UNIVERSITÄT
DES
SAARLANDES

Machine Learning Solutions for High Dynamic Range Video Processing and Image Quality Evaluation

Dissertation zur Erlangung des Grades des Doktors der Ingenieurwissenschaften der Fakultät für Mathematik und Informatik der Universität des Saarlandes

Vorgelegt von
Uğur Çoğalan

Saarbrücken, 2024

Dean:	Prof. Dr. Roland Speicher
Date:	21.10.2024
Chair:	Prof. Dr. Thorsten Herfet
Reviewers:	Prof. Dr. Karol Myszkowski
	Prof. Dr. Hans-Peter Seidel
	Prof. Dr. Ana Serrano
Academic Assistant:	Dr. Colin Groth

Abstract

Conventional imaging sensors notoriously fall short in capturing real-world scenes by clamping image details in dark and bright scene regions. Longer exposures improve dark region depiction but often result in excessive blur for hand-held cameras, which is further aggravated for highly dynamic scenes. Conversely, shorter exposures reduce blur but at the expense of noisy images. In practice, it is often impossible to strike a balance between all those factors, and even for advanced computational photography techniques that today employ machine learning image enhancement techniques, it is difficult to obtain satisfactory, most importantly, veridical, non-hallucinated depiction.

Multi-exposure sensors enable different exposures for neighboring pixels, where such exposures can be freely adapted to the dynamic range of the captured scene. In this thesis, we observe that multi-exposure sensors enable the development of more robust learning-based techniques for denoising and motion blur removal because less noisy and less blurred neighboring reference pixels are readily available due to their different exposure. At the same time, filling gaps in the spatial domain for such differently exposed neighboring pixels is a trivial super-resolution task so that full-resolution differently exposed images can be reconstructed from a single multi-exposure shot. This, in turn, enables merging such exposures into a high-dynamic range (HDR) image. In the context of video, we demonstrate that motion blur in longer exposed pixels provides important information to improve the quality of optical flow computation, where even complex non-linear motion between two captured frames can be reconstructed. This enables high-quality video frame interpolation (VFI) to produce high-framerate videos that can be played in slow-motion mode, where HDR scenes can also be handled for the first time. Overall, our work demonstrates that alternative sensor designs, such as multi-exposure sensors, can often be better aligned with the strengths of machine-learning solutions, where additional information provided by such sensors simplifies more complex tasks such as HDR image reconstruction and VFI. In contrast, deficits of such sensors in terms of spatial resolution are easy to compensate for.

Perceptually meaningful image quality evaluation is an important aspect of computational imaging that warrants continuous progress. In this thesis, rather than devising a novel image quality metric, we seek to develop a coherent methodology to improve traditional metrics like PSNR and SSIM, as well as more recent learning-based LPIPS and DISTS. We achieve this by considering visual masking, an important characteristic of the human visual system that changes its sensitivity to distortions as a function of local image content. Our approach results in enhanced metrics that are more in line with human prediction both visually and quantitatively.

Zusammenfassung

Herkömmliche Bildsensoren sind bei der Erfassung realer Szenen bekanntermaßen unzureichend, da sie Bilddetails in dunklen und hellen Szenenbereichen festhalten. Längere Belichtungszeiten verbessern die Darstellung dunkler Bereiche, führen jedoch bei Handkameras häufig zu übermäßiger Unschärfe, die sich bei hochdynamischen Szenen noch verschlimmert. Umgekehrt reduzieren kürzere Belichtungszeiten die Unschärfe, allerdings auf Kosten verrauschter Bilder. In der Praxis ist es oft unmöglich, ein Gleichgewicht zwischen all diesen Faktoren zu finden, und selbst für fortgeschrittene Computerfotografietechniken, die heute Bildverbesserungstechniken des maschinellen Lernens nutzen, ist es schwierig, eine zufriedenstellende, vor allem wahrheitsgetreue, nicht halluzinierte Darstellung zu erhalten.

Mehrfachbelichtungssensoren ermöglichen unterschiedliche Belichtungen für benachbarte Pixel, wobei diese Belichtungen frei an den Dynamikbereich der aufgenommenen Szene angepasst werden können. In dieser Arbeit beobachten wir, dass Mehrfachbelichtungssensoren die Entwicklung robusterer lernbasierter Techniken zur Rauschunterdrückung und Entfernung von Bewegungsunschärfe ermöglichen, da weniger verrauschte und weniger unscharfe benachbarte Referenzpixel aufgrund ihrer unterschiedlichen Belichtung leicht verfügbar sind. Gleichzeitig ist das Füllen von Lücken im räumlichen Bereich für solche unterschiedlich belichteten benachbarten Pixel eine triviale Superauflösungsaufgabe, sodass unterschiedlich belichtete Bilder in voller Auflösung aus einer einzigen Mehrfachbelichtungsaufnahme rekonstruiert werden können. Dies wiederum ermöglicht die Zusammenführung solcher Aufnahmen zu einem High-Dynamic-Range-Bild (HDR). Im Zusammenhang mit Videos zeigen wir, dass Bewegungsunschärfe in länger belichteten Pixeln wichtige Informationen zur Verbesserung der Qualität der Berechnung des optischen Flusses liefert, bei der sogar komplexe nichtlineare Bewegungen zwischen zwei erfassten Bildern rekonstruiert werden können. Dies ermöglicht die hochwertige Video-Frame-Interpolation (VFI), um Videos mit hoher Bildrate zu produzieren, die im Zeitlupenmodus abgespielt werden können, wobei erstmals auch HDR-Szenen verarbeitet werden können. Insgesamt zeigt unsere Arbeit, dass alternative Sensordesigns, wie z. B. Mehrfachbelichtungssensoren, oft besser auf die Stärken von Lösungen für maschinelles Lernen abgestimmt werden können, bei denen zusätzliche Informationen, die von solchen Sensoren bereitgestellt werden, komplexere Aufgaben wie HDR-Bildrekonstruktion und VFI vereinfachen. Defizite solcher Sensoren in der Ortsauflösung lassen sich dagegen leicht ausgleichen.

Die wahrnehmungsbezogen aussagekräftige Bewertung der Bildqualität ist ein wichtiger Aspekt der computergestützten Bildgebung, der kontinuierliche Fortschritte erfordert. In dieser Dissertation wollen wir keine neuartige Bildqualitätsmetrik entwickeln, sondern eine kohärente Methodik entwickeln, um traditionelle Metriken wie PSNR und SSIM sowie neuere lernbasierte LPIPS und DISTS zu verbessern. Dies erreichen wir durch die Berücksichtigung der visuellen Maskierung, einer wichtigen Eigenschaft des menschlichen visuellen Systems, die ihre Empfindlichkeit gegenüber Verzerrungen in Abhängigkeit vom lokalen Bildinhalt ändert. Unser Ansatz führt zu verbesserten Metriken, die sowohl visuell als auch quantitativ besser mit den

menschlichen Vorhersagen übereinstimmen.

Acknowledgments

I would like to thank all the individuals who contributed to the completion of this thesis. I would also like to extend my deepest gratitude to my advisor, Karol Myszkowski, for his continuous support and advice and for sharing his in-depth knowledge. The success of my dissertation would not have been possible without Mojtaba Bemana's support and nurturing.

I would also like to extend my sincere thanks to Tobias Ritschel for his excellent guidance for the early project. In addition, I would like to express my thanks to Ahmet Oğuz Akyüz, who played a decisive role in shaping this thesis.

I was pleased to meet with amazing researchers, especially Krzysztof Wolski, Bin Chen, and Chao Wang in our research group. A special thanks to Krzysztof Wolski for the proofreading of this thesis. I am deeply indebted to all members of AG4 for enriching my academic and personal life and to Hans-Peter Seidel for providing the best possible opportunities.

Lastly, I would like to express my special thanks to my family members, especially my mother and father, for their positive attitude during this period.

Contents

Chapter 1

Introduction

This thesis explores optimizing digital image acquisition through specialized sensors and enhanced quality metrics by leveraging deep learning. Firstly, Section 1.1 motivates further development in the thesis, then is followed by the main contributions (Section 1.2), and lastly, Section 1.3 presents an overview of the whole thesis.

1.1 Motivation

Photographs play a sophisticated role in the modern world, realizing multiple purposes, such as opening windows into the past, reflecting the present, and expressing the natural world artistically. In the digital age, they have become more reachable through cell phones, which record moments of the daily lives of humans. Such an increase in the number of photographs leads to a development in their digital processing to enhance their visual quality. Photographs are acquired using digital camera sensors, which collect light for each pixel and convert it into electrical signals, later transformed into color images. Each step of the image acquisition pipeline, spanning from analog to digital processing, is prone to optimization, whether through a better design of an image sensor or image processing algorithms.

In this context, the field of computational photography has gained importance by proposing diverse techniques and algorithms to enhance, manipulate, or reconstruct the images captured by digital cameras. The proposed techniques overcome the limitations inherent to camera sensors and lenses, improving the image quality. One common use of computational photography is to reduce noise and blur in images. This involves correcting defects arising from the sensor's architecture or environmental conditions to improve captured images' quality, realism, and clarity. These distortions could severely affect the resulting images; therefore, the existing computational photography algorithms might not be able to handle them due to a lack of additional information.

High Dynamic Range (HDR) imaging is another essential field of computational photography that aims to create a single composite image with an expanded dynamic range because standard single-exposure sensors only capture a limited range of luminance values (LDR). This limitation results in restricted content, in which bright regions get saturated as a result of clamping while dark regions are exposed to severe noise. To overcome this issue traditionally, HDR imaging, which typically relies on merging multiple LDR inputs [Kang et al., 2003; Mangiat and Gibson, 2010; Gryaditskaya et al., 2015; Kalantari et al., 2013; Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019], has gained popularity in recent decades, and many display and editing tasks would greatly benefit from it [Reinhard et al., 2010]. However, the main limitation is the dynamic content of the scene, possibly caused by the camera movement and the motion in the scene, which results in motion blur.

Additionally, the alignment of the exposures gets harder due to missing information in each captured exposure that complicates optical flow computation as precise pixel matches between subsequent frames are compromised.

HDR imaging is crucial for capturing the different luminance levels and high contrast available in real-world scenes. For this reason, it has been used in many applications in computer graphics to create realistic and visually better world depictions. Furthermore, HDR images have been essential in autonomous driving due to better representation of the surroundings in any lighting conditions. While LDR images, as display-referred formats, are better suited for commonly used displays, they may not be optimal for emerging displays, including existing HDR displays. Recently, HDR displays, offering a wider color gamut and richer brightness levels [Reinhard et al., 2006] and providing end users with a more immersive viewing experience, have entered the market to meet the growing demand for displaying HDR content.

Considering the aforementioned problems, recently, new sensor architectures that allow one to configure different levels of exposure for different spatial patterns have appeared on the market. The ultimate design goal is to overcome the motion differences between temporally consecutive exposures by combining them into the same frame. This way, the read-out gap between the different exposures is eliminated, and the capturing time of the different exposures is aligned within the same frame. Programmable sensors with spatially varying exposures become an attractive choice for modern machine vision cameras and smartphones, e.g., CMOSIS CMV12000 [CMV12000, 2021], Sony's Quad Bayer [Sony, 2022], and Samsung's Tetracell/Nonacell [Samsung, 2022] technologies. Intrinsically, this new sensor design expands the dynamic range in a single shot by spatially interleaving different exposures across the sensor. On the other hand, it brings additional issues, such as noise and blur differences, because of the differing capturing times between the exposures, creating a new challenge: combining different exposures with different noise and blur behaviors into a coherent natural image.

Video frame interpolation (VFI) is another important task of computational photography extended into the temporal domain. The key VFI goal is to estimate the new frames between the existing primary frames in a video sequence. This increases the frame rate of the videos and results in a smoother pass in the large-motion regions. VFI enables many exciting applications, ranging from video compression and framerate up-conversion in TV broadcasting to artistic video effects such as speed ramps in professional cinematography. The performance of VFI methods is primarily affected by various factors such as scene lighting conditions, the magnitude and complexity of motion in the scene, the spatial extension of resulting motion blur, the presence of complex occlusions, or thin structures in the scene. Recent VFI methods [Sim et al., 2021; Reda et al., 2022] mainly rely on well-exposed frames in the captured video. Nevertheless, undesired under- and over-exposure effects might appear in the case of high dynamic range scenes captured using traditional single-exposure sensors. The resultant noise and intensity clamping can adversely affect the quality of VFI as finding the pixel correspondence between the frames becomes more ambiguous. Another major challenge is the large and non-uniform motion in the scene that can affect the quality of the resulting interpolated frames. In this context, a multi-exposure sensor provides short and long exposures for spatially interleaved pixel columns in a single shot. Importantly, while the exposure duration differs, the exposure completion is temporally aligned, which enables the recovery of two temporal samples of scene motion that are perfectly spatially registered at the sensor.

On the other hand, for the evaluation of the reconstructions, full-reference image quality metrics (FR-IQMs) are vital components that provide quantitative measures by comparing to its reference. The most commonly used FR-IQMs for evaluating image quality are the mean square error (MSE), mean absolute error (MAE), and PSNR. While these per-pixel metrics are easy to compute, they assess image quality regardless of spatial content, leading to false positive predictions. More specifically, a perceptual metric should provide correct visible error localization in case of uniform distortions such as Gaussian noise and motion blur, which means that noise distortion in high-contrast and textured regions should be penalized less while penalizing more motion blur distortion at high-contrast edges. Due to the estimation of errors perceptually correct in the deformed images, they are commonly used as a cost function in optimizing restoration tasks such as denoising and deblurring. Thus, optimization with perceptually better metrics leads to better reconstructions that are more in line with human perception.

Considering the proposed approaches, this thesis mainly focuses on improving the image acquisition of digital cameras by utilizing a specialized multi-exposure design and leveraging the power of deep learning methods. The deficiencies in the digital image sensors are the source for many different distortions occurring in the resulting images, which can be reduced with a modern sensor design that accounts for distortions within the spatial frame, providing more information. This ensures the images with (*i*) the better quality and (*ii*) expanded dynamic range that is complemented with (*iii*) the high-framerate video reconstruction taking the power of aligned multi exposures within the same frame. Regarding the judgment of the quality of the restoration of the images and guiding the training phase can be achieved with the (*iv*) enhanced FR-IQMs, which are enhancements over the existing metrics rather than a new quality metric.

1.2 Contributions

This thesis has two parts; firstly, it includes image and video enhancement through a new trend in multi-exposure sensor design. The second part proposes a series of enhanced FR-IQMs to provide quality measurements and optimize restoration tasks.

Image and Video Enhancement and HDR Reconstruction Digital images are distorted with different sources due to deficiencies within the sensor design, such as noise and motion blur. Previous LDR work learns to deblur and denoise supervised by pairs of clean and distorted videos [Nah et al., 2019; Tao et al., 2018]. Regrettably, capturing distorted sensor readings is time-consuming; as well, there is a lack of clean HDR videos that can help simulate the multiple exposures that the sensor provides. On the other hand, previous work made simplifying assumptions, such as Gaussian and Poisson noise, that are not applicable to real-world sensors. Moreover, HDR reconstruction with traditional sensors [Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019] relies on the incomplete information between the exposures to derive the optical flow that can be guided inconsistently. Considering all these issues, Chapter 3 presents the following contributions (based on Çoğalan et al. [2022]):

- a non-parametric noise modeling that is based on simple 2D histograms;
- a training dataset preparation relying only on the existing LDR video, where emulating required exposures, including exposure-specific motion blur, using high-speed footage;

- a better denoising and deblurring performance compared to the existing state-of-the-art methods;
- a compact solution that creates a sharp and clean HDR image from a single multi-exposure shot;
- a unique image-capturing framework that extends to HDR video reconstruction;
- a CNN network that merges the flows obtained for different exposures, providing a better sampling of flows for deriving missing exposures.

Video Frame Interpolation for High Dynamic Range Sequences Although recent methods [Reda et al., 2022; Sim et al., 2021] have shown progress in handling significant motion in VFI, they typically heavily rely on the motion linearity assumption that might not hold in practice. Explicit handling of non-linear motion becomes possible by processing more than two subsequent frames [Xu et al., 2019; Park et al., 2021]; however, temporal sampling might still be too sparse for reliable motion reconstruction. Motion blur due to low shutter speed and long exposure times further leads to spatial and temporal loss of image details. For this reason, handling blurry frames is typically treated as a challenge in the VFI task [Shen et al., 2020a; Zhang et al., 2020], while potentially, motion blur encodes continuous temporal information on the magnitude and direction of motion, particularly for large motion. Chapter 4 presents a technique to create high-framerate HDR videos leveraging the additional blur information inherent in multi-exposure sensors (based on Çoğalan et al. [2023]), and includes the following contributions:

- a compact machine learning solution for VFI that can handle HDR content and complex non-uniform motion, enabled by deriving two temporal samples of the scene motion for each frame by joint processing of short and long exposures as captured using a multi-exposure sensor;
- an adopted PWC-Net [Sun et al., 2018] architecture to estimate the motion flow from motion blur in the long exposure, which is combined with the sharp image content in the short exposure;
- a metric of motion complexity that provides insights into existing datasets used in the training of VFI methods;
- and evaluation of the performance of each method with different levels of motion non-linearity.

Enhanced Image Quality Measurement The most commonly used image quality metrics, such as PSNR and SSIM [Wang et al., 2004a], often fall short in predicting visual errors as perceived by the human observer. Modern metrics like LPIPS [Zhang et al., 2018b], DISTS [Ding et al., 2022], and DeepWSD [Liao et al., 2022] strive to assess the perceptual dissimilarity between images by comparing deep features extracted from classification networks [Simonyan and Zisserman, 2015]. However, they are designed to generate a single value per image pair, and the focus of underlying deep features is classification, which makes them less sensitive to some distortions, such as noise. Furthermore, the ground-truth data for visibility errors is hard to obtain, such that utilization of mean opinion scores (MOS) data with human subjects is crucial for supervision. Chapter 5 enhances the existing FR-IQMs penalizing the perceptually visible errors (based on Çoğalan et al. [2024]), and presents the following contributions:

- a methodology predicting the mask that acts as a per-pixel weight, applicable to most of the existing FR-IQMs;
- a lightweight CNN for generating the masks that demand minimal computational resources;
- a learned generic masking model that is capable of identifying various types of distortions and generalization to both deep features and spatial domain;
- a significant enhancement in the accuracy of quality prediction for FR-IQMs across various test MOS datasets;
- an estimated per-pixel error map that visually aligns more closely with human perception compared to the original FR-IQMs;
- two potential applications with the enhanced FR-IQMs as a loss function for training the state-of-the-art image denoising and motion deblurring.

1.3 Outline

This thesis is organized as follows. Chapter 2 reviews and discusses the relevant work to this thesis. Chapter 3 presents the distortion model that samples the sensor readings that are later used for the reconstruction of HDR images and videos. Chapter 4 defines the video frame interpolation technique that increases the framerate of the captured videos. Chapter 5 explains the enhancement methodology for the existing image quality metrics, which improves their performance in image restoration tasks such as denoising and deblurring. The conclusion and future work directions are presented in the Chapter 6.

Chapter 2

Previous Work

This chapter reviews the previous work relevant to this thesis. The image acquisition pipeline for digital cameras is introduced, and proposed algorithms for each stage are discussed in Section 2.1. Secondly, Section 2.2 discusses the related methods for creating HDR images and videos. Thirdly, related video frame interpolation methods are reviewed in Section 2.3. Finally, the image quality metrics are covered in Section 2.4.

2.1 Digital Image Acquisition Pipeline

Digital images undergo several processing steps before being displayed. Raw sensor data is not yet understandable by humans, so it needs additional processing on both the hardware and software sides. This section explains the required steps for converting the sensor data to meaningful for human vision RGB color images (Figure 2.1). The utilization of specific algorithms to process raw images directly impacts image quality.

2.1.1 Photon Collection

Digital camera sensors such as Charge Coupled Devices (CCD) and Complementary Metal-Oxide-Semiconductor (CMOS) function by collecting light photons during a specific period, which is known as exposure time. These photons enter the sensor through the lens and interact with photosites within it, which are sensitive to light. Subsequently, the sensor converts the collected photons into electrical voltage. Internal processes within the sensor cause different noise types that are explained in Section 2.1.4.

2.1.2 Bayer Filter & Demosaicking

A Bayer filter, also known as a Color Filter Array (CFA), is used to collect RGB color information for digital images. Instead of collecting the red, green, and blue color channels separately, the Bayer filter provides color information in 2×2 patterns, where half of each pattern is green due to the higher sensitivity of the human visual system to the electromagnetic wavelengths representing green color. The other half of the pattern is shared between the red and blue channels. However, Bayer filtering leads to resolution loss, so various methods have been developed to reconstruct the full-resolution color channels of red, green, and blue (known as demosaicking). Traditional methods typically rely on numerous forms of interpolation that are content-aware [Gunturk et al., 2002; Lu et al., 2010; Zhang et al., 2011a]. However, they struggle with the strong edges and textured regions that result in zippering and moire artifacts. For this reason, CNN-based methods [Gharbi et al., 2016; Kokkinos

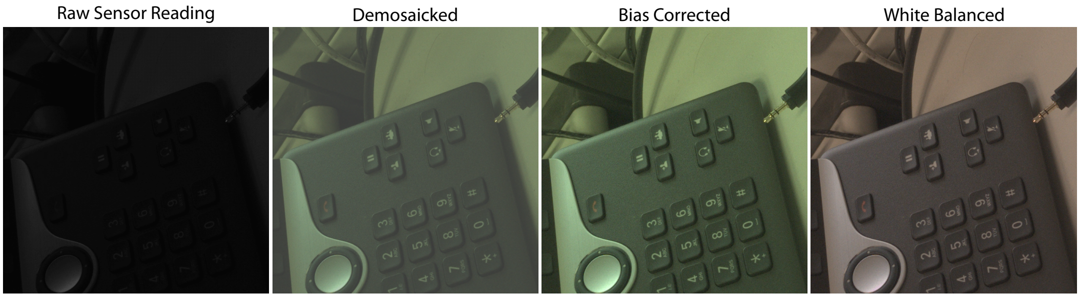


FIGURE 2.1: *Each step in the process of generating an image from raw sensor readings has a specific effect. The raw data is in a Bayer pixel format, meaning that each pixel contains information about only one color. To produce a full-color image, the data must undergo demosaicking to convert it into three color channels. However, when the camera applies a bias, it can create a black shadow over the image. This can be corrected with dark frame subtraction. After that, white balancing is applied to remove any global green casting and make the image ready for display.*

and Lefkimmiatis, 2019; Liu et al., 2020a] have been proposed to manage interpolation better, resulting in further improvements in demosaicking quality.

2.1.3 White Balancing

White balancing is an essential process that ensures the colors of the image look natural according to human perception. After years of evolution, the human visual system adapts to different illumination conditions, such as different times of the day or sky conditions, and can identify the white color. However, digital cameras are not part of this long evolution period; they can produce images with green, blue, or orange casts. On the other hand, most digital cameras have the auto white balancing option that provides the correct weightings of each color channel for different light conditions. However, in the case of multiple light sources, single weights per color channel may not be enough. Recently developed methods [Barron, 2015; Afifi and Brown, 2020] aim to enhance image quality by adjusting colors for different illumination conditions. The most recent method [Afifi et al., 2024] utilizes a multi-exposure sensor in the context of white balancing.

2.1.4 Sensor Noise

Noise types can be grouped as temporal and spatial noise depending on their characteristics [Janesick, 2001]. Temporal noise types, such as photon shot noise, dark current shot noise, and read-out noise, can be eliminated by averaging multiple frames. On the other hand, spatial noise is fixed and can be corrected together with dark frame subtraction and pixel area correction. Temporal noise types lead to flickering between each captured frame while spatial noise remains constant, but they are easy to correct with the careful calibration procedure. This section reviews the different noise types inherent to digital camera sensors. The temporal noise types are traditionally represented with a combination of parametric Poisson and Gaussian distributions. However, this thesis (Chapter 3) aims to model different types of noise with a non-parametric assumption-free solution that better corresponds to real-world sensors.

Photon shot noise The number of photons that the sensor collects in a given time interval follows the Poisson distribution. Random fluctuations in the measured number of photons are called photon shot noise, which is more dominant in low-light regions. Even with a careful electrical circuit design, photon shot noise is inevitable;

however, collecting more photons with longer exposure times helps to reduce the noise to some extent.

Dark current shot noise The randomness of the electrons in charge carriers causes dark current shot noise that follows a Poisson distribution. Environmental conditions such as temperature determine the effect of the dark current shot noise, and cooling the sensor reduces the random variations. The increase in the exposure time increases the effect of the dark current shot noise; therefore, shorter exposure times are suited to reduce this type of noise.

Read-out noise Other noise sources, such as white noise, reset noise, and quantization noise, are combined under the read-out noise, and a Gaussian distribution represents the combination of them. The reasons vary from voltage amplification to analog-to-digital conversion (ADC). These fluctuations are inherent to the sensor's electronics and are independent of the measured amount of photons.

Row noise At short exposures, more structured forms of noise can become important, one of them being row noise. This is not to be mistaken with fixed-pattern noise that is frequently spatially correlated but much easier to correct. In row noise, pixels do not change independently; rather, all pixels in a row change in correlation, i. e., the entire row is darkened or brightened. This is a well-known problem for CMOS sensors. It is typically caused by random noise that shifts the voltage level(s) in ADC, which, in turn, affects the ADC slope and results in an offset that is consistent for all pixels in the row. Sensor manufacturing tries to solve this problem by collecting on-the-fly statistics for surrounding rows to find voltage level offsets in ADC [Oten and Li, 2011], which is prone to errors due to variable image content.

Fixed pattern noise The electrical signals can go below zero due to the fluctuations in the analog-to-digital conversion phase. To prevent such issues, a positive offset is applied to every pixel. However, this offset can not be spatially uniformed because of the variations in the columns and pixels, which leads to fixed pattern noise. It is independent of the amount of light collected and depends on the exposure time and temperature of the sensor. The dark frame under zero light and identical conditions, such as the same exposure time and temperature, captures the internal bias of the sensor together with variations, and subtracting it from the raw measurements corrects the introduced fixed pattern noise.

Per-Pixel non-uniformity Even with a careful manufacturing process of the sensors, imperfections in the area of the pixels cause the measured number of photons to vary between each pixel. This leads to different illumination in the different parts of the resulting image. These differences could be identified easily, capturing a totally uniform region, and later, it is helpful for correcting the mismatched pixel areas.

2.1.5 Denoising and Deblurring

Imperfections in the sensors cause distortions in the captured images that impair the quality of the captured images as mentioned in Section 2.1.4. One of the central distortions is noise and blur, which have become the main focus of the image processing field for years. The first key challenge here is modeling the characteristics of sensor noise and blur that need to match the real sensor data. The second challenge is the correction of these distortions by estimating the missing information to enhance the quality of the images. This section reviews several algorithms for denoising and deblurring proposed in recent years.

Noise modeling Different factors affect the sensor's noise behavior depending on the sensor's imperfections and environmental conditions. Noise modeling of digital camera sensors is crucial in many aspects, such as understanding the sensor's quality, optimization, calibration, and correction. Classic solutions involve fitting Gaussian and Poisson [Healey and Kondepudy, 1994; Liu et al., 2007; Foi et al., 2008; Foi, 2009] or more involved [Plötz and Roth, 2017] distributions, sometimes under extreme conditions [Chen et al., 2018], to many pairs of clean and distorted images. Generative adversarial networks have recently been employed to map between different ISO levels, and the requirement for paired data is lifted [Bernardo Henz, 2021]. While parametric noise models are routinely used as mathematically tractable priors, this thesis (Chapter 3) uses more expressive non-parametric models, as the only need is to generate distorted training data.

Denoising Denoising aims to correct the pixel fluctuations without corrupting the image's content. Denoising has traditionally been performed directly on noisy images using state-of-the-art algorithms such as BM3D [Dabov et al., 2007], non-local means [Buades et al., 2005], and Nuclear Norms [Gu et al., 2014]. Most deep denoisers [Chen et al., 2018; Zhang et al., 2018a; Zhang et al., 2017; Burger et al., 2012; Mao et al., 2016; Chen et al., 2018a; Guo et al., 2019; Lefkimmiatis, 2018; Jia et al., 2019] are trained on pairs of noisy and clean images, while some work is trained without pairs [Ulyanov et al., 2018; Lehtinen et al., 2018; Krull et al., 2019; Laine et al., 2019; Krull et al., 2020; Batson and Royer, 2019; Quan et al., 2020; Moran et al., 2020; Xu et al., 2020], using GANs [Chen et al., 2018b] or self-supervision [Wu et al., 2020]. The usefulness of neural networks in denoising for real sensors has been disputed [Plötz and Roth, 2017; Chen et al., 2018].

Blur modeling Video obtained with a high-speed camera [Su et al., 2017; Nah et al., 2017; Nah et al., 2019] accurately represents the motion blur (MB) due to the presence of correct motion continuously in each frame. However, the video's frame per second (FPS) affects the quality of simulated motion blur, especially in large motion areas. Beam splitters [Zhong et al., 2020] enable motion blur synthesis for generating training data using gyroscope-acquired [Mustaniemi et al., 2020] or random [Mildenhall et al., 2018] motion.

Deblurring Non-blind deconvolution methods [Zoran and Weiss, 2011; Schuler et al., 2013; Sun et al., 2014; Schmidt et al., 2013; Xu et al., 2014; Cho et al., 2011; Whyte et al., 2010] restore sharp images given the blur kernel. Blind deconvolution methods attempt to derive the kernel based on various priors on either the sharp latent image or the blur kernel [Fergus et al., 2006; Levin et al., 2009; Xu and Jia, 2010; Michaeli and Irani, 2014; Gong et al., 2017; Sun et al., 2015; Chakrabarti, 2016]. Explicit kernel derivation can be avoided in end-to-end training, where the sharp image is derived directly [Nah et al., 2017; Tao et al., 2018], by self-supervision [Liu et al., 2020b] or adversarial training [Kupyn et al., 2018; Kupyn et al., 2019]. Video deblurring additionally capitalizes on inter-frame relationships while assuring temporal coherence of the result [Kim and Lee, 2015; Kim et al., 2017; Zhou et al., 2019; Zhong et al., 2020; Su et al., 2017]. Deblurring can be combined either with spatial [Zhang et al., 2019] or temporal [Purohit et al., 2019; Jin et al., 2018; Jin et al., 2019] super-resolution. The presence of noise, clamping, and multiple exposures, as in the condition of this thesis (Chapter 3), adds a further challenge. Methods such as [Pan et al., 2021] model general distortions using CycleGAN [Zhu et al., 2017] but have not been demonstrated to perform denoising.

2.1.6 Multi-Image Denoising

Several solutions have been proposed to capture multiple images of the same content to provide more information for ill-posed denoising.

Fixed-exposure burst photography Burst photography combines a handful of low-exposure frames into a high-quality LDR result using efficient hand-crafted solutions deployed in cellphones [Liu et al., 2014; Hasinoff et al., 2016; Liba et al., 2019; Liba et al., 2019], based on learning of recurrent architectures [Wieschollek et al., 2017], unordered sets [Aittala and Durand, 2018], or per-pixel filter kernels [Mildenhall et al., 2018]. The problem of read noise that accumulates from each contributing frame can be avoided in quanta burst photography that employs binary single-photon cameras to capture high-speed sequences [Ma et al., 2020].

Low/high exposure image pairs Short-exposure images are sharp but noisy, while long-exposure images are blurry but free of noise. Such exposure pairs have been used for non-uniform kernel deblurring [Yuan et al., 2007; Whyte et al., 2010]. Along a similar line, Mustaniemi et al. [2020] and Chang et al. [2022] jointly learn how to denoise and deblur exposure pairs supervised by synthetic training data. Although all these methods produce LDR output, this thesis (Chapter 3) aims for HDR creation.

2.2 HDR Imaging

HDR imaging aims to expand the dynamic range of images beyond the capacity of sensors. Digital cameras use imaging sensors that have limited capability to capture the full range of brightness in everyday scenes. As a result, captured images may have areas that are too bright and appear washed out, while dark areas may appear noisy or completely black, resulting in a loss of detail in the image. To eliminate such problems, a wide range of methodologies aim to handle dynamic range expansion, either employing specialized hardware or additional information with different requirements. This section reviews such methods depending on their assumptions beforehand.

2.2.1 Multi-Shot

Typical digital camera sensors can capture a wide range of luminances, but not within one shot. For this reason, an exposure sequence, i.e., time-sequential capture of one scene at different exposure settings, can be merged into a single HDR image [Mann and Picard, 1995; Mitsunaga and Nayar, 1999; Debevec and Malik, 1997] by recovering the camera response function (CRF) for linearization of the input images. Additionally, the following works [Robertson et al., 2003; Granados et al., 2010] aim to reduce the noise in the resulting HDR image. Alternative methods such as [Mertens et al., 2007; Prabhakar et al., 2017; Mustaniemi et al., 2020] create the LDR image directly instead of fusing the multiple exposures into an HDR image. However, these methods assume that the objects and the camera are static, meaning there is no occurring motion in the scene. This way, these methods do not need to handle the registration between the multiple exposures.

On the other hand, multi-exposure techniques are adopted for real-life conditions with dynamic content, which is more challenging to handle because of a lack of information possibly caused by under- and over-exposed regions in the captured LDR exposures. HDR reconstruction is typically performed by aligning the multiple exposures. As each exposure is employed in a sequence of reconstructed HDR frames, missing exposures must be aligned with moving content there, which can be achieved using optical flow [Kang et al., 2003] or homography registration followed by rank minimization where possible misalignments are treated as sparse outliers [Oh et al., 2015]. Even better for handling occlusions, non-rigid and fast motion are block- [Mangiat and Gibson, 2010; Gryaditskaya et al., 2015] and PatchMatch-based [Kalantari et al., 2013] motion estimation methods. Recently, neural networks have been employed for optical flow computation and exposure blending [Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019]. Non-flow-based networks have also been considered [Wu et al., 2018; Yan et al., 2020]. Yan et al. [2020] can handle large motions by exploiting the non-local correlation in input images. Chen et al. [2021] consider a coarse-to-fine neural network architecture to handle the alignment of the temporal exposures in the video sequence. Wu et al. [2018] use homography background registration followed by direct blending of the input exposures. The key problem for all those methods is alignment between subsequent frames at larger exposure ratios [Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019; Yan et al., 2020], or in the presence of massively saturated, meaning that either clamped or occluded regions. The alignment of the exposures can degrade the final reconstruction, and temporal artifacts such as popping and flickering can occur. This thesis (Chapter 3) demonstrates that such problems can be alleviated using an exposure ratio of 1:4 to achieve a significantly higher dynamic range.

2.2.2 Single-Shot

Capturing exposure sequences takes time, and their alignment is challenging, particularly for video. This can be alleviated by single-shot solutions relying on custom optics and sensors. The logarithmic response does not require any exposure control [Seger et al., 1999] but remains prone to noise in dark regions. Spatially-varying exposure (SVE) techniques place a fixed [Nayar and Mitsunaga, 2000; Schöberl et al., 2012; Schöberl et al., 2012; Serrano et al., 2016; Aguerrebere et al., 2014] or adaptive [Nayar and Branzoi, 2003; Nayar et al., 2004] mask of variable optical density in front of the sensor, but face problems with resolution and aliasing. Such problems can be reduced when a neural network is used for mask learning and HDR image reconstruction [Alghamdi et al., 2021]. Beam splitting preserves resolution with different exposures [Tocci et al., 2011; Aggarwal and Ahuja, 2001; Kronander et al., 2013] but requires involved optics. Dual-ISO sensors, e.g., Gpixel GMAX and some of the Canon EON sensors, enable varying analog signal gain for odd and even scanlines. Their key advantage is that variable blur between scanlines is avoided, as the exposition is fixed for the whole sensor. On the other hand, instead of collecting more photons in the long exposure and reducing noise this way, only a noisy short exposure is taken, and the long exposure is emulated by increasing ISO, which leads to further noise amplification. Therefore, denoising and deinterlacing are the critical challenges for processing dual-ISO frames [Hajisharif et al., 2014; Go et al., 2019], including data-driven solutions such as learned artifact dictionaries [Choi et al., 2017], and CNNs that rely on Gaussian noise models when synthesizing training data [Çoğalan and Akyüz, 2020]. Dual-gain sensors in high-end Canon and professional cinematographic Alexa (ARRI) cameras employ a similar idea but generate two full frames with different analog gains to improve the ratio of read noise to the signal in the high-gain image. Large photosites reduce the noise inherent to short exposures, which is needed to avoid highlight clipping.

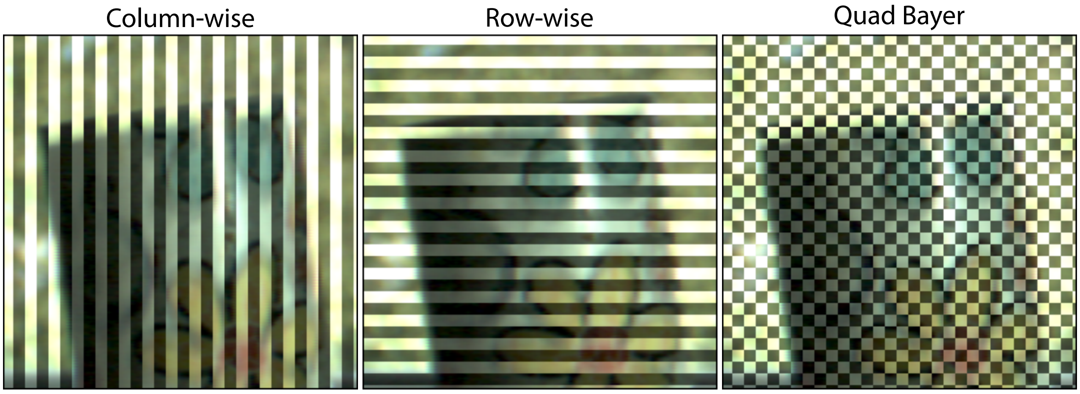


FIGURE 2.2: *Different existing multi-exposure sensor patterns. All patterns result in the same number of missing pixels for each exposure, regardless of whether the multiple exposures span column-wise, row-wise, or pixel-wise in Quad Bayer technology.*

Multi-Exposure Sensors

Multi-exposure CMOS sensors enable varying exposures for odd and even scanlines (some Aptina AR and Sony IMX sensors [IMX, accessed on Sept. 17, 2021]) or columns (CMOSIS CMV12000 [CMV12000, 2021]). The same approach has been applied to various sensors with different spatial patterns (Figure 2.2) in Sony's Quad Bayer [Sony, 2022] and Samsung's Tetracell/Nonacell [Samsung, 2022] technologies. These kinds of sensors are currently used in cell phones and recent machine vision cameras [Basler Dual Exposure, accessed on March 7, 2024]. The number of captured exposures varies between two to three depending on the design of the sensor. The aim is to combine different exposures into a single frame to skip the alignment of exposures. However, it brings additional problems of loss of resolution due to interleaved exposures within a single shot. Moreover, noise is a crucial problem in short exposures, while long exposures contain motion blur due to longer exposure time. In this context, Gu et al. [2010] perform flow-compensated interpolation for subimage deinterlacing to obtain differently exposed, full-resolution images. Cho et al. [2014] directly calibrate scanlines using bilateral filters followed by motion blur removal [Lenzen and Scherzer, 2011] and sharpening. Along similar lines, Heide et al. [2014] propose an end-to-end optimization, which jointly accounts for demosaicking, deinterlacing, denoising, and deconvolution. Lastly, An and Lee [2017] restore under- and over-exposed pixels using a CNN, but no results for real sensor data are demonstrated.

Without loss of generality, this thesis (Chapter 3) considers a specific sensor design where every even column is captured with a short exposure and every odd column with a long exposure [CMV12000, 2021]. This results in distortions that are specific to such kinds of sensors. One distortion type is the pixel noise within the image that no longer follows a single model but is correlated with exposure. Furthermore, different exposures lead to extra noise, which means that the short exposures have high noise but are not clamped, while the high exposures have less noise but suffer from clamping. Due to the design choice applied to such sensors, they suffer from increased levels of row noise, so orthogonal to the exposure layout, entire rows of pixels change coherently and differently for different exposures. Lastly, and most distant from other sensors, the additional exposure levels also lead to varying forms of motion blur. Not only does motion blur lead to spatially-varying blur, but this blur rapidly alternates between odd and even columns. Short exposures have low motion blur, while high exposures suffer from vital motion blur. Successful handling of such distortions guides HDR images that are sharp and noise-free.

An HDR image reconstructed using such sensors would provide not only a greater illumination range than general low dynamic range images but also better clarity of detail in highlights and dark regions, which can potentially improve the performance of many vision and graphics tasks such as object detection, depth estimation, scene segmentation, etc. Novel sensor designs, offering great flexibility of in-pixel processing, have been successfully used in video compressive imaging [Martel et al., 2020; Iliadis et al., 2020], depth from defocus [Martel et al., 2017], feature classification [Chen et al., 2017], HDR imaging [Martel et al., 2020] and motion deblurring [Nguyen et al., 2022]. The functionality of novel sensor designs in the context of HDR imaging and its possible applications have been comprehensively discussed in the survey [Wang and Yoon, 2022].

2.2.3 Tonemapping

HDR images are typically represented with 32-bit values that do not fit the dynamic range of the commonly used low dynamic range display devices. LDR displays are limited to the 8-bit range, meaning that HDR content with a wider range of luminance values can not be correctly displayed. To properly visualize a wider dynamic range of HDR images in LDR displays, the tonemapping technique is used to compress the HDR images' dynamic range while preserving the details and contrast. Global tonemapping operators [Reinhard et al., 2002; Drago et al., 2003; Kim and Kautz, 2008] consider the global statistics of the HDR images that are applied uniformly to the entire image. Although they are computationally cheaper, the resulting tonemapped images can lack contrast and details. On the other hand, other approaches [Ashikhmin, 2002; Durand and Dorsey, 2002; Fattal et al., 2002; Krawczyk et al., 2005] aim to locally preserve the contrast and details while compressing the HDR images. Apart from traditional approaches, modern methods [Rana et al., 2020; Wang et al., 2022a] utilize deep features to enhance the contrast of the tonemapped images realistically.

2.2.4 Inverse Tonemapping

Inverse tone mapping is a procedure to revert LDR images to their HDR representation. LDR images have a limited dynamic range clipped to 8 bits; therefore, conversion to HDR requires estimating or hallucinating the loss of information to represent them in higher bit depth. This can be challenging because either estimated regions do not match the original content or the reconstruction has artifacts in under or over-exposed regions. Although immense progress has been made recently based on CNNs and GANs [Marnerides et al., 2018; Endo et al., 2017; Eilertsen et al., 2017; Santos et al., 2020; Liu et al., 2020c; Wang et al., 2023] results do not yet match the quality of multi-exposure techniques or dedicated sensors.

2.3 Video Frame Interpolation

Video frame interpolation has been extensively studied and has gathered the attention of researchers. Interpolating missing frames between key frames is crucial for many applications, such as novel view interpolation and framerate conversion. This section examines previous methods that use different assumptions and input sequences in the context of frame interpolation.

2.3.1 Sharp Video Frame Interpolation

Most Video Frame Interpolation (VFI) techniques assume that the motion in the input video is uniform. However, there are a few methods explicitly designed without this assumption. Therefore, VFI techniques are categorized based on uniform and non-uniform assumptions. The methods in both these categories discard motion blur that can happen even within a single captured image due to possible longer exposure times.

Uniform Assumption

Video frame interpolation methods rely on the existing two keyframes that are naturally based on uniform motion assumption because they have no chance to track motion. Traditional methods such as [Werlberger et al., 2011; Yu et al., 2013] have considered optical flow to synthesize intra-frames utilizing conventional techniques. In the era of machine learning, the focus has shifted towards convolutional neural networks (CNNs). To this extent, SepConv [Niklaus et al., 2017] merges flow estimation and frame warping into a single convolution step. They predict spatially-varying 1D kernels and convolve with them input frames to interpolate new frames. SuperSlowMo [Jiang et al., 2018] uses bi-directional flows and an occlusion map to synthesize intermediate frames at arbitrary times. DAIN [Bao et al., 2019] utilizes additional interpolation kernels and depth maps for blending the input frames. A cycle consistency loss is introduced to learn frame interpolation with fewer training pairs [Liu et al., 2019] or without any supervision [Reda et al., 2019]. BMBC [Park et al., 2020] warps the input frames with a proposed bilateral motion model and combines them using learned dynamic blending filters. CAIN [Choi et al., 2020] uses a channel attention module to interpolate video frames without the need for estimation of motion. SoftSplat [Niklaus and Liu, 2020] proposes differentiable forward warping via softmax splatting and shows its benefits for VFI. AdaCoF [Lee et al., 2020] proposes a warping module in which a target pixel can refer to not only one but many pixels at any location in the reference. XVFI [Sim et al., 2021] presents a high-speed (1000fps) video dataset and proposes a multi-scale recursive approach to handle large motion in the scene. Recently, FILM [Reda et al., 2022] has introduced a unified framework that achieves superior results for large and complex motions by balancing the motion range distribution in the training dataset. Combining large and strongly non-uniform motions might lead to highly objectionable artifacts for all methods discussed here.

Non-Uniform Assumption

It is highly possible that non-uniform motion, such as rotatory motion, could be captured in the video sequences. This challenging case can directly affect the quality of the synthesized frames due to possible wrong positioning caused by the bi-directional warping of existing keyframes. For this reason, recent methods focused on handling non-uniform motion such that QVI [Xu et al., 2019] is one of the first video interpolation methods to model curvilinear motion with the quadratic equation using four temporal frames. Chi et al. [2020] extend QVI by introducing an additional cubic term that accounts for the change in acceleration. ABME [Park et al., 2021] handles the non-uniform motion in the scene by extending the BMBC [Park et al., 2020] for asymmetric bilateral motion between input frames. In all those methods, more than two consecutive frames are required to capture the non-uniform motion that might be challenging for large and complex motions, both because of temporal sampling deficits and overall reduced flow estimation accuracy. In this thesis (Chapter 4), capturing two exposures in a single frame using a multi-exposure sensor helps to increase the sampling rate twice, so the motion blur inherent to the longer exposure serves as an additional cue to the flow estimation that helps recover non-uniform motion.

2.3.2 Motion Flow Reconstruction From Motion Blur

A combination of longer exposure times and rapid motion in the scene or camera might lead to visible motion blur that typically is considered degradation and eliminated using dedicated image and video deblurring solutions. Extensive surveys on this topic [Koh et al., 2021; Zhang et al., 2022] have analyzed the recent solutions for video deblurring. This section focuses on deblurring solutions that explicitly recover intra-frame optical flow from motion blur. Earlier works [Rekleitis, 1995; Schoueri et al., 2009] assume global motion models that lead to spatially-invariant deblurring kernels. More advanced solutions support spatially-varying kernels that are approximated by linear motion [Hyun Kim and Mu Lee, 2014; Dai and Wu, 2008]. Gong et al. [2017] propose a deep-learning approach to handle heterogeneous blur; however, they simulate motion flows with a set of constrained flow magnitudes and directions to generate the training pairs. The following work [Argaw et al., 2021] alleviates this issue by deploying available synthetic and real scene blur datasets without any restrictive motion assumptions and estimating a dense optical flow directly from motion blur in the image. However, their estimation may be subject to ambiguity in predicting the correct direction of flow, which is crucial in the case of this thesis. Beyond restoring latent sharp images, a joint estimate of the 3D shape and motion is feasible, but highly motion-blurred images are required [Qiu et al., 2019; Rozumnyi et al., 2022]. While these methods aim to recover the motion flow from blur, they assume that the input blurry image is mostly well-exposed. However, longer exposure times lead to considerable saturated pixels in the long blurry exposure. This problem could be alleviated using the sharp short exposure that also bypasses the image deblurring task.

2.3.3 Joint Deblurring and Interpolation

Recent works demonstrate that joint deblurring and frame interpolation greatly improves the resulting VFI quality over an independent treatment of these tasks. One of the recent methods [Jin et al., 2019] adopts a joint optimization scheme to extract sharp keyframes within a frame by processing four consecutive blurry frames and then smoothly interpolating the in-between frame using the extracted keyframes. The following works [Shen et al., 2020a; Shen et al., 2020b] simultaneously remove the motion blur and interpolate the in-between frames by employing a recurrent pyramid framework to aggregate the temporal information efficiently. Another method [Gupta et al., 2020] relaxes the strong assumption that all the input frames in a captured video are blurry and adapts attention mechanisms to decide on deblurring each frame based on the information from the neighbor frames. While these methods mainly attempt to remove the motion blur in the VFI task, the inherent motion blur can potentially reveal information about the magnitude and direction of the motion, especially in the case of large non-uniform motion. Along these lines, Zhang et al. [2020] propose a VFI solution closest to the work presented in this thesis. They first extract two sharp keyframes corresponding to the start and the end of a blurry frame, and then, by taking two consecutive frames, they compute the optical flow between the resulting four keyframes. By employing a quadratic motion formulation, they can handle non-uniform motion. However, in this approach, the inaccuracy in predicting the keyframes affects the quality of the flow estimation, which in turn is prone to error, especially for large motion, whereas the methodology in this thesis (Chapter 4) benefits from the less blurred short exposure in each frame to make the flow estimation more reliable. This enables consideration of more intra- and inter-frame flows that are independently estimated, and this processing is carried across subsequent stages of the multi-network pipeline using a multiresolution approach. This thesis (Chapter 4) aims to create a novel HDR VFI so that dealing with extensive saturated regions in the blurry long exposure plays an important role.

2.3.4 High-Speed Video Datasets

High-speed cameras are expensive, and collecting high-speed videos is difficult because daily live cameras mostly do not support higher framerates. High-speed videos play a crucial role in the development of methodologies such as deblurring, frame interpolation, and obtaining ground-truth optical flow. Some of the recent examples of such high frame rate datasets are Adobe240 [Su et al., 2017], GoPro[Nah et al., 2017], X4K1000FPS [Sim et al., 2021], and SlowFlow [Janai et al., 2017]. While Adobe240 and GoPro are the standardized datasets, SlowFlow is captured with a better camera supporting a higher resolution, and it was originally used to obtain ground truth optical flow. The most recent dataset, X4K1000FPS, which has the highest resolution with the highest framerate (Table 2.1), gives a great opportunity to test frame interpolation in case of very high framerates. The magnitude and non-uniformity of the motion are variable in each scene as a result of a diverse set of object movements. Comparably, the lower framerates provide discrete motion blur simulation that could possibly diverge from the camera readings. For this reason, the X4K1000FPS dataset is more valuable for the simulation of camera motion blur due to the more continuous readings.

Dataset	FPS	Resolution	# of Scenes
Adobe240	240	1280 × 720	133
GoPro	240	1280 × 720	33
SlowFlow	240	1530 x 928	41
X4K1000FPS	1000	4096 × 2160	110

TABLE 2.1: *Commonly used high-speed datasets with varying framerates and different resolutions. They are captured with the high frame rate cameras, providing more continuous scene readings compared to conventional cameras. They include various types of scenes to enrich their diversity.*

The first purpose of this thesis (Chapter 4) for using high-speed datasets is the simulation of sensor readings, e.g., camera motion blur, due to continuous captured motion. Furthermore, they provide ground-truth labels for video frame interpolation that extra in-between frames indicate the true motion for the missing frames. Lastly, this thesis focuses on the motion non-uniformity metric that analyzes the motion linearity within the existing frames. This analysis is useful for measuring the performance of the existing VFI methods and the proposed HDR VFI methodology under different levels of motion non-uniformity.

2.4 Image Quality Metrics

Image quality metrics have been important since the early days of image processing. They quantitatively assess the quality of images by considering aspects such as sharpness, contrast, and color accuracy. Depending on whether a reference image exists, image quality metrics can be categorized as either full-reference (FR-IQM) or no-reference (NR-IQM). This section reviews the existing image quality metrics that employ different strategies for image quality assessment.

2.4.1 Full-Reference Metrics

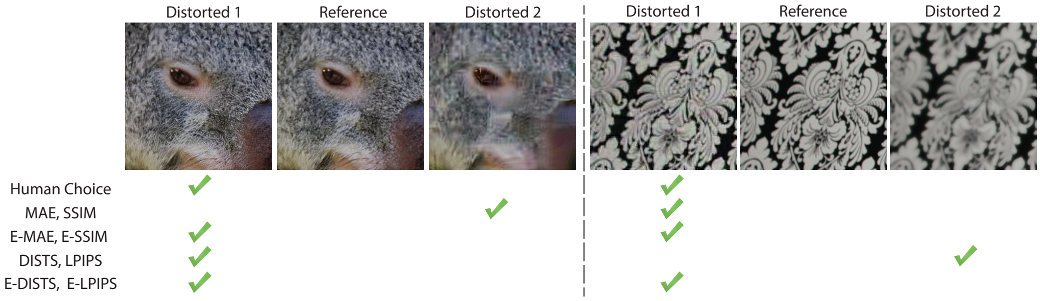
FR-IQMs can be categorized into classical metrics, which perform the computation directly in the image space, and learning-based metrics, which leverage deep feature models to assess image quality.

Classical Metrics

Basic FR-IQMs, such as MSE, RMSE, and MAE, compute the per-pixel difference to quantify image distortion. While these metrics are straightforward to calculate, their consistency with human vision is typically low. Such perceptual consistency can be improved by considering relative instead of absolute error, as in PSNR and the symmetric mean absolute percentage error (SMAPE) [Vogels et al., 2018]. To account for the spatial aspects of the human visual system, alternative metrics such as SSIM [Wang et al., 2004a] are introduced, which consider image patches and measure local differences in luminance, contrast, and structural information. SSIM is further extended to multi-scale MS-SSIM [Wang et al., 2003] and complex wavelet CW-SSIM [Sampat et al., 2009] versions that capture both global and local structural information. FSIM [Zhang et al., 2011b] decomposes the image into multiple subbands using Gabor filters and compares subband responses between the reference and distorted images. By assuming that natural images have a specific distribution of pixel values, models based on information theory [Sheikh and Bovik, 2005; Sheikh and Bovik, 2006] measure the mutual information between images by comparing their joint histograms and taking into account the statistical dependencies between neighboring pixels. Classical metrics can offer either a single overall quality score or a visibility map indicating the distortion intensity. Watson-DCT [Watson, 1993], VDM [Lubin, 1995], VDP [Daly, 1993], HDR-VDP [Mantiuk et al., 2011a], and FovVideoVDP [Mantiuk et al., 2021] measure either the visibility of distortions or perceived distortions magnitude, or both by considering various visual aspects such as luminance adaptation, contrast sensitivity, and visual masking. A more recent metric, FLIP [Andersson et al., 2020], emphasizes color differences, and it is sensitive to even subtle distortions by emulating flipping between the compared image pair.

Deep Learning-Based Metrics

In recent years, research in FR-IQM has been placing greater emphasis on perceptual comparisons in deep feature space rather than image space to enhance the alignment with human judgments. Prashnani et al. [2018] are among the first to utilize deep feature models learned from human-labeled data to predict perceptual errors. Zhang et al. [2018b] demonstrate that internal image representations from classification networks can be used for image comparison. They propose the Perceptual Image Patch Similarity (LPIPS) index, which quantifies image similarity by measuring the ℓ_2 distances between pre-trained VGG features. To further improve the correlation



	Distorted 1	Reference	Distorted 2	Distorted 1	Reference	Distorted 2
Human Choice	✓			✓		
MAE, SSIM			✓	✓		
E-MAE, E-SSIM	✓			✓		
DISTS, LPIPS	✓					✓
E-DISTS, E-LPIPS	✓			✓		

FIGURE 2.3: *Agreement of metric predictions with human judgments. The classic (MAE and SSIM) and learning-based (LPIPS and DISTS) metrics are considered, and their prediction is compared to their enhanced versions (E-MAE, E-SSIM, E-DISTS, and E-LPIPS) using the approach in this thesis. On the left is the situation where MAE and SSIM favor JPEG-like artifacts over slightly resampled textures. On the right, a scenario is accounted for where LPIPS and DISTS prefer blur over a subtle color shift. Metric versions with the proposed technique in this thesis are better aligned with human choice. The images have been extracted from the PIPAL dataset [Jinjin et al., 2020].*

with human judgments, they learn per-channel weights for selected VGG features using their collected perceptual similarity dataset. Recognizing that simple ℓ_p-norm measures fail to consider the statistical dependency of errors across different locations, [Ding et al., 2022] introduces the DISTS, which aims to measure the texture and structure similarity between feature pairs by comparing their global mean, variance, and correlations in the form of SSIM. Building upon this work, A-DISTS [Ding et al., 2021a] extended the approach to incorporate local structure and texture comparisons. Czolbe et al. [2020] incorporate their extended Watson-DCT model [Watson, 1993] as a measure of VGG feature distance. Moving away from deterministic point-wise feature comparisons, DeepWSD [Liao et al., 2022] compares the overall distributions of features using the Wasserstein distance, a statistical measure for comparing two distributions. Nevertheless, the majority of the proposed IQM metrics are targeted toward producing a single quality score and are not primarily designed to generate per-pixel error maps. In this regard, Wolski et al. [2018] employ a custom CNN model trained in a fully supervised way using coarse user marking data to predict an error visibility map that highlights the regions where distortions are more likely to be noticeable.

This thesis (Chapter 5) extends the classic and deep learning-based full-reference metrics, as can be seen in (Figure 2.3), by introducing a learnable component trained on perceptual MOS data in a self-supervised way. By implicitly analyzing local image content, the trained model derives per-pixel maps that mimic visual masking, effectively modeling the visual significance of distortions.

2.4.2 No-Reference Metrics

NR-IQMs asses the quality of the images without relying on the reference images. When ground-truth data is unavailable, certain metrics can still be useful, but estimating errors becomes more difficult without reliable references. Conventionally, different perspectives have been employed in the no-reference metrics as comprehensively reviewed in [Chandler et al., 2014]. BIQI [Moorthy and Bovik, 2010] proposes a methodology that uses the statistics of distorted images based on natural image statistics. BLIINDS-II [Saad et al., 2012] introduces discrete cosine wavelet coefficients to estimate the features of the distorted images. Recently, deep learning-based no-reference metrics such as KonCept512 [Hosu et al., 2020], HYPERIQA [Su et al., 2020], MUSIQ [Ke et al., 2021] and MANIQA [Yang et al., 2022] have been proposed.

While NR-IQM methods often report impressive performance, their practical applicability remains limited. FR-IQM metrics are still predominant in CG applications, as the reference images are typically readily available. The focus of this thesis is not the NR-IQMs, but they are included in the comparisons (Chapter 5) to see how they are positioned among the FR-IQMs and their enhanced versions.

2.4.3 Visual Masking Model

There have been several efforts in the past towards incorporating the perceptual aspects of human vision, specifically visual masking [Legge and Foley, 1980; Foley, 1994; Wilson and Gelb, 1984], into FR-IQM methods [Lubin, 1995; Daly, 1993; Mantiuk et al., 2011a; Mantiuk et al., 2021]. In simple words, visual masking refers to the phenomenon in which certain components of an image (in the case of this thesis, distortions) may be less visible to the viewer due to the presence of other visual elements in the same image. Visual masking can affect image quality perception, making some image distortions less visible to the viewer [Ferwerda et al., 1997; Zeng et al., 2002]. However, existing visual masking models are typically hand-crafted and struggle to generalize effectively across various distortions. Although learning a visual masking model appears to be a natural solution, the lack of reliable ground truth data for visual masking makes direct supervision impractical. In this thesis (Chapter 5), a self-supervised approach is proposed to predict visual masking using a dataset of images featuring a variety of distortions of different magnitudes whose quality has been evaluated in the mean opinion scores (MOS) experiment with human subjects [Lin et al., 2019]. This thesis aims to improve the quality prediction of existing metrics to align more closely with human judgment by detecting the presence and evaluating the magnitude of visible distortion in each pixel.

2.4.4 Image Quality Datasets

Considerable amounts of datasets have been proposed to test the prediction performance of the quality metrics. These datasets later served as a purpose of training for learning-based approaches. They mainly consist of reference images grouped with different distortions at different levels. Each reference image is coupled with different types of distortions so that the diversity in the dataset is achieved. Although the earlier datasets have traditional distortions such as Gaussian noise, Gaussian blur, and JPEG artifacts, the newer datasets (e.g., PIPAL [Jinjin et al., 2020]) contain additional artifacts resulting from different types of convolutional neural network reconstructions. This dataset enables the quality assessment of recently faced artifacts. Here is the list of the most recent image quality datasets CSIQ [Larson and Chandler, 2010], TID2013 [Ponomarenko et al., 2015], KADID [Lin et al., 2019], and PIPAL [Jinjin et al., 2020] and the details of the each dataset is reported in Table 2.2. CSIQ, TID2013, and KADID datasets are MOS datasets, meaning that humans rate how strongly the distortion is visible to them. On the other hand, the most recent dataset, PIPAL, employs a statistic-based Elo rating system by collecting people's opinions.

In addition to these datasets, Wolski et al. [2018] proposed a dataset that indicates the probability of visibility of the errors in the image regions. However, the experimental procedure is very time-consuming compared to MOS datasets due to the careful localization of the visual errors by humans, making it harder to repeat for the new types of distortions, such as CNN artifacts.

Dataset	# of Reference Images	# of Distortion Types	Total # of Distorted Images
CSIQ	30	6	866
TID2013	25	24	3000
KADID	81	25	10125
PIPAL	200	40	23200

TABLE 2.2: *The number of images in the recently proposed image quality datasets. As the older datasets have less diversity in the number of distortion types, the newer datasets come with a richer set of distortion types that helps assess how image quality metrics are generalizable.*

This thesis (Chapter 5) uses KADID to derive visual masks in a self-supervised manner. The proposed methodology is evaluated on the CSIQ, TID2013, and PIPAL datasets to determine how the enhanced metrics correlate with mean opinion scores.

Chapter 3

Image and Video Enhancement and HDR Reconstruction

This chapter uses multi-exposure sensors together with a CNN approach to improve image and video quality by removing CMOS sensor distortions and expanding dynamic range. High dynamic range (HDR) video reconstruction using conventional single-exposure sensors can be achieved by temporally alternating exposures (Section 2.2.1). This, in turn, requires computing exposure alignment, which is difficult to achieve due to the exposure differences that notoriously create problems for moving content, particularly in larger saturated and dis-occluded regions. An attractive alternative is multi-exposure sensors that capture, in a single-shot, differently exposed and spatially interleaved half-frames so that they are perfectly spatially and temporally (up to varying motion blur) aligned by construction (Section 2.2.2). In this chapter, it is demonstrated that reduced spatial resolution and aliasing in such sensors are successfully compensated, and overall, the quality and dynamic range of reconstructed HDR video with respect to single-exposure sensors are improved for a given number of alternating exposures. Specifically, low, mid, and high exposures are considered, and the mid exposure is captured for every frame that serves as a spatial and temporal reference. Here, neural networks for denoising, deblurring, and upsampling tasks are capitalized so that two clean, sharp, and full-resolution exposures for every frame are obtained effectively, which are then complemented by warping a missing third exposure. High-quality warping is achieved by learning optical flow that merges the individual flows found for each specific exposure. Such flow merging is instrumental in handling saturated/dis-occluded image regions, while dense temporal sampling of mid-exposure improves motion quality reproduction between more sparsely sampled exposures. It is also demonstrated that by capturing only a limited amount of sensor-specific data and a novel use of histograms instead of common parametric noise statistics, that makes it possible to generate synthetic training data that lead to a better denoising and deblurring quality than can be achieved by existing state-of-the-art methods. As there is not enough high-quality HDR video available, the method is devised to learn from LDR video instead. The proposed approach compares favorably to several strong baselines and can boost existing HDR image and video methods when they are re-trained on the used training data.

3.1 Introduction

Common single-exposure sensors only capture a limited range of luminance values, while many display and editing tasks would greatly benefit from capturing a higher range [Reinhard et al., 2010]. Multi-exposure techniques [Mann and Picard, 1995; Mitsunaga and Nayar, 1999; Debevec and Malik, 1997] allow for HDR image and

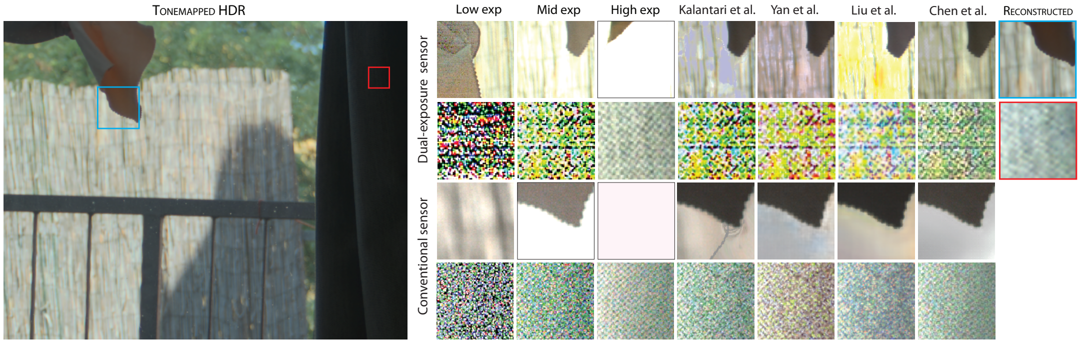


FIGURE 3.1: *HDR video reconstruction of a high-contrast scene (left) using three temporally alternating exposures. Different reconstruction methods are considered for the scene captured with multi-exposure (rows 1 and 2) and conventional single-exposure (rows 3 and 4) sensors. Insets are locally tone-mapped. Rows 1 and 3 show a bright part of the scene (outdoor sun), while rows 2 and 4 show a dark part (indoor shadow). As different sensors are used, the scene is not fully identical, so insets cannot be compared pixel-wise. The first three columns show that any single exposure on any sensor does either not capture the bright or the dark part. Note that the waving tissue is dynamic across those exposures in rows 1 and 3. Other methods, designed for conventional sensors, have artifacts discussed in the main text, such as noise, motion blur, saturation, and color shifts (rows 3 and 4). When using these methods on the multi-exposure sensor in single-exposure mode, artifacts remain (rows 1 and 2). The proposed method provides high-quality results across the entire brightness regime (right, columns).*

video reconstruction for such sensors [Kang et al., 2003; Mangiat and Gibson, 2010; Gryaditskaya et al., 2015; Kalantari et al., 2013; Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019], but a notorious problem here is spatial and temporal exposure alignment for moving content, in particular, in the presence of large saturated and occluded/disoccluded regions. Modern multi-exposure sensors, such as some CMOSIS CMV and Sony IMX sensors, allow one to configure different levels of exposure for different spatial patterns [CMV12000, 2021; IMX, accessed on Sept. 17, 2021]. This allows the expansion of the dynamic range in a single shot by spatially interleaving different exposures across the sensor [Gu et al., 2010; Cho et al., 2014]. The challenge is to combine different exposures into a coherent natural image Figure 3.2. It is even more challenging to not only spatially but also temporally alternate exposures to expand the dynamic image and video range even further (Figure 3.1). This chapter addresses those challenges by capitalizing on the strength of modern machine learning (ML) methods in compensating for distortions inherent to multi-exposure capturing. Moreover, it is also demonstrated that additional data provided by such sensors is instrumental in improving the ML method's performance beyond what might be possible for single-exposure sensors.

3.2 Overview

This chapter proposes a processing pipeline to create HDR video from a sequence of frames with spatially-interleaved multi-exposures as shown in Figure 3.3. The pipeline is composed of three networks **`denoise`**, **`flowMerger`**, and **`blend`** (denoted with purple font) that are cascaded with non-learnable, known from the literature components: `demosaicking` [Malvar et al., 2004], `flowEstimation` [Teed and Deng, 2020], differentiable `warp` operator (which consists of backward warping `bwarp` [Paszke et al., 2019] and forward warping `fwarp` [Xu et al., 2019]) and `makeHDR` [Debevec and Malik, 1997] (green font). Input to the algorithm is three Bayer raw frames that are captured at times: $\mathtt{I}_{t-1}$, $\mathtt{I}_{t}$, and $\mathtt{I}_{t+1}$. First, joint debluring, denoising, and upsampling are performed using a Siamese architecture with three **`denoise`** networks

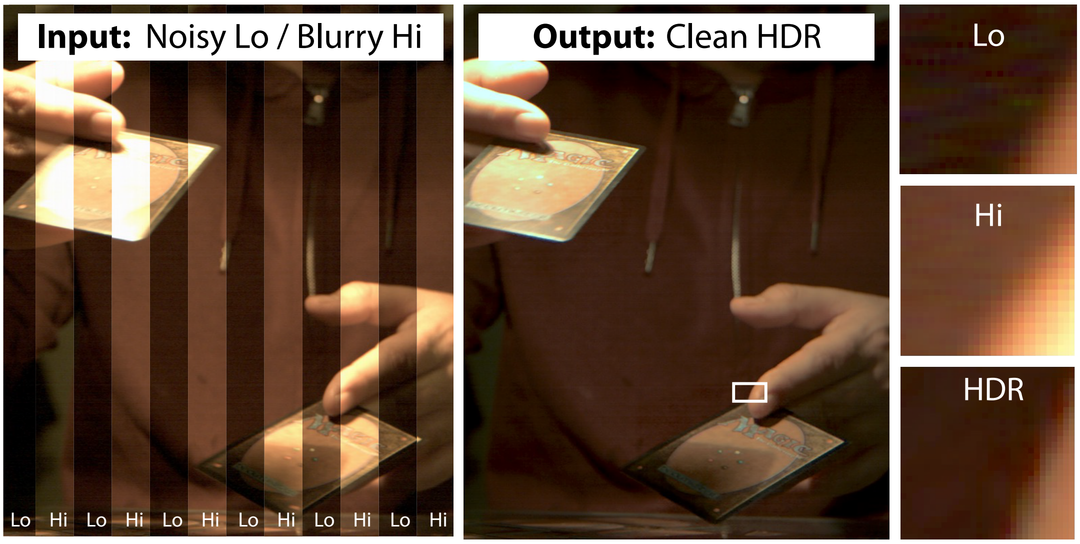

Input: Noisy Lo / Blurry Hi
Lo Hi Lo Hi Lo Hi Lo Hi Lo Hi Lo Hi
Output: Clean HDR
Lo
Hi
HDR

FIGURE 3.2: *Example of spatially-varying exposures and reconstructed HDR. Exposure varies between odd and even columns.*

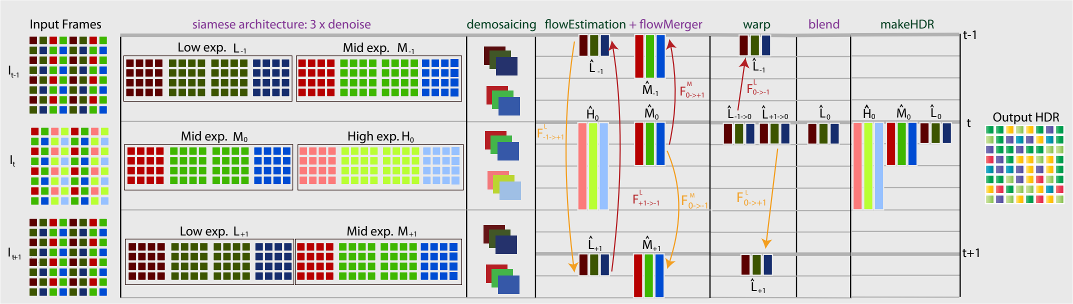


FIGURE 3.3: *The pipeline for temporally-alternating exposures takes three neighboring frames* $\mathtt{I}_{t-1}$, $\mathtt{I}_t$*, and* $\mathtt{I}_{t+1}$ *as the inputs and creates an HDR image aligned with the frame* $\mathtt{I}_t$*. Each input frame is a Bayer raw image with two interleaving exposures. In this scheme, the frame* $\mathtt{I}_t$ *contains mid and high exposures, and each of the frames* $\mathtt{I}_{t-1}$ *and* $\mathtt{I}_{t+1}$ *contains the low and mid exposures. The input exposures are deblurred, denoised, and upsampled using Siamese network* **`denoise`***. The resulting exposures undergo* `demosaicing`*. Then, non-learnable* `flowEstimation` *derives optical flows between the corresponding exposure pairs, as denoted with the line arrows, and finally, network* **`flowMerger`** *aggregates them into optically consistent forward and backward flows. Afterward, this way, learned flows are used to* `warp` *exposures* $\hat{\mathtt{L}}_{-1}$ *and* $\hat{\mathtt{L}}_{+1}$ *to the position of* $\mathtt{I}_t$*, and then missing exposure* $\hat{\mathtt{L}}_0$ *is learned by network* **`blend`***. Finally, all three exposures* $\hat{\mathtt{L}}_0$*,* $\hat{\mathtt{M}}_0$*, and* $\hat{\mathtt{H}}_0$ *are used to reconstruct the output HDR image. Note that the length of the bars indicates the exposure time of each exposure.*

to reconstruct clean and full-resolution exposures. Those exposures then undergo `demosaicing` to finally derive low and mid exposure $\hat{\mathtt{L}}_{-1}$ and $\hat{\mathtt{M}}_{-1}$ for frame $\mathtt{I}_{t-1}$, mid and high exposure $\hat{\mathtt{M}}_0$ and $\hat{\mathtt{H}}_0$ for frame $\mathtt{I}_t$, and low and mid exposure $\hat{\mathtt{L}}_{+1}$ and $\hat{\mathtt{M}}_{+1}$ for frame $\mathtt{I}_{t+1}$. Effectively, either low or high exposure is missing for each frame, as mid-exposures are captured for every frame. In the right part of Figure 3.3, the focus is on reconstructing such a missing exposure $\hat{\mathtt{L}}_0$ for frame $\mathtt{I}_t$ at time t. Note that the time axis is shown along the vertical direction, so the exposure duration and its overlap in the temporal domain can be seen. After computing optical flows between available exposure pairs using `flowEstimation` as denoted by the line arrows, network **`flowMerger`** learns optimized forward and backward flows that they are employed by `warp` to align exposures $\hat{\mathtt{L}}_{-1}$ and $\hat{\mathtt{L}}_{+1}$ with frame $\mathtt{I}_t$. Network **`blend`** combines the resulting warped exposures into the missing $\hat{\mathtt{L}}_0$, which is then submitted along with $\hat{\mathtt{M}}_0$ and $\hat{\mathtt{H}}_0$ to `makeHDR` that reconstructs an output HDR frame.

One of the key problems here is to obtain training data, which is synthesized using only a limited amount of captured data to account for complex sensor-specific noise characteristics. Pre-captured high-speed videos are also used to model exposure-dependent motion blur (MB). In Section 3.3, the sensor reading simulation as required to train the network **`denoise`** is presented. In Section 3.4, the details on networks **`flowMerger`** and **`blend`** are provided, as well as the complete pipeline for HDR video reconstruction. Section 3.6 evaluates the performance of **`denoise`** and HDR video reconstructions. Noise analysis of different sensors is presented in Section 3.7, and additional materials are provided in Section 3.8 and Section 3.9. Section 3.10 discusses the limitations of the proposed techniques, and finally, this chapter is concluded in Section 3.11.

3.3 Deblurring and Denoising

Image deblurring and denoising are typically solved by supervising a CNN with pairs of CLEAN and DISTORTED videos to implement DISTORTED→CLEAN restoration. For the multi-exposure sensor, it is difficult, as capturing DISTORTED sensor readings is time-consuming, and there is also a lack of CLEAN video. For this reason, the methodology is proposed to overcome both limitations.

Addressing the first, instead, a different function is learned: CLEAN→DISTORTED, which generates samples containing correlated pixel and row noise, as well as motion blur from CLEAN sensor readings. Previous work has made simplifying assumptions, such as Gaussian or Poisson noise, none of which apply to the problem of a multi-exposure sensor. A *non-parametric noise model* is proposed that is expressive yet can be trained on a low number of CLEAN-DISTORTED pairs. While simple histograms are referred to here, they have not been used so far in deep denoising applications.

Second, LDR video is supervised instead because there are not enough CLEAN samples that require HDR video. Unfortunately, this LDR video does not have the same type of MB as found in HDR sensor readings. Hence, high-speed LDR video is used to simulate column-alternating MB in the multi-exposure sensor.

The proposed approach has two steps: learning a model to synthesize distortions to train on (Section 3.3.1; an example result in Figure 3.4) and learning to remove distortions (Section 3.3.2).

3.3.1 Clean-to-Distorted

There are three distortion steps described in the order of the underlying physics (Figure 3.5): motion blur (Section 3.3.1), pixel noise (Section 3.3.1), and row noise (Section 3.3.1). For all steps, the analysis is evaluated from noisy sensor readings to devise a statistical model for inference from DISTORTED, and a synthesis step to apply it to CLEAN.

Motion Blur

With different exposures in different columns, their MB is also different. For example, at exposure ratio $r = 4$, MB is also four times longer, and the image is a mix of sharp and blurry columns. As getting reference data without MB, in particular HDR, is difficult, the focus turns to existing LDR high-speed video footage to simulate multi-exposure MB.

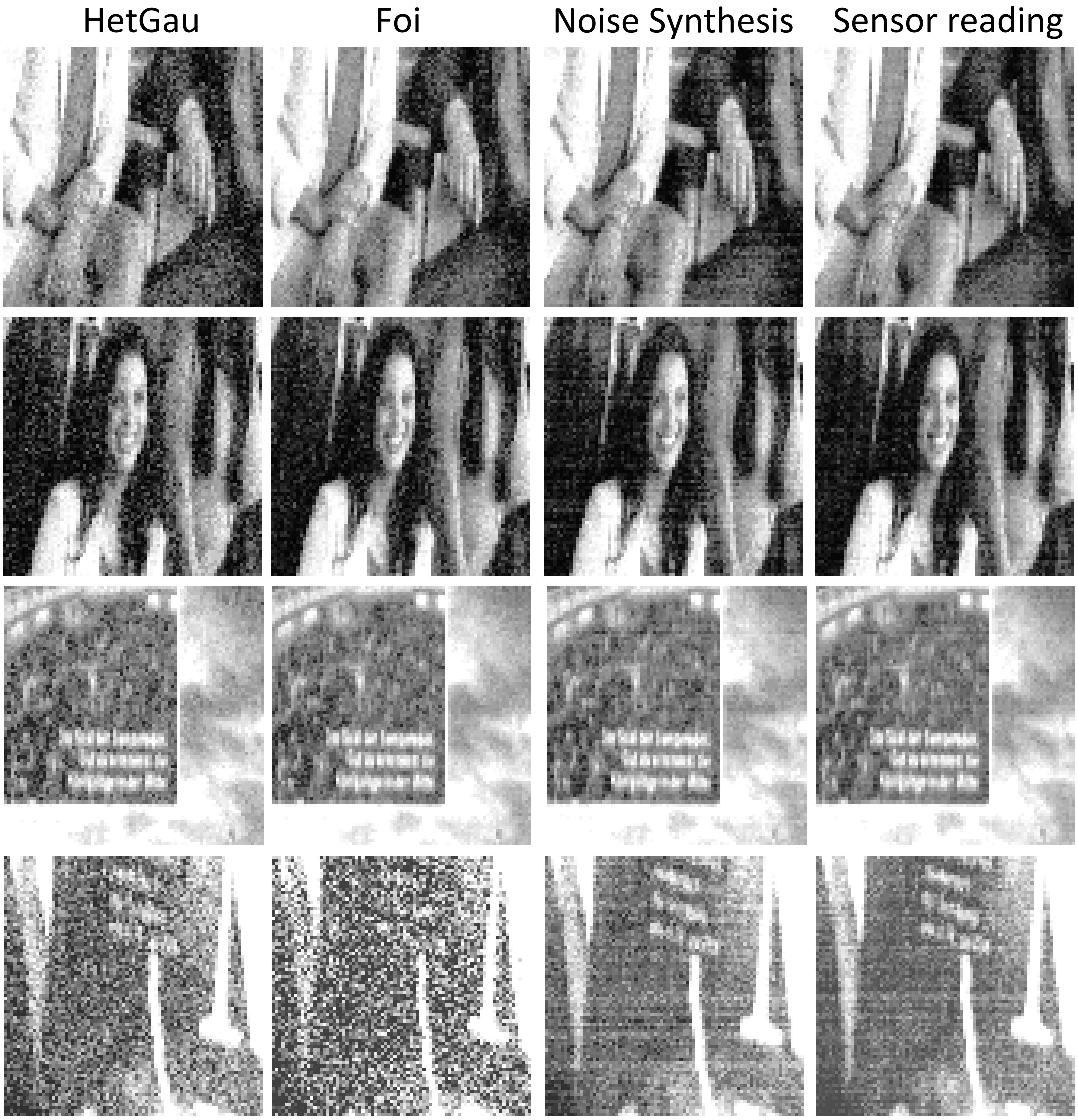


FIGURE 3.4: *Given a noise-free low-exposure reference (not shown), the proposed noise synthesis is compared as well as parametric heteroscedastic Gaussian [Foi et al., 2008] (HetGau) and [Foi, 2009] noise models with a real sensor reading reference* ***(right column).*** *Note the strong correlation between the noise synthesis and the reference.*

Data The high-speed video dataset from [Janai et al., 2017], which has no, or negligible, inherent MB, is used. Note that these are not captured with the multi-exposure sensor, and they have limited dynamic range. They are neither input to nor output from the proposed approach and only provide supervision.

Synthesis Synthesis starts from a random frame of 8-bit LDR high-speed video I_{LDR}. It is converted to a floating point, and an inverse gamma is applied at $\gamma = 2.2$. This is called the *low frame* image, denoted $I_{\mathrm{L}} = I_{\mathrm{LDR}}^{\gamma}$. Since the multi-exposure sensor assures that the low and high exposures are ending at the same time [CMV12000, 2021], to simulate the high frame exposure, four subsequent I_{L} are averaged, then scaled by the exposure ratio, and clamped as in

$$I_{\mathrm{H}} = \texttt{clamp}(r \times \mathbb{E}_{t\in\{0,1,2,3\}}[I_{\mathrm{L}}(t)]).$$

Finally, the low-frame pixels are inserted into the even columns and the high frames into the odd ones, resulting in the motion-blurred image I_{MB}.

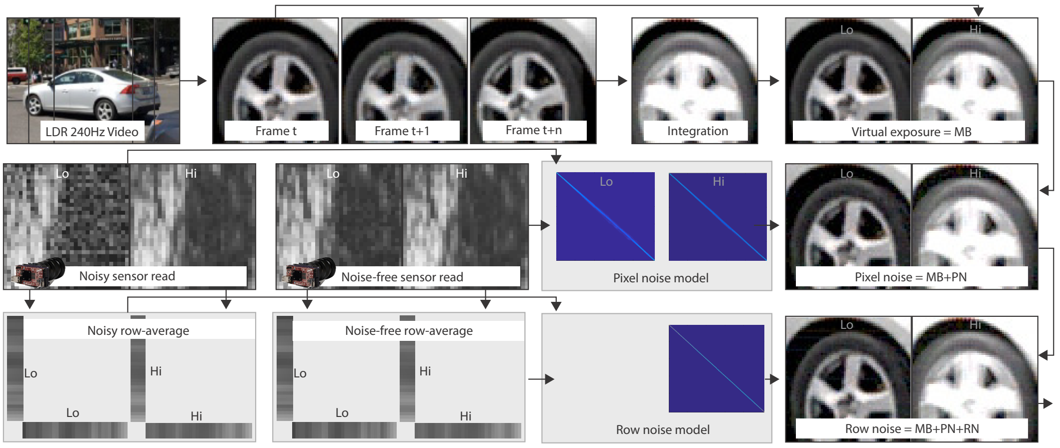


FIGURE 3.5: *The proposed HDR distortion generation pipeline: LDR 240 Hz video is utilized in the top left, from which frames t to $t + n$ are extracted, integrated, and virtually exposed to produce an image with MB* ***(first row)****. Next, pairs of noisy and time-averaged noise-free sensor readings are taken, and a non-parametric noise model (histogram) for low and high exposure is produced. This noise model is added to the virtual exposure image MB* ***(second row)****. Finally, a model of row noise is extracted by averaging vertically or horizontally; this can be added to the pixel noise image, producing the final image with all distortions present* ***(third row)****.*

Pixel Noise

Pixel noise occurs in the sensor and is applied after motion blur, which happens in the sensors. Instead of employing a parametric noise model that has the strengths as priors and for analysis, the non-parametric histograms are used to capture a noise model well-suited for generation. Prior to the noise model derivation, the fixed pattern noise is removed for the sensor, apart from the fixed pixel noise, which also includes the fixed column noise [Janesick, 2001].

Data Assuming that there is a limited amount of GT sensor readings available. In practice, no more than 30 pairs of images (not video) are used that are captured with the target sensor of everyday scenes. Ground truth is acquired by averaging the result of 100 captures of the scene at a very low exposure (so as to make clipping effects negligible) and using a very long exposure.

Analysis The noise is different for different exposures and also for different color channels. The non-parametric model $p_{c,e}(x|y)$ is built, the probability that when the GT value is y, the sensor will read x for channel c and exposure e. A separate model is maintained for every channel in every exposure, leading to six models for three color channels and two exposures although the noise models are similar for different channels at the same exposure. Histograms $H_{c,e}[x][y]$ are used to represent the probability distribution over x for each y in channel c at exposure e. To construct all histograms, every pair of sensor readings and their ground truth, as well as every pixel and every channel, are iterated. Bin x for histogram y is incremented when the GT pixel is y, and the sensor reading is x for channel c and exposure e. The number of histogram bins depends on the bit depth, typically 12 bits, resulting in 4096 bins. After analysis, all histograms are converted into inverse cumulative histograms $C_{c,e}[x][y]$, allowing to sample from them in constant time.

Synthesis Noise synthesis is applied to I_{MB}, the image with simulated MB. Every pixel and every channel of the MB image I_{MB} is iterated to obtain a GT value y. A random number $\phi_{c,e}$ is used to look up the respective cumulative histogram $C_{c,e}$ to produce a simulated sensor value x. Combining all pixels, channels, and exposures results in a virtual synthetic image I_{PN} involving MB and pixel noise.

Row Noise

At short exposures, more structured forms of noise can become important, one of them being *row* noise. This is not to be mistaken with fixed-pattern noise that is frequently spatially correlated but much easier to correct. In row noise, pixels do not change independently; rather, all pixels in a row change in correlation, i. e., the entire row is darkened or brightened. This is a well-known problem for CMOS sensors that are typically caused by the random noise that shifts the voltage level(s) in the Analog-to-Digital Converter (ADC) that, in turn, affects the ADC slope and results in an offset that is consistent for all pixels in the row (refer to [Oten and Li, 2011] and references therein). As the CMOSIS CMV12000 (global shutter) sensor, which is used in this chapter, features row-by-row pixel read-out, the row noise can be clearly observed (Figure 3.4). Similar row noise (dynamic streak noise) has been observed on many sensors, e.g., Canon 5D Mark III and Grasshopper3 GS3-U3-32S4C. Along with clipped noise distributions, this noise is notorious in low-light conditions that are of interest in HDR image capturing.

Sensor manufacturing tries to solve this problem by collecting on-the-fly statistics for surrounding rows to find voltage level offsets in ADC [Oten and Li, 2011], which is prone to errors due to variable image content. It would be impractical for each such sensor and its electronics instance to derive a parametric model that captures the specifics of their row noise. Even if one distortion followed any assumptions, their combination does not. Instead, a histogram approach is employed, which is more descriptive than a parametric model and simpler to implement without any involved mathematical considerations. This approach is used to synthesize noise and ultimately remove it.

Analysis All pairs of GT and sensor images are again iterated, but instead of working on pixels, the work is done on entire rows. In particular, the eight separate means across every row for every channel and exposure are considered. This mean is denoted as $\bar{x}$ in the sensor image and as $\bar{y}$ in the GT image. The next step is to build a model in the form of a histogram, resulting in the inverse cumulative row noise model $\bar{C}_{c,e}[\bar{x}][\bar{y}]$.

Synthesis Synthesis of row noise starts from the image with synthetic MB and pixel noise I_{PN}. Every row, channel, and exposure is iterated, the row means $\bar{y}_{c,e}$ is computed, and again, random number $\bar{\phi}_{c,e}$ is used to draw from $\bar{C}_{c,e}[\bar{\phi}][\bar{y}]$. To make the row mean match the desired mean, the difference of the means is added to the row, resulting in the final synthetic noisy image I_{All}.

3.3.2 Distorted-to-Clean

The network **`denoise`** (refer to Section 3.5 for implementation details) is trained under an L_1 loss in linear space to derive clean exposures from the multi-exposure sensor capturing. The network fully operates in the raw Bayer domain. Non-learnable `demosaicking` [Malvar et al., 2004] is then used to derive RGB exposures. In Section 3.6.1, the performance of network **`denoise`** is evaluated in denoising and deblurring tasks.

3.4 HDR video reconstruction

To reconstruct HDR video, three consecutive frames, $\mathtt{I}_{\mathtt{t}-1}$, $\mathtt{I}_{\mathtt{t}}$, and $\mathtt{I}_{\mathtt{t}+1}$, are considered, each with a pair of interleaved exposures that partially overlap in the time domain as shown in Figure 3.6. As one such interleaved exposure, each frame includes a common mid exposure to facilitate finding spatial and temporal (amount of motion blur) correspondence between the frames. All frames are captured in a RAW format.

Denoising, deblurring, and upsampling network The Siamese network architecture is used for network **denoise**, as introduced in Section 3.3.2, is repeated for each input frame $\mathtt{I}_{\mathtt{t}-1}$, $\mathtt{I}_{\mathtt{t}}$, and $\mathtt{I}_{\mathtt{t}+1}$. The input frames are jointly denoised, deblurred, upsampled, and split into six separate exposures as follows:

$$\begin{aligned} \mathtt{L}_{-1},\mathtt{M}_{-1} &= \mathbf{denoise}(\mathtt{I}_{\mathtt{t}-1}) \\ \mathtt{M}_{0},\mathtt{H}_{0} &= \mathbf{denoise}(\mathtt{I}_{\mathtt{t}}) \\ \mathtt{L}_{+1},\mathtt{M}_{+1} &= \mathbf{denoise}(\mathtt{I}_{\mathtt{t}+1}), \end{aligned} \tag{3.1}$$

where clean full-resolution low and mid exposure $\mathtt{L}_{-1}$ and $\mathtt{M}_{-1}$ for frame $\mathtt{I}_{\mathtt{t}-1}$, mid and high exposure $\mathtt{M}_0$ and $\mathtt{H}_0$ for frame $\mathtt{I}_{\mathtt{t}}$, and low and mid exposure $\mathtt{L}_{+1}$ and $\mathtt{M}_{+1}$ for frame $\mathtt{I}_{\mathtt{t}+1}$ are reconstructed. The resolution increases here due to upsampling in the horizontal direction by a factor of 2, while the amount of blur is aligned with mid-exposures. The resulting exposures are then directly submitted to a non-learnable `demosaicking` algorithm as proposed in [Malvar et al., 2004]:

$$\begin{aligned} \hat{\mathtt{L}}_{-1} &= \mathtt{demosaicking}(\mathtt{L}_{-1}), & \hat{\mathtt{M}}_{-1} &= \mathtt{demosaicking}(\mathtt{M}_{-1}) \\ \hat{\mathtt{M}}_{0} &= \mathtt{demosaicking}(\mathtt{M}_{0}), & \hat{\mathtt{H}}_{0} &= \mathtt{demosaicking}(\mathtt{H}_{0}) \\ \hat{\mathtt{L}}_{+1} &= \mathtt{demosaicking}(\mathtt{L}_{+1}), & \hat{\mathtt{M}}_{+1} &= \mathtt{demosaicking}(\mathtt{M}_{+1}), \end{aligned} \tag{3.2}$$

where $\hat{\mathtt{L}}_{-1}$, $\hat{\mathtt{M}}_{-1}$, $\hat{\mathtt{M}}_{0}$, $\hat{\mathtt{H}}_{0}$, $\hat{\mathtt{L}}_{+1}$ and $\hat{\mathtt{M}}_{+1}$ denote full-resolution linear RGB exposures. Note that exposure $\hat{\mathtt{L}}_{0}$ is missing at time t that corresponds to frame $\mathtt{I}_{\mathtt{t}}$ (refer to Figure 3.6). Further discussion focuses on the reconstruction of this missing exposure (refer to Figure 3.3, right).

Flow merging network $\hat{\mathtt{L}}_{0}$ can be reconstructed by warping frames $\hat{\mathtt{L}}_{-1}$ and $\hat{\mathtt{L}}_{+1}$ that require the corresponding optical flow computation. In multi-exposure sensors, different exposure pairs $\hat{\mathtt{L}}_{-1}/\hat{\mathtt{M}}_{-1}$, $\hat{\mathtt{M}}_{0}/\hat{\mathtt{H}}_{0}$, and $\hat{\mathtt{L}}_{+1}/\hat{\mathtt{M}}_{+1}$ are perfectly spatially registered, while as an outcome of deblurring, they are temporally aligned as well. Consequently, a flow found between frames with the same exposure immediately applies to another exposure. This alleviates the problem of flow computation for pixels that are saturated at one exposure, which might be challenging for single-exposure sensors [Kalantari and Ramamoorthi, 2017; Kalantari and Ramamoorthi, 2019]. Moreover, the outcome of a state-of-the-art optical flow algorithm can be further refined by checking the flow consistency between different exposures.

The optical flow algorithm `flowEstimation` [Teed and Deng, 2020] is used for estimating flows $F^{\mathtt{M}}_{0\to-1}$, $F^{\mathtt{M}}_{0\to+1}$, $F^{\mathtt{L}}_{+1\to-1}$ and $F^{\mathtt{L}}_{-1\to+1}$, where $F^{\mathtt{E}}_{\mathtt{i}\to\mathtt{j}}$ is the flow from exposure $\mathtt{E}_{\mathtt{i}}$ to $\mathtt{E}_{\mathtt{j}}$. Typically, the most reliable are flows $F^{\mathtt{M}}_{0\to-1}$ and $F^{\mathtt{M}}_{0\to+1}$ as exposures $\hat{\mathtt{M}}_{-1}$, $\hat{\mathtt{M}}_{0}$, and $\hat{\mathtt{M}}_{+1}$ are available for consecutive frames. Nevertheless, saturated image regions in such mid exposures might reduce the quality of derived flows, and then relying on flows $F^{\mathtt{L}}_{+1\to-1}$ and $F^{\mathtt{L}}_{-1\to+1}$ between low exposures might be beneficial. As the latter flows are computed between exposures $\hat{\mathtt{L}}_{-1}$ and $\hat{\mathtt{L}}_{+1}$, and thus are separated by two frames, they need to be properly split to the time position of missing exposure

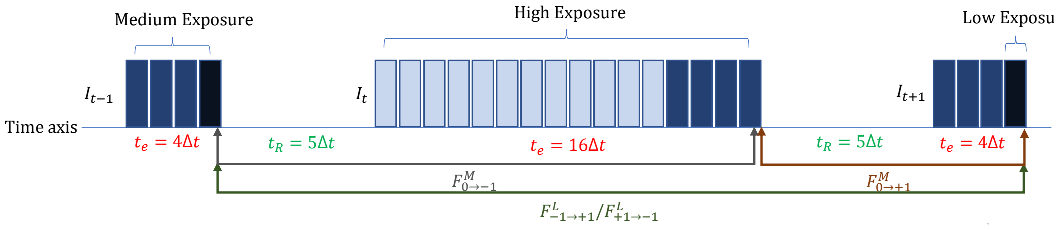


FIGURE 3.6: *Optical flow computation for multi-exposure sensor and temporally interleaving exposures. In each frame, two exposures partially overlap in the temporal domain, and a ratio of four between the exposure duration is assumed. Following the multi-exposure sensor specification, the end point of those exposures is temporally aligned. The time interval* $\Delta\mathtt{t}$ *corresponds to low exposure duration (black bar). Each frame contains mid exposure (dark blue bars that overlap with black bars in* $\mathtt{I}_{\mathtt{t}-1}$ *and* $\mathtt{I}_{\mathtt{t}+1}$*). Frame* $\mathtt{I}_\mathtt{t}$ *also includes high exposure (bright blue bars). The starting and ending points for flows* $\mathtt{F}^{\mathtt{M}}_{0\to-1}$*,* $\mathtt{F}^{\mathtt{M}}_{0\to+1}$*,* $\mathtt{F}^{\mathtt{L}}_{-1\to+1}$*, and* $\mathtt{F}^{\mathtt{L}}_{-1\to+1}$ *in the temporal domain are marked with arrows.* $\mathtt{t_e}$ *denotes the exposure time, and* $\mathtt{t_R}$ *=5*$\Delta\mathtt{t}$ *is the read-out time that roughly follows multi-exposure sensor specification.*

$\hat{\mathtt{L}}_0$ in frame $\mathtt{I}_\mathtt{t}$ (refer to Figure 3.6). This task is relegated to network **flowMerger** that derives refined flows $\mathtt{F}^{\mathtt{L}}_{0\to-1}$ and $\mathtt{F}^{\mathtt{L}}_{0\to+1}$ by consistently merging information from all input flows:

$$\begin{aligned}\mathtt{F}^{\mathtt{L}}_{0\to-1},\mathtt{F}^{\mathtt{L}}_{0\to+1} = \mathbf{flowMerger}(&\hat{\mathtt{M}}_0,\hat{\mathtt{M}}_{-1},\\ &\hat{\mathtt{M}}_{+1},\hat{\mathtt{L}}_{-1},\hat{\mathtt{L}}_{+1},\mathtt{F}^{\mathtt{M}}_{0\to-1},\mathtt{F}^{\mathtt{M}}_{0\to+1},\\ &\mathtt{F}^{\mathtt{L}}_{-1\to+1}.\mathtt{F}^{\mathtt{L}}_{+1\to-1}),\end{aligned} \tag{3.3}$$

Exposure $\hat{\mathtt{L}}_0$ blending network The flows $\mathtt{F}^{\mathtt{L}}_{0\to-1}$ and $\mathtt{F}^{\mathtt{L}}_{0\to+1}$ are used, a non-learnable warping algorithm [Paszke et al., 2019] is applied to derive two low-exposure estimates $\hat{\mathtt{L}}_{-1\to0}$ and $\hat{\mathtt{L}}_{+1\to0}$ that are aligned with frame $\mathtt{I}_\mathtt{t}$:

$$\hat{\mathtt{L}}_{-1\to0} = \mathtt{bwarp}(\hat{\mathtt{L}}_{-1},\mathtt{F}^{\mathtt{L}}_{0\to-1}) \qquad \hat{\mathtt{L}}_{+1\to0} = \mathtt{bwarp}(\hat{\mathtt{L}}_{+1},\mathtt{F}^{\mathtt{L}}_{0\to+1})$$

The warped exposures $\hat{\mathtt{L}}_{-1\to0}$ and $\hat{\mathtt{L}}_{+1\to0}$ are then supplied to the **blend** network:

$$\hat{\mathtt{L}}_0 = \mathbf{blend}(\hat{\mathtt{M}}_0,\hat{\mathtt{L}}_{-1\to0},\hat{\mathtt{L}}_{+1\to0}) \tag{3.4}$$

which is responsible for creating the missing low exposure $\hat{\mathtt{L}}_0$. Here, exposure $\hat{\mathtt{M}}_0$ provides an additional reference for spatial positioning. Finally, $\hat{\mathtt{L}}_0$, $\hat{\mathtt{M}}_0$, and $\hat{\mathtt{H}}_0$ are combined into the output HDR $\mathtt{Y}_0$ frame using a non-learnable `makeHDR` technique similar to [Debevec and Malik, 1997].

The methodology so far, as well as the scheme in Figure 3.3, focused on reconstructing missing low exposure $\hat{\mathtt{L}}_0$ for frame $\mathtt{I}_\mathtt{t}$. Nevertheless, missing high exposures $\hat{\mathtt{H}}_{-1}$ and $\hat{\mathtt{H}}_{+1}$ for frames $\mathtt{I}_{\mathtt{t}-1}$ and $\mathtt{I}_{\mathtt{t}+1}$ are reconstructed in the same way. Once trained, the network is employed to derive both low and high exposures.

3.4.1 Loss Function

The training starts from network **denoise**; then its weights are fixed and cascaded with network **flowMerger**, which is trained next. Finally, this procedure is repeated with network **blend** that completes the whole system training. L_1 loss is used to calculate the following loss functions:

$$\begin{aligned} \mathtt{L_d} = {} & ||\mathtt{L}_{-1} - \mathtt{L}'_{-1}||_1 + ||\mathtt{M}_{-1} - \mathtt{M}'_{-1}||_1 + ||\mathtt{M}_0 - \mathtt{M}'_0||_1 + \\ & ||\mathtt{H}_0 - \mathtt{H}'_0||_1 + ||\hat{\mathtt{L}}_{+1} - \mathtt{L}'_{+1}||_1 + ||\mathtt{M}_{+1} - \mathtt{M}'_{+1}||_1, \end{aligned} \tag{3.5}$$

where $\mathtt{L}'_{-1}$, $\mathtt{M}'_{-1}$, $\mathtt{M}'_0$, $\mathtt{H}'_0$, $\mathtt{L}'_{+1}$ and $\mathtt{M}'_{+1}$ denote the ground truth Bayer RGGB channels, and

$$\mathtt{L}_{\mathtt{f}_1} = ||\hat{\mathtt{L}}_{-1\to 0} - \mathtt{L}''_0||_1 + ||\hat{\mathtt{L}}_{+1\to 0} - \mathtt{L}''_0||_1 \tag{3.6a}$$

$$\mathtt{L}_{\mathtt{f}_2} = ||\hat{\mathtt{L}}_{-1\to +1} - \mathtt{L}''_{+1}||_1 + ||\hat{\mathtt{L}}_{+1\to -1} - \mathtt{L}''_{-1}||_1 \tag{3.6b}$$

$$\mathtt{L}_{\mathtt{f}_3} = ||\hat{\mathtt{M}}_{+1\to 0} - \mathtt{M}''_0||_1 + ||\hat{\mathtt{M}}_{-1\to 0} - \mathtt{M}''_0||_1 \tag{3.6c}$$

$$\mathtt{L}_\mathtt{f} = \mathtt{L}_{\mathtt{f}_1} + \mathtt{L}_{\mathtt{f}_2} + \mathtt{L}_{\mathtt{f}_3} \tag{3.6d}$$

$$\mathtt{L_b} = ||\hat{\mathtt{L}}_0 - \mathtt{L}''_0||_1, \tag{3.7}$$

where $\mathtt{L}''_0$, $\mathtt{L}''_{-1}$, $\mathtt{M}''_0$ and $\mathtt{L}''_{+1}$ denote the ground truth RGB channels, while

$$\begin{aligned} \hat{\mathtt{L}}_{-1\to +1} &= \mathtt{bwarp}(\hat{\mathtt{L}}_{-1}, \mathtt{fwarp}(\mathtt{F}^\mathtt{L}_{0\to -1} - \mathtt{F}^\mathtt{L}_{0\to +1}, \mathtt{F}^\mathtt{L}_{0\to +1})) \\ \hat{\mathtt{L}}_{+1\to -1} &= \mathtt{bwarp}(\hat{\mathtt{L}}_{+1}, \mathtt{fwarp}(\mathtt{F}^\mathtt{L}_{0\to +1} - \mathtt{F}^\mathtt{L}_{0\to -1}, \mathtt{F}^\mathtt{L}_{0\to -1})) \\ \hat{\mathtt{M}}_{-1\to 0} &= \mathtt{bwarp}(\hat{\mathtt{M}}_{-1}, \mathtt{F}^\mathtt{L}_{0\to -1}) \\ \hat{\mathtt{M}}_{+1\to 0} &= \mathtt{bwarp}(\hat{\mathtt{M}}_{+1}, \mathtt{F}^\mathtt{L}_{0\to +1}), \end{aligned} \tag{3.8}$$

where `fwarp` [Xu et al., 2019] is used to align flows to the correct position. $\mathtt{L}_d$, $\mathtt{L}_f$ and $\mathtt{L}_b$ are the losses for networks **denoise**, **flowMerger**, and **blend**, respectively. As discussed, **denoise** is a part of a Siamese network architecture that considers all input channels in loss function $\mathtt{L}_d$. Eq. 3.6a ensures that forward and backward flows meet at the missing exposure $\mathtt{L}''_0$ while Eq. 3.6b keeps refined flows consistent between low exposures $\hat{\mathtt{L}}_{-1}$ and $\hat{\mathtt{L}}_{+1}$. It is made sure that refined flows also work on previous $\hat{\mathtt{M}}_{-1}$ and next $\hat{\mathtt{M}}_{+1}$ mid exposures by introducing Eq. 3.6c. As such mid exposures are captured for every frame, typically they lead to a more precise motion reconstruction, possibly except bright image regions, where flows derived from low exposures might be more precise. Ablation results of each component of $\mathtt{L}_f$ are presented in Section 3.6.2. Finally, $\mathtt{L}_b$ compares the output blended $\hat{\mathtt{L}}_0$ with $\mathtt{L}''_0$.

3.5 Implementation

A similar network architecture is used for all three networks: **denoise**, **flowMerger**, and **blend**. However, the input and output channels are adjusted according to the type of the network. The convolutional network with residual connections [He et al., 2016] that consists of 12 dilated convolutional layers is used. More details on the number of input and output channels and dilation factors are provided in Table 3.4 in Section 3.8. Residual connections are symmetrically used between corresponding convolutional layers for a direct gradient flow through the network. Each layer has a filter size of 3 × 3. After convolutional layer 9, a non-learnable bilinear upsampling increases the horizontal resolution of internal channels by a factor of 2.

Each network is trained separately as discussed in Section 3.4. The end-to-end system takes as the input raw Bayer RGGB channels extracted from three dually-exposed frames: $\mathtt{I}_{t-1}$, $\mathtt{I}_t$, and $\mathtt{I}_{t+1}$. During the training time, the patches of size 256 × 128 × 24 are fed, and the networks output a missing exposure of size 512 ×

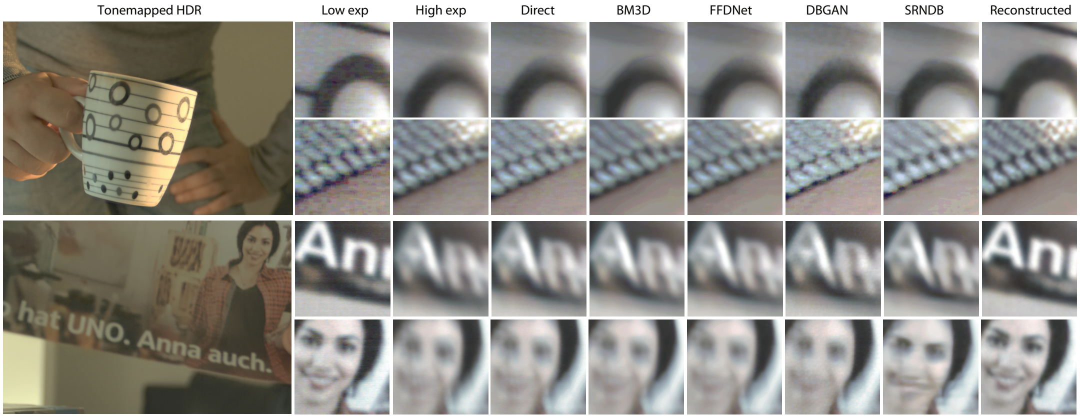


FIGURE 3.7: *Comparison of different methods **(columns)** on two scenes **(rows)**. Please see the text for discussion.*

512 × 3. Note that the convolutional network can handle images of any size at the inference time. 35 videos from the high-speed video dataset [Janai et al., 2017] are used, where 40,000 patches are selected from 8,000 frames.

The ADAM optimizer [Kingma and Ba, 2015] is used for faster convergence with suggested parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is 10^{-4}, and the batch size of 4 is used. The networks are implemented using PyTorch [Paszke et al., 2019] with GPU support. An NVIDIA Tesla P100 GPU is used for training and testing phases. Training takes approximately 1 day for each of the learnable components **`denoise`**, **`flowMerger`**, and **`blend`** in the end-to-end system. The running time of the whole algorithm is 1.81, 0.63, and 0.43 secs for image sizes of 1024 × 1024, 512 × 512, and 256 × 256, respectively. Such relatively long runtimes can be attributed to the full-resolution that is employed across all layers in the network solution. As discussed in Section 3.8, U-Net (encoder-decoder) type architectures, although potentially much faster, lead to a lower quality of HDR reconstruction.

3.6 Results

The quantitative and qualitative evaluation on image deblurring/denoising (Section 3.6.1), flow computation (Section 3.6.2), and video reconstruction (Section 3.6.3) tasks are presented.

All test images have been captured using an Axiom-beta camera with a CMOSIS CMV12000 sensor [CMV12000, 2021] and a Canon EFS 18-135 mm lens at resolution 4096 × 3072 RAW 12-bit pixels, using the lowest gain with exposure ratio 4 (or 16 when explicitly mentioned) and (low) exposure time varying from 0.25 to 16 ms. Although the noise model is created for a given fixed ratio, the exposure times for the two discrete exposures can vary continuously. All results are shown after gamma correction and photographic tone mapping [Reinhard et al., 2002]. CMOSIS CMV12000 sensor [CMV12000, 2021] is a CMOS sensor that features a global shutter, large pixel sizes, low dark current noise, and is relatively inexpensive in comparison with CCD sensors with similar performance. Therefore, the sensor is suitable for demanding computer vision applications, and it is offered by many well-known industrial camera makers [Basler, accessed on Sept. 17, 2021; Omnivision, accessed on Sept. 17, 2021; vision, accessed on Sept. 17, 2021].

3.6.1 HDR Image Denoising/Deblurring Evaluation

Now, the combination of the proposed method and the synthetic training data, as well as other ways to obtain training data and other methods for denoising and deblurring, are evaluated. This section focuses on denoising and deblurring of differently exposed frames and HDR images resulting from such different exposure merging [Debevec and Malik, 1997].

Methods The evaluation considers eight methods (color-coded; "Method" in Table 3.1): Direct is a non-learned direct, physics-based fusion of the low and high frame, with bicubic upsampling [Debevec and Malik, 1997]. Next, BM3D [Dabov et al., 2007] is a gold-standard, non-deep denoiser. When BM3D is "trained" this means performing a grid search on the training data in order to find the standard deviation parameter with the optimal PSNR. FFDNet [Zhang et al., 2018a] is a state-of-the-art deep denoiser. DBGAN [Kupyn et al., 2019] and SRNDB [Tao et al., 2018] are recent deblurring approaches. LSD [Mustaniemi et al., 2020] is a deep multi-exposure method that produces denoised and deblurred LDR images. The final method is Heide [Heide et al., 2014], which is a general image reconstruction method capable of working with multiplexed exposures. Note that, in order to run the single-image denoising and deblurring methods on the multi-exposure frames, each exposure first is extracted from the raw interleaved frame (refer to Figure 3.2). Each exposure is then upsampled in horizontal dimension by a factor of two to compensate for the missing columns. Finally, demosaicing and gamma correction are applied to the raw Bayer exposures to get the full-resolution sRGB exposures.

Training data Each method is studied to see how it performs when trained with different data ("Train. data" in Table 3.1). Each type of training data has a different symbol. If the authors provide a pre-trained version, it is denoted as "Theirs" (▼). "Sensor" (▲) means training on the image for which paired training data is available directly without the proposed re-synthesis. Please note that this training is not applicable to tasks that involve removing MB, as the supervision inevitably contains MB. Next, heteroscedastic Gaussian noise, "HetGau" (●), and noise simulation from [Foi, 2009] "Foi" (✺) are studied, which refers to taking the training data, fitting a linear model of Gaussian parameters of the error distribution and then re-synthesizing training. Finally, four ablations of the training data generation are studied: only row noise ("OnlyRN", ✦), only pixel noise ("OnlyPN", ★), only motion blur ("OnlyMB", ✷), and finally ("All", ✹) as presented in Section 3.3.2.

Metrics The measurements are conducted using PSNR, SSIM[Wang et al., 2004b], and HDR-VDP-3, which is the latest version of [Mantiuk et al., 2011b], where more is better. Table 3.5 in the Section 3.9 also shows the standard deviation values for the mean PSNR measurements in Table 3.1.

Test set The test set for the quantitative comparisons in Table 3.1 consists of both sensor and synthesized data. The sensor data contains 30 static scenes captured with the multi-exposure camera in two different exposure settings, and the noise-free references are obtained by averaging 45 frames per scene and exposure. The sensor data is only used to evaluate the proposed method in the denoising task that is uniquely denoted with the subscript "SENSOR" in Table 3.1. As the sharp and noise-free reference frames are hard to obtain using the multi-exposure sensor for dynamic scenes, another test set is created by simulating noise and motion blur as discussed in Section 3.3.1. Such synthetic data enables the proper evaluation of the proposed method when noise and motion blur are simultaneously present in differently exposed frames in the context of HDR frame reconstruction (last four columns). The synthesized data contains 204 frames from 19 different high-speed videos [Galoogahi et al., 2017] and features variable dynamic range.

Tasks In Table 3.1, the proposed method is consistently compared with the deblurring and denoising techniques for different tasks (five rightmost columns in Table 3.1). First, the input is a noisy low exposure only, and the reference is a clean low exposure (tasks Lo2Lo$_{\text{SENSOR}}$ and Lo2Lo that involve evaluations for the sensor reading and synthetic test sets, respectively). In the second task (Hi2Hi-MB), the input is a high exposure only, and the reference is a sharp and noise-free high exposure. Since the high exposures are typically less noisy while the MB is more pronounced, denoising methods are evaluated in this task to see their performance in removing the noise in the presence of MB. Third is a task where the input is both exposures and the output is an HDR image without noise, while the MB remains intact (LoHi2HDR). The fourth task also consumes low and high exposures and removes both noise and MB to output HDR (LoHi2HDR-MB). For the third and fourth tasks, the denoising and deblurring methods are applied to each exposure separately, and then their outputs are fused for the final HDR reconstruction [Debevec and Malik, 1997]. Note in the Lo2Lo$_{\text{SENSOR}}$ task, only the denoising methods are evaluated, while in the Lo2Lo task, the deblurring methods are also evaluated so that the resulting low exposures can be used for the final HDR reconstruction in the LoHi2HDR and LoHi2HDR-MB tasks.

Table 3.1: *Performance of different methods and different training data* (rows) *for different tasks* (columns). *Different icon shapes denote different training; colors map to different methods. The proposed method is comprehensively evaluated for the denoising and deblurring tasks and additionally, training is conducted with different parametric noise models in order to evaluate the proposed noise modeling. For denoising comparisons, BM3D [Dabov et al., 2007] and FFDNet [Zhang et al., 2018a] are compared, and for the deblurring task, recent deblurring approaches DBGAN [Kupyn et al., 2019] and SRNDB [Tao et al., 2018] are compared. LSD$_2$ [Mustaniemi et al., 2020] is a deep multi-exposure method that produces denoised and deblurred HDR images. Heide [Heide et al., 2014] is a general image reconstruction method capable of working with multiplexed exposures. In the bottom rows, the proposed method of noise synthesis is compared with existing parametric methods: Foi [2009] and heteroscedastic Gaussian noise [Foi et al., 2008] (HetGau). Four ablations of the training data generation are also included: only row noise OnlyRN, only pixel noise OnlyPN, only motion blur OnlyMB, and finally All.*

				Task		
		Lo2Lo$_{\text{SENSOR}}$	Lo2Lo	Hi2Hi-MB	LoHi2HDR	LoHi2HDR-MB
	Input: Low Exp.	✓	✓	✗	✓	✓
	Input: High Exp. with MB	✗	✗	✓	✓	✓
	MB removed	✗	✗	✓	✗	✓
	Output: LDR	✓	✓	✓	✗	✗
	Output: HDR	✗	✗	✗	✓	✓
Train. data	Method	Error (PSNR\SSIM\HDR-VDP-3)				
▼ Theirs	Direct [Debevec and Malik, 1997]	37.66\0.911\9.24	33.27\0.843\8.78	28.86\0.858\9.06	34.23\0.926\9.39	33.26\0.890\9.13
▼ Theirs	BM3D [Dabov et al., 2007]	36.37\0.924\9.53	33.98\0.940\8.78	28.93\0.918\8.93	34.24\0.960\9.60	33.03\0.947\9.19
▲ Sensor		39.74\0.956\9.20	34.47\0.943\9.01	—	34.35\0.962\9.63	—
● HetGau		39.12\0.948\9.31	34.97\0.945\8.92	29.05\0.923\9.01	34.63\0.965\9.63	33.30\0.954\9.25
✹ All		39.76\0.956\9.20	34.99\0.946\8.90	29.05\0.921\9.02	34.69\0.964\9.64	33.33\0.953\9.25
▼ Theirs	FFDNet [Zhang et al., 2018a]	38.50\0.938\8.87	34.44\0.924\8.94	28.98\0.914\9.01	33.78\0.956\9.66	32.33\0.943\9.28
▲ Sensor		38.59\0.939\8.90	34.37\0.944\9.07	—	34.13\0.959\9.66	—
✹ All		38.67\0.940\8.94	34.23\0.944\8.62	28.96\0.922\8.99	34.37\0.960\9.64	33.14\0.950\9.27
▼ Theirs	DBGAN [Kupyn et al., 2019]	—	28.97\0.894\8.78	26.76\0.902\9.14	31.16\0.941\9.35	30.35\0.934\9.18
▼ Theirs	SRN-DB [Tao et al., 2018]	—	31.53\0.934\8.94	27.72\0.913\9.17	32.14\0.955\9.49	31.38\0.948\9.25
▼ Theirs	LSD$_2$ [Mustaniemi et al., 2020]	—	—	—	29.94\0.935\9.07	32.09\0.951\9.20
▼ Theirs	Heide et al. [2014]	—	—	—	—	34.12\0.895\9.32
▲ Sensor	denoise	33.79\0.929\9.32	28.05\0.826\8.63	—	29.01\0.868\8.63	—
✺ Foi		41.85\0.963\9.44	37.72\0.941\9.69	35.92\0.950\9.58	39.41\0.972\9.80	39.01\0.963\9.61
● HetGau		40.98\0.956\9.22	36.39\0.941\9.37	35.68\0.942\9.56	38.71\0.967\9.73	37.58\0.961\9.57
✦ OnlyRN		39.58\0.945\9.11	33.94\0.891\9.28	32.19\0.900\9.37	35.32\0.937\9.67	35.24\0.919\9.40
★ OnlyPN		38.48\0.950\9.52	35.02\0.926\9.50	31.97\0.922\9.50	36.08\0.947\9.62	36.24\0.926\9.49
✷ OnlyMB		39.97\0.951\9.22	36.06\0.910\9.53	34.90\0.930\9.56	37.91\0.955\7.37	37.62\0.938\9.58
✹ All		**42.56\0.967**\9.30	**38.11\0.950**\9.60	**36.22\0.952\9.61**	**39.71\0.975\9.89**	**39.07\0.966\9.62**

Discussion Results are shown in Table 3.1. The proposed method trained on the synthetic training data (✹) performs best on all tasks. The ablations (✦, ★ and ✷) all perform worse than the full method, indicating all additions are relevant. The proposed method also consistently performs better on both the sensor and synthesized test sets for the denoising task, as shown in the columns Lo2Lo$_{\text{Sensor}}$ and Lo2Lo. It can also be observed that improvement in the performance of other methods when trained on synthesized data using the distortion model (✻ and ✻) compared to being trained on their original data (▼ and ▼); however, none can compete with the full method (✹). The only exception is the HDR-VDP-3 response for the original BM3D method ▼ in the Lo2Lo$_{\text{Sensor}}$ task. The proposed network **`denoise`** is also trained with other noise models, such as using sensor data directly (▲), heteroscedastic Gaussian noise [Foi et al., 2008] (●) and noise simulation from [Foi, 2009] (❋), but none of these was able to capture the combination of motion blur, pixel noise, and row noise, resulting in larger errors. As a sanity check, BM3D [Dabov et al., 2007] and FFDNet [Zhang et al., 2018a] are also tuned on sensor data (▲ and ▲), but it did not lead to a superior result compared to tuned on the synthesized data (✻ and ✻). A further test is to compare to Direct [Debevec and Malik, 1997] (▼), which is not learned or doing anything except up-sampling and fusion; this should be a lower bound for any method or task. Finally, the approach compares favorably to [Heide et al., 2014] (▼), a general and flexible imaging framework that can work on multi-exposure images. Looking at performance over different tasks, when they get more involved, i. e., removing MB or producing HDR, other methods start to perform more similarly, but the proposed method tends to win by a larger margin. When denoising methods such as BM3D and FFDNet are evaluated, as expected, they fail to remove the blur in the high exposure. As a result, the low and high exposures are not aligned with each other, which, in turn, explains the worse performance in the final HDR reconstruction. In contrast, the full method (✹) successfully removes the noise and blur in the low and high exposures. The better performance is achieved for the combined task LoHi2HDR-MB even compared to individual tasks. This is mainly because, in the HDR reconstruction, the high exposure contributes more to the dark regions, which are challenging for the low exposure. In summary, using the right training data helps the proposed methods and others to solve multiple aspects of multiple tasks.

The quantitative results from above are complemented by the qualitative ones in Figure 3.7. The first row shows the proposed method (✹) complete image. The second and third rows show selected low and high input patches, which suffer from noise or blur, respectively. Directly (▼) fusing both into HDR, as in the fourth column, reduces noise and blur but cannot remove them. The BM3D (✻) and FFDNet (▼) columns show that individual frames can be denoised, but blur remains. This is most visible in moving parts, such as the cup and brochure in the first and third rows respectively. Using deblurring, as in DBGAN (▼) or SRNDB (▼), can not completely reduce the blur. The proposed joint method (✹) performs best on these.

3.6.2 Flow Comparison

In Figure 3.8, the quality of warped images using a direct flow computation [Teed and Deng, 2020] as well as the proposed network **`flowMerger`** (refer to Section 3.4) are compared. First, a missing low exposure is warped using flow $F^{M}_{0\to-1}$, which is unreliable for large saturated regions. Second, the flow $F^{L}_{-1\to+1}$ is employed, whose magnitude is adjusted based on the linear motion assumption, and denoted as flow $\widetilde{F}^{L}_{0\to-1}$, to frame $\mathtt{I}_t$ (refer to Figure 3.6). Finally, the network **`flowMerger`** results in

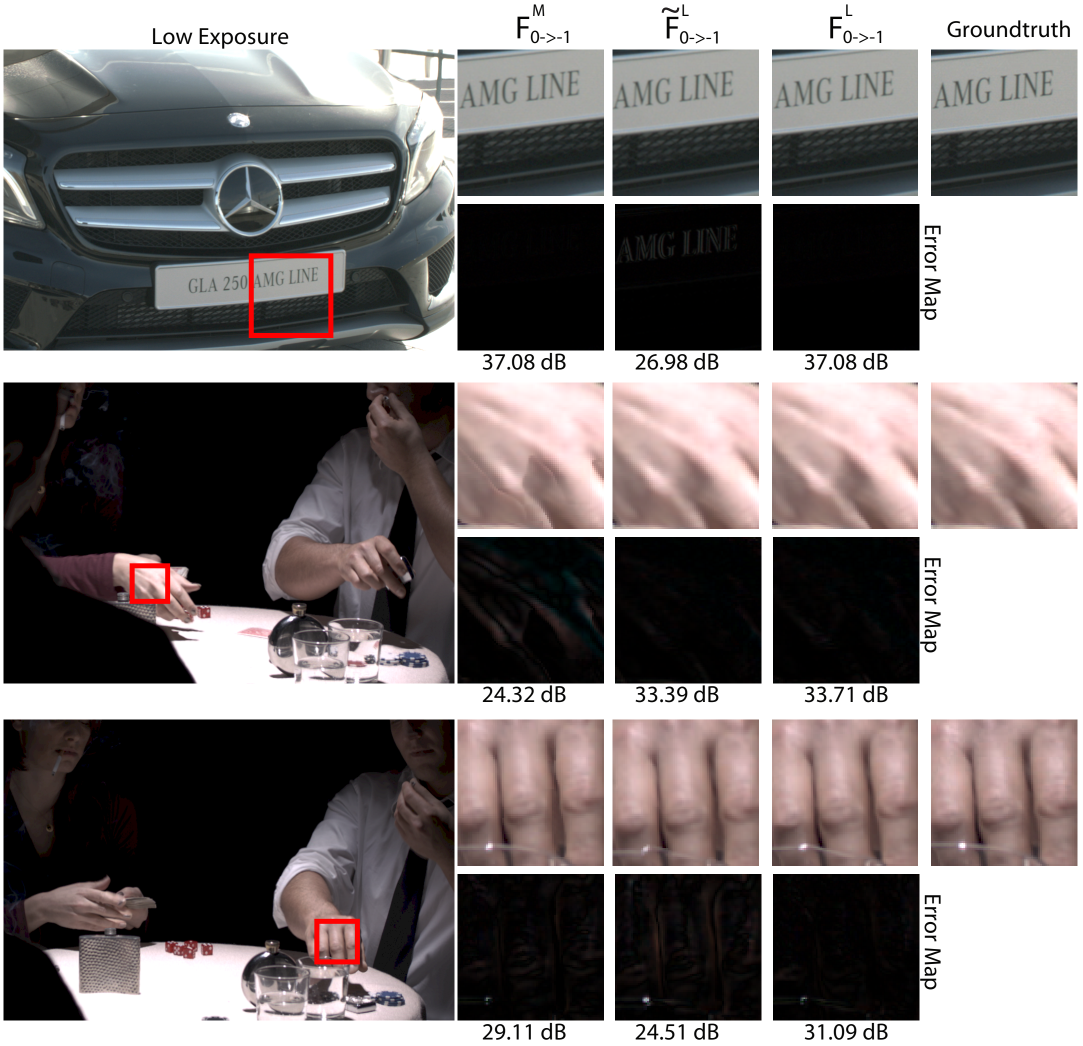


FIGURE 3.8: *Flow quality comparison. Left: an example of warped low exposure. Right: cropped patches with moving hand from the warped exposure for different flows: $F^{M}_{0\rightarrow-1}$, $\widetilde{F}^{L}_{0\rightarrow-1}$, and $F^{L}_{0\rightarrow-1}$ as discussed in Section 3.4 and Section 3.6.2. Under each inset, a PSNR measure with respect to the ground truth in the right column is presented.*

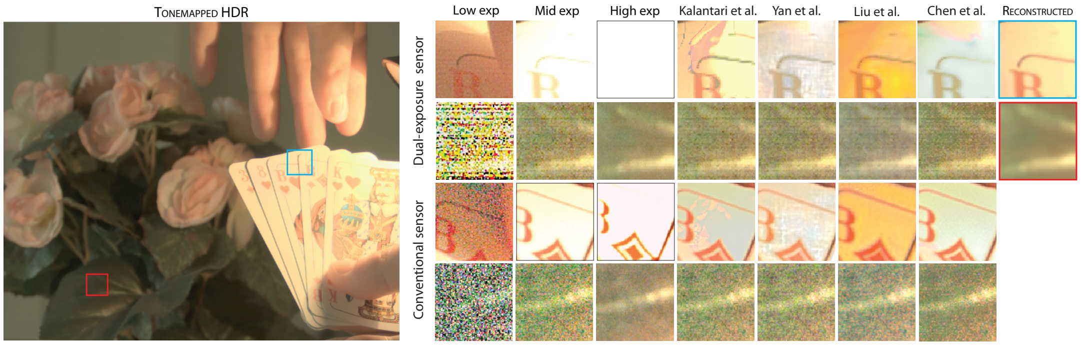


FIGURE 3.9: *Comparison of different HDR video reconstruction methods: Kalantari and Ramamoorthi [2017], Yan et al. [2020], Liu et al. [2020c], Chen et al. [2021], and the proposed method for a dynamic scene with moving hands and cards. The figure layout follows the one in Figure 3.1.*

refined flow $F^{L}_{0\rightarrow-1}$ that leads to a more precise motion reconstruction than its linear interpolation $\widetilde{F}^{L}_{0\rightarrow-1}$ between frames $\mathtt{I}_{t-1}$ and $\mathtt{I}_{t+1}$.

The upper scene in Figure 3.8 demonstrates the case where flow $F^{M}_{0\rightarrow-1}$ is reliable,

TABLE 3.2: *Ablation results of the loss components of Eq. 3.6d.*

Loss function	PSNR
L_{f_1}	43.36
$L_{f_1} + L_{f_2}$	43.57
$L_{f_1} + L_{f_2} + L_{f_3}$	**43.63**

and as a result, the **`flowMerger`** successfully maintains this flow information, leading to a perfect alignment. In the second example, due to large saturation in the hand, $F^{M}_{0\to-1}$ fails to align; however, **`flowMerger`** this time uses the flow information from $\widetilde{F}^{L}_{0\to-1}$ that better aligns with the ground truth. The last example is the special case where both $F^{M}_{0\to+1}$ and $\widetilde{F}^{L}_{0\to-1}$ fail, and the refined flow $F^{L}_{0\to-1}$ successfully corrects the alternative flows and improves the alignment with the ground truth.

The effect of each component of the loss function in Eq. 3.6d (refer to Section 3.4.1) are ablated. The effect of different loss components is measured as specified in Table 3.2 by re-training the network **`flowMerger`**. PSNR measurements have been done using 70 frames extracted from the HDR video dataset provided in [Froehlich et al., 2014]. Adding subsequent terms in the loss function (refer to Eq. 3.6d) improves the flow quality. Note that respective PSNR values for flows between mid exposures $F^{M}_{0\to-1}$ and low exposure exposures $\widetilde{F}^{L}_{0\to-1}$ are 42.61 and 42.37. Each variant of the proposed method outperforms the state-of-the-art flow algorithm as presented in [Teed and Deng, 2020].

3.6.3 HDR Video Denoising/Deblurring Evaluation

Methods Three methods (color-coded; "Method" in Table 3.3) are used for the comparisons: Kalantari [Kalantari and Ramamoorthi, 2017], which blends all exposures by first aligning the low and high exposure to the mid exposure. Next, Yan [Yan et al., 2020], which directly fuses all three exposures without estimating the flow between them. Chen [Chen et al., 2021] is the state-of-the-art HDR video reconstruction method that takes temporally alternating LDR exposures captured by conventional sensors as an input and adopts an elaborated coarse-to-fine scheme to output the final HDR image.

Training data Similar to Table 3.1, also each method is evaluated when trained with different data ("Train. data" in Table 3.3). It is denoted as "Theirs" (▼) if the authors provide a pre-trained version. Since all other methods are trained by single-exposure sensors with different noise characteristics models, they are re-trained with the synthetic training data, and it is referred to as ("All", ✷) in Table 3.3.

Metrics The following metrics are used: PSNR, SSIM [Wang et al., 2004b], and HDR-VDP-3, which is the latest version of [Mantiuk et al., 2011b], where more is better.

Tasks Here, two tasks (two columns in Table 3.3) are studied: First, MB remains in the output HDR (LOMIDHI2HDR). Second, noise and MB are removed (LOMIDHI2HDR-MB). Note that in both tasks, Liu [Liu et al., 2020c] takes only the mid exposure as an input.

The exposure ratio of 1:4:16 is considered between the three exposures in all tasks. The test set for all tasks contains 16 frames from 7 different video sequences from High-speed Video Dataset [Sim et al., 2021] and [Janai et al., 2017].

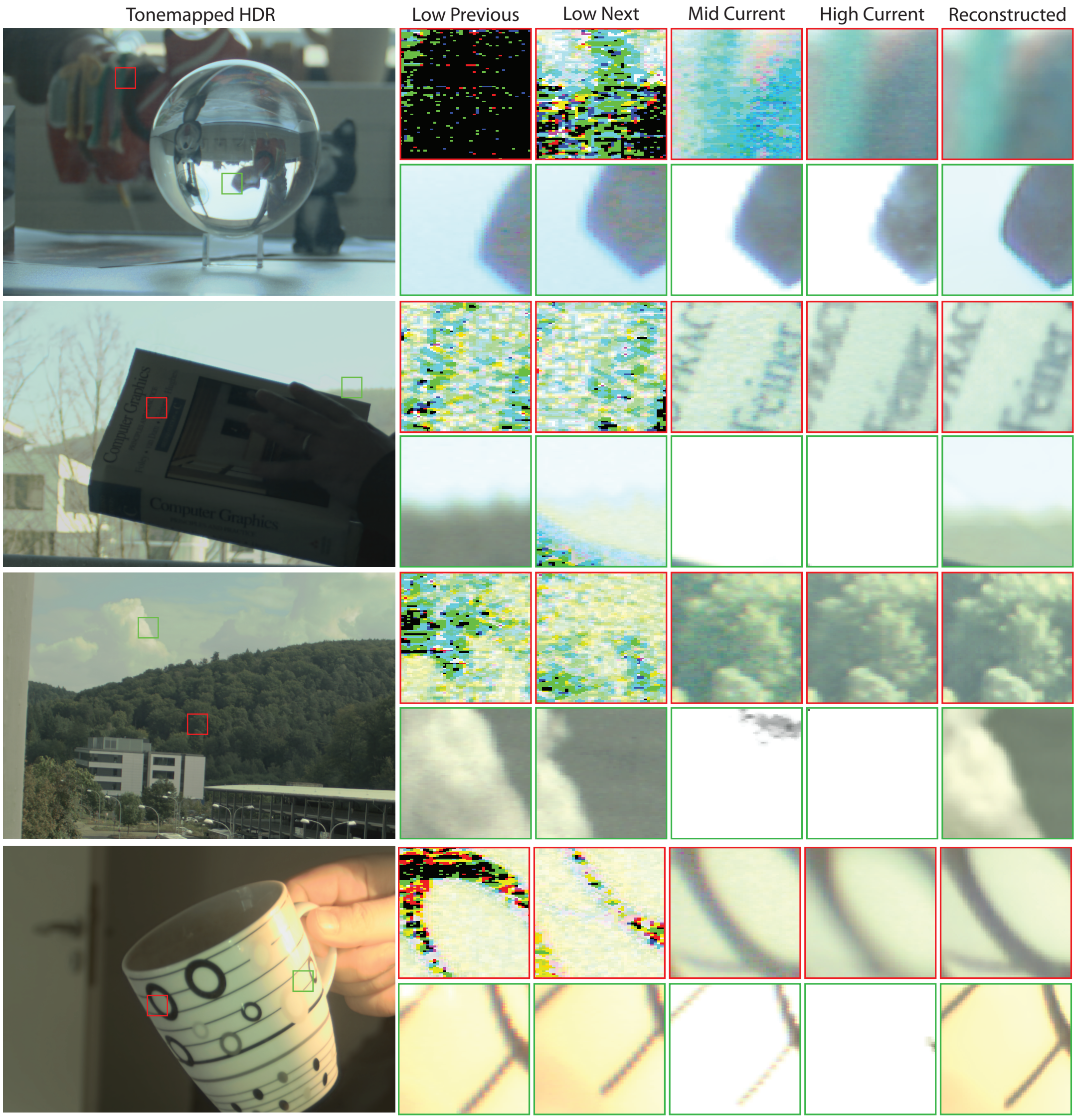


FIGURE 3.10: *HDR frame reconstruction for video sequences with large motion magnitudes (first column). In the presented setup, only mid and high exposures are available for the current frame, while the missing low exposure is reconstructed from the previous and next frames. The difference between the low exposures illustrates the magnitude of motion. Note that all three exposures are upsampled to match their size with the HDR reconstruction (last column).*

Discussion Quantitative results are shown in Table 3.3. The proposed method trained on the synthetic training data (✷) performs best on all tasks. In addition to this, Kalantari [Kalantari and Ramamoorthi, 2017] and Yan [Yan et al., 2020] rely heavily on the mid exposure, which causes problems in dark regions as can be seen in Figure 3.1 and Figure 3.9. As a result, other methods give better results when they are compared with the HDR ground truth that is blur-free. Moreover, when other methods are trained exclusively with the proposed training data, their performance increases (LoMidHi2HDR-MB).

The quantitative results from above are complemented by the qualitative ones in Figure 3.1 and Figure 3.9. In both figures, rows 1 and 3 show insets taken from a bright part of the scene, while rows 2 and 4 show a dark part. Since other methods were originally trained with single-exposure sensors, similar to the ones used in rows 3 and 4, the single-exposure mode is also used on the multi-exposure sensor for those methods (rows 1 and 2). Only the proposed method uses multi-exposure mode (last

TABLE 3.3: *Measured PSNR, SSIM and HDR-VDP-3 results between the multi exposure techniques.*

		Task	
		LoMidHi2HDR	LoMidHi2HDR-MB
	MB removed	✗	✓
	Output: HDR	✓	✓
Train. data	Method	Error (PSNR\SSIM \HDR-VDP-3)	
▼ Theirs	Kalantari and Ramamoorthi [2017]	31.12\0.896\8.47	39.67\0.952\9.55
✹ All		—	38.19\0.975\9.81
▼ Theirs	Yan et al. [2020]	30.20\0.846\8.26	32.48\0.848\8.95
✹ All		—	37.36\0.960\9.95
▼ Theirs	Liu et al. [2020c]	29.53\0.793\8.78	27.53\0.817\8.97
▼ Theirs	Chen et al. [2021]	32.66\0.927\9.70	37.40\0.949\9.90
✹ All	The Proposed Method	**39.67\0.975\9.77**	**41.38\0.982**\9.85

column). Because of capturing dynamic scenes, the waving tissue in Figure 3.1 and the moving hands and cards in Figure 3.9, with two sensors, the captured scenes are not fully identical, so insets cannot be compared pixel-wise. As can be seen, any single (Low, Mid, and High) exposure on any sensor does not capture the bright or dark part. In Figure 3.1 Kalantari [Kalantari and Ramamoorthi, 2017], Yan [Yan et al., 2020] and Chen [Chen et al., 2021] fail to reconstruct the occluded region properly (rows 1 and 3) due to the saturation in mid and high exposures. Liu [Liu et al., 2020c] that takes only the mid exposure as an input fails to recover the saturated region (row 1). In Figure 3.9 Kalantari [Kalantari and Ramamoorthi, 2017] leads to some shading discontinuity artifacts, while Yan [Yan et al., 2020] and Chen [Chen et al., 2021] additionally result in color distortions (rows 1 and 3). A consistent color shift towards yellow can be observed in Liu [Liu et al., 2020c] (rows 1 and 3). All of the compared methods can not exploit the high exposure to properly reconstruct the dark region in the curtain (Figure 3.1) and leaves (Figure 3.9), which results in extensive noise (rows 2 and 4). In contrast, the proposed method handles both the saturated regions in the background and the dark region (last column).

Figure 3.10 further demonstrates the importance of low, mid, and high exposures for HDR frame reconstruction in the presence of high-magnitude motion. The red insets present darker image regions where the mid and high exposures directly contribute to the currently reconstructed HDR frames. The green insets present bright image regions, where the mid and high exposures are mostly saturated so that the current HDR frame reconstruction is mostly based on the low exposures from the previous and next frames that are first warped and then blended (refer to Figure 3.3).

3.6.4 Application: HDR Illumination Reconstruction

A key application of HDR is to use it for illumination [Debevec and Malik, 1997]. The mirror ball is captured, motion blur and noise are removed using the proposed method (✹) and re-rendered using Blender's [Community, 2020] path tracer with 512 samples and automatic tone and gamma mapping. The resulting image is seen in Figure 3.11. It is found that the non-linear mapping of MC rendering amplifies structures, making noise, particularly row noise, more visible. Using only the high exposure removes noise but cannot capture the dynamic range, resulting in washed-out shadows. The proposed method succeeds in removing it, particularly row noise, resulting in sharp shadows and noise-free reflections. Note that all images contain

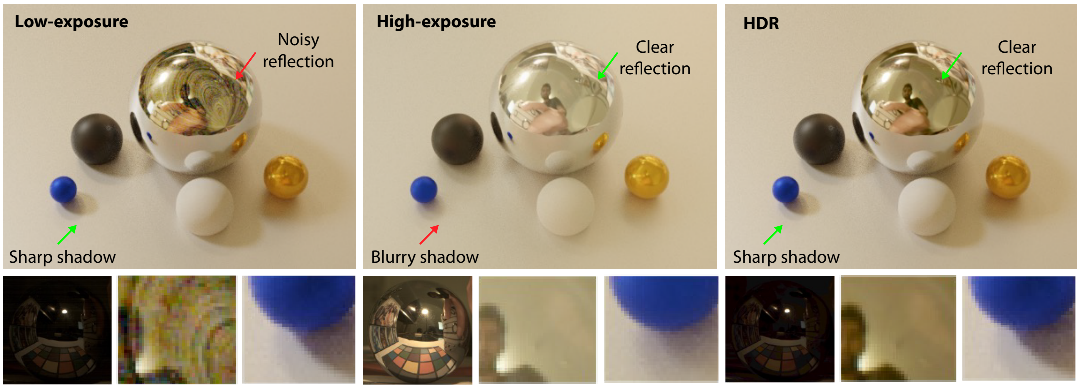


FIGURE 3.11: *Rendering from a spherical illumination map captured at a low exposure **(left)**, a high exposure **(middle)** and using the proposed approach **(right)**. The illumination is seen as an inset on the left for each approach. For the low exposure, the shadows are sharp, as the light source did not saturate, but the dark regions are clipped and massively noisy. For the high exposure, the dark regions are reproduced, slightly noisy, but the light source is clamped, leading to a loss in dynamic range and sharp shadows. However, the proposed method reproduces both. Note that visible overall brightness differences are expected, as clamping is present in some images, which does not conserve energy.*

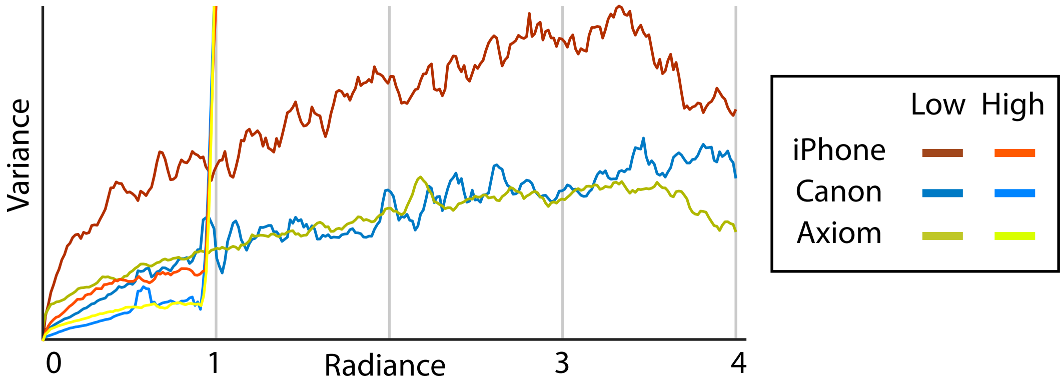


FIGURE 3.12: *Noise for contemporary sensors at different exposures and intensity: The horizontal axis is unit radiance. The vertical axis is variance (less is better). Different hues depict different sensors. Bright colors are high, and dark colors are low exposure.*

some noise due to the finite MC sample count (all images computed 20 min.). The noise appears less in the high exposure, as reduced contrast results in an easier light simulation problem that leads to an overall incorrect, strongly biased solution.

3.7 Exposure Control on Modern Sensors

To better understand the trade-off between single- and multi-exposure sensors, a pilot experiment is conducted to evaluate the exposure-dependent noise for three different kinds of sensors: iPhone (Apple iPhone 8), Canon (Canon EOS 550D) and Axiom-beta (CMOSIS CMV12000; a full-frame single-exposure setup). 600 images of the same scene have been captured in low- and high-exposure (four times longer) modes for each sensor. A Canon records 14, iPhone, and Axiom-beta 12 bits. All readings are converted to floating point values between 0 and 4. High exposure is divided by four to match the same range. For low exposure, an ideal (as the scene is static) burst fusion is simulated by averaging random four-tuples. The average of all low-exposure frames is considered the reference for each sensor. Note that by construction, the reference of the high and low modes is the same. Then, for every quantized (12- or 14-bit) value L of the reference of each sensor, one pixel is selected

TABLE 3.4: *The network architecture details as discussed in Section 3.5: the number of channels in convolutional layers and dilation factors. Note that layer pairs: 4-6, 3-7, 2-8, and 1-10 are connected with residual connections [He et al., 2016].*

Layers	Input Channels	Output Channels	Dilation Factor
1	8	64	1
2	64	64	2
3	64	64	4
4	64	64	8
5	64	64	16
6	128	64	16
7	128	64	8
8	128	64	4
9	128	64	2
10-11	64	64	1
12	64	8	1

with that value and the variance $\mathrm{Var}(L)$ is computed for all readings in all images. A high value means that a particular sensor for this mode and this absolute radiance has more noise (worse).

Figure 3.12 shows that for all sensors, as expected, noise increases with signal [Granados et al., 2010; Janesick, 2001]. It is further seen that around 0.25, the variance for high exposure diverges (clipping), indicating that these or even higher values cannot be used with long exposure. More importantly, it is also observed that low exposure has a higher variance until the point where the high exposure clips. This trend is true for all sensors, so between 0 and 1: every sensor (hue) at its low exposure mode (brightness) has a higher variance than the high exposure. This can be attributed to the read noise of each burst frame that is accumulated [Ma et al., 2020]. This indicates that combining low exposures, even under the ideal condition of no motion, is no immediate solution. In summary, no single strategy of either averaging low exposures or just using one high exposure is successful across the entire HDR range. Sensors benefit from having access to different exposures at different spatial locations. Given ML strengths, which ideally complement sensor weaknesses, it might be worth revisiting different HDR capturing approaches.

3.8 Network Architecture

Experimental results with U-Net [Ronneberger et al., 2015] resulted in artifacts along the edges even when there is no motion. This could be attributed to the side effects of downsampling and upsampling of the channels inside the network, which cause misalignment problems between high and low-exposure channels. Additionally, replacing U-Net with bilinear upsampling layers with so-called transposed convolutional layers did not cause such misalignment problems, but it ended with vertical stripes. Resulting patches can be observed in Figure 3.13. This led to the avoidance of using U-Net architecture due to its internal structure.

TABLE 3.5: *Standard deviation measurements of the PSNR values in the Table 3.1.*

		Task				
		LO2LO$_{\text{SENSOR}}$	LO2LO	HI2HI-MB	LOHI2HDR	LOHI2HDR-MB
	Input: Low Exp.	✓	✓	✗	✓	✓
	Input: High Exp. with MB	✗	✗	✓	✓	✓
	MB removed	✗	✗	✓	✗	✓
	Output: LDR	✓	✓	✓	✗	✗
	Output: HDR	✗	✗	✗	✓	✓
Train. data	Method	PSNR ± standard deviation				
▼ Theirs	Direct [Debevec and Malik, 1997]	37.66±2.85	33.27±5.04	28.86±7.43	34.23±7.12	33.26±5.96
▼ Theirs	BM3D [Dabov et al., 2007]	36.37±2.74	33.98±6.09	28.93±7.59	34.24±6.89	33.03±5.99
▲ Sensor		39.74±2.81	34.47±5.51	—	34.35±7.05	—
● HetGau		39.12±2.83	34.97±6.01	29.05±7.64	34.63±7.21	33.30±5.96
✹ All		39.76±2.83	34.99±6.05	29.05±7.64	34.69±7.24	33.33±6.00
▼ Theirs	FFDNet [Zhang et al., 2018a]	38.50±3.20	34.44±5.91	28.98±7.59	33.78±6.81	32.33±5.68
▲ Sensor		38.59±3.23	34.37±5.73	—	34.13±6.85	—
✹ All		38.67±3.23	34.23±6.19	28.96±7.60	34.37±6.99	33.14±5.90
▼ Theirs	DBGAN [Kupyn et al., 2019]	—	28.97±4.29	26.76±5.48	31.16±4.88	30.35±4.39
▼ Theirs	SRN-DB [Tao et al., 2018]	—	31.53±4.56	27.72±7.21	32.14±5.77	31.38±5.28
▼ Theirs	LSD$_2$ [Mustaniemi et al., 2020]	—	—	—	29.94±5.72	32.09±5.06
▼ Theirs	Heide et al. [2014]	—	—	—	—	34.12±5.18
▲ Sensor	`denoise`	33.79±2.13	28.05±4.86	—	29.01±6.04	—
✺ Foi		41.85±3.36	37.72±4.39	35.92±5.50	39.41±5.20	39.01±3.80
● HetGau		40.98±3.28	36.39±3.79	35.68±6.98	38.71±5.42	37.58±3.53
✦ OnlyRN		39.58±3.96	33.94±4.94	32.19±6.79	35.32±6.00	35.24±5.10
★ OnlyPN		38.48±3.59	35.02±5.04	31.97±5.31	36.08±6.60	36.24±5.40
✸ OnlyMB		39.97±2.78	36.06±3.89	34.90±5.06	37.91±6.19	37.62±3.50
✹ All		42.56±3.12	38.11±4.46	36.22±6.98	39.71±5.20	39.07±4.16

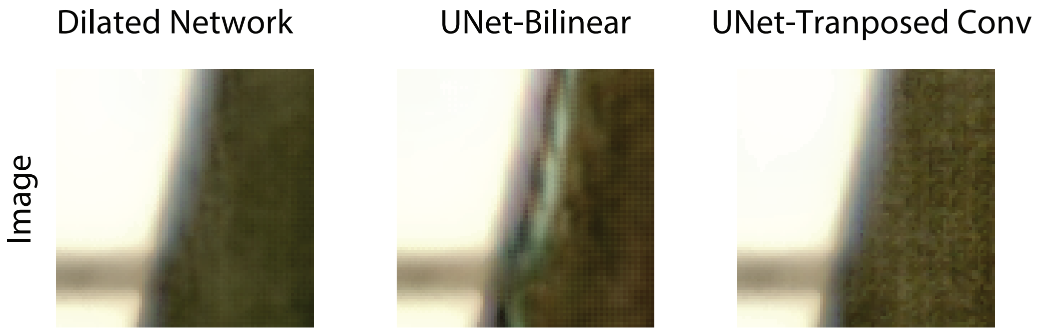


FIGURE 3.13: *U-Net with bilinear up-sampling causes severe artifacts along the edges while the transposed convolutional layer prevents it. On the other hand, U-Net with transposed convolutional layers produced vertical stripes in dark regions. A dilated network that results in artifact-free images both along edges and in dark regions.*

3.9 Additional Results

In Table 3.5, for completeness, the standard deviation of PSNR values is also measured, as presented in Table 3.1. An example scene employed with a ratio of 1:16 is provided in Figure 3.15.

3.10 Limitations

Figure 3.14 shows a failure case in a dynamic water splash sequence that includes both strongly illuminated and shadowed components. As shown in the inset, the mid exposure is mostly saturated in the illuminated region, so the missing low exposure should be interpolated using exclusively captured low exposures. Due to the motion

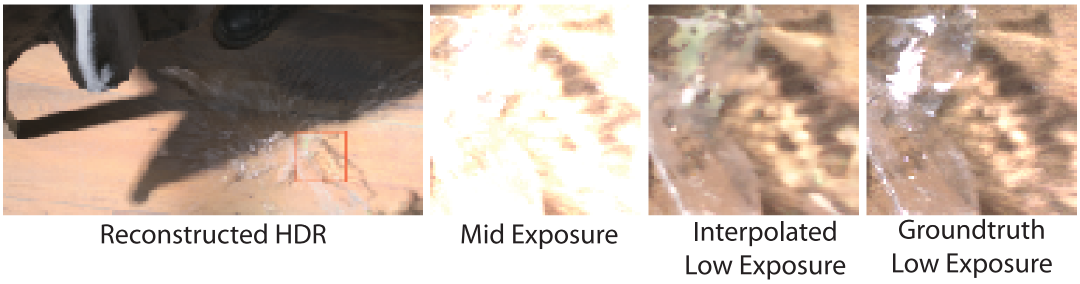


FIGURE 3.14: *Example failure case of* **`flowMerger`** *network in the presence of complex motion.*

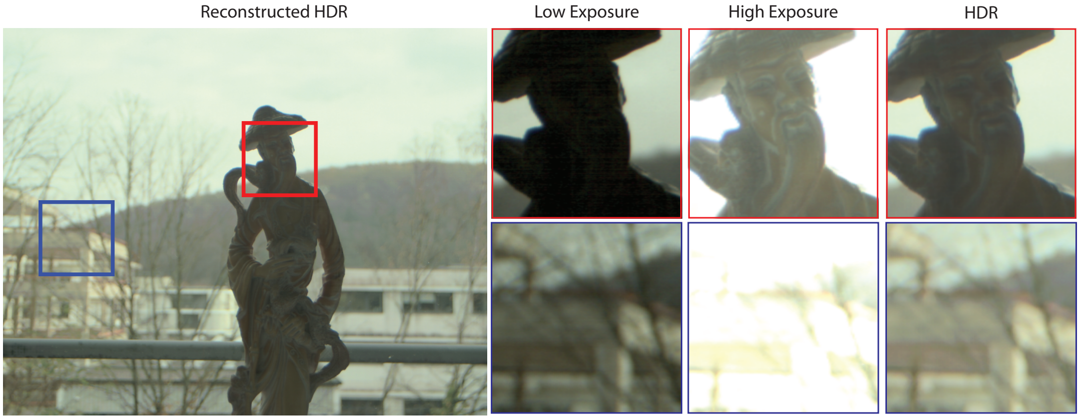


FIGURE 3.15: *Example of scene captured with the ratio of 1:16.*

pattern complexity and sparse temporal sampling, the optical flow computation fails, and clear artifacts can be observed in the reconstructed low exposure. This indicates that the performance of the proposed solution can be reduced for scenes that simultaneously include complex motion in dark and bright regions. As shown in the green insets in Figure 3.10, saturation of mid and high exposures still leads to a reasonable low exposure reconstruction for simple motion patterns.

It is found that the chosen ratio of 1:4 is a good trade-off between the captured dynamic range and the quality of the resulting HDR video, and it is expected that other ratios might be required only sparsely. As increasing gain boosts the noise, re-capturing the training data might be required. To determine the specific requirements for re-capturing the training data, an analysis of how much gain changes affect the proposed noise histograms is needed.

3.11 Conclusion

This chapter presents a CNN solution for HDR image and video reconstruction that is tailored for both single-shot capturing with spatially interleaving exposures and multi-shot capturing with spatially interleaving and temporally alternating exposures. In the single-shot scenario, the proposed solution solves a number of serious problems inherent to multi-exposure sensors by joint processing low and high exposures and taking advantage of their perfect spatial and temporal registration. These include correlated noise and spatially-varying blur, interlacing, and spatial resolution reduction. In the multi-shot scenario, registration of subsequent frames for multi-exposure sensors is greatly facilitated with respect to traditional single-exposure sensors, in particular for large saturated or occlusion regions. The proposed CNN solution capitalizes on such factors by effectively merging optical flows that are originally derived between corresponding exposures. The proposed flow merging, along with capturing mid-exposure half-frames for every video frame, also greatly improves the quality of motion for all exposures, particularly in the presence of non-linear motion. It is demonstrated that synthetic training data is generated by capturing a limited amount of data specific to multi-exposure sensors and using simple histograms to represent the noise statistics. This leads to a better denoising and deblurring quality than achieved by existing state-of-the-art techniques. Moreover, it is shown that using limited sensor-specific data can greatly improve the performance of other techniques. This is for two reasons: First, previous methods did not have access to massive amounts of training data for multi-exposure sensors. A problem is solved here by proposing the first dedicated distortion model that allows synthesizing training data. Second, multi-exposure sensors, in combination with proper CNN-based denoising and deblurring, provide much richer data that can be fused. Finally, the application of captured HDR environment maps for 3D scene re-lighting is presented, where the denoising and deblurring improve the quality of Monte Carlo (MC) rendering. An exciting area of future work is a more systematic investigation of how CNN solutions can compensate for sensor weaknesses and vice versa, which should lead to novel sensor designs and new challenges for CNN techniques to process such data.

Chapter 4

Video Frame Interpolation for High Dynamic Range Sequences

This chapter introduces a video frame interpolation (VFI) methodology that aims to increase the frame rate of captured videos while reconstructing high dynamic range (HDR) content. Video frame interpolation enables many important applications, such as slow-motion playback and frame rate conversion. However, one major challenge in using VFI is accurately handling high dynamic range scenes with complex motion. To this end, the possible advantages of multi-exposure sensors are that they readily provide sharp short and blurry long exposures that are spatially registered and whose ends are temporally aligned. This way, motion blur registers temporally continuous information on the scene motion that, combined with the sharp reference, enables more precise motion sampling within a single camera shot. It is demonstrated that this facilitates a more complex motion reconstruction in the VFI task and HDR frame reconstruction, which has so far been considered only for the originally captured frames, not in between interpolated frames. The designed neural network trained in these tasks clearly outperforms existing solutions. The proposed metric of scene motion complexity provides important insights into the performance of VFI methods at test time.

4.1 Introduction

Video frame interpolation enables many interesting applications ranging from video compression and frame rate up-conversion in TV broadcasting to artistic video effects such as speed ramp in professional cinematography. The performance of VFI methods is largely affected by various factors such as scene lighting conditions, the magnitude and complexity of motion in the scene, the spatial extension of resulting motion blur, the presence of complex occlusions, or thin structures in the scene. Popular VFI methods [Jiang et al., 2018; Bao et al., 2019; Sim et al., 2021] mostly rely on well-exposed frames in the captured video. Nevertheless, in the case of high dynamic range scenes captured using traditional single-exposure sensors, undesired under- and over-exposure effects might appear. The resultant noise and intensity clamping can adversely affect the quality of VFI as finding the pixel correspondence between the frames becomes more ambiguous. Another major challenge is the large and non-uniform motion in the scene. Although recent methods [Reda et al., 2022; Sim et al., 2021] have shown progress in handling large motion, they typically heavily rely on the motion linearity assumption that might not hold in practice. Explicit handling of non-linear motion becomes possible by processing more than two subsequent frames [Xu et al., 2019; Park et al., 2021]; however, temporal sampling might still be too low for reliable motion reconstruction. Motion blur due to low shutter speed and long

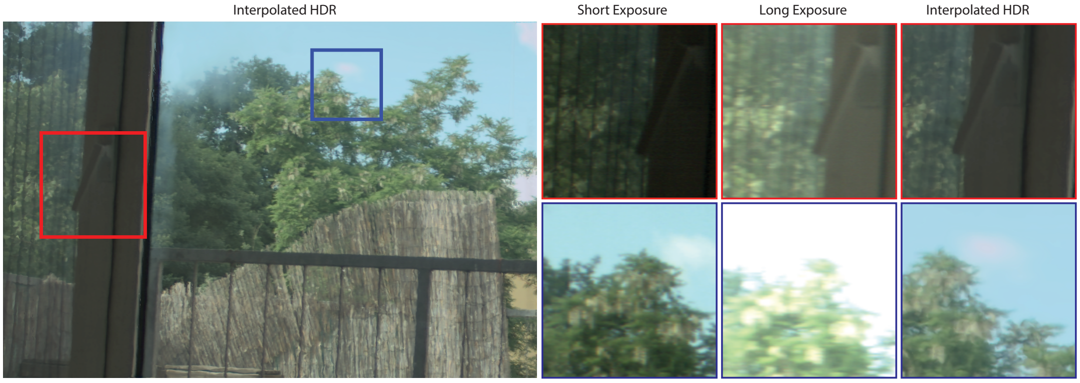


FIGURE 4.1: *This section proposes a method for high dynamic range video frame interpolation for multi-exposure sensors that have gained popularity due to their use in recent smartphones. The first column shows an interpolated HDR frame, while the insets focus on the dark and bright scene details. Note that the short and long exposures, as captured by the sensor (middle columns), are shifted with respect to the interpolated HDR frame (right column). The dark region (the upper row) requires a long exposure duration and features significant motion blur due to camera motion. The proposed methodology employs temporally continuous information on the scene motion that is encoded in motion blur to improve the VFI quality. At the same time, the short exposure avoids pixel saturation in the sky region (the bottom row) and enables its reconstruction in the interpolated HDR frame.*

exposure times further leads to spatial and temporal loss of image details. For this reason, handling blurry frames is typically treated as a challenge in the VFI task [Shen et al., 2020a; Zhang et al., 2020], while potentially, motion blur encodes continuous temporal information on the magnitude and direction of motion, particularly for large motion.

In this chapter, such sensor capabilities, as in Chapter 3, are explored to improve the motion estimation accuracy in VFI. In particular, a multi-exposure sensor is considered that captures short and long exposures for spatially interleaved pixel columns in a single shot [CMV12000, 2021]. Importantly, while the exposure duration differs, the exposure completion is temporally aligned, which enables the recovery of two temporal samples of scene motion that are perfectly spatially registered at the sensor. It is shown that such an increased temporal sampling rate substantially improves the accuracy of complex motion interpolation, as motion non-linearity can readily be reconstructed for two subsequent frames. Furthermore, the short exposure typically leads to a sharp image, while the long exposure results in substantial motion blur, providing additional insights into the motion direction and magnitude (Figure 4.1). This is of particular importance in dark scene regions, where the short exposure might be strongly underexposed and noisy, and the long exposure becomes the only reliable measurement of scene motion. As in other works, a multi-exposure technique is employed to reconstruct HDR video frames, but for the first time, simultaneous VFI is performed that can handle complex, non-linear motion in the scene. The end-to-end convolutional network is trained to achieve those goals. Additionally, the metric for motion non-linearity is proposed to analyze the existing high-speed videos and the performance of VFI methods as a function of motion complexity.

In the following Section 4.2, the HDR-VFI method for HDR sequences is presented. In Section 4.3, the metric of scene motion uniformity that enables meaningful comparison of existing VFI methods is introduced. Then, Section 4.4 provides implementation details of the proposed network. Section 4.5 contrasts the HDR-VFI technique with existing works in a performance comparison and reports an outcome of ablation studies. Finally, this chapter is concluded in Section 4.6.

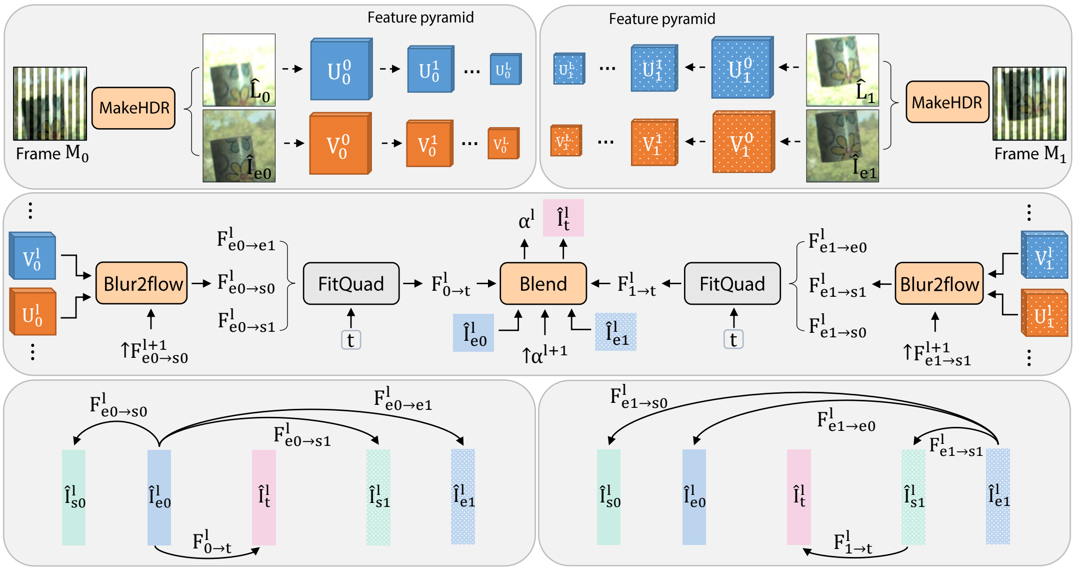


FIGURE 4.2: *Overview of the proposed HDR VFI pipeline.* Upper row: *Two subsequent frames* M_0 *and* M_1*, as captured using the multi-exposure sensor, are independently processed by a learned* **MakeHDR** *network so that the output sharp HDR frames* $\hat{I}_{ei}$ *aligned with the end of the long exposure (the suffix* **e** *stands for the end) and blurry long exposure frames* $\hat{L}_i$ *are obtained. Next, each frame is fed separately to a feature extractor to build feature pyramids.* Middle row: *At each pyramid scale l, given the features* V^l_i *and* U^l_i *along with* $\uparrow F^{l+1}_{ei\to si}$ *upsampled from the previous scale, the intra-frame flow* $F^l_{ei\to si}$*, the flow between the start (denoted with the suffix* **s***) and end of the long exposure in each frame is recovered using a learned* **Blur2Flow** *network. Then, bidirectional flows* $F^l_{e0\to e1}$ *and* $F^l_{e1\to e0}$ *estimated between* $\hat{I}^l_{e0}$ *and* $\hat{I}^l_{e1}$ *(which are the sharp HDR frames* $\hat{I}_{e0}$ *and* $\hat{I}_{e1}$ *downsampled by* 2^l*) are found using the state-of-the-art flow estimation method Raft [Teed and Deng, 2020]. Next, given two estimated flows for each frame, the additional flows of* $F^l_{e0\to s1}$ *and* $F^l_{e1\to s0}$ *are also derived. The motion flow triplets* $(F^l_{e0\to e1}, F^l_{e0\to s0}, F^l_{e0\to s1})$ *as well as* $(F^l_{e1\to e0}, F^l_{e1\to s1}, F^l_{e1\to s0})$ *are independently fed to a non-learnable* **FitQuad** *module to calculate the forward flows* $F^l_{0\to t}$ *and* $F^l_{1\to t}$ *that are parametrized using a quadratic motion model for a position t (refer to the two bottom insets). Finally, using the module* **Blend**, $\hat{I}^l_{e0}$ *and* $\hat{I}^l_{e1}$ *are fused with the forward flows* $F^l_{0\to t}$ *and* $F^l_{1\to t}$ *and a soft occlusion map* $\uparrow \alpha^{l+1}$ *upsampled from the previous scale to reconstruct the intermediate frame* $\hat{I}^l_t$ *at scale l. This procedure is repeated until the scale of the original input frames is reached.* Bottom row: *A schematic presentation of all involved flows and their relation to the input and interpolated frames.*

4.2 Method

This section proposes a VFI method that reconstructs HDR frames in the continuous-time domain. Figure 4.2 summarizes the processing pipeline, and the following paragraphs provide a more detailed description of its key components.

The proposed methodology takes as input two subsequent video frames M_0 and M_1 that are captured using the multi-exposure sensor and produces a sharp HDR frame $\hat{I}_t$ for any position t between M_0 and M_1. Each captured frame M_i, where with the suffix i any input frame is denoted, contains a pair of spatially interleaved short and long exposures and is processed by the **MakeHDR** network to produce a sharp HDR frame $\hat{I}_{ei}$ that is aligned with the end of the long exposure (the suffix **e** stands for the end), and a blurry long exposure frame $\hat{L}_i$. Both frames are decomposed into their respective multi-resolution feature pyramids, and from this stage, the whole processing is performed at different scales, where, as shown in the middle row in Figure 4.2, information reconstructed at a lower-resolution scale $l+1$ contributes to the higher-resolution scale l. Here, for brevity, the scale index l is omitted. The feature pyramids are fed to the **Blur2Flow** network to predict the flow $F_{ei\to si}$ that extracts the flow between the start (denoted with the suffix **s**) and the end of the long exposure.

Next, the flows $\mathtt{F}_{\mathtt{e0}\rightarrow\mathtt{e1}}$ and $\mathtt{F}_{\mathtt{e1}\rightarrow\mathtt{e0}}$ are computed between the sharp HDR frames $\hat{\mathtt{I}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}_{\mathtt{e1}}$ in both directions using an off-the-shelf flow estimation method such as Raft [Teed and Deng, 2020]. This way, the flows $\mathtt{F}_{\mathtt{e0}\rightarrow\mathtt{s0}}$ and $\mathtt{F}_{\mathtt{e0}\rightarrow\mathtt{e1}}$ that are aligned with $\hat{\mathtt{I}}_{\mathtt{e0}}$ are obtained, then additionally, the flow $\mathtt{F}_{\mathtt{e0}\rightarrow\mathtt{s1}}$ is derived, and all three flows are employed to fit a quadratic motion model using a non-learnable **FitQuad** module. This process is repeated for the flows $\mathtt{F}_{\mathtt{e1}\rightarrow\mathtt{s1}}$, $\mathtt{F}_{\mathtt{e1}\rightarrow\mathtt{e0}}$, and $\mathtt{F}_{\mathtt{e1}\rightarrow\mathtt{s0}}$ that are aligned with $\hat{\mathtt{I}}_{\mathtt{e1}}$. Refer to the bottom row in Figure 4.2 for the depiction of the discussed flows. Next, to warp the keyframes $\hat{\mathtt{I}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}_{\mathtt{e1}}$ to a novel temporal position t, first the forward flows $\mathtt{F}_{\mathtt{0}\rightarrow\mathtt{t}}$ and $\mathtt{F}_{\mathtt{1}\rightarrow\mathtt{t}}$ are found, and then the backward flows $\mathtt{F}_{\mathtt{t}\rightarrow\mathtt{0}}$ and $\mathtt{F}_{\mathtt{t}\rightarrow\mathtt{1}}$ are computed using differentiable flow reversal as introduced in [Xu et al., 2019]. Finally, using a multi-scale blending scheme **Blend**, the warped images are combined with a soft occlusion weight at different scales to synthesize the frame $\hat{\mathtt{I}}_{\mathtt{t}}$. More details are now provided on all the processing steps discussed here.

HDR reconstruction: `makeHDR` The input video is acquired using a multi-exposure sensor [CMV12000, 2021] that simultaneously captures each frame's short and long exposure. In the proposed capturing setup, the exposure time for the long exposure is four times higher than the short exposure. Each exposure is stored at odd and even columns in the sensor. As a result, both exposures are provided as half-resolution images, and they need to be upsampled in the horizontal direction. Moreover, the short exposure exhibits strong noise in dark scene regions and requires denoising. On the other hand, the long exposure is less noisy, while it might contain considerable motion blur and requires deblurring. To do so, the network design and the training strategy introduced in Section 3.4 are employed to jointly deblur, denoise, and upsample the input frames $\mathtt{M}_{\mathtt{i}}$ to produce sharp, clean, and full-resolution short and long exposures. Both exposures are combined using a non-learnable technique, similar to [Debevec and Malik, 2008], to produce a sharp HDR frame $\hat{\mathtt{I}}_{\mathtt{ei}}$. Also, the network output is extended to produce an additional full-resolution blurry long exposure $\hat{\mathtt{L}}_{\mathtt{i}}$.

Motion from blur: Blur2Flow Motion blur can potentially reveal information about the motion in the scene. This idea is pursued, and the **Blur2Flow** network is proposed that derives the motion flow $\mathtt{F}_{\mathtt{ei}\rightarrow\mathtt{si}}$ that is associated with the blur pattern in the long exposure $\hat{\mathtt{L}}_{\mathtt{i}}$. The sensor design ensures that the short and long exposures are completed precisely at the same time, and in the HDR reconstruction, the sharp frame $\hat{\mathtt{I}}_{\mathtt{ei}}$ is aligned with the short exposure. Given $\hat{\mathtt{L}}_{\mathtt{i}}$ and $\hat{\mathtt{I}}_{\mathtt{ei}}$ provided in each frame, one can employ a standard motion estimation method to estimate the intra-frame flow. However, in the case of multi-exposure frames, the two inputs overlap in time, and finding the correct correspondence of $\hat{\mathtt{I}}_{\mathtt{ei}}$ in the long exposure $\hat{\mathtt{L}}_{\mathtt{i}}$ is ambiguous. Therefore, an existing method such as PWC-Net [Sun et al., 2018] cannot be adopted as is, so the following modification to the PWC-Net architecture is applied, tailoring it to the available inputs. In the original PWC-Net, the two nearby frames are fed to the same feature extractor to build the feature pyramids. Then, at each pyramid scale l, the feature of the second frame is warped to the position of the first frame using the upsampled flow, and a cost volume is created to compare the features of the first frame with the warped features from the second one. In the case of multi-exposure frames, as the sharp HDR frame and long exposures are different in type, they are processed with two independent feature extractors, and multi-scale features $\mathtt{V}^{\mathtt{l}}_{\mathtt{i}}$ and $\mathtt{U}^{\mathtt{l}}_{\mathtt{i}}$ are created that correspond to the sharp HDR frame $\hat{\mathtt{I}}_{\mathtt{ei}}$ and long exposure $\hat{\mathtt{L}}_{\mathtt{i}}$, respectively. Then, at each scale l, the intra-frame flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{ei}\rightarrow\mathtt{si}}$ is estimated as follows:

$$\mathtt{F}^{\mathtt{l}}_{\mathtt{ei}\rightarrow\mathtt{si}} = \mathbf{Blur2Flow}(\mathtt{V}^{\mathtt{l}}_{\mathtt{i}}, \mathtt{U}^{\mathtt{l}}_{\mathtt{i}}, \uparrow \mathtt{F}^{\mathtt{l+1}}_{\mathtt{ei}\rightarrow\mathtt{si}}) \tag{4.1}$$

where $\texttt{Blur2Flow}$ is a multi-layer CNN with DenseNet connections [Sun et al., 2018; Huang et al., 2017] and $\uparrow \mathtt{F}^{\mathtt{l}+1}_{\mathtt{ei}\rightarrow\mathtt{si}}$ is the upsampled flow from the previous layer. Note that at each scale, the features of the sharp HDR frame do not need to be warped; hence, no cost volume must be computed. This process is repeated until a desired scale l_0 is reached.

Quadratic motion model: FitQuad Such multi-scale processing procedure continues with the non-learnable quadratic motion modeling. Given the intra-frame flows $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s0}}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s1}}$ that are recovered by **Blur2Flow** separately for each frame, the inter-frame flows $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{e0}}$ between the HDR frames $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$ (downscaled to a given scale l) are also found using a state-of-the-art flow estimation method as proposed in [Teed and Deng, 2020]. While in practice, a quadratic motion model that is aligned with $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ can be derived with only two flows ($\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s0}}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}}$), another possible flow is established, namely $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s1}}$, which corresponds to the flow between $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{s1}}$. It is computed as follows:

$$\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s1}} = \mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}} + \mathtt{warp}(\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s1}}) \tag{4.2}$$

where $\mathtt{warp}$ is a differentiable warping operator using bilinear sampling [Jaderberg et al., 2015]. Here, the flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s1}}$ is aligned with the frame $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$; therefore, $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s1}}$ needs to be warped using the flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}}$ to become aligned with $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ (refer to the bottom row in Figure 4.2). Since the two flows are opposite in their directions, the flows are summed up instead of subtracting them. Similarly, for the frame $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$, the additional flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s0}}$ is computed as:

$$\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s0}} = \mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{e0}} + \mathtt{warp}(\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{e0}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s0}}) \tag{4.3}$$

Now, for warping $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ to a novel time t, a quadratic motion flow is derived as:

$$\mathtt{F}^{\mathtt{l}}_{0\rightarrow\mathtt{t}} = \frac{1}{2}\mathtt{a}_0 \times \mathtt{t}^2 + \mathtt{v}_0 \times \mathtt{t} \tag{4.4}$$

where a_0 and v_0 express the acceleration and velocity of a non-uniform motion, and they are derived from $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s0}}$, $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{e1}}$, and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s1}}$ using the least square fit. Note that the derived model explains the non-uniform motion for the entire range of $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{s0}}$ to $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$. For a curvilinear motion, e.g. a rotatory motion, these parameters can be considered as the first two terms in the Taylor approximation of the curvilinear motion. Similarly, the flow $\mathtt{F}^{\mathtt{l}}_{1\rightarrow\mathtt{t}}$ is computed:

$$\mathtt{F}^{\mathtt{l}}_{1\rightarrow\mathtt{t}} = \frac{1}{2}\mathtt{a}_1 \times \mathtt{t}^2 + \mathtt{v}_1 \times \mathtt{t} \tag{4.5}$$

where the parameters a_1 and v_1 are calculated from the triplet of flows $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s1}}$, $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{e0}}$, and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s0}}$ using a least square fit. Existing VFI methods with non-uniform motion assumptions usually require more than two frames as the input. However, this enforces that the parameters of non-uniformity (acceleration and velocity) are fixed along multiple frames, which might not hold in practice. In contrast, the proposed method only relies on two immediate frames, and as a result, such constraints are imposed in a closer temporal range that allows the modeling of more complex non-uniform motion. Moreover, providing the additional flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s1}}$ not only allows the approximation of a higher order motion, e.g., a cubic motion model, but also incorporates the motion flow information from the other frame to increase flow consistency between $\mathtt{F}^{\mathtt{l}}_{1\rightarrow\mathtt{t}}$ and $\mathtt{F}^{\mathtt{l}}_{0\rightarrow\mathtt{t}}$. In Section 4.5.4, the effect of including $\mathtt{F}^{\mathtt{l}}_{\mathtt{e0}\rightarrow\mathtt{s1}}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{e1}\rightarrow\mathtt{s0}}$ in the motion model is ablated. Since the time interval between $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{s1}}$ and

$\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$ is shared when computing the motion model for the frame pairs $\mathtt{M}_0$ and $\mathtt{M}_1$, and then $\mathtt{M}_1$ and $\mathtt{M}_2$, the temporal consistency is also preserved.

Multiscale blending: Blend In the last step, a multi-scale blending scheme is introduced to reconstruct the final interpolated image $\hat{\mathtt{I}}_{\mathtt{t}}$. Specifically, at each scale l, given the forward flows $\mathtt{F}^{\mathtt{l}}_{0\to\mathtt{t}}$ and $\mathtt{F}^{\mathtt{l}}_{1\to\mathtt{t}}$, the backward flows $\mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to0}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to1}$ are computed using the flow reversal introduced in QVI [Xu et al., 2019]. Then the sharp HDR frames $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$ are warped to the novel position t using the backward flows as:

$$\hat{\mathtt{I}}^{\mathtt{l}}_{0\to\mathtt{t}} = \mathtt{warp}(\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to0}) \text{ and } \hat{\mathtt{I}}^{\mathtt{l}}_{1\to\mathtt{t}} = \mathtt{warp}(\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to1}) \tag{4.6}$$

where $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{e1}}$ are the input frames $\hat{\mathtt{I}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}_{\mathtt{s0}}$ downsampled by 2^l. Afterward, the soft occlusion weight α^l is predicted to control the contribution of input warped images $\hat{\mathtt{I}}^{\mathtt{l}}_{0\to\mathtt{t}}$ and $\hat{\mathtt{I}}^{\mathtt{l}}_{1\to\mathtt{t}}$:

$$\alpha^l = \mathtt{\textbf{Blend}}(\hat{\mathtt{I}}^{\mathtt{l}}_{0\to\mathtt{t}}, \hat{\mathtt{I}}^{\mathtt{l}}_{1\to\mathtt{t}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to0}, \mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to1}, \uparrow \alpha^{l+1}) \tag{4.7}$$

where **Blend** is a multilayer CNN and $\uparrow \alpha^{l+1}$ is the upsampled weight from the previous scale. Note the input flows $\mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to0}$ and $\mathtt{F}^{\mathtt{l}}_{\mathtt{t}\to1}$ aid the network in reasoning about the occlusion regions. Given the occlusion weight, the warped images are combined as follows:

$$\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{t}} = \frac{(1-t)\alpha^l \odot \hat{\mathtt{I}}^{\mathtt{l}}_{0\to\mathtt{t}} + t(1-\alpha^l) \odot \hat{\mathtt{I}}^{\mathtt{l}}_{1\to\mathtt{t}}}{(1-t)\alpha^l + t(1-\alpha^l)} \tag{4.8}$$

where $\hat{\mathtt{I}}^{\mathtt{l}}_{\mathtt{t}}$ is the synthesized intermediate frame at scale l, as required in the loss computation (Eq. 4.11). The operator $\odot$ stands for per-pixel multiplication. Finally, at the finest scale l_0, the interpolated frame $\hat{\mathtt{I}}_{\mathtt{t}}$ is derived.

Loss function The loss function comprises three components targeted to train the **MakeHDR**, **Blur2Flow**, and **Blend** networks. First, the output of the **MakeHDR** network is supervised with the ground truth $\mathtt{I}_{\mathtt{ei}}$ and $\mathtt{L}_{\mathtt{i}}$ (refer to Section 4.5.1 on details of how the ground truth frames are acquired from high-speed video datasets) using the reconstruction loss:

$$\mathrm{L}_{\mathrm{hdr}} = \sum_{i=0,1} \|\mathtt{I}_{\mathtt{ei}} - \hat{\mathtt{I}}_{\mathtt{ei}}\|_1 + \|\mathtt{L}_{\mathtt{i}} - \hat{\mathtt{L}}_{\mathtt{i}}\|_1 \tag{4.9}$$

As the ground truth flow is not available, a multiscale image loss is employed to supervise the **Blur2Flow** network:

$$\mathrm{L}_{\mathrm{flow}} = \sum_{i=0,1} \sum_{l=l_0}^{L} \|\mathtt{I}^{\mathtt{l}}_{\mathtt{ei}} - \mathtt{warp}(\mathtt{I}^{\mathtt{l}}_{\mathtt{si}}, \mathtt{F}^{\mathtt{l}}_{\mathtt{ei}\to\mathtt{si}})\|_1 \tag{4.10}$$

where $\mathtt{I}^{\mathtt{l}}_{\mathtt{ei}}$ and $\mathtt{I}^{\mathtt{l}}_{\mathtt{si}}$ are the ground truth frames $\mathtt{I}_{\mathtt{ei}}$ and $\mathtt{I}_{\mathtt{si}}$ downsampled by 2^l. At each scale l, $\mathtt{I}^{\mathtt{l}}_{\mathtt{si}}$ is warped using the predicted flow $\mathtt{F}^{\mathtt{l}}_{\mathtt{ei}\to\mathtt{si}}$ and compared with $\mathtt{I}^{\mathtt{l}}_{\mathtt{ei}}$. Note that this loss component will try to align the warped image and the input frames for all regions in an image, including the occluded part. However, it is argued that this is not a significant issue because the intra-frame motion captured in the long exposure is relatively small compared to the inter-frame motion. Hence, the small disoccluded areas within a frame are handled, and the only degradation that can occur is over-smoothed flow at occlusion boundaries, which can be resolved with a more sophisticated occlusion treatment. Lastly, the output of the **Blend** network is

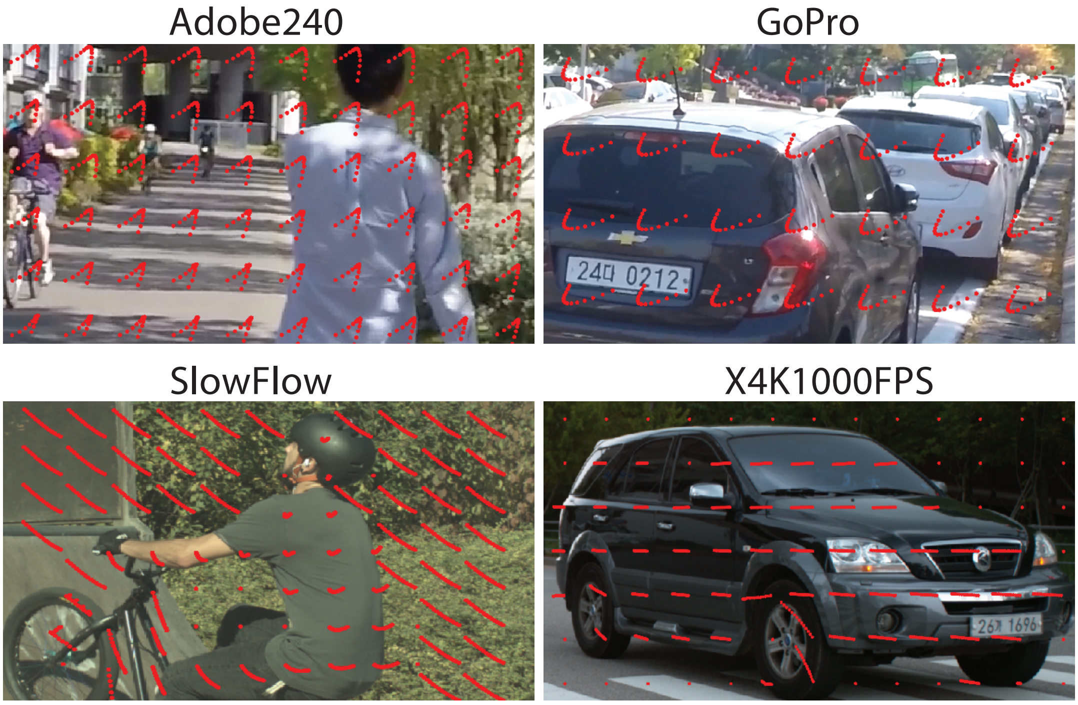


FIGURE 4.3: *Trajectories of pixels (red dots) for 16 consecutive frames in a sample scene from four different datasets. The scenes in Adobe240 and GoPro datasets mostly have globally non-uniform motion due to non-uniform camera motion, while in datasets such as X4K1000FPS and SlowFlow, the scenes mostly contain locally non-uniform motion.*

supervised using the reconstruction loss at each scale:

$$\mathtt{L}_{\text{synth}} = \sum_{l=l_0}^{L} \|\mathtt{I}_{\mathtt{t}}^{\mathtt{l}} - \hat{\mathtt{I}}_{\mathtt{t}}^{\mathtt{l}}\|_1 \tag{4.11}$$

where $\mathtt{I}_{\mathtt{t}}^{\mathtt{l}}$ is the corresponding ground truth for interpolated frame $\hat{\mathtt{I}}_{\mathtt{t}}^{\mathtt{l}}$ at each scale l. The final loss $\mathtt{L}_{\text{total}}$ is then computed as:

$$\mathtt{L}_{\text{total}} = \mathtt{L}_{\text{hdr}} + \mathtt{L}_{\text{flow}} + \mathtt{L}_{\text{synth}} \tag{4.12}$$

It is worth mentioning that based on the observation, optimizing the network based solely on the final loss would create ambiguity as to whether the network should improve **`Blur2Flow`** or **`Blend`** network to decrease the loss; therefore, intermediate supervision (Eq. 4.10 and Eq. 4.11) is essential to train each component properly.

4.3 Motion Non-Uniformity Analysis

In order to properly validate the proposed method, it must be ensured that the dataset contains diverse examples of scene motion non-uniformity. To this end, the motion non-uniformity in some popular high-speed video datasets, including Adobe240 [Su et al., 2017], GoPro[Nah et al., 2017], X4K1000FPS [Sim et al., 2021], and SlowFlow [Janai et al., 2017] are analyzed. The procedure is as follows: For each pixel in a given frame, Raft [Teed and Deng, 2020] is used to track the corresponding pixels for N consecutive frames. The number of consecutive frames $N = 8$ is chosen for the Adobe240, GoPro, and SlowFlow datasets as they are captured with 240FPS,

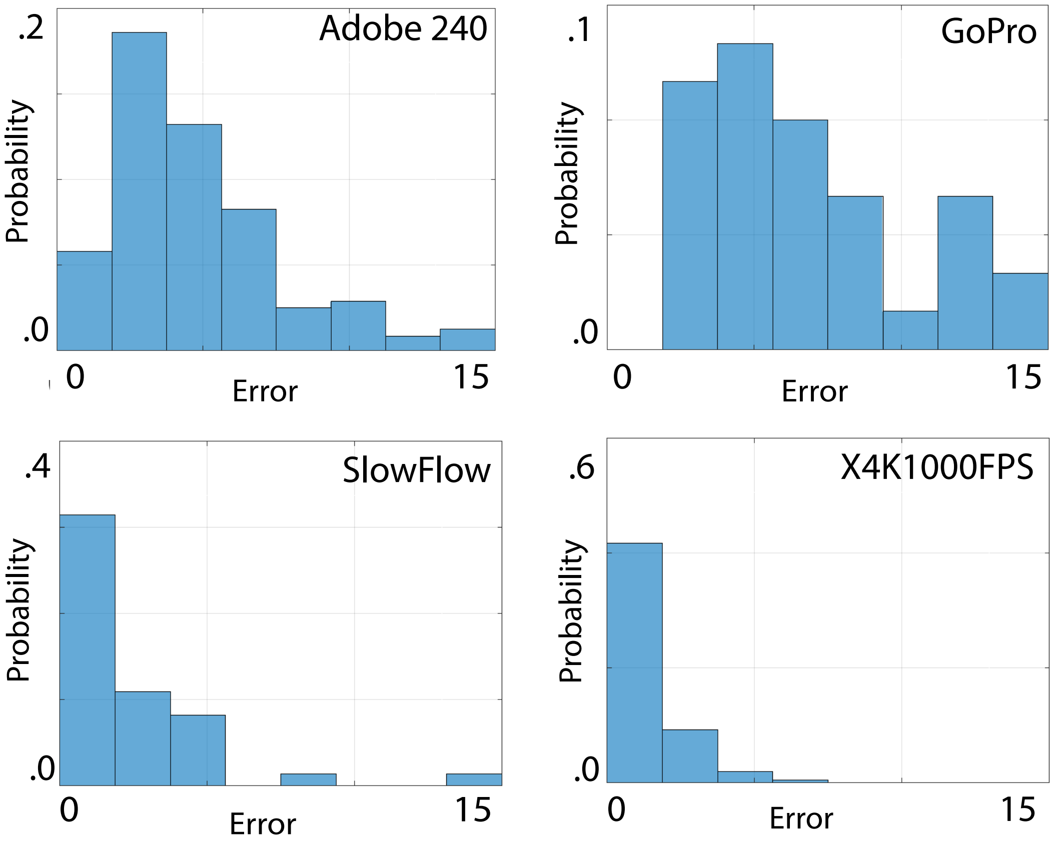


FIGURE 4.4: *The histogram of measured non-uniform motions for different datasets, where the horizontal axis shows the normalized motion error ($\times 10^{-2}$) with respect to the linear motion fit, and the vertical axis denotes the probability of the observed frames given an error value.*

and eight frames represent the time gap between two consecutive frames in a 30FPS video, and $N = 33$ is chosen for X4K1000FPS containing 1000FPS videos. Note that in some cases, such tracking might fail due to occlusions and textureless regions. The occlusion regions are found by applying a forward-backward flow consistency check [Jonschkowski et al., 2020] between the first and last frames and excluding them in the measurements. Likewise, as the estimated flow in the textureless regions is usually erroneous, the flow is clipped to zero if its value is less than one pixel. Figure 4.3 shows the trajectories of pixels for four sample scenes that contain regions with non-uniform motion. In the next step, a linear model is found, in the least square sense, that fits the motion trajectory for each pixel. Then, the mean square error is considered with respect to such a linear fit, where higher errors indicate more motion non-uniformity. Note that for each pixel, the error value is normalized by the aggregated pixel displacement across the consecutive frames. Since the error is calculated for individual pixels, the amount of motion non-uniformity is measured in a frame by taking the 50th percentile of the calculated error over all pixels. This procedure is then repeated for non-overlapping sets of N consecutive frames in each scene in each dataset. Figure 4.4 shows the histogram of measured non-uniform motions for each dataset, where the horizontal axis denotes the error of the linear fit ($\times 10^{-2}$) divided into eight discrete bins, and the vertical axis is the probability of observing the scene for a given error value. The Adobe240 and GoPro datasets feature significant percentages of non-uniform motion as they are captured with a handheld camera. Although large motions are present in the X4K1000FPS dataset, the camera moves along mostly linear trajectories.

4.4 Implementation

The network architecture of **MakeHDR** follows the one introduced in Section 3.4. The network output is provided in the Bayer domain, and demosaicking is applied using OpenCV [Bradski, 2000], followed by a gamma correction to create the final short and long exposures in the sRGB format. The **Blur2Flow** network employs an architecture similar to the PWCNet [Sun et al., 2018], and also outputs the motion flow at a quarter resolution and employs the context network for refining the flow. Then, bilinear interpolation is applied to obtain the half- and full-resolution flows. The **Blend** network is implemented as a 12-layer conventional neural network with dilated convolutions and skip connections. During training, the patch size of 768×768 is used; nevertheless, at the inference time, the convolutional network and all non-learnable components scale with resolution.

4.5 Results

In this section, the training and evaluation datasets are introduced first. Then, the quantitative and qualitative comparisons of the proposed method with existing VFI methods are shown. Finally, the ablation is provided to justify the proposed method's training set and different components.

4.5.1 Dataset

As it is impossible to capture ground truth high-speed HDR videos using the multi-exposure sensor, and third-party high-speed HDR videos are unavailable, the training and evaluation datasets using existing LDR high-speed videos are synthesized. In the experiments, the scenes from X4K1000FPS [Sim et al., 2021] and SlowFlow [Janai et al., 2017] are taken as the training datasets, and Adobe240 [Su et al., 2017] and GoPro [Nah et al., 2017] are considered as the evaluation datasets. The training and testing video sequences are defined as follows: 16 consecutive frames are taken in a high-speed video, where the 1st and 4th frames are the sharp beginning and ending frames ($\hat{\mathtt{I}}_{\mathtt{s0}}$ and $\hat{\mathtt{I}}_{\mathtt{e0}}$). The four neighboring frames starting from 1 to 4 are summed up to simulate the long exposure $\hat{\mathtt{L}}_{\mathtt{0}}$. Then, 9 frames are skipped to simulate the camera readout gap. Similarly, the 13th and 16th frames are taken as the $\hat{\mathtt{I}}_{\mathtt{s1}}$ and $\hat{\mathtt{I}}_{\mathtt{e1}}$, and the frames from 13 to 16 are summed up to create the long exposure $\hat{\mathtt{L}}_{\mathtt{1}}$. Frames 7 and 10 are considered as the target frames for the reconstructions. Note that in the simulation of long exposures, the aggregated pixel intensity is clipped if it exceeds the value of 255. In the simulation, each patch is ignored if more than 20% of its content is already saturated in the original high-speed video. In order to make the HDR-VFI method robust to high blur and saturation, data augmentation is performed by creating different amounts of blur and different amounts of saturation. For the test set, the proposed method is evaluated against the other methods for different ranges of non-uniformity; hence, all scenes in the Adobe240 [Su et al., 2017] and GoPro [Nah et al., 2017] datasets are split into four different categories of Easy, Medium, Difficult, and Extreme based on the error magnitude of the linear fit derived in Section 4.3. Specifically, the entire histogram range (15×10^{-2} here) is divided into four equal segments (expressing the four motion non-uniformity categories), and 125 sample frames are drawn both for the Adobe240 and GoPro datasets per each category.

TABLE 4.1: *Quantitative comparison of the proposed method with state-of-the-art VFI methods. The ABME and QVI methods are designed to handle non-uniform motions, while the XVFI and FILM methods rely on a linear motion assumption but can handle large motions. Methods are indicated with * when they are trained from scratch with the training set.*

	Adobe240		**GoPro**	
Methods	PSNR	SSIM	PSNR	SSIM
ABME [Park et al., 2021]	31.28	0.83	30.98	0.82
QVI [Xu et al., 2019]	31.30	0.86	30.80	0.84
QVI* [Xu et al., 2019]	31.16	0.86	30.70	0.84
XVFI [Sim et al., 2021]	31.07	0.83	30.75	0.82
XVFI* [Sim et al., 2021]	30.66	0.83	30.41	0.82
FILM [Reda et al., 2022]	31.11	0.83	30.75	0.82
FILM* [Reda et al., 2022]	31.04	0.83	30.74	0.82
HDR-VFI	**34.82**	**0.93**	**35.01**	**0.92**

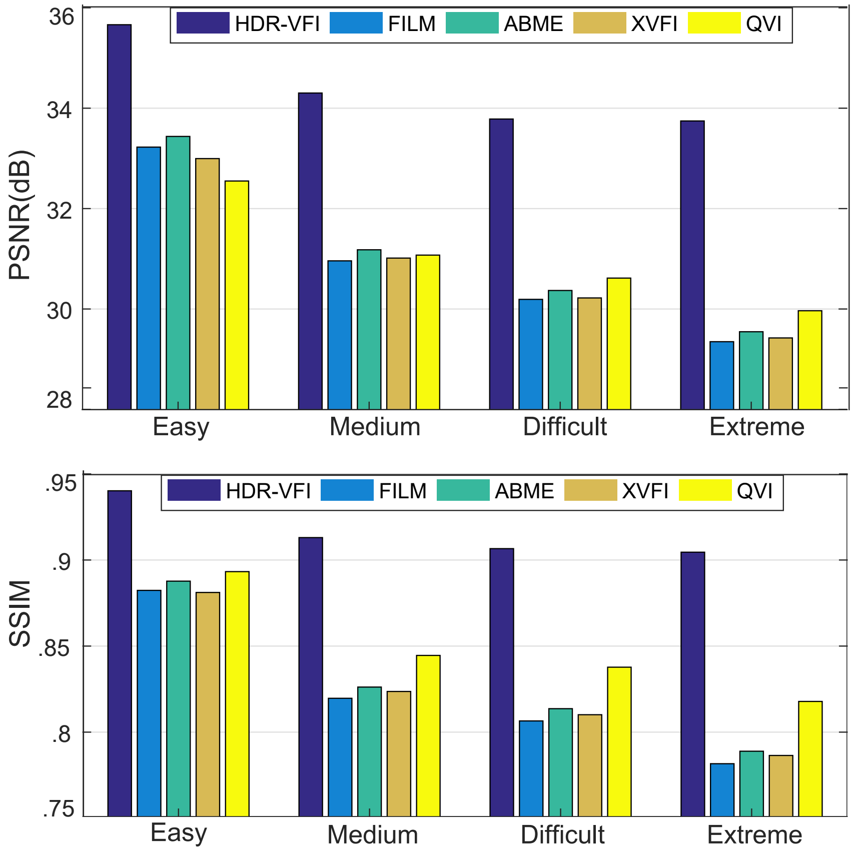


FIGURE 4.5: *Quantitative comparison of the proposed method with state-of-the-art VFI methods for four different motion non-uniformity categories (refer to Section 4.5.1). Each bin reports the average reconstruction error over 250 sample frames per category for a given method.*

4.5.2 Quantitative Comparison

The proposed method is compared with state-of-the-art sharp VFI methods: FILM [Reda et al., 2022] and XVFI [Sim et al., 2021], which rely on a uniform motion assumption, and QVI [Xu et al., 2019], and ABME [Park et al., 2021] which explicitly support the non-uniform motion. QVI employs four consecutive frames as the input,

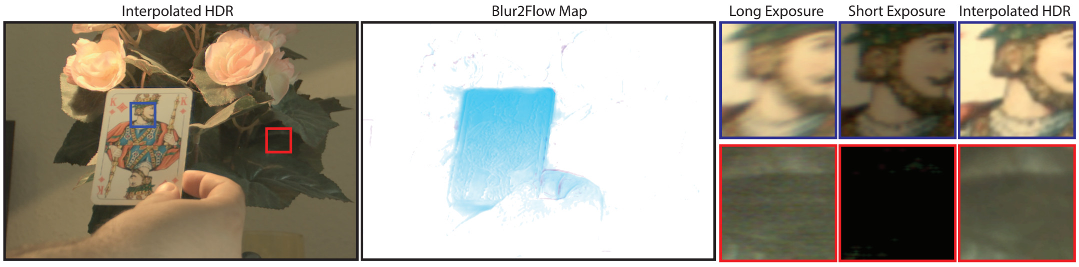


FIGURE 4.6: *Visualization of flow maps reconstructed by the proposed* **`Blur2Flow`** *network. Otherwise, the figure layout follows the one in Figure 4.1.*

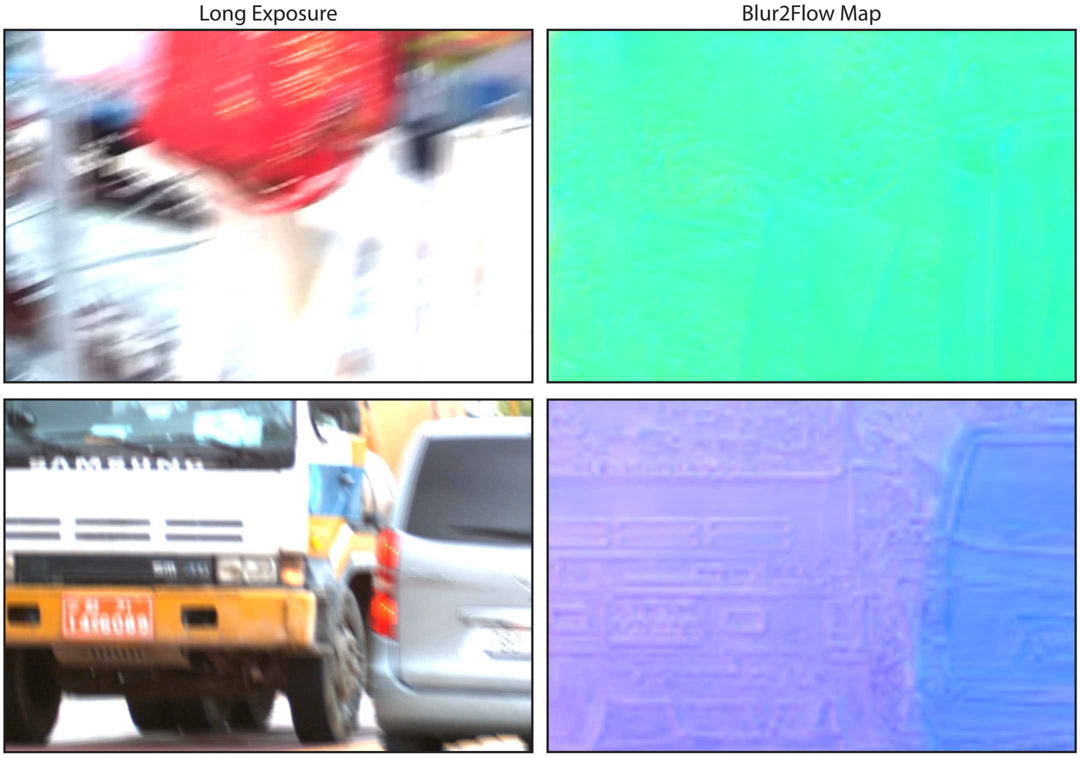


FIGURE 4.7: *Visualization of flow maps of* **`Blur2Flow`** *network for the synthetically created motion blur.*

and FILM and XVFI require just two frames. While ABME also uses only two frames as input, it relaxes the uniform motion constraint by first estimating symmetric bilateral motion fields and then refining them to become asymmetric. As the LDR (sRGB) images in the high-speed dataset are used to synthesize the training and evaluation set, they can directly be fed as input to the VFI methods. For the proposed method, though, they are fed along with the simulated long exposure as described in Section 4.5.1. Note that it is not possible to compare the reconstructions with the blurry VFI methods, as they require well-exposed blurry input frames (effectively, blurry HDR frames) while long exposure in multi-exposure typically contains a considerable amount of saturation that poorly handled by these methods. Table 4.1 summarizes the comparisons with the VFI methods (used with their pre-trained weights) for each of the test datasets (Adobe240 and GoPro) separately as specified in Section 4.5.1. Note that XVFI uses almost the same training set as HDR-VFI while applying extra data augmentation, and a method such as FILM carefully prepared their dataset to include all the possible motion ranges, with a much larger training data size than the considered dataset in this chapter. Nevertheless, for a fair comparison, XVFI,

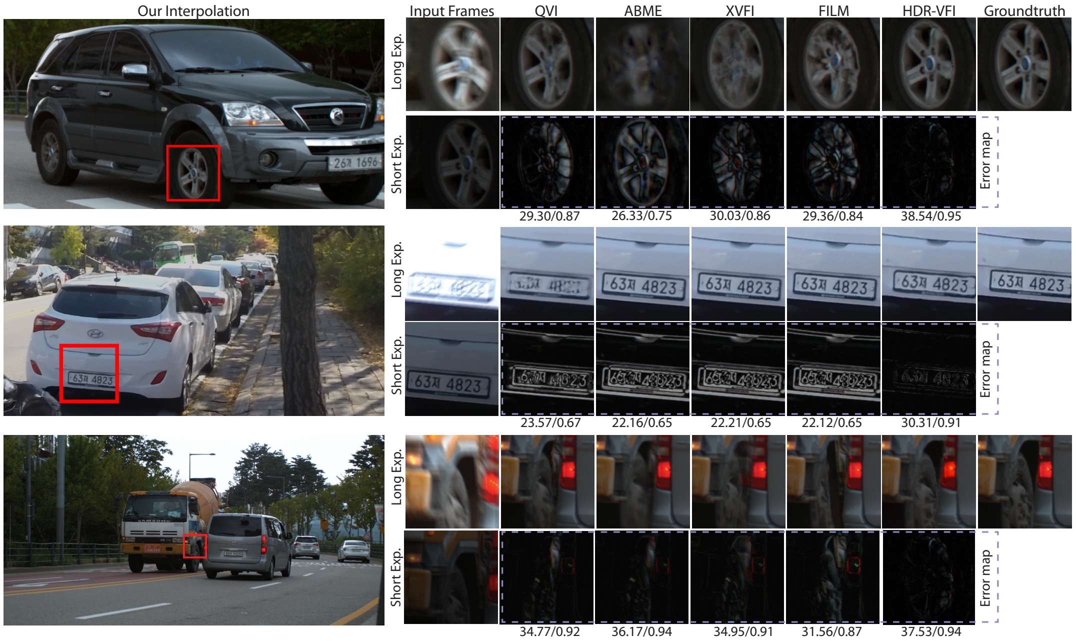


FIGURE 4.8: *Visual comparisons of the proposed HDR-VFI method with the state-of-the-art VFI methods using the synthetic dataset described in Section 4.5.1. For each of the three scenes, the first row of insets shows the performance of respective VFI methods, while the second row presents the corresponding per-pixel error maps between the interpolated results and the ground truth. The PSNR/SSIM values written below each error map are computed for each inset rather than the entire image. In the upper scene taken from the X4K1000FPS test set, the wheel moves in a non-linear trajectory, and the existing VFI methods struggle to position the wheel correctly for the interpolated frames, while the HDR-VFI leads to a good alignment with the ground truth. In the middle scene taken from the GoPro dataset, the camera is moving with an extremely non-uniform motion as shown in Figure 4.3. While the existing VFI methods produce visually plausible results, they are not correctly aligned with the ground truth, as the error map reveals. The bottom scene, taken again from the X4K1000FPS test set, contains a combination of camera and object movements. In this case, the existing VFI methods fail to properly handle occlusion boundaries.*

FILM, and QVI are retrained using the training set of HDR-VFI (indicated with * in Table 4.1) and observed a lower performance. Unfortunately, the training code for ABME is not publicly available. Moreover, Figure 4.5 provides a deeper insight into each method performance when those datasets are aggregated and split into four different categories with respect to motion complexity (Section 4.3). Overall, the competing VFI methods perform similarly for more uniform motion, while the QVI method clearly has advantages for more complex motion. In all cases, the proposed method HDR-VFI outperforms the existing VFI methods by a large margin. It is also more stable in the interpolation quality for higher motion non-uniformity. It can be hypothesized that this stability could be attributed to the quadratic motion fitting part, which has no learnable parameters and only relies on the accuracy of flows, which might drop off slightly at higher non-uniform motion. Other VFI solutions that mostly learn how to handle non-uniform motion might impose higher requirements on the training set.

4.5.3 Qualitative Comparison

The examples of HDR scenes captured in daylight and dark conditions are first visualized in Figure 4.1 and Figure 4.6. The flow map reconstructed by the **`Blur2Flow`** module in Figure 4.6 and the motion blur magnitude in the long exposures indicate

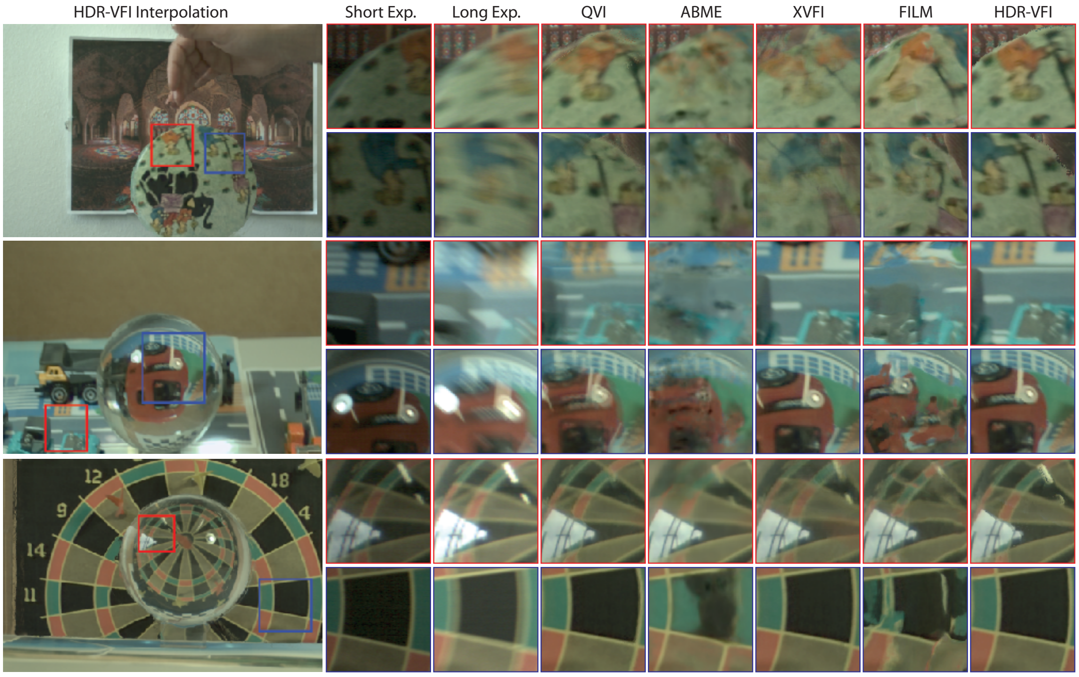


FIGURE 4.9: *The visual comparisons of the interpolation results for three scenes captured using the camera with a multi-exposure sensor. HDR-VFI is able to correctly interpolate the frames in the scenes with the challenging cases of a rolling disc (the upper scene), a rotary camera motion (the middle scene), and a moving object behind a refractive object (the bottom scene).*

the complexity of motion. Additionally, the flow maps for the synthetic data are visualized in Figure 4.7. In Figure 4.1, the HDR-VFI method benefits from additional information that is encoded in the motion blur pattern to improve the interpolation quality. Visual comparisons are then provided with the state-of-the-art VFI methods for three synthesized scenes with ground truth in Figure 4.8. Moreover, HDR-VFI is compared to other methods using the captured sequences in Figure 4.9. All the capturing processes were done with the Axiom-beta camera with a CMOSIS CMV12000 sensor [CMV12000, 2021]. In both setups, the exposure ratio of 4 is used between the short and long exposures. Since the frames captured using a multi-exposure sensor cannot be fed directly to the other VFI methods, the sharp HDR images $\hat{\mathtt{I}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}_{\mathtt{e1}}$ are first reconstructed using the **MakeHDR** network. They are then tonemapped using Reinhard-Global 2002 [Reinhard et al., 2002] and gamma-corrected are fed to the LDR VFI methods. The upper scene in Figure 4.9 shows an example of a rolling disc in which the existing VFI methods, even the ones designed to deal with non-uniform motion such as ABME and QVI, fail to properly interpolate an intermediate frame due to non-uniform motion caused by the rotatory motion of the disc. In the next examples, the crystal ball is captured while the camera is rapidly rotating (the middle scene) or an object is moving behind the crystal ball (the bottom scene). In these challenging examples where even a uniform motion in the scene might appear non-uniform in the refracted image, other methods struggle to correctly reconstruct an in-between frame. In all cases, it is observed that the proposed method faithfully reconstructs the in-between frames even in difficult conditions where there are reflections on the crystal ball (the middle and bottom scenes).

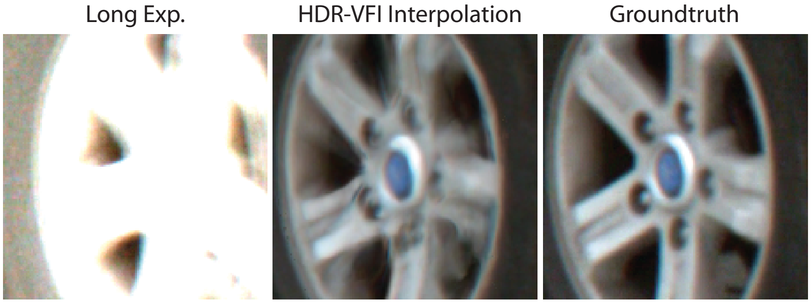


FIGURE 4.10: *The interpolation failure example in a case where the moving content is highly saturated in the long exposure.*

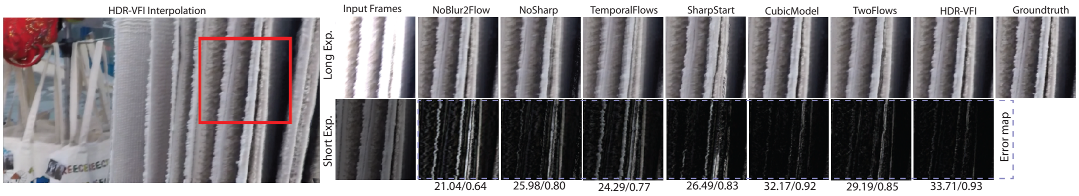


FIGURE 4.11: *Ablation results. The figure layout is similar to Figure 4.8. Refer to Section 4.5.4 for more details on each ablation scenario.*

4.5.4 Ablation Study

A series of ablations are performed to show the contributions of each key component in the proposed method, and the alternative solutions are analyzed. The obtained results are summarized in Figure 4.11 and Table 4.2, where each ablation component is denoted with a unique label that is also included in the related paragraph title.

Impact of `Blur2Flow` network: NoBlur2Flow The contribution of the **`Blur2Flow`** network is analyzed where it is attempted to reconstruct the intermediate frames using only the backward and forward flows between the sharp HDR frames $\hat{\mathtt{I}}_{\mathtt{e0}}$ and $\hat{\mathtt{I}}_{\mathtt{e1}}$ using Raft [Teed and Deng, 2020]. This experiment suggests the version of the proposed HDR-VFI method that makes the linear assumption, in which the flow is split linearly at any position t between the frames; however, this leads to large positional errors in the interpolated content, as seen in Figure 4.11. The results clearly indicate the effectiveness of including the **`Blur2Flow`** network in the proposed pipeline (Table 4.2).

Impact of sharp HDR frame: NoSharp The effect of including the sharp HDR frame $\hat{\mathtt{I}}_{\mathtt{ei}}$, along with the long blurry exposure $\hat{\mathtt{L}}_{\mathtt{i}}$, is investigated on the accuracy of motion from blur derivation. To do so, $\hat{\mathtt{L}}_{\mathtt{i}}$ is considered as the only input to the **`Blur2Flow`** network, and $\hat{\mathtt{I}}_{\mathtt{ei}}$ (note that $\hat{\mathtt{I}}_{\mathtt{ei}}$ is still available for other components in the pipeline) is excluded. As it can be seen in Figure 4.11, the availability of $\hat{\mathtt{I}}_{\mathtt{ei}}$ reduces geometric image distortions, and $\hat{\mathtt{I}}_{\mathtt{ei}}$ compensates for the lack of information for saturated pixels that are inherent for $\hat{\mathtt{L}}_{\mathtt{i}}$ in the setup with a multi-exposure sensor. Following this observation, it is expected that replacing the **`Blur2Flow`** network with a solution, where the intra-frame flow is extracted solely based on $\hat{\mathtt{L}}_{\mathtt{i}}$ [Zhang et al., 2020] should lead to a similar outcome as this ablation.

Quadratic model with temporal flows: TemporalFlows Considering more than two consecutive frames involves a larger time span; as a result, fine-grained motion cannot be properly handled. Such observations are made when comparing the HDR-VFI method with a method like QVI, which uses four frames to compute the quadratic model. Nonetheless, to highlight the advantage of the intra-flow $\mathtt{F}_{\mathtt{ei} \to \mathtt{si}}$ estimated from the **`Blur2Flow`** module, an ablation is conducted where the quadratic motion

TABLE 4.2: *The ablation results indicate the performance of alternative solutions for major design choices in the proposed method. Refer to Section 4.5.4 where more details are provided on each ablation.*

	PSNR	SSIM
NoBlur2Flow	30.97	0.82
NoSharp	30.28	0.82
TemporalFlows	31.93	0.85
SharpStart	34.35	0.91
CubicModel	33.84	0.92
TwoFlows	34.00	0.90
HDR-VFI	**34.92**	**0.93**

is fit using the temporal flows extracted from four consecutive HDR frames (similar to QVI); however, a lower performance is observed than HDR-VFI with two frames, while it still has a better performance compared to QVI.

Alternative approach to `Blur2Flow` network: SharpStart Instead of directly recovering the motion flow from the blur, a 12-layer conventional neural network is employed with dilated convolutions to predict the sharp frame $\hat{\mathtt{I}}_{\mathtt{si}}$ aligned with the beginning of the frame, then the Raft [Teed and Deng, 2020] is used to estimate the intra-frame flow $\mathtt{F}_{\mathtt{ei}\to\mathtt{si}}$ between the $\hat{\mathtt{I}}_{\mathtt{ei}}$ and predicted $\hat{\mathtt{I}}_{\mathtt{si}}$. This ablation demonstrates that the particular method of deriving the intra-frame flow from motion is less important under the condition that sharp, saturation-free reference $\hat{\mathtt{I}}_{\mathtt{ei}}$ is available. Still, the proposed method leads to slight quality improvement.

Quadratic vs. cubic motion model: CubicModel Since the HDR-VFI method provides three estimated flows in each frame, it is possible to approximate a higher-order motion, e.g., cubic. Hence, ablation is performed where the quadratic motion model derived in Section 4.2 is replaced with a cubic model. Overall, the obtained results are comparable in terms of the SSIM prediction, but the quadratic model is slightly better in terms of PSNR and visual results (Figure 4.11). A key difference is that while the cubic model involves a closed-form solution, the quadratic model is derived in a least-squares fashion that allows for the correction of slight errors in the derived flows.

Two vs. three flows: TwoFlows To see the effect of including the additional flows $\mathtt{F}_{\mathtt{e1}\to\mathtt{s0}}$ and $\mathtt{F}_{\mathtt{e0}\to\mathtt{s1}}$ in the derivation of the quadratic motion model, they are excluded from the input to the **`FitQuad`** module. The obtained results (Table 4.2) indicate that including an independent estimate of the third flow contributes toward correcting for potential inconsistencies in the other two flows. For example, in Figure 4.11, ghosting artifacts along higher contrast edges are clearly visible when only two flows are employed.

4.5.5 Limitations

Saturation is inevitable in long exposure for bright scene regions. In the case of a local motion blur that is fully covered with saturation, the predicted flow using the **`Blur2Flow`** network becomes less accurate. Figure 4.10 shows an example of this case where the saturation is increased synthetically in the long exposure for the wheel example shown in Figure 4.8, and the HDR-VFI method fails to correctly reconstruct the intermediate frame. However, in case of a local motion blur with partial saturation or a global camera motion, even with fully saturated regions, as shown in Figure 4.8

and Figure 4.11, the HDR-VFI method can recover the flow by propagating the flow information from the unsaturated regions.

The dynamic range that can be reconstructed is limited by the exposure ratio of four that is assumed in this chapter. For larger ratios, the accuracy of HDR frame reconstruction by the **MakeHDR** network might be reduced (Chapter 3), which could adversely affect the accuracy of HDR video interpolation. Moreover, when capturing an HDR scene, the lowest exposure time is adjusted in such a way that the long exposure is not very saturated so that there is enough valuable blurry information.

4.6 Conclusion

This chapter presents a method for high-dynamic-range video frame interpolation using multi-exposure sensors. The proposed method outperforms the existing VFI methods both in terms of quantitative metrics as well as visual results for the challenging scenes containing non-uniform motions. In particular, high-precision alignment of scene motion with the ground truth is achieved, where other methods clearly fail, although they may produce visually plausible results. The HDR-VFI method can handle complex motion with consistently high performance as it depends little on explicitly training this reconstruction aspect. Instead, the increased temporal sampling rate due to motion reconstruction from blurred information is capitalized. Also, the HDR-VFI is less dependent on scene lighting conditions, whereas other methods designed for single-exposure sensors may suffer from image saturation in bright regions or excessive noise in dark conditions.

Chapter 5

Enhanced Image Quality Measurement

Full-reference image quality metrics (FR-IQMs) aim to measure the visual differences between a pair of reference and distorted images, with the goal of accurately predicting human judgments. However, existing FR-IQMs, including traditional ones like PSNR and SSIM and even perceptual ones such as HDR-VDP, LPIPS, and DISTS, still fall short in capturing the complexities and nuances of human perception. Rather than devising a novel IQM model, this chapter seeks to improve upon the perceptual quality of existing FR-IQM methods. This is achieved by considering visual masking, an important characteristic of the human visual system that changes its sensitivity to distortions as a function of local image content. Specifically, for a given FR-IQM metric, a methodology is proposed to predict a visual masking model that modulates reference and distorted images in a way that penalizes the visual errors based on their visibility. Since the ground truth visual masks are difficult to obtain, it is demonstrated how they can be derived in a self-supervised manner solely based on mean opinion scores (MOS) collected from an FR-IQM dataset. The proposed approach results in enhanced FR-IQM metrics that are more in line with human prediction both visually and quantitatively.

5.1 Introduction

Full-reference image quality metrics, which take as an input a pair of reference and distorted images, play a crucial role in a wide range of applications in digital image processing, such as image compression and transmission, as well as in evaluating the rendered content in computer graphics and vision. They are commonly used as a cost function in optimizing restoration tasks like denoising, deblurring, and super-resolution [Ding et al., 2021b].

Consequently, it is critical to develop FR-IQMs that accurately reflect the visual quality of images in accordance with the characteristics of the human visual system (HVS). The most commonly used FR-IQMs for evaluating image quality are the mean square error (MSE) or mean absolute error (MAE). While these per-pixel metrics are easy to compute, they assess image quality regardless of spatial content, leading to false positive predictions. This can be seen in Figure 5.1a, where Gaussian noise is less noticeable in textured regions, while MAE predicts uniformly distributed error. Similarly, a depth-of-field blur is primarily visible on high-contrast fonts Figure 5.1b, while MAE predicts the blur visibility also in smooth gradient regions. Other classic metrics like SSIM [Wang et al., 2004a], while accounting for spatial content, often result in false positive predictions (the JPEG artifact and image-based rendering (IBR) artifact in Figure 5.1c-d, respectively). A recent hand-crafted metric FLIP [Andersson

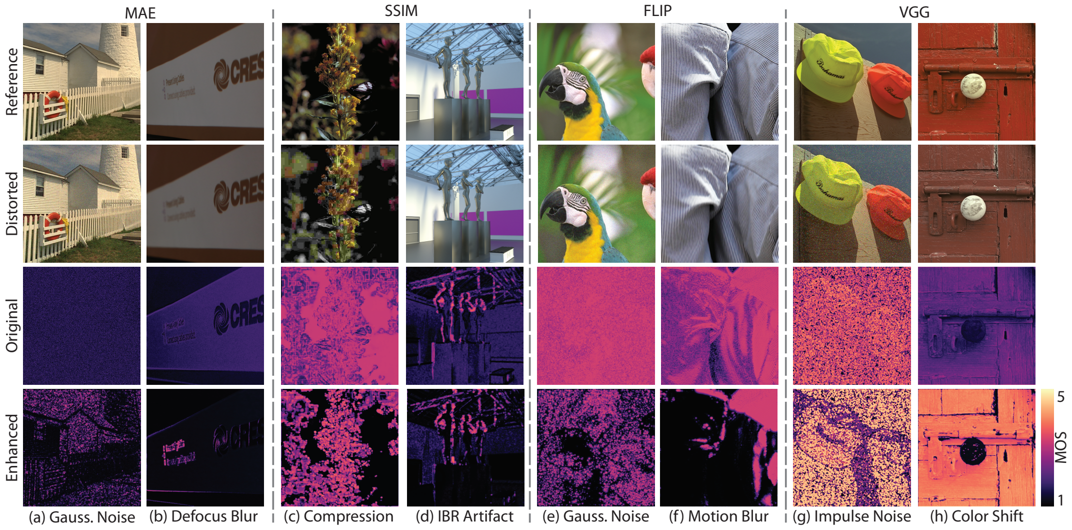


FIGURE 5.1: *This chapter introduces self-supervised visual masking that enhances image quality prediction for existing quality metrics such as MAE, SSIM, FLIP, and VGG. The work is inspired by the well-known characteristic of the Human Visual System (HVS), visual masking, which results in locally varying sensitivity to image artifact visibility that reduces with increasing contrast magnitude of the original image pattern. The experimental results show that the learned masking clearly outperforms its traditional hand-crafted versions and better adapts to specific distortion patterns. The reference and distorted images are shown in the first two rows, while the third and fourth rows show the error maps as predicted by the original metrics and their enhanced versions using our masking approach. As can be seen, mask-enhanced metrics better predict the visibility of local distortion by the human observer. Each error map is scaled to fit within the mean opinion scores (MOS) range for a more intuitive comparison. In this color scale, darker indicates less visible distortion.*

et al., 2020] is specifically designed to predict the visual differences in time-sequential image-pair flipping, which can make it too sensitive for side-by-side image evaluation, e.g., noise is less visible in high-contrast texture (Figure 5.1e) or motion blur is not equally visible across different parts of an image (Figure 5.1f). Recognizing that hand-crafted image features may not adequately capture the HVS complexity, modern metrics [Zhang et al., 2018b] strive to assess the perceptual dissimilarity between images by comparing deep features extracted from classification networks [Simonyan and Zisserman, 2015]. These metrics appear to better account for the HVS characteristics; however, they are designed to generate a single value per image pair and cannot provide correct visible error localization, as can be seen in the impulse noise example (Figure 5.1g). Moreover, the features learned through training the classification networks tend to be less sensitive to global distortions, such as moderate color and brightness changes (Figure 5.1h) that have less impact on reliable classification.

This chapter extends the classic and deep learning-based full-reference metrics by introducing a learnable component trained on perceptual MOS data in a self-supervised way. By implicitly analyzing local image content, the proposed model derives per-pixel maps that mimic visual masking, effectively modeling the visual significance of distortions. The self-supervised masking methodology is introduced in Section 5.2 and Section 5.3 present a comparison of existing FR-IQMs and their enhanced versions together with an outcome of ablation studies. Section 5.4 provides the source code of the proposed method. Finally, this chapter is concluded in Section 5.5.

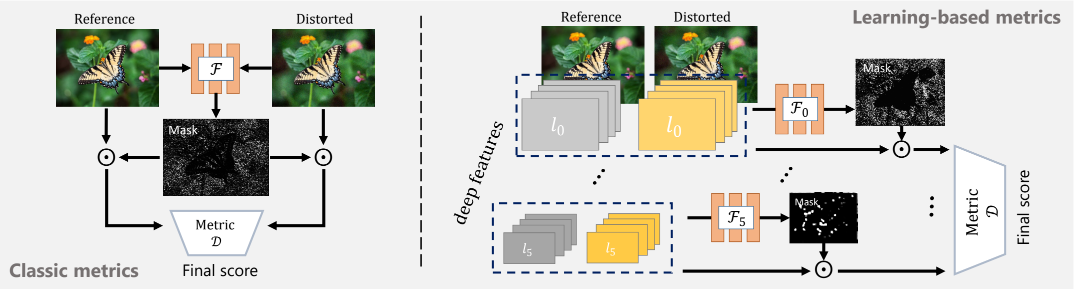


FIGURE 5.2: *The proposed visual masking for enhancing classic metrics such as MAE and SSIM (left) and learning-based metrics such as DISTS or LPIPS (right). For classic metrics, the input to mask predictor network $\mathcal{F}$ are sRGB images, while for learning-based metrics, the inputs are the VGG features extracted from the images. The visual masks are learned in a self-supervised fashion by minimizing the difference between the metric final score and human scores collected from an FR-IQM dataset.*

5.2 Self-Supervised Visual Masking

This section elaborates on the proposed methodology for perceptually calibrating the existing FR-IQMs. Given a reference and distorted pair (X and Y) $\in R^{H\times W\times C}$, firstly, a visual mask is learned, $M \in R^{H\times W\times 1}$, which has the same spatial dimensions as the inputs. For classical metrics (Figure 5.2-left), the input X and Y are sRGB images ($C = 3$), while for learning-based metrics such as LPIPS, DISTS, or DeepWSD, the input X, and Y are the VGG features extracted from the images and C is the number of channels in a given VGG layer (Figure 5.2-right). The predicted mask is then element-wise multiplied with X and Y before being fed into an FR-IQM, $\mathcal{D}$. Note that, for learning-based metrics, the direct modulation of the input sRGB images by a mask M would distort their content and consequently reduce the VGG performance as it is originally trained on complete, non-masked images. The proposed solution with VGG feature modulation draws inspiration from classic FR-IQMs [Lubin, 1995; Daly, 1993; Mantiuk et al., 2011a; Mantiuk et al., 2021], where the response from hand-crafted filter banks is transduced using a fixed, perception-motivated masking model [Legge and Foley, 1980; Foley, 1994; Wilson and Gelb, 1984]. In this approach, the response from pre-trained VGG filters is modulated with a learned per-pixel mask M, where perception modeling is learned from the MOS data. The visual mask M is estimated by utilizing a lightweight CNN denoted as $\mathcal{F}$, which takes both X and Y as input. Mathematically, this can be expressed as:

$$M = \mathcal{F}(X, Y) \tag{5.1}$$

It is important to note that the network $\mathcal{F}$ is trained specifically for a metric $\mathcal{D}$. In the case of metrics such as LPIPS, DISTS, and DeepWSD, their specific architecture is followed, and a mask is computed for each layer using a separate $\mathcal{F}$, then the same mask is applied for all channels in a given layer (Figure 5.2-right). The original spatial pooling is preserved for each metric, such as ℓ_1 distance in LPIPS, structural similarity in DISTS, or Wasserstein distance in DeepWSD. Since the output of the mask generator network can not be directly supervised, a self-supervised approach is adopted to train it using an IQM dataset with a single quality score. The network's parameters are optimized by minimizing the ℓ_2 difference between the metric output value and human scores. The loss function is formulated as follows:

$$Loss = \|\mathcal{G}(\mathcal{D}(M \odot X, M \odot Y)) - q\|_2^2 \tag{5.2}$$

Here, $q \in [0, 1]$ represents the normalized mean opinion score when comparing

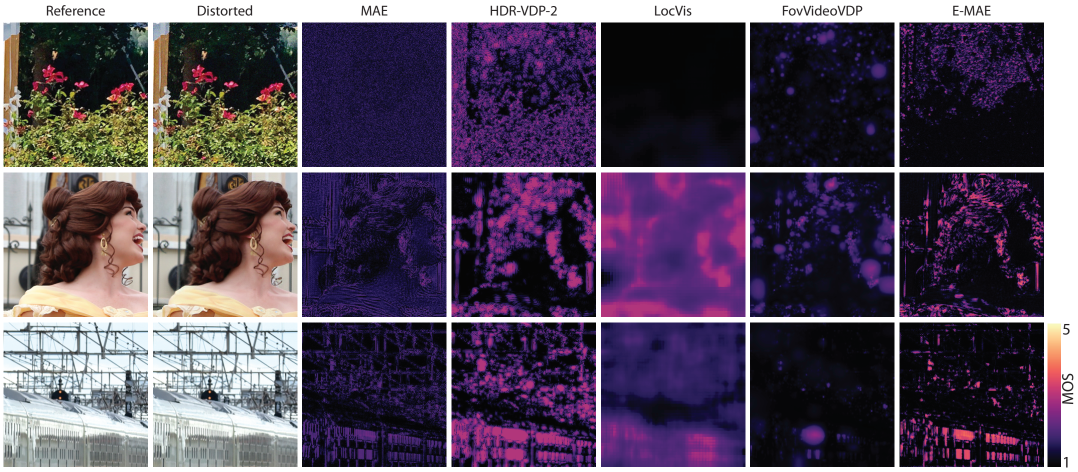


FIGURE 5.3: *Visual comparisons of distortion visibility maps for Gaussian noise (upper row) and superresolution artifacts (middle and bottom rows). The distortion examples are taken from the PIPAL dataset. The first two columns present the reference and distorted images, followed by the respective metric predictions: MAE, HDR-VDP-2 [Mantiuk et al., 2011a], LocVis [Wolski et al., 2018], FovVideoVDP [Mantiuk et al., 2021], and E-MAE. Here, the MAE map is additionally visualized to understand each distortion's characteristics better. As can be seen, the existing metrics tend to either underestimate or overestimate the distortion visibility. Note that LocVis and E-MAE have not seen distorted images with superresolution artifacts in their training.*

the images X and Y. As the metric response can vary in an arbitrary range, following a similar approach in [Zhang et al., 2018b], a small network $\mathcal{G}$ is jointly trained to map the metric response to the human ratings.

5.2.1 Training and Network Details

For training, the KADID dataset [Lin et al., 2019] is used, which comprises 81 natural images that have been distorted using 25 types of traditional distortions, each at five different levels, making roughly 10k training pairs. Note that the mask generator network $\mathcal{F}$ is trained for all the distortion categories together rather than for one specific category. It is experimentally found that a lightweight CNN with three convolutional layers, each consisting of 64 channels, suffices for successful training. ReLU activation is applied after each layer, while Sigmoid activation is used for the final layer to keep the mask values in the range between 0 and 1. The computation overhead of the mask generator network is very negligible, and it takes only 12 ms to compute the mask on a 1080Ti GPU with an input resolution of $768 \times 512 \times 3$. The mapping network $\mathcal{G}$ consists of two 32-channel fully connected (FC) ReLU layers, followed by a 1-channel FC layer with Sigmoid activation. The batch size for training is set to 4. The Adam optimizer [Kingma and Ba, 2015] is employed with an initial learning rate of 10^{-4} and a weight decay of 10^{-6}.

5.3 Results

In this section, the experimental setup is first presented, which is used to evaluate the proposed method, and the ablations of different training strategies follow.

5.3.1 Experimental Setup

The visual masking approach is employed to enhance some of the classical metrics (MAE, PSNR, SSIM, MS-SSIM, FLIP, and fovVideoVDP) and recent learning-based

TABLE 5.1: *Performance comparison of existing FR-IQMs and their enhanced versions using the proposed approach (specified by the prefix E) on three standard IQM datasets. The prefix R denotes the original metric retraining on the KADID dataset, while the prefix S refers to employing a visual saliency mask instead of the visual mask. At the bottom, the corresponding results for NR-IQMs are included. Higher values of SRCC, PLCC, and KRCC indicate better quality prediction. The first and second best metrics for each dataset are indicated in bold and underlined, respectively. Additionally, the version with superior correlation is highlighted in dark gray for each metric.*

	CSIQ			**TID**			**PIPAL**		
Metric	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
FSIM	0.900	0.913	0.740	0.847	0.789	0.611	0.651	0.617	0.441
VIF	0.826	0.841	0.642	0.820	0.813	0.616	0.584	0.538	0.378
HDR-VDP-2	0.761	0.886	0.704	0.715	0.753	0.571	0.514	0.503	0.354
PieAPP	0.827	0.840	0.653	0.832	0.849	0.652	**0.729**	**0.709**	**0.521**
MAE	0.819	0.801	0.599	0.639	0.627	0.409	0.458	0.443	0.304
S-MAE	0.656	0.697	0.493	0.498	0.496	0.347	0.369	0.365	0.248
E-MAE	0.871	0.917	0.738	0.857	0.863	0.673	0.597	0.606	0.429
PSNR	0.851	0.837	0.645	0.726	0.714	0.540	0.468	0.456	0.314
E-PSNR	0.901	0.910	0.728	0.855	0.844	0.656	0.637	0.629	0.446
SSIM	0.848	0.863	0.665	0.697	0.663	0.479	0.550	0.534	0.373
E-SSIM	0.869	0.910	0.732	0.842	0.868	0.677	0.671	0.656	0.469
MS-SSIM	0.826	0.841	0.642	0.820	0.813	0.616	0.584	0.538	0.379
E-MS-SSIM	0.862	0.895	0.709	0.806	0.825	0.621	0.642	0.634	0.453
FLIP	0.731	0.724	0.527	0.591	0.537	0.413	0.498	0.442	0.306
E-FLIP	0.871	0.902	0.715	0.859	0.858	0.666	0.621	0.612	0.434
FovVideoVDP	0.795	0.821	0.632	0.742	0.727	0.544	0.565	0.509	0.358
E-FovVideoVDP	0.841	0.882	0.685	0.830	0.816	0.623	0.662	0.626	0.449
VGG	0.938	0.952	0.804	0.853	0.820	0.639	0.643	0.610	0.432
E-VGG	0.914	0.938	0.776	0.895	0.889	0.710	0.695	0.675	0.485
LPIPS	0.944	0.929	0.769	0.803	0.756	0.568	0.640	0.598	0.424
R-LPIPS	0.931	0.917	0.756	0.898	0.886	0.697	0.670	0.640	0.447
E-LPIPS	0.922	0.933	0.771	0.884	0.876	0.689	0.705	0.678	0.490
DISTS	0.947	0.947	0.796	0.839	0.811	0.619	0.645	0.626	0.445
E-DISTS	0.938	0.925	0.754	**0.903**	**0.915**	**0.725**	0.725	0.697	0.507
Watson-VGG	0.944	0.940	0.785	0.808	0.763	0.573	0.627	0.606	0.429
E-Watson-VGG	0.917	0.936	0.776	0.886	0.895	0.716	0.697	0.678	0.488
DeepWSD	0.949	**0.961**	0.821	0.879	0.861	0.674	0.593	0.584	0.409
R-DeepWSD	**0.955**	**0.961**	**0.823**	0.895	0.88	0.695	0.654	0.633	0.449
E-DeepWSD	0.937	0.937	0.775	0.905	0.892	0.710	0.704	0.672	0.485
HYPERIQA	0.769	0.757	0.573	0.679	0.662	0.489	0.325	0.363	0.250
MANIQA	0.874	0.827	0.642	0.784	0.760	0.572	0.404	0.407	0.276

methods (VGG, LPIPS, DISTS, Watson-VGG, and DeepWSD). Note for MS-SSIM, the same $\mathcal{F}$ is used across all scales, while the inputs are images at different scales. Moreover, the metric called VGG is computed by simply taking the ℓ_1 difference between VGG features for the same layers as originally chosen for LPIPS and DISTS. Deploying the masking model to PieAPP or any other metrics that create new CNN architectures from scratch is not practical as there is no intermediate component to which the masking model can be applied. Thus, the main focus remains on mainstream metrics that use features extracted from pre-trained networks for quality prediction. The performance of the proposed approach is assessed on three well-established IQM datasets: CSIQ [Larson and Chandler, 2010], TID2013 [Ponomarenko et al., 2015], and PIPAL [Jinjin et al., 2020]. The first two datasets mainly consist of synthetic distortions, ranging from 1k to 3k images. On the other hand, PIPAL is the most comprehensive IQM dataset due to its diverse and complex distortions, consisting of 23k images. Each reference image in this dataset was subjected to 116 distortions, including 19 GAN-type distortions. For evaluation, following [Ding et al., 2022], the smaller side resolution of input images is resized to 224 while maintaining the aspect ratio. Note that rescaling is only performed on the test datasets to match the image resolution in which the MOS data were collected. The proposed approach does not require rescaled inputs, and all visual figures in this section are processed in their original resolution. For each dataset, three metrics are used for evaluation: Spearman's rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and the Kendall rank correlation coefficient (KRCC). The PLCC measures the accuracy of the predictions, while the SRCC indicates the monotonicity of the predictions, and the KRCC measures the ordinal association. The PLCC measures linear correlation, requiring both variables (metric output and MOS) to be on the same scale. Hence, the metric scores are mapped to the MOS values using a four-parameter logistic function, consistent with established IQM methods [Ding et al., 2022; Liao et al., 2022]. The network $\mathcal{G}$ is not used for PLCC remapping; otherwise, a specific $\mathcal{G}$ needs to be trained for each metric on a given test set. Importantly, SRCC and KRCC scores do not require additional remapping, thus directly reflecting the correlation between metric output and MOS data.

5.3.2 Evaluations

This section presents the outcome of the quantitative (agreement with the MOS data) and qualitative (the quality of error maps) evaluation of the proposed method. The mask content is also analyzed and related to perceptual models of contrast and blur perception. Finally, the error map prediction of different distortion levels is analyzed, and the potential use of enhanced E-MAE metric as a loss in denoising and deblurring image restoration tasks is investigated.

Quality prediction The experimental results are presented in Table 5.1, and the proposed extension is denoted with the prefix E for each specific IQM. The extension for traditional metrics, such as MAE, PSNR, SSIM, FLIP, and fovVideoVDP, consistently improves their performance for all datasets. This is remarkable as those metrics are commonly used, and the simple extension can make their distortion prediction closer to the human observer. Interestingly, the enhanced E-MAE and E-PSNR outperform recent learning-based VGG, LPIPS, and DISTS in the TID dataset while showing a comparable performance for the PIPAL dataset. Notable improvements are also observed in both datasets for the recent learning-based metrics (E-VGG, E-LPIPS, E-DIST, Watson-VGG, and E-DeepWSD), positioning them at a level comparable to other state-of-the-art IQMs, such as PieAPP [Prashnani et al., 2018]. The only exception is the case of the small-scale CSIQ dataset, where the original learning-based metrics achieve high correlations with the MOS data and leave little space for further improvements.

Retraining LPIPS per-channel weights (denoted as R-LPIPS in Table 5.1) using the KADID dataset is also considered. The strategy improves correlation for TID and PIPAL datasets with respect to the original LPIPS. Compared to E-LPIPS, such retraining is more prone to overfitting; it performs marginally better for the TID dataset, which has more distortion similarities with KADID, while it is significantly worse for the larger and more diverse PIPAL dataset. Similarly, training layer-specific weights for DeepWSD (R-DeepWSD) improves correlation. However, E-DeepWSD achieves better performance. Moreover, channel/layer-wise weighing can not be reasonably applied to image-based metrics (MAE, SSIM, FLIP).

Moreover, the performance of recent NR-IQM methods MANIQA [Yang et al., 2022] and HYPERIQA [Su et al., 2020] that are trained on the KADID dataset is evaluated. As it can be seen in Table 5.1, the NR-IQM methods show significantly lower correlations with the MOS data, particularly for the PIPAL dataset, which indicates that FR-IQM methods can better generalize to unseen distortion types.

Visual saliency methods incorporate semantic information. However, they are not trained to discriminate between dominant distortions and salient features (e.g., faces). This seems to be a limiting factor in the direct saliency use in the proposed image quality evaluation framework. To validate this observation, the predicted saliency map from an off-the-shelf saliency network [Jia and Bruce, 2020] is employed as a mask to the MAE metric that is denoted as S-MAE in Table 5.1. While significantly lower correlations with the MOS data are observed in this simple attempt, the proposed visual masking approach can be complemented by saliency so that effective distortion predictions are narrowed to image regions that are likely to be visually attended.

Analysis of poorer performance for the CSIQ dataset The correlation of each metric across six distortion categories for the CSIQ dataset is further investigated. The proposed approach slightly improves or maintains high correlations for the majority of the distortion categories; the only exception is the *global contrast decrements* category, where a significant decrease in the correlation can be seen across all metrics, resulting in an overall negative impact on the correlation. This can be attributed to the fact that the global contrast change results in strong brightness differences where the proposed masking model apparently can not generalize to this specific unseen distortion category. Figure 5.4 illustrates two sample images from this category where the predicted mask for the E-MAE metric exhibits less sensitivity to changes in brightness, particularly noticeable in the sky regions.

Error map prediction In Figure 5.1, the error maps predicted by various existing IQMs and their enhanced versions for a set of images featuring different types of distortions are shown. As the output of each metric can be in an unbounded range and vary across different metrics and their improved versions with the proposed approach, for a more intuitive and fair comparison, instead of simply normalizing them within the range from zero to one using a Sigmoid function [Andersson et al., 2020], the output of each metric after being scaled to the MOS range using a pre-trained scaling network $\mathcal{G}$ are visualized. Specifically, the KADID dataset is utilized, and a separate $\mathcal{G}$ is trained for each metric to transform their raw response into values that align with human ratings (MOS). Note that for the enhanced version of each metric, the network $\mathcal{G}$ is already provided from the training step. This scaling process is generally akin to mapping the metric scores to the MOS values using a four-parameter function when computing the correlation. As can be seen in Figure 5.1,

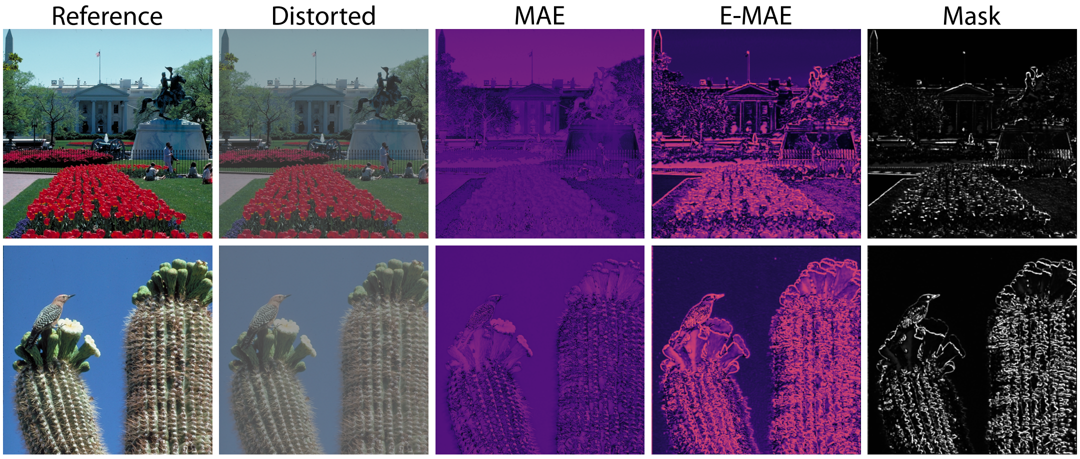


FIGURE 5.4: *Visualizations of error maps for MAE and its enhanced version (E-MAE) alongside the predicted masks for two sample examples within the "global contrast decrements" category from the CSIQ dataset.*

the enhanced error maps using the proposed approach better align with the human perception of the distortion. A notable example is the case of Gaussian noise, where a metric like MAE predicts uniformly distributed error, and the visual masking approach effectively redistributes the error in terms of both their magnitude and local visibility. Additionally, Figure 5.3 showcases three examples where the E-MAE metric achieves better localized error maps compared to well-established visibility metrics such as HDR-VDP-2 [Mantiuk et al., 2011a], LocVis [Wolski et al., 2018], and FovVideoVDP [Mantiuk et al., 2021].

Mask visualization It is also intriguing to see the learned mask, i.e., the output of the network $\mathcal{F}$, and to compare it with a traditional visual contrast masking model, such as the one used in JPEG2000 compression [Zeng et al., 2002]. To this end, Figure 5.5 presents visual masks generated for noise and blur distortions. The same distortion level and three levels of image contrast enhancement (×0.5, ×1, and ×2) are considered. In the case of noise distortion, learned masks predict stronger visual masking in the high-contrast butterfly and better noise visibility in the out-of-focus smooth background. Increasing image contrast (×2) leads to even stronger visual masking in the butterfly area and the plant behind it. Reducing image contrast (×0.5) leads to the inverse effect. Such behavior is compatible with the visual contrast masking model [Zeng et al., 2002; Tursun et al., 2019], where due to self-contrast masking, the higher the contrast of the original signal (e.g., on edges), the stronger the distortion should be to make it visible. Along a similar line, due to neighborhood masking, the higher the contrast texture, the stronger the visual masking as well. In the case of blur distortion, the learned mask predicts its strong visibility on high-contrast edges. The stronger the image contrast (×2), the blur visibility improves. Assigning a higher weight by visual mask to high contrast regions agrees with perceptual models of blur detection and discrimination [Watson and Ahumada, 2011; Sebastian et al., 2015]. Note that each mask is derived by taking as an input both the reference and distorted images; the mask can resolve even per-pixel distortions, as in the case of noise (Figure 5.5), and accordingly informs the E-MAE metric on the perceptual importance of such distortions. What is also remarkable is that the HVS might impose contradictory requirements on hand-crafted visual models that become specific for a given distortion. This is well illustrated in Figure 5.5, where noise can be better masked by strong contrast patterns [Zeng et al., 2002; Tursun et al., 2019] while blur is actually better revealed by strong contrast patterns [Watson and Ahumada, 2011]. The learned

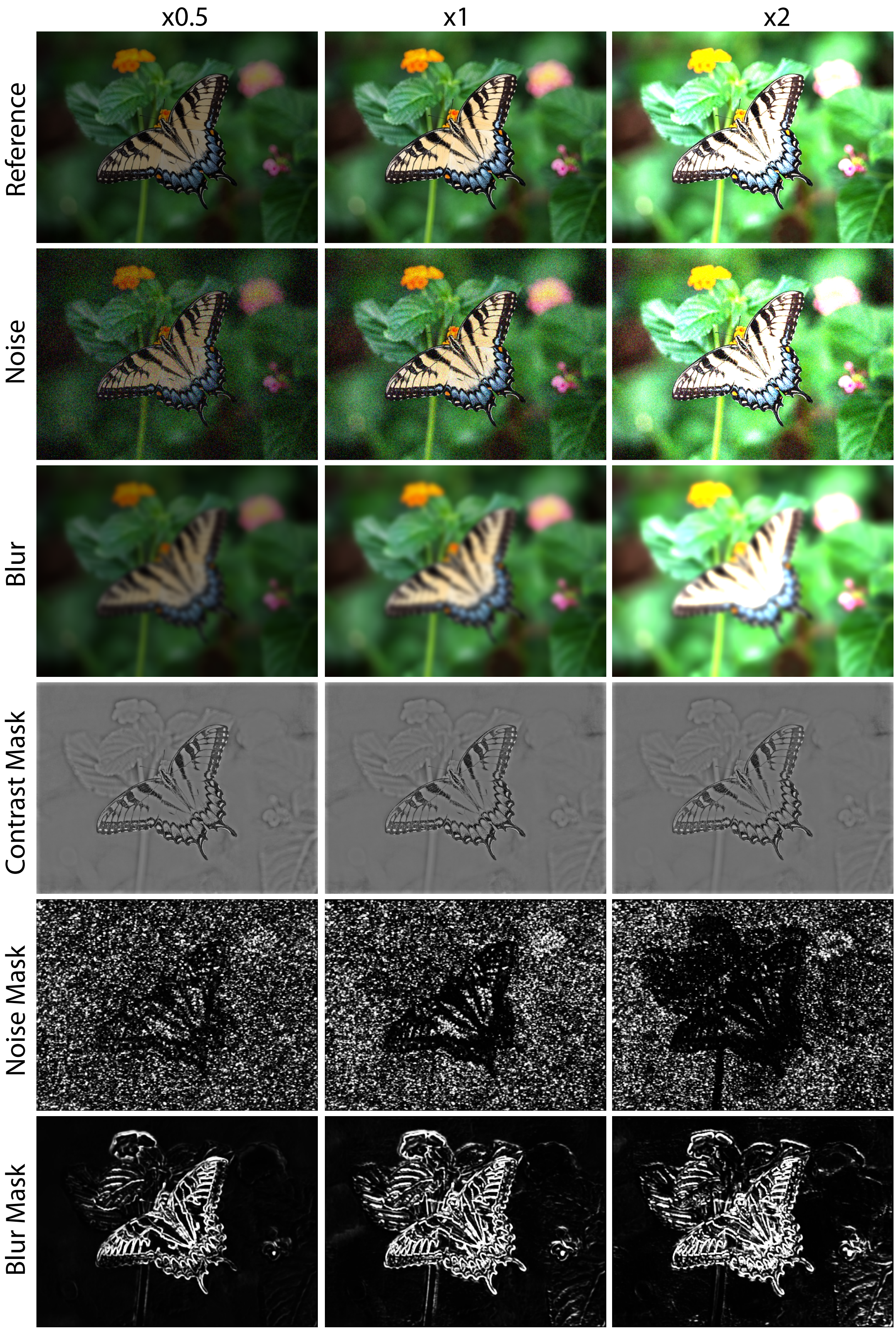


FIGURE 5.5: *Comparison of the E-MAE metric masks for the noise (fifth row) and blur (sixth row) distortions as a function of different image contrast (×0.5, ×1, and ×2). In the fourth row, a map with the human sensitivity to local contrast changes as predicted by a traditional model of visual contrast masking [Tursun et al., 2019, Eq.4] is shown. In all cases, darker means more masking (less sensitive to distortion).*

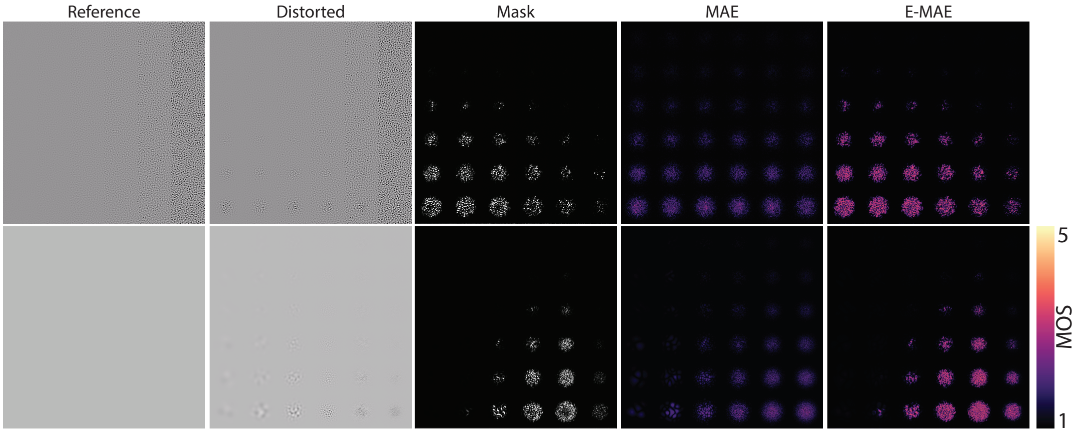


FIGURE 5.6: *Error map prediction for the MAE and E-MAE metrics along with learned weighting masks for two perception patterns from Čadík et al., 2013. These patterns were specifically designed to investigate various perceptual phenomena, including contrast sensitivity and contrast masking. In the first row, the background consists of a high-frequency pattern with increasing contrast toward the right and a stimulus pattern with decreasing contrast from bottom to top (which becomes more apparent when zoomed in). In this scenario, contrast masking is more pronounced with increasing background contrast that, in turn, reduces the stimulus visibility, and E-MAE correctly predicts this effect. The second row presents another example, showing a set of patterns where their spatial frequencies increase toward the right while their contrast decreases toward the top. In this case, the learned masking roughly follows an inverse U-shape characteristic, akin to the contrast sensitivity function (CSF) Daly, 1993; Barten, 1999; Wuerger et al., 2020. The visual masking well approximates the sensitivity drop for high frequencies but tends to excessively suppress the visibility of low-frequency patterns. In spite of this drawback, it is still quite remarkable that the CSF shape becomes apparent in learned visual masks without any explicit training with calibrated near-threshold CSF data.*

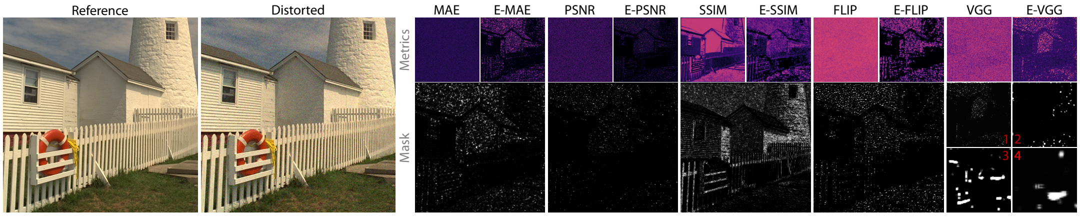


FIGURE 5.7: *Visualisation of predicted mask across different metrics for a given pair of reference and distorted images with Gaussian noise from the TID dataset. Note that the SSIM values have been remapped to 1-SSIM, where lower values indicate less visible errors. In the case of the PSNR, the error map for the measured MSE is shown. For the VGG metric, the predicted mask for all layers is visualized, while the error map is shown only for the first layer.*

E-MAE mask recognizes the distortion context and reacts as expected by penalizing less noise distortion in high-contrast and textured regions while penalizing more blur distortion at high-contrast edges. Interestingly, such local, seemingly contradictory behavior has been learned solemnly by providing multiple pairs of reference and distortion images along with the corresponding quality MOS rating, which is just a single number. No annotation on specific distortion types was required during the training stage. Figure 5.6 shows further examples that the learned masking is also informed about contrast masking by texture [Ferwerda et al., 1997] and the contrast sensitivity function (CSF) [Daly, 1993; Barten, 1999; Wuerger et al., 2020].

Masks vs. metrics analysis Masks typically vary with distortion type, as demonstrated in Figure 5.5 for noise and blur. In Figure 5.7, the predicted masks across various metrics are further illustrated, including MAE, PSNR, SSIM, FLIP, and VGG for a given pair of reference and distorted images with Gaussian noise. As can be seen,

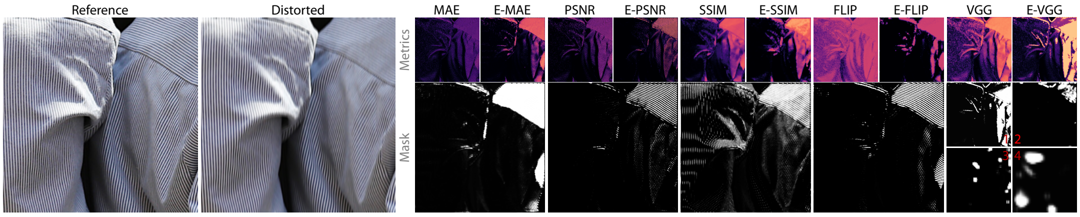


FIGURE 5.8: *Visualisation of predicted mask across different metrics for a given pair of reference and distorted images with motion blur from the PIPAL dataset. The SSIM values have been remapped to 1-SSIM, where lower values indicate less visible errors. In the case of the PSNR, the error map is shown for the measured MSE. For the VGG metric, the predicted mask is visualized for all layers, while the error map is shown only for the first layer.*

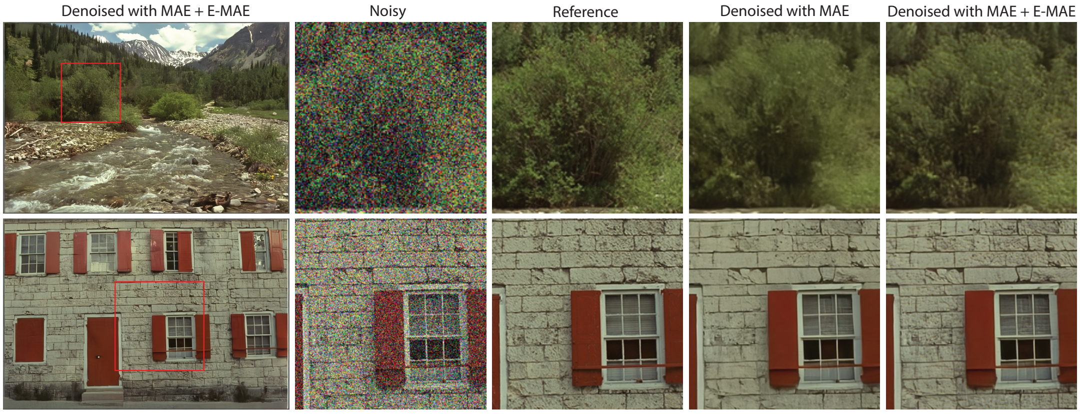


FIGURE 5.9: *Visual results in the image denoising task when employing MAE and MAE+E-MAE as loss functions. Considering that denoiser networks typically reduce noise through smoothing, the objective was to investigate whether the use of the E-MAE loss component could encourage the network to retain or hallucinate details, even if they do not precisely match the reference but their discrepancy from the ground truth is possibly not perceivable. As can be seen, the denoised images with the MAE+E-MAE loss yield sharper content and higher contrast.*

metrics with similar characteristics, such as MAE, PSNR, and FLIP, tend to learn similar maps. For a more perceptually-informed metric like SSIM that partially models visual masking, the predicted mask adjusts its sensitivity by assigning lower weight to regions where SSIM exaggerates the error (e.g., in the grass area) and identity weight when accurately predicting the error magnitude (e.g., the body of lighthouse). When it comes to VGG, the mask learned for the early layer resembles the MAE mask since the initial convolutional layers tend to learn basic image features like edges and textures, whereas for the deeper layers, as the VGG learns more abstract features, the interpretation of the masks become less obvious.

In addition to Figure 5.7, the visual masks are further inspected in Figure 5.8 predicted by the proposed approach across multiple metrics, using an example of motion blur distortions from the PIPAL dataset. As can be seen, the presence of blur is not uniform across the entire image; it becomes particularly noticeable when the direction of the motion blur is different from the pattern of the shirt (the right and upper parts). Here are the similar characteristics of predicted masks for MAE, PSNR, FLIP, and the first layer of VGG metrics, as in Figure 5.7. For SSIM, which already includes a divisive contrast component akin to visual masking modeling, the predicted mask assigns identity weights to regions where SSIM accurately predicts errors and lowers weights in areas where SSIM exaggerates the error.

Employing the enhanced metric as a loss In this part, the benefit of the enhanced IQMs in optimizing image restoration algorithms is investigated. To this end, MAE

TABLE 5.2: *Evaluation of a blind Gaussian denoising task when employing MAE and the equal combination of MAE and E-MAE as loss functions. The performance of the trained models is shown on synthetic Gaussian noise created with four distinct noise levels (σ) averaged across five benchmark datasets, consistent with the ones used in Zamir et al., 2022.*

	$\sigma = 15$			$\sigma = 25$			$\sigma = 50$			$\sigma = 60$		
Loss	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MAE	34.36	**0.94**	0.058	**31.94**	0.90	0.092	**28.82**	**0.84**	0.163	**28.02**	**0.81**	0.182
MAE + E-MAE	**34.37**	**0.94**	**0.055**	31.92	**0.91**	**0.087**	28.71	**0.84**	**0.152**	27.88	**0.81**	**0.167**

TABLE 5.3: *Evaluation of a blind Gaussian denoising task when employing VGG and the equal combination of VGG and E-VGG as loss functions. The performance of the trained models is shown on synthetic Gaussian noise created with four distinct noise levels (σ) averaged across five benchmark datasets, consistent with the ones used in Zamir et al., 2022.*

	$\sigma = 15$			$\sigma = 25$			$\sigma = 50$			$\sigma = 60$		
Loss	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MAE + VGG	34.16	0.936	**0.033**	31.68	0.900	**0.056**	28.51	0.826	**0.111**	27.66	0.797	0.132
MAE + E-VGG	**34.34**	**0.939**	0.049	**31.91**	**0.905**	0.078	**28.78**	**0.835**	0.139	**27.98**	**0.811**	0.155

and MAE+$\lambda\cdot$ E-MAE are employed as loss functions for training image denoising and motion deblurring using the state-of-the-art image restoration method, Restormer [Zamir et al., 2022]. For the denoising task, the images in the BSD400 dataset [Martin et al., 2001] are selected as the training set, and synthetic noise to these images is introduced by applying additive white Gaussian noise with a randomly chosen standard deviation ranging between 0 and 50. Each training is performed with the same number of iterations in an identical setup (e.g., learning rate). Then, the trained models are evaluated on five benchmark datasets, consistent with the ones used in [Zamir et al., 2022]. The evaluation is conducted for various noise levels, and the results are reported in Table 5.2. It can be observed that training just with the MAE loss leads to higher PSNRs, in particular for higher noise levels, but at the same time, image blur and contrast loss can be observed (refer to Figure 5.9). More perceptually inclined quality metrics penalize for such visual quality reduction, e.g., LPIPS is sensitive to excessive blur [Zhang et al., 2018b]. Combining with an E-MAE loss component clearly improves such metrics' scores consistently across various noise levels as well as the visual quality. For the motion deblurring task, the GoPro dataset [Nah et al., 2017] is employed for the training and evaluation. The combination of MAE and E-MAE enhances the deblurring results across different quality metrics (Table 5.4) and leads to a sharper appearance (Figure 5.10). In both tasks, it is empirically found that $\lambda = 1$ gives the best performance. It can also be observed that relying exclusively on the E-MAE loss component leads to worse results, which is expected, as indicated in [Ding et al., 2021b].

Employing the enhanced VGG metric as a loss Following the experiments in optimizing image restoration algorithms, the state-of-the-art image restoration method, Restormer [Zamir et al., 2022], is trained for the image-denoising with MAE + VGG and MAE + E-VGG in identical conditions. The results are reported in Table 5.3. The trained method with VGG shows a better LPIPS score as expected; however, it is found that denoising with E-VGG looks visually better, particularly in smooth low contrast regions (Figure 5.11).

TABLE 5.4: *Evaluation of a motion deblurring task when employing MAE and the equal combination of MAE and E-MAE as loss functions. The performance of the trained models is shown on synthetic blur created using the GoPro dataset Nah et al., 2017.*

Metric	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓
MAE	31.70	0.92	0.1030	0.0192
MAE + E-MAE	**31.78**	**0.93**	**0.1018**	**0.0184**

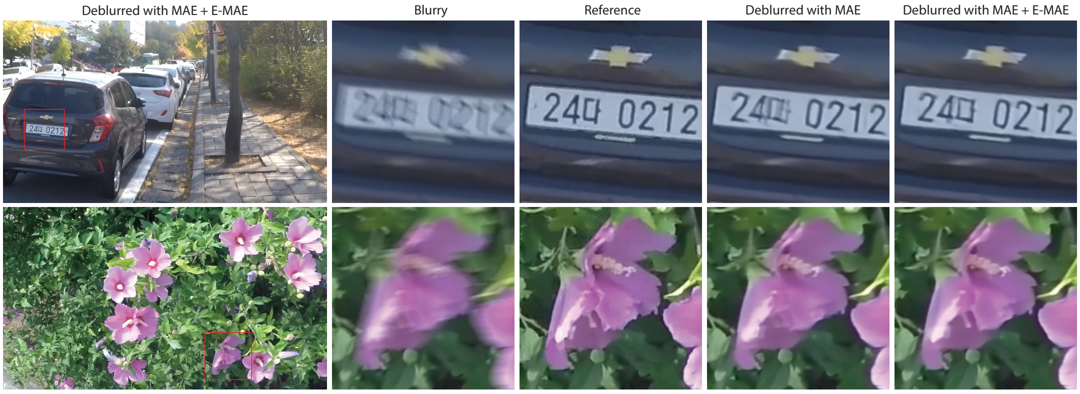


FIGURE 5.10: *Visual results for the motion deblurring task when employing MAE and MAE + E-MAE as loss functions.*

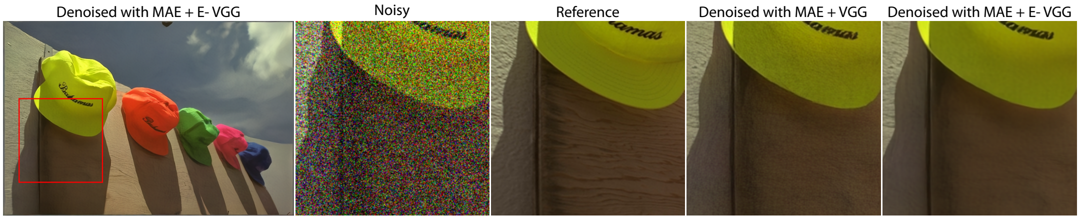


FIGURE 5.11: *Visual results in the image denoising task when employing MAE+VGG and MAE+E-VGG as loss functions. Denoising with MAE+VGG typically remains the noise in the dark region. On the other hand, MAE+E-VGG removes the noise successfully, which matches human perception better.*

5.3.3 Ablations

A set of ablations is performed to investigate the impact of reduced training data in terms of distortion levels, reference image number, and distortion type diversity on the E-MAE metric prediction accuracy.

Distortion levels The first experiment analyzes the importance of incorporating various distortion levels into the training set. In this regard, the E-MAE metric is trained using only one distortion level per category, and the results are reported in Figure 5.12. Interestingly, for all the datasets (except PIPAL), an inverse U-shape trend emerged across five different distortion levels, where the lowest correlation is observed when training with the minimum and maximum distortion levels (levels 1 and 5). Conversely, a moderate amount of distortion (level 3) appears to be sufficiently representative for each distortion category and achieved a comparable correlation to training with all five levels. This behavior can be anticipated because, at the lowest and highest distortion levels, the distortions are either barely visible or strongly visible, leading to the consistent selection of mostly extreme rating scores. Consequently, when the network is exclusively exposed to images with one such extreme distortion and rating levels, it fails to learn to differentiate between them. On the other hand, at moderate distortion levels where distortions are partially visible or invisible, the network has a better opportunity to learn masks that behave differently

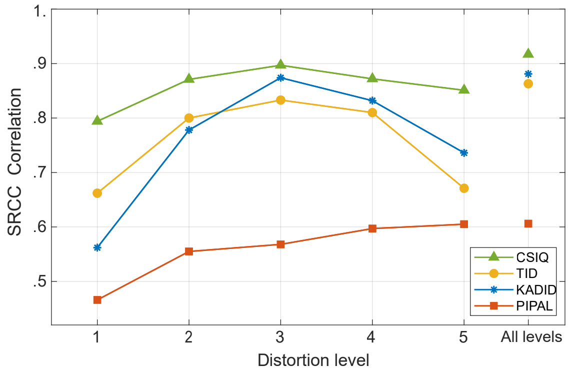


FIGURE 5.12: *Evaluation of E-MAE training performance using only selected distortion levels for each distortion category. The SRCC correlation is measured with the MOS data, and as a reference, the results of complete training with all distortion levels are included.*

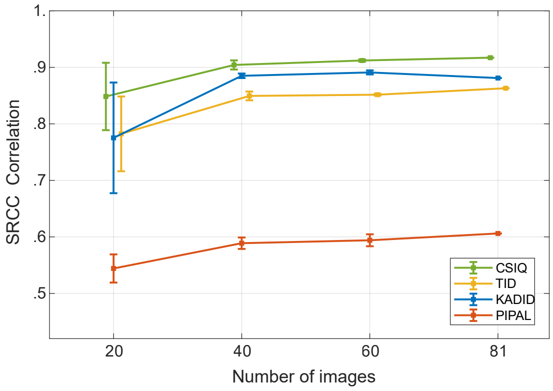


FIGURE 5.13: *Evaluation of E-MAE training performance using different numbers of the reference images (scenes). Multiple training runs have been performed for 20, 40, and 60 randomly selected scenes from the full set of 81 reference images. The data points represent the respective SRCC correlation averages over such runs, while the vertical bars depict the standard deviation.*

for varying spatial locations.

Dataset size Although employing a large-scale KADID dataset in the training (25 distortion types × five distortion levels), the number of reference images is limited to 81. This ablation aims to investigate the training performance by even further reducing the number of reference images. To this end, multiple runs of E-MAE metric training using randomly selected subsets of 20, 40, and 60 reference images are performed. Figure 5.13 presents the SRCC correlations averaged over multiple runs. The correlation differences between 40, 60, and the full set of 81 reference images are minor. In the case of 20 reference images, the performance is slightly lower and the variance higher, which indicates that 20 scenes might not be enough to capture image content variability.

Distortion diversity The impact of separate E-MAE training is investigated on specific distortion subsets such as noise, blur, combined noise, and blur, as well as the complete KADID dataset. At the test time, trained E-MAE versions on noise and blur subsets of the TID dataset, as well as its complete version, are evaluated. The results, presented in Table 5.5, reveal that training solely on the noise category unsurprisingly improves the SRCC correlation within that category; however, it also enhances the overall correlation for the TID dataset with respect to the original MAE. Conversely, training exclusively on blur does not improve the performance within the blur category itself, as the blur distortion already exhibits a strong correlation

TABLE 5.5: *The SRCC correlation with the MOS data for the E-MAE metric trained with specific distortion categories (noise, blur, noise&blur) and the entire (all) KADID dataset, as indicated in the brackets. The TID dataset is used for testing, where the "Category" columns indicate whether only the noise and blur subsets are considered or the entire dataset.*

Metric **Category:**	noise	blur	all
MAE	0.601	0.934	0.545
E-MAE (noise)	0.847	0.927	0.674
E-MAE (blur)	0.732	0.926	0.655
E-MAE (noise & blur)	0.841	0.936	0.726
E-MAE (all)	**0.906**	**0.955**	**0.857**

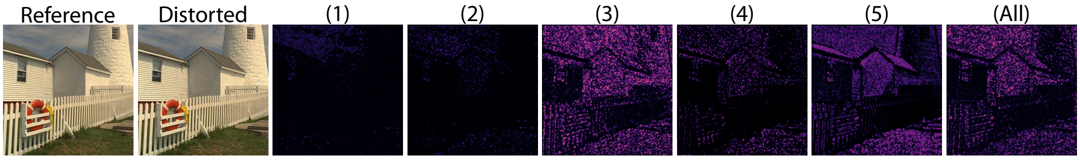


FIGURE 5.14: *Visualization of E-MAE error maps for Gaussian noise when it is trained with different levels of distortion.*

(0.934) for the MAE metric, making any improvement marginal. On the other hand, training with all categories combined significantly improves the correlation in both the noise and blur categories compared to training with only noise or blur categories, which can suggest that exposing the network to a wider range of distortion types enables better generalization.

Mask visualization for the ablation experiments The impact of separate E-MAE training is investigated in terms of how the quantitative measures in ablation experiments are reflected in the predicted error maps. To this end, the E-MAE error map is shown within various experimental setups (refer to Section 5.3.3) for a Gaussian noise distortion example from the TID dataset. Figure 5.14 shows the error maps when the metric is trained with only one distortion level per category. It is observed that the enhanced error maps have less visual similarity compared to training across all five levels when it is trained using the lowest and highest distortion levels, while it has the highest similarity when trained with distortion level 3. This observation is aligned with the correlation measurement in Figure 5.12. Additionally, Figure 5.15 shows the error maps when E-MAE is trained with a subset of images in the training set. Training with 20 reference images appears insufficient in generating accurate visual masking, which is aligned with the findings in Figure 5.13, where a reduction in correlation is observed with just 20 images. Conversely, training with 40 or 60 images closely approximates the results of training with the entire dataset, similarly reflected in the error maps. Lastly, in Figure 5.16, the maps obtained through training with different subsets of distortion categories from the training set are presented. Here, training exclusively with noise and blur can not produce precise masking, and including more categories is necessary to produce more localized masking. This is consistent with the correlation measures reported in Table 5.5.

Additional results for the error maps In addition to the presented results, Figure 5.17 shows the enhanced error maps for the MAE, SSIM, and FLIP metrics. It can be observed that the different enhanced metrics have almost the same error visibility maps.

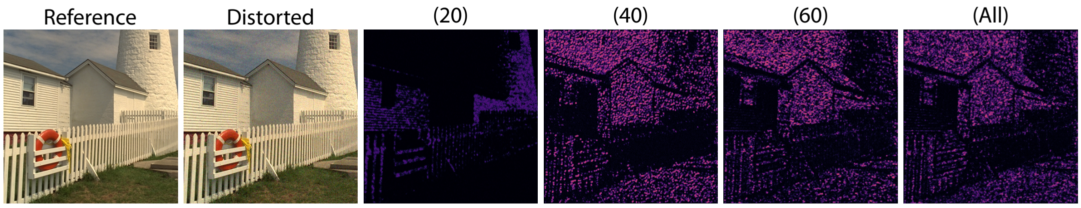


FIGURE 5.15: *Visualization of E-MAE error maps for Gaussian noise when it is trained with a different number of training images.*

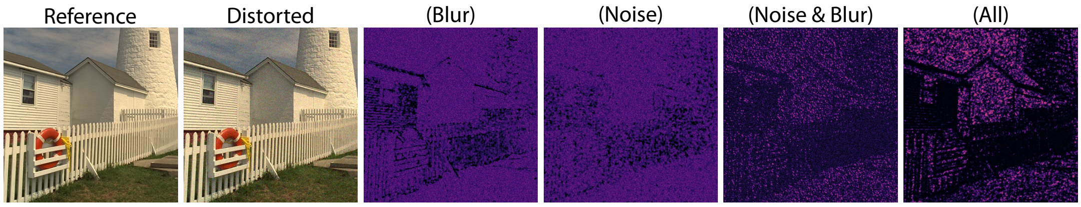


FIGURE 5.16: *Visualization of E-MAE error maps when it is trained with different distortion categories.*

5.4 Source Code

The source code for the enhanced MAE metric using the Pytorch [Paszke et al., 2019] is shared below to illustrate the simplicity of the visual masking approach:

```
class E_MAE(torch.nn.Module):
    # Initialize all the components
    def __init__(self):
        super(E_MAE, self).__init__()
        self.cuda()
        self.chns = [3]
        self.L = len(self.chns)
        self.mask_finder = []
        self.mask_finder_1 = MaskFinder(self.chns[0] * 2).cuda()
        self.mask_finder_1.requires_grad = False
        self.scaler_network = ScalerNetwork()
        model_path = os.path.abspath(os.path.join('weights',
         ↪  'E_MAE.pth'))
        self.load_state_dict(torch.load(model_path,
         ↪  map_location='cpu'), strict=False)

    # Returns the metric score
    def forward(self, y, x, as_loss=True, resize = True):
        mask = self.mask_finder_1(torch.cat([x, y], 1))
        score = ((mask * torch.abs(x - y))).mean()
        return score

    # Outputs the error map
    def E_MAE_map(self, y, x):
        C, H, W = x.shape[0:3]
        masks = self.mask_finder_1(torch.cat([x, y], 1))
        return self.scaler_network((masks*torch.abs(x - y)).mean([1],
         ↪  keepdim=True)) -
         ↪  self.scaler_network(torch.tensor(0.0).cuda().reshape(1,1,1,1)),
         ↪  masks[0]
```

5.5 Conclusion

This chapter presents a new approach to reducing the notorious gap between the existing quality metric prediction and the actual distortion visibility by the human observer. It is achieved by self-supervised training of a metric-specific network using the existing distortion datasets labeled with mean opinion score (MOS). Although overall image quality is rated with a single MOS value in the training data, by securing sufficient diversity of such training, as detailed in the ablation study, the network can leverage global MOS into a meaningful per-pixel mask. The mask, through different weighting of local distortion visibility, which also adapts to specific distortion types, helps a given metric to aggregate such local information into the comprehensive MOS value, as imposed by the training data. The mask can be learned directly in the image space for traditional metrics or in the feature space for recent learning-based metrics. In either case, it is trivial to incorporate into most of the existing metrics. Remarkably, the proposed approach improves the performance of commonly used metrics, such as MAE, PSNR, SSIM, and FLIP on all datasets tested. The prediction accuracy of recent learning-based metrics is typically substantially enhanced.

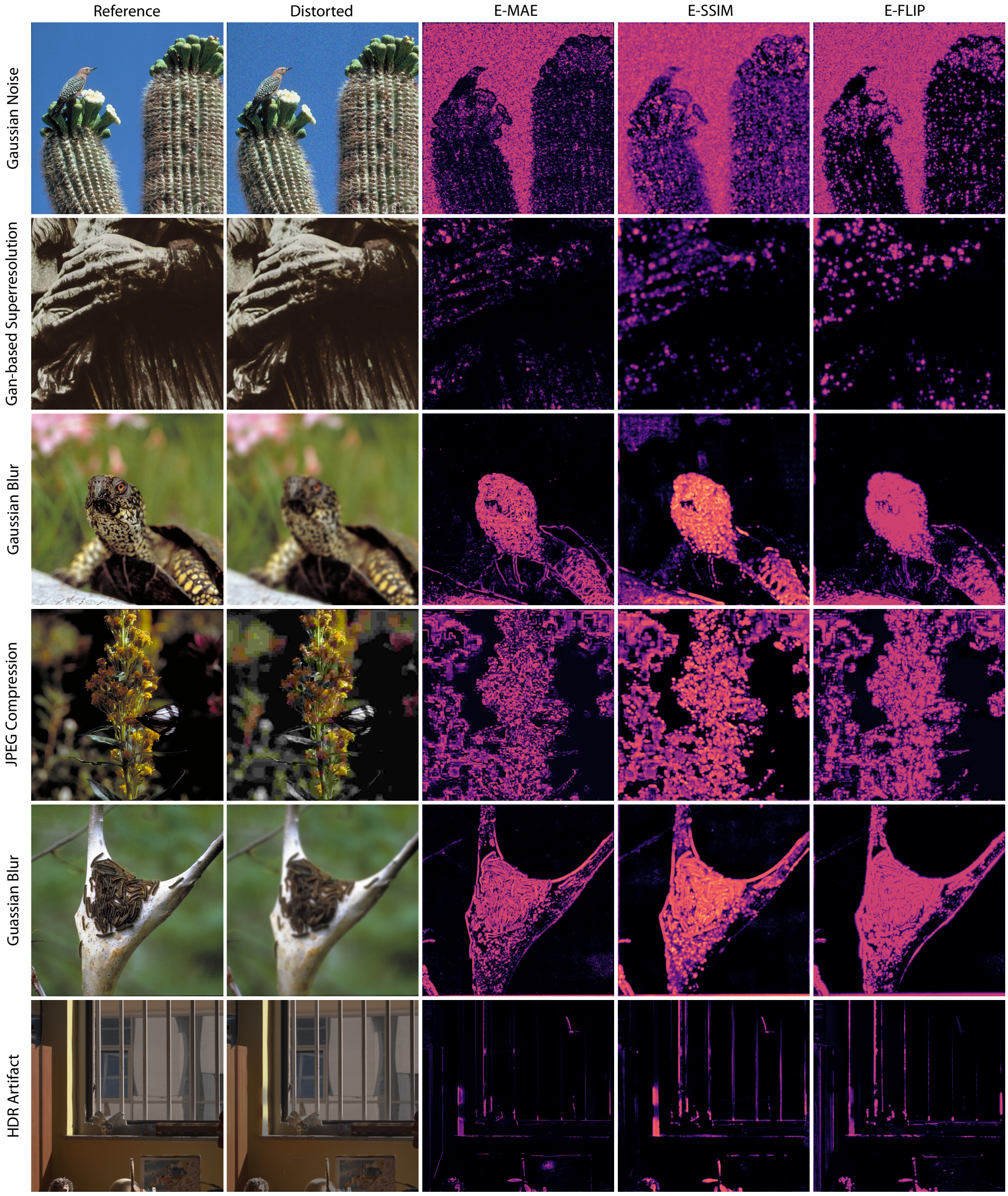


FIGURE 5.17: *Visualization of E-MAE, E-SSIM, and E-FLIP error maps across the same distortion types.*

Chapter 6

Conclusion

This thesis introduces innovative methods to improve image and video quality using a multi-exposure sensor. In addition, the quality of the reconstructions is evaluated with enhanced quality metrics, which further leads to developments in the reconstructions when employed as a loss function.

One of the methodologies described in Chapter 3 presents a non-parametric noise modeling that characterizes the sensor behavior better compared to the existing methods. Furthermore, Chapter 3 proposes an HDR reconstruction scheme that results in HDR images and videos that are free from motion blur. Then, Chapter 4 shows a pipeline that creates jointly high-speed HDR videos for the first time that include multiple interpolated frames enabling high-quality slow-motion display. The existing high-speed videos are examined with the proposed non-uniformity metric, providing valuable insights for future research. Lastly, Chapter 5 presents a methodology that enhances quality prediction of the existing full-reference image quality metrics by utilizing a visual masking approach. The proposed approach results in enhanced quality metrics ranging from traditional to feature-based metrics. Also, the enhanced metrics are employed as a loss function to improve the quality of the state-of-the-art denoising and deblurring algorithms.

Besides the contributions presented in this thesis, the following sections explore insights and potential future directions from each chapter.

6.1 Image and Video Enhancement and HDR Reconstruction

The proposed non-parametric noise modeling for pixel noise relies on a per-pixel matching. The observation leads to separate histograms for different exposures and color channels. Another dimension to discover is the effect of the neighboring pixels on the target pixel. The CMOS sensors distribute the collected voltage charges to the neighboring frames, known as pixel cross-talk [Blanksby and Loinaz, 2000]. Another well-known effect point spread function (PSF) is the measure of how the illumination of the neighboring pixels affects the target pixel. These effects are not considered in the proposed technique but could contribute to better modeling of the sensor noise characteristics.

Current multi-exposure sensors provide two or three exposures within the same spatial layout. Although the multi-exposure sensor with two exposures is utilized in this thesis, more than two exposures can better help the deblurring presented in Chapter 3 due to possible non-saturated overlapping regions. The same benefit can also be obtained in the case of alternating multi-exposures that can result in a better alignment between the exposures, leading to a wider dynamic range. Furthermore, the futuristic sensor design idea is that such a hardware architecture allows each pixel to change the exposure time independently. When combined with a smart capturing algorithm, this novel design can solve many vision tasks, such as denoising, deblurring, and HDR, without much effort. Considering a high dynamic range scene example, conventional sensors fail to capture all the radiance available in the scene, so very dark and bright regions are clipped. Such a sensor design can adjust the exposure time in dark regions to higher to get rid of the clamping and noise problems. In the same way, exposure time can be lowered in bright regions to prevent saturation. This design eventually gives a freedom to control the local regions of the captured scene. On the other hand, the sensor can decrease the motion blur to some extent. For example, if there is rapid motion, exposure time can be lowered for that specific region so that large motion blur, which is dependent on the exposure time, can be avoided.

6.2 Video Frame Interpolation for High Dynamic Range Sequences

The blur difference between the two exposures is successfully handled in Chapter 4, resulting in better frame interpolation, including the challenging scenes with non-uniform motions. The optical flow within the spatial frame is derived using the motion blur, and the defocus blur is not considered while handling the deblurring. Investigating optical blur and finding ways to remove it along with motion blur could be an interesting but challenging future direction. The current state-of-the-art image restoration methods [Zamir et al., 2022; Wang et al., 2022b] still treat them as two separate tasks due to the difficulties in removing the coexisting blur. However, the employed multi-exposure sensor design can significantly facilitate disentangling the motion blur that changes with exposure from the optical blur that remains constant between exposures.

Moreover, when capturing an HDR scene, the lowest exposure time is adjusted so that the long exposure is not very saturated, leaving enough valuable blurry information. This procedure is currently done manually; automatic selection of the optimal exposure time is an interesting future work direction that could lead to further performance improvements. It is important not to saturate local regions with complex motions to retain information. Otherwise, non-linearity can not be captured, as shown in Figure 4.10.

It is also interesting to port the proposed HDR-VFI technique to other multi-exposure sensors that are used in modern smartphones [GSMArena, 2022], such as Sony's Quad Bayer [Sony, 2022] and Samsung's Tetracell/Nonacell [Samsung, 2022] sensors. Such sensors with more than two exposures can expand the dynamic range further. The motion blur information encoded in the medium and long exposures provides continuous motion blur information that makes finding optical flow easier. The proposed methodology could also be applied to conventional sensors without a multi-exposure mode. Even if the misalignment between the long and short exposures decreases the performance of the optical flow derivation from the blurry long exposure, it is interesting to see the overall performance of frame interpolation. On the other hand, capturing short and long exposures starts at different times but ends simultaneously in the multi-exposure setup. By taking advantage of its unique features, HDR-VFI significantly outperforms the current state-of-the-art VFI methods.

The existing high-speed videos are analyzed to measure the non-uniformity levels by proposing a unique metric in Chapter 4. These measurements are later used to test the performance of state-of-the-art VFI methods together with the proposed HDR-VFI method. They were not included in the training phase due to the handling of the non-uniformity with the non-learnable component. However, the metric can be used to balance the training dataset for any method in terms of motion non-uniformity. This strategy has been employed in [Reda et al., 2022] but only for the motion magnitude. It could be interesting to see the effect of the balanced dataset with different levels of motion non-uniformity. That could also be employed for the preparation of the training set in Chapter 3 due to the direct supervision of splitting the optical flows. Increasing the number of samples with non-uniform motion makes the network more robust against such cases.

6.3 Enhanced Image Quality Measurement

As introduced in Chapter 5, the enhanced metrics with the visual masking approach better correlate with the perceptual dataset. However, the training is conducted with one of the commonly used datasets, KADID [Lin et al., 2019], which has an unknown experimental setting due to the collection of mean opinion scores in an uncontrollable remote setup. This condition can possibly affect the error map predictions, which is unknown at this stage. For example, the unknown and possibly inconsistent distance to the screen during the mean opinion score collection could affect the error map prediction for incompatible observer distances.

Another possible approach is to extend the metrics to handle the hallucinated content. Existing image quality metrics are tested with perceptual datasets having a limited number of distortions without covering the hallucinated content. In this context, inverse tone mapping methods [Endo et al., 2017; Liu et al., 2020c; Wang et al., 2023] are the direct targets that hallucinate the non-existing content regarding the realistic look of the final reconstructions. For this reason, it is worth creating a new perceptual dataset that focuses on the hallucinated content with mean opinion scores. This dataset can be used to train the proposed visual masking approach that aligns the enhanced metrics better with human perception in case of hallucinated content.

Additionally, although shown in Section 5.3.3 that reducing from 80 to 40 images the training dataset has limited impact on the visual masking performance, it could be attributed to certain redundancy in image content in the KADID dataset, where all distortions categories are repeated for all images. However, it would be interesting to radically increase the number and variability in the training dataset, which is difficult due to the lack of perceptual MOS data. Here, one inexpensive alternative would be to derive an artificial mean opinion score to increase the number of images. For this purpose, the authors of the KADID generated the KADIS dataset with 140k reference images without assigned mean opinion scores. The distortions are generated using the same parameters so that they match the original KADID dataset. This brings the challenge of finding the closest pairs of reference images that have common perceptual properties. To achieve this problem, the existing full-reference image quality metrics such as LPIPS [Zhang et al., 2018b], E-LPIPS, DISTS [Ding et al., 2022] and E-DISTS or the recently proposed OPENCLIP [Cherti et al., 2023] can be employed. The problem is that all these metrics find the most similar images sticking to the visual properties such as object and color matching, which means that the similarity of the images is evaluated in terms of content similarity. However, correct estimation of the mean opinion score requires identifying the perception of different distortion categories based on the percentage of textured and flat regions. Each distortion has its own characteristics, so using image similarity metrics that measure the distance between two clean images can not be generalized. The metric should evaluate the unique features of each distortion category. One way could be to estimate the distance of mean opinion scores specific to each distortions category. Existing image classifiers [Simonyan and Zisserman, 2015; He et al., 2016; Liu et al., 2022; Tu et al., 2022] with pre-trained weights can be adapted to this problem due to meaningful features in the early layers. The desired CNN should output distance measurement separately for each distortion type using a fully connected layer as in the classification task. This way, the average mean opinion scores of the closest matches can be assigned to the new distortions available in the KADIS dataset. In the end, the dataset with artificially created mean opinion scores should improve the quality estimations of the existing metrics using the same methodology proposed in Chapter 5. Additionally, fully evaluating the distance measurements of different categories could yield an overall similarity measurement. Then, it would also be interesting to check the behavior of the further enhanced metric as a loss function on the tasks of denoising and deblurring [Zamir et al., 2022] and whether it leads to further improvements. Considering all these concepts, this approach could be an exciting future research.

Bibliography - Own Work

Çoğalan, Uğur, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel [June 2022]. "Learning HDR video reconstruction for dual-exposure sensors with temporally-alternating exposures". In: *Comput. Graph.* 105.C, 57–72.

Çoğalan, Uğur, Mojtaba Bemana, Hans-Peter Seidel, and Karol Myszkowski [2023]. "Video Frame Interpolation for High Dynamic Range Sequences Captured with Dual-exposure Sensors". eng. In: *Computer Graphics Forum (Proc. EUROGRAPHICS)* 42.2, pp. 119–131.

Çoğalan, Uğur and Ahmet Oğuz Akyüz [2020]. "Deep Joint Deinterlacing and Denoising for Single Shot Dual-ISO HDR Reconstruction". In: *IEEE Transactions on Image Processing* 29, pp. 7511–7524.

Çoğalan, Uğur, Mojtaba Bemana, Hans-Peter Seidel, and Karol Myszkowski [2024]. "Enhancing image quality prediction with self-supervised visual masking". eng. In: *Computer Graphics Forum (Proc. EUROGRAPHICS)* 43.2, e15051.

Bibliography

Afifi, Mahmoud and Michael S. Brown [2020]. “Deep White-Balance Editing”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1394–1403.

Afifi, Mahmoud, Zhenhua Hu, and Liang Liang [2024]. *Optimizing Illuminant Estimation in Dual-Exposure HDR Imaging*. arXiv: 2403.02449 [cs.CV].

Aggarwal, M. and N. Ahuja [2001]. “Split aperture imaging for high dynamic range”. In: *ICCV*. Vol. 2, pp. 10–17.

Aguerrebere, C., A. Almansa, Y. Gousseau, J. Delon, and P. Musé [2014]. “Single shot high dynamic range imaging using piecewise linear estimators”. In: *ICCP*, pp. 1–10.

Aittala, Miika and Frédo Durand [2018]. “Burst Image Deblurring Using Permutation Invariant Convolutional Neural Networks”. In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*. Springer-Verlag, 748–764. ISBN: 978-3-030-01236-6.

Alghamdi, Masheal, Qiang Fu, Ali Thabet, and Wolfgang Heidrich [2021]. “Transfer Deep Learning for Reconfigurable Snapshot HDR Imaging Using Coded Masks”. In: *Computer Graphics Forum*.

An, V. G. and C. Lee [2017]. “Single-shot high dynamic range imaging via deep convolutional neural network”. In: *APSIPA*, pp. 1768–1772.

Andersson, Pontus, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D. Fairchild [2020]. “FLIP: A Difference Evaluator for Alternating Images”. In: *Proc. ACM Comput. Graph. Interact. Tech.* 3.2, 15:1–15:23.

Argaw, Dawit Mureja, Junsik Kim, Francois Rameau, Jae Won Cho, and In So Kweon [2021]. “Optical flow estimation from a single motion-blurred image”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2, pp. 891–900.

Ashikhmin, Michael [2002]. “A Tone Mapping Algorithm for High Contrast Images”. In: *Eurographics Workshop on Rendering*. The Eurographics Association. ISBN: 1-58113-534-3.

Bao, Wenbo, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang [2019]. “Depth-aware video frame interpolation”. In: *Proc. CVPR*, pp. 3703–3712.

Barron, Jonathan T. [2015]. “Convolutional Color Constancy”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV ’15. USA: IEEE Computer Society, 379–387. ISBN: 9781467383912.

Barten, Peter G.J. [1999]. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE – The International Society for Optical Engineering. ISBN: 0-8194-3496-5.

Basler [accessed on Sept. 17, 2021]. *https://www.baslerweb.com/en/products/cameras/area-scan-cameras/basler-beat/bea400-2kc/*.

Basler Dual Exposure [accessed on March 7, 2024]. *https://www.baslerweb.com/en/news/blaze-dual-exposure-hdr/*.

Batson, Joshua and Loic Royer [2019]. “Noise2Self: Blind Denoising by Self-Supervision”. In: *ICML*, pp. 524–533.

Bernardo Henz Eduardo S. L. Gastal, Manuel M. Oliveira [2021]. "Synthesizing Camera Noise using Generative Adversarial Networks". In: *IEEE Transactions on Visualization and Computer Graphics* 27.3, pp. 2123–2135.

Blanksby, A.J. and M.J. Loinaz [2000]. "Performance analysis of a color CMOS photogate image sensor". In: *IEEE Transactions on Electron Devices* 47.1, pp. 55–64.

Bradski, G. [2000]. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools*.

Buades, Antoni, Bartomeu Coll, and J-M Morel [2005]. "A non-local algorithm for image denoising". In: *CVPR*. Vol. 2, pp. 60–65.

Burger, Harold C, Christian J Schuler, and Stefan Harmeling [2012]. "Image denoising: Can plain neural networks compete with BM3D?" In: *CVPR*, pp. 2392–2399.

Čadík, Martin, Robert Herzog, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel [2013]. "Learning to predict localized distortions in rendered images". In: *Computer Graphics Forum (Proc. Eurographics)*. Vol. 32. 7, pp. 401–410.

Chakrabarti, Ayan [2016]. "A Neural Approach to Blind Motion Deblurring". In: *ECCV*, pp. 221–235.

Chandler, Damon M., Md Mushfiqul Alam, and Thien D. Phan [2014]. "Seven challenges for image quality research". In: *Human Vision and Electronic Imaging XIX*. Ed. by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder. Vol. 9014. International Society for Optics and Photonics. SPIE, p. 901402.

Chang, Meng, Huajun Feng, Zhihai Xu, and Qi Li [2022]. "Low-Light Image Restoration With Short- and Long-Exposure Raw Pairs". In: *IEEE Transactions on Multimedia* 24, pp. 702–714.

Chen, C., Q. Chen, J. Xu, and V. Koltun [2018]. "Learning to See in the Dark". In: *CVPR*, pp. 3291–3300.

Chen, Guanying, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang [2021]. "HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset". In: *Proc. CVPR*, pp. 2502–2511.

Chen, Jianing, Stephen J Carey, and Piotr Dudek [2017]. "Feature extraction using a portable vision system". In: *IEEE/RSJ Int. Conf. Intell. Robots Syst., Workshop Vis.-based Agile Auton. Navigation UAVs*. Vol. 2.

Chen, Jingwen, Jiawei Chen, Hongyang Chao, and Ming Yang [2018a]. "Image Blind Denoising with Generative Adversarial Network Based Noise Modeling". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164.

Chen, Jingwen, Jiawei Chen, Hongyang Chao, and Ming Yang [2018b]. "Image blind denoising with generative adversarial network based noise modeling". In: *CVPR*, pp. 3155–3164.

Cherti, M., R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev [June 2023]. "Reproducible Scaling Laws for Contrastive Language-Image Learning". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 2818–2829.

Chi, Zhixiang, Rasoul Mohammadi Nasiri, Zheng Liu, Juwei Lu, Jin Tang, and Konstantinos N Plataniotis [2020]. "All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling". In: *Proc. ECCV*, pp. 107–123.

Cho, Hojin, Seon Joo Kim, and Seungyong Lee [2014]. "Single-Shot High Dynamic Range Imaging Using Coded Electronic Shutter". In: *Comp. Graph. Forum* 33.7, 329–338.

Cho, S., Jue Wang, and S. Lee [2011]. "Handling outliers in non-blind image deconvolution". In: *CVPR*, pp. 495–502.

Choi, Inchang, Seung-Hwan Baek, and Min H Kim [2017]. “Reconstructing interlaced high-dynamic-range video using joint learning”. In: *IEEE Trans Image Processing* 26.11, pp. 5353–5366.

Choi, Myungsub, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee [2020]. “Channel attention is all you need for video frame interpolation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, pp. 10663–10671.

CMV12000 [2021]. *High Speed Machine Vision Global Shutter CMOS Image Sensor*. URL: https://ams.com/cmv12000 [visited on 09/29/2022].

Community, Blender Online [2020]. *Blender*.

Czolbe, Steffen, Oswin Krause, Ingemar Cox, and Christian Igel [2020]. “A Loss Function for Generative Neural Networks Based on Watson’s Perceptual Model”. In: *Proc. NIPS*.

Dabov, K., A. Foi, V. Katkovnik, and K. Egiazarian [2007]. “Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering”. In: *IEEE Trans. Image Processing* 16.8, pp. 2080–2095.

Dai, Shengyang and Ying Wu [2008]. “Motion from blur”. In: *Proc. CVPR*, pp. 1–8.

Daly, S. [1993]. “The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity”. In: *Digital Image and Human Vision*, pp. 179–206.

Debevec, Paul E. and Jitendra Malik [1997]. “Recovering High Dynamic Range Radiance Maps from Photographs”. In: *Proc. SIGGRAPH*, 369–378.

Debevec, Paul E and Jitendra Malik [2008]. “Recovering high dynamic range radiance maps from photographs”. In: *ACM SIGGRAPH 2008 classes*, pp. 1–10.

Ding, K., K. Ma, S. Wang, and E. P. Simoncelli [May 2022]. “Image Quality Assessment: Unifying Structure and Texture Similarity”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 44.05, pp. 2567–2581. ISSN: 1939-3539.

Ding, Keyan, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma [2021a]. “Locally Adaptive Structure and Texture Similarity for Image Quality Assessment”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. ACM.

Ding, Keyan, Kede Ma, Shiqi Wang, and Eero P Simoncelli [2021b]. “Comparison of full-reference image quality models for optimization of image processing systems”. In: *International Journal of Computer Vision* 129, pp. 1258–1281.

Drago, F., K. Myszkowski, T. Annen, and N. Chiba [2003]. “Adaptive Logarithmic Mapping For Displaying High Contrast Scenes”. In: *Computer Graphics Forum* 22.3, pp. 419–426.

Durand, Frédo and Julie Dorsey [July 2002]. “Fast bilateral filtering for the display of high-dynamic-range images”. In: *ACM Trans. Graph.* 21.3, 257–266. ISSN: 0730-0301.

Eilertsen, Gabriel, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger [2017]. “HDR Image Reconstruction from a Single Exposure Using Deep CNNs”. In: *ACM Trans. Graph.* 36.6.

Endo, Yuki, Yoshihiro Kanamori, and Jun Mitani [2017]. “Deep Reverse Tone Mapping”. In: *ACM Trans. Graph.* 36.6.

Fattal, Raanan, Dani Lischinski, and Michael Werman [July 2002]. “Gradient domain high dynamic range compression”. In: *ACM Trans. Graph.* 21.3, 249–256. ISSN: 0730-0301.

Fergus, Rob, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman [2006]. “Removing Camera Shake from a Single Photograph”. In: *ACM Trans. Graph.* 25.3, 787–794.

Ferwerda, James A, Peter Shirley, Sumanta N Pattanaik, and Donald P Greenberg [1997]. “A model of visual masking for computer graphics”. In: *Proc. ACM SIGGRAPH*, pp. 143–152.

Foi, Alessandro [2009]. “Clipped noisy images: Heteroskedastic modeling and practical denoising”. In: *Signal Processing* 89.12. Special Section: Visual Information Analysis for Security, pp. 2609–2629. ISSN: 0165-1684.

Foi, Alessandro, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian [2008]. “Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data”. In: *IEEE Trans Image Processing* 17.10, pp. 1737–1754.

Foley, J.M. [1994]. “Human luminance pattern-vision mechanisms: masking experiments require a new model”. In: *J. Opt. Soc. Am. A* 11.6, pp. 1710–19.

Froehlich, Jan, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel [2014]. *Creating Cinematic Wide Gamut HDR-Video for the Evaluation of Tone Mapping Operators and HDR-Displays*.

Galoogahi, H., A. Fagg, C. Huang, D. Ramanan, and S. Lucey [Oct. 2017]. “Need for Speed: A Benchmark for Higher Frame Rate Object Tracking”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1134–1143.

Gharbi, Michaël, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand [Dec. 2016]. “Deep joint demosaicking and denoising”. In: *ACM Trans. Graph.* 35.6. ISSN: 0730-0301.

Go, Chihiro, Yuma Kinoshita, Sayaka Shiota, and Hitoshi Kiya [2019]. “An image fusion scheme for single-shot high dynamic range imaging with spatially varying exposures”. In: *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences* 102.12, pp. 1856–1864.

Gong, D., J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi [2017]. “From Motion Blur to Motion Flow: A Deep Learning Solution for Removing Heterogeneous Motion Blur”. In: *CVPR*, pp. 3806–3815.

Gong, Dong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi [2017]. “From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur”. In: *Proc. CVPR*, pp. 2319–2328.

Granados, M., B. Ajdin, M. Wand, C. Theobalt, H. Seidel, and H. P. A. Lensch [2010]. “Optimal HDR reconstruction with linear digital cameras”. In: *CVPR*, pp. 215–222.

Gryaditskaya, Yulia, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel [2015]. “Motion Aware Exposure Bracketing for HDR Video”. In: *Comp. Graph. Forum* 34.4, 119–130.

GSMArena [2022]. *Quad Bayer Sensors: what they are and what they are not*. URL: https://www.gsmarena.com/quad_bayer_sensors_explained-news-37459.php [visited on 09/30/2022].

Gu, J., Y. Hitomi, T. Mitsunaga, and S. Nayar [2010]. “Coded rolling shutter photography: Flexible space-time sampling”. In: *ICCP*, pp. 1–8.

Gu, S., L. Zhang, W. Zuo, and X. Feng [2014]. “Weighted Nuclear Norm Minimization with Application to Image Denoising”. In: *CVPR*, pp. 2862–2869.

Gunturk, B.K., Y. Altunbasak, and R.M. Mersereau [2002]. “Color plane interpolation using alternating projections”. In: *IEEE Transactions on Image Processing* 11.9, pp. 997–1013.

Guo, Shi, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang [2019]. “Toward Convolutional Blind Denoising of Real Photographs”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1712–1722.

Gupta, Akash, Abhishek Aich, and Amit K. Roy-Chowdhury [2020]. “ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM ’20. Association for Computing Machinery, 256–264. ISBN: 9781450379885.

Hajisharif, Saghi, Joel Kronander, and Jonas Unger [2014]. “HDR Reconstruction for Alternating Gain (ISO) Sensor Readout”. In: *Eurographics 2014 - Short Papers*. Ed. by Eric Galin and Michael Wand. The Eurographics Association.

Hasinoff, Samuel W., Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy [2016]. “Burst Photography for High Dynamic Range and Low-Light Imaging on Mobile Cameras”. In: *ACM Trans. Graph.* 35.6.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun [2016]. “Deep residual learning for image recognition”. In: *CVPR*, pp. 770–78.

Healey, Glenn E and Raghava Kondepudy [1994]. “Radiometric CCD camera calibration and noise estimation”. In: *IEEE PAMI* 16.3, pp. 267–276.

Heide, Felix, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli [2014]. “FlexISP: A Flexible Camera Image Processing Framework”. In: *ACM Trans. Graph.* 33.6.

Hosu, V., H. Lin, T. Sziranyi, and D. Saupe [2020]. “KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment”. In: *IEEE Transactions on Image Processing* 29, pp. 4041–4056.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger [2017]. “Densely connected convolutional networks”. In: *Proc. CVPR*, pp. 4700–4708.

Hyun Kim, Tae and Kyoung Mu Lee [2014]. “Segmentation-free dynamic scene deblurring”. In: *Proc. CVPR*, pp. 2766–2773.

Iliadis, Michael, Leonidas Spinoulas, and Aggelos K Katsaggelos [2020]. “Deepbinarymask: Learning a binary mask for video compressive sensing”. In: *Digital Signal Processing* 96, p. 102591.

IMX, Sony [accessed on Sept. 17, 2021]. *https://www.framos.com/en/news/sony-launches-highly-sensitive-4/3-cmos-sensor-for-4k-surveillance*.

Jaderberg, Max, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu [2015]. “Spatial Transformer Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc.

Janai, Joel, Fatma Guney, Jonas Wulff, Michael J Black, and Andreas Geiger [2017]. “Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data”. In: *Proc. CVPR*, pp. 3597–3607.

Janesick, James R. [2001]. *Scientific Charge-coupled Devices*.

Jia, Sen and Neil D.B. Bruce [2020]. “EML-NET: An Expandable Multi-Layer NETwork for saliency prediction”. In: *Image and Vision Computing* 95, p. 103887. ISSN: 0262-8856.

Jia, Xixi, Sanyang Liu, Xiangchu Feng, and Lei Zhang [2019]. “FOCNet: A Fractional Optimal Control Network for Image Denoising”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6047–6056.

Jiang, Huaizu, Deqing Sun, Varan Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz [2018]. “Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9000–9008.

Jin, M., G. Meishvili, and P. Favaro [2018]. “Learning to Extract a Video Sequence from a Single Motion-Blurred Image”. In: *CVPR*, pp. 6334–6342.

Jin, Meiguang, Zhe Hu, and Paolo Favaro [2019]. “Learning to extract flawless slow motion from blurry videos”. In: *Proc. CVPR*, pp. 8112–8121.

Jinjin, Gu, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao [2020]. “PIPAL: A Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration”. In: *Proc. ECCV*, pp. 633–651.

Jonschkowski, Rico, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova [2020]. “What matters in unsupervised optical flow”. In: *Proc. ECCV*, pp. 557–572.

Kalantari, Nima Khademi and Ravi Ramamoorthi [July 2017]. “Deep high dynamic range imaging of dynamic scenes”. In: *ACM Trans. Graph.* 36.4. ISSN: 0730-0301.

Kalantari, Nima Khademi and Ravi Ramamoorthi [2019]. “Deep HDR Video from Sequences with Alternating Exposures”. In: *Comp Graph Forum (Proc. Eurographics)* 38 [2].

Kalantari, Nima Khademi, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen [2013]. “Patch-based high dynamic range video.” In: *ACM Trans. Graph.* 32.6, pp. 202–1.

Kang, Sing Bing, Matthew Uyttendaele, Simon Winder, and Richard Szeliski [2003]. “High Dynamic Range Video”. In: *ACM Trans. Graph.* 22.3, 319–325.

Ke, J., Q. Wang, Y. Wang, P. Milanfar, and F. Yang [2021]. “MUSIQ: Multi-scale Image Quality Transformer”. In: *Proc. ICCV*, pp. 5128–5137.

Kim, Min H. and Jan Kautz [2008]. “Consistent tone reproduction”. In: *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*. CGIM ’08. ACTA Press, 152–159. ISBN: 9780889867208.

Kim, T. and K. Lee [2015]. “Generalized video deblurring for dynamic scenes”. In: *CVPR*, pp. 5426–5434.

Kim, T. H., K. M. Lee, B. Schölkopf, and M. Hirsch [2017]. “Online Video Deblurring via Dynamic Temporal Blending Network”. In: *ICCV*, pp. 4058–4067.

Kingma, Diederik P. and Jimmy Ba [2015]. “Adam: A Method for Stochastic Optimization”. In: *Proc. ICLR*.

Koh, Jaihyun, Jangho Lee, and Sungroh Yoon [2021]. “Single-image deblurring with neural networks: A comparative survey”. In: *Computer Vision and Image Understanding* 203, p. 103134.

Kokkinos, Filippos and Stamatios Lefkimmiatis [2019]. “Iterative Joint Image Demosaicking and Denoising Using a Residual Denoising Network”. In: *IEEE Transactions on Image Processing* 28.8, pp. 4177–4188.

Krawczyk, Grzegorz, Karol Myszkowski, and Hans-Peter Seidel [2005]. “Lightness Perception in Tone Reproduction for High Dynamic Range Images”. In: *Computer Graphics Forum* 24.3, pp. 635–645.

Kronander, Joel, Stefan Gustavson, Gerhard Bonnet, and Jonas Unger [2013]. “Unified HDR reconstruction from raw CFA data”. In: *IEEE International Conference on Computational Photography (ICCP)*, pp. 1–9.

Krull, A., T. Buchholz, and F. Jug [June 2019]. “Noise2Void - Learning Denoising From Single Noisy Images”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 2124–2132.

Krull, Alexander, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug [2020]. “Probabilistic Noise2Void: Unsupervised Content-Aware Denoising”. In: *Frontiers in Computer Science* 2. ISSN: 2624-9898.

Kupyn, Orest, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas [2018]. “DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8183–8192.

Kupyn, Orest, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang [2019]. “Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better”. In: *ICCV*, pp. 8878–8887.

Laine, Samuli, Tero Karras, Jaakko Lehtinen, and Timo Aila [2019]. “High-Quality Self-Supervised Deep Image Denoising”. In: *NiPS*. Vol. 32, pp. 6970–6980.

Larson, Eric C. and Damon M. Chandler [2010]. "Most apparent distortion: full-reference image quality assessment and the role of strategy". In: *J. Electronic Imaging* 19, p. 011006.

Lee, Hyeongmin, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee [2020]. "AdaCoF: Adaptive collaboration of flows for video frame interpolation". In: *Proc. CVPR*, pp. 5316–5325.

Lefkimmiatis, Stamatios [2018]. "Universal Denoising Networks : A Novel CNN Architecture for Image Denoising". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3204–3213.

Legge, Gordon E. and John M. Foley [1980]. "Contrast masking in human vision". In: *J. Opt. Soc. Am.* 70.12, pp. 1458–1471.

Lehtinen, Jaakko, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila [2018]. "Noise2Noise: Learning Image Restoration without Clean Data". In: *Proc. of Machine Learning Research*. Vol. 80, pp. 2965–2974.

Lenzen, Frank and Otmar Scherzer [2011]. "Partial Differential Equations for Zooming, Deinterlacing and Dejittering". In: *Int. J Comp. Vis.* 92, pp. 162–176.

Levin, A., Y. Weiss, F. Durand, and W. T. Freeman [2009]. "Understanding and evaluating blind deconvolution algorithms". In: *CVPR*, pp. 1964–1971.

Liao, Xingran, Baoliang Chen, Hanwei Zhu, Shiqi Wang, Mingliang Zhou, and Sam Kwong [2022]. "DeepWSD: Projecting Degradations in Perceptual Space to Wasserstein Distance in Deep Feature Space". In: *Proceedings of the 30th ACM International Conference on Multimedia*. ACM.

Liba, Orly, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T. Barron, Dillon Sharlet, Ryan Geiss, Samuel W. Hasinoff, Yael Pritch, and Marc Levoy [2019]. "Handheld Mobile Photography in Very Low Light". In: *ACM Trans. Graph.* 38.6.

Lin, Hanhe, Vlad Hosu, and Dietmar Saupe [2019]. "KADID-10k: A Large-scale Artificially Distorted IQA Database". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3.

Liu, Ce, Richard Szeliski, Sing Bing Kang, C Lawrence Zitnick, and William T Freeman [2007]. "Automatic estimation and removal of noise from a single image". In: *IEEE PAMI* 30.2, pp. 299–314.

Liu, Lin, Xu Jia, Jianzhuang Liu, and Qi Tian [2020a]. "Joint Demosaicing and Denoising With Self Guidance". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2237–2246.

Liu, Peidong, Joel Janai, Marc Pollefeys, Torsten Sattler, and Andreas Geiger [2020b]. "Self-Supervised Linear Motion Deblurring". In: *IEEE Robotics and Automation Letters* 5.2, pp. 2475–2482.

Liu, Y., W. Lai, Y. Chen, Y. Kao, M. Yang, Y. Chuang, and J. Huang [June 2020c]. "Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1648–1657.

Liu, Yu-Lun, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang [2019]. "Deep video frame interpolation using cyclic frame generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 8794–8802.

Liu, Ze, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo [2022]. "Swin Transformer V2: Scaling Up Capacity and Resolution". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009.

Liu, Ziwei, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun [2014]. "Fast Burst Images Denoising". In: *ACM Trans. Graph.* 33.6.

Lu, Yue M., Mina Karzand, and Martin Vetterli [2010]. "Demosaicking by Alternating Projections: Theory and Fast One-Step Implementation". In: *IEEE Transactions on Image Processing* 19.8, pp. 2085–2098.

Lubin, J. [1995]. "A visual discrimination model for imaging system design and development". In: *Vision Models for Target Detection and Recognition*. Ed. by E. Peli. World Scientific, pp. 245–283.

Ma, Sizhuo, Shantanu Gupta, Arin C. Ulku, Claudio Brushini, Edoardo Charbon, and Mohit Gupta [July 2020]. "Quanta Burst Photography". In: *ACM Trans Graph* 39.4.

Malvar, Henrique S, Li-wei He, and Ross Cutler [2004]. "High-quality linear interpolation for demosaicing of Bayer-patterned color images". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. IEEE, pp. iii–485.

Mangiat, Stephen and Jerry Gibson [2010]. "High dynamic range video with ghost removal". In: *Applications of Digital Image Processing XXXIII*. Vol. 7798, pp. 307 –314.

Mann, S. and R. W. Picard [1995]. "On Being 'Undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures". In: *ISfT*, pp. 442–448.

Mantiuk, Rafał, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich [2011a]. "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions". In: *ACM Transactions on Graphics (Proc. SIGGRAPH)* 30.4, pp. 1–14.

Mantiuk, Rafał, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich [June 2011b]. "HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions". In: *ACM Trans. Graph.* 30.4. ISSN: 0730-0301.

Mantiuk, Rafał K, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney [2021]. "FovVideoVDP: A visible difference predictor for wide field-of-view video". In: *ACM Transactions on Graphics (Proc. SIGGRAPH)* 40.4, pp. 1–19.

Mao, Xiaojiao, Chunhua Shen, and Yu-Bin Yang [2016]. "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections". In: *NiPS*, pp. 2802–2810.

Marnerides, D., T. Bashford-Rogers, J. Hatchett, and K. Debattista [2018]. "ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content". In: *Comp. Graph. Forum* 37.2, pp. 37–49.

Martel, Julien NP, Lorenz K Mueller, Stephen J Carey, Piotr Dudek, and Gordon Wetzstein [2020]. "Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.7, pp. 1642–1653.

Martel, Julien NP, Lorenz K Müller, Stephen J Carey, and Piotr Dudek [2017]. "High-speed depth from focus on a programmable vision chip using a focus tunable lens". In: *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4.

Martin, D., C. Fowlkes, D. Tal, and J. Malik [2001]. "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics". In: *Proc. 8th Int'l Conf. Computer Vision*. Vol. 2, pp. 416–423.

Mertens, T., J. Kautz, and F. Van Reeth [2007]. "Exposure Fusion". In: *Pacific Graphics*, pp. 382–390.

Michaeli, Tomer and Michal Irani [2014]. "Blind Deblurring Using Internal Patch Recurrence". In: *ECCV*, pp. 783–798.

Mildenhall, B., J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll [2018]. "Burst Denoising with Kernel Prediction Networks". In: *CVPR*, pp. 2502–2510.

Mitsunaga, T. and S. K. Nayar [1999]. "Radiometric self calibration". In: *CVPR*, 374–380 Vol. 1.

Moorthy, Anush Krishna and Alan Conrad Bovik [2010]. "A Two-Step Framework for Constructing Blind Image Quality Indices". In: *IEEE Signal Processing Letters* 17.5, pp. 513–516.

Moran, Nick, Dan Schmidt, Yu Zhong, and Patrick Coady [2020]. "Noisier2Noise: Learning to Denoise from Unpaired Noisy Data". In: *CVPR*, pp. 12064–12072.

Mustaniemi, Janne, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä [2020]. "LSD$_2$-Joint Denoising and Deblurring of Short and Long Exposure Images with Convolutional Neural Networks". In: *BMVC*.

Nah, S., S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee [2019]. "NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study". In: *CVPRW*, pp. 1996–2005.

Nah, Seungjun, Tae Hyun Kim, and Kyoung Mu Lee [2017]. "Deep multi-scale convolutional neural network for dynamic scene deblurring". In: *Proc. CVPR*, pp. 3883–3891.

Nayar and Branzoi [2003]. "Adaptive dynamic range imaging: optical control of pixel exposures over space and time". In: *ICCV*, 1168–1175 vol.2.

Nayar, S. K., V. Branzoi, and T. E. Boult [2004]. "Programmable imaging using a digital micromirror array". In: *CVPR*. Vol. 1, pp. I–I.

Nayar, Shree K. and Tomoo Mitsunaga [2000]. "High Dynamic Range Imaging: Spatially Varying Pixel Exposures". In: *CVPR*, pp. 1472–1479.

Nguyen, C. M., J. P. Martel, and G. Wetzstein [Aug. 2022]. "Learning Spatially Varying Pixel Exposures for Motion Deblurring". In: *2022 IEEE International Conference on Computational Photography (ICCP)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1–11.

Niklaus, Simon and Feng Liu [2020]. "Softmax splatting for video frame interpolation". In: *Proc. CVPR*, pp. 5437–5446.

Niklaus, Simon, Long Mai, and Feng Liu [2017]. "Video frame interpolation via adaptive separable convolution". In: *Proc. ICCV*, pp. 261–270.

Oh, Tae-Hyun, Joon-Young Lee, Yu-Wing Tai, and In So Kweon [2015]. "Robust High Dynamic Range Imaging by Rank Minimization". In: *IEEE PAMI* 37.6, pp. 1219–1232.

Omnivision [accessed on Sept. 17, 2021]. *https://www.ovt.com/sensors/OH02A10/*.

Oten, Remzi and Jim Li [2011]. *Method for reducing row noise with dark pixel data*. US Patent 7,982,784 B2.

Pan, Jinshan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang [2021]. "Physics-Based Generative Adversarial Models for Image Restoration and Beyond". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.7, pp. 2449–2462.

Park, Junheum, Keunsoo Ko, Chul Lee, and Chang-Su Kim [2020]. "BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation". In: *Proc. ECCV*, pp. 109–125.

Park, Junheum, Chul Lee, and Chang-Su Kim [2021]. "Asymmetric bilateral motion estimation for video frame interpolation". In: *Proc. ICCV*, pp. 14539–14548.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith

Chintala [2019]. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035.

Plötz, T. and S. Roth [2017]. “Benchmarking Denoising Algorithms with Real Photographs”. In: *CVPR*, pp. 2750–2759.

Ponomarenko, Nikolay, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo [2015]. “Image database TID2013: Peculiarities, results and perspectives”. In: *Signal Processing: Image Communication* 30, pp. 57–77.

Prabhakar, K. R., V. S. Srikar, and R. V. Babu [2017]. “DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs”. In: *ICCV*, pp. 4724–4732.

Prashnani, Ekta, Hong Cai, Yasamin Mostofi, and Pradeep Sen [2018]. “Pieapp: Perceptual image-error assessment through pairwise preference”. In: *Proc. CVPR*, pp. 1808–1817.

Purohit, Kuldeep, Anshul Shah, and AN Rajagopalan [2019]. “Bringing alive blurred moments”. In: *Proc. CVPR*, pp. 6830–6839.

Qiu, Jiayan, Xinchao Wang, Stephen J Maybank, and Dacheng Tao [2019]. “World from blur”. In: *Proc. CVPR*, pp. 8493–8504.

Quan, Yuhui, Mingqin Chen, Tongyao Pang, and Hui Ji [2020]. “Self2Self With Dropout: Learning Self-Supervised Denoising From Single Image”. In: *CVPR*.

Rana, Aakanksha, Praveer Singh, Giuseppe Valenzise, Frederic Dufaux, Nikos Komodakis, and Aljosa Smolic [2020]. “Deep Tone Mapping Operator for High Dynamic Range Images”. In: *IEEE Transactions on Image Processing* 29, pp. 1285–1298.

Reda, Fitsum, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless [2022]. “FILM: Frame Interpolation for Large Motion”. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Berlin, Heidelberg: Springer-Verlag, 250–266.

Reda, Fitsum A, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro [2019]. “Unsupervised video interpolation using cycle consistency”. In: *Proc. ICCV*, pp. 892–900.

Reinhard, Erik, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski [2010]. *High dynamic range imaging: acquisition, display, and image-based lighting*.

Reinhard, Erik, Michael Stark, Peter Shirley, and James Ferwerda [2002]. “Photographic Tone Reproduction for Digital Images”. In: *ACM Trans. Graph.* 21.3, 267–276.

Reinhard, Erik, Greg Ward, Sumanta Pattanaik, and Paul Debevec [2006]. “05 - Display Devices”. In: *High Dynamic Range Imaging*. Ed. by Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. The Morgan Kaufmann Series in Computer Graphics. San Francisco: Morgan Kaufmann, pp. 167–185. ISBN: 978-0-12-585263-0.

Rekleitis, Ioannis [1995]. “Visual motion estimation based on motion blur interpretation”.

Robertson, Mark A., Sean Borman, and Robert L. Stevenson [2003]. “Estimation-theoretic approach to dynamic range enhancement using multiple exposures”. In: *J Electronic Imaging* 12.2, pp. 219 –228.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox [2015]. “U-net: Convolutional networks for biomedical image segmentation”. In: *MICCAI*, pp. 234–41.

Rozumnyi, Denys, Martin R Oswald, Vittorio Ferrari, and Marc Pollefeys [2022]. "Motion-from-Blur: 3D Shape and Motion Estimation of Motion-blurred Objects in Videos". In: *Proc. CVPR*, pp. 15990–15999.

Saad, Michele A., Alan C. Bovik, and Christophe Charrier [2012]. "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain". In: *IEEE Transactions on Image Processing* 21.8, pp. 3339–3352.

Sampat, Mehul P., Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K. Markey [2009]. "Complex Wavelet Structural Similarity: A New Image Similarity Index". In: *IEEE Transactions on Image Processing* 18.11, pp. 2385–2401.

Samsung [2022]. *ISOCELL GN1 Sensors*. URL: https://news.samsung.com/global/video-introducing-samsungs-versatile-new-image-sensor-the-isocell-gn1 [visited on 09/15/2022].

Santos, Marcel Santana, Ren Tsang, and Nima Khademi Kalantari [July 2020]. "Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss". In: *ACM Trans Graphics (Proc. SIGGRAPH Asia)* 39.4.

Schmidt, U., C. Rother, S. Nowozin, J. Jancsary, and S. Roth [2013]. "Discriminative Non-blind Deblurring". In: *CVPR*, pp. 604–611.

Schoueri, Yasmina, Milena Scaccia, and Ioannis Rekleitis [2009]. "Optical flow from motion blurred color images". In: *2009 Canadian Conference on Computer and Robot Vision*, pp. 1–7.

Schuler, C. J., H. C. Burger, S. Harmeling, and B. Schölkopf [2013]. "A Machine Learning Approach for Non-blind Image Deconvolution". In: *CVPR*, pp. 1067–1074.

Schöberl, M., A. Belz, J. Seiler, S. Foessel, and A. Kaup [2012]. "High dynamic range video by spatially non-regular optical filtering". In: *ICIP*, pp. 2757–2760.

Schöberl, Michael, Alexander Belz, Arne Nowak, Jürgen Seiler, André Kaup, and Siegfried Foessel [2012]. "Building a High Dynamic Range Video Sensor with Spatially Non-Regular Optical Filtering". In: *Proc. SPIE* 8499.

Sebastian, Stephen, Johannes Burge, and Wilson S. Geisler [2015]. "Defocus blur discrimination in natural images with natural optics". In: *Journal of Vision* 15.5, pp. 16–16.

Seger, Ulrich, Uwe Apel, and Bernd Höfflinger [1999]. "HDRC-imagers for natural visual perception". In: *Handbook of Computer Vision and Application* 1, pp. 223–235.

Serrano, Ana, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia [2016]. "Convolutional sparse coding for high dynamic range imaging". In: *Computer Graphics Forum*. Vol. 35. 2. Wiley Online Library, pp. 153–163.

Sheikh, Hamid R and Alan C Bovik [2005]. "Information embedded in images: Joint estimation of information and distortion". In: *Proceedings of the IEEE International Conference on Image Processing*. Vol. 2, pp. II–702.

Sheikh, H.R. and A.C. Bovik [2006]. "Image information and visual quality". In: *IEEE Transactions on Image Processing* 15.2, pp. 430–444.

Shen, Wang, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao [2020a]. "Blurry video frame interpolation". In: *Proc. CVPR*, pp. 5114–5123.

Shen, Wang, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao [2020b]. "Video Frame Interpolation and Enhancement via Pyramid Recurrent Framework". In: *IEEE Trans Image Proc* 30, pp. 277–292.

Sim, Hyeonjun, Jihyong Oh, and Munchurl Kim [2021]. "XVFI: Extreme video frame interpolation". In: *Proc. ICCV*, pp. 14489–14498.

Simonyan, Karen and Andrew Zisserman [2015]. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *Proc. ICLR*.

Sony [2022]. *Quad Bayer Coding*. URL: https://www.sony-semicon.com/en/technology/mobile/quad-bayer-coding.html [visited on 09/30/2022].

Su, Shaolin, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang [2020]. "Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network". In: *Proc. CVPR*, pp. 3664–3673.

Su, Shuochen, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang [2017]. "Deep video deblurring for hand-held cameras". In: *Proc. CVPR*, pp. 1279–1288.

Sun, Deqing, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz [2018]. "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume". In: *Proc. CVPR*, pp. 8934–8943.

Sun, J., Wenfei Cao, Zongben Xu, and J. Ponce [2015]. "Learning a convolutional neural network for non-uniform motion blur removal". In: *CVPR*, pp. 769–777.

Sun, Libin, Sunghyun Cho, Jue Wang, and James Hays [2014]. "Good Image Priors for Non-blind Deconvolution". In: *ECCV*, pp. 231–246.

Tao, Xin, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia [2018]. "Scale-recurrent network for deep image deblurring". In: *CVPR*, pp. 8174–8182.

Teed, Zachary and Jia Deng [2020]. "RAFT: Recurrent all-pairs field transforms for optical flow". In: *Proc. ECCV*, pp. 402–419.

Tocci, Michael D., Chris Kiser, Nora Tocci, and Pradeep Sen [2011]. "A Versatile HDR Video Production System". In: *ACM Trans. Graph.* 30.4.

Tu, Zhengzhong, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li [2022]. "MaxViT: Multi-axis Vision Transformer". In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer-Verlag, 459–479. ISBN: 978-3-031-20052-6.

Tursun, Cara, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk [2019]. "Luminance-Contrast-Aware Foveated Rendering". In: *ACM Transactions on Graphics (Proc. SIGGRAPH)* 38.4.

Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky [2018]. "Deep Image Prior". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454.

vision, Emergent [accessed on Sept. 17, 2021]. *https://emergentvisiontec.com/products/area-scan-cameras/25-gige-area-scan-cameras-hb-series/hb-12000/*.

Vogels, Thijs, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák [2018]. "Denoising with kernel prediction and asymmetric loss functions". In: *ACM Transactions on Graphics (Proc. SIGGRAPH)* 37.4, pp. 1–15.

Wang, C., B. Chen, HP. Seidel, K. Myszkowski, and A. Serrano [2022a]. "Learning a self-supervised tone mapping operator via feature contrast masking loss". In: *Computer Graphics Forum* 41.2, pp. 71–84.

Wang, Chao, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler [2023]. "GlowGAN: Unsupervised Learning of HDR Images from LDR Images in the Wild". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10509–10519.

Wang, Lin and Kuk-Jin Yoon [2022]. "Deep Learning for HDR Imaging: State-of-the-Art and Future Trends". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12, pp. 8874–8895.

Wang, Z., E.P. Simoncelli, and A.C. Bovik [2003]. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2, 1398–1402 Vol.2.

Wang, Zhendong, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li [2022b]. "Uformer: A general u-shaped transformer for image restoration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693.

Wang, Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli [2004a]. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612.

Wang, Zhou, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli [2004b]. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Trans. Image Processing* 13.4, pp. 600–612.

Watson, A.B. [1993]. "Visually optimal DCT quantization matrices for individual images". In: *Proc. DCC '93: Data Compression Conference*, pp. 178–187.

Watson, Andrew B. and Albert J. Ahumada [2011]. "Blur clarified: A review and synthesis of blur discrimination". In: *Journal of Vision* 11.5, pp. 10–10.

Werlberger, Manuel, Thomas Pock, Markus Unger, and Horst Bischof [2011]. "Optical Flow Guided TV-L1 Video Interpolation and Restoration". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 273–286. ISBN: 978-3-642-23094-3.

Whyte, O., J. Sivic, A. Zisserman, and J. Ponce [2010]. "Non-uniform deblurring for shaken images". In: *CVPR*, pp. 491–498.

Wieschollek, P., M. Hirsch, B. Schölkopf, and H. Lensch [2017]. "Learning Blind Motion Deblurring". In: *ICCV*, pp. 231–240.

Wilson, Hugh R. and Douglas J. Gelb [1984]. "Modified line-element theory for spatial-frequency and width discrimination". In: *J. Opt. Soc. Am. A* 1.1, pp. 124–131.

Wolski, Krzysztof, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K Mantiuk [2018]. "Dataset and metrics for predicting local visible differences". In: *ACM Transactions on Graphics* 37.5, pp. 1–14.

Wu, Shangzhe, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang [2018]. "Deep High Dynamic Range Imaging with Large Foreground Motions". In: *ECCV*, pp. 120–135.

Wu, Xiaohe, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo [2020]. "Unpaired Learning of Deep Image Denoising". In: *ECCV*. Springer, pp. 352–368.

Wuerger, Sophie, Maliha Ashraf, Minjung Kim, Jasna Martinovic, María Pérez-Ortiz, and Rafał K. Mantiuk [2020]. "Spatio-chromatic contrast sensitivity under mesopic and photopic light levels". In: *Journal of Vision* 20.4, pp. 23–23.

Xu, Jun, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao [2020]. "Noisy-as-Clean: Learning Self-Supervised Denoising From Corrupted Image". In: *IEEE Transactions on Image Processing* 29, pp. 9316–9329.

Xu, Li and Jiaya Jia [2010]. "Two-Phase Kernel Estimation for Robust Motion Deblurring". In: *ECCV 2010*, pp. 157–170.

Xu, Li, Jimmy SJ Ren, Ce Liu, and Jiaya Jia [2014]. "Deep convolutional neural network for image deconvolution". In: *NiPS*, pp. 1790–1798.

Xu, Xiangyu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang [2019]. "Quadratic video interpolation". In: *Advances in Neural Information Processing Systems* 32.

Yan, Qingsen, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang [2020]. "Deep HDR Imaging via A Non-Local Network". In: *IEEE Trans. Image Processing* 29, pp. 4308–4322.

Yang, S., T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang [June 2022]. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1190–1199.

Yu, Zhefei, Houqiang Li, Zhangyang Wang, Zeng Hu, and Chang Wen Chen [2013]. "Multi-Level Video Frame Interpolation: Exploiting the Interaction Among Different Levels". In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.7, pp. 1235–1248.

Yuan, Lu, Jian Sun, Long Quan, and Heung-Yeung Shum [July 2007]. "Image deblurring with blurred/noisy image pairs". In: *ACM Trans. Graph.* 26.3, 1–es. ISSN: 0730-0301.

Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang [2022]. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proc. CVPR*, pp. 5728–5739.

Zeng, Wenjun, Scott Daly, and Shawmin Lei [2002]. "An overview of the visual optimization tools in JPEG 2000". In: *Signal Processing: Image Communication* 17.1, pp. 85–104.

Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang [2017]. "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising". In: *IEEE Trans. Image Processing* 26.7, pp. 3142–3155.

Zhang, Kai, Wangmeng Zuo, and Lei Zhang [2018a]. "FFDNet: Toward a fast and flexible solution for CNN-based image denoising". In: *IEEE Trans. Image Procssing* 27.9, pp. 4608–4622.

Zhang, Kaihao, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li [2022]. "Deep image deblurring: A survey". In: *International Journal of Computer Vision* 130.9, pp. 2103–2130.

Zhang, Lei, Xiaolin Wu, Antoni Buades, and Xin Li [2011a]. "Color demosaicking by local directional interpolation and nonlocal adaptive thresholding". In: *Journal of Electronic Imaging* 20.2, p. 023016.

Zhang, Lin, Lei Zhang, Xuanqin Mou, and David Zhang [2011b]. "FSIM: A Feature Similarity Index for Image Quality Assessment". In: *IEEE Transactions on Image Processing* 20.8, pp. 2378–2386.

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang [June 2018b]. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 586–595.

Zhang, Xinyi, Hang Dong, Zhe Hu, Wei Sheng Lai, Fei Wang, and Ming Hsuan Yang [2019]. "Gated fusion network for joint image deblurring and super-resolution". In: *BMVC*.

Zhang, Youjian, Chaoyue Wang, and Dacheng Tao [2020]. "Video frame interpolation without temporal priors". In: *Advances in Neural Information Processing Systems* 33, pp. 13308–13318.

Zhong, Zhihang, Ye Gao, Yinqiang Zheng, and Bo Zheng [2020]. "Efficient Spatio-Temporal Recurrent Neural Network for Video Deblurring". In: *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, pp. 191–207.

Zhou, S., J. Zhang, J. Pan, W. Zuo, H. Xie, and J. Ren [Nov. 2019]. "Spatio-Temporal Filter Adaptive Network for Video Deblurring". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 2482–2491.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros [2017]. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *ICCV*, pp. 2223–32.

Zoran, D. and Y. Weiss [2011]. “From learning models of natural image patches to whole image restoration”. In: *ICCV*, pp. 479–486.