appear in the training set. We measure test accuracy for the inlier classification performance and AUROC for the outlier detection performance. Our method (SSB) achieves superior performance in both tasks.

12.1 Introduction

Semi-supervised learning (SSL) has achieved great success in improving model performance by leveraging unlabeled data [Lee13, LA17, TV17, MMKI18, BCG+19, BCC+20, SBL+20, XDH+20, FKS21, ZWH+21, WCF+22]. However, standard SSL assumes that the unlabeled samples come from the same set of categories as the labeled samples, which makes them struggle in open-set settings [OOR+18], where unlabeled data contain out-of-distribution (OOD) samples from novel classes that do not appear in the labeled set (see Fig. 12.1). In this chapter, we study this more realistic setting called *open-set semi-supervised learning*, where the goal is to learn both a good closed-set classifier to classify inliers and to detect outliers as shown in Fig. 12.1.

Recent works on open-set SSL [HFC+21, CZLG20, SKS21, YIIA20, GZJ+20, HHLY22, HHYY22, HYG22] have achieved strong performance [WBH19, MR, HA04, AY01] through a multi-task learning framework, which consists of an inlier classifier, an outlier detector, and a shared feature encoder, as shown in Figure 12.2. The outlier detector is trained to filter out OOD data from the unlabeled data so that the classifier is only trained on inliers. However, this framework has two major drawbacks. First, detector-based filtering often removes many

inliers along with OOD data, leading to suboptimal classification performance due to the low utilization ratio of unlabeled data. Second, the inlier classifier which shares the same feature encoder with the outlier detector can have an adverse effect on the detection performance as shown in Table 12.1.

To this end, we contribute a **S**imple but **S**trong **B**aseline, **SSB**, for open-set SSL with three ingredients to address the above issues. (1) In contrast to detector-based filtering aiming to remove OOD data, we propose to incorporate pseudo-labels with high inlier classifier confidence into the training, *irrespective of whether a sample is an inlier or OOD*. This not only effectively improves the unlabeled data utilization ratio but also includes many useful OOD data that can be seen as natural data augmentations of inliers (see Fig. 12.5). (2) Instead of directly sharing features between the classifier and detector, we add non-linear transformations for the task-specific heads and find that this effectively reduces mutual interference between them, resulting in more specialized features and improved performance for both tasks. (3) In addition, we propose pseudo-negative mining to further improve outlier detector training by enhancing the data diversity of OOD data with pseudo-outliers. Despite its simplicity, SSB achieves significant improvements in both inlier classification and OOD detection. As shown in Fig. 12.1, existing methods either struggle in detecting outliers or have difficulties with inlier classification while SSB obtains good performance for both tasks.

12.2 Related Work

In this section, we discuss prior work on the open-set semi-supervised learning. We will not revisit the topic of semi-supervised learning and open-set learning as previously discussed in Chapter 2.

First shown by [OOR+18], standard SSL methods suffer from performance degradation when there are out-of-distribution (OOD) samples in unlabeled data. Since then, various approaches have been proposed to address this challenge [CZLG20, GZJ+20, YIIA20, SKS21, HFC+21, PYJS22, HHLY22, HHYY22, HYG22]. Existing methods seek to alleviate the effect of OOD data by filtering them out in different ways so that the classification model is trained with inliers only. For example, [CZLG20] uses model ensemble [Sch90] to compute soft pseudo-labels and performs filtering with a confidence threshold; [GZJ+20] proposes a bi-level optimization to weaken the loss weights for OOD data; [YIIA20] assigns an OOD score to each unlabeled data and refines it during the training; [SKS21] leverages one-vs-all (OVA) classifiers [SS21] for OOD detection and propose a consistency loss to train them; [HFC+21] proposes a cross-modal matching module to detector outliers. [HYG22] employs adversarial domain adaptation to filter unlabeled data and find recyclable OOD data to improve the performance; [HHLY22] uses energy-discrepancy to identify inliers and outliers. In contrast, we show that if the representations of the inlier classifier and the outlier detector are well-separated, OOD data turns out to be a powerful source to improve the inlier classification without degrading the detection performance. So, instead of filtering OOD data, we use a simple confidence-based pseudo-labeling to incorporate them into the training.

12.3 Method

In this section, we first present the problem setup of open-set semi-supervised learning (SSL). Then, we give an overview of our method SSB in Section 12.3 before presenting details of the three simple yet effective ingredients used in our method in Section 12.3.2, 12.3.3, and 12.3.4.

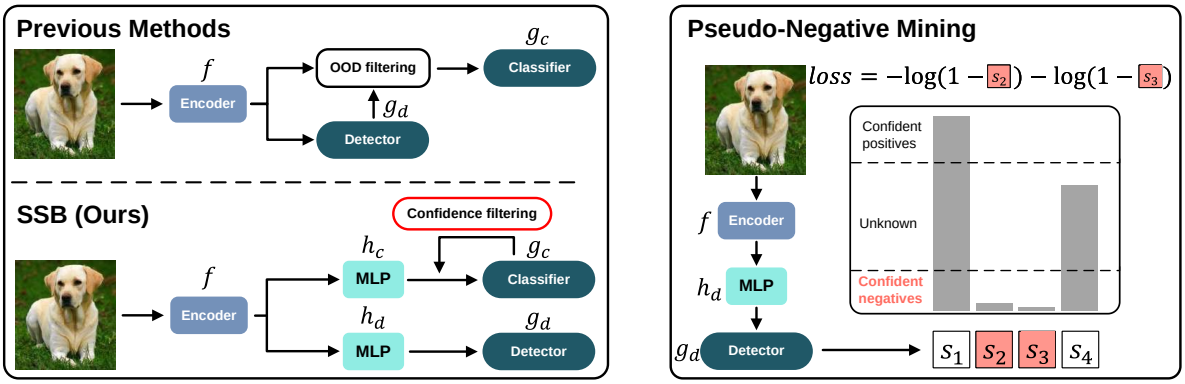


Figure 12.2: **Left:** Our baseline for open-set SSL consists of an inlier classifier g_c, an outlier detector g_d, and a shared feature encoder f whose features are separated from the task-specific heads by two projection heads h_c and h_d. Unlike the detector-based filtering, we adopt confidence-based pseudo-labeling by the inlier classifier to leverage useful OOD data for classifier training. For detector training, we train one-vs-all (OVA) classifiers as in OpenMatch [SKS21]. **Right:** Given the inlier scores (s_1 to s_4), pseudo-negative mining selects confident negatives (s_2 and s_3 in the figure), whose inlier scores are lower than a pre-defined threshold, as pseudo-outliers to help the outlier detector training.

Problem setup and notations: As shown in Fig. 12.1, open-set SSL generalizes the settings of standard SSL and out-of-distribution (OOD) detection. It considers three disjoint sets of classes: $\mathcal{C}$ corresponds to the inlier classes that are partially annotated, $\mathcal{U}_\mathcal{S}$ contains the outlier classes seen during training but without annotations, and lastly, $\mathcal{U}_\mathcal{U}$ is composed of the classes that are not seen during training (only seen at test time). The training data contains a small labeled set $\mathcal{D}_{\text{labeled}} = \{(\mathbf{x}_i^l, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{C}$ and a large unlabeled set $\mathcal{D}_{\text{unlabeled}} = \{(\mathbf{x}_i^u)\}_{i=1}^M \subset \mathcal{X}$, where $\mathcal{X}$ is the input space. While the labeled set only consists of samples of inlier classes, the unlabeled set contains both samples from $\mathcal{C}$ and $\mathcal{U}_\mathcal{S}$. Thus, the the ground-truth label of $\mathbf{x}^u$ is from $\mathcal{C} \cup \mathcal{U}_\mathcal{S}$ with $\mathcal{C} \cap \mathcal{U}_\mathcal{S} = \emptyset$.

The goal of open-set SSL is to train a model that can perform good inlier classification as well as detecting both seen and unseen outliers. Without loss of generality, consider a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times (\mathcal{C} \cup \mathcal{U}_\mathcal{S} \cup \mathcal{U}_\mathcal{U})$, where $\mathcal{C} \cap \mathcal{U}_\mathcal{U} = \emptyset$ and $\mathcal{U}_\mathcal{S} \cap \mathcal{U}_\mathcal{U} = \emptyset$. The learned model should be able to correctly classify inliers $\{(\mathbf{x}_i | y_i \in \mathcal{C})\}$ and detect outliers from $\{(\mathbf{x}_i | y_i \in \mathcal{U}_\mathcal{S})\}$ as well as $\{(\mathbf{x}_i | y_i \in \mathcal{U}_\mathcal{U})\}$, which is crucial for practical applications.

12.3.1 Method Overview

Following [HFC+21, CZLG20, SKS21, YIIA20, GZJ+20, HHLY22, HHYY22, HYG22], we adopt a multi-task learning framework for open-set SSL, which performs inlier classification and outlier detection. As shown in Fig. 12.2, SSB comprises four components: (1) An inlier classifier g_c, (2) an outlier detector g_d, (3) a shared feature encoder f, and (4) importantly, two projection heads h_c and h_d. Inspired by [SKS21], the outlier detector g_d consists of $|\mathcal{C}|$ one-vs-all (OVA) binary classifiers, each of which is trained to distinguish inliers from outliers for each single class. Given a batch of labeled data $\mathbf{X}^l = \{(\mathbf{x}_i^l, y_i)\}_{i=1}^{B_l}$ and unlabeled data $\mathbf{X}^u = \{(\mathbf{x}_i^u)\}_{i=1}^{B_u}$, the total loss for training the model is:

$$L_{total} = L_{cls}(\mathbf{X}^l, \mathbf{X}^u; f, h_c, g_c) + L_{det}(\mathbf{X}^l, \mathbf{X}^u; f, h_d, g_d) \tag{12.1}$$

where L_{cls} and L_{det} are the classification and detection losses, respectively. For the sake of brevity, we will drop the dependencies of the loss function on f, h_c, g_c, h_d, and g_d in the following. The complete algorithm of SSB is summarized by Alg. 1 in Appendix D.

During inference, the test image is first fed to the inlier classifier to compute the class prediction. Then, the corresponding detector is used to decide whether it is an inlier of the predicted class or an outlier. We explain the details of SSB in the following three sections.

12.3.2 Boosting Inlier Classification with Classifier Pseudo-Labeling

Existing methods for open-set SSL [HFC+21, CZLG20, SKS21, YIIA20, GZJ+20] aim to eliminate OOD data from the classifier training. This is typically accomplished by training outlier detectors that can filter out OOD data from unlabeled data, as shown in Fig. 12.2. However, as we will see in Table 12.3, detector-based filtering often removes many inliers along with OOD data, which leads to a low utilization ratio of unlabeled data and hinders inlier classification performance.

In this work, instead of using detector-based filtering, we propose to incorporate unlabeled data with confident pseudo-labels (as generated by the *inlier classifier*) into the training, *irrespective of whether it is inlier or OOD data*. This not only effectively improves the unlabeled data utilization ratio but also includes many useful OOD data as natural data augmentations of inliers into the training (see Fig. 12.5). Inspired by [SBL+20], we train the model with pseudo-labels from the inlier classifier whose confidence scores are above a pre-defined threshold. Specifically, for each unlabeled sample $\mathbf{x}_i^u$, we first predict the pseudo-label distribution as $\hat{p}_i^u = \text{softmax}(h_c(g_c(f(\mathbf{x}_i^u))))$. Then, the confidence score of the pseudo-label is computed as $\max \hat{p}_i^u$. Finally, the cross-entropy loss is calculated for samples whose pseudo-labels have confidence scores greater than a pre-defined threshold τ as:

$$L_{cls}^u(\mathbf{X}^u) = \frac{1}{B_u} \sum_{i=1}^{B_u} \mathbb{1}(\max \hat{p}_i^u \geq \tau) H(\hat{p}_i^u, \hat{y}_i^u) \tag{12.2}$$

where $H(\cdot, \cdot)$ denotes the cross-entropy, $\hat{y}_i^u = \text{argmax} \hat{p}_i^u$, and $\mathbb{1}(\cdot)$ is the indicator function which outputs 1 when the confidence score is above the threshold τ.

The total classification loss is computed as the summation of a labeled data loss and the unlabeled data loss as:

$$L_{cls}(\mathbf{X}^l, \mathbf{X}^u) = L_{cls}^l(\mathbf{X}^l) + L_{cls}^u(\mathbf{X}^u) \tag{12.3}$$

where L_{cls}^l is a standard cross-entropy loss for labeled data.

Despite its simplicity, we obtain a substantial performance improvement in inlier classification through classifier confidence-based pseudo-labeling as shown in Table 12.1. Our method is conceptually different from previous methods as we aim to leverage OOD data rather than remove them. On the one hand, our method effectively improves the unlabeled data utilization ratio as shown in Table 12.3, which leads to great inlier classification performance improvement. On the other hand, our method provides an effective way of leveraging useful OOD data for classifier training. In fact, many OOD data are natural data augmentations of inliers and are beneficial for classification performance if used carefully. As shown in Fig. 12.5, the selected OOD data present large visual similarities with samples of inlier classes, and, thus, significantly enhance the data diversity, leading to improved generalization performance.

12.3.3 Non-Linear Feature Boosting

In previous methods, simply including OOD samples into the classifier training harms detection performance since the inlier classifier and the outlier detector use the same feature representation [SKS21, YIIA20, HFC+21]. On the one hand, the classifier uses OOD data as pseudo-inliers, thus mixing their representations in the feature space. On the other hand, the outlier detector is trained to distinguish inliers and outliers, which leads to separated representations in the feature space. As a result, the contradiction between the classifier and the outlier detector ultimately adversely affects each other, which limits the overall performance, as shown in Table 12.1.

In this work, we find empirically that simply adding non-linear transformations between the task-specific heads and the shared feature encoder can effectively mitigate the adverse effect. Given a sample $\mathbf{x}_i$, two multi-layer perceptron (MLP) projection heads h_c and h_d are used to transform the features from the encoder. The output of the network is thus $h_c(g_c(f(\mathbf{x}_i)))$ for the classifier and $h_d(g_d(f(\mathbf{x}_i)))$ for the outlier detector. Compared to the previous methods, the non-linear transformations effectively prevent mutual interference between the classifier and detector, resulting in more specialized features and improved performance in both tasks. In Table 12.1, while the OOD detection performance degenerates when adding OOD data for classifier training for the model without the projection heads, SSB, in contrast, still exhibits excellent performance in detecting outliers with the help of the projection heads. Moreover, the efficacy of the non-linear projection head also generalizes to other frameworks. We show in the experiment section that it is compatible with various SSL backbones and open-set SSL methods and leads to performance improvement.

12.3.4 Outlier Detection with Pseudo-Negative Mining

In this section, we first describe the outlier detector used in SSB and then introduce a simple yet effective technique called pseudo-negative mining to improve the outlier detector training.

Following [SKS21], we adopt $|\mathcal{C}|$ one-vs-all (OVA) binary classifiers for OOD detection, where each OVA classifier is trained to distinguish between inliers and outliers for each individual inlier class. Given a labeled sample $\mathbf{x}_i^l$ from class y_i, it is regarded as an inlier for class y_i and an outlier for class $k, k \neq y_i$. Therefore, the OVA classifiers can be trained using binary cross-entropy loss on the positive-negative pairs constructed from the labeled set as:

$$L_{det}^{l}(\mathbf{X}^l) = -\frac{1}{B_l}\sum_{i=1}^{B_l} log(p_{y_i}(\mathbf{x}_i^l)) + \frac{1}{K}\sum_{k\neq y_i} log(1 - p_k(\mathbf{x}_i^l)) \tag{12.4}$$

where $p_k(\mathbf{x}_i^l)$ is the inlier score of $\mathbf{x}_i^l$ for class k computed by the k-th detector and $K = |\mathcal{C}| - 1$.

However, due to data scarcity, it is difficult to learn good representations for outliers with labeled data only. To this end, we propose pseudo-negative mining to further improve the outlier detector training by leveraging confident negatives as pseudo-outliers to enhance the data diversity of OOD data. As shown in Fig. 12.2, given an unlabeled sample $\mathbf{x}_i^u$, we consider it as a pseudo-outlier for class k if the inlier score for class k is lower than a pre-defined threshold. Then, $\mathbf{x}_i^u$ is used as a negative sample to calculate the cross-entropy loss of class k. The final loss for $\mathbf{x}_i^u$ is the summation over all classes using it as the negative sample:

$$L_{det}^{u}(\mathbf{x}_i^u) = -\frac{1}{\sum_k \mathbb{1}(p_k < \theta)}\sum_{k=1}^{|\mathcal{C}|} \mathbb{1}(p_k < \theta) log(1 - p_k(\mathbf{x}_i^u)) \tag{12.5}$$

where p_k is the inlier score from the k-th detector and $\mathbb{1}(\cdot)$ is the indicator function which outputs 1 when the confidence score is less than the threshold θ. This increases the data diversity of outliers and improves generalization performance as shown in Table 12.5. Compared to standard pseudo-labels, pseudo-outliers have much higher precision because we specify which classes the sample does not belong to rather than which class it belongs to. The latter is a more difficult task than the former. Therefore, pseudo-negative mining is less susceptible to inaccurate predictions while increasing data utilization.

Our final loss for detector training also includes Open-set Consistency (OC) loss [SKS21] and entropy minimization (EM) [GB05] because they can lead to further improvement. The overall loss for training the detector is as follows:

$$L_{det}(\mathbf{X}^l, \mathbf{X}^u) = L^l_{det}(\mathbf{X}^l) + \lambda^u_{det} L^u_{det}(\mathbf{X}^u) + \lambda^u_{OC} L^u_{OC}(\mathbf{X}^u) + \lambda^u_{em} L^u_{em}(\mathbf{X}^u) \quad (12.6)$$

where λ^u_{det}, λ^u_{OC}, and λ^u_{em} are loss weights; L^u_{OC} is the soft open-set consistency regularization loss, which enhances the smoothness of the OVA classifier with respect to input transformations; L^u_{em} is the entropy minimization loss, which encourages more confident predictions.

12.4 Experiments

In this section, we first compare SSB with existing methods in Section 12.4.1, and then provide an ablation study and further analysis in Section 12.4.2.

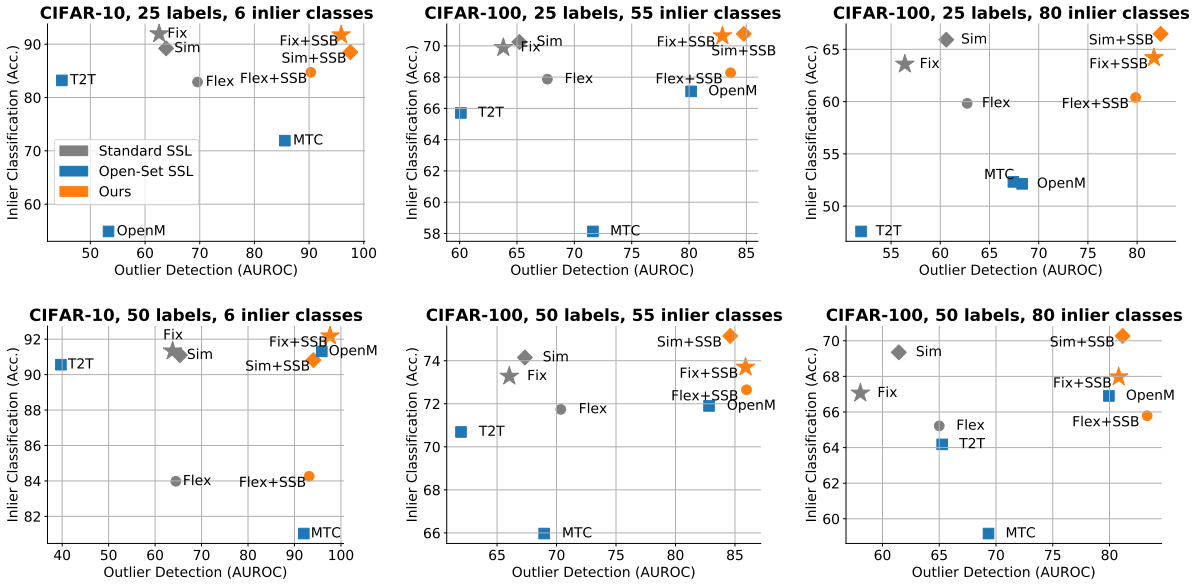


Figure 12.3: **Classification and detection performance on CIFAR-10 and CIFAR-100 with varying numbers of inlier classes and labeled data.** We measure test accuracy for the inliers classification performance and AUROC for the outlier detection performance. While standard SSL methods suffer in outlier detection and open-set SSL methods suffer in inlier classification, SSB achieves good performance in both tasks. Noted that the reported outlier detection performance is the *average AUROC in detecting both seen and unseen outliers*. Please see Appendix A for a detailed breakdown of the results in tables and results on more benchmarks.

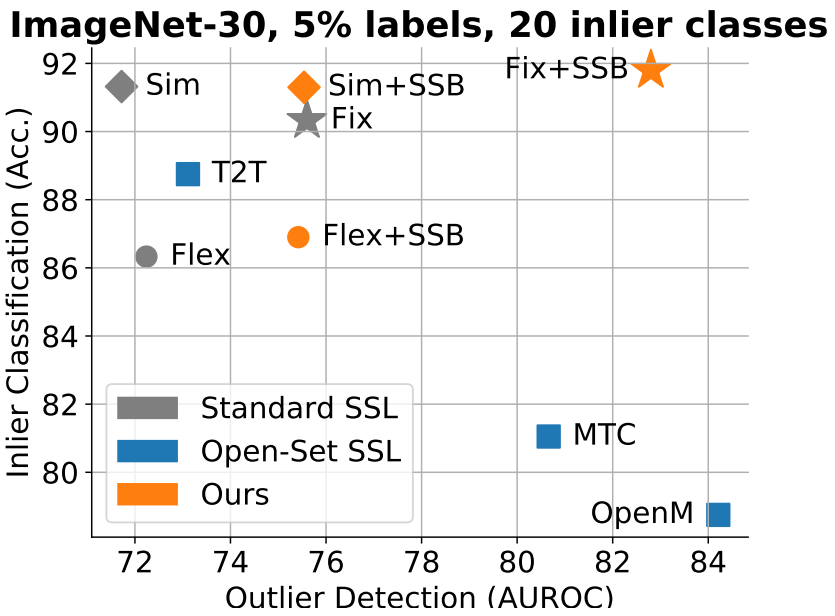


Figure 12.4: **Classification performance versus the outlier detection performance on ImageNet-30.** SSB achieves good performance in both inlier classification and OOD detection. Please see Appendix A for a detailed breakdown of the results in tables.

12.4.1 Main Results

Datasets & Evaluation. As mentioned in Section 13.3, the goal of open-set SSL is to train a good inlier classifier as well as an outlier detector that can identify both seen and unseen outliers. Therefore, we need to construct three class spaces: inlier classes $\mathcal{C}$, seen outlier classes $\mathcal{U}_S$, and unseen outlier classes $\mathcal{U}_U$. For each setting: the labeled set contains samples from $\mathcal{C}$ only; the unlabeled set contains samples from $\mathcal{C}$ and $\mathcal{U}_S$; the test set contains samples from $\mathcal{C}$, $\mathcal{U}_S$, and $\mathcal{U}_U$. The inlier classification performance is evaluated on $\mathcal{C}$ using test accuracy as in standard supervised learning. The OOD detection performance is measured by AUROC following [SKS21] and we report the **average performance in detecting seen outliers and unseen outliers** (see Appendix A for separate AUROC on seen outliers and unseen outliers).

Following [SKS21], we evaluate SSB on CIFAR-10 [KH09], CIFAR-100 [KH09], and ImageNet [DDS+09] with different numbers of labeled data. For CIFAR-10, the 6 animal classes are used as inlier classes, and the rest 4 are used as seen outlier classes during the training. Additionally, test sets from SVHN [NWC+11], CIFAR-100, LSUN [YSZ+15], and ImageNet are considered as unseen outliers, and used to evaluate the detection performance on unseen outliers. For CIFAR-100, the inlier-outlier split is performed on super classes, and two settings are considered: 80 inlier classes (20 outlier classes) and 55 inlier classes (45 outlier classes). Similar to CIFAR-10, test sets from SVHN, CIFAR-10, LSUN, and ImageNet are used to evaluate the detection performance on unseen outliers. For ImageNet, we follow [SKS21] to use ImageNet-30[HG16], which is a subset of ImageNet containing 30 distinctive classes. The first 20 classes are used as inlier classes while the rest 10 are used as outlier classes. Stanford Dogs [KJYFF11], CUB-200 [CZSL20], Flowers102 [NZ08], Caltech-256 [GHP07], Describable Textures Dataset [cim], LSUN are used as unseen outlier classes at test time.

Implementation details. We use Wide ResNet-28-2 [ZK16] as the backbone for CIFAR experiments and ResNet-18 [HZRS16] for ImageNet experiments. As standard SSL models do not

have the notion of OOD detection, we adopt the method in [HG16], where the OOD score of an input image $\mathbf{x}$ is computed as $1 - \max \text{softmax}(f(\mathbf{x}))$ and f denotes the model. Thus, the input image is considered as an outlier if the OOD score is higher than a pre-defined threshold. For other open-set SSL methods, we directly employ the authors' implementations and follow their default hyper-parameters.

For SSB, we use two two-layer MLPs with ReLU [NH10] non-linearity to separate representations for all settings. The hidden dimension is 1024 for CIFAR settings and 4096 for ImageNet settings. For classifier training, we follow [SBL+20] and set the threshold τ as 0.95. For outlier detector training, we set λ^u_{det} as 1 for all settings and follow [SKS21] for the weights of OC loss and entropy minimization. The threshold θ is 0.01 for all experiments (see ablation in Appendix C). Following [SKS21], we train our model for 512 epochs with SGD [KW52] optimizer. The learning rate is set as 0.03 with a cosine decay. The batch size is 64. Additionally, we defer the training of the outlier detector until epoch 475 to reduce the computational cost as we find empirically the deferred training does not comprise the model performance. The ablation on the deferred training is in Appendix C.

When combined with standard SSL methods (e.g. SSB + FlexMatch), we replace the classifier training losses in Equation 1 with the corresponding losses of different methods while keeping the outlier detector the same. When combined with open-set SSL methods (e.g. MTC + SSB), we make three modifications. First, we separate the outlier detector branch from the classifier branch using the proposed MLP projection head. Second, we replace the outlier detector training losses with our loss from Equation 6. Third, we do not filter unlabeled data with the outlier detector for classifier training.

Results. We compare SSB with both standard SSL and open-set SSL methods. Fig. 12.3 and 12.4 summarize the inlier test accuracy and outlier AUROC for CIFAR datasets and ImageNet, respectively. Considering the goal of open-set SSL is to achieve *both good inlier classification accuracy and outlier detection*, SSB greatly outperforms standard SSL methods in outlier detection, and open-set SSL methods in inlier classification. For example, on CIFAR-10 with 25 labels, the AUROC of our best method is 11.97% higher than the best method excluding ours. Moreover, when combined with standard SSL algorithms, our method demonstrates consistent improvement in OOD detection, and in most cases, better test accuracy for inlier classification. This suggests the flexibility of our method, which makes it possible to benefit from the most advanced approaches. Note that the performance improvement of SSB can not be simply explained by the increased number of parameters introduced in the projection heads. Please see Fig. 12.7 for a comparison between SSB and other methods + MLP heads.

Additionally, SSB is more robust to the number of labeled data than others. We achieve reasonable performance given a small number of labeled data while other methods fail to generalize. For example, on CIFAR-10 with 6 inlier classes, OpenMatch has similar inlier accuracy as ours at 50 labels. When the number of labeled data is halved, their performance decreases to 54.88% while our method still has a test accuracy of 91.74%. Please see Appendix B for comparisons on more benchmarks.

12.4.2 Ablation Study

In this section, we analyze the design choices of SSB and show their importance through ablation experiments. If not specified, we use CIFAR-10 with 25 labeled data as our default setting for ablation. The same data split is used for fair comparison.

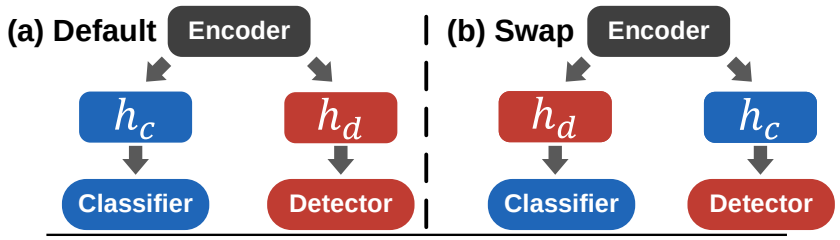
Importance of non-linear projection heads. As mentioned in Section 13.3, we use 2-layer MLPs to mitigate the adverse effect between the inlier classifier and outlier detector. Here we study

the effect of the projection heads in Table 12.1. As we can see, incorporating confidence filtering yields a significant improvement in inlier classification performance (resulting in a 12.23% to 13.18% increase). However, the OOD detection performance experiences a substantial decline when the projection heads are missing (AUROC from 89.67% to 63.46%). This is because the classifier tends to mix the features of inliers and outliers with the same pseudo-labels in a shared feature space, which contradicts the goal of the outlier detector. The addition of the projection heads not only restores the OOD detection performance but also achieves superior results when combined with confidence filtering. Adding the projection heads in combination with confidence filtering not only restores the OOD detection performance but achieves even better performance, which indicates the importance of representation separation. Note that it is important to have two independent projection heads for the inlier classifier and outlier detector. A shared projection head does not restore the OOD detection performance as shown in Table 12.1. Moreover, we show in Table 12.2 that both classification and detection performance degrade when swapping the task-specific features of a pre-trained model with the fixed encoder. In particular, when re-training the detector (just a fully-connected layer) on top of classification features, seen AUROC drops from 89.18% to 53.99%, which suggests our model learns more task-specific features. Therefore, the utilization of the projection heads separates concerns between the classifier and detector, which eases the difficulties of the task and allows them to be trained jointly without adversely affecting each other. The effect of the depth and the width of the projection head is studied in Appendix C.

Proj. head	Conf. filter	Inlier Cls. (Acc.)	Outlier Det. (AUROC)
		78.05	89.67
shared		76.75	91.92
separate		78.47	90.92
	✓	90.28	63.46
shared	✓	90.93	63.87
separate	✓	**91.65**	**94.76**

Table 12.1: **Effect of the projection head and confidence-based pseudo-labeling for classifier training.** We use a 2-layer MLP as the projection head. All models are trained with pseudo-negative mining on the same data split.

Improving data utilization with confidence-based pseudo-labeling. Here we study the effect of different classifier training strategies. We compare three unlabeled data filtering methods for classifier training: (1) *det.* selects pseudo-inliers with the outlier detector as in [SKS21]; (2) *det. (tuned)*, where we choose the selection threshold in detector-based filtering so that the recall of actual inlier samples matches ours; (3) *conf.* uses unlabeled data whose confidence is higher than a pre-defined threshold, which is our method. As shown in Table 12.3, although *det.* successfully removes many OOD data, it also eliminates many inliers, resulting in a low utilization ratio of unlabeled data (0.29% unlabeled data are used in training). In contrast, our method includes pseudo-labels with high classifier confidence into the training, irrespective of whether a sample is out-of-distribution, which leads to a high utilization ratio of unlabeled data (94.22%), thus, outperforming *det.* with a large margin. Moreover, our method also outperforms *det (tuned)* whose data selection threshold is tuned for better performance. This is because we incorporate a significant amount of OOD data in the training process (40.16% v.s. 16.90%). In fact, many OOD data are natural data augmentation of inliers, which can substantially improve



Nearest Neighbor	Inlier Cls. (Acc.)	Outlier Det. (AUROC)
default (a)	**55.04**	**99.43**
swap cls. & det. features (b)	53.70	77.89

Table 12.2: **Classification and detection performance using features of different heads.** We fix the encoder and MLP heads and evaluate the classification and detection performance using nearest neighbors on labeled set. Our model learns specialized features since swapping h_c and h_d leads to inferior performance in both tasks.

closed-set classification if used carefully. When removing pseudo-labeled OOD data using an oracle during the training. The inlier classification accuracy decreases by 3.37% on CIFAR-10 with 25 labels (from 91.65% to 88.28%), which suggests pseudo-labeled OOD data are helpful for inlier classification. In Fig. 12.5, we visualize top-5 confident OOD samples predicted for three inlier classes from *conf.* on CIFAR-100. We can see that the selected samples are related to the inlier classes and contain the corresponding semantics despite being outliers. For example, OOD data selected for *sea* are images with sea background (more examples in Appendix E).

Filter method		det.	det (tuned)	conf. (ours)
Test	Inlier Clf. (Acc.)	47.20	86.53	**91.65**
	Outlier Det. (AUROC)	57.72	87.87	**94.76**
Train	Utilization ratio of:			
	- Unlabeled	0.29	58.09	94.22
	- OOD data	0.04	16.90	40.16
	Prec. of pseudo-inliers	95.17	86.53	58.30
	Recall of inliers	0.47	93.86	92.14

Table 12.3: **Effect of different OOD filtering methods for classifier training.** We compare three filtering methods: *conf.* denotes the confidence-based pseudo-labeling; *det.* uses the outlier detector to select pseudo-inliers for classifier training; *det. (tuned)* is a tuned version of *det.* that matches the recall of inliers with our method. We compare the performance as well as the data utilization ratio, precision, and recall of the inliers from unlabeled data during training. All models are trained with pseudo-negative mining and the projection head on the same data split.

Effect of pseudo-negative mining. Table 12.5 shows the effect of pseudo-negative mining. We compare our pseudo-negative mining with standard pseudo-labeling which predicts artificial labels for unlabeled data and uses confident predictions with labeled data loss. While standard pseudo-labeling does not help the OOD detection performance further, pseudo-negative

Figure 12.5: **OOD samples can be used as data augmentation to improve the generalization performance.** The figure shows three semantic classes from labeled data (wolf, road, and sea), and top-5 confident OOD samples predicted for those classes. The ground-truth semantic class of the OOD sample is on the top of each image. We can see that OOD data with high confidence present large visual similarities to the corresponding semantic classes.

mining improves the seen AUROC by 4.73% over the model without pseudo-negative mining. Compared to standard pseudo-labeling, pseudo-negative mining not only includes more unlabeled data into the training, but also presents high precision for the selected pseudo-outliers as shown in Fig. 12.6.

As mentioned in Section 12.3.4, we utilize unlabeled data with low inlier scores as pseudo-outliers to enhance the data diversity of outlier classes. An unlabeled sample is used as a pseudo-outlier only if its confidence score is less than a pre-defined threshold θ. Table 12.4 compares the results of different thresholds. We can see that our method achieves similar performance as long as θ takes a relatively small value, which suggests the good robustness of our method against this hyper-parameter. We provide more ablation on loss weight and data augmentation in Appendix C.

Threshold θ	Inlier Cls. (Acc.)	Outlier Det. (seen AUROC)
0.2	91.87	92.96
0.1	**92.03**	93.16
0.05	91.97	94.21
0.01	91.65	**94.76**
0.005	91.52	94.75
0.001	91.70	94.15

Table 12.4: **Effect of different thresholds θ for pseudo-negative mining.** Our method shows good robustness against a wide range of thresholds. We use CIFAR-10 with 25 labeled data here.

Ablation on outlier detectors. Here, we compare the performance of different outlier detection methods. Specifically, we choose three schemes from recent works, including the binary

Pseudo-labeling	Inlier Cls. (Acc.)	Outlier Det. (AUROC)
None	91.52	90.03
Standard	91.63	89.69
Pseudo-neg.	**91.65**	**94.76**

Table 12.5: **Effect of pseudo-negative mining for OOD detection.** All models are trained with confidence-based pseudo-labeling and a 2-layer MLP projection head on the same data split.

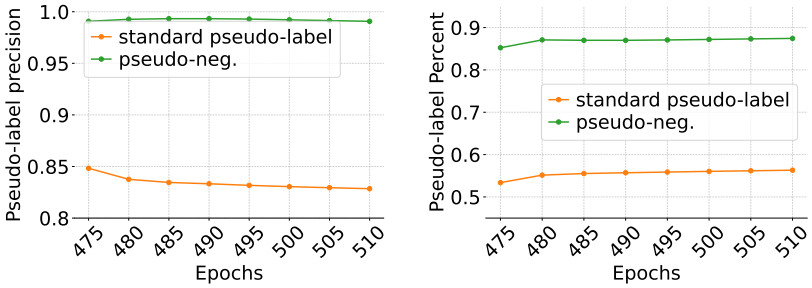


Figure 12.6: Compared to standard pseudo-labeling, pseudo-negative mining has not only higher prediction precision, but also higher data utilization rate.

classifier from MTC [YIIA20], cross-modal matching from T2T [HFC+21], and OVA classifiers from OpenMatch [SKS21]. As shown in Table 12.6. While all methods show reasonable performance, OVA classifiers exhibit the best performance in both inlier classification and OOD detection. Hence, we use OVA classifiers as the outlier detector in our final model.

OOD Detector	Inlier Cls. (Acc.)	Outlier Det. (AUROC)
binary classifier [YIIA20]	70.93	76.12
cross-modal matching [HFC+21]	69.27	75.99
OVA classifiers [SKS21]	**71.00**	**82.62**

Table 12.6: **Comparison between different outlier detectors.** The experiment is conducted on CIFAR-100 with 55 inlier classes and 25 labels per class.

Compatibility with other open-set SSL methods. We evaluated the compatibility of our method with other open-set SSL techniques in Table 12.7. Our results indicate that our method is highly compatible, as all existing methods showed improved performance in both inlier classification and outlier detection when combined with our approach. This demonstrates the flexibility of our method and suggests that it can be easily integrated into existing frameworks as a plug-and-play solution.

Equal-parameter comparison. As mentioned in Section 12.4.1, the performance improvement of SSB can not be simply explained by the increased number of parameters introduced in the projection heads. Here we compare SSB with other methods + MLP heads so that they have the same number of parameters as SSB. As shown in Fig. 12.7, adding MLP heads improves the performance of other methods, but SSB still greatly outperforms all of them, indicating that the performance improvement of our method can not be merely explained by the increase

	Inlier Cls. (Acc.)	Outlier Det. (AUROC)
MTC	60.24	69.88
MTC + Ours	**60.42**	**74.98**
T2T	64.78	52.93
T2T + Ours	**66.98**	**69.50**
OpenMatch	68.53	80.00
OpenM. + Ours	**71.00**	**82.62**

Table 12.7: **Integrating our method with other open-set SSL methods improves performance.** The setting is CIFAR-100 with 55 inlier classes and 25 labels per class.

of the model capacity.

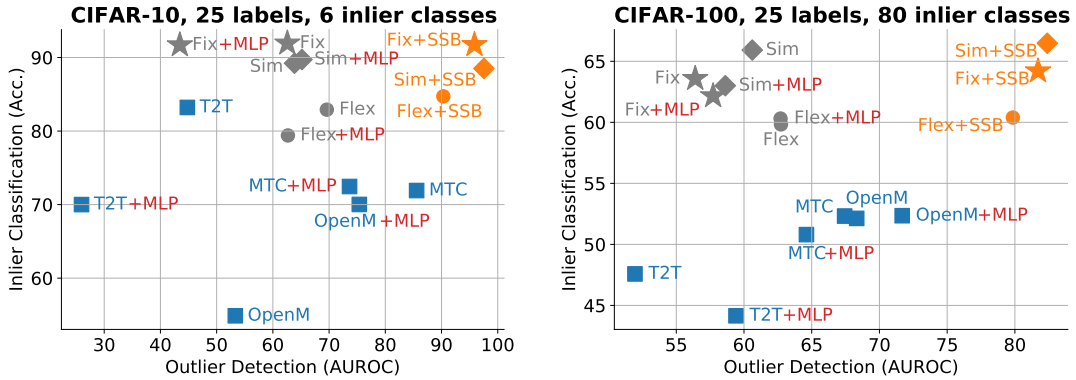


Figure 12.7: **Comparison between SSB and other methods with the same model parameters.** The performance improvement of SSB can not be simply explained by the increased number of parameters.

12.5 Conclusion

In this chapter, we study a realistic and challenging setting, open-set SSL, where unlabeled data contains outliers from categories that do not appear in the labeled data. We first demonstrate that classifier-confidence-based pseudo-labeling can effectively improve the unlabeled data utilization ratio and leverage useful OOD data, which largely improves the classification performance. We find that adding non-linear transformations between the task-specific head and the shared features provides sufficient decoupling of the two heads, which prevents mutual interference and improves performance in both tasks. Additionally, we propose pseudo-negative mining to improve OOD detection. It uses pseudo-outliers to enhance the representation learning of OOD data, which further improves the model's ability to distinguish between inliers and OOD samples. Overall, we achieve state-of-the-art performance on several benchmark datasets, demonstrating the effectiveness of the proposed method.

Nonetheless, SSB has potential limitations. Despite the improved overall performance, the outlier detector suffers from overfitting as the performance gap between detecting seen outliers and unseen outliers is still very large. Therefore, in the future, more regularizations need to be

considered to improve generalization. Another drawback is that our method is not able to deal with long-tail distributions, which is also realistic in practice. Presumably, our method will have difficulty distinguishing inliers of tail classes and OOD data due to the data scarcity at tail.

In the next chapter, Chapter 13, we investigate an unconstrained environment where the objective is not solely to detect and reject unseen classes, but rather to discover and classify novel categories. This scenario, commonly referred to as the open-world scenario, we explore leveraging pretrained vision-language models on egocentric action recognition task.

13 X-MIC: Cross-Modal Instance Conditioning for Egocentric Action Generalization

Contents

In this chapter, we delve into the open-world scenario, leveraging vision-language (VLMs) pretrained models, as there has been growing interest in adapting VLMs to image and third-person video classification due to their success in zero-shot recognition. However, the adaptation of these models to egocentric videos has been largely unexplored. To address this gap, we propose a simple yet effective cross-modal adaptation framework, which we call X-MIC. Using a video adapter, our pipeline learns to align frozen text embeddings to each egocentric video directly in the shared embedding space. Our novel adapter architecture is specifically tailored to egocentric videos, focusing attention on the hand region. This results in an enhanced alignment of text embeddings to the egocentric video domain, leading to a significant improvement in cross-dataset generalization. We evaluate our approach on the Epic-Kitchens, Ego4D, and EGTEA datasets for fine-grained cross-dataset action generalization, demonstrating the effectiveness of our method.

This chapter is based on [KSR+24]. As the first author, Anna Kukleva conducted all experiments and was the main writer.

13.1 Introduction

Egocentric action recognition has recently become a popular research topic due to the rising interest in augmented reality and robotics. Recently, two large-scale egocentric datasets Epic-Kitchens [DDF+22] and Ego4D [GWB+22b], capturing the daily activities of users have been introduced. While there is a growing interest in studying action recognition on egocentric datasets, evaluations primarily occur within the same dataset; lacking cross-dataset evaluations that is crucial for real-world deployment of recognition models. Testing models on different datasets presents several challenges, such as encountering unfamiliar environments, different users, and previously unseen objects and their corresponding actions, all of which can significantly impact performance. Recently, vision-language models [JYX+21, RKH+21, ZJM+22] such as CLIP [RKH+21] have demonstrated remarkable performance across diverse third-persons datasets like Kinetics-600 [KCS+17] and ImageNet [ima], showcasing their ability to

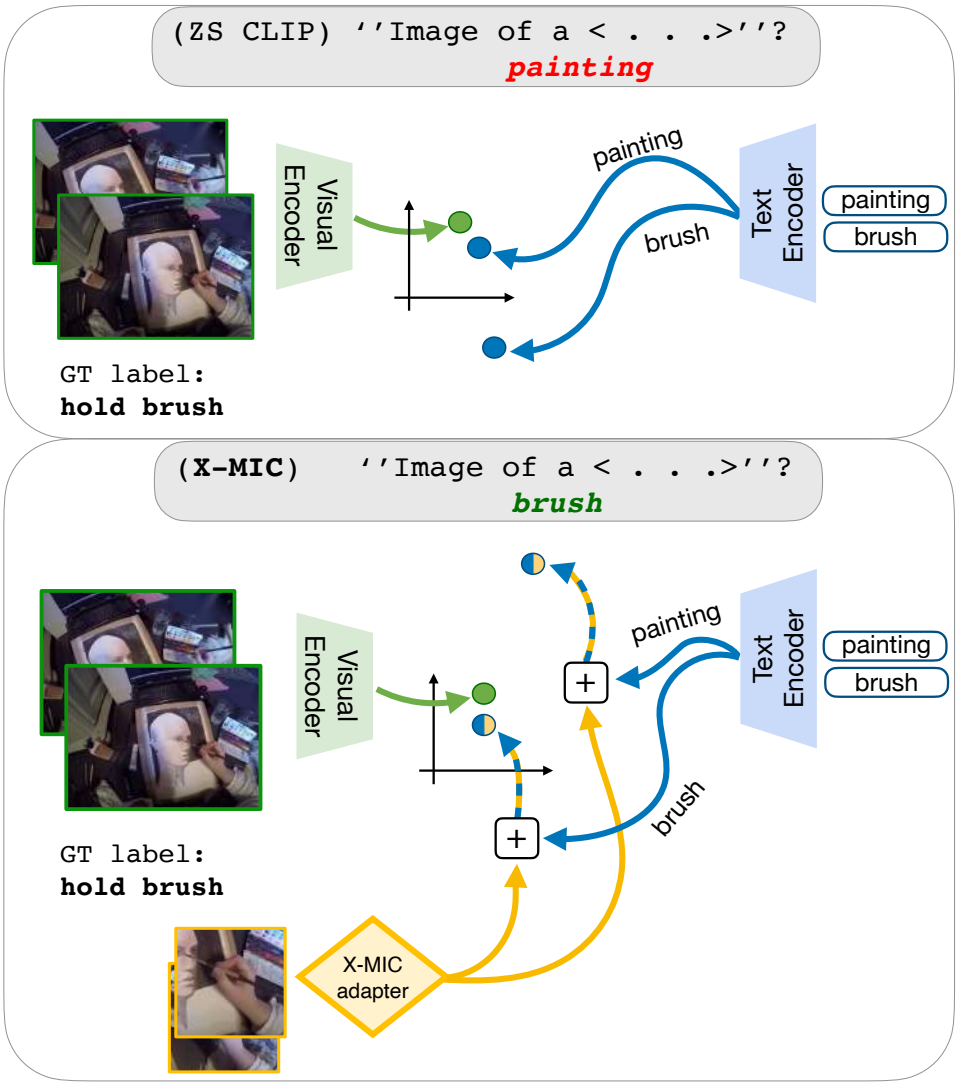


Figure 13.1: **Egocentric video classification with VL models. Top:** Standard zero-shot CLIP. As the dominant object in the scene is painting, the model predicts class "painting" while the object of interest is "brush". **Bottom:** CLIP model with our X-MIC adaptation directly in the shared VL embedding. X-MIC vectors adapt focus of the CLIP model to the hand area, guiding text modality to capture egocentric domain-specific information.

generalize effectively and achieving zero-shot performance of 59.8% and 76.2%, respectively. However, their zero-shot performance drops significantly when applied to egocentric datasets like Epic-Kitchens, with noun and verb recognition reaching only 8.8% and 5.9%, respectively; highlighting the domain gap between third-person and egocentric data.

CLIP's zero-shot generalization to new datasets leverages learning a shared embedding space for text and visual modalities. To enhance generalization to new domains, a prominent research direction [ZYLL22b] explores adapting the text encoder by appending trainable *prompt* tokens to class tokens, modifying the class-text input from "a photo of an apple" to "<learnable prompt> apple". As an alternative approach, recent work has proposed to train feature *adapters* on both the visual and textual domains [CDW+22, GGZ+23], drawing insights from the NLP

works [HGJ+19, SM19]. Despite their promising results, these methods overlook the inherent characteristics of the egocentric video domain. To overcome this, we propose a simple yet effective adapter architecture, injecting egocentric video-specific knowledge into a frozen VL embedding space, depicted in Fig. 13.1. Our method transforms each video through an adapter into a vector for cross-modal instance conditioning of text — referred to as X-MIC-vector. Our cross-modal adaptation directly in the embedding space results in significantly improved efficiency during training & testing. Moreover, a new adapter module disentangles frozen visual encoder from the visual temporal modeling through cross-modal adaptation. Each X-MIC-vector is video-specific, therefore, allowing us to align any frozen text to each input video individually. Finally, to align the text embedding to the video, we simply sum the X-MIC-vector to the text embedding vectors.

We extensively evaluate our approach on Epic-Kitchens [DDF+22], Ego4D [GWB+22b] and EGTEA [LLR18] datasets, demonstrating superior generalization compared to SOTA VL-adaptation methods.
Our contributions can thus be summarized as:

- Addressing the task of egocentric cross-dataset and zero-shot action recognition with VLMs that is designed for real-world applications, e.g. AR, addressing the impracticality of collecting data from every new environment;
- A simple yet effective framework, referred to as X-MIC, for cross-modal adaption of VL models directly in the pre-trained VL embedding space; our module disentangles temporal modeling from the frozen visual encoder;
- A novel egocentric spatial-temporal attention module enhances information around hands, thereby improving egocentric action recognition performance;
- Thorough comparisons with respect to image and video state-of-the-art VL adaptation methods which demonstrate the effectiveness of our approach.

13.2 Related Work

In this section, we discuss prior work on the egocentric action generalization and adaption of vision-language models to video domain. We will not revisit the topic of vision-language large-scale adaptation methods for the image domain as previously discussed in Chapter 2.

Egocentric Action Generalization. While egocentric vision gained attention with datasets like Epic-Kitchens [DDF+22] and Ego4D [GWB+22b], current state-of-the-art [ZK22, XANS22, PCA+21, WLM+22, KNZD19, XLG+20, ZAC+19, SCS+] mainly focus primarily on intra-dataset evaluation, which limits their applicability to real-world scenarios. Several methods fine-tuned CLIP on egocentric datasets, [ZMKG23, LWS+22, PSN+23], yet generalization on fine-grained verbs and nouns recognition remains underexplored. Our work comprehensively investigates both intra-dataset and inter-dataset generalization on both verbs and nouns.

Adapting VLMs to Videos. Recent advancements in prompt learning extend to third-person videos. A5/A6 [JHZ+22] uses textual prompts and a temporal module atop the visual encoder while keeping both encoders frozen. EVL [LGZ+22] discards the text encoder, relying solely on temporally encoded frame features from CLIP image encoder. Vita-CLIP [was23] uses shallow prompts on the text encoder, similar to [ZYLL22b], and introduces deep temporal prompts for the visual encoder. OAP [CSMY24] generalizes the verbs observed during training to an open vocabulary of objects with a prompt-based object encoder. Our method builds on existing work

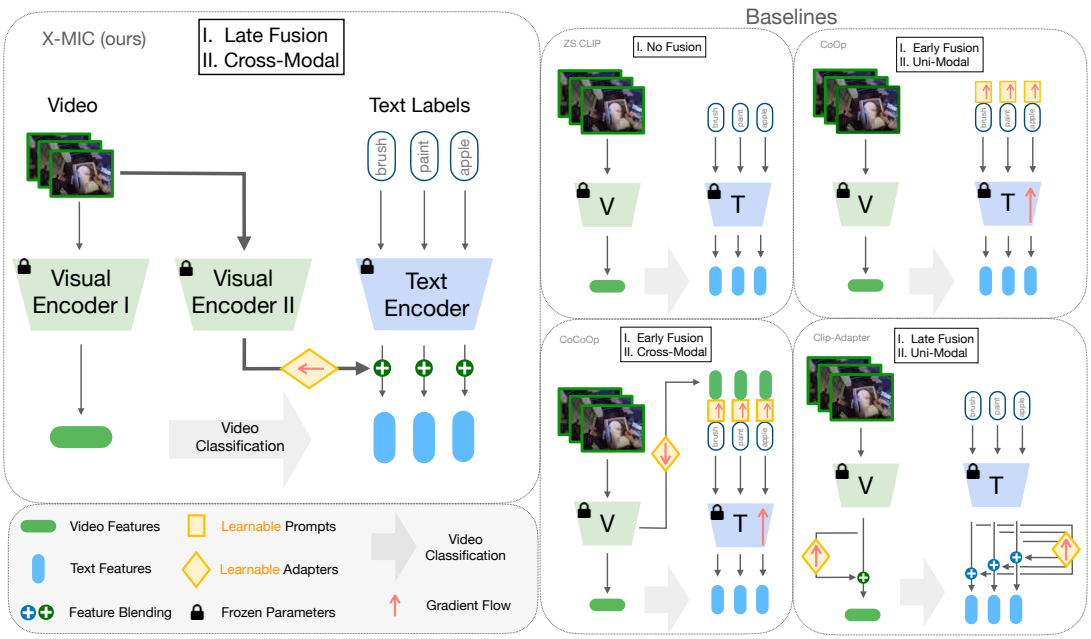


Figure 13.2: **Overview of our X-MIC method and previous adaptation methods of VLMs.** Baselines: **No Fusion** is a standard zero-shot video classification method. The average of the frame representations is compared to text representations in the shared VL embedding space. **Early Fusion & Uni-Modal** is a prompt learning method, where the learnable parameters are concatenated to text tokens and optimized through the text encoder. Subsequently, the text encoder is adapted to the new domain. **Early Fusion & Cross-Modal** is a prompt learning approach, an extension of Early Fusion & Uni-Modal method, where additional learnable parameters are introduced in the form of an adapter. This adapter maps video presentations to embedding space of text tokens, which are then concatenated to learnable prompts and text tokens. Memory consumption, required for forward-backwards pass through the text encoder, expands with respect to all combinations of all text-labels and videos in the batch. **Late Fusion & Uni-Modal** is a method, where adaptation of both encoders is based on the feature blending of original text and video representations with the adapted corresponding representations. Ours: **X-MIC** adaptation method falls in *Late Fusion & Cross-Modal* category. Adapted video features are blended with the original text features. Simple yet efficient adaptation of text modality to each individual video is efficient as it does not require gradient propagation through text or video encoders. Additionally, we propose to employ Visual Encoder II, offering flexibility in utilizing various types of visual features for conditioning purposes.

while introducing an adapter architecture specifically tailored to egocentric domain, resulting in superior performance.

13.3 Method

We begin by introducing the preliminaries such as classification with VLMs like CLIP and different types of VL adaptation in Sec. 13.3.1. Then, in Sec. 13.3.2, we give an overview of our adapter method for text conditioning and present our egocentric-spatio-temporal attention module.

13.3.1 Preliminaries and Baselines on VL Adaptation

Vision-language models (VLMs), such as CLIP, demonstrate effective zero-shot generalization across various downstream tasks for image recognition and third-person video recognition. However, certain domains, like egocentric videos, still face challenges due to a significant gap between web-collected and egocentric data.

Below, we provide an overview existing prompt learning and adapter-based methods.

Video Classification with VL Dual Encoders. Trained on hundreds of millions of text and visual data pairs, VL dual encoders bring the two modalities together in a shared embedding space. When evaluating models pretrained on extensive web data, a crucial metric is their ability to transfer to other downstream tasks without additional fine-tuning, a process commonly known as zero-shot evaluation. To perform zero-shot classification, one needs to propagate a set of C predefined classes in the form of text, denoted as $t =$ "Image of a <class>" through a pre-trained text encoder $T(\cdot)$. This process extracts individual text embeddings, represented as $e_t = T(t) \in \mathcal{R}^{1\times D}$ for each class. Subsequently, these vectors undergo $l2$ normalization, resulting in $\bar{e}_t = \frac{e_t}{||e_t||}$ (hereafter, the overline symbol $\bar{e}$ indicates $l2$ normalization of vector e). Then a matrix $\bar{E}_T \in \mathcal{R}^{C\times D}$ is constructed, representing a simple linear classifier, and is thus referred to as the text-based classifier. To classify an input video v, we sample N frames, denoting the sampled frames as $v' = \{z_i\}^N$, where z_i represents a frame from the video v. Subsequently, all sampled frames are mapped using the frozen visual encoder $V(\cdot)$ to the shared VL embedding. Applying average pooling over the embeddings of the frames yields a single-vector video representation: $\bar{e}_{v'} = avg\_pool(\{\overline{V(z_i)}\}^N) \in \mathcal{R}^{1\times D}$. The video vector $\bar{e}_{v'}$ is then classified using the text-based classifier $\bar{E}_T$.

No Fusion. We refer to frozen dual encoders $T(\cdot)$ and $V(\cdot)$ without additional adaptation as "No Fusion" baseline.

Early Fusion and Uni-Modal Adaptation. A prompt learning-based method, CoOp [ZYLL22b], introduces P learnable vectors appended to all C input text classes in the token embeddings of the textual encoder (see Fig. 13.2). To optimize these prompts, gradients are propagated through the frozen text encoder for $C \times P \times D$ adaptable parameters, where D is the dimensionality of the tokens. This optimization remains independent of the batch size of the visual input.

Early Fusion and Cross-Modal Adaptation. A follow-up work, CoCoOp [ZYLL22a], extends learnable text prompts to cross-modal prompts by introducing an adapter module of the frozen visual encoder to the token embedding space (see Fig. 13.2). In this architecture, each of the C class-tokens are appended not only with P learnable text prompts but also with individual input-conditioned prompts generated by the adapter. Optimizing these prompts for a batch of size B involves propagating $B \times P \times D \times C$ gradients, making training inefficient and slow as shown in [ZYLL22a].

Late Fusion and Uni-Modal Adaptation. CLIP-Adapter [GGZ+23] adopts a late fusion approach as an alternative to early fusion adaptation. The text and visual encoders are followed by uni-modal adapter modules that generate adapted uni-modal feature vectors. These adapted features are then fused with the corresponding original features in the VL embedding space, subsequently optimized with the standard classification loss. This optimization is efficient due to the lightweight nature of adapters, eliminating the need for heavy text-encoder gradient propagation.

13.3.2 X-MIC Adaptation

Overview. We aim at achieving generalization in egocentric action recognition across domains and to novel action classes. Our X-MIC-adaptation framework is designed to improve the alignment between frozen text representations and the egocentric visual domain directly within the VL embedding space. To adapt the text modality to the egocentric domain, we introduce a simple yet effective cross-modal text conditioning operation based on the input videos. Specifically, each X-MIC-vector serves as an adapted video representation. We align any frozen text representation to each individual input video by a simple addition operation with the X-MIC-vector. Consequently, text representations are adapted to individual input videos, and these adapted text embeddings are further utilized for the classification of corresponding videos into fine-grained noun and verb classes. Moreover, by introducing an egocentric-spatio-temporal attention module, we aggregate temporal information between video frames and emphasize areas around hands to enhance hand-object interactions. X-MIC-vectors offer dual benefits: a simple and efficient cross-modal conditioning approach, and the decoupling of domain-specific knowledge from the frozen VL embedding, resulting in improved generalization on egocentric videos.

X-MIC Adaptation. Our adaptation method, X-MIC, aligns frozen text class embeddings directly to the new domain in the shared VL embedding space. During training and inference, our approach resembles zero-shot classification, as we classify frozen video representations from the original visual backbone $V(\cdot)$ using an adapted text-based classifier $\bar{E}_T$ tailored to each input video v. This enables efficient domain adaptation without the need for fine-tuning the entire model, categorizing our method as late fusion with cross-modal adaptation.

Specifically, for an input video v, we sample N frames to form a sparse video sequence $v' = \{z_i\}^N$. The video sequence v' is then decoded using the original visual encoder $V(\cdot)$, resulting in a single vector $\bar{e}_v$. Additionally, we encode the C classes into the text-based classifier $\bar{E}_T$ as detailed in Sec. 13.3.1.

To generate the X-MIC-vector, we introduce a second frozen visual encoder, denoted as $V_{II}(\cdot)$. This secondary encoder can either be an identical copy of the original encoder $V(\cdot)$ or a distinct pre-trained encoder. In Sec. 13.4.1, we demonstrate that incorporating a different type of $V_{II}(\cdot)$ can result in significant generalization improvements. For instance, DINO [ODM+23], which is uni-modal, captures distinct characteristics [PKH+22] of the visual input compared to multi-modal CLIP like models that focus solely on main objects.

We employ the second encoder $V_{II}(\cdot)$ to produce an intermediate representation of frames, denoted as $x_v = \{V_{II}(z_i)\}^N$. Before adapting the intermediate representation, we apply $l2$-normalization to the vector. See Sec. 13.4.4 for a detailed analysis of the impact of this normalization. Our video adapter $A(\cdot)$ incorporates a temporal aggregation module. By feeding these intermediate representations into this module, we obtain the final X-MIC-vector for adaptation, represented as $a_v = A(\bar{x}_v)$.

Finally, to adapt the frozen text-based classifier $\bar{E}_T$ to the video v, we simply sum X-MIC-vector with each class representation in the embedding space: $\overline{e_t + a_v} \in \mathcal{R}^D$, and when combined, these updated vectors form an adapted text-based classifier $\overline{E}_T^{a_v}$. Subsequently, we classify the video representation $\bar{e}_v$ with the adapted text-based classifier $\overline{E}_T^{a_v}$. The process of classification with X-MIC-adaptation can be summarized as follows:

$$c = \text{argmax}_t < \overline{e_t + A(\overline{V_{II}(x_v)})}, \bar{e}_v >, \tag{13.1}$$

where c represents the class with the highest similarity between the adapted text-based classifier

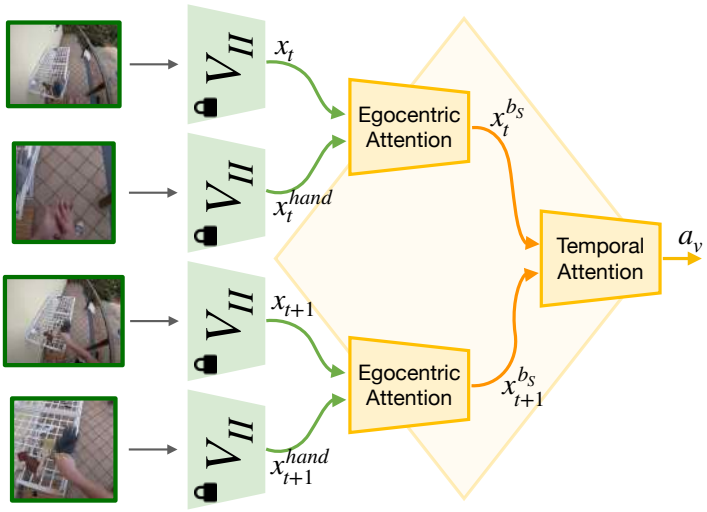


Figure 13.3: **Ego-Spatio-Temporal Attention Module.** Our video-adapter module for X-MIC adaptation of egocentric videos. It takes a sequence of full frames interleaved with hand crops in input, and outputs X-MIC vector a_v, representing video v as a single vector for text conditioning in the shared VL embedding space.

$\overline{E}_T^{a_v}$ and the video v, and $< \cdot, \cdot >$ denotes dot product.

Ego-Spatio-Temporal Attention Module. Our adaptation module consist of two transformer blocks $b_S(\cdot)$ and $b_T(\cdot)$ designed to aggregate different types of information. This module not only adapts each video to the shared VL embedding space but also captures egocentric video-specific spatial and temporal information, through $b_S(\cdot)$ and $b_T(\cdot)$ respectively. To capture hand-object interactions better, we introduce an attention block that focuses on regions around hands, see Fig. 13.3. This involves applying self-attention between frame crops covering hands and full frames, guiding the model to emphasize information around hands, a crucial region of interest in egocentric videos. To aggregate temporal information from the video, we employ a temporal self-attention block, similar to [JHZ+22] that updates the frame representations. Our X-MIC-vector for adaptation is derived by applying average pooling over all updated frame representations.

Egocentric videos may include diverse backgrounds and involve significant camera motion. By focusing on the region around hands through cropping and applying self-attention to both the full frame and the cropped region, we guide our model to prioritize attention on the hands. More specifically, for each frame, we use the full frame z_i and the cropped hand region z_i^{hand} from the same frame. We get the intermediate representations through the video encoder, for the full frame $x_i = V_{II}(z_i)$ and the hand region $x_i^{hand} = V_{II}(z_i^{hand})$. We concatenate the two to obtain an intra-frame sequence $[x_i; x_i^{hand}] \in \mathcal{R}^{2\times D}$. We derive the intra-frame representation $x_i^{b_S}$ by averaging the updated representations from both the full and cropped frames:

$$x_i^{b_S} = \text{avg\_pool}(b_S([x_i; x_i^{hand}])). \tag{13.2}$$

To capture the temporal relations across frames, we apply self-attention between all frames of the video. Specifically, we use the second transformer $b_T(\cdot)$ block to update $x_i^{b_S}$ frame

representations, which we aggregate with average pooling into our X-MIC-vector:

$$a_v = \text{avg\_pool}(b_T([x_1^{b_S}, x_2^{b_S}, \cdots, x_N^{b_S}])). \tag{13.3}$$

In this way, our adaptation module effectively incorporates egocentric video-specific spatial and temporal information into the frozen vision-language embedding space, enhancing generalization to novel classes.

13.4 Experiments

Our X-MIC method is mainly evaluated on the cross-dataset setting between two large-scale egocentric video datasets Ego4D [GWB+22b] and Epic-Kitchens [DDF+22] datasets. Additionally, we evaluate generalization performance on the small-scale EGTEA [LLR18] dataset.

13.4.1 Datasets

Ego4D [GWB+22b]. We a subset of Ego4D [GWB+22b] annotated with fine-grained noun and verb labels, specifically from the FHO benchmark task. In this benchmark the training set consists of 64K video clips, while the testing set comprises 33K clips. The average clip duration is 8 seconds, resulting in a total of approximately 215 hours of videos, excluding irrelevant background clips. The dataset contains 521 noun classes and 117 verb classes.

Epic-Kitchens [DDF+22]. We use the Epic-Kitchens100, comprising 67K video clips for training and 10K video clips for testing. The average clip length is 3.5 seconds, totaling about 70 hours, excluding irrelevant background clips. The dataset features annotations for 300 noun classes and 97 verb classes, focusing on kitchen-related topics.

EGTEA [LLR18] We use this dataset solely for model testing, given its training set of 8,000 video clips with clip length of 3.2 seconds. During inference, we combine three test splits resulting in 6K video clips in total. The dataset is annotated with fine-grained 20 verb and 54 noun classes.

13.4.2 Implementation Details

We evaluate the generalization performance based on the adaptation of the pre-trained CLIP ViT-B/16 model, unless otherwise specified. For model training, we use AdamW with(β_1, β_2) = (0.9, 0.999) and weight decay of 0.01 for 15 epochs with a fixed learning rate of 1e-6. The Transformer module b_1 contains 1 self-attention layer, whereas temporal attention module b_2 includes 2 self-attention layers. During training, we sample 16 random frames, and during evaluation we sample frames uniformly. For detecting the hand regions, we use the 100DOH [SGSF20] detector, extracting bounding boxes for each frame.

Cross-Datasets Evaluation. In this work, we investigate the generalization performance of fine-grained noun and verb recognition across egocentric datasets. Our objective is two-fold: achieving strong performance within the dataset on the corresponding test set and demonstrating robust generalization to a shared dataset. We compute the harmonic mean between these two to gauge the balance of different types of generalization. For instance, we train our models on Ego4D and subsequently evaluate on both the Ego4D and Epic-Kitchens

		Trained on Ego4D (E4D)						Trained on Epic-Kitchens (EK)					
	ta	Nouns			Verbs			Nouns			Verbs		
Evaluation dataset		E4D	EK	hm	E4D	EK	hm	EK	E4D	hm	EK	E4D	hm
ZS CLIP	-	5.89	8.74	7.03	2.18	4.25	2.88	8.74	5.89	7.03	4.25	2.18	2.88
CoOp	-	28.22	10.87	15.70	22.57	20.42	21.44	21.56	9.37	13.06	30.91	13.35	18.64
Co-CoOp	-	30.00	9.51	14.44	21.31	12.99	16.14	24.23	9.27	13.41	34.16	14.17	20.03
CLIP-Adapter	-	30.00	8.95	13.78	22.82	19.94	21.28	33.21	5.73	9.77	36.70	16.09	22.37
CLIP-Adapter*	✓	31.26	10.00	15.16	27.32	22.28	24.54	34.40	4.67	8.22	48.69	15.52	23.54
A5	✓	31.39	7.84	12.55	26.31	22.77	24.41	32.04	3.31	5.99	46.05	17.93	25.81
Vita-CLIP	✓	33.52	10.61	16.11	22.66	25.81	24.13	34.41	9.52	14.91	48.78	13.47	21.11
X-MIC	✓	33.54	15.35	21.06	28.93	26.48	27.65	30.64	12.32	17.57	50.01	18.10	**26.58**
X-MIC+DINO	✓	35.85	18.96	**24.80**	28.27	29.49	**28.86**	44.07	11.45	**18.17**	53.02	16.01	24.60

Table 13.1: **SOTA comparison on within- and cross-dataset evaluation on Ego4d and Epic Kitchens datasets.** Left: Models trained on Ego4D. Right: Models trained no Epic-Kitchens. Evaluation is on noun and verb classes. ta denotes temporal attention in the corresponding method, other methods apply simple average of the frames. hm stands for harmonic mean evaluation. X-MIC+DINO denotes our model with DINO [ODM+23] as Visual Encoder II.

		Trained on E4D, Evaluated on EK						Trained on EK, Evaluated on E4D					
	ta	Nouns			Verbs			Nouns			Verbs		
Eval. subset		shared	novel	hm	shared	novel	hm	shared	novel	hm	shared	novel	hm
ZS CLIP	-	10.38	13.58	11.77	12.32	4.32	6.40	11.38	10.49	10.92	2.73	9.84	4.27
CoOp	-	16.86	16.02	16.43	25.03	5.97	9.64	15.77	10.11	12.32	20.22	5.69	8.89
CoCoOp	-	16.35	11.51	13.51	24.34	0.00	0.00	17.31	11.46	13.79	15.32	6.46	9.09
CLIP-Adapter	-	12.46	5.99	8.09	21.48	3.09	5.40	8.72	7.42	8.02	19.60	3.00	5.20
CLIP-Adapter*	✓	16.24	12.22	13.95	25.29	1.23	2.35	14.67	7.68	10.08	24.17	4.50	7.59
A5	✓	15.25	5.24	7.80	27.90	3.09	5.56	13.54	5.71	8.03	24.29	0.55	1.07
Vita-CLIP	✓	15.84	6.15	8.86	27.22	4.11	7.14	14.60	9.76	11.69	16.46	6.58	9.40
X-MIC	✓	20.04	21.51	20.75	29.01	7.00	**11.27**	19.66	12.24	**15.09**	23.00	7.16	**10.92**
X-MIC+DINO	✓	25.56	20.52	**22.76**	31.92	6.38	10.63	18.91	10.67	13.65	20.55	6.48	9.85

Table 13.2: **Zero-shot action generalization**. Left: The models are trained on Ego4D (E4D) and subsequently evaluated on Epic-Kitchnes (EK) using disjoint subsets of classes (shared and novel). Right: The models are trained on Epic-Kitchens (EK) and then evaluated in a cross-dataset manner on subsets of classes within Ego4D.

	Nouns			Verbs		
Evaluated on	Ego4D	EGTEA	hm	Ego4D	EGTEA	hm
ZS CLIP	5.89	19.70	9.07	2.18	18.71	3.51
CoOp	29.23	23.90	26.29	22.57	26.45	24.35
Co-CoOp	29.85	27.90	28.84	21.31	27.74	24.10
CLIP-Adapter	30.00	21.41	24.98	22.82	26.51	24.52
CLIP-Adapter *	29.18	22.40	25.34	27.32	26.57	26.93
A5	33.50	23.70	27.76	26.31	28.03	27.14
Vita-CLIP	33.52	17.24	22.76	22.66	27.63	24.89
X-MIC	33.54	29.21	**31.21**	28.93	31.41	**30.12**
X-MIC+DINO	35.85	30.32	**32.84**	28.27	32.01	**30.01**

Table 13.3: **SOTA comparison on EGTEA dataset.** The model is trained on Ego4D dataset and evaluated in zero-shot manner on EGTEA. X-MIC+DINO denotes our model with DINO [ODM+23] as Visual Encoder II.

test sets. In Table 13.2, we further analyse zero-shot generalization by identifying disjoint subsets of seen and novel classes across datasets.

13.4.3 X-MIC Comparison to SOTA

In Tables 13.1 and 13.2, we start with comparing our method to CLIP. Notably, on both the Ego4D and Epic-Kitchens datasets, CLIP yields surprisingly low results for both verbs and nouns, in contrast to its strong performance on third-person datasets [KCS+17, SZS12]. Next, we compare our method to other adaptation methods of VLMs which have shown improvements on image and video recognition benchmarks.

Image-based Adaptation Methods. First, we present a comparison to image-based adaptation models, including CoOp [ZYLL22b], CoCoOp [ZYLL22a], and CLIP-Adapter [GGZ+23] which do not contain a temporal component. Our analysis in Table 13.1 shows that early fusion-based models like CoOp and CoCoOp exhibit limited learning capacity, resulting in poorer performance compared to other models for both nouns and verbs when evaluated within-dataset, especially on Epic-Kitchens. This aligns with earlier findings shown in [CGT+22]. A late fusion-based framework, CLIP-Adapter, improves the within-dataset scores but demonstrates weaker generalization on nouns for cross dataset evaluation. However, a more detailed analysis of generalization performance in Table 13.2 reveals that CoOp [ZYLL22b] demonstrates robustness even in the absence of any temporal attention module. We hypothesize that other models may be more prone to overfitting on the shared classes due to a larger number of parameters.

Video-based Adaptation Methods. In Table 13.1 we evaluate the performance of recent third-person video adaptation models, specifically A5 [JHZ+22] and Vita-CLIP [was23], in an egocentric scenario. Additionally, we enhance the CLIP-Adapter model by incorporating temporal attention and evaluate its effectiveness as a video model. We notice that the inclusion or exclusion of a temporal component, beyond simple averaging, has a relatively minor impact on noun recognition using CLIP-Adapter. To illustrate, when trained on the Epic-Kitchens dataset, CLIP-Adapter, with (denoted as CLIP-Adapter*) and without a temporal attention module, exhibits comparable performance in noun recognition within the dataset (EK), with scores of 33.21% and 34.40%, respectively. However, the role of temporal attention becomes crucial in enhancing verb recognition performance, as evidenced by consistent improvements across both datasets and all models. A5 [JHZ+22], which combines both early fusion and

	Nouns			Verbs		
	E4D	EK	hm	E4D	EK	hm
F	31.68	14.20	19.61	27.19	24.02	25.51
H	31.35	14.02	19.37	26.32	26.59	26.46
F+H	33.54	15.35	**21.06**	28.93	26.48	**27.65**

Table 13.4: **Influence of Ego-Spatial-Temporal attention.** F denotes full frames, H denotes hand crops. F+H correspond to our proposed attention module. All models share the same architecture of the temporal attention module.

	Nouns			Verbs		
	E4D	EK	hm	E4D	EK	hm
w/o	31.41	13.31	18.69	22.65	20.19	21.34
w/	33.54	15.35	**21.06**	28.93	26.48	**27.65**

Table 13.5: **Influence of temporal attention.** Replacing the temporal module with a simple average decreases verb and noun recognition. The models share the same architecture for Ego-Spatial attention module.

temporal attention, shows poor cross-dataset generalization on nouns for both datasets, aligning with the findings reported by its authors [JHZ+22] in the context of cross-dataset third-person video generalization. The recent SOTA model on third-person video generalization, Vita-CLIP [was23], demonstrates enhanced noun recognition on both datasets but exhibits lower verb recognition on Ego4D. In contrast to other video adaptation models, we decouple temporal attention from the frozen backbone and introduce X-MIC-vector, encapsulating all temporal information. Moreover, employing cross-modal adaptation, we introduce video-specific classifiers. For each video, we create an individual text-based classifier, which is adapted with our X-MIC-vector. Our approach demonstrates state-of-the-art generalization performance while maintaining high performance on within-dataset evaluation. Moreover, by leveraging DINO pretrained model [ODM+23] as visual encoder V_{II}, we observe significant improvements on within-dataset evaluation. In Table 13.3, we present our evaluation on EGTEA. Overall, we note consistent trends across all methods.

13.4.4 Ablations

In this section, we evaluate the effectiveness of our design choices. For all ablations, we train models on Ego4D and evaluate on Ego4D and Epic-Kitchens datasets. As backbone, we use CLIP ViT-B/16, unless otherwise specified.

Ego-Spatial-Temporal Attention. In Table 13.4, we demonstrate the impact of utilizing full frames, that usually includes scene context, and hand crops on the performance of egocentric videos. We observe that concentrating solely on hand regions enhances verb generalization, whereas the utilization of full images proves marginally more advantageous for noun generalization. When employing our proposed ego-spatial-temporal attention mechanism, we achieve a notable improvement in the harmonic mean. Specifically, there is a 1.45% increase for nouns and a 2.14% boost for verbs compared to using full frames. By guiding the model to consider context in relation to hand areas, our attention approach not only enhances performance within the dataset but also showcases improved cross-dataset performance.

Larger backbone. In Table 13.6, we assess the effectiveness of our method using a bigger CLIP

			Nouns			Verbs		
Evaluation dataset			E4D	EK	hm	E4D	EK	hm
CLIP	ViT-L/16	Zero-Shot	5.89	8.74	7.03	2.18	4.25	2.88
		X-MIC	33.54	15.35	21.06	28.93	26.48	27.65
	ViT-L/14	Zero-Shot	8.40	13.88	10.46	8.57	9.70	9.10
		X-MIC	33.75	22.46	26.97	28.13	28.93	28.52
Lavila	ViT-L/14	Zero-Shot	24.99	31.06	27.69	6.19	15.74	8.88
		X-MIC	35.18	34.97	35.08	12.28	24.66	16.37

Table 13.6: **Influence of different backbones.** We compare the performance of CLIP ViT-L/14 with ViT-L/16. Additionally, we provide a comparison of CLIP backbone, pretrainied on text-image pairs, to Lavila backbone, pretrained on pairs of egocentric videos and narrations from full Ego4D.

norm	Nouns		
	E4D	EK	hm
n1	33.54	15.35	**21.06**
none	32.64	14.34	19.92
n2,n3	32.74	14.59	20.19
n1,n2,n3	31.99	14.49	19.95
n1,n2	15.81	12.3	13.83
n1,n3	12.12	11.34	11.71

Table 13.7: **Influence of feature normalization**. [n1] corresponds to the normalization of visual features after the V_{II} encoder and before the adapter and demonstrates an optimal balance between normalization and no normalization. [n2] corresponds to the normalization of the X-MIC vector before summation with text representation. [n3] corresponds to the normalization of text representation before summation with the X-MIC vector.

	Nouns					Verbs				
	ZS		X-MIC			ZS		X-MIC		
prompts	E4D	EK	E4D	EK	hm	E4D	EK	E4D	EK	hm
<class>	5.89	**8.74**	33.54	15.35	21.06	2.18	4.25	28.93	26.48	**27.65**
Image of a <class>	**10.52**	6.75	32.31	14.81	20.31	3.28	5.40	28.56	25.98	27.21
Video of a <class>	10.32	6.80	32.62	14.77	20.33	2.93	5.97	28.14	22.70	25.13
Egocentric image a <class>	9.61	7.11	32.09	15.65	21.04	2.98	3.83	28.58	24.02	26.10
Image of a hand holding a <class>	10.09	6.32	32.92	14.28	19.92	**3.29**	**9.87**	27.53	19.33	22.71
Egocentric image of a hand holding <class>	9.23	6.86	33.29	15.83	**21.45**	2.41	6.24	27.66	16.94	21.01

Table 13.8: **Influence of prompting the frozen text model with additional context.** ZS denotes zero-shot CLIP evaluation. Noun recognition is robust to contextual variations, while verb recognition performs best without additional context.

model, specifically comparing the performance of CLIP ViT-L/14 with ViT-L/16. While we do not observe performance gains for within the dataset evaluations, a compelling trend emerges in cross-dataset generalization, particularly on nouns. Notably, employing the larger model ViT-L/14 results in a significant improvement of over 7% in noun and 2.45% in verb generalization on Epic. This encouraging outcome underscores the potential of vision transformers and suggests that further exploration and refinement of these models could yield even more substantial gains in cross-dataset generalization.

Egocentic VL backbone. Table 13.6 presents an evaluation on X-MIC-model performance using backbones CLIP and Lavila [ZMKG23], which is pretrained on text-video pairs from the Ego4D dataset in a contrastive manner. Note that the Lavila backbone initializes its model from CLIP pretrained models. We first compare the zero-shot results from the original CLIP backbone and Lavila. Lavila demonstrates a significant improvement in noun recognition by 16.59% on the Ego4D dataset and noun generalization to Epic by 17.18%. While Lavila shows a decrease in verb recognition accuracy within the dataset by 2.38% , its generalization to Epic verbs increases by 6.04%. This outcome is surprising, as we initially expected Lavila to generalize better on verbs due to its training on an egocentric dataset, indicating a strong bias toward object-oriented pretraining strategies. We observe similar trends when our model utilizes CLIP versus Lavila as a backbone, where noun generalization increases significantly, while verb generalization slightly decreases.

Prompts for text encoder. In Table 13.8, we evaluate the performance of zero-shot CLIP and our model by prompting the frozen text model for classification with additional context. Our experiments include specific details like "Video of a" or indications of hands and an egocentric view. We find that zero-shot noun performance is the best with the standard “Image of a ” context for Ego4D and without context for Epic-Kitchens. However, zero-shot verb recognition benefits from an additional context, achieving 3.29% and 9.87% on Ego4d and Epic-Kitchens, respectively. With our X-MIC adaptation, we observe that noun recognition remains robust to these changes, while verb recognition is sensitive and performs best when no additional context is provided highlighting the complexity of incorporating contextual information in egocentric scenarios.

Importance of normalization. We investigate the significance of feature normalization in the embedding space in Table 13.7. n1 represents our default choice, involving the normalization of visual features after the V_{II} encoder and before the adapter. n2 indicates the normalization of X-MIC-vector, i.e., visual features after our video-adapter module, prior to summation with frozen text features. Lastly, n3 denotes the normalization of frozen text features before summation with X-MIC-vector. The ’none’ corresponds to no normalization. [n1] demonstrates the optimal balance between regularization and no regularization. Configurations [n1], [n2,n3], [n1,n2,n3], and ’none’ all yield symmetric feature magnitudes before the summation of frozen text features and X-MIC-vector and marginally change the harmonic mean. Conversely, variations such as [n1,n2] and [n1,n3] result in imbalances during the summation of different modalities, leading to suboptimal performance.

13.5 Conclusion

In this chapter, we introduce X-MIC, a simple yet effective cross-modal adaptation framework for VLMs, that injects egocentric video information into the frozen VL embedding, achieving significant improvements in fine-grained cross-dataset egocentric recognition of nouns and verbs. Moreover, X-MIC vectos offer decoupling of the domain-specific knowledge from the

frozen VL embedding. This allows to explore different visual backbones for text conditioning directly in the embedding space, showing improved generalization.

It is important to note that our explorations in this chapter focus solely on video classification and do not encompass text-vision tasks like text-to-video retrieval, which would necessitate using text-conditioned videos instead of our video-conditioned text representations. We plan to explore this direction in future work.

14 Conclusion and Future Work

Contents

Recent advances in deep learning have significantly increased its demand across various applications that become integral in our daily lives. Tools developed from pre-trained deep learning models have seamlessly integrated into our everyday routines. However, the efficacy of these advances depends on the availability of data and corresponding annotations tailored to specific tasks. The process of gathering annotations is labour-intensive and time-consuming, especially when aiming for large-scale, high-quality datasets. Furthermore, the dynamic nature of the world makes it impossible to label every aspect of the world. The prohibitive costs associated with annotations present a barrier to scaling up annotations for diverse tasks. Therefore, to facilitate the expansion of datasets and address a wider range of tasks, it is important to explore ways to mitigate annotation and data collection costs.

14.1 Key Insights and Conclusions

In this thesis, we investigated challenges arising when full supervision for training models is not available and explored four complimentary research directions: First, in Part I, we investigate learning without labels, referred to as self-supervised and unsupervised methods, to better understand video and image representations. To learn from data without labels, it is intuitive to inject priors such as invariance, speed of the actions in the videos, or semantic information granularity to obtain powerful data representations. We found that formulating the objective for self-supervised learning that reflects the inherent biases of the data can significantly enhance performance on the downstream tasks. Moreover, gaining insights into the learning dynamics of contrastive learning methods proved beneficial for training on underexplored long-tail data and improving representations in the latent space. Then, in Part II, we considered scenarios involving reduced supervision levels. We proposed to automate annotation of videos on large scale leveraging large-language models. To reduce annotation costs, we proposed to omit precise annotations for one of the modalities in multimodal learning, namely in text-video and image-video settings, and transfer available knowledge to large copora of unlabeled video data. Moreover, we studied semi-supervised learning scenario, where only a subset of annotated data alongside unlabeled data, and explored influence of different regularization techniques on pseudo-labeling of unlabeled data. In Part III, we contributed towards the scenarios where parts of available data is inherently limited due to privacy and security reasons or naturally rare, which not only restrict annotations but also limit the overall data volume. Training with imbalanced data typically results in severe imbalance in predictions. However, by employing

our calibration, space reservation, and margin enhancement techniques, the network can substantially enhance performance on underrepresented samples while maintaining strong performance on classes represented with abundant samples. Finally, in Part IV, we explored scenarios where the model is asked to generalize beyond the given set of object or action classes. In an open-set setting, our framework decouples the tasks of detecting outliers and classifying the known classes into two parallel branches. This approach notably enhances both tasks, as we discovered that optimizing these tasks over the same representations can lead to conflicting outcomes. Furthermore, there has been increasing interest in adapting vision-language models (VLMs) for image and third-person video classification, driven by their success in zero-shot recognition. However, the adaptation of these models to egocentric videos remains largely unexplored. We proposed a new approach to investigate the domain gap between pretrained VLMs and egocentric videos, along with a straightforward yet powerful framework for adapting these models to the new domain in open world setting.

In the following, we revisit the contributions of individual chapters in more detail, before discussing future work in Section 14.2.

Part I, Learning Without Supervision: In the first part, our research is focused on unsupervised and self-supervised learning to develop effective image and video recognition models using powerful and robust representations.

In Chapter 3, we explored the unsupervised learning of temporal action segmentation in untrimmed instructional videos, videos where content is designed to teach viewers how to perform some activity or task. We proposed a novel approach to learn an embedding that captures the inherent order of actions in instructional videos by learning the relationships between frames based on their relative timing. This reflects the natural flow of instructional videos, where actions typically have a specific sequence (e.g., pouring water into a mug requires holding the mug first). To demonstrate the effectiveness of our learned representations, we introduced an unsupervised decoding method that segments videos into coherent action segments based on an ordered clustering of the embedded features.

In Chapter 4, similarly, we leveraged the inherent biases present in videos, where the behavior of actions is predefined by the task at hand. Unintentional actions refer to instances where something does not proceed as intended, such as a person stumbling instead of continuing forward. This behavior can be simulated through temporal transformations like changes in motion direction and speed. We specifically formulated these temporal transformations for unintentional actions to bias the model towards capturing abrupt changes in behavior. By pretraining the model to predict these transformations in a self-supervised manner, we observed that our task-specific representations notably enhance performance across various tasks involving unintentional actions.

In Chapter 5, we aimed to improve representation learning on long-tail data by enhancing discrimination of the underrepresented classes. We investigated the behavior of one of the most popular variants of self-supervised methods, specifically contrastive methods, and uncovered correlations between the temperature parameter and the creation of semantic clusters. We discovered that by adjusting the temperature parameter τ, we can influence the type of semantic information captured in the embedding space. Traditionally, τ has been treated as a constant hyperparameter. However, in this study, we proposed the use of a dynamic τ and demonstrate that a simple cosine schedule can lead to significant improvements in the learned representations.

Part II, Learning with Limited Supervision: In the second part, we explored various scenarios that allow to reduce the level of supervision necessary for training models.

In Chapter 6, we investigated scenarios where only a subset of annotated data, alongside

unlabeled data, is available—a scenario commonly referred to as semi-supervised learning. Traditionally, available labels are extended to unlabeled data through pseudo-labeling. However, obtaining accurate pseudo-labels necessitates training a generalizable model on the limited labeled data. To enhance generalizability, we explored various regularization techniques. Among these, consistency regularization, which enforces invariant representations, is one of the most effective and efficient methods for semi-supervised learning. Nevertheless, our findings indicate that encouraging equivariance instead, achieved by increasing the feature distance, further enhances performance. By applying different types of regularization to different features in the embedding, we were able to improve both the generalization of the model and the accuracy of pseudo-labeling.

In Chapter 7, we addressed the challenge of downstream fine-tuning on video-text pairs, which typically requires abundant densely annotated in-domain video data. To bypass the high annotation costs associated with this task, we introduced a novel setting: text-video retrieval with uncurated & unpaired data. In this new setting, training solely relies on text queries and uncurated web videos, without any paired text-video data. Additionally, we proposed the In-Style method, which facilitates the transfer of the style from the target text queries to the vast collection of video data. We evaluated our pseudo-generated annotated text-video pairs on downstream tasks and observed performance comparable to fully supervised training. This represents a significant step toward achieving low-cost targeted downstream fine-tuning.

In Chapter 8, we further explored low-cost video annotations, specifically discovering that web images serve as an excellent source of pseudo-labels transferable to videos However, bridging the gap between web images and videos involves two distinct challenges: a spatial domain shift between web images and video frames, and a modality gap between image and video data. To tackle these challenges, we proposed a cyclic-based approach aimed at iteratively addressing each challenge in isolation. Utilizing a spatial model, we narrowed the gap between images and video frames by enforcing consistency. Additionally, employing a spatio-temporal model, we learned temporal information that is necessary for understanding videos.

In Chapter 9, we considered large-scale multimodal instructional video pretraining datasets. A common method for collecting such video datasets at scale involves web crawling to gather relevant videos. Typically, these videos come with free, noisy annotations in the form of automatically transcribed speech converted into subtitles using automatic speech recognition systems. These subtitles are then utilized for pretraining models. However, the alignment between this speech and the videos is often suboptimal, resulting in less-than-ideal training supervision. To address this issue, we proposed leveraging large language models to enhance the quality and alignment of these subtitles with their corresponding videos. With our automatic prompting pipeline, we can generate human-like video captions on a large scale without requiring human supervision. This improvement in the quality of text-video pairs used for pretraining results in significantly enhanced representations for various downstream video tasks.

Part III, Learning with Limited Data: In the third part, we considered settings that reflect scenarios where there are only a limited number of training samples available per task, and the model is required to dynamically adapt to the new environment, relying only on a few labeled samples.

In Chapter 10, we placed particular focus on learning balanced representations for both overrepresented and underrepresented data while processing only a portion of the data at any given time. The main challenge of this task is to integrate new information relying solely on a limited number of data samples while retaining previously acquired knowledge from abundant data. We discovered that classification learned with overrepresented data tends to

result in overly confident predictions. In order to accommodate novel classes, we introduced a base-normalized cross-entropy approach that dynamically adjusts the cross-entropy loss based on input samples, facilitating faster acquisition of new information. To recalibrate learning after training on novel data, we employed balanced replay, incorporating both previous and novel classes, which leads to calibrated performance.

In Chapter 11, we tackled the task of continual learning with limited data from the metric learning perspective. First, we found that combining self-supervised learning with supervised learning significantly improves the learned representations, enabling effective generalization to subsequent tasks with limited data. Furthermore, partitioning the representation space into distinct subspaces for each class allowed us to accommodate novel classes without sacrificing performance on previously learned classes. By explicitly enhancing the margins between classes, we promoted compact representations of each class, resolving conflicts between conflicting classes and leading to more effective representations.

Part IV, Learning Beyond Supervision: In the fourth part, we shifted our attention towards learning beyond the boundaries of the data presented during training at any moment of the training. One of the prominent challenges lies in finding a balance between memorizing patterns from the observed data and generalizing to unseen classes.

In Chapter 12, we investigated a realistic scenario where the available data at training time comprises labeled few-shot samples and an additional unlabeled set of data with unknown classes. The objective is twofold: correctly classify test samples belonging to known classes and simultaneously detect outliers representing unknown classes. Surprisingly, we discovered that the classification performance can be significantly improved by incorporating unlabeled data, even when some samples do not belong to the respective classification classes. However, in this situation, the task of detecting unknown classes conflicts with the classification task. To address this, we proposed disentangling the two tasks using linear layers, which significantly enhanced performance for both objectives.

In Chapter 13, we further explored the open-world setting with the help of vision-language pretrained models. In particular, we addressed the challenges of egocentric action recognition, highlighting the challenge of generalization due to the absence of pretraining web data in this domain. Hence, we introduced a simple yet effective cross-modal adaptation framework enabling the adaptation of a general multimodal classifier to each input video individually. Furthermore, we focused specifically on the egocentric domain and leveraged spatial information around hands to enhance the attention of the model. We showed that our framework yields remarkable generalization to the new domain and unseen classes.

14.2 Future Directions

In the following, we provide a discussion on potential future directions within the scope of this thesis. Our focus lies on areas we find especially important and exiting, particularly in light of foundational models such as large language models or multimodal models. These models have opened up a wide range of possibilities across various applications. We prioritize topics that diverge from conventional settings, such as few-shot learning, where models are trained from scratch on minimal data, which may not be as relevant. Instead, we explore learning, evaluation and adaptation of foundational models with multimodal or limited data scenarios. To achieve this vision, below we outline a few future directions we plan to pursue.

14.2.1 Learning Representations

Learning representations that accurately reflect the intricate underlying structure of training data is crucial for deep learning models to achieve improved performance, generalization, and robustness across diverse tasks and domains. Hence, it is essential to develop models that address the inherent complexity of the data.

Multi-modal class imbalance in self-supervised learning: Learning robust representations encompasses capturing semantic information across various levels of granularity. Different tasks may require different degrees of detail to be captured effectively. However, natural data often adheres to a long-tail distribution, where certain events occur only in specific circumstances. When learning from such data, it becomes important to understand whether the algorithm adequately captures rare events or exhibits bias towards overrepresented samples. To tackle this challenge, in future work, we are interested in investigating the biases present in large-scale pretrained models due to inherent natural distributions. Additionally, we aim to explore methods that can help mitigating these pitfalls in the future, ensuring that models learn from diverse and representative data while avoiding skewed representations. Extending our understanding in this area will be vital for developing more inclusive and accurate machine learning models across a range of applications.

Large-scale meta-learning for self-supervised methods: Current self-supervised large-scale learning requires extensive datasets, with models expected to perform well across a variety of tasks. We believe that the exploration of meta-learning methods at a large scale holds significant potential. Specifically, we propose exploring adaptive self-supervised learning techniques capable of dynamically adjusting representations based on available data and task demands. This approach would entail simulating diverse scenarios during training, thereby streamlining adaptation to more specific domains.

Improving alignment and quality of pretraining data: Expanding on the ideas presented in Chapter 9, we are interested in enhancing the alignment and overall quality of pretraining multimodal data. This could encompass several tasks, such as exploring strategies to integrate supplementary contextual visual, language, sound, potentially touch cues into the captioning procedure could prove beneficial. Investigating methods to alleviate noise and errors inherent in pretraining data sets is another promising direction. This would require through analysis of the potential biases and noise in the data with the following methods to mitigate them. Moreover, the development of automatic speech recognition systems capable of contextual conditioning on both short-term and long-term video content for generating descriptive narratives in real time is an interesting direction for the future work.

14.2.2 Reducing the Annotation Costs

The high cost and labor involved in collecting and annotating large datasets pose a significant challenge. To address this bottleneck, a key area of research lies in developing strategies that reduce reliance on extensive labeled data. This includes leveraging weaker annotation methods across multiple modalities and effectively utilizing limited sets of available labeled samples across modalities.

Mutimodal generalization to unseen tasks and domains with limited data: Exploring strategies to enhance the generalization of multimodal representations across diverse tasks and domains is an additional direction, particularly for scenarios where labeled data is scarce. This could involve developing methods to disentangle task-specific and task-agnostic features within the learned representations, exploring methods for obtaining transferable features robust to task and domain variations, investigating approaches for few-shot learning capable of tailoring multi-modal general representations to novel tasks with minimal labeled data, or adapting models trained on one modality or domain to excel in another modality or domain despite limited labeled samples.

Mutlimodal methods to reduce annotation costs: Extending the approach proposed in Chapter 7 to address multimodal learning scenarios beyond text-video retrieval is a further direction. This could involve exploring methods for leveraging unpaired data in other modalities, such as audio-video or image-text pairs, and developing techniques to effectively align and fuse information from different modalities for downstream tasks such as captioning, retrieval, or generation.

Transfer learning across modalities: Research on transfer learning techniques that can transfer knowledge from a source modality with abundant data to a target modality with limited data will be important. This could involve exploring methods for adapting pre-trained models from one modality to another, developing techniques for learning shared representations across modalities, investigating approaches for domain adaptation or fine-tuning in multi-modal settings, or exploring data augmentation techniques specifically designed for improving generalization across modalities with limited or abundant data. This could also involve developing modal-specific data augmentation strategies, such as temporal language shift relative to the corresponding video [MZA+19].

14.2.3 Understanding and Adaptation of the Representations

Investigating the relationship between different representations is important, alongside expanding this understanding to different data types to enhance model performance and reveal general learning principles. Moreover, by improving standard benchmarks to accurately reflect real-world scenarios and adapting methods to specific domains and tasks, novel low-cost learning approaches will address real-world challenges effectively.

Understanding of the learned representations: In our thesis, we uncovered a strong correlation between a specific hyperparameter and the semantic information learned by our model, as outlined in Chapter 5. Further exploring how learning algorithms behave across different types of data is essential. Recent methods like masked image modeling [HCX+22] and diffusion models [RBL+21] offer promising starting points for models, but it is crucial to understand any potential problems they may bring. Expanding our understanding in this area not only improves model performance but also helps uncover general principles guiding learning across various datasets.

Evaluation and benchmarking of low-cost learning approaches in real-world scenarios: Improve the existing standard benchmarks to accurately reflect the challenges faced by current foundational models in adapting to diverse low-cost scenarios. This could include comparing

the performance of different methods under various conditions, such as different levels of supervision, domain shifts, or data modalities, and identifying the strengths and limitations of each approach in different scenarios. Furthermore, these benchmarks must include rigorous testing of methods in settings characterized by data scarcity or class imbalance, thereby evaluating their efficacy in dynamic and evolving environments. Additionally, the benchmarks should compare the performance of these methods against established state-of-the-art approaches to understand their relative strengths and limitations comprehensively. Improving standard benchmarks to is crucial for development of novel effective methods for applying foundational models in real-world contexts.

Application to the specific domains and tasks: Methods in few-shot and incremental learning scenarios are driven by real-world demands, such as medical imaging or sensor data analysis. Yet, current benchmarks rely on standard object recognition pipelines using datasets like ImageNet, failing to capture the intricacies of real-world data. Therefore, we think that it is necessary to assess these methods using domain-specific data and, moreover, develop methods specifically for the respective domains. By addressing domain-specific challenges, this research can yield practical applications and profound impacts across diverse fields.

Domain-specific adaptation techniques: Explore domain-specific adaptation techniques to improve the performance of models in specific application domains or tasks. This could involve developing tailored adaptation methods that leverage domain-specific information, such as spatial or temporal cues in egocentric action recognition, medical imaging data, or natural language processing tasks. Additionally, research could focus on developing techniques for fine-tuning pretrained models on domain-specific data with limited labeled samples, potentially incorporating transfer learning or meta-learning approaches to improve adaptation efficiency.

Egocentric action recognition: In Chapter 13, we found that current general-purpose Vision-Language Models (VLMs) underperform when applied to egocentric videos. Yet, investigations into egocentric video analysis remain relatively limited, requiring a more diverse exploration of fine-grained recognition in this evolving domain. Notably, recognition of third-person activities surpasses that of egocentric activity recognition by a significant margin [ZG23, RkM$^+$23b]. It is important to understand the challenges, discrepancies, and influential factors affecting performance to move advancements towards achieving more adaptive real-world assistance in interactive settings.

Long-term activity and verb recognition: Current vision-language models rely on recognizing objects in static images or simple actions during the training from the datasets collected from the web, therefore, these models lack understanding of long-term context and fine-grained verbs [MCN$^+$23]. To address this, we are interested in developing algorithms that adapt these models for domains requiring a deeper understanding of the sequence of events, like egocentric videos. By enabling fine-grained activity recognition beyond just objects, these models can unlock real-world applications in augmented reality (AR) glasses that understand user interactions or robots capable of performing complex tasks like preparing a meal. Ultimately, overcoming these limitations promotes a future where vision-language models seamlessly integrate with our daily lives.

14.2.4 A Broader View on the Topic

Our long-term goal is to create methods that facilitate model adaptation to the ever-changing world, encompassing new environments and diverse tasks, while minimizing the associated data collection and annotation overhead. Real-world scenarios often present specialized tasks with scarce, diverse, imbalanced and noisy data, highlighting the limitations of traditional, fully-supervised learning on downstream tasks. To overcome this challenge, it is natural to rely on foundational models that offer a powerful starting point for various downstream tasks, eliminating the need to build models entirely from scratch for each specific task. However, to adapt models in the data limited scenarios, we should explore alternative supervision signals and methods. Exploring the influence of different modalities (vision, language, audio, touch) on human perception will significantly enhance the generalizability of AI models. By incorporating more than just sight and sound, the models will improve the understanding of the real world by perceiving and reacting to situations from multiple perspectives. Therefore, I am interested in developing efficient learning algorithms that integrate these diverse information sources in different learning scenarios like self-supervised learning, few-shot learning and semi-supervised learning. With that we can achieve cost-effective and robust training protocols, enabling AI models to succeed in real-world environments and adapt with limited data.

Bibliography

[ABA+16] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on pages 32, 34, 35, 37, 38, 40, 41, and 43.

[ABARB21] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *The AAAI Conference On Artificial Intelligence (AAAI)*, 2021. Cited on page 20.

[ABD+22] Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on page 4.

[AC19] Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 25.

[ADC+22] Touqeer Ahmad, Akshay Raj Dhamija, Steve Cruz, Ryan Rabinowitz, Chunchun Li, Mohsen Jafarzadeh, and Terrance E. Boult. Few-shot class incremental learning leveraging self-supervised features. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 173.

[ADD+22] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 16.

[AEHKL+16] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *Arxiv Preprint Arxiv:1609.08675*, 2016. Cited on pages 19 and 136.

[AHWS+17] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2017. Cited on pages 112 and 114.

[AKS24] Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast foro few-shot class-incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2024. Cited on pages 12, 14, and 171.

[ALT+23] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions On Computational Social Systems*, 2023. Cited on page 22.

[AOA+20] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference On Neural Networks (IJCNN)*, 2020. Cited on pages 18, 83, 91, and 92.

[ARV20] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 16 and 68.

[ARZV21] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 5.

[ATRPP17] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 2017. Cited on page 21.

[AW20] Ali Ayub and Alan R Wagner. Cognitively-inspired model for incremental learning using a few examples. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 24.

[AY01] Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *Proceedings Of The 2001 ACM Sigmod International Conference On Management Of Data*, 2001. Cited on page 188.

[AYQ+21] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 113.

[BAP14] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2014. Cited on pages 18, 84, 86, and 103.

[BB15] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2015. Cited on page 8.

[BB16] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on page 25.

[BBC+19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *Arxiv Preprint Arxiv:1912.06680*, 2019. Cited on page 1.

[BCC+20] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 18, 19, 20, 83, 90, 91, 92, 101, 102, and 188.

[BCG+19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on pages 18, 83, 89, 90, 91, 92, and 188.

[BCJ+22] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 107.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances In Neural Information Processing Systems (NeurIPS)*, 2016. Cited on page 1.

[BDLR06] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. *Semi-supervised Learning*, 2006. Cited on page 19.

[BDPW21] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *Arxiv Preprint Arxiv:2106.08254*, 2021. Cited on page 17.

[Bel66] Richard Bellman. Dynamic programming. *Science*, 1966. Cited on page 87.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference On Fairness, Accountability And Transparency*, 2018. Cited on page 1.

[BHA+21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *Arxiv Preprint Arxiv:2108.07258*, 2021. Cited on page 107.

[BHB19] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 84.

[Bis06] Christopher M. Bishop. *Pattern Recognition And Machine Learning (information Science And Statistics)*. 2006. Cited on page 7.

[BIS+23] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *Arxiv Preprint Arxiv:2304.12210*, 2023. Cited on pages 4 and 69.

[BKVM15] Xavier Bouthillier, Kishore Konda, Pascal Vincent, and Roland Memisevic. Dropout as data augmentation. *Arxiv Preprint Arxiv:1506.08700*, 2015. Cited on page 22.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 1 and 6.

[BNVZ21] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 6, 19, 107, 113, 114, 141, 142, 143, and 144.

[BP19] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 23, 24, 157, and 172.

[BP20] Eden Belouadah and Adrian Popescu. scail: Classifier weights scaling for class incremental learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. Cited on page 24.

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *Arxiv Preprint Arxiv:2004.05150*, 2020. Cited on page 53.

[BPL22] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on page 16.

[BRS+22] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on page 5.

[BW08] Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research (JMLR)*, 2008. Cited on page 25.

[BZK+17] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 16 and 69.

[C+] Ego4D Consortium et al. Egocentric live 4d perception (ego4d) database: A large-scale first-person video database, supporting research in multi-modal machine perception for daily life activity. Cited on page 107.

[CB23] Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *Arxiv Preprint Arxiv:2303.11504*, 2023. Cited on page 20.

[CBJD18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 16.

[CBL+21] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 107.

[CD11] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings Of The 49th Annual Meeting Of The Association For Computational Linguistics: Human Language Technologies*, 2011. Cited on pages 112, 137, 140, and 142.

[CDAT18] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on pages 23 and 172.

[CDW+22] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *Arxiv Preprint Arxiv:2205.08534*, 2022. Cited on page 204.

[CFGH20] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *Arxiv Preprint Arxiv:2003.04297*, 2020. Cited on pages 16 and 68.

[CFN+22] Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference On Machine Learning (ICML)*, 2022. Cited on pages 173, 176, and 183.

[CGR+24] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2024. Cited on page 1.

[CGT+22] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. Cited on pages 26 and 212.

[CH21] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 4 and 16.

[cim] Cited on page 194.

[CJL+19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 74.

[CKA+19] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 123, 127, 128, and 129.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Arxiv Preprint Arxiv:2002.05709*, 2020. Cited on pages 3, 16, 18, 68, 70, 75, 78, 87, 89, 126, and 173.

[CKS+20] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *Arxiv Preprint Arxiv:2006.10029*, 2020. Cited on pages 16, 18, 84, 87, and 89.

[CKSP15] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2015. Cited on page 25.

[CL21] Kuilin Chen and Chi-Guhn Lee. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference On Learning Representations (ICLR)*, 2021. Cited on pages 155, 157, 160, 163, 164, and 165.

[CLB+20] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 127.

[CLK+19] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *Arxiv Preprint Arxiv:1904.04232*, 2019. Cited on pages 7, 23, and 172.

[CLL21] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 18, 69, and 79.

[CLL+23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *Large Model Systems Organization*, 2023. Cited on pages 20, 136, 139, 142, and 145.

[CLSZ20] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 26.

[CLTB21] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 20.

[CLW+24] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances In Neural Information Processing Systems (NeurIPS)*, 2024. Cited on page 6.

[CMJG+18] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on pages 22, 159, 166, 172, and 177.

[CMM+20] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 16, 68, and 78.

[CMWP23] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on page 16.

[CNL11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings Of The Fourteenth International Conference On Artificial Intelligence And Statistics*, 2011. Cited on page 90.

[CQS+20] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 25.

[CRC+20] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference On Machine Learning (ICML)*, 2020. Cited on page 17.

[CRD+21] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 20, 142, and 144.

[CRF+21] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 7, 23, 172, and 177.

[CSCH20] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. Cited on page 123.

[CSDS21] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 113, 142, 143, and 144.

[CSMY24] Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. Opening the vocabulary of egocentric actions. *Advances In Neural Information Processing Systems (NeurIPS)*, 2024. Cited on pages 27 and 205.

[CSSH20] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 123, 127, 128, and 129.

[CSZ09] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions On Neural Networks*, 2009. Cited on page 5.

[CTF+23] H Chen, R Tao, Yue Fan, Y Wang, M Savvides, J Wang, B Raj, X Xie, and Bernt Schiele. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *International Conference On Learning Representations (ICLR)*, 2023. Cited on page 5.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 16.

[CW16a] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference On Machine Learning (ICML)*, 2016. Cited on page 86.

[CW16b] Taco S Cohen and Max Welling. Steerable cnns. *Arxiv Preprint Arxiv:1612.08498*, 2016. Cited on page 86.

[CWG+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 101.

[CWHL21] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. Spatial-temporal causal inference for partial image-to-video adaptation. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2021. Cited on pages 123, 127, and 128.

[CXH21] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 3.

[CZ17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 18, 38, and 127.

[CZJW20] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 107.

[CZLG20] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2020. Cited on pages 26, 188, 189, 190, and 191.

[CZSL20] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 90, 115, and 194.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 50, 53, 56, 59, and 107.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. Cited on pages 17 and 20.

[DCO+20] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 159 and 166.

[DCRS20] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 159 and 162.

[DDF+22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal Of Computer Vision (IJCV)*, 2022. Cited on pages 203, 205, and 210.

[DDM+21] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Pdan: Pyramid dilated attention network for action detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on page 47.

[DDM+23] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference On Machine Learning (ICML)*, 2023. Cited on page 5.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. imagenet: a Large-scale Hierarchical Image Database. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2009. Cited on pages 68, 92, 127, 160, and 194.

[DGB18] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances In Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 25.

[DGE15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2015. Cited on page 3.

[DGF16] Emily Denton, Sam Gross, and Rob Fergus. Semi-supervised learning with context-conditional generative adversarial networks. *Arxiv Preprint Arxiv:1611.06430*, 2016. Cited on page 19.

[DGRS21] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Arxiv Preprint Arxiv:2101.07974*, 2021. Cited on page 17.

[DHT+21] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2021. Cited on page 23.

[DJ21] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 19.

[DKKP21] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 107.

[DKS22] Enea Duka, Anna Kukleva, and Bernt Schiele. Leveraging self-supervised training for unintentional action recognition. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 10, 13, and 47.

[DLAM+21] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2021. Cited on page 1.

[DMCD19] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 1.

[DT17] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *Arxiv Preprint Arxiv:1708.04552*, 2017. Cited on pages 22 and 90.

[DWZW22] Zexing Du, Xue Wang, Guoqing Zhou, and Qing Wang. Fast and unsupervised action boundary detection for action segmentation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 33.

[DX17] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on page 33.

[DY22] Guodong Ding and Angela Yao. Leveraging action affinity and continuity for semi-supervised temporal action segmentation. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 33.

[DYC+20] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 16.

[DZQ+23] Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, and Yinghuan Shi. Towards semi-supervised learning with non-random missing labels. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on page 19.

[DZX+20] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 21 and 122.

[ECV20] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 48, 49, 56, 57, 59, and 60.

[ELRS22] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2022. Cited on page 9.

[EP04] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings Of The Tenth ACM Sigkdd International Conference On Knowledge Discovery And Data Mining*, 2004. Cited on page 159.

[EV20] Dave Epstein and Carl Vondrick. Video representations of goals emerge from watching failure. *Arxiv Preprint Arxiv:2006.15657*, 2020. Cited on page 49.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference On Machine Learning (ICML)*, 2017. Cited on pages 21, 22, and 23.

[FBGG17] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on page 17.

[FDKS22] Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 14, 100, 101, and 102.

[Fei20] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 121.

[FG19] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 33.

[FGO+15] Basura Fernando, Efstratios Gavves, José M. Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2015. Cited on pages 18 and 38.

[FHSJ14] Emily B. Fox, Michael C. Hughes, Erik B. Sudderth, and Michael I Jordan. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals Of Applied Statistics*, 2014. Cited on page 33.

[FHW16] Lydia Fischer, Barbara Hammer, and Heiko Wersing. Optimal local rejection for classifiers. *Neurocomputing*, 2016. Cited on page 25.

[FKDS23a] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Revisiting consistency regularization for semi-supervised learning. *International Journal Of Computer Vision (IJCV)*, 2023. Cited on pages 11, 13, and 83.

[FKDS23b] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Ssb: Simple but strong baseline for boosting performance of open-set semi-supervised learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on pages 13, 14, and 187.

[FKS21] Yue Fan, Anna Kukleva, and Bernt Schiele. Revisiting consistency regularization for semi-supervised learning. In *The German Conference on Pattern Recognition (GCPR)*, 2021. Cited on pages 11, 13, 83, 85, and 188.

[FOS20] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *Corr*, 2020. Cited on page 84.

[FP22] Chen Feng and Ioannis Patras. Adaptive soft contrastive learning. In *2022 26th International Conference On Pattern Recognition (ICPR)*, 2022. Cited on page 16.

[FSL+21] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 26.

[FV18] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 16 and 69.

[FWD+19] Yanwei Fu, Xiaomei Wang, Hanze Dong, Yu-Gang Jiang, Meng Wang, Xiangyang Xue, and Leonid Sigal. Vocabulary-informed zero-shot and open-set learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2019. Cited on page 25.

[FXM+21] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *Arxiv Preprint Arxiv:2104.11227*, 2021. Cited on page 49.

[FXXC21] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *Arxiv Preprint Arxiv:2106.11097*, 2021. Cited on pages 107 and 108.

[GB05] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2005. Cited on page 193.

[GCDZ19] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 49.

[GDCG17] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *Arxiv Preprint Arxiv:1707.07418*, 2017. Cited on page 25.

[GDG+17] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *Arxiv Preprint Arxiv:1706.02677*, 2017. Cited on page 4.

[GEF+17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Icassp*, 2017. Cited on page 113.

[GENL+23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 2.

[GEY17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances In Neural Information Processing Systems (NeurIPS)*, 2017. Cited on page 25.

[GFA07] Gutemberg Guerra-Filho and Yiannis Aloimonos. A language for human action. *Computer*, 2007. Cited on page 33.

[GFC20] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *Arxiv Preprint Arxiv:2012.15723*, 2020. Cited on page 26.

[GGZ+23] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal Of Computer Vision (IJCV)*, 2023. Cited on pages 26, 27, 204, 207, and 212.

[GH10] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings Of The Thirteenth International Conference On Artificial Intelligence And Statistics*, 2010. Cited on pages 7 and 173.

[GHP07] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. Cited on page 194.

[GHZ+18] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 121.

[GK18] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 155, 156, 161, and 162.

[GK19] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 168.

[GL15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference On Machine Learning (ICML)*, 2015. Cited on pages 125 and 127.

[GLC+21] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *Arxiv Preprint Arxiv:2111.05610*, 2021. Cited on page 107.

[GLL18] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. 2018. Cited on page 161.

[GMX+13] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Arxiv Preprint Arxiv:1312.6211*, 2013. Cited on pages 23 and 172.

[GS22] Matthew Gwilliam and Abhinav Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 69.

[GSA+20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 3, 4, and 16.

[GSAS20] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 107, 112, and 144.

[GSDG16] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *European Conference On Computer Vision (ECCV)*, 2016. Cited on pages 21 and 122.

[GSK18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference On Learning Representations (ICLR)*, 2018. Cited on pages 16, 20, 88, and 89.

[GSN17] Chuang Gan, Chen Sun, and Ram Nevatia. Deck: Discovering event composition knowledge from web images for zero-shot event detection and recounting in videos. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2017. Cited on page 121.

[GTM19] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 19.

[GUA+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 2016. Cited on pages 125, 127, 128, and 129.

[GVM+22] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 20 and 107.

[GWB+22a] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 5 and 107.

[GWB+22b] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin,

Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 203, 205, and 210.

[GWL21] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 18.

[GYP20] Gunshi Gupta, Karmesh Yadav, and Liam Paull. Look-ahead meta learning for continual learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 22.

[GYY+16] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on pages 21 and 122.

[GZJ+20] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference On Machine Learning (ICML)*, 2020. Cited on pages 9, 188, 189, 190, and 191.

[GZPB20] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Arxiv Preprint Arxiv:2011.00597*, 2020. Cited on page 47.

[HA04] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 2004. Cited on page 188.

[HBD+20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on page 110.

[HCM+19] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 22.

[HCX+22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 17 and 222.

[her] Cited on pages 7, 23, 24, 172, and 180.

[HFC+21] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 188, 189, 190, 191, 192, and 199.

[HFFN16] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference On Computer Vision (ECCV)*, 2016. Cited on page 33.

[HFW+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 3, 16, 18, 70, and 75.

[HG16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference On Learning Representations (ICLR)*, 2016. Cited on pages 194 and 195.

[HG17] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2017. Cited on page 22.

[HGJ+19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference On Machine Learning (ICML)*, 2019. Cited on pages 26 and 205.

[HHLY22] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 188, 189, and 190.

[HHYY22] Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2022. Cited on pages 188, 189, and 190.

[HKS18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 56.

[HLK17] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *Arxiv Preprint Arxiv:1711.10125*, 2017. Cited on page 25.

[HPL+19] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on pages 22, 140, 157, 163, 164, 165, 168, and 172.

[HRE+20] K Han, SA Rebuffi, S Ehrhardt, A Vedaldi, and A Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on page 9.

[HSF18] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International Conference On Learning Representations (ICLR)*, 2018. Cited on page 85.

[HSW+22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on pages 9 and 27.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Arxiv Preprint Arxiv:1503.02531*, 2015. Cited on pages 22, 159, and 172.

[HVZ19] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 25 and 26.

[HWVDI+19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings Of The 2019 Chi Conference On Human Factors In Computing Systems*, 2019. Cited on page 1.

[HXZ19] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 17.

[HXZ20] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 49, 56, and 57.

[HXZ22] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 6, 20, 136, 140, 141, 142, and 144.

[HYG22] Zhuo Huang, Jian Yang, and Chen Gong. They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions On Multimedia*, 2022. Cited on pages 26, 188, 189, and 190.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on pages 24, 75, 92, 127, 128, and 194.

[ICP+21] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 173, 176, and 183.

[ima] Cited on pages 5 and 203.

[ITAC18] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Mining on manifolds: Metric learning without labels. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 16, 18, and 69.

[ITAC19] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on pages 91 and 92.

[IWW+21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. Cited on page 5.

[IZKA15] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. *Arxiv Preprint Arxiv:1511.06811*, 2015. Cited on pages 17 and 18.

[Japoo] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *In Proceedings Of The International Conference On Artificial Intelligence*, 2000. Cited on page 101.

[JCMW21] Ziyu Jiang, Tianlong Chen, Bobak Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International Conference On Machine Learning (ICML)*, 2021. Cited on pages 69, 74, and 75.

[JDC+20] Mohsen Jafarzadeh, Akshay Raj Dhamija, Steve Cruz, Chunchun Li, Touqeer Ahmad, and Terrance E Boult. A review of open-world learning and steps toward open-world learning without labels. *Arxiv Preprint Arxiv:2011.12906*, 2020. Cited on pages 8, 9, and 25.

[JHZ+22] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 9, 27, 205, 209, 212, and 213.

[JNDV18] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *The British Machine Vision Conference (BMVC)*, 2018. Cited on page 123.

[Joa03] Thorsten Joachims. Transductive learning via spectral graph partitioning. In *International Conference On Machine Learning (ICML)*, 2003. Cited on page 19.

[JOE20] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Arxiv Preprint Arxiv:2006.14613*, 2020. Cited on page 17.

[JQ19] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 22.

[JTC+22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 4.

[JW19] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 22.

[JXAN20] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions Of The Association For Computational Linguistics*, 2020. Cited on page 26.

[JYX+21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference On Machine Learning (ICML)*, 2021. Cited on pages 5, 19, 136, and 203.

[KAG22] Bulat Khaertdinov, Stylianos Asteriadis, and Esam Ghaleb. Dynamic temperature scaling in contrastive self-supervised learning for sensor-based human activity recognition. *IEEE Transactions On Biometrics, Behavior, And Identity Science*, 2022. Cited on pages 16 and 69.

[KAS14] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2014. Cited on pages 32 and 37.

[KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference On Learning Representations (ICLR)*, 2015. Cited on pages 100 and 115.

[KBS+23] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *International Conference On Learning Representations (ICLR)*, 2023. Cited on pages 11, 13, and 67.

[KCS+17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *Arxiv Preprint Arxiv:1705.06950*, 2017. Cited on pages 127, 128, 160, 203, and 212.

[KDGM+19] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 155, 160, 161, 162, and 163.

[KGS16] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2016. Cited on page 37.

[KH09] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. Cited on pages 68, 74, 90, 99, 179, and 194.

[KHP+20] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, SungJu Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 99, 100, 101, and 102.

[KHSM22] Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on pages 23 and 172.

[KJG+11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2011. Cited on page 127.

[KJYFF11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2011. Cited on page 194.

[KK18] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. In *International Conference On Learning Representations (ICLR)*, 2018. Cited on page 172.

[KKS21a] A Kukleva, H Kuehne, and B Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. in 2021 ieee. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 172, 177, 179, and 180.

[KKS21b] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 12, 14, and 155.

[KKSG19] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on pages 10, 13, 31, and 47.

[KLX+20] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on page 69.

[KMHK20] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 18, 84, 86, 91, 92, and 103.

[KMRW14] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2014. Cited on page 19.

[KNZD19] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on page 205.

[Kot22] Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Arxiv Preprint Arxiv:2206.04041*, 2022. Cited on page 8.

[KPR+17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings Of The National Academy Of Sciences*, 2017. Cited on pages 7, 23, 159, and 172.

[KR21] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 25.

[KRG17] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision And Image Understanding*, 2017. Cited on page 33.

[KRM+23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 26.

[KS20] Andrew Kae and Yale Song. Image to video domain adaptation using web supervision. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. Cited on pages 122 and 123.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. Cited on pages 1, 24, and 38.

[KSM+21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 5.

[KSP+20] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 16 and 69.

[KSR+24] Anna Kukleva, Fadime Sener, Edoardo Remelli, Bugra Tekin, Eric Sauser, Bernt Schiele, and Shugao Ma. X-mic: Cross-modal instance conditioning for egocentric action generalization. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2024. Cited on pages 13, 14, and 203.

[KTS+14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2014. Cited on pages 8, 128, and 161.

[KTW+20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 173, 174, 175, and 183.

[KTZ+21] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 123, 127, and 129.

[Kuh55] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. Cited on page 177.

[KW52] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals Of Mathematical Statistics*, 1952. Cited on page 195.

[KW18] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *Arxiv Preprint Arxiv:1810.09305*, 2018. Cited on pages 20, 112, 114, 116, and 141.

[KXR+20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on page 101.

[KZG+17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal Of Computer Vision (IJCV)*, 2017. Cited on pages 113 and 143.

[KZN17] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on page 32.

[LA17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference On Learning Representations (ICLR)*, 2017. Cited on pages 5, 18, 83, 84, 86, 103, and 188.

[LANZ19] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *The British Machine Vision Conference (BMVC)*, 2019. Cited on page 107.

[LARC21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. Cited on page 26.

[LAV21] Iro Laina, Yuki M Asano, and Andrea Vedaldi. Measuring the interpretability of unsupervised representations via quantized reversed probing. In *International Conference On Learning Representations (ICLR)*, 2021. Cited on pages 16 and 69.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. Cited on page 1.

[LBPL19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 20.

[LCC+20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. Cited on page 19.

[LCL+20] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 22.

[LCY+22] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 174.

[LDF+20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2020. Cited on page 19.

[Lee13] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference On Machine Learning (ICML)*, 2013. Cited on pages 18, 19, 83, and 188.

[LFV+17] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on page 33.

[LFV20] Iro Laina, Ruth Fong, and Andrea Vedaldi. Quantifying learnability and describability of visual concepts emerging in representation learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 16 and 69.

[LGC+22] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 23.

[LGZ+22] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 27 and 205.

[LH17a] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2017. Cited on pages 22, 159, 166, and 172.

[LH17b] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *Arxiv Preprint Arxiv:1711.05101*, 2017. Cited on page 56.

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference On Learning Representations (ICLR)*, 2019. Cited on page 115.

[LHGM21] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 69.

[LHSY17a] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2017. Cited on page 17.

[LHSY17b] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequence. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2017. Cited on page 38.

[LHW+20] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *ACM Multimedia*, 2020. Cited on pages 123 and 129.

[Liu23] Yaoyao Liu. Learning from imperfect data: Incremental learning and few-shot learning. 2023. Cited on page 21.

[LJS+20] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *Arxiv Preprint Arxiv:2002.06353*, 2020. Cited on pages 19 and 108.

[LJZ+22] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. Cited on pages 107, 109, 113, 115, 138, 142, and 143.

[LKP+23] Wei Lin, Anna Kukleva, Horst Possegger, Hilde Kuehne, and Horst Bischof. Taec: Unsupervised action segmentation with temporal-aware embedding and clustering. *Ceur Workshop Proceedings*, 2023. Cited on pages 14 and 33.

[LKS+22] Wei Lin, Anna Kukleva, Kunyang Sun, Horst Possegger, Hilde Kuehne, and Horst Bischof. Cycda: Unsupervised cycle domain adaptation to learn from image to video. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 11, 14, and 121.

[LLK07] Benjamin Laxton, Jongwoo Lim, and David Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2007. Cited on page 32.

[LLLS17] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations (ICLR)*, 2017. Cited on page 25.

[LLR18] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on pages 205 and 210.

[LLSH23a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference On Machine Learning (ICML)*, 2023. Cited on pages 5 and 19.

[LLSH23b] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference On Machine Learning (ICML)*, 2023. Cited on page 20.

[LLSL19] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 159 and 166.

[LLWJ18] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. *Advances In Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 25.

[LLX+19] Tiange Luo, Aoxue Li, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on page 162.

[LLXH22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference On Machine Learning (ICML)*, 2022. Cited on pages 5, 19, 26, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 118, 138, 143, and 145.

[LLZ+21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 107.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference On Computer Vision (ECCV)*, 2014. Cited on pages 83, 113, and 143.

[LMH+18] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference On Pattern Recognition (ICPR)*, 2018. Cited on pages 23 and 172.

[LMRS19] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 22.

[LMS17] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 3 and 17.

[LMZ+19] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 69.

[LOW+20] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International Journal Of Computer Vision (IJCV)*, 2020. Cited on pages 21, 122, and 123.

[LPB+22] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 20, 141, and 144.

[LRC+23] Vladislav Lialin, Stephen Rawls, David Chan, Shalini Ghosh, Anna Rumshisky, and Wael Hamza. Scalable and accurate self-supervised multimodal representation learning without aligned video and text data. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. Cited on page 20.

[LSL+20] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 159.

[LT21] Jun Li and Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 31 and 33.

[LWHL19] Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. Deep metric transfer for label propagation with limited annotated data. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on page 19.

[LWL21] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 126.

[LWOL20] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 9.

[LWS+18] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 2018. Cited on page 129.

[LWS+22] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. Cited on page 205.

[LWY+23] Chaomeng Lu, Xufeng Wang, Aimin Yang, Yikai Liu, and Ziao Dong. A few-shot based model-agnostic meta-learning for intrusion detection in security of internet of things. *IEEE Internet Of Things Journal*, 2023. Cited on page 22.

[LWZK17] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Attention transfer from web images for video recognition. In *ACM Multimedia*, 2017. Cited on pages 21, 121, 123, 127, and 128.

[LXX+22] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 20, 107, and 108.

[LYF+23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2023. Cited on page 9.

[LYZ+23] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2023. Cited on page 69.

[LZL+18] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 19.

[LZLF20] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 22.

[MAS+20] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 20, 107, 136, and 144.

[MBZ+17] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 2017. Cited on pages 21, 122, and 127.

[MC89] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology Of Learning And Motivation*. 1989. Cited on pages 23 and 172.

[MC21] Atefeh Mahdavi and Marco Carvalho. A survey on open set recognition. In *2021 IEEE Fourth International Conference On Artificial Intelligence And Knowledge Engineering (AIKE)*, 2021. Cited on page 24.

[MCN+23] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on pages 4 and 223.

[MD20] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 123, 127, and 129.

[MFD19] Davide Moltisanti, Sanja Fidler, and Dima Damen. Action recognition from single timestamp supervision in untrimmed videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 33.

[MGB21] Sudhanshu Mittal, Silvio Galesso, and Thomas Brox. Essentials for class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on page 168.

[MGL+18] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 2018. Cited on page 3.

[MKL+24] Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Advances In Neural Information Processing Systems (NeurIPS)*, 2024. Cited on page 5.

[MLNM19] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. Cited on pages 112, 114, and 116.

[MM20] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 3.

[MMKI18] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2018. Cited on pages 83 and 188.

[MNXA21] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 25 and 26.

[MR] Oded Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer. Cited on page 188.

[MRA+23] Miranda X Morris, Aashish Rajesh, Malke Asaad, Abbas Hassan, Rakan Saadoun, and Charles E Butler. Deep learning applications in surgery: Current uses and future directions. *The American Surgeon*, 2023. Cited on page 1.

[MRRB+23] Hossein Mohammad-Rahimi, Rata Rokhshad, Sompop Bencharit, Joachim Krois, and Falk Schwendicke. Deep learning: A primer for dentists and dental researchers. *Journal Of Dentistry*, 2023. Cited on page 1.

[MSMD21] Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on page 9.

[MSR21] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2021. Cited on pages 7, 23, and 172.

[MSWL23] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal Of Computer Vision (IJCV)*, 2023. Cited on page 1.

[MZA+19] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 2, 3, 5, 19, 20, 21, 107, 108, 112, 113, 114, 136, 137, 140, 141, 142, 145, and 222.

[MZH16] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Unsupervised learning using sequential verification for action recognition. In *European Conference on Computer Vision (ECCV)*, 2016. Cited on page 17.

[NALV18] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 16.

[NBMS21] Alexander Newman, Yuen Lam Bavik, Matthew Mount, and Bo Shao. Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology*, 2021. Cited on page 1.

[NBZA21] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *EEE/CVF International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 49 and 53.

[Nes83] Yu Nesterov. A method of solving a convex programming problem with convergence rate $o(k^2)$. *Doklady Akademii Nauk*, 1983. Cited on page 18.

[NF16] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference On Computer Vision (ECCV)*, 2016. Cited on pages 3 and 16.

[NH10] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference On Machine Learning (ICML)*, 2010. Cited on page 195.

[NOF+18] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 25.

[NSS+22] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 6, 19, 113, 116, 141, 142, 143, and 144.

[NWC+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *In NeurIPS Workshop On Deep Learning And Unsupervised Feature Learning*, 2011. Cited on pages 90 and 194.

[NXP+22] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *Arxiv Preprint Arxiv:2201.10005*, 2022. Cited on page 20.

[NZ08] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference On Computer Vision, Graphics & Image Processing*, 2008. Cited on page 194.

[Ode16] Augustus Odena. Semi-supervised learning with generative adversarial networks. *Arxiv Preprint Arxiv:1606.01583*, 2016. Cited on page 19.

[ODM+23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Arxiv Preprint Arxiv:2304.07193*, 2023. Cited on pages 2, 16, 208, 211, 212, and 213.

[OKB11] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances In Neural Information Processing Systems (NeurIPS)*, 2011. Cited on pages 113 and 143.

[OKK22] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 100.

[OLB+23] Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Homin Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: A survey on image-based generative and discriminative training. *Transactions on Machine Learning*, 2023. Cited on pages 3 and 15.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *Arxiv Preprint Arxiv:1807.03748*, 2018. Cited on pages 16, 18, 69, 70, 107, 110, 138, 173, 174, and 175.

[OOR+18] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2018. Cited on pages 90, 99, 100, 101, 188, and 189.

[OP19] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on pages 9 and 25.

[Ope23] OpenAI. Gpt-4 technical report, 2023. Cited on page 6.

[PAD18] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances In Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 25.

[PCA+21] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 205.

[PCAN20] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2020. Cited on pages 123 and 129.

[PCJ21] Shu Liu Pengguang Chen and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 16.

[PDB+] Z Peng, L Dong, H Bao, Q Ye, and F Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arxiv 2022. *arXiv preprint arXiv:2208.06366*. Cited on page 17.

[PHA+21] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference On Learning Representations (ICLR)*, 2021. Cited on page 113.

[PKH+22] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations (ICLR)*, 2022. Cited on page 208.

[PMJ+20] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 25.

[PP19] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 25.

[PQOBTM21] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference*, 2021. Cited on pages 107, 113, and 143.

[PSN+23] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on page 205.

[PTD20] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. gdumb: A simple approach that questions our progress in continual learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 24.

[PXDL20] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 18 and 83.

[PYJS22] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 189.

[PZW+22] Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C Lovell. Few-shot class-incremental learning from an open-set perspective. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on pages 8 and 23.

[QBL18] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 21, 22, 155, 161, 162, 163, 164, 165, 180, and 181.

[QMG+21] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 17.

[QZCH21] Yuanyuan Qing, Yijie Zeng, Qi Cao, and Guang-Bin Huang. End-to-end novel visual categories learning via auxiliary self-supervision. *Neural Networks*, 2021. Cited on page 25.

[RBBNZM21] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *Arxiv Preprint Arxiv:2104.10972*, 2021. Cited on page 56.

[RBH+15] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2015. Cited on page 18.

[RBH+21] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021. Cited on pages 20 and 144.

[RBL+21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 222.

[RCSJ20] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 16 and 69.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal Of Computer Vision (IJCV)*, 2015. Cited on pages 83 and 179.

[Ree01] William J Reed. The pareto, zipf and other power laws. *Economics Letters*, 2001. Cited on pages 68 and 69.

[REH+20] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Semi-supervised learning with scarce annotations. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 18 and 89.

[RFKL19] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 22.

[RKG17] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 32, 33, and 40.

[RKG18] Alexander Richard, Hilde Kuehne, and Juergen Gall. Temporal action labeling using action sets. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 33.

[RKH+20] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml : An incremental task-agnostic meta-learning approach. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 159.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference On Machine Learning (ICML)*, 2021. Cited on pages 2, 5, 19, 26, 106, 107, 108, 109, 114, 118, 136, 138, and 203.

[RKIG18] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 32, 33, and 41.

[RKM+23a] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 26.

[RkM+23b] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 223.

[RKP18] Shafin Rahman, Salman Khan, and Fatih Porikli. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions On Image Processing*, 2018. Cited on page 25.

[RKSL17a] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. Icarl: Incremental classifier and representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 22 and 172.

[RKSL17b] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: incremental classifier and representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 159, 160, 163, 164, 165, 166, and 168.

[RKX+23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference On Machine Learning (ICML)*, 2023. Cited on page 136.

[RL16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference On Learning Representations (ICLR)*, 2016. Cited on pages 22, 90, and 92.

[RLFZ19] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. Cited on pages 155, 156, 160, 161, 162, and 168.

[RPY+22] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. *Arxiv Preprint Arxiv:2210.04142*, 2022. Cited on page 16.

[RRBV19] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference On Learning Representations (ICLR)*, 2019. Cited on page 159.

[RRS] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In *The German Conference on Pattern Recognition (GCPR)*. Cited on pages 111, 112, 137, 140, and 142.

[RRS+18] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations (ICLR)*, 2018. Cited on page 22.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. Cited on page 20.

[RSTY22] Rahul Rahaman, Dipika Singhania, Alexandre Thiery, and Angela Yao. A generalized and robust framework for timestamp supervision in temporal action segmentation. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 33.

[RSY+21] Joshua David Robinson, Li Sun, Ke Yu, kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 69, 70, 71, 72, and 73.

[RTMFF15] Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. Learning temporal embeddings for complex video analysis. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2015. Cited on pages 18 and 38.

[RTR+18] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference On Learning Representations (ICLR)*, 2018. Cited on page 160.

[RVG+19] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference On Learning Representations (ICLR)*, 2019. Cited on page 9.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *Openai Blog*, 2019. Cited on pages 20 and 107.

[RZ11] Gabriel Radvansky and Jeffrey Zacks. Event perception. *Wiley Interdisciplinary Reviews. Cognitive Science*, 2011. Cited on page 1.

[SBL+20] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 5, 18, 19, 26, 83, 84, 86, 87, 88, 90, 91, 92, 93, 100, 101, 102, 103, 188, 191, and 195.

[Sch90] Robert E Schapire. The strength of weak learnability. *Machine Learning*, 1990. Cited on page 189.

[SCL+18] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference On Machine Learning (ICML)*, 2018. Cited on pages 23 and 172.

[SCN+20] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *Arxiv Preprint Arxiv:2010.10864*, 2020. Cited on page 57.

[SCR+22] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 16, 19, 20, 107, 110, 112, 113, 114, 115, 144, and 145.

[SCS+] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. Cited on page 205.

[Scu65] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions On Information Theory*, 1965. Cited on page 18.

[SDGS18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. Cited on pages 19 and 116.

[SES+19] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 161.

[SFH17] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2017. Cited on page 85.

[SGSF20] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 210.

[SHG+22] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 2.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014. Cited on page 18.

[SJT16] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2016. Cited on pages 5, 18, 84, 86, and 103.

[SKH+24] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale. In *European Conference On Computer Vision (ECCV)*, 2024. Cited on pages 12, 14, and 135.

[SKL+21] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2021. Cited on page 5.

[SKS21] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 26, 188, 189, 190, 191, 192, 193, 194, 195, 196, and 199.

[SKSK23] Nina Shvetsova, Anna Kukleva, Bernt Schiele, and Hilde Kuehne. In-style: Bridging text and uncurated videos with style transfer for text-video retrieval. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on pages 11, 13, and 105.

[SLS+20] Jonathan C Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A Ross. Learning video representations from textual web supervision. *Arxiv Preprint Arxiv:2007.14937*, 2020. Cited on pages 19 and 136.

[SM13] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Ubicomp*, 2013. Cited on pages 32 and 37.

[SM19] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference On Machine Learning (ICML)*, 2019. Cited on page 205.

[SM21] Jong-Chyi Su and Subhransu Maji. The semi-supervised inaturalist-aves challenge at fgvc7 workshop. *arXiv preprint arXiv:2103.06937*, 2021. Cited on page 99.

[SMS+21] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 33.

[SMV+19] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 20 and 49.

[SPA+19] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference On Machine Learning (ICML)*, 2019. Cited on pages 18 and 69.

[SPK+23] Nina Shvetsova, Felix Petersen, Anna Kukleva, Bernt Schiele, and Hilde Kuehne. Learning by sorting: Self-supervised learning with group ordering constraints. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on page 14.

[SRLI+20] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. Cited on page 26.

[SRSB13] Walter Scheirer, Anderson Rocha, Archana Sapkota, and Terrance Boult. Toward open set recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2013. Cited on page 24.

[SRY22] Dipika Singhania, Rahul Rahaman, and Angela Yao. Iterative contrast-classify for semi-supervised temporal action segmentation. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2022. Cited on page 33.

[SS21] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 189.

[SSP+21] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 123, 127, 128, and 129.

[SSS+20] Xiahan Shi, Leonard Salewski, Martin Schiegg, Zeynep Akata, and Max Welling. Relational generalized few-shot learning. In *The British Machine Vision Conference (BMVC)*, 2020. Cited on pages 155, 160, 161, 162, and 168.

[SSSK21] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 2021. Cited on page 3.

[SSSN15] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM Multimedia*, 2015. Cited on pages 21 and 123.

[SSZ17] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. Cited on pages 7, 21, 22, 23, 161, 163, 164, 165, and 172.

[SVB+21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Arxiv Preprint Arxiv:2111.02114*, 2021. Cited on pages 5, 113, 114, 136, and 143.

[SWUH18] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 129.

[SXL18] Lei Shu, Hu Xu, and Bing Liu. Unseen class discovery in open-world classification. *Arxiv Preprint Arxiv:1801.05609*, 2018. Cited on page 9.

[SY18] Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 32, 33, 34, 35, 37, 38, 39, 41, and 43.

[SY21] Abhilash Reddy Shankarampeta and Koichiro Yamauchi. Few-shot class incremental learning with generative feature replay. In *The International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2021. Cited on page 23.

[SYZ+18] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on page 22.

[SZC+20] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 19 and 20.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. Cited on pages 127, 160, 162, and 212.

[TAB+23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *Arxiv Preprint Arxiv:2312.11805*, 2023. Cited on page 1.

[TB19] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. Cited on page 20.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2015. Cited on page 161.

[TCG21] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 4.

[TGZ+23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center For Research On Foundation Models*, 2023. Cited on page 20.

[THC+20] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 7, 23, 24, 155, 157, 160, 165, 172, and 179.

[THvdO21] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 69.

[TKI20] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 75.

[TKM+24] Julian Tanke, Oh-Hun Kwon, Felix B Mueller, Andreas Doering, and Juergen Gall. Humans in kitchens: A dataset for multi-person human motion forecasting with scene context. *Advances In Neural Information Processing Systems (NeurIPS)*, 2024. Cited on page 1.

[TLI+23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *Arxiv Preprint Arxiv:2302.13971*, 2023. Cited on pages 1, 6, and 142.

[TLL+24] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 2024. Cited on page 7.

[TSP+20] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 78.

[TSSW22] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 107.

[TV17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2017. Cited on pages 18, 83, 90, 91, and 188.

[TWK+20] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 7, 24, and 158.

[TWSM20] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference On Learning Representations (ICLR)*, 2020. Cited on pages 18 and 69.

[TZIC16] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 2016. Cited on page 127.

[VBL+16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. Cited on pages 22 and 160.

[VCS+15] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2015. Cited on page 110.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. Cited on page 130.

[VGVGVG21] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 4 and 16.

[VHVZ22] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022. Cited on page 25.

[VLK+19] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. Cited on pages 18 and 83.

[VSK+21] Rosaura G VidalMata, Walter J Scheirer, Anna Kukleva, David Cox, and Hilde Kuehne. Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on pages 14 and 31.

[was23] Vita-clip: Video and text adaptive clip via multimodal prompting. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on pages 27, 205, 212, and 213.

[WBD+23] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: beit pretraining for vision and vision-language tasks. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 17.

[WBH19] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 2019. Cited on page 188.

[WBW+11] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. Cited on page 179.

[WC19] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2019. Cited on page 86.

[WCF+22] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, Heli Qi, Zhen Wu, Yu-Feng Li, Satoshi Nakamura, Wei Ye, Marios Savvides, Bhiksha Raj, Takahiro Shinozaki, Bernt Schiele, Jindong Wang, Xing Xie, and Yue Zhang. Usb: A unified semi-supervised learning benchmark for classification. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. Cited on page 188.

[WCH+23] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference On Learning Representations (ICLR)*, 2023. Cited on page 19.

[WCL+22] Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, and Charless Fowlkes. Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. Cited on pages 31 and 33.

[WCW+19] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on pages 22, 24, 157, 159, 166, 172, and 177.

[WDH+23] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 4.

[WFX+22] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 3.

[WG15] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2015. Cited on pages 17 and 18.

[WGHH18] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 21, 22, and 156.

[WHL+23] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference On Learning Representations (ICLR)*, 2023. Cited on pages 6 and 20.

[WI20] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference On Machine Learning (ICML)*, 2020. Cited on pages 69 and 70.

[WJE19] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 17.

[WKZ+23] Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *IEEE Transactions on Image Processing*, 2023. Cited on page 16.

[WL21a] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 69, 70, 71, and 79.

[WL21b] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 69.

[WLF+23] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 107.

[WLM+22] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 205.

[WLYL23] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2023. Cited on page 9.

[WLZF18] Donglai Wei, Joseph Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 17 and 18.

[WRH17] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances In Neural Information Processing Systems (NeurIPS)*, 2017. Cited on page 69.

[WS13] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2013. Cited on pages 18 and 38.

[WSL+20] Fangyu Wu, Jeremy S Smith, Wenjin Lu, Chaoyi Pang, and Bailing Zhang. Attentive prototype few-shot learning with capsule network-based embedding. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 22.

[WSM+21] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 99, 100, 101, and 102.

[WSS21] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 121.

[WWC+19] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 112 and 114.

[WWW+21] Guangrun Wang, Keze Wang, Guangcong Wang, Philip HS Torr, and Liang Lin. Solving inefficiency of self-supervised representation learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on pages 16, 18, and 69.

[WXLVG17] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on pages 33 and 121.

[WXSL18] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2018. Cited on pages 16 and 70.

[WXY+22] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Openauc: Towards auc-oriented open-set recognition. *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. Cited on page 25.

[WYY+22] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference On Learning Representations (ICLR)*, 2022. Cited on page 5.

[WZW+19] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 2019. Cited on page 25.

[WZWL21] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 47.

[WZY21] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 107.

[XANS22] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *Arxiv Preprint Arxiv:2206.09852*, 2022. Cited on page 205.

[XDH+20] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. Cited on pages 18, 19, 83, 84, 86, 91, 103, and 188.

[XGF16] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference On Machine Learning (ICML)*, 2016. Cited on pages 25 and 26.

[XHZ+22] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2022. Cited on pages 6, 20, 21, and 136.

[Xia20] Yongqin Xian. Learning from limited labeled data-zero-shot and few-shot learning. 2020. Cited on page 21.

[XKD+20] Yongqin Xian, Bruno Korbar, Matthijs Douze, Bernt Schiele, Zeynep Akata, and Lorenzo Torresani. Generalized many-way few-shot video classification. In *ECCV 2020 Workshops*, 2020. Cited on pages 160 and 162.

[XLG+20] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *Arxiv Preprint Arxiv:2001.08740*, 2020. Cited on page 205.

[XLSA18] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2018. Cited on page 8.

[XMYR16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on pages 19, 112, 137, 140, and 142.

[XSH+18] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 114.

[XSY+21] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 5.

[XSZ12] Ye Xu, Furao Shen, and Jinxi Zhao. An incremental learning vector quantization algorithm for pattern classification. *Neural Computing And Applications*, 2012. Cited on pages 163, 164, and 165.

[XXZ+19] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 17.

[YBZ+22] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 16.

[YGG17] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *Arxiv Preprint Arxiv:1708.03888*, 2017. Cited on page 179.

[YGX22] Yijun Yang, Ruiyuan Gao, and Qiang Xu. Out-of-distribution detection with semantic mismatch under masking. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 25.

[YHH+22] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision (ECCV)*, 2022. Cited on pages 16, 18, and 69.

[YHO+19] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on page 22.

[YHZS20] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on page 22.

[YIIA20] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on pages 188, 189, 190, 191, 192, and 199.

[YJ06] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2006. Cited on page 22.

[YJK+11] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2011. Cited on page 127.

[YKK18] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 107.

[YPB16] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2016. Cited on page 16.

[YSK+19] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 25.

[YSKX22] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions On Knowledge And Data Engineering*, 2022. Cited on page 5.

[YSL+18] Jufeng Yang, Xiaoxiao Sun, Yu-Kun Lai, Liang Zheng, and Ming-Ming Cheng. Recognition from web data: A progressive filtering approach. *Tip*, 2018. Cited on page 121.

[YSZ+15] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *Arxiv Preprint Arxiv:1506.03365*, 2015. Cited on page 194.

[YTL+20] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 163, 164, and 165.

[YWCD19] Feiwu Yu, Xinxiao Wu, Jialu Chen, and Lixin Duan. Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning. *Tip*, 2019. Cited on pages 21, 122, 123, 127, and 128.

[YWSD18] Feiwu Yu, Xinxiao Wu, Yuchao Sun, and Lixin Duan. Exploiting images for video recognition with hierarchical generative adversarial networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. Cited on pages 21, 122, 123, 127, and 128.

[YWV+22] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Arxiv Preprint Arxiv:2205.01917*, 2022. Cited on page 19.

[YX20] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 69.

[YXS+20] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 47 and 121.

[YYL+23] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *International Conference On Learning Representations (ICLR)*, 2023. Cited on pages 7, 8, 23, 24, 172, 179, and 180.

[YZLL21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *Arxiv Preprint Arxiv:2110.11334*, 2021. Cited on page 25.

[YZW+22] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *Arxiv Preprint Arxiv:2212.04979*, 2022. Cited on page 143.

[ZAC+19] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2019. Cited on page 205.

[ZCDLP18] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference On Learning Representations (ICLR)*, 2018. Cited on page 18.

[ZCG+18] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *Advances In Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 22.

[ZCS+23] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2023. Cited on pages 20 and 142.

[ZDJZ20] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference On Computer Vision (ECCV)*, 2020. Cited on page 126.

[ZFC21] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. Cited on page 26.

[ZFG22] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Self-supervised learning for unintentional action prediction. In *The German Conference on Pattern Recognition (GCPR)*, 2022. Cited on page 49.

[ZFK+21] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2021. Cited on pages 23 and 172.

[ZG22] Xiaojin Zhu and Andrew B Goldberg. *Introduction to semi-supervised learning*. Springer Nature, 2022. Cited on page 5.

[ZG23] Olga Zatsarynna and Juergen Gall. Action anticipation with goal consistency. In *2023 IEEE International Conference On Image Processing (ICIP)*, 2023. Cited on page 223.

[ZGZ+20] Yingying Zhang, Yuxin Gong, Haogang Zhu, Xiao Bai, and Wenzhong Tang. Multi-head enhanced self-attention network for novelty detection. *Pattern Recognition*, 2020. Cited on page 9.

[ZHT+16] Jianguang Zhang, Yahong Han, Jinhui Tang, Qinghua Hu, and Jianmin Jiang. Semi-supervised image-to-video adaptation for video action recognition. *IEEE Trans Cybern*, 2016. Cited on pages 21 and 123.

[ZIE16] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference On Computer Vision (ECCV)*, 2016. Cited on page 17.

[ZJM+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference On Machine Learning (ICML)*, 2021. Cited on page 16.

[ZJM+22] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning For Healthcare Conference*, 2022. Cited on page 203.

[ZK16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *The British Machine Vision Conference (BMVC)*, 2016. Cited on pages 91, 92, 94, 100, and 194.

[ZK22] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference On Computer Vision (ECCV)*, 2022. Cited on page 205.

[ZKHB22] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on page 5.

[ZLH+21] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 5, 20, 21, and 136.

[ZLL+17] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: A weakly-supervised deep learning framework for learning from web data. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2017. Cited on page 121.

[ZLX+23] Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on pages 23, 172, 179, and 180.

[ZLZ+23] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on page 9.

[ZMKG23] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2023. Cited on pages 3, 6, 20, 107, 205, and 215.

[ZOKB19] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on pages 20 and 89.

[ZSL+21] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 23, 24, 172, 179, and 180.

[ZTC+22] Y Zhong, H Tang, J Chen, J Peng, and Y-X Wang. Is self-supervised learning more robust than supervised learning? In *Proc ICML Workshop on Pre-training*, 2022. Cited on page 69.

[ZWBG21] Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. Temperature as uncertainty in contrastive learning. *Arxiv Preprint Arxiv:2110.04403*, 2021. Cited on page 79.

[ZWH+21] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances In Neural Information Processing Systems (NeurIPS)*, 2021. Cited on pages 5, 19, 85, 91, 92, 93, 101, 102, and 188.

[ZWY+22] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 23, 172, 179, and 180.

[ZXC18] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *The AAAI Conference On Artificial Intelligence (AAAI)*, 2018. Cited on pages 19, 111, 112, 136, 137, 140, and 142.

[ZXG+20] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 23 and 172.

[ZXL+20] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2020. Cited on pages 3 and 16.

[ZYKW18] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference On Computer Vision (ECCV)*, 2018. Cited on page 126.

[ZYL+19] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2019. Cited on page 126.

[ZYLL22a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 8, 9, 26, 27, 207, and 212.

[ZYLL22b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal Of Computer Vision (IJCV)*, 2022. Cited on pages 8, 26, 27, 204, 205, 207, and 212.

[ZYM+22] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions On Pattern Analysis And Machine Intelligence (TPAMI)*, 2022. Cited on pages 7 and 180.

[ZYW+22] Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *International Conference On Machine Learning (ICML)*, 2022. Cited on page 69.

[ZZ21] Guangtao Zheng and Aidong Zhang. Few-shot class-incremental learning with meta-learned class structures. In *2021 International Conference On Data Mining Workshops (ICDMW)*, 2021. Cited on page 23.

[ZZL+21a] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on page 9.

[ZZL+21b] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *IEEE/CVF International Conference On Computer Vision (ICCV)*, 2021. Cited on page 16.

[ZZLL22] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. Cited on page 23.

[ZZP+22] Chaoning Zhang, Kang Zhang, Trung X Pham, Axi Niu, Zhinan Qiao, Chang D Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2022. Cited on pages 16, 18, 69, and 79.

[ZZW+21] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, 2021. Cited on pages 23 and 172.