

UNIVERSITY OF
CAMBRIDGE

Improving Photometric Camera Accuracy and Image Quality in High Dynamic Range Imaging

Param Hanji



Girton College

This dissertation is submitted on July, 2023 for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Param Hanji
July, 2023

Abstract

Improving Photometric Camera Accuracy and Image Quality in High Dynamic Range Imaging

Param Hanji

Cameras have long grappled with the challenge of capturing the vast range of light intensities present in the real world and reproducing them on a medium with much lower dynamic range. This challenge persisted from the era of print photography and continued with the adoption of standard dynamic range (SDR) displays, such as CRT and LCD displays. However, the emergence of high dynamic range (HDR) displays, such as OLED and dual-modulation LCD-LED, has enabled more accurate reproduction of a wider range of light intensities.

To fully leverage the capabilities of these HDR displays, it is essential to capture images that faithfully represent the range of light intensities in the real world. In this dissertation, I propose improvements to the HDR capture and reconstruction pipeline by developing noise-aware statistical estimation methods that use physically accurate parametric models to approximate the noise distribution of pixels. These methods address inaccuracies in existing merging algorithms and do not rely on camera-specific parameters that can be difficult to calibrate accurately. Accurate HDR reconstruction also requires accounting for small inaccuracies in the reported camera metadata caused by changes in lighting or rounding by the camera firmware.

The proposed HDR reconstruction pipeline has the potential to enhance existing HDR datasets. To demonstrate its effectiveness, I evaluate the improved pipeline using two datasets designed for different applications. The reconstruction pipeline also plays a crucial role in an ultra-realistic capture-display-render system built to faithfully reproduce our 3D world. The psychophysical experiments conducted on this prototype display, which showed that the reconstructed 3D scenes are indistinguishable from their real counterparts, serve as the most comprehensive validation of the HDR pipeline, as they focus on reconstructing real-world scenes in a physically accurate manner.

Additionally, I investigate the effectiveness of learning-based methods on the challenging task of predicting HDR information from a single 8-bit photograph. This problem

requires the hallucination of underexposed and overexposed pixels, making it more difficult than merging an exposure stack. Through a controlled subjective experiment, I found that state-of-the-art deep convolutional networks are only partially successful at reconstructing accurate HDR pixels. I highlight the limitations of existing evaluation protocols, which explain the poor correlations between widely-used HDR image quality metrics and subjective data. Building upon these findings, I propose a simple correction that significantly improves the predictions of quality metrics for this task.

One observation from the study is that the single-image HDR problem is treated as a one-to-one mapping by most existing works. In response to this, I present an alternative approach that utilises a generative network to model the one-to-many mapping for global operations. This approach offers several advantages, including the ability to generate multiple plausible HDR images from a single input photograph, thereby enhancing the accuracy of HDR reconstruction.

By tackling the challenges associated with capturing and reconstructing HDR information from still scenes, as well as predicting HDR information from single images, this research has the potential to make significant advancements in the field of HDR imaging. Through these advancements, we can take a significant step forward in bridging the gap between the dynamic ranges observed in the real world and those present in our digital content.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr Rafał K. Mantiuk, for his guidance, patience, and support throughout this journey. His expertise and dedication have been invaluable in shaping my research and helping me navigate the challenges of my PhD. Rafał's unwavering belief in my abilities and his willingness to push me beyond what I thought was possible has been instrumental in my growth as a researcher.

I would like to acknowledge the funding support from the European Research Council (ERC), which has made this research possible.

I am deeply grateful for the guidance and mentorship provided to me by my second supervisor, Dr. Cengiz Öztireli, throughout my academic journey. Cengiz's deep knowledge and expertise in his field have been a constant source of inspiration for me. I would also like to thank my assessors Prof. Peter Robinson, and Prof. Alan Blackwell, for their insightful feedback and constructive criticism, which have greatly improved the quality of this work.

Next, I am grateful to my colleagues in the Rainbow Group at Cambridge, who have provided a stimulating academic environment and offered valuable insights and suggestions. In particular, I would like to thank Fangcheng Zhong, Akshay Jindal, Aamir Mustafa, Aliaksei Mikhailiuk, Dingcheng Yue, Dan Andrei Iliescu, Gyorgy Denes as well as senior researchers in our group Minjung Kim, Maryam Azimi, and Özgür Yöntem. I will forever cherish our insightful discussions, late dinners, and football games. I also extend my gratitude to my external collaborators, Prof. Jonas Unger, Gabriel Eilertsen, Saghi Hajisharif at Linköping.

I owe a debt of gratitude to my family and friends, who have provided immense support and encouragement throughout my journey. Last but not least, I would like to express my deepest appreciation to my partner, Simii, whose love and support have been instrumental in my success. Her belief in me and my abilities has been an endless source of motivation, and her unwavering presence made the long nights of research feel less daunting.

Contents

1 Introduction **15**

1.1 Motivation . . . 15
1.2 Contributions and impact . . . 16
1.3 Outline of the dissertation . . . 17
1.4 Publications . . . 17
1.4.1 Code and datasets . . . 18

2 Background **19**

2.1 Image formation pipeline . . . 19
2.1.1 Sensor model . . . 19
2.1.2 Sources of noise . . . 20
2.1.3 RAW linear images . . . 21
2.1.4 ISP pipeline . . . 22
2.2 HDR estimation . . . 23
2.2.1 Static scenes . . . 23
2.2.2 Dynamic scenes . . . 25
2.2.3 Single-image HDR . . . 25
2.3 Maximum likelihood estimation . . . 26
2.3.1 MLE for known distributions . . . 26
2.3.2 Modelling unknown distributions . . . 27
2.3.3 MLE for regression . . . 28
2.4 Summary . . . 30

3 Practical camera noise modelling **31**

3.1 Noise models . . . 31
3.1.1 Per-gain and per-colour parameters . . . 32
3.1.2 Unified parametric model . . . 32
3.1.3 Black-box models . . . 32
3.2 Camera noise calibration . . . 33
3.3 Validation . . . 34

3.4 Additional cameras . . . 37
3.5 Summary . . . 37

4 Noise-aware HDR merging for still scenes 39
4.1 HDR radiance estimation . . . 39
4.1.1 Simple or heuristic estimators . . . 40
4.1.2 Estimators based on MLE . . . 40
4.1.3 Estimators and noise parameters . . . 43
4.2 Experiments with synthetic data . . . 43
4.2.1 Comparison of Estimators . . . 43
4.2.2 Gain modulation . . . 45
4.2.3 Robustness to noise . . . 45
4.3 Validation on Real Images . . . 48
4.3.1 Large HDR dataset . . . 48
4.3.2 Qualitative comparison . . . 49
4.3.3 Quantitative comparison . . . 51
4.4 Summary . . . 51

5 Automatic exposure compensation 53
5.1 Motivation . . . 53
5.2 Related work . . . 56
5.3 Banding due to inaccurate exposure . . . 57
5.4 Exposure estimation in real images . . . 57
5.4.1 Homoskedastic pixels . . . 59
5.4.2 Heteroskedastic pixels . . . 60
5.4.3 Camera noise calibration . . . 61
5.4.4 Reducing the linear system . . . 61
5.4.4.1 Spatial balance: Tiling . . . 62
5.4.4.2 Exposures balance: Spanning trees . . . 62
5.4.5 Handling outliers . . . 64
5.5 Results and applications . . . 64
5.5.1 Validation on a synthetic dataset . . . 65
5.5.2 Performance on real captures . . . 67
5.5.3 Deghosting for scenes with motion . . . 68
5.5.4 Measuring display MTF . . . 70
5.6 Summary . . . 71

6 Application: Perceptual realism 73
6.1 Introduction . . . 73

6.2 Apparatus overview . . . 74
6.2.1 HDR displays . . . 74
6.2.2 Focal planes and optics . . . 76
6.2.3 Real-scene box . . . 76
6.2.4 Data camera for light field capture . . . 77
6.3 HDR-MF-S imaging & rendering system . . . 77
6.3.1 HDR light field capture . . . 78
6.4 Results . . . 79
6.5 Summary . . . 80

7 Application: Dataset for computer vision 83
7.1 Related work . . . 84
7.2 Capture setup . . . 85
7.2.1 Lights . . . 86
7.2.2 Camera . . . 88
7.2.3 Motorized camera slider . . . 88
7.3 Scenes . . . 88
7.3.1 Faces . . . 89
7.3.2 Tunnel . . . 90
7.3.3 Street . . . 90
7.4 Processing of the frames . . . 90
7.4.1 Distribution of pixel values . . . 91
7.5 Performance of computer vision methods . . . 91
7.5.1 Effect of changing illumination . . . 93
7.5.2 Comparison of TFs . . . 95
7.6 Limitations . . . 97
7.7 Summary . . . 100

8 High dynamic range from a single photograph - Evaluation 101
8.1 Problems involved . . . 101
8.1.1 Inverse tone-mapping . . . 103
8.1.2 Single-image HDR . . . 103
8.1.3 The evaluation problem . . . 104
8.2 Assessing the quality of SI-HDR reconstructions . . . 104
8.3 Related work . . . 105
8.4 SI-HDR dataset . . . 106
8.4.1 Multi-exposure stacks . . . 106
8.4.2 Camera simulation . . . 106
8.4.3 SI-HDR methods . . . 107

8.5 CRF correction . . . 107
8.6 Subjective evaluation . . . 108
8.6.1 Details of the experiment . . . 109
8.6.2 Results: SI-HDR benchmark . . . 110
8.7 Quality metrics for SI-HDR . . . 111
8.7.1 Existing metrics . . . 111
8.7.2 Existing validation protocols . . . 113
8.7.3 Improved protocol for SI-HDR evaluation . . . 115
8.7.4 Validation of quality metrics . . . 118
8.7.5 Minimum measurable increment of quality . . . 119
8.8 Summary . . . 120

9 Modelling style for inverse tone-mapping 121
9.1 Methodology . . . 121
9.1.1 Approach . . . 122
9.1.2 The latent distribution model . . . 123
9.1.3 Network architecture . . . 123
9.1.4 Optimization and inference . . . 124
9.2 Results . . . 126
9.2.1 Dataset . . . 126
9.2.2 Other dimensionality reduction techniques . . . 126
9.2.3 Quantitative comparison . . . 127
9.3 Summary . . . 128

10 Conclusions 129
10.1 Key contributions . . . 129
10.2 Future work . . . 130

Bibliography 131

A Proof of MVUE 149

B Moments of a transformed random variable 151

Glossary

ADC analog-to-digital converter.

BTF brightness transfer function.

CDF cumulative distribution function.

CFA colour filter array.

CNN convolutional neural networks.

CRF camera response function.

CRLB Cramér-Rao lower bound.

CRT cathode ray tube.

CV computer vision.

EE endpoint error.

EM expectation maximisation.

EXIF exchangeable image file format.

GAN generative adversarial network.

HDR high dynamic range.

HDR-MF-S high-dynamic-range multi-focal stereo.

IID independent and identically distributed.

INN invertible neural network.

IQA image quality assessment.

ISP image signal processor.

iTM inverse tone-mapping.

JOD just objectionable difference.

LCD liquid crystal display.

MAP maximum a posteriori.

MC Monte Carlo.

mIOU mean intersection over union.

MLE maximum likelihood estimation.

MLP multi-layer perceptron.

MST minimum spanning tree.

MTF modulation transfer function.

MVUE minimum-variance unbiased estimator.

NLL negative log-likelihood.

OLED organic light emitting diode.

OLS ordinary least squares.

PCA principal component analysis.

PDF probability density function.

SDR standard dynamic range.

SI-HDR single-image HDR.

TF transfer function.

VAE variational autoencoder.

WLS weighted least squares.

Chapter 1

Introduction

1.1 Motivation

Recent surveys have shown that people are spending an increased amount of time looking at photographs and videos on digital screens [118, 164]. However, there are significant differences between the visual experience on digital screens and in the natural world. Factors such as the lack of 3D information, limited resolution, and low dynamic range can greatly affect the overall visual experience. In this dissertation, my focus is on addressing one of these key factors, namely the low dynamic range, by demonstrating how existing imaging systems can be exploited to capture high dynamic range (HDR) information. By leveraging the capabilities of current imaging technologies, we can improve the quality of digital images and videos to provide a more immersive and realistic visual experience.

To fully appreciate the challenge at hand, it is important to first understand the vast dynamic range that our eyes are capable of perceiving. Our visual system has evolved to detect and process an incredible range of light intensities, from the dim light of a starry night ($0.001\,\mathrm{cd/m^2}$) to the brightness of the midday sun (over $10{,}000\,\mathrm{cd/m^2}$), as depicted in Figure 1.1. This is accomplished through the remarkable ability of our visual system to adjust to varying levels of ambient light, allowing us to see clearly and accurately across an astonishing dynamic range of 10,000,000:1.

Individual scenes rarely exhibit such a massive dynamic range because starlight or

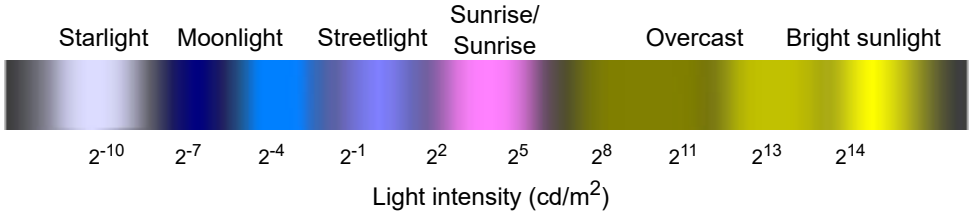


Figure 1.1: Range of light intensities of the natural world.

moonlight is completely masked by the Sun during the day. However, high dynamic range (HDR) scenes, with dynamic ranges of 100,000:1 are not uncommon. Day-to-day examples exhibiting such a dynamic range include indoor scenes with windows exposed to bright sunlight, or nighttime scenes with blindingly bright lights. If the goal of our imaging and display systems is to reproduce the natural world, this is the relevant dynamic range.

In contrast, the dynamic range of a typical modern camera is limited, ranging from 10 to 14 stops (where each stop doubles the dynamic range). This translates to a dynamic range ratio of between 1 : 1024 for lower-end mobile phone cameras to 1 : 16384 for high-end DSLRs. As a result, capturing scenes with a high dynamic range using these cameras often leads to noisy or underexposed pixels in dark regions and saturated pixels in bright regions. This can greatly diminish the overall quality and realism of the resulting images.

The primary reason for this reduced camera dynamic range is that the captured images and videos were typically displayed on media with a limited dynamic range. For example, glossy printed material spans only 6 stops while CRT and older LCD displays span about 8 stops. Modern-day OLED displays, on the other hand, have a much higher dynamic range and can show bright highlights right next to perfect blacks. To cater to modern displays and more closely match the vast dynamic range of the natural world, there is thus a critical need to improve the dynamic range of digital imaging systems, such as cameras. By doing so, we can ensure that digital images and videos more accurately reflect the visual richness and complexity of the real world, providing a more immersive and realistic viewing experience.

1.2 Contributions and impact

In this dissertation, I seek to accurately capture the full dynamic range of real-world scenes, which is crucial for faithfully reproducing HDR scenes on future display systems. Rather than relying on custom modified lens systems [84, 153] or expensive event-based cameras [45], I focus on *how to accurately capture HDR scenes with existing imaging systems*. To achieve this, I utilise statistical methods based on maximum likelihood estimation (MLE) to estimate the scene radiance directly (Chapters 4 and 9), or to estimate other parameters relevant to HDR merging (Chapter 5). My work is thus directly applicable to regular images captured by today's cameras. Furthermore, the algorithms are motivated by a well-known camera noise model but they do not rely on calibration-sensitive parameters. These computational methods are particularly beneficial for low-quality sensors with limited dynamic ranges, such as those found in mobile phones and computer vision cameras.

I also compare single-image HDR (SI-HDR) methods to convert existing SDR digital

media to HDR and propose an improved evaluation protocol to drive future research. Finally, I demonstrate how to generate HDR videos from legacy SDR content by matching the style of skilled colour artists.

1.3 Outline of the dissertation

I start by discussing the HDR imaging pipeline Chapter 2, paying close attention to noise added at various stages. Since I cast several challenges to classical statistical estimation problems in later chapters, the background also includes some MLE-based estimators. In Chapter 3, I discuss parametric noise models and describe how to calibrate commercial cameras.

Then, in Chapters 4 and 5, I show how to exploit the noise model to produce physically accurate HDR reconstructions by merging a stack of images. The former provides an estimator for scene radiance, while the latter estimates capture parameters to align exposure stacks. Both methods share the advantage of being motivated by a physically-accurate noise model without being sensitive to camera-specific calibration. The improved reconstructions have real-world applications, as evidenced in Chapters 6 and 7. Finally, I turn to the challenging problem of style-preserving inverse tone-mapping in Chapter 9 after benchmark state-of-the-art SI-HDR methods in Chapter 8.

1.4 Publications

The following is the chronological list of publications that are included this dissertation:

- **Param Hanji**, Fangcheng Zhong, and Rafał K. Mantiuk. "Noise-aware merging of high dynamic range image stacks without camera calibration". In Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2020.
  Contribution: Derived noise-based estimators and analysed them in simulation and on real image stacks.

- **Param Hanji**, Muhammad Z. Alam, Nicola Giuliani, Hu Chen and Rafał K. Mantiuk. "HDR4CV: High dynamic range dataset with adversarial illumination for testing computer vision methods". In *Journal of Imaging Science and Technology.* 2021. Presented at *London Imaging Meeting (LIM), 2021*
  Contribution: Setup semi-realistic scenes, created the dataset and ran the preliminary study comparing encoding functions for different algorithms.

- Gabriel Eilertsen, Saghi Hajisharif, **Param Hanji**, Apostolia Tsirikoglou, Rafał K. Mantiuk, and Jonas Unger. "How to cheat with metrics in single-image HDR

reconstruction". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021.
Contribution: Created the dataset to test various methods.

- Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, **Param Hanji**, Simon J. Watt, and Rafał K. Mantiuk. "Reproducing Reality with a High-Dynamic-Range Multi-Focal Stereo Display". In *ACM Transactions on Graphics, 40(6)*. 2021.
Contribution: Responsible for the capture pipeline including controlling the camera slider, merging image stacks and initial image processing operations.

- **Param Hanji**, Rafał K. Mantiuk., Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. "Comparison of single image HDR reconstruction methods — the caveats of quality assessment". In *ACM SIGGRAPH Conference Proceedings*. 2022.
Contribution: Collected subjective scores, compared correlations with a large number of image quality metrics and analysed the data.

- Aamir Mustafa, **Param Hanji**, and Rafał K. Mantiuk. "Distilling style for image pairs for forward and inverse tone mapping". Best paper in *ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*. 2022.
Contribution: Implemented the VAE baseline and applications relevant to inverse tone mapping.

- **Param Hanji**, and Rafał K. Mantiuk. "Robust estimation of exposure ratios in multi-exposure image stacks". In *IEEE Transactions on Computational Imaging*. 2023.
Contributions: Derived and implemented the noise-aware linear system and compared it with the existing method.

1.4.1 Code and datasets

Along with the published papers, some code and datasets have been made publicly available:

- HDRutils: A collection of utility functions to merge HDR images stacks, correct differences in exposure ratios and simulate real camera noise.
Link: `https://github.com/gfxdisp/HDRutils`

- SI-HDR dataset: Consists 181 HDR scenes with ground truth reconstructions and synthetic exposure stacks.
Link: `https://www.repository.cam.ac.uk/handle/1810/340049`

- HDR4CV dataset: HDR video dataset consisting of 4 sequences (100 frames each) to evaluate the performance of computer vision methods under challenging illuminations.
Link: `https://www.repository.cam.ac.uk/handle/1810/324068`

Chapter 2

Background

This dissertation identifies limitations and problems with existing high dynamic range (HDR) imaging pipelines and addresses them with rigorous statistical methods. I devote the first section of the background to explaining the imaging pipeline of modern cameras, highlighting the probabilistic nature of noise in images. This is followed by a discussion of related work relevant to HDR imaging. Finally, I introduce some statistical tools used in later chapters, along with relevant probabilistic motivations.

2.1 Image formation pipeline

I first describe the camera capture process, starting with photons incident on the imaging sensor, and describe how the generated electrons are processed. This segues into a discussion about the nature of noise in images and helps identify its potential sources. Then, I briefly discuss the multi-stage image signal processor (ISP) pipeline for producing 8-bit images present in most cameras.

2.1.1 Sensor model

When a camera captures an image, photons from the scene enter the camera lens and pass through a colour filter array (CFA) before reaching the imaging sensor, which could be a charge-coupled device (CCD) or complementary metal-oxide semiconductor (CMOS) sensor, as shown in Figure 2.1. The *aperture* of the lens determines the amount of light that enters the camera system. When exposed for a fixed interval, depending on the camera's *exposure time* or *shutter speed*, the imaging sensor converts some of the incident photons into electrons. The number of electrons generated is proportional to the number of incident photons. These electrons accumulate to produce a voltage, which is then processed by analogue electronics. The voltage is amplified, with the amount of amplification controlled by the camera's *gain*. Finally, an analog-to-digital converter (ADC) converts the signal

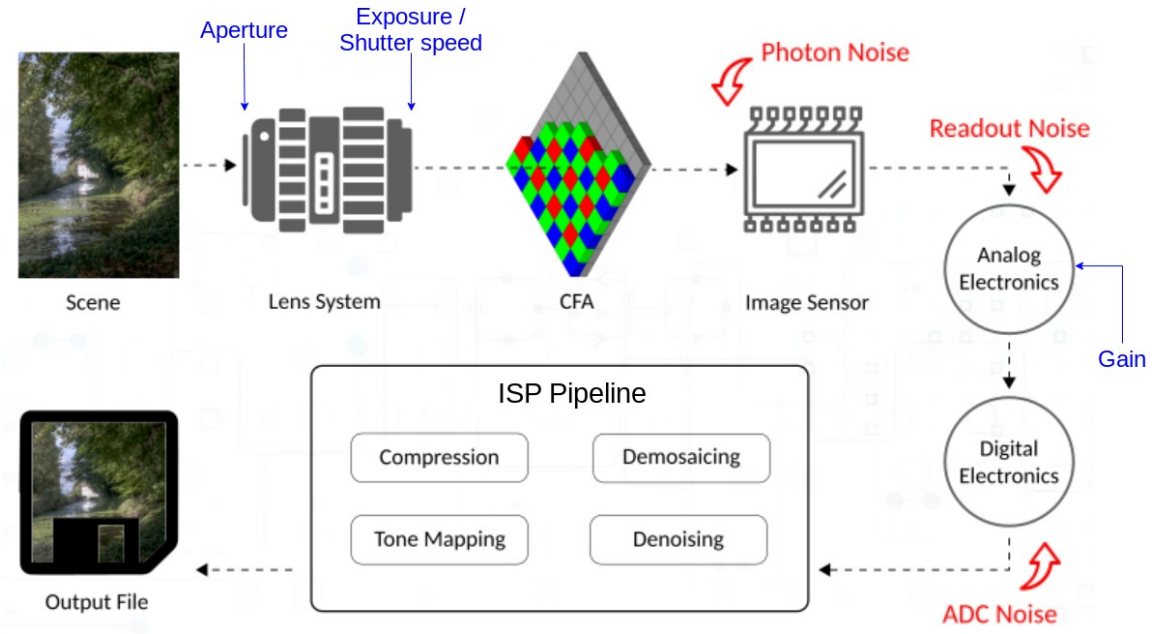


Figure 2.1: The process of converting photons from a scene into digital images is known as the image formation pipeline, which involves multiple stages. The lens and camera settings such as aperture, shutter speed and gain (shown in blue) control the RAW pixel values, indicated in blue. However, various stages introduce noise (shown in red) that can cause variations in the RAW pixel values. Eventually, the RAW image data is processed by the image signal processor (ISP) to produce a usable image.

into discrete pixel intensities. The resulting image is in a RAW format and has a linear relationship with the scene radiance, which is an essential prerequisite for the methods introduced in later chapters. However, since the image is a measurement of the scene by an imperfect device (the camera), it contains noise from various sources.

2.1.2 Sources of noise

Detailed studies [43, 54, 83] indicate that camera noise consists of signal-dependent and signal-independent (or static) components. The various sources of noise are highlighted in red in Figure 2.1.

The process of photon registration is affected by the inherent randomness of photons arriving at the sensor which is accurately modelled by a Poisson distribution [59]. This leads to the first component of camera noise, photon shot noise, which is signal-dependent. Because the signal is amplified by some *gain** before reaching the ADC, the signal-independent noise is typically split into pre-amplifier and post-amplifier components [58]. Readout noise captures the voltage fluctuations while accumulating electrons and is amplified along with the signal. The last component, analogue-to-digital noise, is added after amplification and is attributed to the quantisation errors. Previous studies [43, 83]

*This is determined by ISO set by the user

show that these signal-independent components are normally-distributed.

Digital sensors also exhibit fixed-pattern noise due to photo-response and dark-current non-uniformity [3]. These sources of noise, however, are easy to compensate for as they are fixed for every sensor and often removed by camera firmware from RAW images. Fixed-pattern noise was not present in the images captured by cameras I tested; hence, I do not model it when accounting for noise in Chapter 4 and Chapter 5. Other random sources of noise, such as temperature-dependent dark-current shot noise [24], are accommodated in the signal-independent components.

Let the image be taken with an exposure time t, aperture a and gain g. Further, let $Y(p)^{\dagger}$ be a random variable representing the final recorded value of the unknown scene radiance $\phi(p)$ at pixel p. The different sources of noise, depicted in Figure 2.1, motivate the decomposition of this random variable into a sum of three independent random variables. The first random variable is sampled from a Poisson distribution with a parameter equal to the number of incoming photons; the other two are sampled from zero-mean normal distributions. The first normally distributed component accounts for readout noise and has a standard deviation equal to σ_{read} and the second component, parameterised by σ_{adc}, captures amplifier and quantisation noise. Assuming that the pixel is not saturated,

$$Y(p) \sim \underbrace{\text{Pois}(\phi(p)\, t\, a)\, g\, k_c}_{\text{photon noise}} + \underbrace{\mathcal{N}(0, \sigma_{\text{read}})\, g\, k_c}_{\text{readout noise}} + \underbrace{\mathcal{N}(0, \sigma_{\text{adc}})\, k_c}_{\text{ADC noise}} . \tag{2.1}$$

Each colour channel has a different quantum efficiency for photon-to-electron conversion due to differences in the sensitivity of the sensor across the light spectrum. This is accounted for by the colour coefficient k_c where $c \in \{r, g, b\}$. Gain affects the Poisson random variable and the first normal random variable, whereas the k_c is a multiplier on all three terms.

2.1.3 RAW linear images

Modern digital cameras provide access to uncompressed, minimally processed data directly from the electronic imaging sensor in the form of RAW images. These pixels are linearly related to the radiance but are affected by the various sources of noise. When the scene's radiance is close to zero, readout and ADC noise can result in underestimating the true radiance, effectively making some pixel values negative. Camera manufacturers typically add an offset, called black-level, to ensure that RAW pixel values are positive. The first step to restoring linearity is, thus, black-level subtraction.

All the methods described in this dissertation focus on accurately recovering HDR radiance values. Omitting further digital operations and working with RAW values

[†]Throughout this dissertation, capital letters represent random variables and small letters represent individual measurements or other scalars.

is preferable due to the linear relationship. Moreover, accurate noise modelling is a crucial feature of most methods and nonlinear operations of the digital pipeline alter the distribution of noise.

2.1.4 ISP pipeline

The RAW image from the sensor passes through the digital ISP pipeline and includes black-level subtraction, demosaicing, colour correction, denoising, tone-mapping and compression [17]. Most of these stages perform nonlinear operations and alter the original readings significantly. Any attempt to recover the scene radiance should thus omit these operations. I include this overview of the various problems involved for completeness.

Demosaicing RAW images store colour information at reduced resolutions through a process called colour filter array (CFA) subsampling. For instance, in the commonly used Bayer CFA pattern (RGGB), the green channel is stored at half-resolution, while the red and blue channels are stored at quarter-resolution, as depicted in Figure 2.1. In order to recover the complete colour information and generate a 3-channel RGB image, a process called demosaicing or debayering is necessary. This process involves interpolating the missing colour information based on the available data from neighbouring pixels. Since the green channel is better sampled, the most widely used demosaicing methods first recover it through spatial [95, 109] or frequency domain filtering [8, 62]. Demosaicing is a well-studied topic [88] and various methods from simple bilinear interpolation to deep convolutional neural networks (CNN) [147, 149] have been proposed.

Colour Correction Colours in captured images are often incorrect because, unlike the human visual system, a camera lacks the ability to compensate for different illuminations. As a result, images captured under different lighting conditions may have a colour cast, where the overall colour of the image appears too warm or too cool. A simple way to address this issue, is by applying white balance correction, which involves scaling the colour channels to ensure that white objects in the scene appear white in the captured image as well. White balance correction is typically performed immediately before or after demosaicing.

While simple white balance correction effectively eliminates colour casts, it may not suffice to accurately represent the colours of complex scenes. For camera-agnostic rendering, it becomes essential to depict the capture in a standard colour space such as CIE-XYZ. This process of colour correction involves estimating and applying a 3×3 matrix that transforms from the camera-native RGB space into a standard colour space.

Denoising The demosaiced RGB image contains noise as described in Eq. (2.1). Image denoising is the process of removing noise, regardless of its source, and is another well-researched problem. The challenge here is to remove unwanted noise while retaining image details. Most denoising algorithms assume that noise can be approximated by a Gaussian distribution [129, 178] and aim to enforce specific image priors. Convolving with specially designed (and sometimes nonlinear) filters, wavelet-thresholding, [129] and block matching algorithms [25] are a few successful approaches. More recently, synthetic and natural image datasets have resulted in several data-driven methods [69, 70, 171, 172].

Tone-mapping The final major step, tone-mapping, is a technique to efficiently encode pixels in a limited number of bits. The goal of tone-mapping is to compress the dynamic range of the captured scene in a visually plausible manner while taking special care of both highlights and low-luminance regions. There are different approaches to tone-mapping, including inexpensive global operators like gamma and logarithm, as well as local operators that map pixels to an appropriate value by considering their respective spatial neighbourhoods. Several sophisticated tone-mapping operators have been proposed, including adaptive logarithmic compression [30], perceptually motivated methods [125, 132], gradient domain operator [39] and histogram-based display adaptive operator [102].

After the image has undergone all the necessary processing steps, it is usually quantised and compressed to reduce its file size. Finally, after compression using standard JPEG or advanced codecs, the image is saved to the file for storage or transmission.

2.2 HDR estimation

The standard imaging pipeline from the previous section falls short for HDR scenes since the dynamic range of modern cameras is far lesser than the range of light intensities present in the natural world. Strategies for capturing HDR information and subsequent reconstruction algorithms depend on whether camera and object motion needs to be accounted for.

2.2.1 Static scenes

Accurately estimating HDR scene information in the absence of camera and objection motion involves capturing multiple standard dynamic range (SDR) images with changing exposure time [4, 49, 58, 101, 155], or gain [53] or capturing a burst of images with the same parameters [57]. The stack of images is later combined in post- processing as part of the digital pipeline [46].

HDR processing typically begins with RAW images in which pixels are distributed according to Eq. (2.1). Although RAW pixels are linearly related to the scene radiance,

they represent inaccurate measurements due to saturation and noise since the dynamic range of cameras is limited. Before merging (averaging) the exposure stack, we need to convert samples to relative radiance units by compensating Eq. (2.1) for exposure time t, gain g, aperture f-number a, and per-color transmission coefficient k_c:

$$X(p) = \frac{Y(p)}{t\, g\, a\, k_c}, \quad \text{where } a = \pi \left(\frac{f}{2\, a}\right)^2 . \tag{2.2}$$

Here, f is the focal length of the lens. At each pixel p, $x(p)$ is a noisy measurement of the hidden radiance $\phi(p)$, and serves as the realisation of the random variable $X(p)$. The capture parameters t, g and a are typically obtained from exchangeable image file format (EXIF) metadata. However, as I demonstrate in Chapter 5, they may be inaccurate due to various reasons, and estimating the values directly from the images is preferable.

The compensated images from Eq. (2.2) are combined (merged) to produce the required HDR estimate. Early HDR reconstruction methods [6, 27, 101, 114, 136] focused on inverting the CRF. This is because camera manufacturers did not historically provide access to unprocessed and uncompressed RAW images. Most estimators proposed were weighted averages of the linearised pixel values, where the weights were functions of the inverse-CRF or the derivative of the CRF. All these methods do not account for camera noise and therefore provide sub-optimal estimations of HDR pixel values.

The first HDR estimation method that used a noise model was proposed by Tsin et al. [155]. They combined images with weights equal to the ratios of the corresponding exposure times and standard deviations, measured directly from the images. In a later work, Granados et al. [49] showed that Tsin et al.'s method was sub-optimal under a compound-normal noise assumption as pixels near the saturation point were given smaller than ideal weights despite having the highest signal-to-noise ratio (SNR). Debevec and Malik's hat-shaped weighting function [27] also suffers from this limitation of under-weighting pixels close to the saturation point. Kirk and Andersen [82] proposed a maximum likelihood estimation (MLE)-based weighting scheme using a simple noise model. This was later extended to a more complete model incorporating noise from several sources by Hasinoff et al. [58]. Granados et al. [49] noted that the true MLE-based estimator does not have an analytical solution and used the expectation maximisation (EM) algorithm for a more accurate estimation than other similar works.

Aguerrebere et al. [4] compared the previously mentioned HDR reconstruction methods and analysed how far each of their variances was from the theoretical Cramér-Rao lower bound (CRLB). They concluded that the variance of MLE-based estimators, such as that of Granados et al. [49], were close to the CRLB but the estimation could be easily affected by errors in noise parameter calibration.

2.2.2 Dynamic scenes

When objects are not stationary, HDR scenes can be captured using modified imaging systems [84, 153]. However, multiple sensors with beam-splitters make these systems bulky and prohibitively expensive for widespread use. Event cameras with computational pipelines [130, 161] offer another way to capture rapidly changing HDR content. Event cameras, however, remain a work in progress and too expensive for mass adoption. In practice, most imaging systems reconstruct the HDR content from a stack of images that may contain misaligned pixels.

One approach to achieve this is to detect and ignore misaligned pixels. Jacobs et al. [68] attempted to identify such pixels by monitoring their variances across the image stack. Granados et al. [50] employed a noise model similar to Eq. (2.1) by dynamically computing the noise parameters for every scene. Their calibrated noise model identified outliers due to motion.

Another set of methods registers pixels before merging to prevent loss of information. Homography alignment [154] and optical flow [72] are only moderately successful for this task. Sen et al. [143] treat alignment and merging as a joint optimisation problem at the patch level. On the contrary, Hu et al. [65] treat the issues separately and first reconstruct every image in the stack by finding dense correspondences at the patch level.

Recently, deep approaches have successfully been adapted for aligning HDR stacks. Kalantari and Ramamoorthi [71] aligned input images with classical optical flow [90] and trained a network to eliminate warping artefacts. Following the success of end-to-end trained networks, Wu et al. [165] removed explicit optical flow from the pipeline and let the network perform pixel alignment. Finally, a few works relied on attention maps [22, 167] to further improve results.

2.2.3 Single-image HDR

While HDR deghosting methods consider the entire stack, another approach relies on predicting HDR content from a single standard dynamic range (SDR) image. single-image HDR (SI-HDR) (also referred to as inverse tone-mapping (iTM)) methods were first introduced to display conventional images on HDR displays [13, 14, 110, 134]. Estimating HDR pixel values from a single image is an under-constrained problem due to missing information in under and over-exposed regions. Recently, deep neural networks have demonstrated significant improvements based on a learned high-level understanding of the image content. As a result, a large number of methods have been proposed, each incorporating a different architecture and training strategy [33, 35, 76, 86, 87, 92, 107, 139, 170]

From a the single input, Endo et al. [35] generate a stack of SDR images corresponding

to different exposures with an autoencoder. In an attempt to model the ill-posed nature of the problem, Lee et al. [86] train a conditional generative adversarial network (GAN) for the same task. While their approach is sound, GANs are not the most suitable latent distribution model because they are unable to capture the output distribution as well.

In contrast to generating a stack of images, a few methods directly predict the HDR image. Eilertsen et al. [33] focus on the recovery of saturated pixels with a UNet [137]. Marnerides et al. [107] propose another CNN that separately handles global and local image features. Liu et al. [92] separate the various sub-problems by modelling them with three specialised CNNs. Finally, Santos et al. [139] introduce feature masking that reduces the contribution of features extracted from saturated regions to improve end-to-end HDR predictions.

2.3 Maximum likelihood estimation

At its core, determining unknown quantities such as scene radiance values (Chapters 4 and 9) or capture parameters (Chapter 5) from noisy measurements is a statistical estimation problem. In this last section of the background, I introduce maximum likelihood estimation (MLE), a probabilistic approach designed for solving such problems. The idea is to first obtain an expression for the joint probability of the observed data, the *likelihood*. Then, we can maximise this expression either with closed form solutions or iterative optimization such as gradient descent.

If the distribution associated with the forward operation is known, MLE is rather straightforward. For unknown distributions, we can still rely on MLE by fitting a parametric invertible model between the target distribution and a known latent distribution [29, 135].

2.3.1 MLE for known distributions

Let $\boldsymbol{z} = \{z_1, z_2, ..., z_k\}$ be k independent and identically distributed (IID) samples of a known distribution with parameters $\boldsymbol{\theta}$. A distribution is *known* if its probability density function (PDF), $p(z_i \mid \boldsymbol{\theta})$ is known. The likelihood of observing the data is defined as the joint probability of the data. Since the samples are independent, the joint is simply the product of individual probabilities:

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{z}) = p(z_1, z_2, ..., z_k \mid \boldsymbol{\theta}) = \prod_{i=1}^{k} p(z_i \mid \boldsymbol{\theta})\,. \tag{2.3}$$

MLE aims to find the parameters $\hat{\boldsymbol{\theta}}$ that maximise the $\mathcal{L}$. In practice, it is easier to maximise the log-likelihood. Since the log transform is monotonic, this does not change

the result:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{z}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{k} \log p(z_i \mid \boldsymbol{\theta}) \,. \tag{2.4}$$

For example, if each z_i is drawn from a normal distribution with parameters $\boldsymbol{\theta} = \mu, \sigma$, the likelihood is:

$$\log \mathcal{L} = \sum_{i=1}^{k} -\log \sqrt{2\pi\sigma^2} - \frac{(z_i - \mu)^2}{2\sigma^2} \,. \tag{2.5}$$

Differentiating w.r.t. the first parameter (mean) and equating to 0,

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \mu} &= \sum_{i=1}^{k} \frac{z_i - \mu}{\sigma^2} = 0 \\ \implies \hat{\mu} &= \frac{1}{k} \sum_{i-1}^{k} z_i \end{aligned} \tag{2.6}$$

MLE thus states that the average of all available samples is a good estimate of the hidden parameter μ.

A few works discussed in Section 2.2.1 approximate the Poisson component of Eq. (2.1) by a normal distribution. In Chapter 4, I summarise this as well as other approximations and derive MLE-based estimators for the scene radiance under different assumptions.

2.3.2 Modelling unknown distributions

Direct MLE is only possible for known distributions[‡]. The simplest examples of known distributions are those discussed in Section 2.3.1 where the PDF is a straightforward function of a handful of parameters to be estimated (the normal, Poisson, exponential to name a few). For example, the one dimensional normal distribution has two parameters — mean and variance.

Change of variables A distribution is also said to be known if it is linked to another known distribution via an invertible function. Let Z be a random variable corresponding to the known distribution from Section 2.3.1 with parameters $\boldsymbol{\theta}$. Let $Y = f(Z)$ be a new random variable related through the invertible function f. Since f is monotonic, the cumulative distribution function (CDF) of Y is:

$$\begin{aligned} p_Y(Y \leq y) &= p_Z(f(Z) \leq y) \\ &= p_Z(Z \leq f^{-1}(y)) \,. \end{aligned} \tag{2.7}$$

[‡]A distribution is known if we can derive an exact expression for its probability density function.

Differentiating both sides,

$$p_Y(y) = p_Z(f^{-1}(y)) \underbrace{\left|\frac{\mathrm{d}f^{-1}(y)}{\mathrm{d}y}\right|}_{\text{volume correction}} . \tag{2.8}$$

The above equation is known as the change of variables formula and is based on the conservation of probability mass. Notice that it represents an exact expression for the PDF of Y. Eq. (2.8) has a straightforward extension to higher dimensions by replacing the *volume correction* term with the determinant of a Jacobian,

$$p_Y(\boldsymbol{y}) = p_Z(f^{-1}(\boldsymbol{y})) \left|\det\left(\frac{\partial f^{-1}(\boldsymbol{y})}{\partial \boldsymbol{y}}\right)\right| . \tag{2.9}$$

This simple equation forms the crux of a powerful class of deep generative models called normalizing flows [29, 135]. I adapt conditional normalizing flows to solve the inverse HDR problem in Chapter 9.

2.3.3 MLE for regression

Finally, I show that the error function for estimating unknowns of an overconstrained linear system can also be derived using MLE [16, Chapter 3]. The widely used least-squares solution minimises the sum of squared errors for the system $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$[§], where $\boldsymbol{x}$ represents the unknowns to be estimated. To provide a probabilistic interpretation of least-squares, let's define the forward model for the system as,

$$b_i = \boldsymbol{a_i}\boldsymbol{x} + \epsilon_i \tag{2.10}$$

where $\boldsymbol{a_i}$ is the i^{th} row of $\boldsymbol{A}$, b_i is the i^{th} measurement, and ϵ_i is the error associated with the measurement. Different assumptions about the error provide different MLE estimators.

Ordinary least squares If the various measurements are homoskedastic (equally reliable), all the errors are IID samples of the same distribution. Additionally, let's assume this underlying distribution is a zero-mean normal, i.e, each ϵ_i is drawn from $\mathcal{N}(0, \sigma)\ \forall\, i$, where σ is unknown but constant for all measurements. The log-likelihood of observing the data is:

$$\log \mathcal{L}_{\text{homo}} = \sum_i -\log\sqrt{2\pi\sigma^2} - \frac{(b_i - \boldsymbol{a_i}\boldsymbol{x})^2}{2\sigma^2} . \tag{2.11}$$

[§]Bolded capital letters are reserved for matrices and bolded small letters for vectors.

The best estimates of the unknowns are values that maximise Eq. (2.11),

$$\begin{aligned}\hat{\boldsymbol{x}}_{\text{OLS}} &= \arg\max_{\boldsymbol{x}} \log \mathcal{L}_{\text{homo}} \\ &= \arg\min_{\boldsymbol{x}} \sum_i (b_i - \boldsymbol{a_i x})^2 \\ &= \arg\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\,\boldsymbol{x}\|_2^2 \,.\end{aligned} \tag{2.12}$$

The final expression thus indicates that the MLE solution for ordinary least squares (OLS) is equivalent to minimising the sum of squared residual errors. Note that the maximisation changed to a minimisation due to the negative sign in Eq. (2.11). The OLS solution that minimises the residual errors, obtained by differentiating and equating to 0, is:

$$\hat{\boldsymbol{x}}_{\text{OLS}} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{b}\,. \tag{2.13}$$

Weighted least squares To model heteroskedastic linear systems where different measurements are not equally reliable, let each ϵ_i from Eq. (2.10) be drawn from a different normal distribution $\mathcal{N}(0, \sigma_i)\ \forall i$. The log-likelihood changes from Eq. (2.11) to

$$\log \mathcal{L}_{\text{hetero}} = \sum_i -\log\sqrt{2\pi\sigma_i^2} - \frac{(b_i - \boldsymbol{a_i x})^2}{2\sigma_i^2}\,. \tag{2.14}$$

Maximising the expression as before,

$$\hat{\boldsymbol{x}}_{\text{WLS}} = \arg\max_{\boldsymbol{x}} \log \mathcal{L}_{\text{hetero}} = \arg\min_{\boldsymbol{x}} \boldsymbol{W}\,\|\boldsymbol{b} - \boldsymbol{A}\,\boldsymbol{x}\|_2^2\,. \tag{2.15}$$

where $\boldsymbol{W}$ is a diagonal matrix with each diagonal entry given by the inverse-variance of noise, $w_{i,i} = 1/\sigma_i^2$. Now, the weighted least squares (WLS) solution is:

$$\hat{\boldsymbol{x}}_{\text{OLS}} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{W}\,\boldsymbol{b}\,. \tag{2.16}$$

Introducing regularisation with MAP Least squares, both ordinary and weighted, are widely used in many applications, but they suffer from a well-known limitation: sensitivity to outliers. When outliers are present, these methods may produce large errors for some unknowns, which can negatively impact the accuracy of the overall solution. To prevent the occurrence of extreme values, regularisation can be introduced, particularly when an initial estimate $\boldsymbol{x}_0$ is known. This process can also be interpreted probabilistically by treating the unknowns as random variables.

The exact expression to minimise depends on the assumptions we make about the prior distribution of the unknowns. The prior distribution represents the probability distribution of the parameters or unknowns in a model before taking into account any observed data.

If this prior is a normal distribution (with known covariance matrix $\boldsymbol{\Sigma}$), the posterior probability according to Bayes' rule is:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{b}) &\propto \underbrace{p(\boldsymbol{x} \mid \boldsymbol{x}_0, \boldsymbol{\Sigma})}_{\text{prior}} \cdot \underbrace{p(\boldsymbol{b} \mid \boldsymbol{A}, \boldsymbol{x})}_{\text{likelihood}} \\ &= p(\boldsymbol{x} \mid \boldsymbol{x}_0, \boldsymbol{\Sigma}) \cdot \prod p(b_i \mid \boldsymbol{a}_i, \boldsymbol{x}) \,. \end{aligned} \tag{2.17}$$

Maximising the posterior with similar calculus, we have the maximum a posteriori (MAP) estimate:

$$\begin{aligned} \boldsymbol{x}_{\text{MAP}} &= \arg\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \boldsymbol{\Sigma} \|\boldsymbol{x} - \boldsymbol{x}_0\| \\ &= (\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{A}^\mathsf{T}\boldsymbol{b} + \boldsymbol{\Sigma}\,\boldsymbol{x}_0) \,. \end{aligned} \tag{2.18}$$

Incorporating a normal prior in the optimisation is commonly referred to as *ridge-regression*. This is also the most common usage of *Tikhonov regularisation*. The final solutions for exposure estimation (Chapter 5) and camera response function (CRF) correction (Chapter 8) directly use Eq. (2.18).

2.4 Summary

The most reliable way to capture HDR scenes with existing camera technologies is by merging a stack of aligned SDR images. In this chapter, I provided background on the standard imaging pipeline, paying close attention to noise added by analog and digital processes in the first few stages. I reviewed heuristic as well as statistically motivated methods to merge image stacks of still scenes. I also described the two broad approaches to handle dynamic scenes: HDR deghosting and SI-HDR. In the second part of the background, I described the MLE approach to model distributions and estimate unknown parameters from noisy measurements. The best methods for merging stacks rely on per-pixel MLE while considering the noise characteristic of the camera, as I will show in Chapter 4. However, this simple approach is insufficient for SI-HDR where the challenge is to model complex, multidimensional distributions (see Chapter 9).

Chapter 3

Practical camera noise modelling

> "The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic"
>
> *Ronald Fisher*

HDR reconstruction involves estimating the unknown radiance from one or more images. The probabilistic photon registration and electronic processing in the camera will result in some variation or noise in the values recorded in each pixel, as I detailed in Section 2.1.2. Modelling camera noise is thus an important subproblem for accurate reconstruction, as evidenced by many earlier works discussed in Section 2.2.1. A secondary motivation for modelling noise, that is not specific to HDR imaging, is being able to better simulate real cameras. When comparing different methods for image reconstruction problems, it is preferable to use images generated by such physically accurate simulations rather than simpler normally or uniformly distributed noise.

I focus on fitting the parameters of a statistical noise model to real cameras in this chapter. Most practical models are derived from the Poisson-normal distribution given by Eq. (2.1). I discuss two such parametric approximations and a third deep-learning approach that can model more complex noise. Then, I describe a procedure to calibrate cameras with a variety of sensors and show that the fitted parameters simulate real camera noise.

3.1 Noise models

Let me first introduce parametric noise models to predict the variance of pixels in an image.

3.1.1 Per-gain and per-colour parameters

Several existing works [43, 54, 83] recommend fitting separate parameters for each colour channel and each gain setting (specified by ISO). Due to constant gain, the two normally distributed components of Eq. (2.1) can be combined by adding their variances. Thus, the random variable representing pixel values $Y(p)$ is split into a (1) signal-dependent and (2) signal-independent or static component. The variance of the random variable $Y(p)$ is:

$$\mathbb{V}[Y(p)] = \alpha\,\mathbb{E}[Y(p)] + \beta \tag{3.1}$$

where α and β represent noise parameters that need to be estimated. Since these parameters are unique for each gain and each colour channel, a single camera requires a large lookup table (LUT), e.g, the *Nikon D7500* with 38 ISO settings requires $38 \times 3 \times 2 = 228$ parameters.

3.1.2 Unified parametric model

To better summarise the noise distribution of a camera with a minimal number of free parameters, I include gain and colour channel dependencies to the camera model given by Eq. (2.1). The expected value and variance of each $Y(p)$ can be written as:

$$\begin{aligned}
\mathbb{E}[Y(p)] &= \phi(p)\,t\,a\,g\,k_c \\
\mathbb{V}[Y(p)] &= \phi(p)\,t\,a\,g^2\,k_c^2 + \sigma_{\text{read}}^2\,g^2\,k_c^2 + \sigma_{\text{adc}}^2\,k_c^2 \\
&= \underbrace{\mathbb{E}[Y(p)]\,g\,k_c}_{\substack{\text{signal-dependent} \\ \text{noise}}} + \underbrace{\sigma_{\text{read}}^2\,g^2\,k_c^2 + \sigma_{\text{adc}}^2\,k_c^2}_{\text{static noise}}.
\end{aligned} \tag{3.2}$$

Here, $\{\sigma_{\text{read}}, \sigma_{\text{adc}}, k_r, k_g, k_b\}$ are camera-specific noise parameters that require calibration. Notice that, as with Eq. (3.1), the variance can be conveniently represented as a function of the expected value, i.e., $\mathbb{V}[Y(p)] = f(\mathbb{E}[Y(p)])$. I refer to the signal-independent component of the variance as static noise. Static noise is the same for all pixels of an image.

A minor difference between the noise model given by Eq. (3.2) and others found in the literature is that it jointly models all three colour channels, thus reducing the number of free parameters. Other models have a separate set of parameters either for each colour channel [3, 49, 58] or even for each gain setting as shown in the previous section.

3.1.3 Black-box models

More recently, deep generative networks such as normalizing flows [2] and generative adversarial networks [21] have been employed to learn the distribution of noise directly from images. Although these black-box approaches model spatially-varying components of

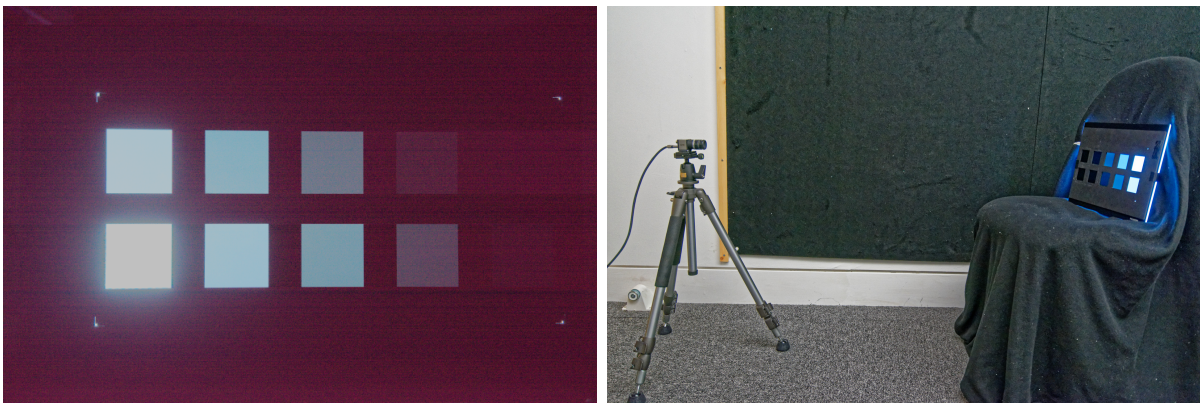


Figure 3.1: The calibration target (left) and capture setup (right) used for camera noise calibration. All pixels within each square region of the target measure the same radiance. The measured values, however, vary slightly due to sensor noise.

real camera noise better, the Poisson-normal statistical model and its normal approximation are better suited for parametric noise modelling as they offer a convenient algebraic form.

3.2 Camera noise calibration

A parametric noise model needs to be fitted for each camera, a process called noise calibration. I show how to calibrate camera-specific parameters for the unified model (Section 3.1.2). This approach was chosen over fitting more complex, deep-learning models because I utilized these noise models to assist in deriving HDR estimators. Derivations in both Chapters 4 and 5 depend on having a straightforward noise model. Therefore, a simple noise model with an explicit formula is more crucial than achieving marginally improved accuracy in reproduction.

Equipment Noise parameters for the unified model (Section 3.1.2) can be estimated using a number of procedures [3, 49, 54]. I relied on simple off-the-shelf components to calibrate four representative cameras with a wide variety of sensor sizes, listed in Table 3.1. This included 2 full-frame sensors (Sony α7r1 and Sony α7r3), one APS-C sensor (Canon T1i) and one mobile phone sensor (Sony IMX345). The calibration target shown in Figure 3.1 was constructed by overlaying a uniform light source (Huawei L4S LED light pad) with neutral density (ND) filters (299 1.2 and 210 0.6 Neutral Density Rosco E-Colour+ lighting filter) of different transmittance values such that each square region emits a different radiance. The variance of pixel values within each captured square provides an empirical measure of noise for a specific value of radiance.

Measurement Several extracted data points (RAW expected value vs. standard deviation) are plotted using crosses in Figure 3.2. The mean of a large number of pixels

Table 3.1: Noise parameters fitted for the tested cameras

Camera	Sensor Size (mm)	Pitch (μm)	Color Coefficients			σ_{read}	σ_{adc}
			k_r	k_g	k_b		
Sony α7r1	35.9×24	4.86	0.33	0.33	0.32	0.7	0.04
Sony α7r3	35.6×23.8	4.5	0.42	0.38	0.39	0.7	3.03
Canon T1i	22.3×14.9	4.69	1.36	1.18	1.15	0.93	5
Sony IMX345	8.27×5.51	1.4	0.3	0.31	0.32	1.06	2.37

contained within each square serves as a substitute for the expected value to fit Eq. (3.2). To minimize the error in the expected value of the RAW pixel intensity of each square, I averaged all pixels within each square from a set of five images captured using the same settings. I captured several such sets starting with a base ISO of 100 and an appropriate exposure time such that none of the pixels was saturated. Note that ISO 100 corresponds to the gain of 1. To capture subsequent sets of images, the ISO was repeatedly doubled and exposure time halved. This maintained the same mean intensity while ensuring complete use of the camera sensor's dynamic range. The number of image sets varied from camera to camera and was typically between five and seven.

Estimation A nonlinear solver [85] was used to estimate the 5 parameters of the noise model from Eq. (3.2). Very noisy samples with SNR less than one were excluded. Estimated parameters for some calibrated cameras are given in Table 3.1.

Other models Most noise models follow a similar procedure of computing the variance of pixels in real images and then fitting with linear or nonlinear solvers. The open-source project *darktable* [54] contains an extensive database consisting of per-gain, per-color calibrated parameters for 313 cameras (Section 3.1.1). The project relies on fitted LUTs for traditional image denoising [100].

Black-box neural networks (Section 3.1.3) on the other hand cannot be fit with simple solvers. The weights for these deep generative models are learned by training on a large dataset of images such as SIDD [1].

3.3 Validation

The fit for different cameras, shown in Figure 3.2, demonstrates that the model can well explain the noise found in the tested cameras. Each plot shows the measured relative noise against recorded digital intensity, as well as the fitted noise model for all colour channels. For a better comparison, the digital values from each camera were rescaled such that the maximum pixel value of every image was set to $2^{14} - 1$ where 14 is the maximum

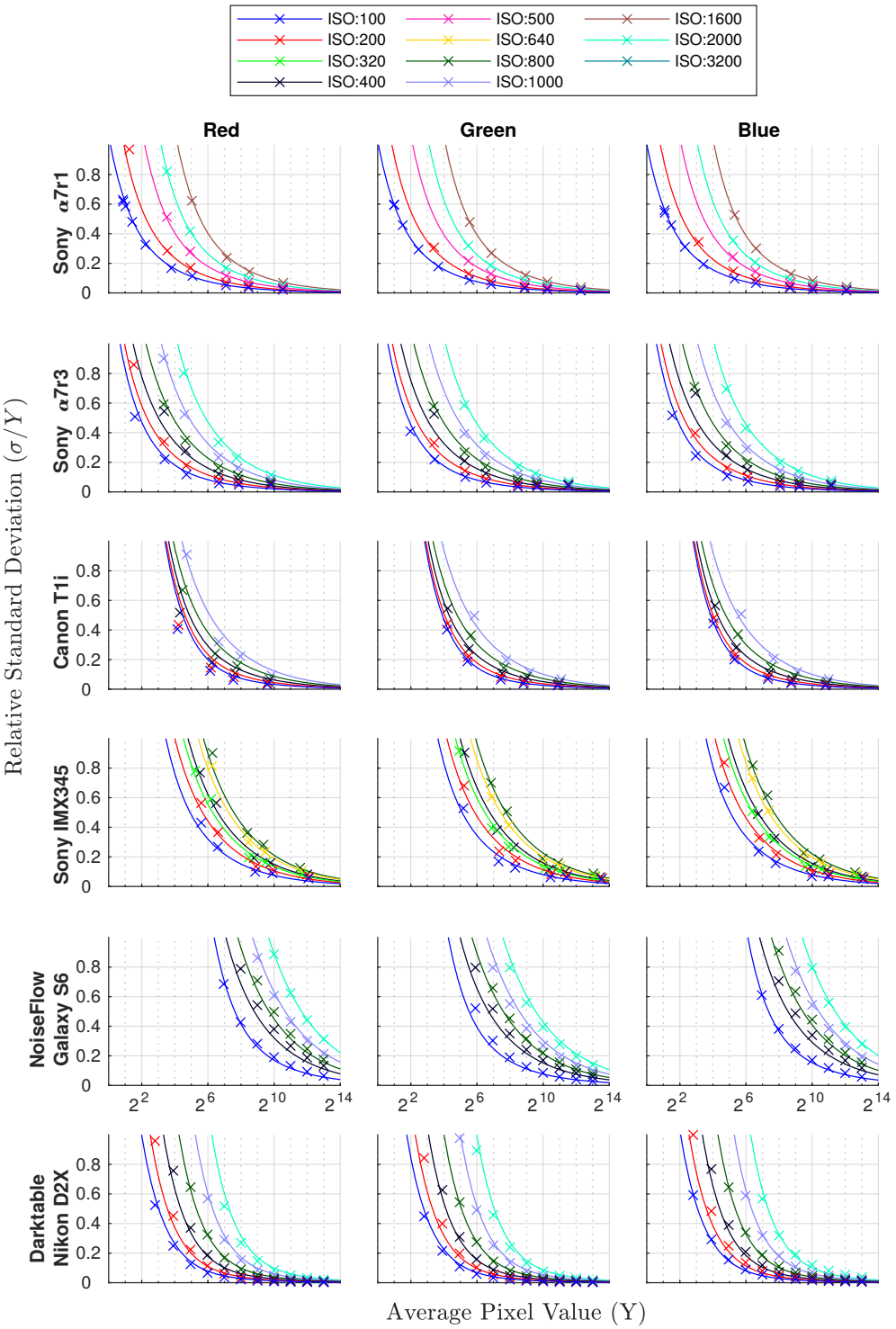
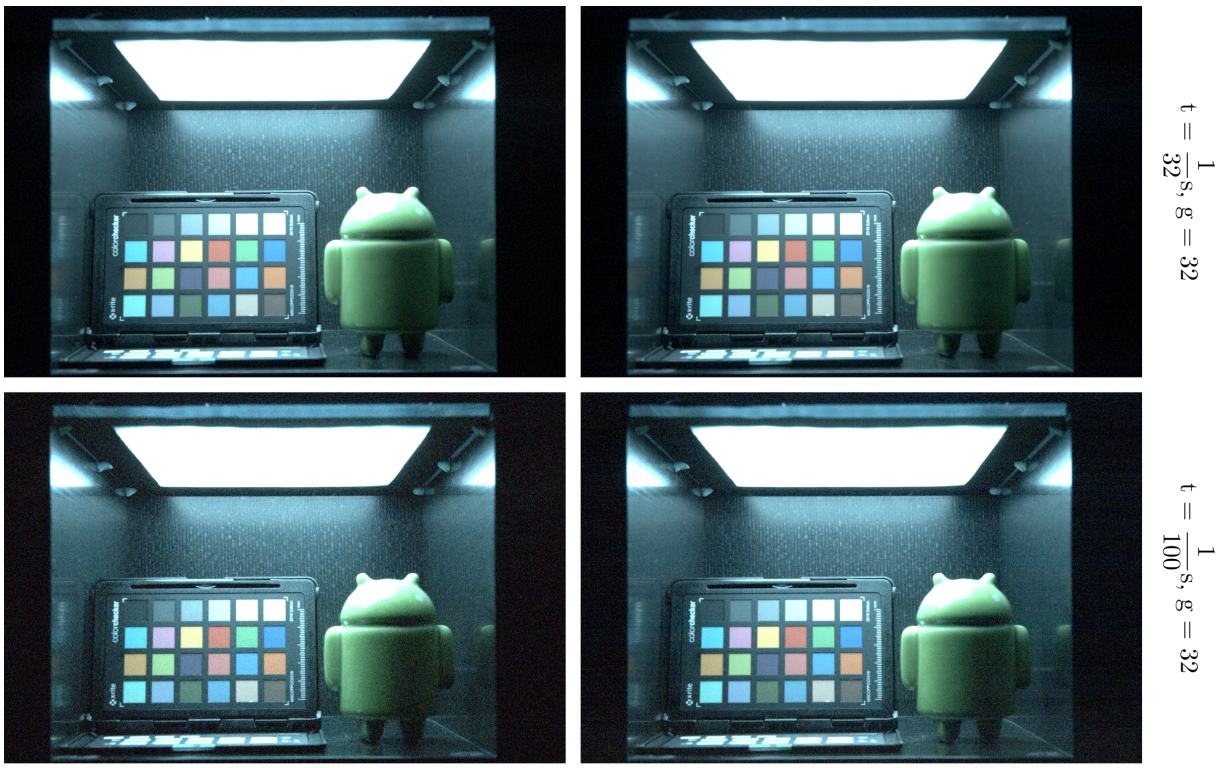


Figure 3.2: Fitted noise models for all colour channels of different cameras. Relative standard deviation is plotted against the average pixel value recorded on a logarithmic scale. The *crosses* represent measurements and the *lines* are the model predictions. The RAW pixel values of all the camera sensors have been scaled to 14-bit values to enable the comparison of different sensors. The first four rows show fits obtained by the procedure described in Section 3.2 using the noise calibration target (Figure 3.1). Rows five and six show that the noise model fits simulated data (using *NoiseFlow* [2]) and *darktable* [54] data respectively.)



(a) Simulated (b) Real

Figure 3.3: The left column (a) contains images simulated by Eq. (2.1), while the right images (b) were captured by a real camera. Each row represents a different camera setting. The images demonstrate that simulated noise is very close to real noise captured by the camera. All the images were captured using the Canon T1i and presented without white balance and using gamma encoding ($\gamma = 2.2$) to depict noise in RAW images.

bit-depth among all tested cameras.

A few interesting observations can be made about the measured cameras. If we define the dynamic range as the ratio between the largest registered value and the smallest value whose SNR is 1 (corresponding to $\sigma/Y = 1$ in the plots in Figure 3.2), we notice that the dynamic range differs substantially between the sensors. However, it should be noted that these sensors differ in their pixel pitches and resolutions (see Table 3.1). Hence, the effective amount of noise in images from the different sensors can vary even more when rescaled to the same resolution. For example, the measured noise in low-quality computer vision and smartphone sensors (rows 4 and 5) is considerable even at larger pixel values. For these sensors, notice that all curves are shifted to the right due to the scaling. The dynamic range of every sensor is also reduced with increased gain (ISO).

I also qualitatively compare real versus simulated noise for different exposure times in Figure 3.3. This comparison further confirms the accuracy of the computed noise parameters and shows that the model (Eq. (3.2)) can effectively simulate the camera capture process. Since ground truth data is unavailable for real images, I relied on such simulations for various experiments detailed in later chapters.

Table 3.2: Noise parameters fitted for *NoiseFlow* networks

NoiseFlow index	Camera	Color Coefficients			σ_{read}	σ_{adc}
		k_r	k_g	k_b		
1	Pixel	1.3	1.29	1.29	0	11.73
2	Nexus 6	5.93	5.95	5.9	0.07	13.91
3	Galaxy S6	2.77	2.76	2.79	0.08	16.16
4	LG G4	3.62	3.66	3.62	0.065	14.16

3.4 Additional cameras

First, I fit the model to noisy images generated by a realistic camera noise simulation, offered by *NoiseFlow* networks [2]. I created a virtual simulation target (similar to Figure 3.1) with squares of constant radiance and used them as inputs to the networks. The advantage of simulating cameras with *NoiseFlow* is that the ground truth radiance is known and the method offers a more faithful reproduction of noise than simple parametric noise models. As shown in row five of Figure 3.2, the simulated noise is well modelled by the proposed noise model. Table 3.2 lists the parameters corresponding to different pre-trained *NoiseFlow* networks. The variance of pixels in images generated by the pre-trained iPhone 7 *NoiseFlow* network could not be fit with the model and thus, I omit this network from all experiments.

The next set of simulated cameras was derived from *darktable* [54], which includes a large database consisting of noise parameters for 312 cameras at different ISO levels. However, the available parameters were obtained by calibration based on a simplified noise model by approximating the Poisson signal-dependent component with a normal distribution. Such an approximation is valid since, at large values, the Poisson distribution is well-approximated by the normal distribution [3]. I updated these parameters to fit Eq. (2.1) and used these in the experiments. Row six in Figure 3.2 shows that the simpler model, with only five parameters per camera, is consistent with the large LUT provided by *darktable*.

3.5 Summary

In this chapter, I introduced statistical models that capture the noise distribution of pixels. I focus on simple parametric models based on Eq. (2.1) for my experiments to improve HDR reconstructions in Chapters 4 and 5. This is because the Poisson-normal statistical model and its normal approximation are better suited for HDR imaging as they offer a convenient algebraic form under the right assumptions.

I showed how to fit the models to real-world cameras and validated the fits by comparing

simulated images to real ones. Once fitted, it is possible to rely on physically accurate simulations either for training large deep learning models (Chapter 9) or for validating results when ground truth is difficult to obtain (Chapters 4 and 5).

Chapter 4

Noise-aware HDR merging for still scenes

Averaging pixel values across a stack of captured images is a fundamental operation in many imaging problems, including HDR reconstruction. I showed in Chapter 2 that to extend the dynamic range of a reconstructed image, the stack typically includes images captured with varying exposure times [4, 46, 49, 58, 101, 155]. However, it is also possible to improve the SNR, especially at low light levels, by capturing the stack of images with different gain (ISO) or the same exposure and gain settings [53, 57, 60]. The key problem common to all these applications is the selection of an accurate HDR estimator, which provides the best estimate of radiance from a stack of noisy camera pixel values.

In this chapter, I argue that proper noise modelling is essential to identify the best HDR radiance estimator. I demonstrate how to derive several scene radiance estimators by MLE.

Further, I show how different approximations to the model result in different estimators and compare them both on simulated and real data. This chapter is restricted to static, well-aligned images and I do not consider the problems of pixel alignment and deghosting [73, 157].

4.1 HDR radiance estimation

Any method attempting to accurately estimate the scene radiance from multiple images strives to increase the dynamic range while simultaneously reducing noise. Given a stack of RAW images $i = 1, 2, \ldots, N$, let $y_i(p)$ represent the RAW value of pixel p and image i captured using a corresponding exposure time t_i and gain g_i. Further, let $x_i(p)$ be the scaled pixel value which represents a sample of the i^{th} scaled random variable $X_i(p)$ according to Eq. (2.2).

4.1.1 Simple or heuristic estimators

Early works [27, 101, 114] focused their efforts on accurately inverting the camera response function (CRF). They then combined linear pixels with simple estimators.

Uniform estimator The simplest estimator is the arithmetic mean of all the available samples and is referred to as the *Uniform* estimator. This is obviously a poor estimator as pixel values from different images in the stack are sampled from different distributions and have different SNRs.

Hat-shaped estimator The widely-used weighting scheme proposed by Debevec and Malik [27] assigns higher weights to image pixel values in the middle of the intensity range. The weights are functions of tone-mapped pixel values that pass through the whole pipeline. Let the smallest and largest intensities that can be recorded by a particular sensor be $y_{\min}$ and $y_{\max}$ respectively. Approximating the CRF by the gamma function, the weights are equal to

$$w_i(p) = \begin{cases} y_i(p)^{1/\gamma} - y_{\min}^{1/\gamma} + \epsilon, & y_i(p)^{1/\gamma} \leq \dfrac{y_{\min}^{1/\gamma} + y_{\max}^{1/\gamma}}{2} \\ y_{\max}^{1/\gamma} - y_i(p)^{1/\gamma} + \epsilon, & y_i(p)^{1/\gamma} > \dfrac{y_{\min}^{1/\gamma} + y_{\max}^{1/\gamma}}{2} \end{cases}, \tag{4.1}$$

where $\epsilon = 10^{-10}$ is a small constant added to ensure that the weights are strictly positive and prevent division by zero. γ is generally set to 2.2. Pixels with values close to the noise floor or the saturation point are assigned lower weights. The weights are applied to the linearized pixel values to obtain the *Hat-shaped* estimator:

$$\hat{\phi}_{\text{hat}}(p) = \frac{\sum_{i=1}^{N} w_i(p)\, x_i(p)}{\sum_{i=1}^{N} w_i(p)}\,. \tag{4.2}$$

4.1.2 Estimators based on MLE

Given the probabilistic image formation model described by Eq. (2.1), the best estimators are based on MLE, which has been shown to be near-optimal for this problem [4]. The recipe is simple and is identical to what I showed in Section 2.3.1 — derive an analytical expression for the joint likelihood (or an approximation in case the associated density or mass function is intractable), and maximise the likelihood to obtain an expression for the unknown radiance.

MLE-based solutions for an exponential family (such as Poisson and normal) are weighted averages where weights tend to be functions of the variances. Thus, it is

convenient to express the variance of $X_i(p)$ as $\sigma_i^2(p)$ after scaling Eq. (3.2):

$$\sigma_i^2(p) = \mathbb{V}[X_i(p)] = \frac{\mathbb{V}[Y_i(p))]}{t_i^2\, g_i^2\, a_i^2\, k_c^2} = \frac{\phi_i(p)}{t_i\, a_i} + \underbrace{\frac{\sigma_{\text{read}}^2}{t_i^2\, a_i^2} + \frac{\sigma_{\text{adc}}^2}{t_i^2\, a_i^2\, g_i^2}}_{\sigma_{s_i}^2\text{: static noise}}, \tag{4.3}$$

where $\sigma_{s_i}^2$ is the scaled static noise, common for all pixels in image i.

The problem is that the probability density function (PDF) of each $X_i(p)$ is a convolution of the three independent density or mass functions (see Eq. (2.1)), and does not have a closed-form expression. We cannot compute the infinite sum in the following expression for the PDF of each X_i*:

$$p(X_i) = \frac{\exp\left(\frac{1}{2\,\sigma_{s_i}^2} - \phi\, t_i\, a_i\right)}{\sqrt{2\,\pi\,\sigma_{s_i}^2}} \sum_{z=0}^{\infty} \frac{\exp\left(-(x_i - z - \phi)^2\right)(\phi\, t_i)^{z\, t_i\, a_i}}{(z\, t_i\, a_i)!}\,. \tag{4.4}$$

Unsurprisingly, it is not possible to derive weights with closed-form expressions by maximising Eq. (4.4). Thus, different simplifying assumptions with an aim to overcome this problem result in different MLE weights.

Variance-weighted estimator The noise-based estimator introduced in [82] and extended in [58] assumes that the Poisson component of each $X_i(p)$ can be approximated by a normal distribution. Eq. (2.1) then simplifies to the sum of three normally distributed random variables, which is also normally distributed, with mean $\mathcal{N}(\phi(p), \sigma_i^2(p))$ and (variance given by Eq. (4.3)). Performing MLE as described in Section 2.3.1, the log-likelihood function to be maximized simplifies to:

$$\log \mathcal{L}_{\mathcal{N}}(\phi(p)) = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\,\pi\,\sigma_i^2(p)}} + \sum_{i=1}^{N} \frac{-(x_i(p) - \phi(p))^2}{2\,\sigma_i^2(p)}\,. \tag{4.5}$$

where $\sigma_i^2(p)$ is the variance of $X_i(p)$ from Eq. (4.3). When $\sigma_i^2(p)$ and $\phi(p)$ are independent, the MLE has a simple form:

$$\hat{\phi}_{\text{var}}(p) = \frac{\sum_{i=1}^{N} \frac{x_i(p)}{\sigma_i^2(p)}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2(p)}}\,. \tag{4.6}$$

The reality, however, is that $\sigma_i^2(p)$ and $\phi(p)$ are not independent since $\sigma_i^2(p)$ is a function of $\phi(p)$ due to photon noise (refer to Eq. (4.3)). This requires another simplifying assumption, that $\sigma_i^2(p)$ can be estimated using a single observation $y_i(p)$. However, when $E[Y_i(p)]$ is approximated by $y_i(p)$, the variance computed using Eq. (4.3), can be zero or

*The pixel index (p) is dropped for brevity

negative. This may happen because the incorrectly estimated photon noise can be negative and larger in magnitude than the static noise, i.e., $y_i/t_i^2 g_i a_i^2 k_c < -\sigma_{s_i}^2$. In my implementation, the weights $\frac{1}{\sigma_i^2(p)}$ are replaced in such instances with a small value of $\epsilon = 10^{-10}$.

Iterative Expectation Maximization (EM) A simple method that produces a computationally efficient MLE solution for Eq. (4.5) is the expectation maximisation (EM) algorithm. In this iterative approach, proposed by Granados et al. [49], $\phi(p)$ is initialized based on the mean value of $y_i(p)$ computed across the stack. Then variances for every exposure are calculated according to Eq. (4.3) and a better estimate of $\phi(p)$ is found using the new variances according to Eq. (4.6). This alternating procedure is repeated until convergence.

Full MLE The MLE for Eq. (4.5) can be estimated without making any additional assumptions using a non-linear optimization method. Although this is the most accurate estimator, running such a solver for each pixel is too computationally expensive to be used in practice. I include this estimator to show how close other estimators are to the true MLE solution.

Normal Photon Noise Estimator (NPNE) Eq. (4.5) has an analytical solution if we assume that the variance is only due to photon noise and there is no static noise. Such a simplification is justified because the overall noise is dominated by the photon noise component for the usable range of $\phi(p)$. Setting static noise to zero and maximizing Eq. (4.5) yields the estimator:

$$\hat{\phi}_{\text{npne}}(p) = \frac{\sqrt{\sum_{i=1}^{N} x_i^2(p)\, t_i \cdot \sum_{i=1}^{N} t_i + N^2} - N}{\sum_{i=1}^{N} t_i}\,. \tag{4.7}$$

Poisson Photon Noise Estimator (PPNE) Setting static noise to zero also enables us to simplify the PDF of each $X_i(p)$ and derive an estimator that maximizes the likelihood without the normal approximation to the Poisson distribution. This means that the random variables are sampled from:

$$X_i(p) \sim \frac{\text{Pois}(\phi(p)\, t_i)}{t_i}\,. \tag{4.8}$$

And the new log-likelihood function to be maximized is:

$$\ln \mathcal{L}_{\text{Pois}}(\phi(p)) = \sum_{i=1}^{N} x_i\, t_i\, \ln \phi\, t_i\, - \sum_{i=1}^{N} \phi\, t_i - \sum_{i=1}^{N} \ln\,(x_i\, t_i)!\,. \tag{4.9}$$

The last sum does not depend on $\phi(p)$ and can safely be ignored. The resulting Poisson Estimator takes on the simple form:

$$\hat{\phi}_{\text{ppne}}(p) = \frac{\sum_{i=1}^{N} x_i(p)\, t_i}{\sum_{i=1}^{N} t_i} \tag{4.10}$$

Such an estimator is a classical choice in the imaging industry [4]. Here, I demonstrated how it can be derived from the assumption of Poisson noise. Additionally, in Appendix A, I show that *PPNE* is the unique minimum-variance unbiased estimator (MVUE) by applying the *Lehmann-Scheffe theorem* under the assumption of zero static noise. This is an important result since estimators produced by MLE are generally not guaranteed to be unbiased or to have the minimum variance.

4.1.3 Estimators and noise parameters

An advantage of *NPNE* and *PPNE* estimators that ignore static noise is that they do not require knowledge of the noise parameters to provide accurate estimates. Eq. (4.7) and Eq. (4.10) are functions of $x_i(p)$, which depends only on the parameter k_c. And, since k_c is the same for all exposures, it is effectively a constant multiplier, which does not affect relative radiance values.

4.2 Experiments with synthetic data

4.2.1 Comparison of Estimators

The empirically compare the calibration-independent estimators, *NPNE* and *PPNE*, to the classical *Uniform* and *Hat-shaped* estimators [27] and state-of-the-art MLE-based estimators, *Variance-weighted* [58] and *Iterative EM* [4, 49]. I rely on Monte Carlo (MC) methods and simulate 10,000 HDR exposure stacks. Each stack consists of three exposures separated by five stops in exposure time and a constant gain of 8. Starting with 100 logarithmically-spaced ground truth values of radiance, spanning a dynamic range of 24 stops, I simulated the probabilistic camera capture process described in Chapter 3. The noisy samples were generated using the parameters of calibrated cameras from Table 3.1 and Table 3.2.

Figure 4.1 shows relative errors of the compared estimators for some calibrated cameras. I include contributions to the error due to both bias and variance. The plots show relative

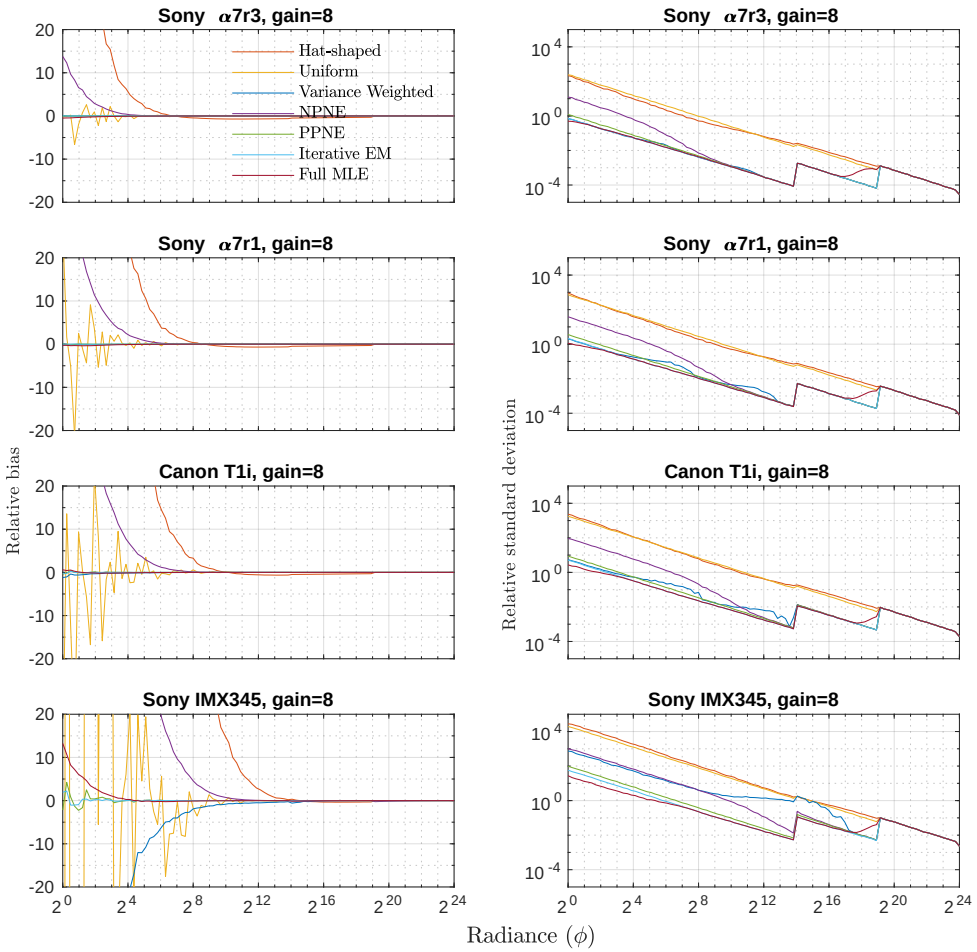


Figure 4.1: MC simulation results: relative errors of the different estimators (*colour lines*) for logarithmically spaced values of the ground truth scene radiance, ϕ (*x-axis*). The errors arise due to nonzero biases (left column of plots) and nonzero standard deviations (right column of plots). The simulation was performed for three exposures, spaced five stops apart.

quantities because they better correspond with perceived magnitudes (Weber's law),

$$\begin{aligned} \mathrm{B}_r(p) &= \frac{E[\hat{\phi}(p)] - \phi(p)}{\phi(p)} \qquad \text{(Relative bias)}\,, \\ \sigma_r(p) &= \frac{\sqrt{E[(\hat{\phi}(p) - E[\hat{\phi}(p)])^2]}}{\phi(p)} \qquad \text{(Relative SD)}\,. \end{aligned} \tag{4.11}$$

Both these measures are combined and reported as relative RMSE (Root Mean Squared Error) in later experiments,

$$\begin{aligned} \mathrm{E}_r(p) &= \sqrt{\mathrm{B}_r^2(p) + \sigma_r^2(p)} \\ &= \frac{\sqrt{E[(\hat{\phi}(p) - \phi(p))^2]}}{\phi(p)} \qquad \text{(Relative RMSE)}\,. \end{aligned} \tag{4.12}$$

In Figure 4.1, notice that when the radiance ϕ is large, all the estimators are unbiased, and the error is solely due to the relative standard deviation. As expected, both the bias and standard deviation are highest for the smartphone sensor (Sony IMX345), and smallest for the full-frame mirrorless camera (Sony α7r3). The sawtooth patterns visible in the standard deviation plots are due to the three exposures: the highest five stops are captured only in images with the shortest exposure time and, therefore, the error drops at $\phi \approx 2^{19}$ and $\phi \approx 2^{14}$ when data from the second and third images become available.

The results in Figure 4.1 confirm that the estimators that do not account for noise (*Uniform* and *Hat-shaped*) result in the highest bias and variance. The bias of the *Uniform* estimator appears unstable at low radiance values because the estimate is dominated by noise. The popular *Variance-weighted* estimator performs reasonably well for the high-quality sensor (Sony α7r3) but results in a noisy estimate with a negative bias for the two other sensors. The negative bias is due to the clamping of weights since variances can not be negative. As expected, the *Full MLE* achieves the best performance. It is, however, much more computationally expensive. Notice that, the *Iterative EM* estimator does not converge to the same solution as the *Full MLE* for lower radiance values.

The most interesting results are seen for the two estimators that account only for photon noise. The estimator that assumes photon noise is normally distributed (*NPNE*) introduces a positive bias and results in a higher standard deviation than the estimator that assumes Poisson photon noise distribution (*PPNE*). Overall, the error of the *PPNE* is comparable with that of the *Iterative EM* estimator, which is the best estimator used in practical applications. The relative standard deviation of noise is only marginally higher than that of the *Iterative EM* but only for very low pixel values. However, it has a major advantage over the *Iterative EM* as it does not require knowledge of the noise parameters. Therefore, it can be used with any camera without prior calibration, as long as the camera noise characteristic can be well explained by the model given by Eq. (2.1).

4.2.2 Gain modulation

An alternate acquisition strategy for capturing HDR scenes is modulating gain while maintaining the same exposure time and consistent motion blur across images [53]. Here, I compare the estimators on a stack of gain-modulated captures. The plots on the top of Figure 4.2 indicate that the performance of most estimators is very similar, except for the failure of the *Hat-shaped* estimator for medium radiance values.

4.2.3 Robustness to noise

I validated the performance of the estimators for exposure stacks captured with increased camera noise. The noise in input image stacks was artificially increased by amplifying the

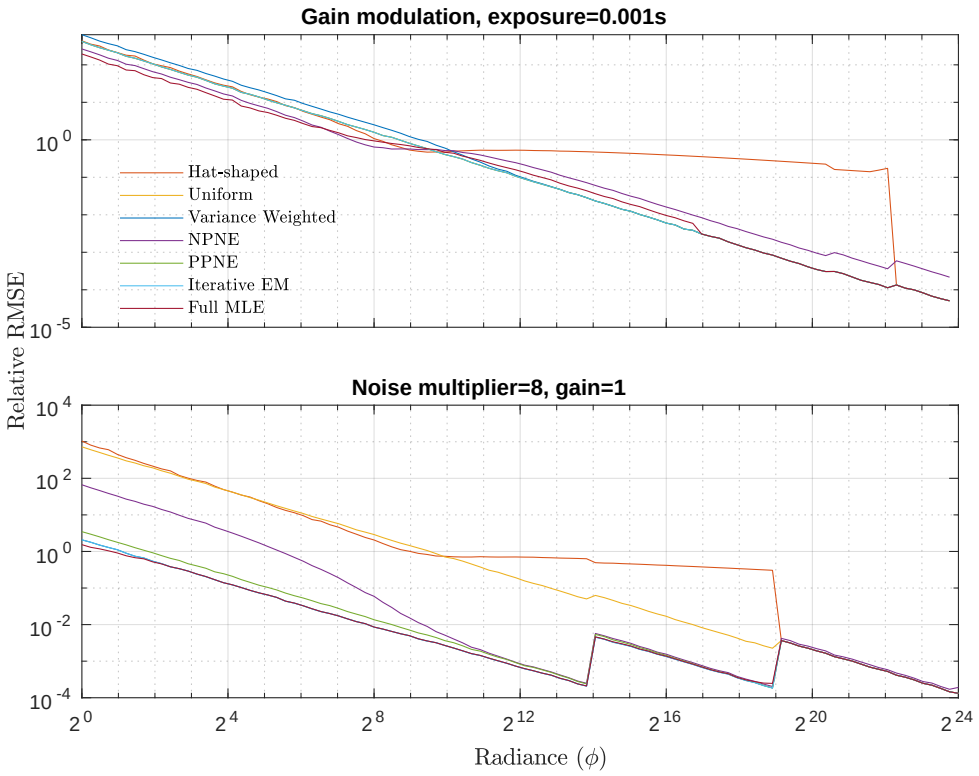


Figure 4.2: Relative errors of estimators for two additional scenarios for the Sony α7r3: using a stack of images with different gain and same exposure time (top) and the effect of increasing the static noise by 8× (bottom). The increased error of the *Hat-shaped* estimator is due to a small negative bias (see left column of Figure 4.1).

contribution of static noise for the Sony α7r3 sensor. See the bottom of Figure 4.2 for the simulation with 8× static noise. These additional MC simulations, confirm that the relative performance of *PPNE* does not noticeably degrade with noise. Its relative error is very similar to the calibration-sensitive *Iterative EM* estimator even when the input is very noisy.

In reality, high noise can be expected in images when the camera gain used is large. The calibration fits depicted in Figure 3.2 include only a few gain values. For further validation, I used similar parameters estimated by the open-source project *darktable* [54] for 312 cameras. In Figure 4.3, we see that the performance of PPNE is almost identical to *Variance-weighted* and *Full-MLE* estimators which are dependent on accurate noise calibration. A clear separation is visible between estimators that account for noise (*Hat-shaped* and *Uniform*) and the MLE-based estimators motivated by a noise model.

From the plots, we can conclude that *PPNE* is robust to camera noise for most working values of gain. The benefit of accurate calibration is observed only for values exceeding 256 (corresponding to ISO 25 600). In the real world, it is not customary to capture HDR scenes using such high gain and *PPNE* is a near-optimal estimator, which does not require knowledge of camera noise parameters and is not affected by any inaccuracy of such noise calibration.

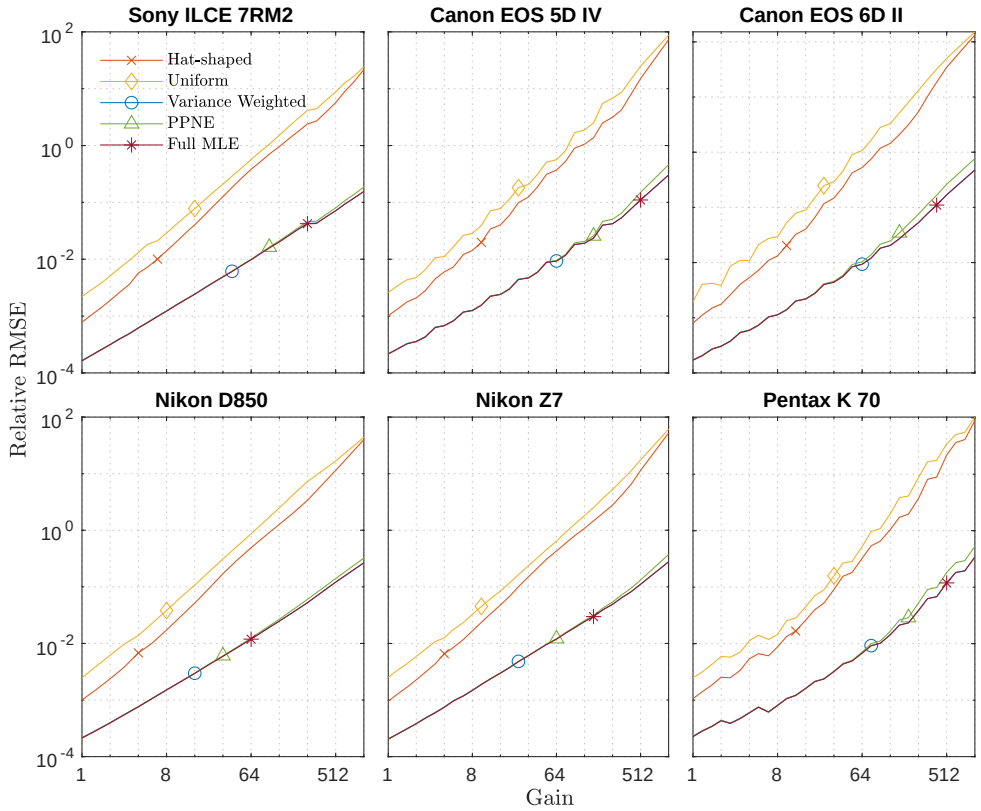


Figure 4.3: High gain simulation using *darktable* [54] cameras. Although 6 cameras with the widest range of gains are depicted here, a similar pattern is observed for the 312 cameras in the database. The *PPNE* curve starts separating from the calibration-sensitive estimators only for gains greater than 256 (ISO greater than 25 600), which are offered by only a few cameras.

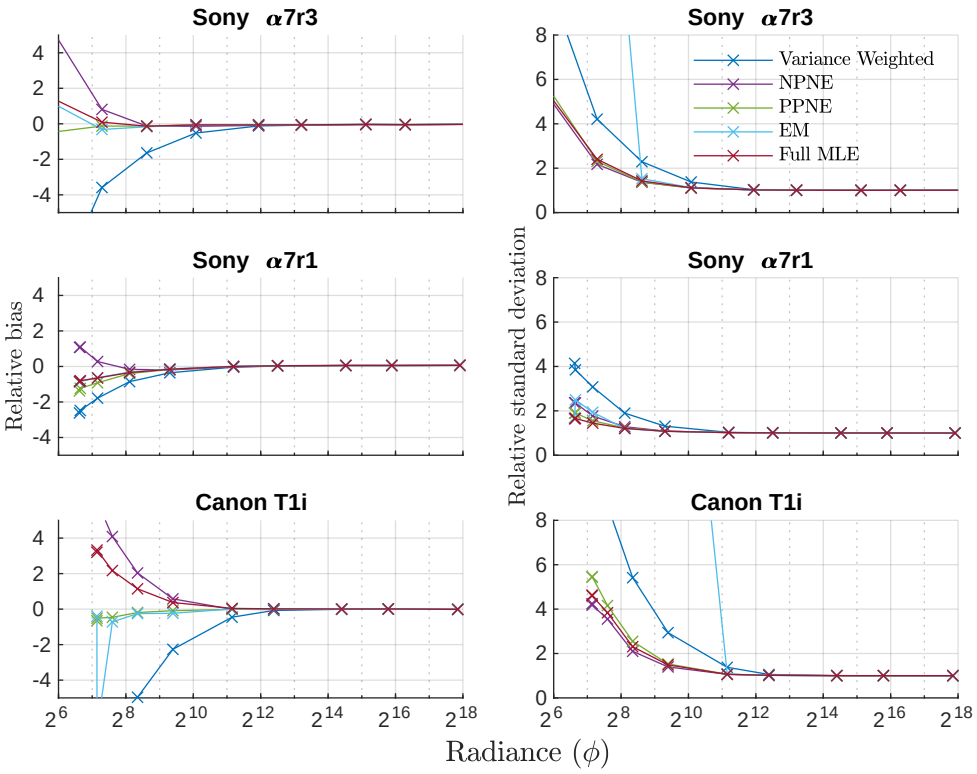


Figure 4.4: Results for real images: observed relative biases and standard deviations for the green channel of three cameras. The sudden increase in standard deviation at $\phi \approx 2^{11}$ for the *Iterative EM* estimator in some plots is due to the non-convergence of the EM algorithm for low inputs.

4.3 Validation on Real Images

To make sure that simulation results are not the outcome of wrong assumptions about camera noise, I measured the errors of different estimators on real data using a stack of images of the calibration target (see Figure 3.1-left). The stack is composed of images captured with different exposure times and gains, but the same average intensity. The reference radiance ϕ was calculated as the average of all pixels in each square of the target. Figure 4.4 shows the error, due to bias and standard deviation, of the HDR estimations for calibrated cameras. Here we see a similar pattern as in Figure 4.1, where the standard deviation of the *Variance-weighted* estimator is much higher than other MLE-based estimators and it has a large negative bias at low radiance. The performance of the analytical *PPNE* is very similar to that of the *Iterative EM* and the *Full MLE*.

4.3.1 Large HDR dataset

Next, I applied the estimators on 73 images from *The HDR Photographic Survey* [37], a database of image stacks captured using a Nikon D2x. From the 106 exposure stacks, I discarded 33 of them as they contained scene motion and produced ghosting [73, 157].

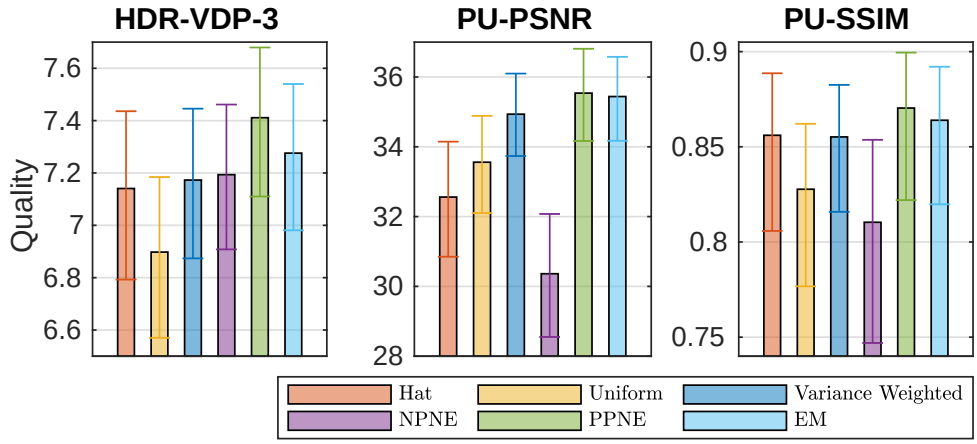


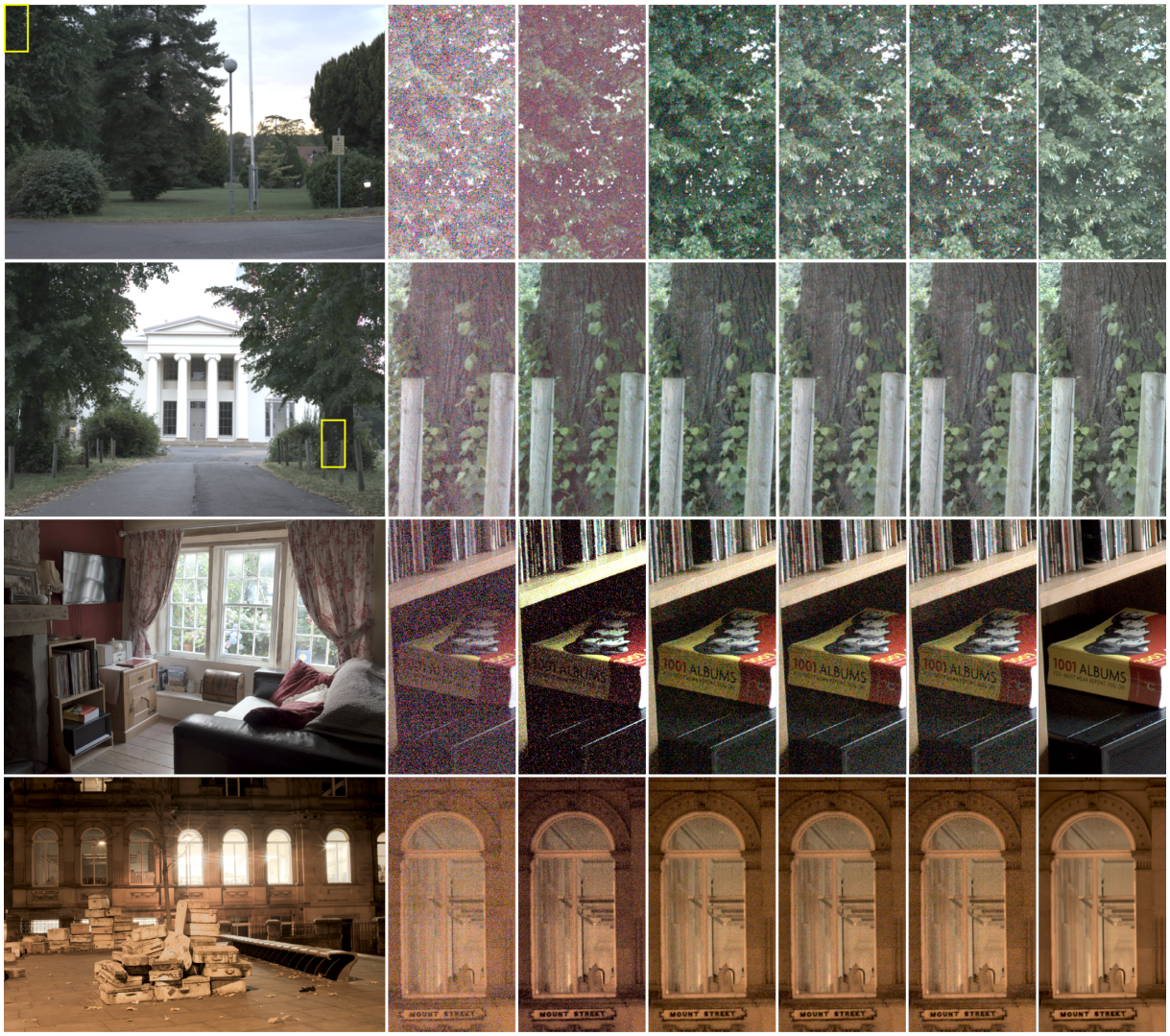
Figure 4.5: The average image quality of the reconstructions along with 95% confidence intervals using different estimators for 73 well-aligned scenes of *The HDR Photographic Survey* [37]. In all the plots, a higher value indicates better performance.

Figure 4.5 summarises the qualities of the reconstructions using the same three HDR image quality assessment (IQA) as used in Section 4.3.3. The ground truth was reconstructed from all nine exposures using the *EM* estimator, while the reconstructions of various estimators were computed on a reduced stack containing the three shortest exposures since they represented the noisiest estimates. All the images in this database were captured with ISO 100.

As before, we observe that the reconstruction of *PPNE* is of very high quality and is closest to the reference. Some of the estimators discussed in Section 4.1.2 require accurate noise parameters. Since calibration parameters are unavailable for the original camera, I used values for the same make and model from the *darktable* [54] database. This could explain why the iterative *EM* estimator performs slightly worse, contradicting results from Section 4.2. Sensitivity to noise parameters is a well-known limitation of other noise-aware estimators [3] and motivates the use of *PPNE*. The above experiments show that such a substitution can be made without having to sacrifice the final image quality.

4.3.2 Qualitative comparison

For final validation, I show visual differences in HDR images due to the choice of the estimator in challenging conditions. I captured several scenes with three exposure times and spaced two stops apart. The images, processed with different estimators, show substantial differences, as depicted in Figure 4.6. The *Hat-shaped* weights resulted in a noisy image with a positive bias (Figure 4.6a) in dark regions. *NPNE* substantially reduced the amount of noise but still produced the bias (Figure 4.6b). The *Variance-weighted* estimator produced a noisier image than *NPNE* (Figure 4.6c) and also introduced a negative bias. This made some pixel values darker than they should be, resulting in an accidental increase in contrast. Merged images of *PPNE* (Figure 4.6d) and the *EM* estimator (Figure 4.6e)



(a) Hat-shaped (b) NPNE (c) Variance-weighted (d) Iterative EM (e) PPNE (f) Ground truth

Figure 4.6: HDR reconstructions of scenes "Trees", "House", "Cottage" and "Street", using different estimators given exposure stacks of 3 images captured by the Sony α7r3 at ISO 6400 and gamma-encoded for visualisation ($\gamma = 2.2$). The positive bias of the *Hat-shaped* estimator and *NPNE*, as well as the negative bias of the *Variance-weighted* estimator, are visible in the dark regions. The images produced by the iterative EM estimator and *PPNE* are almost identical. In all scenes, the shortest exposure times were deliberately small to test the robustness of the estimators.

Table 4.1: The reconstruction error for the scenes depicted in Figure 4.6. The error is computed using HDR IQA metrics: PU-PSNR, PU-SSIM [11] and HDR-VDP-3 [106] (v3.0.6, Q-values). For all the metrics, a higher value denotes higher quality. In each column, the highest value has a gold background, the second best has a silver background and the third has a bronze background. Overall, *PPNE* is the second-best and is narrowly outperformed by the calibration-sensitive *EM* estimator.

Estimator	PU-PSNR				PU-SSIM				HDR-VDP-3			
	House	Trees	Cottage	Street	House	Trees	Cottage	Street	House	Trees	Cottage	Street
Hat-shaped	20.266	13.241	18.516	9.019	0.636	0.489	0.51	0.302	5.719	5.39	5.013	4.638
Var-weighted	16.412	18.191	22.476	13.267	0.624	0.497	0.578	0.512	6.049	5.946	5.846	5.581
NPNE	24.326	17.451	13.588	13.331	0.681	0.543	0.374	0.289	6.027	6.114	5.391	5.304
PPNE	24.428	18.876	22.336	13.361	0.72	0.538	0.607	0.432	6.255	6.253	5.904	5.802
EM	24.922	18.339	23.059	13.962	0.72	0.52	0.614	0.524	6.273	6.37	5.983	5.928

show the least amount of noise and smallest bias. Apart from the independence of noise parameters, another advantage over the *Iterative EM* estimator is the reduced computation time of *PPNE* due to its analytical form, which is important for high-resolution images captured by a DSLR.

4.3.3 Quantitative comparison

I used the images shown in Figure 4.6 to calculate quality scores for three HDR IQA metrics: PU-PSNR, PU-SSIM [11] and HDR-VDP-3 [106]. The test images were obtained from three exposures, captured at a high ISO setting, and merged with each estimator whereas the reference images were obtained by merging five ISO 100 exposures with the *EM* estimator (the most accurate). The results, shown in Table 4.1, confirm the findings of other experiments; *PPNE* produces results that are only marginally worse than those of *EM*, even though *EM* was used to generate the reference images.

4.4 Summary

Although the state-of-the-art HDR reconstruction methods advocate using MLE solvers that require accurate camera parameters, in this chapter, I demonstrated that they provide little advantage over the simple Poisson noise estimator (PPNE), which does not require camera noise calibration. I showed that the Poisson noise estimator is unbiased and its standard deviation is only marginally higher that of the near optimal MLE solution for very low pixel values. Such a difference is unlikely to be noticed in complex images. For a simplified noise model, the Poisson estimator is provably MVUE. Furthermore, I showed how each estimator can be derived making different simplifying assumptions about the camera noise model, and I illustrate the relative errors of the estimators using gain modulation and under increased static noise. In all experiments, the Poisson noise estimator was consistently among the best performing estimators.

A distinguishing feature of PPNE is that it is motivated by a statistical noise model, but is insensitive to camera-specific parameters. This makes it applicable to a wide range of cameras without noise calibration.

Chapter 5

Automatic exposure compensation

Irrespective of the HDR estimator considered in the previous chapter, regions of reconstructed HDR images that lack texture and appear smooth often exhibit banding artefacts, as illustrated in Figure 5.1. While these artefacts resemble compression artefacts, their root cause in this context differs significantly. This chapter begins by elucidating that banding arises from inaccurately reported capture parameters. Subsequently, it introduces a robust procedure aimed at eliminating banding.

The contributions of this chapter are particularly significant for subsequent chapters such as Chapters 6 and 8, which include subjective experiments. This is because humans are sensitive to banding artefacts in controlled studies. Besides producing visually appealing, artefact-free images, addressing inaccurate capture parameters is crucial for reconstructing a physically accurate HDR representation of any real scene—the primary objective of this dissertation.

5.1 Motivation

The reason for banding in merged exposure stacks is a mismatch between the actual capture parameters and those reported by the camera, typically in the exchangeable image file format (EXIF) header. Such inaccuracies could be caused by

- limited accuracy of the (mechanical) aperture and shutter
- wrongly reported EXIF data due to rounding (e.g. $1/60$ exposure time could be reported instead of $1/64$)
- changes in scene illumination, for example, due to flickering lights or overcast skies with moving clouds

It is important to understand that the artefacts shown in Figure 5.1 are not due to spatial misalignment because the image stacks were carefully selected and contain no camera or



Figure 5.1: The first row depicts patches from images of overcast skies taken from large HDR datasets [37, 56]. Misalignment between the reported camera settings results in banding artefacts (pointed at by the red arrows) at the boundary between two exposures. This problem is not limited to the sky and affects other regions too, as shown by patches in the second row. Note that RAW exposure stacks are carefully picked and are spatially-aligned. Thus, they do not require CRF inversion or motion compensation.

object motion. Further, since the input images are linear RAW images, the artefacts are not due to incorrect CRF inversion.

The inaccuracies of exposure time found in the EXIF metadata are substantial, reaching 40%, as shown in Figure 5.2. To prepare this figure, I computed the relative error between the reported EXIF exposure times and the times estimated by the proposed method. The two plots show similar error distribution for two datasets of multi-exposure HDR image stacks [37, 56]. The errors of such magnitude can easily introduce banding artifacts, as shown in Figure 5.1.

Correctly estimating exposures is also relevant for HDR merging with deghosting, which involves detecting and registering pixels that belong to moving objects [127, 157] after scaling by exposure time, gain and aperture. Erroneous capture parameters make this task more challenging as pixel differences could be caused by either object movement or incorrect scaling.

In this chapter, I describe a method to estimate exposure ratios directly from a stack of images, which eliminates banding artefacts from the merged HDR images. The problem is efficiently solved as a sparse linear system while accounting for heteroskedastic camera

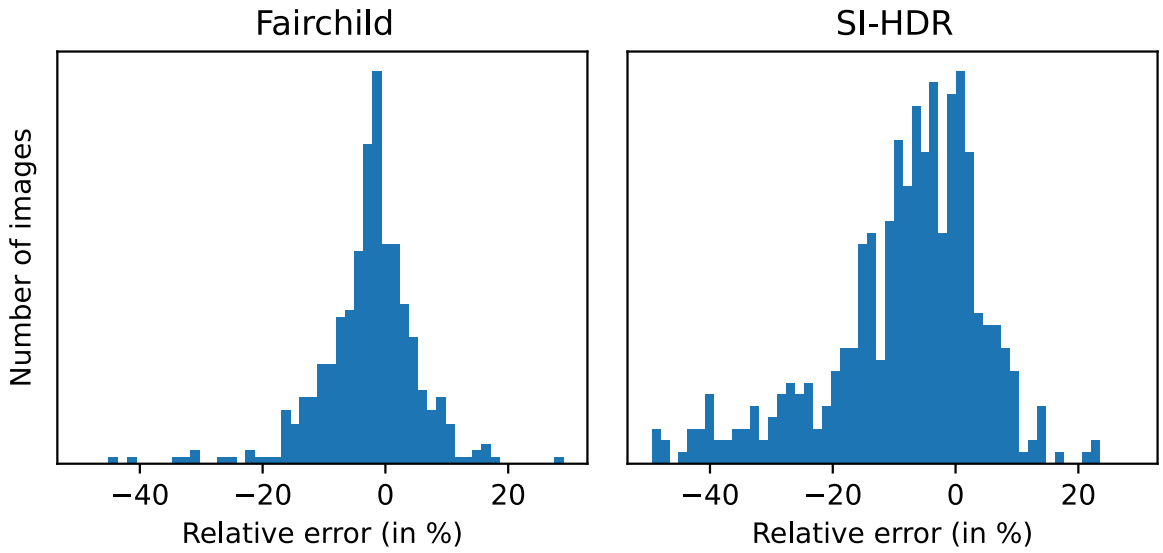


Figure 5.2: Histograms of relative errors for two popular HDR datasets. The computed sample standard deviations of relative errors are 9% for the Fairchild survey [37] and 14% for the SI-HDR dataset [56].

noise with known or unknown noise parameters.

The proposed exposure estimation can be used to improve ground truth reconstructions of several existing stack-based HDR datasets. Further, algorithms that enhance merged reconstructions by iterative optimisation [97], as well as deep neural networks that use HDR images for other related tasks [23, 169] stand to be improved by the proposed, more accurate exposures.

Correctly estimating exposures is also relevant for HDR merging with deghosting, which involves detecting and registering pixels that belong to moving objects [127, 157]. Erroneous capture parameters make this task more challenging, as pixel differences could be caused by either object movement or incorrect exposure time.

Although banding artefacts are visible only in smooth image regions in Figure 5.1, inaccurate exposures affect all pixels in the image. Incorrect pixel intensities pose a serious complication for applications that utilise cameras in place of expensive light measurement instruments. Multi-image exposure stacks often serve as substitutes for more accurate measurements from instruments such as spectrophotometers. In Section 5.5.4, I highlight potential discrepancies in the measured modulation transfer function (MTF) of an HDR display when the exposure stack of a slant-edge [138] is merged with incorrectly reported capture parameters.

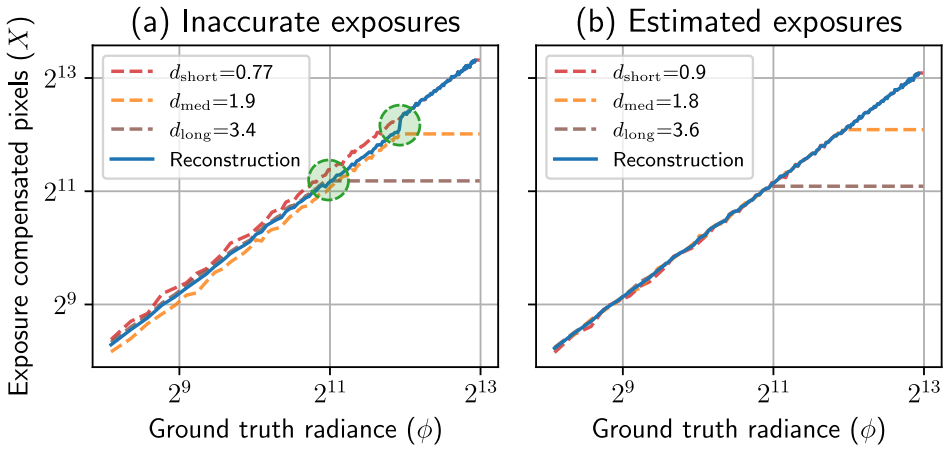


Figure 5.3: A linear gradient spanning 13 stops is captured with exposure times $[0.5, 2, 4]$ (in seconds) by simulating the Sony $\alpha 7r1$ at ISO 3200. Each capture is quantised to 8 bits for easy visualisation. The left plot depicts exposure-compensated pixels that represent samples from $X(p)$ (according to Eq. (2.2)) with inaccurate exposures (red, yellow and brown dashed lines). The reconstructed gradient (blue line) is jagged around $\phi = 2^{11}$ and $\phi = 2^{12}$ (green circles) due to misaligned exposures caused by biased exposure values. It is possible to reconstruct the smooth gradient by correctly aligning exposure ratios as shown in the right plot.

5.2 Related work

The problem of inaccurate capture parameters was identified in very early works in HDR merging [27, 114]. However, most of these focused on the challenging task of inverting the camera response function (CRF) under the assumption of film or sensor *reciprocity* [151]. Mitsunaga and Nayar [114] used a polynomial model to jointly estimate the CRF and exposure times. Then, Grossberg and Nayar [51] demonstrated how to recover the brightness transfer function (BTF), a function that describes how the brightness transfers from one image to another. They recovered the BTF from image histograms and stipulated that it can be used to estimate exposure ratios if the CRF is known.

Based on [51], Cerman and Hlavac [20] assumed a linear CRF by relying on RAW pixel values. They computed the brightness transferred image histograms and used them to weight a system of equations in the linear pixel domain.

In contrast to their work, I solve a weighted linear system in the logarithmic domain to estimate exposure ratios. Weights in the system of equations are derived from a popular statistical camera noise model and ensure a noise-optimal solution. The proposed approach systematically handles heteroskedastic camera measurements and provides accurate estimates even when many pixels are under-exposed or affected by noise.

5.3 Banding due to inaccurate exposure

To better understand the reason for banding, consider the one-dimensional linear gradient depicted in Figure 5.3 (left). The noisy measurements (dashed lines), obtained by simulating captures using calibrated noise parameters of the Sony $\alpha 7r1$ (from Table 3.1), become misaligned when scaled with inaccurate exposure times. Merging such images results in a jagged reconstruction (solid blue line), causing banding in an otherwise smooth output. Notice that although the reconstruction deviates from the ground truth for almost all pixels, artefacts will be visible only at transition points when an image in the stack saturates as highlighted by green circles.

While Figure 5.3 demonstrates the problem with simple averaging, banding is further exaggerated when sophisticated algorithms based on physical noise models [49, 55, 58] are used. This is because pixels of longer exposures are more reliable (due to smaller noise variance), but they saturate at lower physical values. Thus, these estimators weight longer exposure pixels more, and there is a sharp change upon saturation of any image in the stack.

Estimating relative exposure ratios w.r.t one of the images correctly, enables us to align all exposures to produce the reconstruction in Figure 5.3 (right). Note that the exposure-aligned reconstruction may still not coincide with the ground truth since the correct baseline is unavailable. However, accurately estimating relative ratios is sufficient to align the capture parameters and eliminate banding.

5.4 Exposure estimation in real images

To prevent banding and obtain physically accurate pixel values, the exposures of all images in the stack are aligned by estimating all scaling constants d_i. However, computing them directly from input pixels is impossible since the correct absolute measurements that represent observations of $X(p)$ from Eq. (2.2) are not known. Such unknowns are thus eliminated by estimating the ratio of exposures between any two images in the stack instead. For a scene with constant illumination and no scene motion, this ratio should be the same for all pixels from a given pair of images. Its expected value is:

$$d_{ij} = \mathbb{E}_p \left[\frac{Y_i(p)}{Y_j(p)} \right] , \tag{5.1}$$

where i and j index different images in the exposure stack, and $\mathbb{E}_p$ indicates the expected value computed over all M pixels. To allow for fast computation using linear solvers, it is

preferable to operate on logarithmic values. Thus, let

$$e_{ij} = \log d_{ij} \quad \text{and}\, L_i(p) = \log Y_i(p)\,. \tag{5.2}$$

Although the density function of $L_i(p)$ is unknown because of the log transformation, it is possible to approximate the expected value of any transformed random variable using its Taylor expansion. The results, detailed in Eq. (B.2) in Appendix B, are only applicable to normally distributed random variables. It is thus imperative to use only well-exposed pixels (I will show how to do this in Section 5.4.4), since the Poisson photon noise component of such pixels is well-approximated by a normal distribution. Applying the result for $\log Y(p)$ we have an approximate expression for the expected value:

$$\begin{aligned} e_{ij} &= \log \mathbb{E}_p \left[\frac{Y_i(p)}{Y_j(p)}\right] \approx \mathbb{E}_p \left[\log \frac{Y_i(p)}{Y_j(p)}\right] \\ &= \mathbb{E}_p \left[L_i(p) - L_j(p)\right]\,. \end{aligned} \tag{5.3}$$

The equality approximately holds for the operating range of pixel values, as I will detail in Section 5.4.2, Eq. (5.8). This allows computation of the expected value over all pixels by setting up and solving a linear system. For the most reliable estimate, information from all the available exposures should be utilised (all possible values of i and j). Unlike previous work [20], I thus consider not only ratios between neighbouring exposures but between all pairs. This results in a large yet sparse linear system:

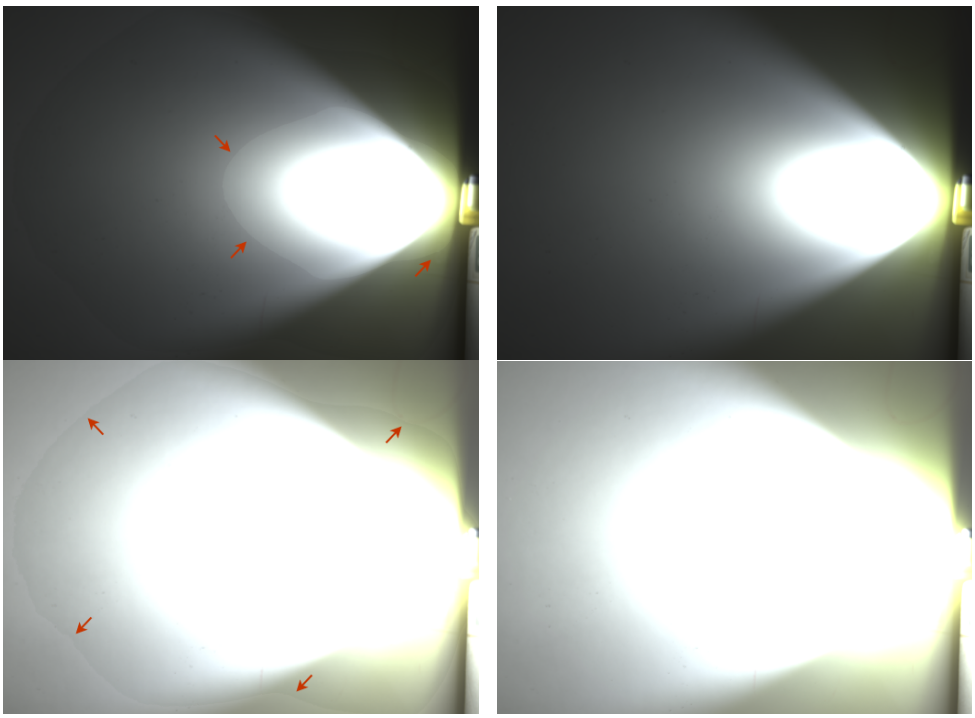
$$\sqrt{\boldsymbol{W}} \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \sqrt{\boldsymbol{W}} \begin{bmatrix} L_1(1) - L_2(1) \\ L_1(2) - L_2(2) \\ \vdots \\ L_1(1) - L_3(1) \\ \vdots \\ L_{N-1}(M) - L_N(M) \end{bmatrix}, \tag{5.4}$$

or more compactly,

$$\sqrt{\boldsymbol{W}}\boldsymbol{O}\boldsymbol{e} = \sqrt{\boldsymbol{W}}\boldsymbol{m}\,, \tag{5.5}$$

where $\boldsymbol{W}$ is a diagonal weight matrix denoting the relative importance of each row of the system. In practice, the weighted system does not always provide a good solution for shorter exposures. Since a rough estimate of exposure values $\boldsymbol{e}_0$ can be extracted from the image metadata, a Tikhonov penalty with weight λ can be introduced to get:

$$\begin{aligned} \hat{\boldsymbol{e}}_{\text{WLS}} &= \arg\min_{\boldsymbol{e}} \boldsymbol{W}\, \|\boldsymbol{O}\boldsymbol{e} - \boldsymbol{m}\|_2^2 + \lambda \|\boldsymbol{e} - \boldsymbol{e}_0\|_2^2 \\ &= (\boldsymbol{O}^\mathsf{T}\boldsymbol{W}\boldsymbol{O} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{O}^\mathsf{T}\boldsymbol{W}\boldsymbol{m} + \lambda \boldsymbol{e}_0)\,. \end{aligned} \tag{5.6}$$



(a) Parameters from EXIF **(b)** Estimated parameters

Figure 5.4: The left column shows exposures (gamma-encoded for visualisation, $\gamma = 2.2$) of the HDR image reconstructed with EXIF parameters, while the right column shows exposures (encoded with $\gamma = 2.2$) of the HDR image reconstructed by solving the linear system represented by Eq. (5.4) with $\boldsymbol{W} = \mathbb{I}$. Banding artefacts are visible in dark (top row) and bright (bottom row) regions of the reconstruction using EXIF metadata (a). Aligning exposures according to Eq. (5.6) fixes the problem (b).

5.4.1 Homoskedastic pixels

If the noise levels of most pixels are much smaller than the signal, it is possible to get a good estimate of exposures by assuming that all the rows of Eq. (5.4) are equally reliable or homoskedastic (affected by the noise with constant variance).

To demonstrate the effectiveness of exposure alignment, I captured an exposure stack of a simple HDR scene consisting of a bright light shining at an angle, which produces an inverse-square fall-off in intensity. Since most pixels are well-exposed, as shown in Figure 5.4, we could assume that they are equally reliable and set $\boldsymbol{W} = \mathbb{I}$. This works well for the carefully controlled scene and eliminates banding artefacts that would have appeared when merging with EXIF parameters. However, the constant noise assumption breaks down for real-world HDR stacks since the noise in camera pixels is heteroskedastic [42, 83] and thus, different pixels provide different amounts of information.

5.4.2 Heteroskedastic pixels

To determine noise-optimal weights for images of real cameras, we need to derive an expression for the variance of each row of Eq. (5.4). Then, we can populate $\boldsymbol{W}$ with inverse-variance weights (i.e., each weight is given by the reciprocal of variance) to obtain the weighted least-square estimate $\hat{\boldsymbol{e}}_{\mathbf{WLS}}$. This is equivalent to MLE under the assumption that the system of equations models additive, normally distributed noise, where the variance is different for different rows (see Section 2.3.3, Eq. (2.14)). Since each entry of the output vector $\boldsymbol{m}$ is a difference between two random variables, the inverse-variance weight for each row of the system is given by:

$$w_{k,k} = \frac{1}{\mathbb{V}[L_i(p) - L_j(p)]} = \frac{1}{\mathbb{V}[L_i(p)] + \mathbb{V}[L_j(p)]}\,. \tag{5.7}$$

Here, k indexes each row of Eq. (5.4) since $\boldsymbol{W}$ is diagonal. For instance, the second row corresponds to $k = 2, i = 1, j = 2$.

The detailed studies of the noise characteristics of cameras [3, 83], discussed in Chapter 3, indicate that real camera noise follows a compound Poisson-normal distribution. For the working range of pixels in most images, this can be approximated by zero-mean additive noise that follows a normal distribution. For simplicity, I will use the simplified per-gain, per-colour noise model given by Eq. (3.1) for $\mathbb{V}[Y(p)]$.

The real quantity of interest, however, is $\mathbb{V}[L(p)]$. Since $L(p)$ is a random variable obtained by applying the log transformation to $Y(p)$, its PDF does not have an exact expression. This is because the domain of $Y(p)$ includes negative values for which the log function is undefined. As I will show in Section 5.4.4, it is preferable to operate on a fraction of pixels that have the lowest relative variances, whose intensities tend to be much greater than 0. In Eq. (B.3) in the Appendix, I show how to approximate the variance when any random variable is transformed by an invertible function. Here, I apply the result for $Y(p)$ when it is transformed by the log function,

$$\begin{aligned} \mathbb{E}[L(p)] &\approx \log \mu_Y(p)\,, \\ \mathbb{V}[L(p)] &\approx \frac{\sigma_Y^2(p)}{\mu_Y^2(p)}\,, \end{aligned} \tag{5.8}$$

where $\sigma_Y^2(p)$ is the variance of the noise from Eq. (3.1) and $\mu_Y^2(p)$ is the expected pixel value, which can be approximated by $\mathbb{E}[Y(p)] = \mu_Y(p) \approx y(p)$. Again, this result is valid because I work with well-exposed pixels, that are approximately normal. After substituting the computed variance in Eq. (5.7), the diagonal weights become:

$$w_{k,k} = \left(\frac{\alpha y_i(p) + \beta}{y_i^2(p)} + \frac{\alpha y_j(p) + \beta}{y_j^2(p)} \right)^{-1}\,. \tag{5.9}$$

5.4.3 Camera noise calibration

While the inverse-variance weights help compute noise-optimal estimates of the exposure ratios, a fundamental limitation is their dependence on calibrated noise parameters α and β. In the considered noise model, these are camera and gain specific and may not be available at the time of HDR merging. Moreover, the quality of HDR reconstructions is highly sensitive to accurate noise parameters [3], motivating the need for methods that do not rely on them.

If noise parameters are unavailable or inaccurate, it is still possible to solve the weighted linear system by assuming that the static noise parameter β is 0 as we work with well-exposed pixels (see Section 5.4.4). Since the entries of $\boldsymbol{W}$ determine the relative importance of the different rows of Eq. (5.4), the common signal-dependent constant α can be eliminated as well. The new weights which no longer depend on calibration-sensitive parameters are:

$$w'_{k,k} = \left(\frac{1}{y_i(p)} + \frac{1}{y_j(p)} \right)^{-1} . \tag{5.10}$$

In practice, we can get away with using camera-independent $\boldsymbol{W'}$ instead of $\boldsymbol{W}$ from Eq. (5.9) as I will show in Section 5.5.1.

5.4.4 Reducing the linear system

The linear system given by Eq. (5.4) contains up to $M\binom{N}{2}$ equations corresponding to all pixels and all pairs of exposures in the stack. Solving such a large system may be impossible for large images or deep exposure stacks due to computation and memory limitations. However, this system is strongly overdetermined as only $N-1$ exposure ratios need to be estimated. Therefore, only a small percentage of equations, k, are needed for efficiently solving Eq. (5.6). Another reason for reducing the system is to eliminate underexposed pixels pairs because they do not satisfy the assumptions made in the derivations of previous sections. For example, dark pixels may be negative or poorly approximated by a normal distribution.

A logical reduction strategy is to select pixels pairs with the highest weights since they are least affected by noise. However, such a strategy is heavily biased towards longer exposures because pixel pairs that include these images will have the highest weights. The shorter exposures will then be poorly represented and, thus, poorly estimated. Another issue is related to the spatial location of bright objects in the scene, such as the Sun or other light sources. If a small fraction of pixels (say 5% of the pixels per image) are selected, all of them are likely to be concentrated in one portion of the image, corresponding to these bright objects. If those objects happen to be in motion, the exposure estimation will fail.



Figure 5.5: Consider the exposure multigraph in (a) where different coloured vertices and edges represent different pixels. The minimum spanning tree (MST) may connect pairwise exposures (b) or connect the longest exposure to all others (c). Algorithm 1 will provide different solutions for different inputs. For example, the MST for the blue horizontal tile near the Sun in (d) is the pairwise-connected spanning tree connected at pixels marked by the two triangles in plot (e). However, for the orange tile, the MST is given by the longest-exposure spanning tree at the pixel marked by the star in plot (e).

I propose two orthogonal design choices to balance the system of equations and handle both these biases in estimated exposure ratios.

5.4.4.1 Spatial balance: Tiling

To ensure that the linear system contains samples from all parts of a scene, the input image stack is split into $t \times t$ pixel tiles (16×16 in my experiments). We can then select a fixed number of pixel pairs from each tile and pixels from a few bright objects will not dominate the system of equations. The tiled processing provides a convenient way to vectorize the construction of the reduced linear system. Several tiles can be processed in parallel for faster execution with multi-core or multi-threaded systems.

The noise-based solution (see Eq. (5.9)) helps provide a robust estimation even though some tiles may contain noisy pixels corresponding to dimly-lit parts of the scene. Variance-optimal balancing of exposures is crucial for such tiles, as I will show in the next section.

5.4.4.2 Exposures balance: Spanning trees

The next objective is guaranteeing that all $N-1$ exposure ratios are correctly estimated. Within each spatial tile, pixels from all exposure pairs need to be included, including noisy short exposures despite their relatively smaller weights.

Consider the *exposure graph*: an undirected weighted multigraph (a graph that contains more than one edge between two vertices as shown in Figure 5.5a), whose vertices represent the N exposures and edges link pairs of co-located pixels from two exposures. Each edge corresponds to one row of Eq. (5.4), with its weight given by Eq. (5.9) or Eq. (5.10). The different coloured vertices and edges in Figure 5.5a represent different pixel locations. Reducing the linear system from Eq. (5.4) is then equivalent to removing edges from this dense multigraph.

The optimal subset contains $(N-1)k$ edges with the largest weights that connect all exposures in a balanced manner. Such a subset can be found by computing the k highest-weighted spanning trees of the multigraph, where the weight of a spanning tree is the sum of the weights of its edges. The previous work directly linked pixels from neighbouring exposures [20], resulting in pairwise connectivity as shown in Figure 5.5b. However, this solution is sub-optimal for some inputs (such as the orange tile in Figure 5.5d) because using edges linked to the longest exposures, as shown in Figure 5.5c), results in higher weights and, therefore better estimates.

An optimal solution would sequentially extract k MSTs using *Kruskal's* or *Prim's* algorithm. Better algorithms [36, 75] extract the k highest spanning trees more efficiently. However, explicitly creating a large multigraph and computing MSTs is both memory and computationally expensive. Below I show that the optimal solution can be found by a simpler greedy algorithm.

Algorithm 1: Greedy MST algorithm

input : $y_1, \ldots, y_N$ (Images sorted by exposure time)
output: MST (Maximimum spanning Tree)

```
1  Function greedyMST(y[ ])
2      N ← length(y)
3      MST ← [ ]
4      for i ← 1 to N − 1 do
           /* Iterate over all exposures starting from the shortest          */
5          mask ← isValid(y_i) and isValid(y_{i+1})
6          p_* ← maxWeight(y_i[mask], y_{i+1}[mask])      // Location of the highest
            weighted edge connecting images y_i and y_{i+1}
7          for j ← N to j + 1 do
8              if isValid(y_j(p_*)) then
9                  MST.addEdge(i •—• j)
10                 break
11     return MST
```

Greedy MST solution Start by extracting all valid pixels — those that are unsaturated and sufficiently above the noise floor. Iterate through all exposures from the shortest to the longest. For each exposure i, identify p_*, the pixel location that contains the highest-weighted edge between exposures i and $i+1$. This is typically the brightest pixel that is not saturated in exposure $i+1$. Then, find the longest exposure $j > i$ in which pixel p_* is not saturated. Add an edge between i and j. By selecting the longest exposure, we ensure that the weight of the edge is maximised.

This procedure, summarised in Algorithm 1, can be repeated k times to extract the k highest-weighted spanning trees without explicitly constructing a graph. If all longer

exposures $y_N(p_*), y_{N-1}(p_*), \ldots y_{i+2}(p_*)$ are invalid due to saturation, the solution reduces to pairwise connectivity depicted in Figure 5.5b.

Note that edges of the MST do not need to come from the same pixel location. Consider the 1-dimensional blue tile in the tonemapped image in Figure 5.5d and its horizontal scanline in Figure 5.5e. The MST is given by two edges: an edge between the medium and long exposures and an edge between the medium and short exposures (marked by triangles in the plot). These are the brightest unsaturated pixels in the long and medium exposures. The MST for the blue tile is thus the pairwise connected spanning tree. The limited dynamic range of the orange tile means that a single pixel, marked by the star in Figure 5.5e, aligns all three exposures for the orange tile, resulting in the spanning tree in which the longest exposure is linked to all other exposures.

5.4.5 Handling outliers

The weighted least-squares solution Eq. (5.6) using rows of the reduced system will provide an accurate estimate only if there is no movement in the scene and all pixels are well aligned. If there is motion across exposures, for example, due to camera shake or object motion, those pairs of pixels will affect the estimates of exposure ratios, as shown in Figure 5.6. Notice the artefacts surrounding bright regions such as the windows and red lights in Figure 5.6a due to camera motion. Here, the exposures are estimated using the original stack and then registered the images with homography alignment [154] before merging. Thus, any visible artefacts are due to incorrect exposure alignment and not due to ghosting.

The spatial tiling described in Section 5.4.4.1 ensures that motion in a few bright objects does not adversely affect the estimated exposures, i.e., the described procedure is robust to localised object motion. In case of widespread pixel misalignment (e.g., due to camera motion), reasonable exposure estimates can be provided if the scene contains uniform image patches. *Usable* tiles are identified by collecting k MSTs within a tile and solving the smaller per-tile system of equations separately. If the predictions of a tile deviate too much from EXIF values, that tile is treated as an outlier and excluded the corresponding MSTs in the final system.

The added overhead of solving per-tile linear systems is negligible because each system contains only k equations ($k = 50$ in my experiments). This outlier removal is much faster than an iterative algorithm such as the one described in [20].

5.5 Results and applications

The primary application the described method is merging exposure stacks with varying exposure times or gains. I first validate results on synthetically generated stacks, for



(a) Regular $\hat{\boldsymbol{e}}_{\mathbf{WLS}}$ **(b)** $\hat{\boldsymbol{e}}_{\mathbf{WLS}}$ with outlier removal

Figure 5.6: For an exposure stack containing some spatially misaligned pixels, solving the weighted linear system Eq. (5.5) results in inaccurate estimates. In this scene, camera motion causes pixels of the building to be misaligned due to camera motion, producing artefacts near the windows in Figure 5.6a. However, pixels from other areas, such as the sky, can still be utilised for exposure estimation. Figure 5.6b demonstrates the advantage of using the iterative procedure described in Section 5.4.5 to improve robustness to outliers.

which we know the ground truth, and then show qualitative comparisons on real captures. Finally, in Section 5.5.4, I show how the proposed method can improve the estimate of the MTF of a display. All results use $\lambda = 10$ for the Tikhonov penalty term described in Eq. (5.6).

5.5.1 Validation on a synthetic dataset

First, I rely on synthetic exposure stacks to compare the accuracy of different methods for exposure estimation. For this, I simulated HDR exposure stacks using the noise parameters of the *Canon PowerShot S100* with exposure times [1/64, 1/8, 1, 8] (in seconds). All images were quantised to 14 bits to match the bit-depth of the camera, and the experiment was

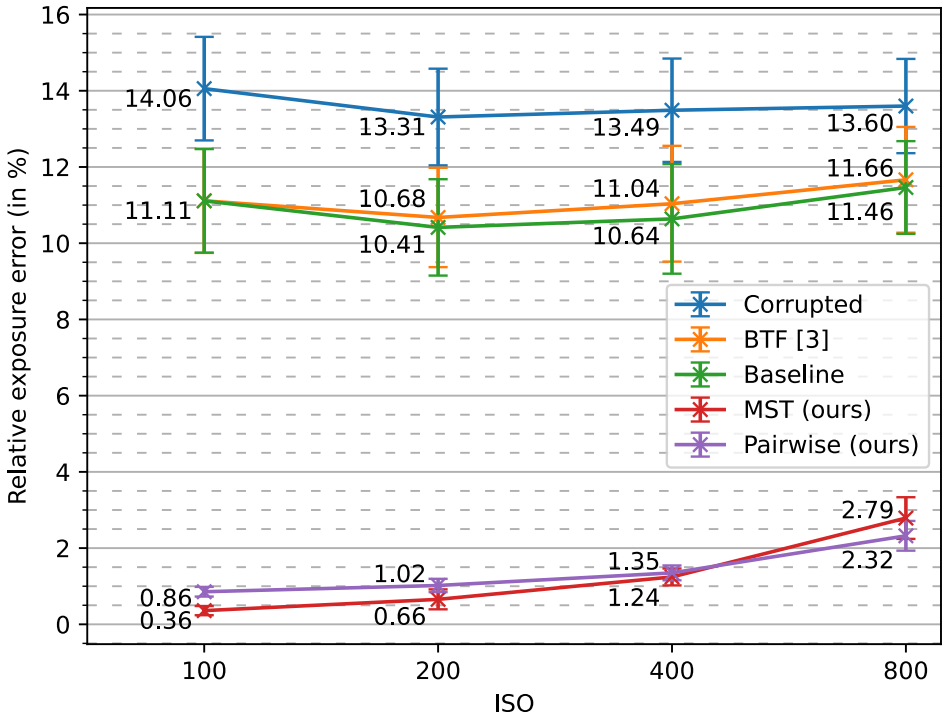


Figure 5.7: Accuracies of exposure estimation methods with increasing ISO computed by simulating captures on a large existing HDR dataset [37]. Error bars indicate one standard deviation of uncertainty computed across all images. The blue line represents corrupted exposures according to Eq. (5.11), while the baseline, in green, was obtained by directly estimating exposures by solving a system based on Eq. (5.1).

repeated for ISO settings between 100 and 800. All captured images were quantised to 14 bits to match the bit-depth of the camera. Source HDR images were taken from the Fairchild photographic survey [37] containing 182 scenes at a resolution of 1888×1280. All simulated captures are linear related to the HDR image, i.e., I do not apply a CRF since it is likely to introduce artefacts. Noise is simulated using parameters of calibrated cameras as described in Chapter 3.

Before merging the images, all exposure times were intentionally corrupted by introducing a small amount of normally distributed noise:

$$e_i' = e_i + \eta_i \quad \text{where } \eta_i \sim \mathcal{N}(0, e_i \times 0.15)\,. \tag{5.11}$$

The relative standard deviation factor 0.15 was selected to match real camera EXIF errors shown in Figure 5.2.

Figure 5.7 plots the relative root mean squared error (RMSE) (in percent) of the exposure ratios, for different ISO levels. Note that we can compute such relative errors only for synthetic datasets. I include results of a simple baseline by solving a linear system that realises Eq. (5.1) without relying on a noise model. Note that the baseline includes other components described in Section 5.4, such as spatial balance via tiling and exposure

balance via MSTs. The blue line shows the error of noisy exposures according to Eq. (5.11). The synthetic exposure stacks represent a challenging estimation problem with exposures three stops apart. As a result, the simple baseline (green) and the histogram-based brightness transfer function (BTF) method [20] (orange) are unable to provide good results due to the adverse impact of pixels in shorter exposures. A key limitation of both methods is not explicitly modeling camera noise, which we address in our formulation.

Figure 5.7 depicts results for the two versions of our weighted least-squares solution (Eq. (5.5)), showing both pairwise connectivity (in red) and the greedy MST heuristic (in purple). Both our solutions use weights that model the noise characteristics of the camera. Thus, they result in better estimates with smaller errors. Overall, the greedy MST heuristic is the best-performing estimator of exposure ratios.

Execution times Histogram matching for high-resolution images is computationally expensive, resulting in an average execution time of 2.11 seconds. Further, I observed that the expensive iterative procedure for removing outliers is needed for good estimates increasing time to 8.54 seconds. The proposed tiled reduction (Section 5.4.4.1) results in a significantly faster average execution time of 0.265 seconds for pairwise connected exposures and 0.29 seconds for the greedy MST solution. The reported times were computed on an *Intel i7-8700 CPU*.

5.5.2 Performance on real captures

Next, I show the advantage of the proposed exposure estimation over naive merging using EXIF data and compare it with the histogram-based BTF approach. The RAW images have a bit-depth of 14 and are linearly related to the scene radiance. I did not apply CRF or tone-mapping for any image shown in this section. Thus, all artifacts visible are due to incorrect exposures. All results in this section are computed with calibration-free weights given by Eq. (5.10).

Banding artefacts are visible for scenes containing a smooth gradient at the transition point when one of the input exposures saturates. This frequently occurs very close to bright light sources such as the Sun during the day or street lights at night, as indicated by red arrows in Figure 5.8a. Pixels close to light sources tend to be unsaturated only in shorter exposures where other parts of the image are strongly affected by noise. Aligning exposures under these conditions is challenging, resulting in the poor performance of histogram-based BTF weights [20]. Thus banding artefacts persist at the same locations in Figure 5.8b (top and bottom) or are introduced in other locations in Figure 5.8b (middle). By accounting for heteroskedastic noise with inverse-variance weights in Figure 5.8c, all the exposures can be correctly aligned to produce banding-free results.

Even in the absence of point light sources, banding can appear at natural smooth



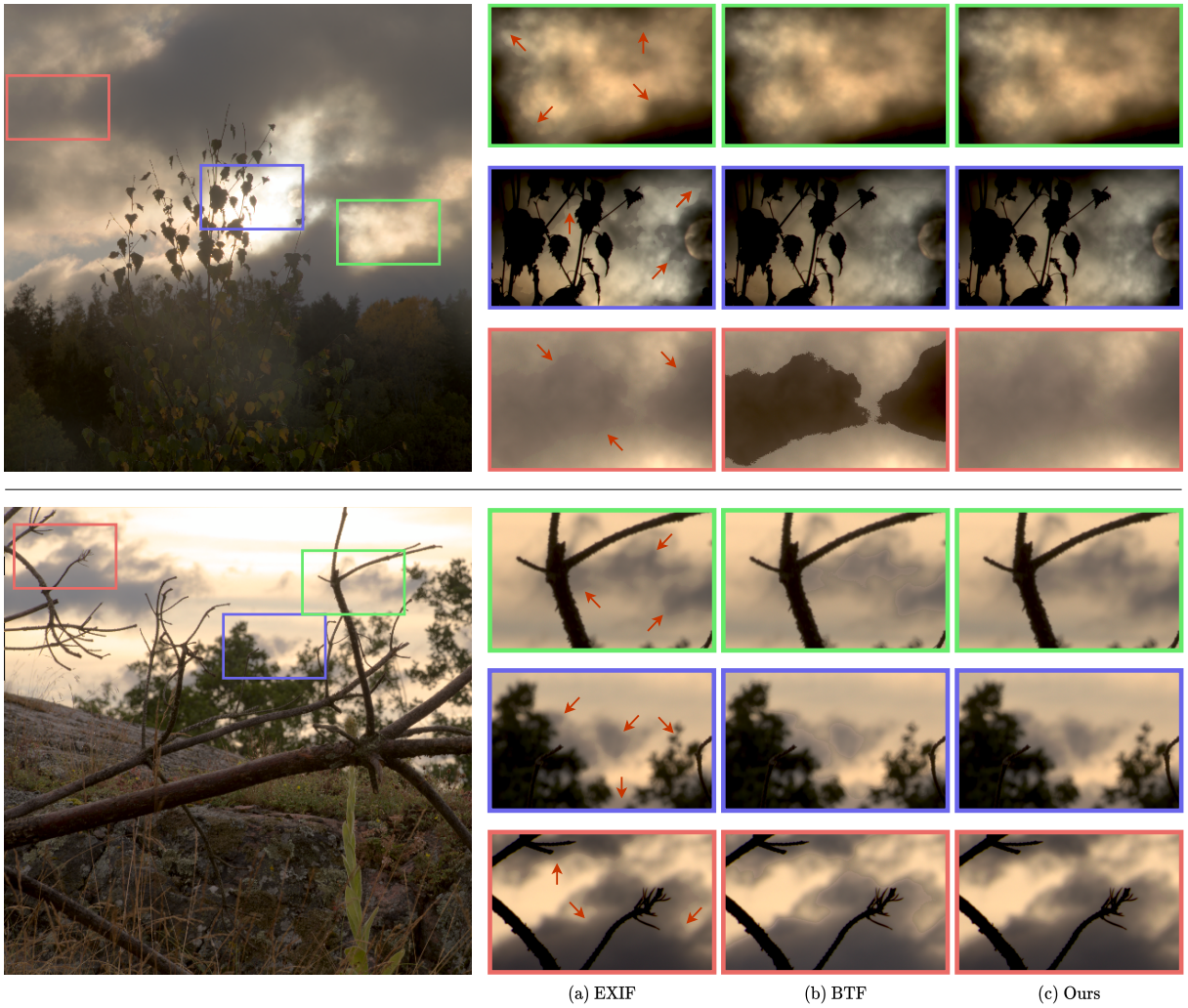
(a) EXIF **(b)** BTF [20] **(c)** $\hat{\boldsymbol{e}}_{\mathbf{WLS}}$ (Ours)

Figure 5.8: Zoomed in patches of light sources, appropriately exposed to highlight banding close to the source due to inaccurately reported EXIF exposures (first column). With histogram weights based on BTF (middle column), only some exposures can be aligned. By modelling camera noise, all exposures are simultaneously aligned (right column).

gradients or due to defocus blur, as shown in Figure 5.9. Large regions of the images are well-exposed for such scenes, and most methods work reasonably well. However, BTF weights can still sometimes fail to recover the exposure ratio for shorter exposures (for example, see the red patch in the middle column of Figure 5.9). The proposed noise-model motivated approach consistently aligns all exposures to produce banding-free reconstructions across the various scenes.

5.5.3 Deghosting for scenes with motion

When multi-image stacks with misaligned exposures are fused, the resulting HDR reconstructions often exhibit accentuated ghosting artefacts. Even recent deep learning-based methods [71, 160, 168] are unable to account for inaccurate exposure ratios. Figure 5.10b illustrates the artefacts that arise when exposures are corrupted according to Eq. (5.11). It is important to note that these artefacts are in addition to the ghosting artefacts visible in Figure 5.10a. The proposed algorithm, which includes tile-based outlier detection to account for scene motion, effectively aligns these exposures.



(a) EXIF (b) BTF (c) Ours

Figure 5.9: Fixing banding in smooth image gradients when images contain sufficient well-exposed pixels to populate the linear system. The proposed inverse-variance solution (last column) is more robust and produces better reconstructions.



(a) Aligned exposures **(b)** Corrupted exposures **(c)** Reference

Figure 5.10: Comparison of HDR reconstructions of an attention-based deep network [160] with aligned (first column) and corrupted (middle column) exposures. These insets are from the HDR deghosting dataset [71], which the deghosting network was trained on.

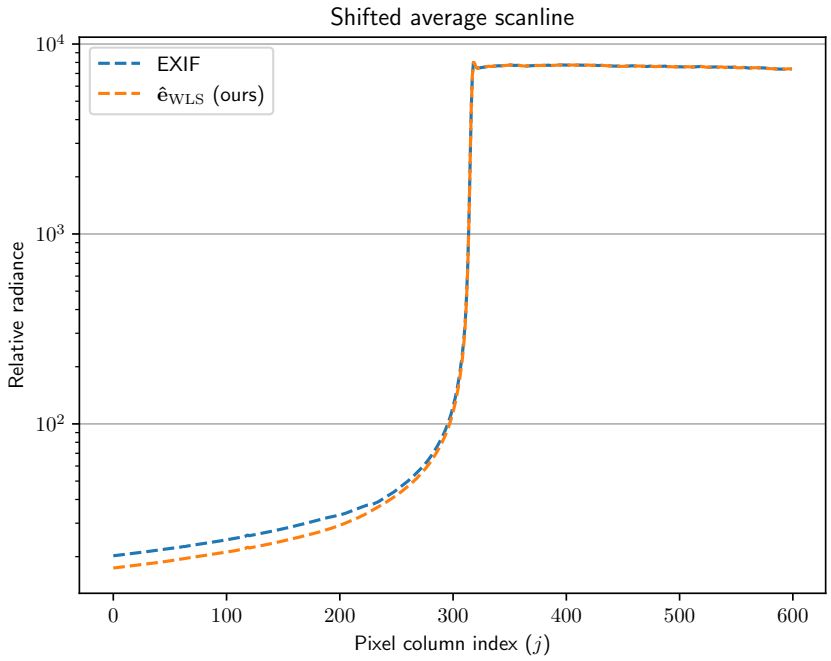


Figure 5.11: Edge profile of an HDR display computed from a slant edge captured with an exposure stack. Errors in EXIF metadata lead to deviation from the actual edge spread function, affecting MTF estimation. Notice the banding artefacts close to the captured slant-edge when the image was merged with EXIF exposures.

5.5.4 Measuring display MTF

Digital cameras are often used as inexpensive light measuring instruments, for example, for measuring the spatial characteristic of an electronic display, such as spatial uniformity of a modulation transfer function (MTF). Inaccuracies in pixel values due to exposure misalignment may lead to errors in the estimated display characteristics. Here, I demonstrate the need for exposure alignment for measuring display MTF.

The MTF measurement involves taking an image of a slanted edge [138], where one side has its pixels set to 0 and the other side to the maximum pixel value. Modern displays can reach very high contrast levels, so such a slanted-edge image can be faithfully captured only using an exposure stack. The inaccuracies of exposure times can easily introduce bias to the measurements.

Figure 5.11 shows the edge profile computed with (blue) and without (orange) exposure estimation. The difference in reconstructions is due to inaccurate EXIF exposure values, visible as banding in the merged image. This translates to the bias at low-radiance pixels (pixel index between 200 and 300) in Figure 5.11 and shows that errors in exposure times can easily lead to biases in the final MTF estimates.

5.6 Summary

In this chapter, I proposed an approach to estimate relative exposures directly from images of an HDR exposure stack instead of relying on inaccurate camera EXIF metadata. The problem can be formulated as a weighted linear system of equations with Tikhonov regularisation. My formulation considers the noise characteristics of the simplified camera from Chapter 3, Section 3.1.1 to derive weights. I showed that good performance on a large number of scenes is possible without the need for camera- and gain-specific noise parameters.

I also describe an efficient approach based on the union of maximum spanning trees of the exposure multigraph to reduce the size of the system for large images. When applied to multi-image HDR stacks, the proposed method eliminates banding artefacts at smooth image gradients close to light sources or due to defocus blur. The exposure-aligned images are closer to physical quantities, making this work essential for using a camera as a light-measurement instrument.

Chapter 6

Application: Perceptual realism

> "Seeing-is-believing is a blind spot in man's vision"
>
> *R. Buckminster Fuller*

Estimating accurate high dynamic range radiance values from a stack of exposure and gain modulated images was the focus of the last two chapters. I detailed how to reconstruct artefact-free images that capture rich HDR scenes.

Such accurate capture is a sine qua non for perceptual realism—producing imagery indistinguishable from real-world 3D scenes. In this chapter, I introduce the capture and display apparatus we built to deliver all necessary visual cues for perceptual realism. This was a team effort, with several people involved. I restrict the contents of this chapter to the parts of the pipeline I was involved in such as calibrating the display (Section 6.2.1), designing the real-scene box (Section 6.2.3), and capturing and reconstructing the HDR light field (Sections 6.2.4 and 6.3.1). However, I also include some other details crucial to better present the context of the project.

The end-to-end capture-render-display system allows side-by-side comparisons with a real object. Thus, it serves as the ultimate validation of the HDR reconstruction methods described in Chapters 4 and 5. Please see our paper [177] and the project page for further details: `https://www.cl.cam.ac.uk/research/rainbow/projects/hdrmfs/`.

6.1 Introduction

We designed the display and associated imaging system to capture and reproduce a small three-dimensional object positioned 50-80 cm from the eye. The setup also includes a real scene box aligned with the display for direct comparisons between virtual and real scenes. The main objective of the design of our high-dynamic-range multi-focal stereo (HDR-MF-S) display is to maximise the visual quality and realism of the displayed images

for all the following capability dimensions: physical luminance, dynamic range (contrast), colour gamut, binocular and focal depth cues. The goal is to deliver all these capabilities altogether with sufficient qualities rather than focusing on maximising a single one.

While there are several fundamentally distinct approaches to 3D display architectures, the current state-of-the-art of these technologies does not achieve the field of view, colour accuracy, resolution, or dynamic range required for perceptual realism. Reproducing a 4-dimensional light field of sufficient size and quality with these technologies requires control over billions of pixels, which is currently infeasible. However, by either stabilising or tracking the viewing position, the subspace of a light field that needs to be reproduced is much smaller, making it possible to build a display of the required capabilities.

6.2 Apparatus overview

One approach to producing the required light distribution, given either fixed or known eye position, is to use a stereoscopic multi-focal display [5]. In such displays, the eye sees the sum of light from multiple superimposed planes at different focal distances. Such displays can effectively drive accommodation to any point between the planes if the plane separation is small enough ($\sim$0.6 D to $\sim$0.9 D) [98, 99], while retaining the desirable capabilities of conventional displays (resolution, colour gamut). Moreover, this uncomplicated design, without any refractive or diffractive optical components in the viewing path, generates images without additional optical distortions. This is in contrast with varifocal displays [31] or near-eye light field displays [66], which are likely to introduce noticeable aberrations.

The apparatus comprises three main components as shown in Figure 6.1: a Wheatstone stereoscope with four high-dynamic-range displays and two focal planes; a real-scene box in front of the observer that is seen through a pair of beam-splitters; and a motorized camera slider capable of capturing dense horizontal light fields of the real-scene box. In this setup, a small physical scene is arranged in the real-scene box. This box normally faces the observer but can be rotated to face the camera rig in order to capture its light field as shown in Figure 6.1. When facing the observer, the real scene and its rendered counterpart are spatially superimposed. We can instantly switch between the real and displayed scenes by controlling the lights in the real-scene box and the display.

6.2.1 HDR displays

The key feature of our display is the capability of reproducing a high dynamic range, with a peak luminance of 3000 cd/m^2 and a black level much below 0.01 cd/m^2. Such a low black level practically eliminates any stray light in areas of an image that should remain black. The HDR reproduction is delivered by four projector-based dual-modulation displays, similar in design to those used in one of the first HDR displays [142].

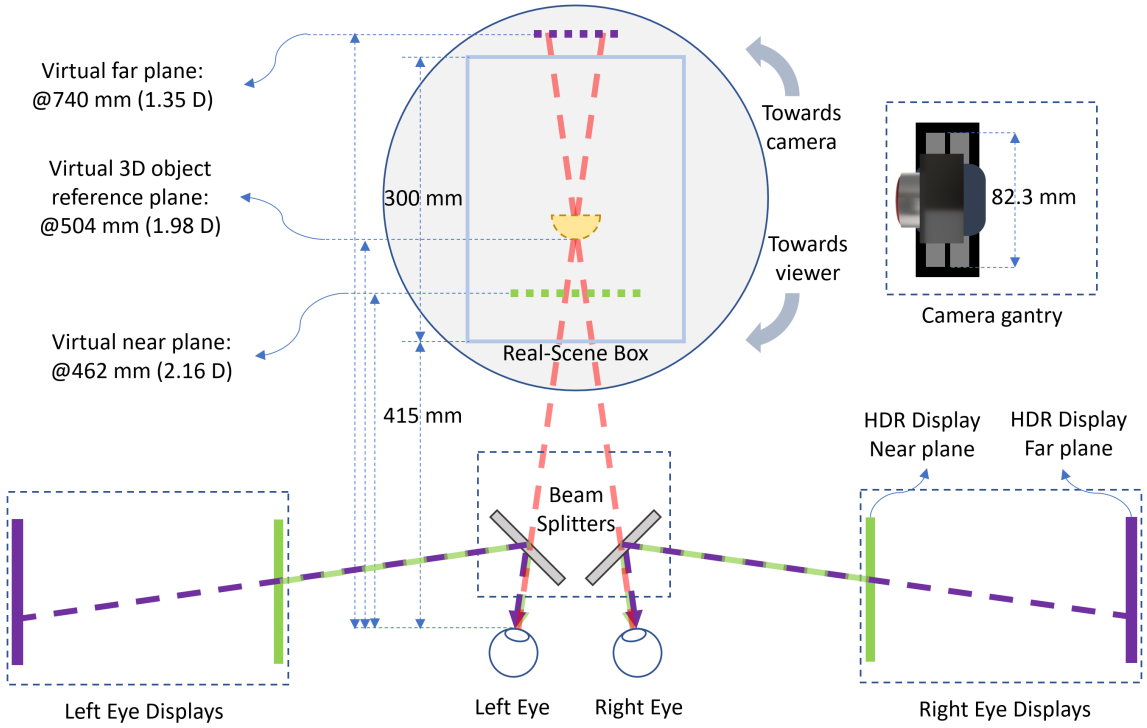


Figure 6.1: Schematic of the high-dynamic-range multi-focal stereo display apparatus. (Note that to simplify the schematic, not all the folding mirrors and beam splitters are shown.) The apparatus creates two image planes (green and blue dashed lines inside the real-scene box) per eye and the observer sees respective images via beam splitters. The real-scene box is observed through the same beam splitters. The real-scene box is on a manually rotating platform moving toward a fixed capturing position or a fixed display position. The camera gantry is on another manually movable platform (not shown in the figure) which can move towards or away from the real-scene box allowing coarse adjustment of the field of view.

The software for controlling each display implemented the standard two-spatial-modulator factorisation algorithm [142] running on a GPU. However, we took special care to achieve accurate geometric alignment and high colour accuracy. The geometric alignment was achieved by taking images with a DSLR of a calibration pattern (a grid of points) displayed separately on the LCD and the DLP and then aligning them using homography and mesh-based warping. The point-spread function of the DLP was measured for the same grid of points and approximated with a Gaussian function. The colourimetric calibration was achieved by measuring the colour ramps with a spectro-radiometer (Specbos 1211) and fitting a gamma-offset-gain model to the LCD panel and using a dense look-up table for the DLP. The dense look-up table was necessary as the response of the projector was non-monotonic after removing the colour wheel.

6.2.2 Focal planes and optics

To vary the focal distance, similar to a multi-focal display [5], our display can generate images at two focal planes, at the distances of 462 mm (2.16 D) and 740 mm (1.35 D) from the viewer, providing a 0.81 D separation between the planes. The separation was selected to ensure that the images shown on two planes provide cues for accommodation for any distance between the two planes [98]. Such distances also ensure a resolution of at least 85 pixels per degree for the observer. These distances are adjustable by moving the HDR displays on their mounting rails.

6.2.3 Real-scene box

The real-scene box has the inner dimensions of 200 mm × 160 mm × 300 mm (width × height × depth)*, matching the physical size of the LCD screens and accommodating the optical separation distance between the far and near plane HDR display apertures. It was made of black acrylic, which was covered on the inside with high absorption blackout material. The ceiling was fitted with an LED array light source with 225 individually addressable RGB LEDs (WS2812B). The real-scene box was fixed on a platform, supported by ball transfer units, allowing it to be freely rotated towards the observer for viewing, or towards the camera for light-field capture, as shown in Figure 6.2.

The real-scene box also included eight cross-shaped calibration markers placed outside the box, as shown in Figure 6.2-left. The markers were used as reference points to register the camera pose. The markers were carved on the two foam boards and illuminated by an RGB LED (WS2812B) with a diffuser to improve their visibility.

*Note the these dimensions have since been updated and are currently larger.

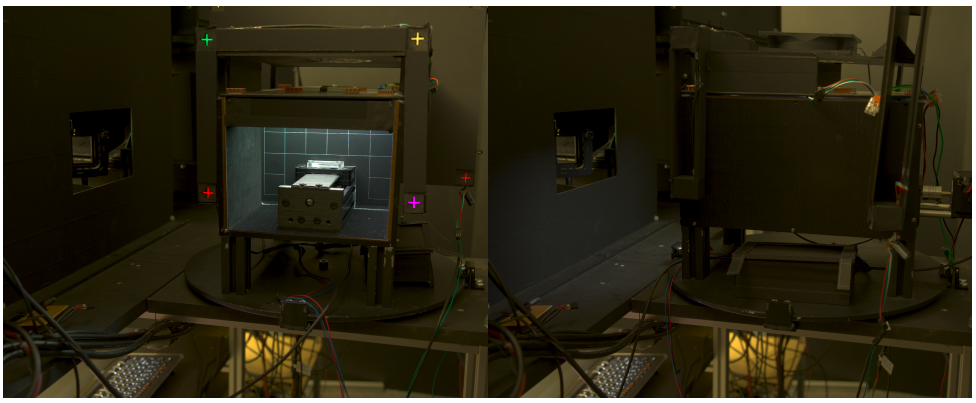


Figure 6.2: The front (facing the camera) and side (facing the viewer) view of the real-scene box.

6.2.4 Data camera for light field capture

To capture a horizontal light field of the real-scene box, we mounted a Sony α7R3 mirrorless camera with a Sony G OSS zoom lens (focal length 24-105 mm) on a motorised camera slider at a distance of 415 mm from the real-scene box, similar to the distance from the viewing position to the real-scene box. The distance was larger than that of the viewing distance to increase the depth-of-field of the captured scene. The camera slider traversed a baseline of 82.3 mm with an accuracy of 5 µm.

6.3 HDR-MF-S imaging & rendering system

To achieve perceptual realism, we need not only a display capable of reproducing all relevant cues, but also an imaging and rendering system, which can capture a real scene and display it with sufficient quality. Most importantly, the rendered scene should match the viewpoint of the observer. Our system is currently limited to processing scenes of relatively simple or known geometry but can handle complex non-Lambertian materials and high-dynamic-range illumination.

Figure 6.3 shows a diagram of the HDR-MF-S imaging and rendering system. We start with the capture of a horizontal HDR light field, which is colour-calibrated for the spectra of the scene illumination. Next, we employ photogrammetry to perform 3D reconstruction and estimate camera matrices. After that, we apply a differentiable rasteriser to register a proxy mesh of the main object with its silhouette in each HDR light field image, so we can project the fitted mesh to each light field image to obtain a view-dependent UV map and texture. Finally, we integrate lumigraph view synthesis with linear depth filtering [5] to render the final scene on our HDR-MF-S display. Note that the differentiable rasteriser and view synthesis modules are not discussed here, as they were developed by my collaborators.

We found lumigraph to be the most suitable 3D representation for our purpose as it

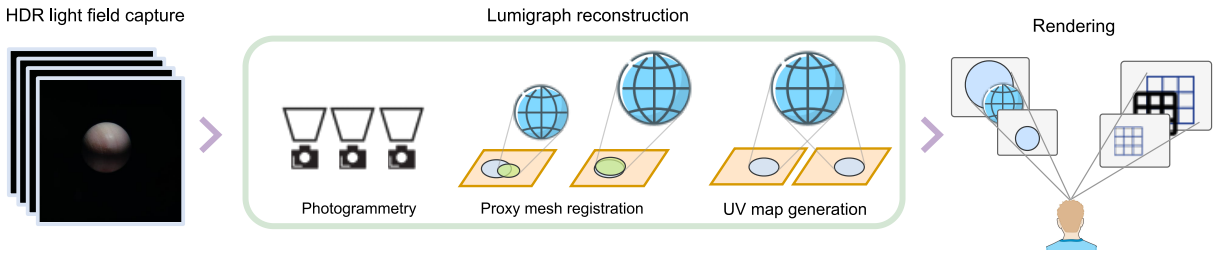


Figure 6.3: Pipeline for reconstructing and rendering 3D scenes with a lumigraph [48]. The system involves merging multi-exposure image stacks from multiple viewpoints, camera pose estimation with photogrammetry, proxy mesh registration with differentiable rasterisation, lumigraph view synthesis with view-dependent UV maps, and multi-focal rendering with linear depth filtering.

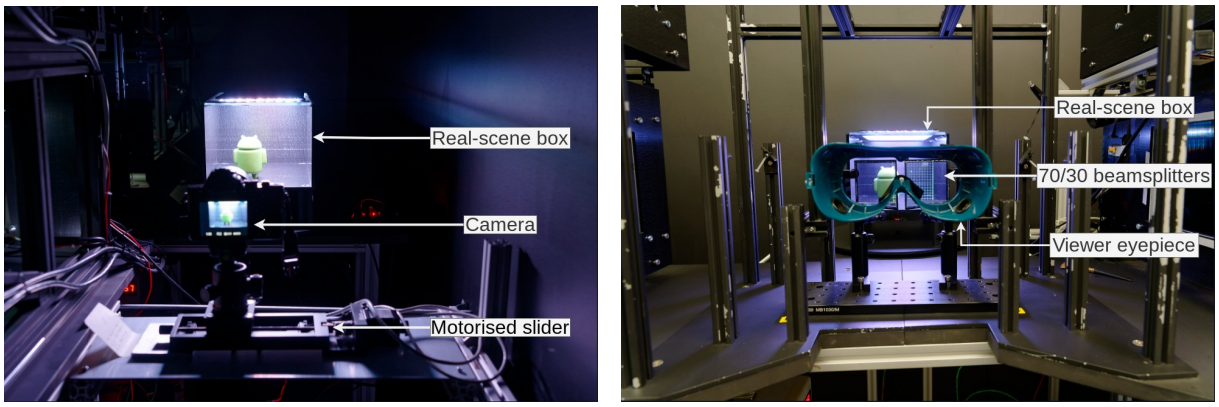


Figure 6.4: The side (left) and front (right) views of our capture-render-display system. The camera mounted on a motorised slider captured an HDR light field. After HDR reconstruction and demosaicing, colour calibration compensated for the spectral transmission of the 70/30 beam-splitters that observers see objects through (right).

models non-Lambertian surfaces, is robust to processing high-resolution textures, and performs rendering in real-time. We have also experimented with dense light fields, either captured or reconstructed using neural radiance fields [112], but they did not match the quality required for perceptual realism, which poses a lower tolerance for artefacts (such as blur, noise, and distortion) and a higher demand for capacity to process high-resolution (8k) images. Instead, we combine photogrammetry and differentiable rendering to align known geometry with the captured HDR images to reconstruct a lumigraph.

6.3.1 HDR light field capture

Using our data camera discussed in Section 6.2.4, we first captured a high-resolution (7360×4912 pixels) light field consisting of 16 views with a separation of 5 mm between them. The camera was moved by a custom motorised slider (Figure 6.4-left). For each camera view, we captured an HDR exposure stack consisting of up to five RAW images spaced two stops apart in exposure time and ISO of 100.

Merging and demosaicing We merged the RAW images to increase the dynamic range and decrease noise. To minimise noise in the reconstructed images, we used the Poisson photon noise estimator (Eq. (4.10), Chapter 4). Next, we demosaiced the merged images using the *DDFAPD algorithm* [109] due to its superior quality sans aliasing and zippering artefacts, as well as its efficient implementation [†]. To avoid any artefacts, it's essential to accurately identify saturated pixels for all colour channels during the HDR merging process. Thus, we merge each stack of images before demosaicing.

Other digital operations We did not need to additionally denoise our images since HDR merging provided us with pixels that were free of noise. All other steps of the digital pipeline described in 2.1, such as tone mapping and compression were skipped to retain the linear relationship between the scene radiance and captured pixels. We also avoided auto white balancing algorithms as observers are sensitive to small inaccuracies in colours.

Colour calibration To calibrate for colours, we measured the spectra of a colour checker passport (X-Rite) positioned inside the real-scene box with a spectroradiometer (Specbos 1211, Jeti). Then, we compensated for the measured spectral transmission of the 70/30 beam-splitter (shown in Figures 6.1 and 6.4-right) and recovered trichromatic coordinates using the CIE XYZ 1931 colour matching functions. The XYZ colour coordinates were used to find the 3×3 matrix that transforms from native camera linear RGB space into CIE XYZ and which results in the smallest RMSE of DeltaE 2000 colour differences. The white patch in the colour checker was used for white balance. Finally, we applied the matrix to convert the merged HDR images from their native camera linear RGB space to the BT.709 space used by our display.

6.4 Results

Although it is difficult to convey the three-dimensionality and colour appearance of the scenes shown on our display using photographs, I include a few in this section to demonstrate some of its unique capabilities. Images of several displayed and real objects were captured using a Sony α7R3 camera with a 55 mm lens (SEL55F18Z).

Figure 6.5 demonstrates a close perceptual match between the real and virtual objects achieved by our system. The accurate spatial alignment of the virtual object overlaying the physical object demonstrates the perceptual match in geometry (Figure 6.5(a)). The display also achieved a close match in appearance and shading (see the overlapping shadows and specular reflections in Figure 6.5(a) and the side-by-side comparison in Figure 6.5(c)).

[†]Code maintained at `https://www.colour-science.org/colour-demosaicing/`

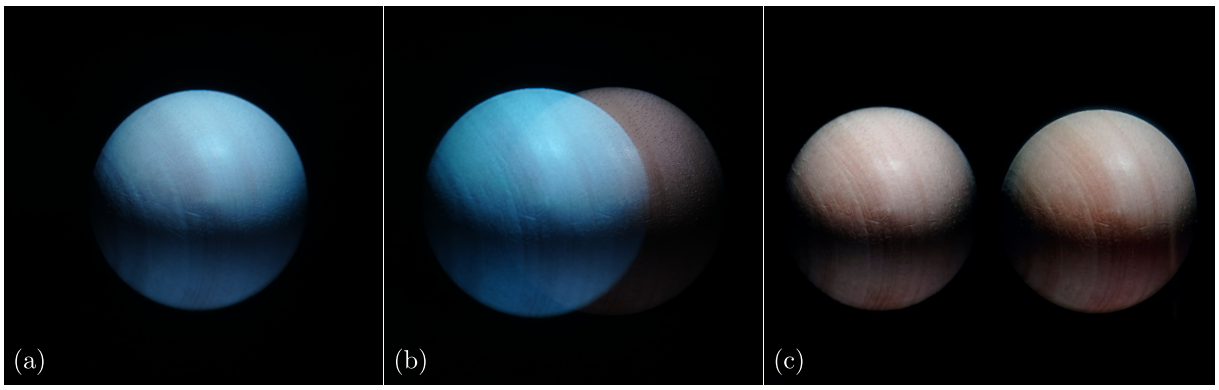


Figure 6.5: (a) Photograph of a virtual object displayed on top of the real object. Changing the hue of the texture demonstrates a mixed-reality effect. (b) The real object can be seen more clearly with the displayed object slightly shifted away. (c) Photograph of the displayed object (right) next to the real object (left). The small white strip visible on the bottom right corner of the right object is not a display artefact but a reflection from the background.

With such a level of precision, it becomes possible to show many mixed-reality effects such as changing the hue of the physical object without changing the shadows or textures.

To evaluate the resolution limit and the aforementioned incorrect defocus blur (when the virtual object is placed between the two focal planes) of our display, we reproduced a 1951 USAF resolution test chart (ThorLabs, R3L3S1P, positive, 3" ×3") and photographed it in comparison with the physical chart (Figure 6.6). We built a custom lightbox to illuminate the chart from the back, producing a high-contrast resolution pattern. We displayed either the real or rendered virtual chart at one of three distances[‡]: 500 mm (near), 577 mm (middle), and 654 mm (far). The camera focus was also set to one of these distances. To reduce the Moiré pattern resulting from the interference of the LCD and camera sensor pixel grids, we reduced the aperture to F16 and processed the images using *DxO PhotoLab 4.3.0* with only Moiré filtering enabled. Note that the Moiré pattern was not visible to naked eyes. Assuming that the resolution limit is the point at which the lines blend together and cannot be regarded as separate, our display can reproduce up to 4.0 $^{\text{lp}}/_{\text{mm}}$ at 500 mm (0.58 $^{\text{lp}}/_{\text{arcmin}}$), 2.83 $^{\text{lp}}/_{\text{mm}}$ at 577 mm (0.48 $^{\text{lp}}/_{\text{arcmin}}$) and 4.0 $^{\text{lp}}/_{\text{mm}}$ at 654 mm (0.76 $^{\text{lp}}/_{\text{arcmin}}$). This shows a dip for the middle distance, at which the displayed image is a superimposition of two defocused focal planes (Figure 6.6, $2^{\text{nd}}row$, 2nd column).

6.5 Summary

The main objective of our work is to build an end-to-end system that can acquire the texture and geometry of a small 3-dimensional object and reproduce it faithfully with all the necessary visual cues on a display. Being able to do so is an important step for

[‡]For this evaluation, the near focal plane was moved close to the near distance and the far focal plane close to the far distance.

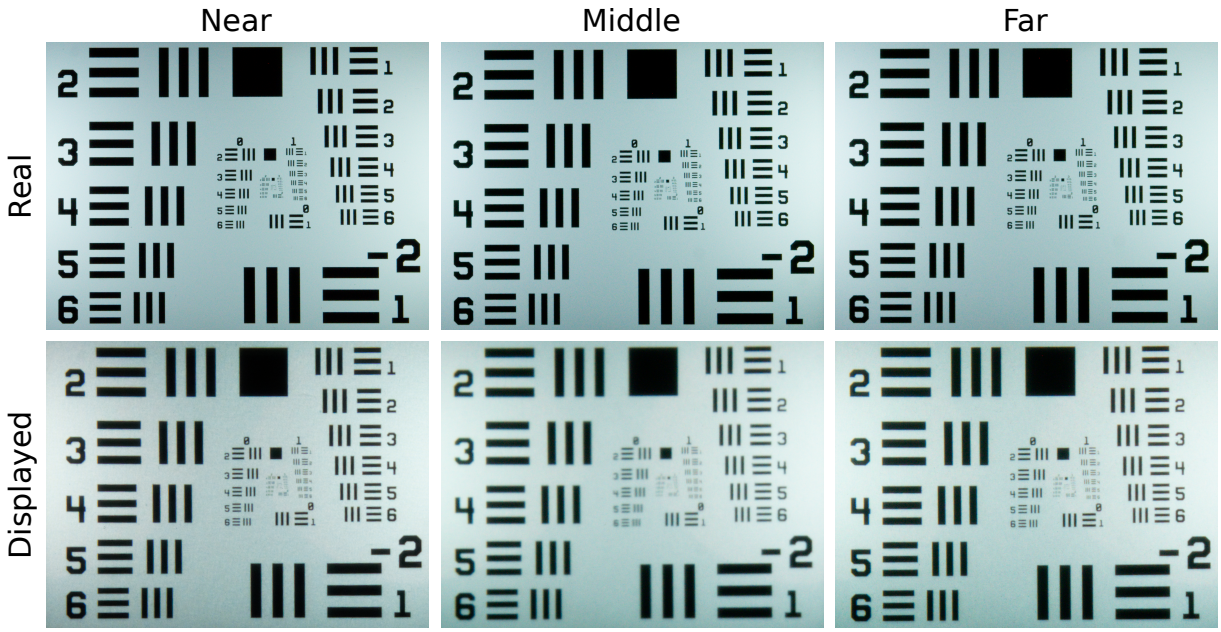


Figure 6.6: Photographs of a physical 2D resolution chart (top) in comparison with its displayed counterpart (bottom) placed at different depths (columns).

perceptually realistic graphics, in which the depicted imagery is indistinguishable from the real world. The exposure merging algorithms from earlier chapters played a key role in reconstructing the HDR light field of the objects.

We also designed an experiment to test whether participants can distinguish between real and virtual objects shown by our system. Please see the paper [177] for full details regarding the user-specific calibration and experiment procedure. The results suggest that our display can deliver virtual imagery that is in only 44% of the cases detected as different from its real-world counterpart.

A significant limitation of the system is its inability to render dynamic scenes. While the 3D reconstruction and rendering of such scenes pose a considerable computational bottleneck, acquisition also presents challenges. Quick acquisition involves capturing or reconstructing HDR details at real-time video frame rates. This will be addressed in Chapters 8 and 9.

Chapter 7

Application: Dataset for computer vision

Existing HDR datasets can be improved by correctly estimating scene radiance (Chapters 4 and 5). In this chapter, I describe HDR for computer vision (HDR4CV) *, one such dataset used to test the robustness of computer vision (CV) methods under adversarial illumination conditions. Since widely used CV benchmark datasets [18, 140] often contain images obtained in relatively "easy" conditions, in which the illumination is mostly uniform across the scene and there is little camera noise, methods that perform well on such datasets tend to fail under real-world, challenging illuminations. An increase in the dynamic range of a scene is particularly problematic for 8-bit images. The HDR4CV dataset automatically helps to determine the robustness of different CV methods.

The dataset is composed of video sequences captured for the same scene but under several different illuminations. The frames were captured from the same camera position with the same scene arrangement, while a set of artificial lights were configured to mimic one of four illumination conditions: (1) an "easy" uniform illumination, (2) a low-light night condition, (3) a high dynamic range (HDR) condition, and (4) a condition with a bright light source that induces strong glare. The main advantage of this approach is that the uniform condition produces reference labels for any CV method. Subsequently, these labels help determine the relative degradation in performance of the method under other illumination conditions. This saves us the painstaking work of manually labelling the dataset for each CV application. To demonstrate the utility of the proposed dataset, I show how several face detection, object detection, and optical flow methods are affected by adversarial illumination (Section 7.5.1).

*The dataset is publicly available at `https://doi.org/10.17863/CAM.71285`.

7.1 Related work

Publicly available image datasets serve as the main means of evaluating and comparing CV methods. Some established datasets, such as Middlebury stereo [140, 141] and optical flow datasets [12] contain well-illuminated images captured in controlled laboratory conditions. A similar trend can be observed for higher-level detection and recognition tasks, for which commonly used evaluation datasets are COCO [89] for objects and LFW [67], CelebA [93] and FFHQ [74] for faces. Although such datasets are highly influential and essential for evaluating CV methods, they attract criticism since they may not sufficiently reflect the varied illuminations found in real-world scenes.

In-the-wild datasets Very large datasets with a good variation of illumination conditions have been captured with monitoring cameras [94, 144] or car-cabin cameras [108]. However, such datasets are intended for a single application, and extending them to other applications requires tedious manual labelling (e.g. 11 man-months of work in some cases [94]). The HDR4CV dataset cannot match the size of these specialised datasets but can be used across diverse CV applications.

Multi-illumination datasets Scenes with the same composition but varying illumination can be captured using a motorised photographic flashlight on a camera by taking multiple images [117, 120]. The light bouncing off the walls and the ceiling illuminates objects from different directions, providing a large variation in illumination required for image relighting methods. Such an approach, however, is suitable only for indoor scenes without motion, results in a rather artificial structure of the reflected flashlight, and is unlikely to produce high-contrast HDR illumination. Instead, I carefully composed scenes with multiple light sources to achieve challenging illumination conditions.

Abdelhamed et al. [1] captured the smartphone image denoising dataset (SIDD), a dataset for testing denoising methods intended for smartphone cameras. They captured a series of images for 10 different static scenes, each under four conditions by varying the camera gain settings, colour temperature, and brightness of the light sources. In contrast, the HDR4CV dataset consists of video sequences with motion parallax, captured with a CV camera, in which the illumination was specifically designed to challenge a range of CV methods.

Tracking and detection under challenging illumination conditions Underexposed and saturated regions pose problems to object detection [10] and tracking methods [7]. To address this, Atoum [10] proposes illumination-aware CNNs to improve object detection. Alismail et al. [7] adapt illumination-invariant binary descriptors to achieve

Table 7.1: Configuration of lights used for different illumination conditions. The first two lights were controlled using the DMX512 protocol, allowing a value between 0-255. The remaining lights could only be turned on or off using relays connected to an Arduino board.

Light	Night	HDR	Glare	Uniform
DMX area	0	63	255	255
DMX spotlight	10	0	255	0
Photographic light	off	on	off	off
Tunnel/street LEDs	on	off	off	on
Background	on	on	on	on

photometric invariance in tracking. HDR4CV provides an evaluation dataset for these and other related problems.

Relighting Image-based relighting can be used to generate novel images of a scene under arbitrary illumination conditions [61, 166]. However, most relighting methods either require several images of the same scene captured under different lighting, or they are restricted to a single application, such as portrait relighting [146]. Since several images need to be captured anyway, we are better off capturing the images under the desired illumination and avoiding potential artefacts of the relighting method used.

Rendering Large datasets with automatically generated labels can be produced with computer graphics methods [19, 44, 156]. Game engines or offline rendering can be used to render photorealistic scenes in arbitrary illumination conditions. However, obtaining highly realistic rendering results, comparable to camera images, requires a substantial amount of effort by skilled professionals. Furthermore, rendered images tend to differ substantially from those captured by cameras. This may introduce a bias in training, validation and result in a model that underperforms *in-the-wild* [176]. This problem is addressed by domain adaptation methods [176], which involve training on a mixture of real and computer-generated images in a fully or semi-supervised manner or adversarial training on the source and target domains [158].

7.2 Capture setup

In this section, I discuss the common illumination setup shared across all scenes under different illumination conditions. Figure 7.1 depicts the various components of the capture setup.



Figure 7.1: The *Street* scene along with the capture and setup. Controllable lights are present on either side of the scene to simulate various lighting conditions. The right image shows the spotlight behind the scene, which shines through the diffuser and serves as a source of glare.

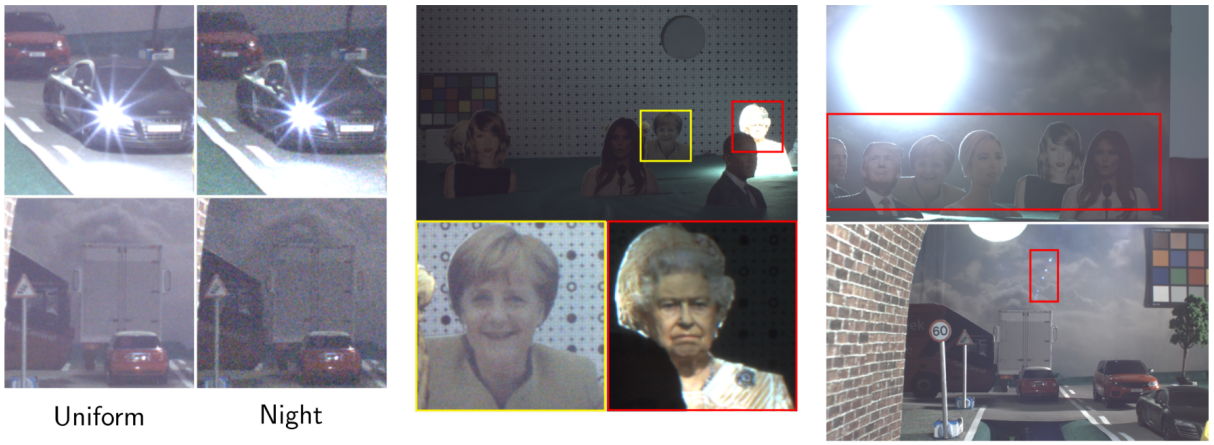
7.2.1 Lights

Each illumination condition was simulated by toggling or dimming a combination of lights according to the configurations detailed in Table 7.1. The lights included a compact spotlight (Cameo Q-SPOT 40 TW) and a photographic box light (Astora SF 120), both controlled over the DMX512 protocol using the QLC+ software [†]. *Selenium webdriver* [‡] was used to control QLC+ software using its web interface. Additionally, background illumination was provided by an LED video light (Neewer NL480). All the lights were tested with a custom high-frequency light meter to ensure that they were flicker-free up to 8.9 kHz.

Some example images from the captured scenes under each illumination conditions are shown in the columns of Figure 7.3. These conditions are described below:

Uniform For each scene, a uniform condition was obtained by manipulating the lights to illuminate as much of the scene as possible. The camera parameters were selected to produce well-exposed frames with low levels of noise. The CV methods tested in all experiments (Section 7.5.1) perform very well on images captured under the uniform condition. Thus, it serves as a reference condition for generating labels used to test other, more challenging illumination conditions. The advantage of such a framework is that the prevention of tedious, manual labelling of the data for different applications. Note, however, that the labels obtained are not ground truth labels and their only purpose is to

[†]Q Light Controller+: `https://www.qlcplus.org`
[‡]Selenium with Python: `https://selenium-python.readthedocs.io`



(a) Image patches of the night illumination captured with exposure time $t = \frac{1}{150}$ secs and gain $g = 16$, compared to a uniform illumination image from the same camera location.

(b) HDR image along with appropriately exposed insets of dark and bright regions. Notice the noise in the left inset and saturation in parts of the right inset

(c) Glare or blooming (top) causing contrast reduction and lens flare (bottom). Notice that flare may occur due to lights outside the field of view.

Figure 7.2: Gamma encoded ($\gamma = 2.2$) images of the HDR4CV dataset showing the various problems or imperfections in images under non-ideal but realistic illumination conditions.

test the relative change in performance under adversarial illuminations.

Night Night illumination simulates low-light conditions by using only a few selected lights and dimming them as needed. In low-light imaging, it is customary to use a high camera gain to capture video, and thus, I used a gain of 16 (the camera's maximum gain). The high gain introduces significant noise in the images as depicted by the zoomed-in patches in Figure 7.2a. Using longer exposure times is often not an option for videos, as it results in motion blur and a low frame rate. Another characteristic of the night condition is bright lights such as tunnel lights in the *Tunnel* and street lights in the *Street* scenes.

High dynamic range Here, the lights were arranged to achieve a very high contrast between the dark and bright parts of the scene to simulate HDR illumination conditions (see Figure 7.2b). Such high contrast was achieved either by illuminating one of the faces with the spotlight (for the *Faces* scene) or by using a foam board to produce sharp shadows from a bright source of light. In the *Tunnel* scene, the ceiling lights of the tunnel were turned off, while photographic lights simulated the bright outdoors. As depicted in Figure 7.2b and Figure 7.3 column (b), dark regions of the scene were barely visible when the image was display-encoded using a regular gamma encoding. These image regions were also affected by noise much more than their brighter counterparts.

Glare It is often unavoidable to have very strong light sources, such as car headlights or the Sun, in an image. Such strong light sources introduce visible glare, also known as blooming, caused by unwanted scattering and reflections of light inside the camera lens. The glare causes a reduction in contrast in an otherwise well-exposed image, as shown

in Figure 7.2c (top). Depending on the configuration of the scene, I used either three bright LED bulbs fixed in a photographic frame and connected to a relay, or the compact spotlight controlled by DMX512 protocol to induce glare in the images (see Figure 7.1 right).

Lens flare is another problem caused by stray reflections. Lens flare can be seen in Figure 7.2c (bottom) and is caused by the tunnel lights present in the scene.

7.2.2 Camera

All scenes were captured using an IDS UI-3860CP-C-HQ computer vision camera, which has a Sony IMX290 1/2.8" CMOS sensor of resolution 1936×1096 pixels and pixel size 2.9 µm. The camera was remotely controlled to capture 12-bit RAW images. For each video frame, I captured a stack of 13 images with increasing exposure times and a distance of 1 stop between them. A higher gain of 16 was used for some conditions such as night and HDR. The final images selected for each illumination condition depend on the scene-specific lighting configuration and lens aperture used. These are selected from the captured exposure stacks. An advantage of the captured exposure stacks is that the HDR scenes can be accurately reconstructed. Subsequently, pretrained generative models [2] or calibrated camera parameters (described in Chapter 3) can be used to simulate other cameras, generating additional realistic images of the same scenes captured with different camera settings.

7.2.3 Motorized camera slider

The camera was mounted on a camera slider was powered by a stepper motor driver (Wantai DQ542MA) and controlled with a custom Arduino board connected to a PC. To ensure that the frames were captured from the same viewpoint for each illumination condition, I cycled through all illumination conditions before moving the camera to the next position on the slider. Each scene was captured from 100 different camera positions, simulating smooth camera motion. The total length of the movement was 808.89 mm and thus, the distance between each frame of the sequence was 8.09 mm. The slider was positioned parallel, perpendicular, or diagonal to the background plane of the physical scene, to simulate different motions.

7.3 Scenes

All scenes are semi-realistic scenarios simulated in the lab as shown in Figure 7.3. They consisted of printed foam board cutouts and models in 1:24 scale. All objects were placed at different distances from the camera to introduce parallax. The background was made



(a) Optimal (b) HDR (c) Night (d) Glare

Figure 7.3: The scenes (rows) captured under different illumination conditions (columns). All images have been gamma encoded ($\gamma = 2.2$) for visualisation.

out of a large foam board with printed photographs or patterns glued on it. All the scenes also included an X-Rite Classic Colour Checker Chart, used for white-balance and colour calibration.

7.3.1 Faces

Face detection and recognition are important and well-studied problems. Most face datasets [67, 74, 93] contain prominent, well-exposed faces under ideal illumination conditions. To evaluate the robustness of state-of-the-art face detection and recognition methods, I constructed a scene composed of cardboard cutouts of the faces of popular figures. I also included cutouts of a tiger and a gorilla to introduce the possibility of false positives. The source of glare was introduced by cutting a circular hole in the foam board that supported the background and covering it with a diffuser film. As depicted in Figure 7.1, the compact spotlight was placed on the other side of the foam board and directed towards the camera.

For this scene, the camera slider moved horizontally, parallel to the scene. The dynamic range of the scene was increased by pointing the focused spotlight on one of the cutouts (the Queen in Fig. 7.2b) while keeping the background light dim.

7.3.2 Tunnel

In this scene, the motion of the camera slider simulated the movement of the car as it exited a tunnel and approached an intersection. Unlike the other scenes, the camera moved in the direction it was facing, perpendicular to the plane of the background. I used a wide-angle lens, similar to those used in dash cameras. The scene consisted of objects built from cardboard and foam board cutouts (the tunnel, truck, bus, traffic signs, etc.), models of cars in 1:24 scale, and a model of a tree. For the night illumination condition, the tunnel was illuminated with controllable LED lights.

This scene had a large dynamic range as it included low-luminance regions inside the tunnel and bright regions outside. Lens flare caused by the headlights of oncoming cars and tunnel lights provided additional adversarial illumination.

7.3.3 Street

The final scene depicted a crowded street with cars, a bus, and pedestrians. For the night condition, the street was partially illuminated with small model lamp posts, controlled by the Arduino board. The pedestrians were model people in 1:25 scale from a model train collection (the closest matching scale). The camera moved either along the street (labelled *Street-parallel*) or on a diagonal 45° path (labelled *Street-diagonal*), simultaneously moving along and towards the street. The source of glare had the shape of a set of construction-site lights with the DMX spotlight behind the diffuser film.

7.4 Processing of the frames

I prepared linear colour camera frames, stored as 16-bit PNG files, and HDR frames stored as OpenEXR images. The RAW camera frames were first demosaiced using the DDFAPD algorithm [109] and then white-balanced. To perform white balance, I multiplied RGB values so that all colour channels had equal intensity for the white patch in the colour checker chart. Recall that every scene included the X-Rite chart in some views.

The 16-bit PNG images were created by selecting an appropriate exposure from the captured exposure stacks. For every scene and illumination, I selected the exposure that avoided saturated pixels, and hence, was representative of typical capture conditions. For example, I chose sequences with high gain for night conditions, and exposures with high contrast and almost no saturation for the HDR scene. Saturation of the pixels was unavoidable for the glare conditions, containing bright light sources. However, I selected exposures in which most objects were visible and not saturated.

The HDR image stacks were merged using the Poisson photon noise estimator (Eq. (4.10), Chapter 4), which reduces estimation error in the presence of noise. I used 13 exposures,

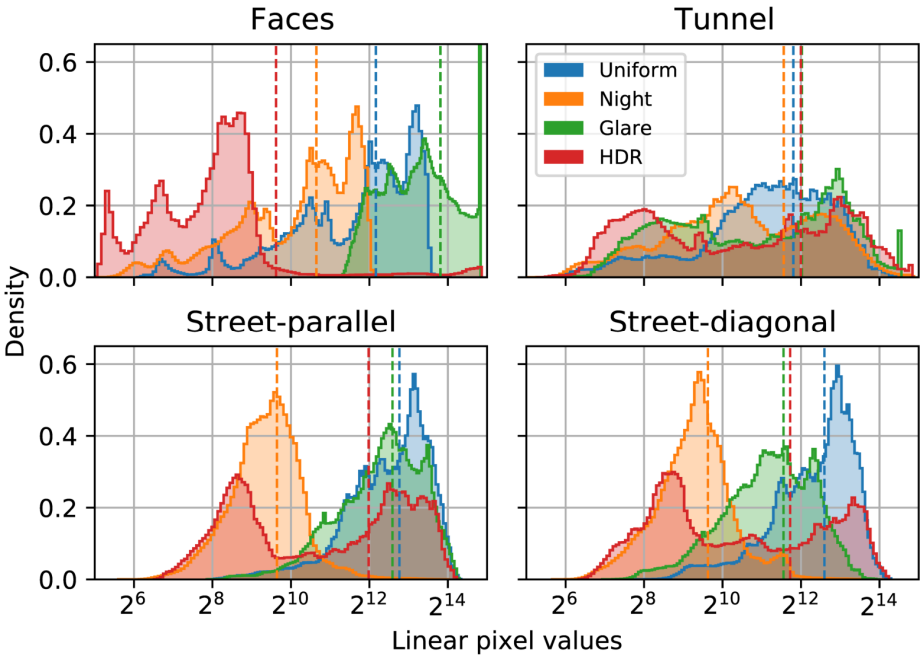


Figure 7.4: Distribution of pixel values under different illumination conditions for all scenes in the HDR4CV dataset. Dashed vertical lines show the average pixel values under each illumination for each scene.

from 0.2 ms to 819.2 ms, separated by 1 stop. Demosaicing and white-balance were performed after merging RGGB sub-pixels. The multi-exposure merging procedure resulted in HDR images with a minimal amount of noise and without any saturated regions. Such images can be used to simulate other cameras with different noise characteristics as shown in Chapter 3.

7.4.1 Distribution of pixel values

To visualise the differences between the illumination conditions, histograms of pixel values are plotted for each scene in Fig. 7.4. It should be noted that the pixel values under uniform illumination are in the upper part of the dynamic range, making these images well-exposed. The HDR condition resulted in the widest histograms, often with two distinct modes. The histograms under the night illumination are shifted to the left, making images darker and more affected by noise. And finally, the histograms are shifted to the right in the glare condition due to the scattered light.

7.5 Performance of computer vision methods

To demonstrate the usefulness of the new HDR4CV dataset, I perform two experiments: first, I test how much the performance of selected CV methods degrades due to challenging

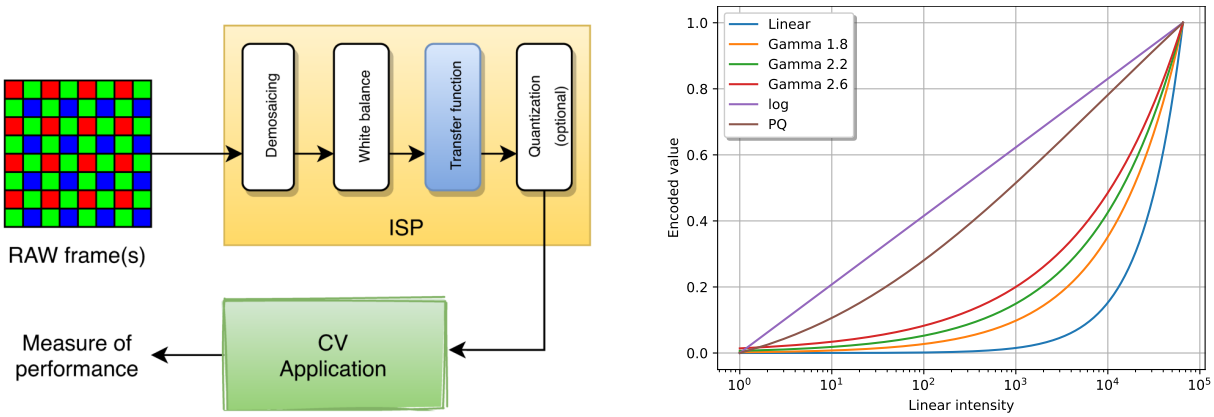


Figure 7.5: Left: The ISP pipeline used for evaluating the performance of CV methods. The intermediate frames are processed by a TF, followed by an optional quantisation. This serves as an input to the CV application being tested. **Right:** TFs used in the experiments to map linear colour values to the encoded values, suitable for CV methods.

Figure 7.6

illumination and then test whether their performance can be improved when different TFs are used to encode frames. A suitable TF maps linear RGB pixel values to non-linear values, which can be represented at lower bit-depths and tend to be more perceptually uniform. [§] These experiments form a preliminary study intended to confirm the selection of the adversarial illumination conditions and are not meant to be a comprehensive evaluation of all possible TFs.

Traditional image signal processing pipelines are designed for best visual quality and are not optimised for CV algorithms [17, 122]. Many steps in the image signal processor (ISP) pipeline may be redundant for some CV algorithms, and may even degrade their performance. Previous work [17] has shown that only two stages of the traditional ISP pipeline are critical in terms of machine vision, namely demosaicing and gamma encoding (or gamma compression). Following this observation, I simulate a simplified camera pipeline, shown in Figure 7.5 (left).

All experiments were performed on publicly available implementations for three CV applications — optical flow, object detection, and face detection. I used the uniform condition encoded with the gamma transfer function (with an exponent of $1/2.2$) to generate reference labels for evaluating methods. As described in Section 7.2.1, the uniform condition consists of well-exposed frames with low levels of noise. I created individual labels for each tested CV method and thus each face detection and optical flow method had its own set of reference labels. I inspected the reference labels and found that almost all objects were correctly detected in the uniform condition and therefore no manual labeling was necessary.

[§]Note that these methods are not labelled *tone mapping* operators, as tone mapping is typically used to produce visually pleasing images for human consumption rather than machine vision.

It should be noted, however, that the reference condition should not be considered as the ground truth and is only meant to show the relative degradation in the performance due to the adversarial illumination.

I used dense optical flow implementations of a polynomial expansion algorithm [38], a coarse-to-fine algorithm [90] and a state-of-the-art deep neural network RAFT [152]. To compare predicted pixel-wise optical flow field $V_{\text{pred}}(p)$ to the reference $V_{\text{ref}}(p)$, I used the endpoint error (EE) metric:

$$\text{EE} = \frac{1}{N} \sum_{p=1}^{N} \left| V_{\text{pred}}(p) - V_{\text{ref}}(p) \right| . \tag{7.1}$$

where N is the number of pixels.

For object detection I used a pretrained YOLOv3 [131] network. For face detection I used a pretrained single-shot detector (SSD) [91], face detector based on histogram of oriented descriptors (HOG) [26], and maximum margin object detection with CNN features (MMOD) [79]. Bounding boxes of both detection tasks were evaluated using the mean intersection over union (mIOU) metric:

$$\text{mIoU} = \frac{area(R_{\text{pred}} \bigcap R_{\text{ref}})}{area(R_{\text{pred}} \bigcup R_{\text{ref}})} , \tag{7.2}$$

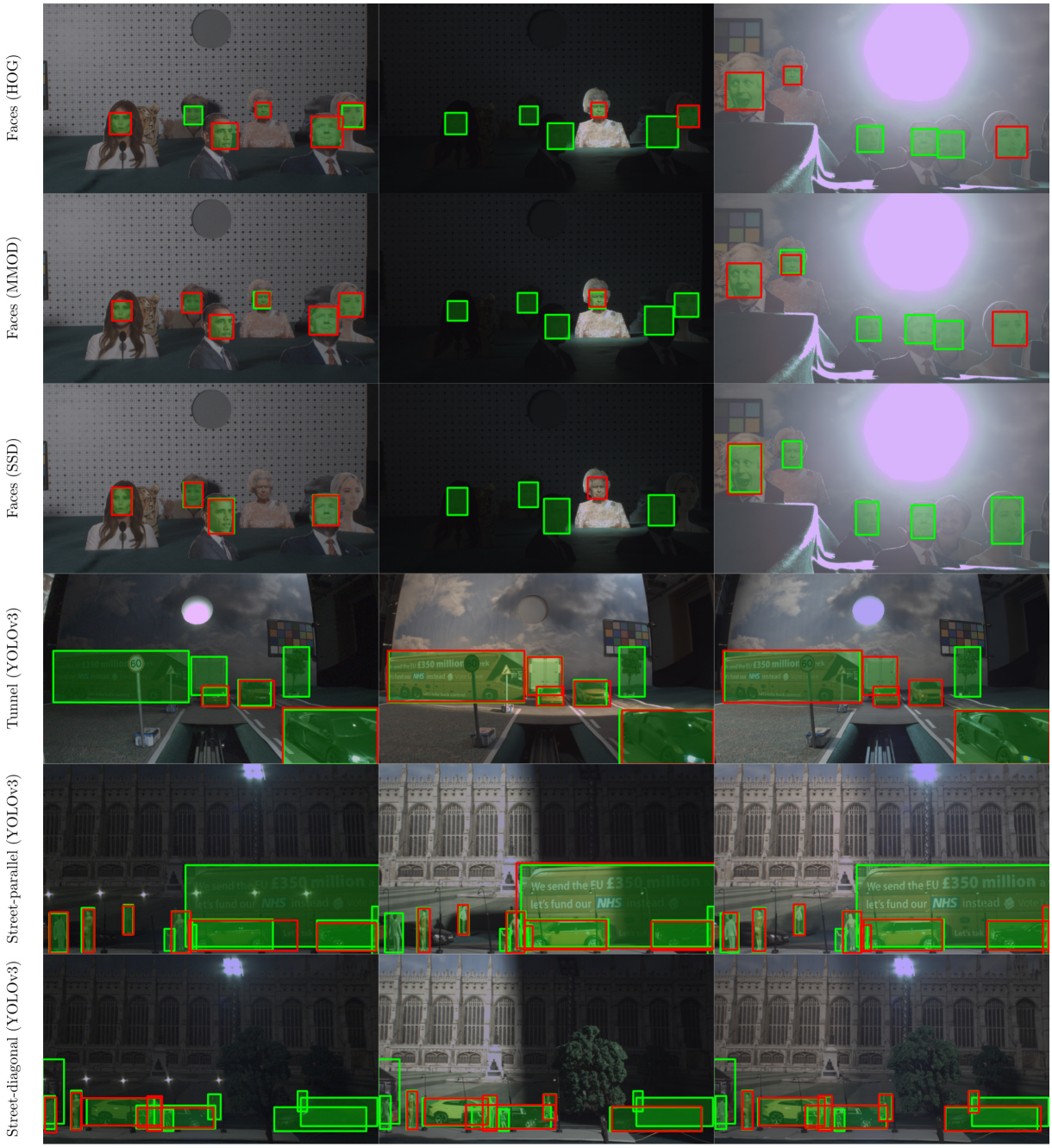
where $R_{\text{pred}} = \bigcup_{i=1}^{N} B_{\text{pred}}(i)$ is the union of all N bounding boxes predicted and $R_{\text{ref}} = \bigcup_{j=1}^{M} B_{\text{ref}}(j)$ is the union of all M reference bounding boxes.

7.5.1 Effect of changing illumination

All images I_{in} were encoded with the gamma encoding, $I_{\text{out}} = I_{\text{in}}^{1/\gamma}$ where $\gamma = 2.2$. Such a gamma encoding is the most widely used TF in CV cameras, which produces images that can be directly viewed on standard monitors, as it approximates the TF used in the sRGB colour space. Here, I test how the performance of the CV methods is degraded in adversarial illumination conditions.

The results of this experiment are listed in Table 7.2. We observe a considerable difference in the results of the tested methods with changing illumination conditions. Figure 7.7 shows a qualitative comparison for the different conditions for the face and object detection methods.

The low scores of all face detection methods for the HDR illumination (see Table 7.2) can be explained by the first three images of the second column of Fig. 7.7. The inefficient encoding of the input images results in 5 out of 6 faces being underexposed. In the glare condition (last column of Fig. 7.7), the faces present under the bright light source are not detected by any of the methods. The MMOD and HOG detectors are slightly more robust



(a) Night (b) HDR (c) Glare

Figure 7.7: Bounding boxes for several face and object detection methods (rows) for three adversarial illumination conditions (columns), each gamma encoded (with $\gamma = 2.2$). Faces and objects detected by the various methods for the reference uniform illumination condition are shown as shaded green rectangles, while the results of the tested methods for other illumination conditions are shown as red rectangles.

Table 7.2: The performance of popular methods under different illumination conditions, averaged over 100 images in each sequence. All input images were encoded using a gamma encoding ($\gamma = 2.2$). The reference labels for each application were obtained from uniform conditions.

Application	Method	Faces			Tunnel			Street-parallel			Street-diagonal		
		Night	HDR	Glare*	Night	HDR	Glare	Night	HDR	Glare	Night	HDR	Glare
Face detection (mIOU) ↑	HoG	0.75	0.15	0.66	✗	✗	✗	✗	✗	✗	✗	✗	✗
	MMOD	0.93	0.16	0.44	✗	✗	✗	✗	✗	✗	✗	✗	✗
	SSD	0.56	0	0.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
Object detection (mIOU) ↑	YOLOv3	0.92	0.63	0.76	0.52	0.85	0.9	0.71	0.76	0.88	0.78	0.82	0.9
Optical flow (EE) ↓	PolyExp	0.97	5.91	3.1	2.52	2.78	2.35	4.11	3.56	1.34	2.07	2.47	1.84
	Coarse2Fine	0.13	0.64	0.35	1.03	6.21	6	0.57	0.18	0.09	0.35	0.28	0.17
	RAFT	54.71	40.49	1.82	22.6	0.32	0.27	0.68	0.15	0.1	0.48	0.2	0.15

to glare and are able to detect faces further away from the source of glare. This explains their better performance for the same inputs.

Glare is not as much of a problem in the other scenes (*Tunnel* and *Street*) because the objects of interest are further away from the bright light source. In general, the methods produce lower mIOU scores for very dark objects in night illumination. This is likely a consequence of underexposed pixels rather than noise since underexposed objects pose a problem even in HDR conditions (for example, the bus and truck in *Tunnel* and the car and person in the bottom-right shaded region in *Street-diagonal*).

7.5.2 Comparison of TFs

Next, I evaluate the performance of the CV methods when the input images are encoded using different TFs. These include a *linear* function (no TF, linear colour values), *gamma* encoding (see Section 7.5.1) using three different exponents ($1/1.8$, $1/2.2$ and $1/2.6$), the *log* function, and the perceptual quantiser (PQ), which is commonly used for encoding HDR content [113]. All the TFs are plotted using log-linear coordinates in Figure 7.6 (right). The advantage of this visualisation is that the slopes of the plotted functions correspond to the compression or expansion of contrast at a particular intensity value. The plots show that the TFs intended for HDR images (*log* and PQ) better preserve contrast across the entire range of values, while *gamma* and *linear* functions compress contrast at lower intensity values.

Optical flow: As shown in the top three rows of Figure 7.8, all tested optical flow methods show similar characteristics: they all benefit from HDR TFs (*log* and PQ) for most conditions. It should however be noted that there are also important differences between the methods. The Coarse2Fine method seems to be robust to noise in the night condition but affected by the large dynamic range in the HDR condition, especially for the *tunnel* scene (note the absolute values of the EE). PolyExp resulted in larger EE across

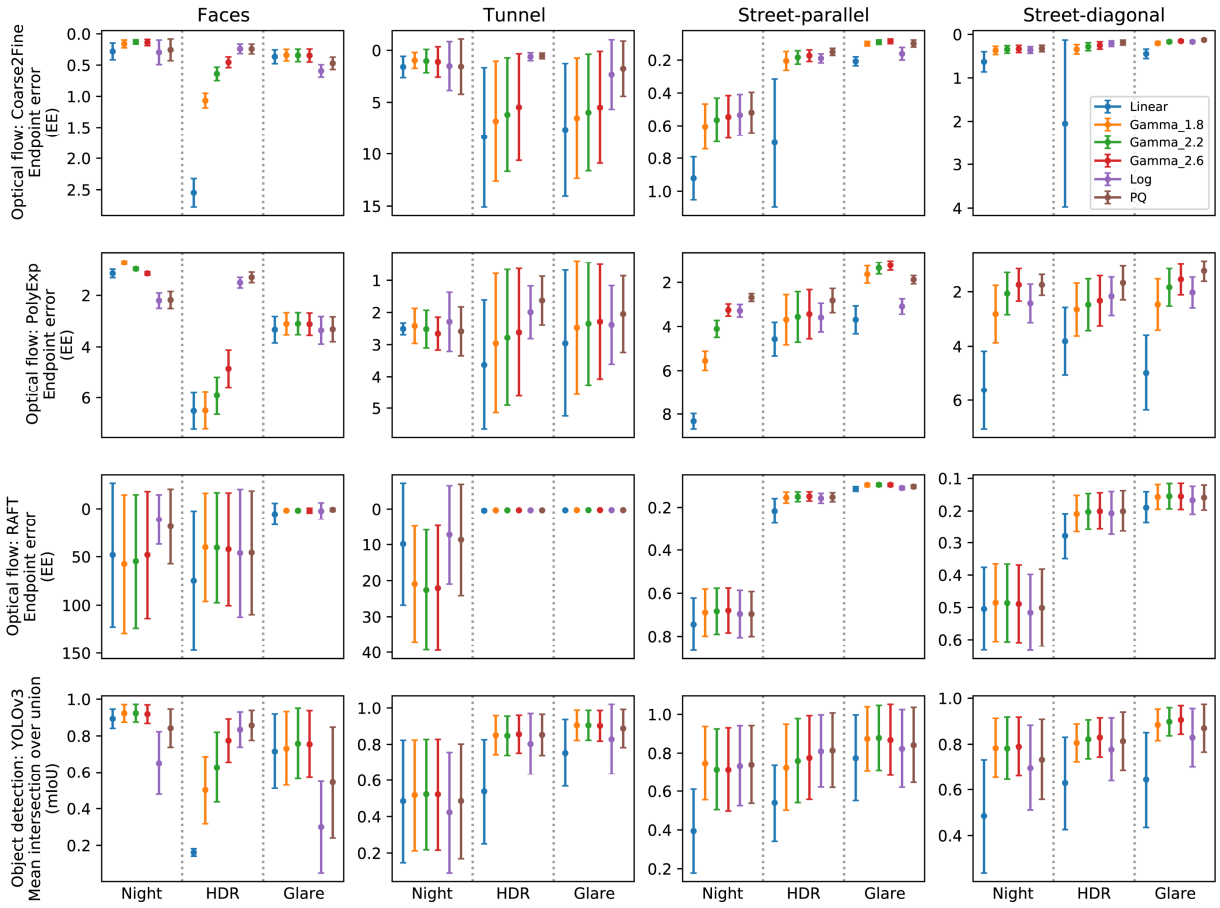


Figure 7.8: The performance of optical flow (top three rows) and object detection (bottom row) for all four scenes (columns) and three adversarial illumination conditions (three sections in each plot). Dots represent the mean and the error bars represent the standard deviation for a sample of 100 frames. The y-axis is reversed for optical flow so that the points located higher on the plot correspond to the better performance.

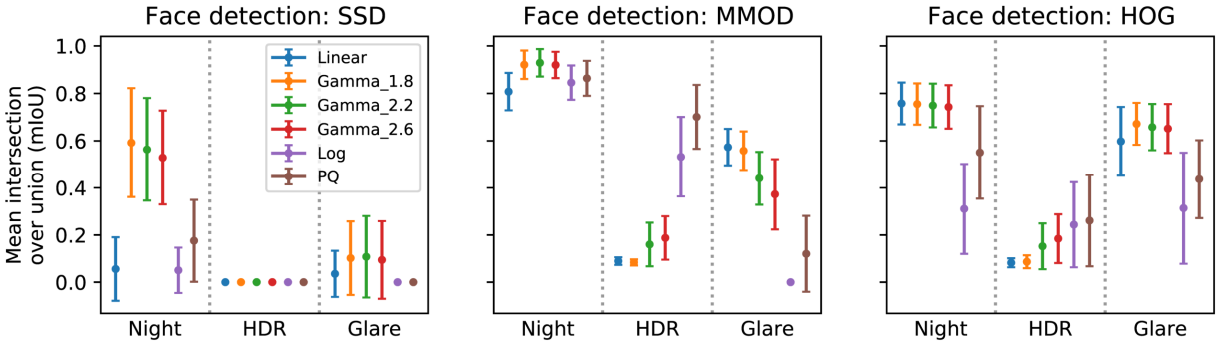


Figure 7.9: The performance of face detection methods for all transfer functions and all illumination conditions. Dots represent sample mIOU, and error bars represent the sample standard deviation over 100 frames. When some transfer functions are used, some methods are unable to detect even a single face. For example, notice the zero entries for the HDR and glare illuminations with the *log* function for SSD and MMOD.

all the conditions and is less robust to noise. Finally, RAFT resulted in huge EE for *Faces* and Tunnel. I investigated this issue and found that the cause was a dark region of the image, which captured a part of our laboratory outside the illuminated part of the scene. RAFT predicted very high velocity for this region, resulting in an EE an order of magnitude higher than for other methods (see Fig. 7.10). For the two *Street* conditions, RAFT was robust in the HDR condition but was affected by noise in the night condition. We can conclude that the Coarse2Fine method is most robust to adversarial conditions and it should be combined with *PQ*.

Object detection: The results for object detection (YOLOv3) shown in the bottom row of Figure 7.8 are strongly affected by the adversarial illumination. The TFs intended for HDR data, *log* and *PQ*, improve performance for HDR condition, but they also reduce performance for the two other conditions. This is because night and glare conditions contain objects of interest in the upper (brighter) part of the dynamic range. Since *gamma* encoding tends to enhance contrast in the upper part of the dynamic range, it achieves better performance for those two conditions but fails for HDR, where it is unable to reproduce contrast in the darker part of the frame, as shown in Figure 7.11.

Face detection: The results plotted in Figure 7.9 show similar trends for all three face detection methods. As was the case with object detection, HDR TFs help in HDR conditions, but can also degrade performance for other conditions. For the same reason, *linear* colour representation unexpectedly performs reasonably well for night and glare, though it still fails for the SSD method. It is also worth noting that the SSD face detector is more affected by the adversarial illumination than the two other methods. We can conclude, that face detection would benefit from an adaptive TF, which selectively reproduces contrast in the darker or brighter part of the dynamic range, depending on image content.

7.6 Limitations

Although the HDR4CV dataset was designed to be possibly realistic, it may not be suitable for applications that rely on accurate material reflectance properties, such as shape-from-shading or relighting. This is because many objects in the model scenes are cardboard cutouts and do not capture the richness of materials in the wild. The dataset could complement computer graphics datasets, which can potentially reproduce more accurate materials but may lack imperfections and artefacts caused by camera sensors (noise, quantization) and optics (glare).

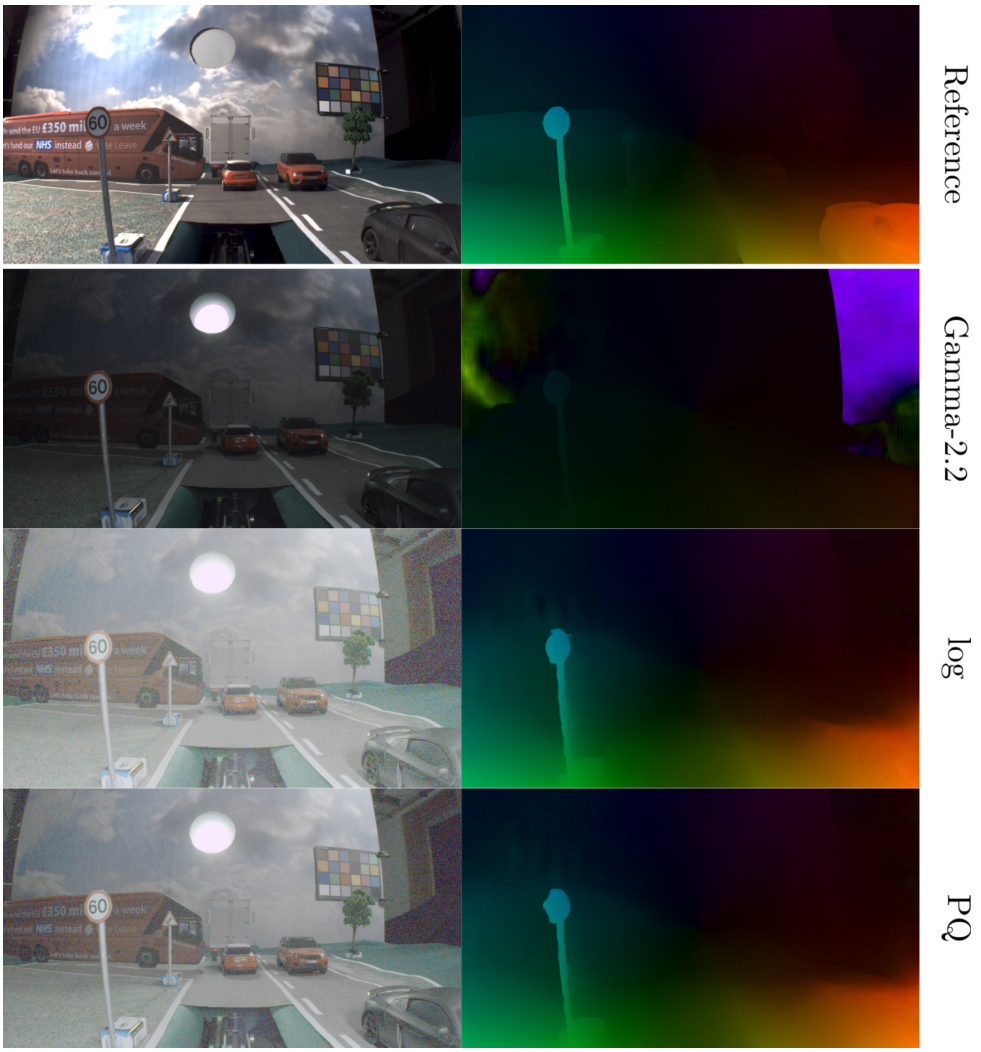


Figure 7.10: Failure of *gamma* encoding for night condition of the *Tunnel* scene when the optical flow is estimated by RAFT. The purple region in optical flow visualisation for *gamma-2.2* indicates high velocity in the south direction, which is most likely caused by the dark region in that corner of the frame. The HDR TFs (*log* and *PQ*) can better handle this sequence. This is also the cause of the large error bars in some plots in Figure 7.8.

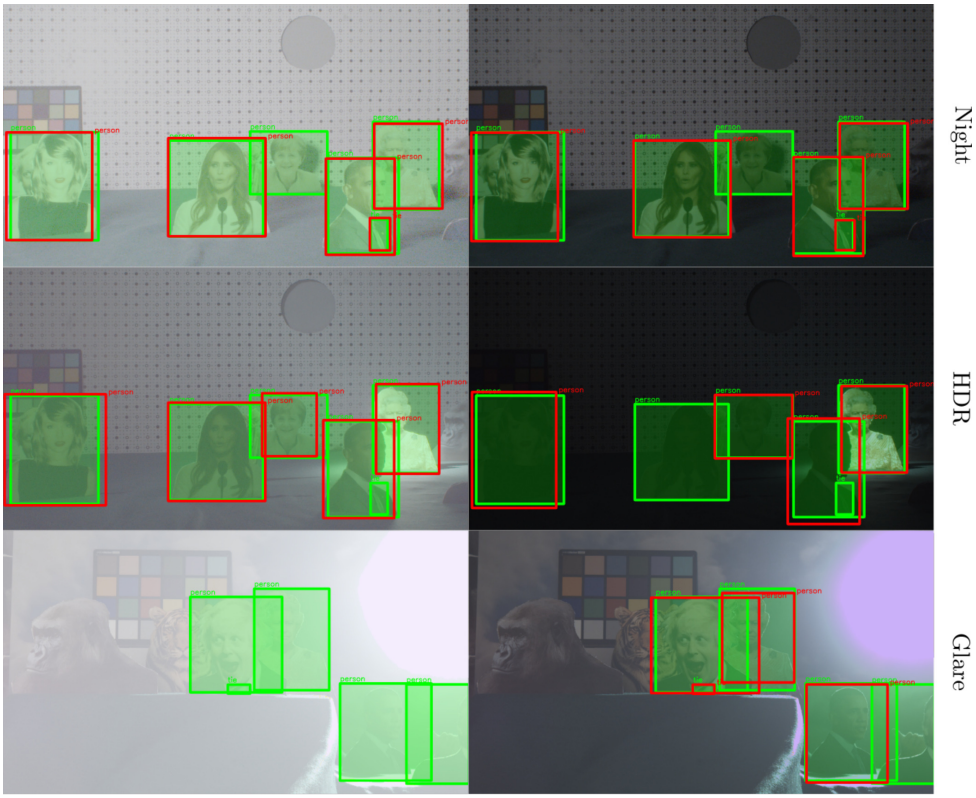


Figure 7.11: Object detection using YOLOv3 for images captured under different illumination conditions (rows) and encoded with *log* (left column) and *gamma* (right column) TFs. The *log* function produces brighter images by devoting more bits to dark regions as indicated by the purple curve in Figure 7.6 (right). This leads to contrast reduction in bright regions posing problems to some object and face detectors for the Glare condition, but this is an effective strategy for other conditions.

7.7 Summary

I created a new HDR4CV dataset consisting of short video clips captured in adversarial illumination conditions. The main advantage of this dataset is that it includes uniformly illuminated scenes which result in images with minimal noise. The uniformly illuminated scenes benefit from physically accurate radiance values reconstructed by considering camera noise (Chapter 4). With these, reference labels can be automatically generated for most CV methods. Such labels can be then used to measure the degradation in the performance of CV methods in adversarial conditions: night (high noise), HDR (high contrast) and glare (stray light in camera lens).

Further, I used the dataset in a preliminary study to evaluate the robustness of popular methods for face detection, optical flow, and object detection under adversarial illumination conditions. I studied whether the performance of these methods can be further improved under challenging lighting conditions by selecting an appropriate transfer function. The results suggest that, as expected, the popular *gamma* encoding is unsuitable for HDR scenes. At the same time, the transfer functions intended for HDR scenes, such as *log* or *PQ*, reduce contrast in bright regions leading to rather poor performance in well-exposed images.

I did not consider sophisticated local and content-aware operators in this preliminary study. These operators can introduce temporal tone differences, potentially impacting the performance of CV methods.[¶] In the context of video datasets, such as the one introduced in this chapter, faster and easier acquisition would enhance the scale and diversity of collected data. The next few chapters discuss an approach to easily reconstruct HDR videos without the need for exposure stacks.

[¶]The effect of such operators is part of a larger study outside the scope of this dissertation.

Chapter 8

High dynamic range from a single photograph - Evaluation

Merging SDR images with correct exposure ratios and MLE-based estimators produces highly accurate reconstructions for tripod-mounted cameras and static scenes. I showed in Chapters 6 and 7 that such HDR accurate reconstructions are important for both human and machine consumption. However, combining exposure stacks results in ghosting artefacts when the scene is dynamic or the camera is hand-held. Eliminating such ghosting artefacts remains a practical and challenging problem [127]. Moreover, using a single camera to capture multi-exposure HDR video imposes constraints on the exposure times due to the need for a fixed frame rate.

In the remaining chapters, I focus on the closely related problems of single-image HDR (SI-HDR) and inverse tone-mapping (iTM), which involves predicting HDR information from a single SDR image. I compared state-of-the-art methods by conducting a carefully constructed subjective experiment. The purpose of the experiment was to identify a reliable SI-HDR method for accurate HDR reconstruction from standard videos. This could significantly enhance the availability of HDR content for both human and machine consumption (see Chapters 6 and 7 for examples). However, I was surprised by the lacklustre performance of recent methods and decided to investigate further.

8.1 Problems involved

Let's start by looking at why SI-HDR and iTM are challenging problems and emphasise the differences between them. The forward model for HDR reconstruction problems describes how a camera obtains an SDR image, L, from a linear HDR ground truth image, H,

$$L(p) = q\left(\min\{1, f(t\,H(p) + \sqrt{\mathbb{V}[H(p)]})\}\right), \tag{8.1}$$

where p denotes the pixel index, t is a simulated exposure time, $\sqrt{\mathbb{V}[\cdot]}$ is signal-dependent and gain-specific camera noise from Chapter 3, q performs quantization to the desired bit depth, and f is a non-linear tone-mapping or CRF. Transforming a standard dynamic range image to high dynamic range (SDR-to-HDR) aims at approximating the inverse of Eq. (8.1). This is a multi-faceted problem that can be divided into the following sub-problems:

R-I **Inversion of the CRF** – Estimating the inverse f^{-1} is a challenging problem and is impossible without the presence of enough contextual information in an image. At the same time, one can argue that this problem is not the most important in SI-HDR reconstruction for at least two reasons. Firstly, a perfect linearization/CRF inversion is not necessary for most applications. This is because some applications are not sensitive to linearization, and for others, some non-linearity will be applied in the end anyway. Secondly, for many applications, linear data is readily available (e.g., RAW images directly from the camera sensor).

R-II **Bit-depth expansion/de-quantization** – Recovering higher bit-depth is typically unnecessary for modern cameras since they are equipped with analog-to-digital converters with sufficient bit-depth. However, when a CRF with shallow slopes needs to be inverted, the process may introduce visible quantization artifacts.

R-III **Reconstructing of under-exposed pixels** – These pixels are mostly affected by noise in Eq. (8.1), and in most cases this problem is similar to image denoising. The details in dark regions can also be lost due to quantization, for example, if the pixels are largely under-exposed and/or aggressively compressed by a CRF.

R-IV **Reconstruction of over-exposed pixels** – The clipping of information due to sensor saturation, modeled by $\min\{\ldots\}$ in Eq. (8.1), is most often directly associated with the SI-HDR reconstruction problem. While the problem shares some similarities with the recovery of under-exposed pixels, it also differs significantly in many aspects. Perhaps the most relevant difference is the fact that over-exposed pixels often carry most of the dynamic range. This results in a long-tailed distribution of luminances because specular highlights and light sources are much brighter than other parts of a scene. For some applications, such as image-based lighting, this information is critical. The problem is also significantly different in terms of solution and is more closely related to inpainting.

R-V **Other** – Other problems could also be part of the reconstruction pipeline, including color correction, denoising, glare removal, removal of compression artifacts, etc. However, these are most often considered separate from the HDR reconstruction problem.

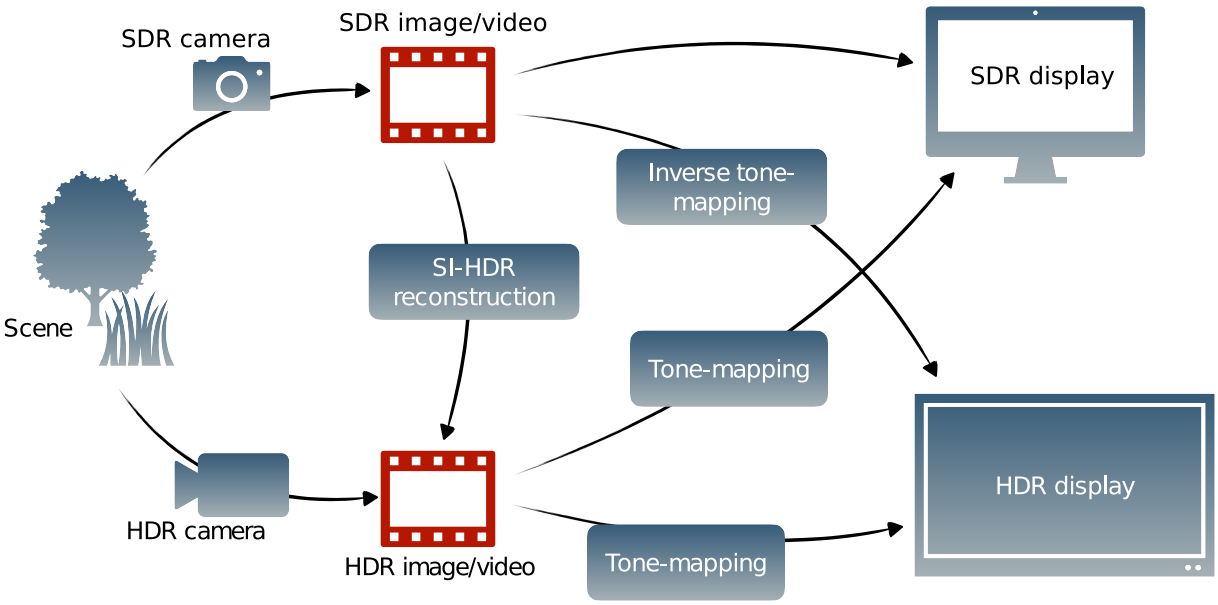


Figure 8.1: Illustration of how the intent of SI-HDR reconstruction and iTM differs. The figure shows SDR/HDR display as the direct application, but the HDR images could also be used for other purposes (IBL, post-processing, etc.). While iTM directly maps SDR content for an HDR display, a similar goal could be achieved by SI-HDR followed by tone-mapping.

The SI-HDR problem is typically understood as the inverse problem of Eq. (8.1). A closely related problem termed iTM aims to enhance an SDR image L but without the goal of matching the original HDR image H. The difference between SI-HDR and iTM is illustrated in Figure 8.1, and can be summarised as follows:

8.1.1 Inverse tone-mapping

This refers to methods that use an iTM operator for maximising subjective quality on an HDR display. In this scenario, colour gamut, exposure, and local contrast should be correctly tuned for an improved viewing experience. Recovering information in highlights is not essential since a rich viewing experience is possible even if some information is clipped to the lower peak luminance of the display. This category includes the earlier heuristic methods [13, 14, 110, 134], as well as more recent learning-based models [77, 78].

8.1.2 Single-image HDR

This refers to methods that aim at recovering the underlying physical light quantities of the captured scene. In this scenario, one of the most important problems is to recover lost information in the saturated areas of the image (R-IV).

It should also be noted that an iTM operator can be composed by taking the result of an SI-HDR method and tone-mapping it for an HDR display, thus potentially disregarding some of the reconstructed information. I unveil an element of confusion in intent in existing

works since some methods are not clear about which objective they aim for. Often, iTM is used to describe methods that aim for SI-HDR reconstruction.

8.1.3 The evaluation problem

The evaluation of iTM is similar to that of conventional tone-mapping, which aims at displaying HDR images on SDR monitors. That is, it is difficult to formulate a numerical evaluation measure that reflects subjective preferences. The evaluation of SI-HDR reconstruction, on the other hand, is closer to multi-exposure HDR reconstruction where the underlying luminance values of the scene are sought. A major problem with recent SI-HDR papers is confusion in intent, causing methods to be presented as iTM operators but evaluated by comparing them to reference HDR images. This, by itself, imposes difficulties in properly assessing a method. However, even if a method is clearly performing SI-HDR reconstruction, it is still problematic to numerically evaluate its performance, for a number of reasons.

8.2 Assessing the quality of SI-HDR reconstructions

While most single-image HDR (SI-HDR) methods are trained to recover the physical HDR radiance values, they were found to perform poorly in this task. Despite this, many SI-HDR and iTM methods provide appreciable image enhancement that improves the quality of resulting images. These two tasks have radically different goals [28], making it difficult to automatically compare competing methods with image quality metrics.

Another issue is that most methods are evaluated on datasets with similar distributions to training images, which provides an unfair advantage. To ensure a fair evaluation, my collaborators and I collected a new SI-HDR dataset — a diverse set of HDR images with SDR counterparts, which were generated using a physically accurate camera model.

With a pairwise comparison experiment on a subset of the SI-HDR dataset, I demonstrate that some SI-HDR methods show promising image enhancement results, even though the reconstructions are far from HDR references. Surprisingly, two out of the six tested methods are more likely to decrease quality than improve it. The existing protocol for automatic evaluation conceals such results because most image quality metrics are more suitable for the assessment of the restoration task rather than the enhancement task. When comparing a reconstructed image with the reference HDR, metrics tend to be highly sensitive to shifts in tone and colour due to incorrect CRF estimation. As a result, artefacts objectionable to humans, particularly in saturated regions of the input image, have less influence on the metric scores.

To remedy this, I propose an improved SI-HDR evaluation protocol that includes CRF correction, which significantly boosts the predictions of most metrics. I then compare an

extensive list of quality metrics and identify those that highly correlate with the results of the subjective experiment. The selected metrics reproduced the subjective rankings within predicted error bounds when run on the larger evaluation dataset. To estimate these bounds, I identified various sources of noise in the predictions and, for selected metrics, estimated the difference in metric scores necessary to claim an improvement above the noise level. Such differences are substantial for most metrics, suggesting that only large quality differences provide a reliable claim in the assessment of SI-HDR methods.

8.3 Related work

SI-HDR methods can be most reliably evaluated in a large-scale perceptual study on an HDR display. However, this is a tedious task, and most works instead rely on objective metrics to show the improvement of a newly proposed method. To measure the difference between the output of an SI-HDR method and the reference, metrics are usually applied directly or after a linear transformation, such as a global scaling with the image median or some other percentile. The comparison could use a dedicated HDR metric, such as HDR-VDP-2 or 3 [106], or SDR metrics, such as PSNR, SSIM [163] or LPIPS [175], on tone-mapped HDR values. Instead of tone-mapping, which may strongly reduce contrast, it has been suggested to use the PU-transform [104] or μ-law transform [71]. Some existing works directly apply SDR metrics on linear HDR images. This should be avoided since linear radiance values are perceptually non-uniform and incorrectly represent perceived differences [103].

There are several potential problems with existing protocols. First, test data, camera simulation, and comparison methods differ substantially between evaluations, making it impossible to compare results between papers. Second, results are easily dominated by the ability of a particular method to invert the CRF (g in Eq. (8.1)). This affects all pixels in an image and obscures the evaluation of recovered HDR information in, e.g., saturated regions of the image. Third, new SI-HDR reconstruction methods are often evaluated on images generated by a similar camera simulation as training images, which is likely to introduce a bias towards the proposed method since it could be better at inverting the CRF. Finally, comparisons are often performed using HDR metrics without properly calibrating the compared images.

A recent attempt at an independent and standardised evaluation was presented as a part of the HDR single and multi-frame imaging challenge in the New Trends in Image Restoration and Enhancement (NTIRE) CVPR workshop [127]. The methods were evaluated on a new HDR dataset, using PSNR on linear and tone-mapped images. Their work overcomes some of the issues with diverse evaluation conditions, but problems around CRF dominance and metric calibration are yet to be addressed. This chapter looks at



Figure 8.2: The subset of HDR images from SI-HDR dataset used for the subjective evaluation, tone-mapped with a global operator [102] for visualisation. A wide variety of content were selected, covering nature, portraits, cities, indoor and outdoor, daylight and night scenes.

the remaining issues by (1) introducing an explicit CRF correction step to prevent CRF dominance, (2) identifying SI-HDR metrics based on a thorough subjective experiment.

8.4 SI-HDR dataset

As most existing HDR datasets were used to train SI-HDR methods, my collaborators and I created a new dataset for the purpose of this evaluation. The new dataset needs to be sufficiently large and diverse in scene content and dynamic range. Most importantly, the data should not be used to train any of the methods compared. In this section, I describe our HDR data collection and SDR simulation, as well as the SI-HDR methods used in comparisons.

8.4.1 Multi-exposure stacks

183 HDR images were captured using a Canon 5D Mark III full-format DSLR camera. The scenes were selected to cover a wide range of image content and lighting conditions. Each scene is composed of up to 7 RAW exposures and merged into an HDR image using the estimator that accounted the photon noise (PPNE from Chapter 4). Next, a simple colour correction was applied using a reference white point and all images were resized to 1920×1280 pixels.

8.4.2 Camera simulation

I simulated the Canon EOS-1Ds with $t = 1/30$ sec and ISO 800 according to Eq. (8.1). I used the measured CRFs from the dataset in [52]. This collection contains a wide variety of CRFs, including measurements from analogue slide film which are not representative of modern digital cameras. Thus, I performed k-means clustering of CRFs with five clusters, and only used the cluster with a mean closest to the overall mean. This cluster contains

85 different CRFs with shapes that are representative of modern cameras. For each SDR simulation, I randomly picked a CRF from this set and individually selected a scene-specific exposure so that either 3% or 5% of pixels were saturated.

8.4.3 SI-HDR methods

I tested six deep-learning SI-HDR methods from the literature, DrTMO [35], HDR-CNN [33], ExpandNet [107], HDR-GAN [86], SingleHDR [92], and Mask-HDR [139]. The selection includes frequently occurring methods in SI-HDR evaluations (DrTMO, HDR-CNN, and ExpandNet), as well as more recent ones (SingleHDR and Mask-HDR), including an adversarially trained network (HDR-GAN). While most methods directly predict linear HDR images as the output of the respective network, DrTMO and HDR-GAN predict exposure stacks. To remain faithful to the original implementations, I merged these stacks with OpenCV's multi-exposure merging (based on [27]).

Ideally, all methods would be retrained on a standardised dataset but unfortunately, such a dataset is not available. Moreover, retraining existing methods typically results in worse performance due to differences in data preparation, choice of camera simulation, augmentation, and hyper-parameter tuning. Instead, the methods were tested on a new unseen dataset using the models that have been trained by the respective authors.

8.5 CRF correction

SI-HDR methods typically fail to invert the camera response function (CRF), making the resulting colours and tones different from the reference HDR images. The inaccuracies in CRF reconstruction are often so large, that the images are too different to be reliably compared in a subjective image quality experiment. Hence, a simple method that improves CRF inversion was developed so that the evaluation can focus on the reconstruction of saturated pixels rather than accurate CRF inversion.

The goal of this step is to find a global smooth colour mapping function, $\mathbb{R}^3 \rightarrow \mathbb{R}^3$, that corrects for the inaccurate CRF reconstruction, without interfering with the reconstruction of saturated pixels. To achieve this, the coefficients of a polynomial basis function were fit. Performing the correction on log RGB values works well in most cases. However, in some situations this produced visible artefacts, especially for pixels with highly saturated colours.

Instead, luma values P encoded using the PQ transfer function [113] were optimised first,

$$P_i = PQ(Y_i)\,, \tag{8.2}$$

where Y_i is the luminance of a pixel. The next step is finding the luminance mapping

between the reference HDR image P and the SI-HDR output $\hat{P}$, by solving:

$$\underset{w_1,\ldots,w_4}{\arg\min} \left|\left| \begin{bmatrix} \hat{P}_1^3 & \hat{P}_1^2 & \hat{P}_1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{P}_N^3 & \hat{P}_N^2 & \hat{P}_N & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} P_1 \\ \vdots \\ P_N \end{bmatrix} \right|\right|, \tag{8.3}$$

where subscripts $1, ..., N$ denote pixel indices, and the four coefficients w_k describe a third-degree polynomial. Then, the mapping of the CIE chromatic coordinates $u'v'$ is optimised:

$$\underset{w_{1,1},\ldots,w_{8,2}}{\arg\min} \left|\left| \begin{bmatrix} \hat{u}_1'^3 & \hat{v}_1'^3 & \hat{u}_1'^2 & \hat{v}_1'^2 & \hat{u}_1'\hat{v}_1' & \hat{u}_1' & \hat{v}_1' & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{u}_N'^3 & \hat{v}_N'^3 & \hat{u}_N'^2 & \hat{v}_N'^2 & \hat{u}_N'\hat{v}_N' & \hat{u}_N' & \hat{v}_N' & 1 \end{bmatrix} \begin{bmatrix} w_{1,1}w_{1,2} \\ \vdots \quad \vdots \\ w_{8,1}w_{8,2} \end{bmatrix} - \begin{bmatrix} u_1' & v_1' \\ \vdots & \vdots \\ u_N' & v_N' \end{bmatrix} \right|\right|, \tag{8.4}$$

where $\hat{u}_i'$ and $\hat{v}_i'$ are the chromaticity coordinates of the SI-HDR reconstruction for pixel i, while u_i and v_i are those of the reference HDR image. The 16 coefficients $w_{k,c}$ describe third-degree polynomials for each channel, with cross-channel dependencies.

In practice, both optimisation problems are solved by minimising the squared error,

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\arg\min} \, ||\mathbf{XW} - \mathbf{Y}||_2^2 + \lambda ||\mathbf{W} - \mathbf{W}_0||_2^2, \tag{8.5}$$

where the first term corresponds to Eq. (8.3) or Eq. (8.4), i.e. where $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Y}$ are the pixel values of the reconstructed image, the polynomial weights, and the pixel values of the reference image, respectively. The formulation includes a Tikhonov regularization term (corresponding to MAP from Section 2.3.3), with strength λ. This is applied to penalise extreme values in the coefficient matrix $\hat{\mathbf{W}}$, which cause colour distortions in the corrected image. The weight penalisation is performed against the point $\mathbf{W}_0$, which is set with the weights that result in identity mapping of colour luminance values (all coefficients are 0 except for the linear term which is set to 1, i.e. $w_3 = 1$ in Eq. (8.3) and $w_{6,1} = 1, w_{7,2} = 1$ in Eq. (8.4)). The solution of the minimisation problem is:

$$\hat{\mathbf{W}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{Y} + \lambda\mathbf{W}_0), \tag{8.6}$$

where $\mathbf{I}$ is the identity matrix.

8.6 Subjective evaluation

The goal of the experiment was to obtain a possibly accurate measure of the gain (or loss) of visual quality that can be achieved by each SI-HDR method. The method of pairwise

comparisons was used, as it was previously shown to provide higher sensitivity than direct rating methods [126].

8.6.1 Details of the experiment

Here are some additional details about the experimental setup and procedure.

Display The images were shown on a 32" Asus ProArt PA32UCX 4k, HDR monitor. The monitor was set to use the PQ1000 response curve and was calibrated with a spectroradiometer (Specbos 1221). The displayed values had luminance 30% lower than expected (mapped PQ values vs. measurements). This was compensated for with a custom colorimetric calibration. The viewing distance was restricted to approximately 80 cm, resulting in an effective display resolution of 77 pixels per degree.

Images I selected a subset of 27 HDR images from the SI-HDR dataset. The images, shown in Figure 8.2, were selected for a wide variety of content: portraits, nature, cities, indoor and outdoor, daylight and night time scenes. When the experiment directly was run directly on the results of SI-HDR methods, images differed too much in tone and colour to be compared. Instead, the images with the corrected CRF (Section 8.5) were used so that the participants judged the ability of the methods to reconstruct saturated pixels rather than their subjective opinion about tone and colour of each image. The images were further cropped to a resolution of 1888×1280 so that two images could be shown side-by-side on a 4K monitor. Their exposure was adjusted so that the saturation point mapped to 100 $\mathrm{cd/m^2}$ on the display. The range of luminance from 100 to 1000 $\mathrm{cd/m^2}$ was used to reproduce the reconstructed pixel values. In total, the quality was assessed for 27 images (contents) × 2 exposures × (6 methods + HDR reference + SDR input) = 432 conditions.

Experimental procedure The participants were shown a pair of images side by side and were asked to select "the image of higher quality — the one that better resembles a natural scene and contains fewer distortions" (the wording from the briefing form). The images were compared within blocks showing the same content and generated using different SI-HDR methods. Both SDR input and HDR reference images were included in the comparison. An active sampling method, ASAP [111], was used to maximise the accuracy of the collected data. ASAP determines a batch of comparisons that maximises the information gain and was shown to outperform heuristics, such as the Swiss chess system.

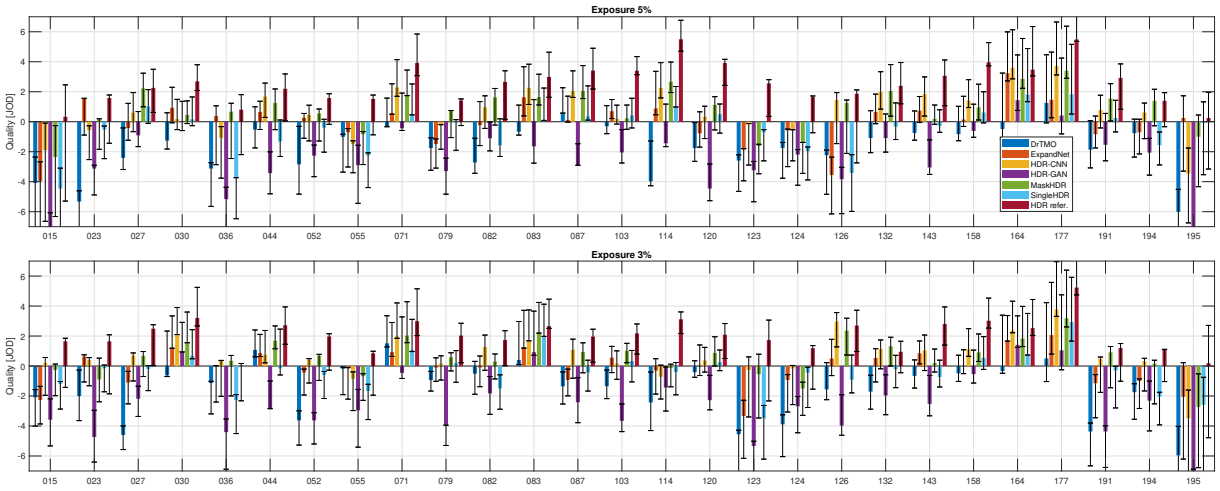


Figure 8.3: Preference of the SI-HDR method results shown per exposure (top/bottom) and for all content (x-axis). The error bars show bootstrapped 95% confidence intervals.

8.6.2 Results: SI-HDR benchmark

Figure 8.3 shows the results of the experiment for each scene and the two exposures corresponding to 5% and 3% clipped pixels. Notice that both exposures show similar rankings of the methods for each scene, but there is a large variance in subjective quality between scenes. For better comparison of the methods, the left plot of Figure 8.4 shows the same results aggregated across all contents and exposures. All results are shifted with respect to the input SDR image, so that positive just objectionable difference (JOD) values indicate improvement in quality and negative values indicate degradation of quality. The right plot indicates the percentage of images that were assessed to be better, same or worse than the SDR input image. Per image JOD of more than 0.5 or less than -0.5 was used to distinguish between better and worse. The results show the rather disappointing performance of SI-HDR methods — many methods more often degrade the quality of an input image rather than improve it. This level of performance shows the difficulty of the task. It is worth noting that high fail rate is not uncommon for similar challenging tasks, such as image inpainting, and such techniques are still useful, but they require manual screening of the results.

I encourage the reader to inspect individual images included in the interactive HTML viewer on the project web page*. Image 177 serves as an example of successful reconstruction. The top row of Figure 8.5 shows the reconstructed portion of this image for all tested methods. Even though the reconstructed colours are quite different from the HDR reference, all the reconstructions improve on the SDR input image with clipped colours. In the middle row of Figure 8.5 is image 052, for which ExpandNet, HDR-CNN and MaskHDR show moderate improvement by boosting the saturated region in a convincing

*`https://www.cl.cam.ac.uk/research/rainbow/projects/sihdr_benchmark/external/index.html`

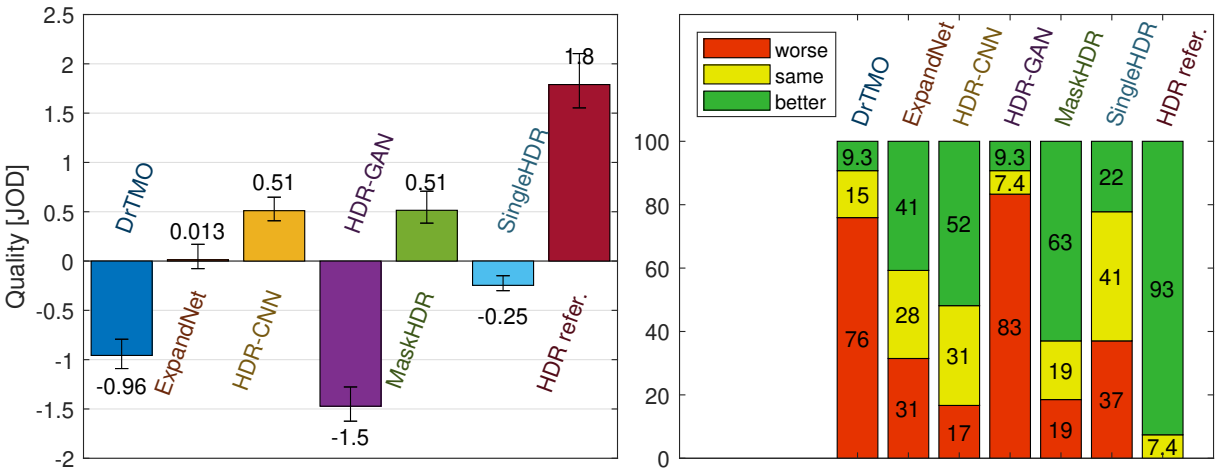


Figure 8.4: Preference of the SI-HDR method results. Left: The bars indicate the preference in JOD units, relative to the source SDR image. Negative values indicate that on average the method produced less preferable result than the non-processed source image. Right: The bars indicate the percentage of images in which the method produced better, same or worse image than the input SDR image.

manner, but DrTMO and HDR-GAN introduce objectionable artefacts. Finally, in the bottom row of Figure 8.5, all methods failed to reconstruct large saturated regions for image 123. The JOD scores in Figure 8.3 reflect the plausibility of hallucinated regions.

8.7 Quality metrics for SI-HDR

To identify an automatic evaluation procedure that correlates with the results from Section 8.6.2, I start by revisiting existing SI-HDR evaluation methods. Then, I propose an improved protocol, which accounts for the random nature of quality assessment and the inaccuracies of quality metrics.

8.7.1 Existing metrics

I identified a range of metrics, both full-reference, and no-reference, listed in Table 8.1. All the metrics either supported direct comparison of HDR images (HDR-VDP-2 and -3, FovVideoVDP) or were suitably adopted. For this, I used the PU21 transform [104], μ-transform [71] or a global tone-mapping operator [133]. PU21 maps linear RGB colour values into approximately perceptually uniform units (accounting for glare and contrast sensitivity). The μ-transform offers a logarithmic scaling with empirically selected parameters. The images passed to all metrics were scaled in absolute units representing physical colour emitted from the display. Finally, values above 1000 cd/m^2 were clamped to match the peak luminance of the display used in the experiment.

Quality can be assessed both with respect to the HDR reference image (reconstruction

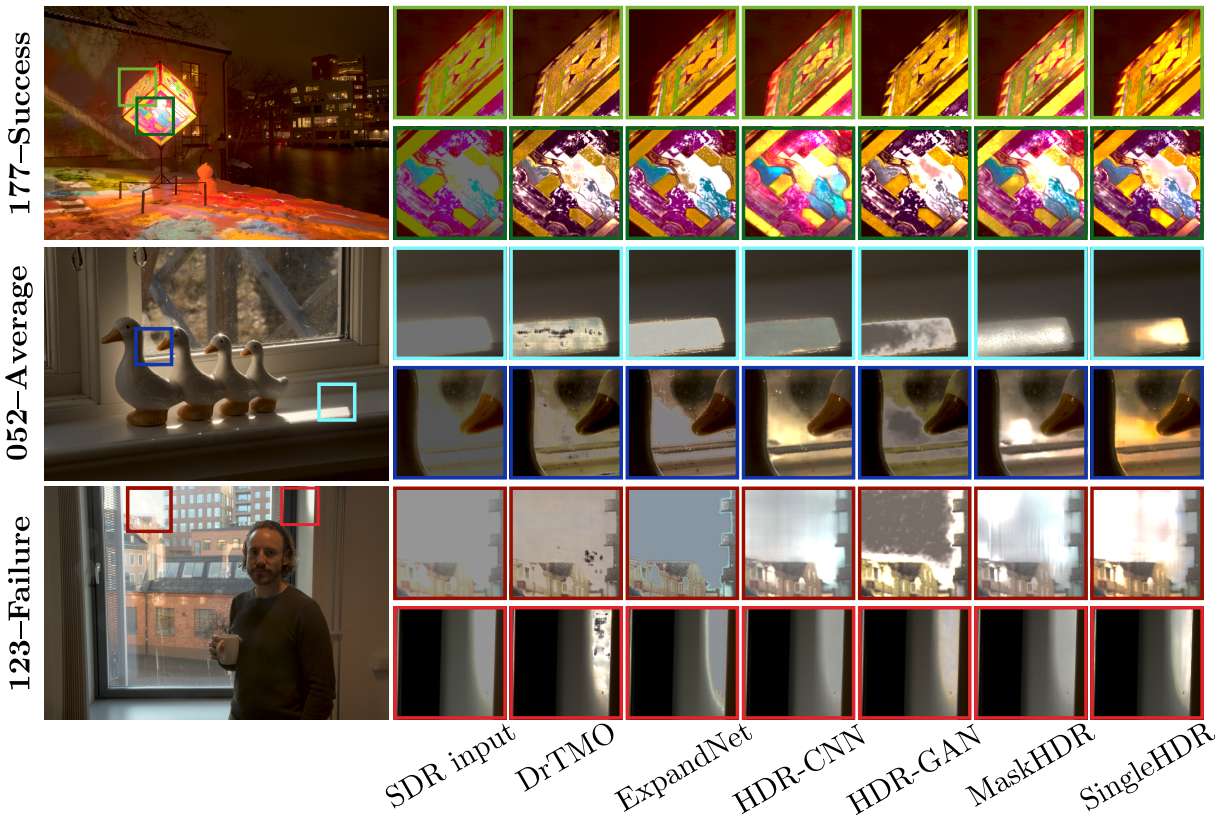


Figure 8.5: Reconstructions by various methods of selected scenes from the subjective data. Please check the interactive HTML viewer on the project page: `https://www.cl.cam.ac.uk/research/rainbow/projects/sihdr_benchmark/external/index.html` to assess the quality of reconstructions for other scenes.

Table 8.1: List of quality metrics used in the evaluation.

Metric	Reference required	HDR metric	Details
PSNR	✓	✗	Widely used ratio to measure noise relative to the signal in log units
SSIM [163]	✓	✗	Popular quality measure that relies on block-wise correlations
MS-SSIM [162]	✓	✗	Multi-scale version of SSIM
VIF [145]	✓	✗	Employs natural scene statistics (NSS) models to measure enhancement
FSIM [173]	✓	✗	Low-level feature similarity index based on the visual system
VSI [174]	✓	✗	Weighted average of local quality maps guided by visual saliency
LPIPS [175]	✓	✗	Perceptual similarity metric across CNN architectures
HDR-VDP-2 and 3 [106]	✓	✓	Low-level vision model, works on HDR images
FovVideoVDP [105]	✓	✓	Low-level vision model for images, video and foveation
BRISQUE [115]	✗	✗	Support vector regression trained on IQA dataset
NIQE [116]	✗	✗	Distance between NSS-based features to those from a database
PIQE [159]	✗	✗	Averaged block-wise distortion estimation
NIMA [148]	✗	✗	Object recognition CNNs re-purposed as a blind metric

task) and relative to the input SDR image (enhancement task). To test whether the latter is more suitable, I included VIF, a metric capable of assessing image enhancement. I append -SDR to the metric name when SDR input is used as a reference and -HDR otherwise.

8.7.2 Existing validation protocols

To test the suitability of popular metrics, I followed the standard protocol used to evaluate image quality metrics [128], and computed the Spearman rank-order correlations between metric predictions and JOD values for individual conditions. This yielded very low correlation values as reported in Table 8.2, especially for no-reference metrics. The highest correlation was 0.47 for PU21-PSNR. When computed for the images after CRF correction (Section 8.5), the highest correlation increased to 0.55 for HDR-VDP-3, which is still too low to provide meaningful predictions. This initial assessment hints that none of the existing metrics is suitable for predicting the quality of images reconstructed by SI-HDR methods.

However, this evaluation protocol does not reflect the standard quantitative evaluation for SI-HDR methods. SI-HDR methods are typically compared by averaging metric scores

Table 8.2: Correlation coefficients (Spearman) between metric predictions and subjective data both without and with CRF inversion, computed for individual conditions.

Metric	Without CRF correction	With polynomial CRF correction
PU21-PSNR	0.472	0.4
PU21-SSIM	0.314	0.303
PU21-LPIPS	0.397	0.394
HDR-VDP-2 (Q)	0.237	0.434
HDR-VDP-3 (Q)	0.452	0.548
FovVideoVDP	0.466	0.366
PU21-MS-SSIM	0.425	0.482
PU21-FSIM	0.473	0.55
PU21-VSI	0.438	0.451
PU21-VIF-HDR	0.393	0.388
PU21-VIF-SDR	0.348	0.324
PU21-PIQE	0.13	0.0791
PU21-BRISQUE	0.0771	0.0405
PU21-NIQUE	0.0632	0.0158
PU21-NIMA	0.377	0.14
μ-PSNR	0.372	0.373
μ-SSIM	0.343	0.304
μ-LPIPS	0.4	0.353
μ-VSI	0.396	0.402
μ-PIQE	0.129	0.0799
Reinhard-PSNR	0.393	0.357
Reinhard-SSIM	0.398	0.26
Reinhard-LPIPS	0.332	0.334
Reinhard-VSI	0.391	0.406
Reinhard-PIQE	0.109	0.0791
Linear-PSNR	0.35	0.271
Linear-SSIM	0.487	0.375
Linear-LPIPS	0.183	0.16
Linear-FSIM	0.513	0.476
Linear-VSI	0.256	0.277
Linear-PIQE	0.13	0.0791

across multiple images and comparing the means. When the correlations of averaged metric scores with subjective JOD values are computed for all the scenes (plotted in Figure 8.4), they are much higher. This is probably due to the poor performance of quality metrics in assessing the absolute level of impairment. Each metric introduces per-content bias, which reduces correlation with the subjective data. However, such biases cancel out when the quality values are averaged across content, and the resulting scores correlate much better with the subjective data.

A major caveat of using averaged data to assess quality metrics is that the correlations (or any other measures of performance) are computed on very few values[†]. A correlation computed on only 7 values is very sensitive to small changes in the subjective data or a different selection of images. To ensure a fair assessment for each metric given our data, 2000 bootstrap samples were generated for each estimated correlation by randomising (sampling with replacement) both the participants and the selection of images [119]. Each sample involved independently scaling the JOD values using a subset of data. The bootstrapping simulated 2000 outcomes of the experiment to capture the variance that is to be expected due to measurement noise.

The correlations for averaged metric scores, shown in Figure 8.6a, are higher especially for the no-reference PU21-PIQE metric ($\rho = 0.83$) and PU21-VSI ($\rho = 0.78$). However, these are not metrics typically used to evaluate SDR methods. Interestingly, the correlations for metrics used in SDR papers (PU21-PSNR, PU21-SSIM, PU21-LPIPS, and HDR-VDP-2, and 3) are all below 0.64. We can thus conclude that the metrics used in the SDR papers were not sensitive enough to differentiate the compared methods and, therefore, could not provide evidence for the improvement of the proposed methods. This observation explains why the participants disagreed with quantitative results in recent SDR papers (see Section 8.6.2).

8.7.3 Improved protocol for SI-HDR evaluation

First, we confirm the observation from [34] that the inaccuracy in the CRF reconstruction is the major reason for the poor metric predictions. When metric scores are computed on images with the CRF correction, we see a major gain in metric performance (compare Figure 8.6a and Figure 8.6b). This is because even small differences in the CRF, sometimes unnoticeable to the human eye, result in large metric errors as it affects all pixels in the image. With the CRF correction, the correlation of PU21-PSNR jumps from 0.62 to 0.79 and a similar trend is seen for other metrics. I also attempted computing metric values only on saturated areas by copying non-saturated pixels from the reference to the test image, but noticed no improvement in metric predictions. The existing metrics do not seem to cope well with localised distortions. I intentionally show distributions of

[†]6 SI-HDR methods + 1 SDR input image = 7 total.

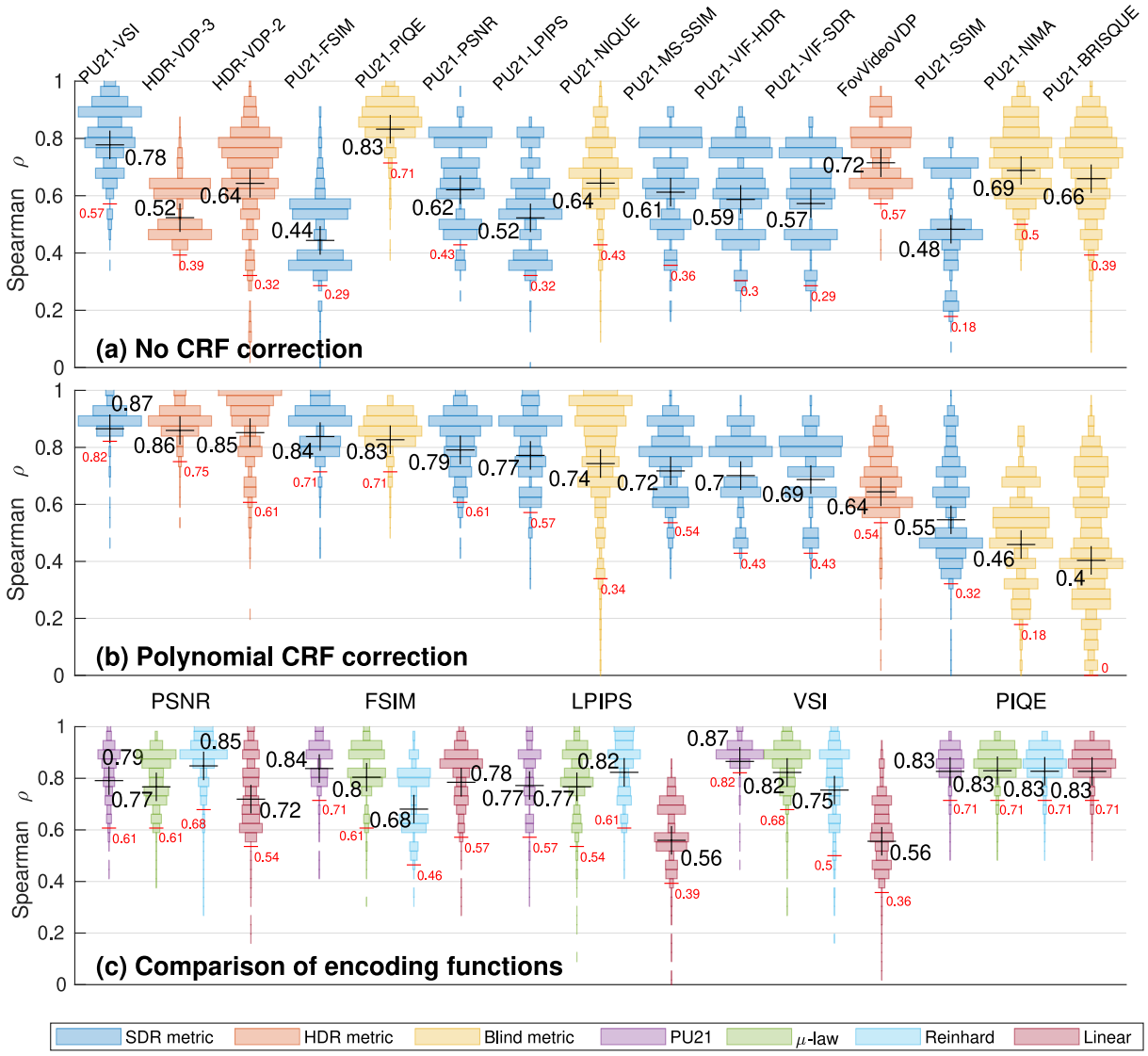


Figure 8.6: Bootstrapped distributions of correlation coefficients for all metrics compared (a) before and (b) after polynomial correction. For selected metrics, (c) demonstrates the effect of different HDR-SDR transforms. "+" denotes unbiased mean correlation [124] and "−" (in red) denotes the 5th percentile (an estimate of the bad-case performance).

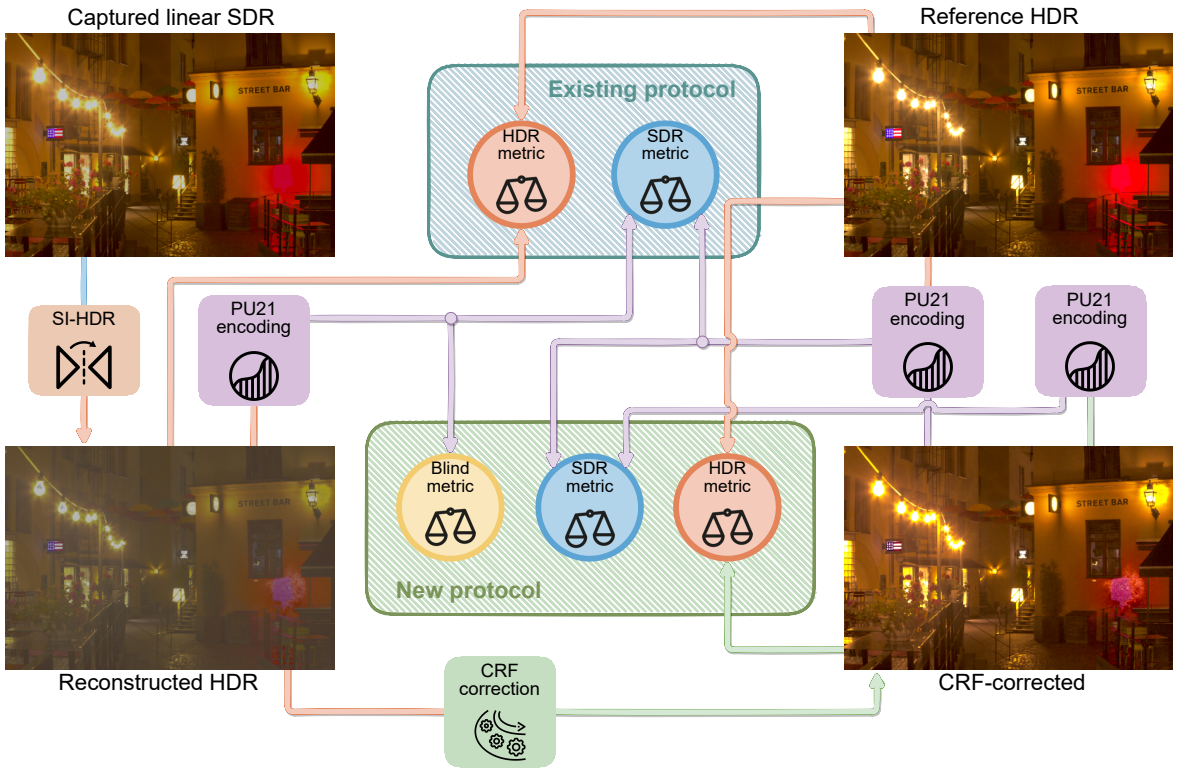


Figure 8.7: Existing protocols for evaluating single-image HDR reconstruction methods directly compare the reconstructed HDR images with the reference, as depicted by the blue shaded rectangle. This is unreliable due to large tone and colour differences between the reference and reconstructed HDR images. I demonstrate that the accuracy of metrics can be much improved if camera-response-curve inversion errors are corrected before computing image quality using existing full-reference metrics as shown in the green shaded rectangle.

correlation coefficients in Figure 8.6 instead of simple scatter plots to stress out that any metric evaluation study comes with a high degree of uncertainty.

It is also relevant to determine what is the best adaptation of the SDR metrics to HDR images: PU21 transform [104], μ-transform [71] or Reinhard et al. [133] global tone-mapping operator or perhaps no adaptation (a metric computed on linear values). From the results in Figure 8.6c, we see that there is insufficient evidence for a significant difference between PU21 and μ-law. However, note that computing metrics on linear values almost always results in significantly worse performance. Using global tone-mapping also reduced performance of several metrics. In the following analysis, I stick to PU21 due to its stronger perceptual basis.

Our recommendation for the new protocol is to report the results for PU21-PSNR, because of its simplicity and good performance, for PU-VSI and HDR-VDP-3, since they are likely to perform well, and for PU21-PIQE as this is a well performing no-reference metric. CRF correction must be performed for all metrics except PU21-PIQE, which performs well without such correction. These results motivate the need for an HDR

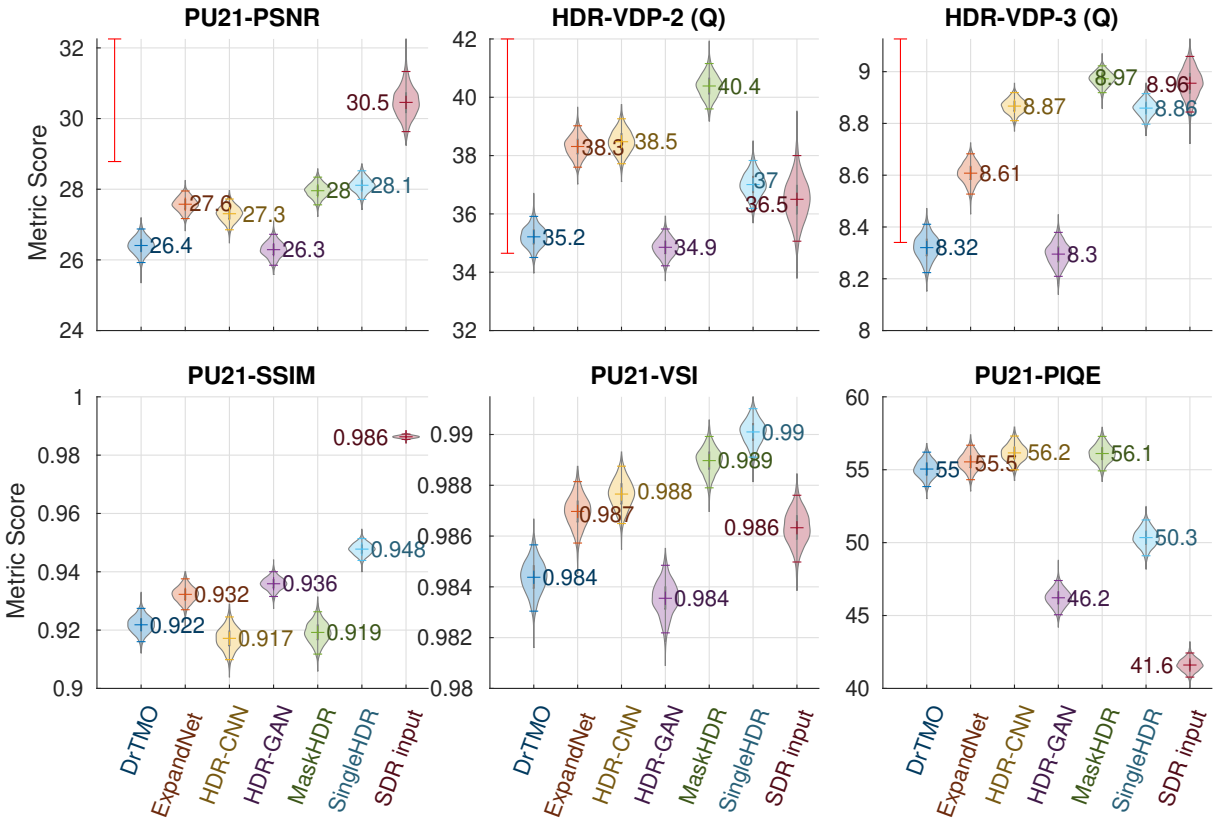


Figure 8.8: Ranking bootstrapped distributions for SI-HDR methods on the validation dataset. For each distribution, small dashes denote 95% confidence intervals, while the red errors bars show the minimum measurable increment in quality for selected metrics (described in Section 8.7.5).

no-reference metric invariant to the type of distortion. A candidate, proposed by [15], is restricted to a single kind of distortion. We do not recommend using PU21-SSIM and PU21-NIMA.

8.7.4 Validation of quality metrics

Here I investigate whether the selected metrics used with the new, improved protocol with CRF correction yield good enough predictions to measure the performance improvement of SI-HDR methods. To evaluate the methods, I used the remaining 156 images from our dataset (excluding images depicted in Figure 8.2 since they were part of the experiment). If the metrics were accurate, the results shown in Figure 8.8 should match the subjective results from Figure 8.4. While most metrics capture the trend, many incorrectly predicted the quality of the input SDR image. Additionally, contrary to the subjective data, a few metrics assigned higher quality to the SingleHDR method. Relying on metrics would thus result in an incorrect ranking of methods with a high likelihood.

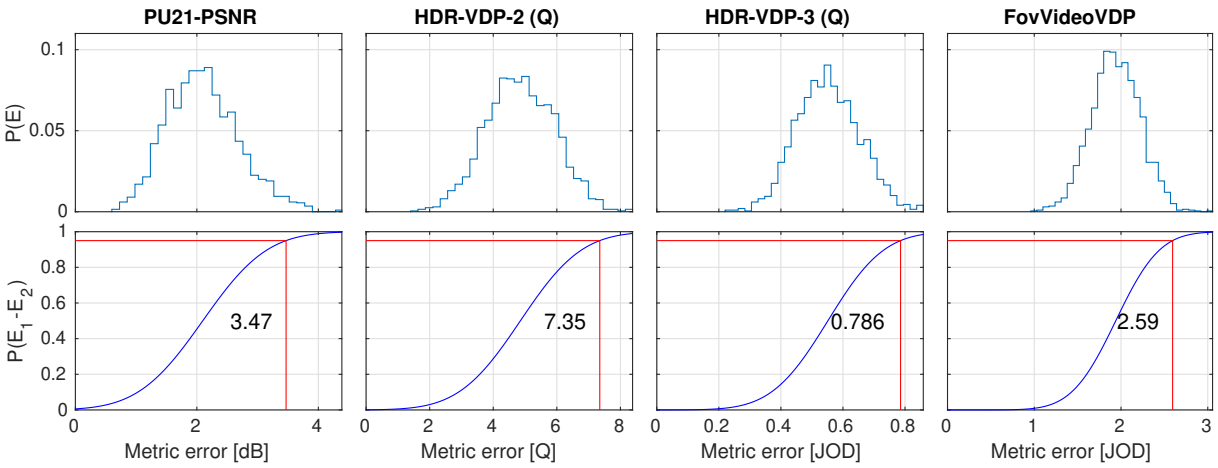


Figure 8.9: Estimated prediction error for 4 selected metrics. Similar to Figure 8.6, the distributions were obtained by bootstrapping the experiment results over the subset of 27 images. The improvement in quality metric values reported in many papers falls below the expected accuracy of each metric.

8.7.5 Minimum measurable increment of quality

Correlation coefficients are difficult to interpret. Even for the metric with the highest expected correlation (PU21-VSI $\rho = 0.87$), it is impossible to say whether this value is good enough to evaluate SI-HDR methods. Instead, we want to determine the smallest difference in metric scores that can tell us with confidence (at $\alpha = 0.05$) that one method is better than the other. To do that, we need to account for all sources of measurement error: (a) due to the selection of images; (b) due to the measurement error in subjective experiment results (error bars in Figure 8.4); and (c) due to the inaccuracy of a metric. Hidden in the metric error (c) is an inherent difficulty related to the ill-posed nature of SI-HDR task. In saturated regions, hallucinated details may be different to the reference, yet perfectly plausible. To obtain a comprehensive measure, I bootstrapped the RMSE values. For each bootstrap sample, I found a linear mapping from JOD values to metric predictions and then computed RMSE in terms of the metric error.

Such an analysis can be applied only to the metrics that produce quality values that are (approximately) linearly related to the perceived magnitude of quality (in JOD). For example, the strongly non-linear relation between SSIM values and MOS means that a small difference in SSIM values has a different impact on perceived quality depending on the absolute SSIM values. Therefore, this analysis includes only PU21-PSNR, HDR-VDP-2, HDR-VDP-3, and FovVideoVDP, which are all designed to be well correlated with the perceived magnitude of distortion.

The distributions of the prediction errors are shown in the top row of Figure 8.9. These distributions could be interpreted as an expected metric error, E, with respect to the subjective scores. To use a metric to compare two methods, we are interested in the

error for the difference between two quality metric scores. Since the distributions are normal (Kolmogorov-Smirnov test, at $\alpha = 0.05$ significance level), the mean and standard deviation for the bootstrapped samples were estimated. The bottom row of Figure 8.9 shows the cumulative distribution for the difference in the error estimates (also normal with the same mean and standard deviation multiplied by $\sqrt{2}$). The plots show that we need a PU21-PSNR difference of at least 3.5 dB to be confident that the method with higher PSNR is on average better (at 5% chance of making an error). The required minimum differences are also high for other metrics. Since the improvement in quality reported in most papers falls below these amounts, it casts doubts on the reliability of the evaluation performed solely with objective quality metrics.

8.8 Summary

The evidence from a controlled subjective quality assessment experiment indicates that the overall progress in single-image HDR (SI-HDR) reconstruction is less impressive than reported in the papers. The results indicate little improvement in the quality of SI-HDR results over the last five years. In fact, a few of the methods were more likely to degrade image quality than improve it.

A critical reason for the lack of visible progress is the wrong use of quality metrics and insufficient subjective evaluation. Many metrics commonly used to evaluate SI-HDR methods have low correlations with subjective scores ($\rho = 0.62$ for PU21-PSNR). Instead, I propose to compute metric scores on images that have been corrected for errors in the CRF inversion. After applying CRF correction, PU21-VSI and HDR-VDP-3 were the best performing metrics. The no-reference metric PU21-PIQE, which does not require CRF correction and PU21-PSNR also perform well and should be included.

Finally, I demonstrate that even the best metrics computed on CRF-corrected images introduce a substantial prediction error. Using non-parametric statistics and subjective results, I estimated the minimum improvement in quality scores required to find the difference between two SI-HDR methods at $\alpha = 0.05$ to be 3.5 dB for PU21-PSNR. Since most papers report much smaller improvement in quality, there is a high risk that the reported improvement is due to random errors rather than actual gains in method performance.

Chapter 9

Modelling style for inverse tone-mapping

Several applications in image and video processing, including inverse tone-mapping, require mapping source to target colours. Since the focus of inverse tone-mapping (iTM) is not the recovery of tone and colours (as explained in Chapter 8), global operations produce good results. However, most existing works either propose hand-crafted recipes [13, 14, 110, 134] or train CNNs [77, 78] to learn the one-to-one mapping using paired data.

The notable conceptual flaw with this approach is that inverse tone-mapping is an ill-posed problem and does not have a single solution. If the task is given to several skilled colour artists, each one is likely to produce a different result that conveys a unique style. Methods that automate this process should respect and preserve such intended style. Because the style information is not present in the input image, existing methods that learn an image-to-image mapping can learn at most a single style [47].

Contrary to existing works, I show that information about the style can be distilled from collections of image pairs and encoded into a 2- or 3-dimensional vector. This gives us not only an efficient representation but also an interpretable latent space for editing the image style. The global colour mapping between a pair of images is represented with a custom normalizing flow, conditioned on a polynomial basis of the pixel colour.

9.1 Methodology

Our objective is to model a global, spatially invariant colour mapping between SDR-HDR pairs of images or video frames. By introducing a low-dimensional *style* vector, a single mapping should generalise to all content. Let $\{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^N$* represent N paired images. Each

*In all previous chapters, x and y represented scaled (by exposure time and gain) and original pixel values respectively. In this chapter alone, $\mathbf{x}$ and $\mathbf{y}$ are used slightly differently, to represent SDR and HDR images respectively, for consistency with other learning-based image-to-image translation works. Notice the different font face to distinguish this usage.

RGB pixel p in those pairs is related by a global mapping $\mathcal{M}$:

$$\mathbf{x}_p^i = \mathcal{M}(\mathbf{y}_p^i, \mathbf{z}^i)\,, \tag{9.1}$$

where i is the image index and $\mathbf{z}^i$ is the low-dimensional vector that encodes *style* parameters specific to each pair. Further, $\mathbf{y}$ represents SDR and $\mathbf{x}$ represents HDR images.

Such a mapping can be effectively expressed as a parametric function using polynomial colour correction (PCC) basis functions [40, 41] for SDR pixels:

$$\mathbf{x}^i = \mathcal{C}(\mathbf{y}^i) \cdot \mathbf{M}^i\,, \tag{9.2}$$

where $\mathcal{C}(\mathbf{y}^i)$ is a matrix with the polynomial basis functions:

$$\mathcal{C}(\mathbf{y}_p^i) = \mathcal{C}([r, g, b]) = [r, g, b, r^2, g^2, b^2, rg, gb, br, \ldots, r^4, g^4, b^4]\,, \tag{9.3}$$

and $\mathbf{M}^i$ is the 34×3 style matrix (a representation of $\mathbf{z}^i$). At least a 4^{th} degree PCC with $34 \times 3 = 102$ parameters was found to be necessary for the diverse set of scenes considered. However, such a large latent representation is not interpretable and cannot be intuitively modified to edit the style. The straightforward to reduce this latent is to use a dimensionality-reduction technique such as principal component analysis (PCA) or variational autoencoder (VAE) [80]. However, both were ineffective for this problem (see Section 9.2.2). Instead, I propose to use a custom adaptation of normalizing flows [135] to represent the style mapping.

9.1.1 Approach

It is useful to cast the problem of finding a suitable approximation of $\mathcal{M}$ from Eq. (9.1) as a generative modelling task using supervised learning. We can learn a conditional mapping from a known latent distribution to the target domain using a parametric model $g(\cdot)$. The approximate model $g(\cdot)$ together with the extracted latent vector $\mathbf{z}^i$ reproduce the required per-pixel SDR-HDR mapping,

$$\mathbf{x}_p^i = g(\mathbf{z}^i; \mathbf{c}_p^i)\,, \tag{9.4}$$

where the conditioning vector contains the polynomial basis functions: $\mathbf{c}_p^i = \mathcal{C}(\mathbf{y}_p^i)$. Intuitively, when the polynomial basis is supplied, the network only needs to learn the appropriate linear combination, similar to matrix $\mathbf{M}^i$ from Eq. (9.2). This dimensionality expansion approach is akin to positional encoding, used to represent multi-dimensional functions in neural radiance fields (NeRF) [112, 150].

9.1.2 The latent distribution model

To approximate the global mapping $\mathcal{M}$ with a parametric per-pixel function, an ideal choice is the class of generative models called normalizing flows [29, 135] introduced in Section 2.3.2, Chapter 2. They employ invertible layers with tractable determinants of jacobian to learn a non-trivial mapping from a known latent to the target distribution. The bidirectionality of normalizing flows makes them particularly suitable since they can both

1. extract a latent *style* vector from a set of pixels (reverse discriminative pass), and
2. reconstruct a pixel from the given input and latent vector (forward generative pass).

Given many samples from the target HDR distribution X, the objective is to train a deep invertible neural network (INN) $g_\theta(\cdot)$[†] using maximum likelihood estimation (MLE) to convert them into samples of the latent distribution Z. Since $g_\theta(\cdot)$ is invertible, the change of variables formula given by Eq. (2.8) expresses the probability of a target sample $\boldsymbol{x} \sim X$ as a function of the probability of the transformed latent $\boldsymbol{z} \sim Z$. The task at hand requires modelling the multidimensional conditional distribution:

$$p_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = p_{Z|Y}\left(g_\theta^{-1}(\boldsymbol{x};\boldsymbol{c})\right) \cdot \left|\det \nabla g_\theta^{-1}(\boldsymbol{x};\boldsymbol{c})\right| . \tag{9.5}$$

Here $\boldsymbol{c}$ is a conditioning vector extracted from SDR pixels, which corresponds to the polynomial coefficients $\boldsymbol{M}$ given by Eq. (9.1).

9.1.3 Network architecture

To successfully learn to map complex distributions, the parametric model needs to be:

- *Invertible*: to both extract latents (reverse) and reconstruct pixels (forward)
- *Differentiable*: for training a large number of parameters by backpropagation
- *Expressive*: to capture the complexity of the target distribution
- *Tractable*: the likelihood (Eq. (9.5)) contains a determinant of a large matrix, which needs to be efficiently computed in linear time

In practice, normalizing flows are constructed by composing a series of invertible operations [9, 29, 81, 96, 135]. The INN used to learn global colour mappings, depicted in Figure 9.1, is composed of a series of 8 invertible blocks each consisting of (1) affine coupling layer, (2) random permutation and (3) batch normalization with ActNorm [81].

[†] θ represents weights of the network trained by backpropagation

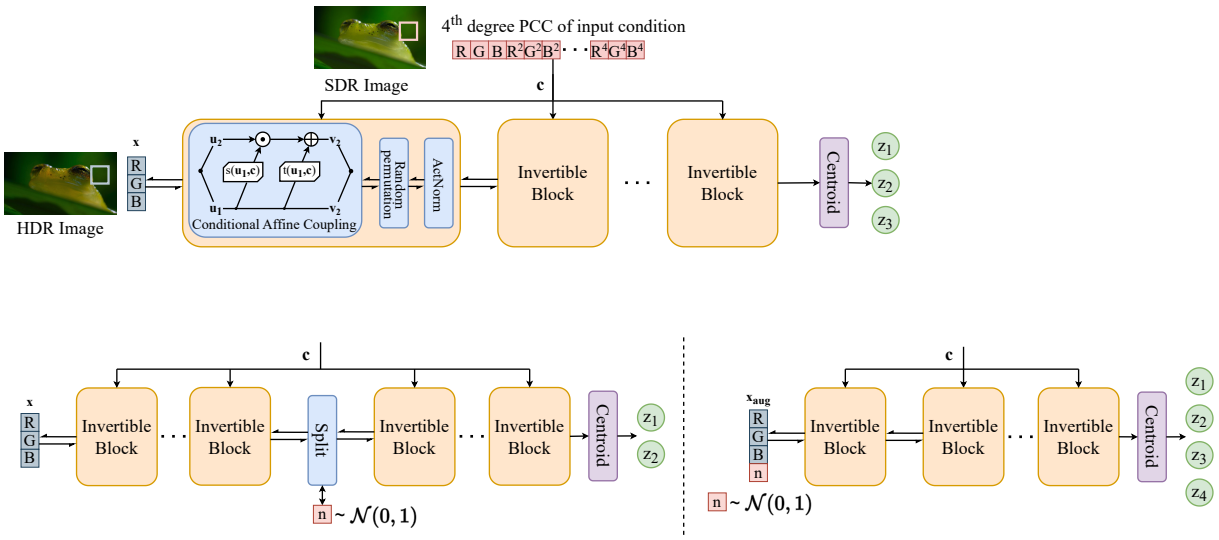


Figure 9.1: The network architecture used for the pixel-wise conditional INN. Here the sub-networks $s(\cdot)$ and $t(\cdot)$ are fully connected networks with 2 hidden layers each. I use 8 invertible blocks for all results.

The coupling layers [29] increase the expressive power of the INN by incorporating complex, non-invertible sub-networks $s(\cdot)$ and $t(\cdot)$. To keep the total number of trainable parameters small, both $s(\cdot)$ and $t(\cdot)$ are limited to simple 2-layer MLPs.

Conditional coupling Affine coupling [29] is a key operation of most normalizing flow architectures that increases the expressibility of the network for unconditional generative modelling. This involves splitting the inputs $\mathbf{u}$ of an intermediate layer into two equal parts $\mathbf{u}_1$ and $\mathbf{u}_2$. The first, $\mathbf{u}_1$ is transmitted unchanged, while $\mathbf{u}_2$ goes through an affine transformation where parameters are some functions of $\mathbf{u}_1$, realised through complex, non-linear sub-networks $s(\mathbf{u}_1)$ and $t(\mathbf{u}_1)$.

However, we are interested in modelling the conditional distribution $p_{X|Y}(\mathbf{x}|\mathbf{c})$, the conditional extension described in [9, 96] is more suited. Here, the conditional vector $\mathbf{c}$ from Eq. (9.3) is appended to the inputs of $s(\cdot)$ and $t(\cdot)$. Now, the output of the conditional coupling layer is given by the concatenated vector $[\mathbf{v_1}, \mathbf{v_2}]$ where

$$\mathbf{v_1} = \mathbf{u_1}\,, \qquad \mathbf{v_2} = \mathbf{u_2} \odot \exp(s([\mathbf{u_1}, \mathbf{c}])) + t([\mathbf{u_1}, \mathbf{c}])\,. \tag{9.6}$$

Since the sub-networks in a coupling block are never inverted themselves, they can be appended with $\mathbf{c}$ without losing the invertibility of the INN.

9.1.4 Optimization and inference

Similar to other normalizing flows, the INN for global iTM is trained by MLE where the likelihood is given by Eq. (9.5). The network thus learns to conditionally map a target

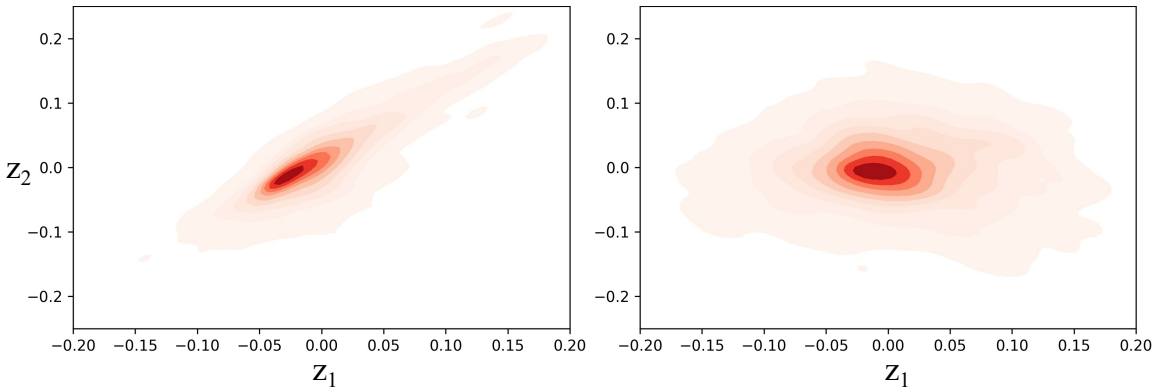


Figure 9.2: The figure shows the distribution of the latent colour representation for different frames from the training set with (left) and without (right) the reconstruction loss $\mathcal{L}_{\text{rec}}$. We see that the latent vectors of the training set follow a normal distribution. Note that the iTM INN was trained to encode the style of frames to 2 dimensions for better visualisation.

HDR pixel to a latent vector. However, when presented with an entire SDR frame, there is no easy way to extract a single low-dimensional vector that captures style. For that, MLE training is augmented with a reconstruction loss to force the per-pixel representations of the same frame to lie closer in the latent space.

Likelihood loss To restrict the magnitude of gradients for backpropagation, it is customary to minimise the negative log-likelihood (NLL) instead of likelihood since the logarithmic transform is monotonic. Assuming a standard normal latent space, where each latent vector $\mathbf{z}_p \sim \mathcal{N}(0, 1)$, the NLL loss is:

$$\begin{aligned} \mathcal{L}_{\text{NLL}}(\mathbf{x}_p, \mathbf{y}_p) &= -\log p_{Z|Y}\left(g_\theta^{-1}(\mathbf{x}_p; \mathbf{c}_p)\right) - \log \left|\nabla g_\theta^{-1}(\mathbf{x}_p; \mathbf{c}_p)\right| \\ &= \frac{\log 2\pi + \left(g_\theta^{-1}(\mathbf{x}_p; \mathbf{c}_p)\right)^2}{2} - \log \left|\nabla g_\theta^{-1}(\mathbf{x}_p; \mathbf{c}_p)\right| \end{aligned} \tag{9.7}$$

When trained using this loss, the iTM INN learns a bijective mapping from the distribution of HDR pixels to the chosen latent, conditioned on the encoded input SDR pixel. Through MLE, per-pixel latents are encouraged to follow a standard normal distribution as confirmed by Figure 9.2 (right).

Reconstruction loss The secondary requirement is for pixels of the same HDR frame to cluster together in the latent space Z. This is achieved by first passing all K pixels of frame i through the INN in reverse, and computing the centroid of their latent representations,

$$\mathbf{z}^i = \frac{1}{K} \sum_{p=1}^{K} g_\theta^{-1}(\mathbf{x}_p^i; \mathcal{C}(\mathbf{y}_p^i)) \,, \tag{9.8}$$

where $\mathbf{z}^i$ is the per-frame *style* vector. Then, the frame can be reconstructed by with K forward passes[‡] of the INN using the single extracted *style* latent $\mathbf{z}^i$ with different conditional inputs. Thus, the reconstruction loss $\mathcal{L}_{\text{rec}}$ for each pixel of an image is:

$$\mathcal{L}_{\text{rec}}(\mathbf{x}_p^i, \mathbf{y}_p^i) = \| \, g_\theta(\mathbf{z}^i; \mathcal{C}(\mathbf{y}_p^i)) - \mathbf{x}_p^i \, \|_2 \tag{9.9}$$

This reconstruction loss enforces a similar value of the style vector $\mathbf{z}^i$ for the entire frame.

Optimisation of the INN is done bi-directionally with the total loss being the sum of the NLL loss $\mathcal{L}_{\text{NLL}}$ and the reconstruction loss $\mathcal{L}_{\text{rec}}$. After successful training, the INN can be used to extract an overall per-frame *style* vector. This is achieved by running a reverse pass for each pixel of a given image and computing the centroid according to Eq. (9.8). Figure 9.2 shows the change in the latent space for the training samples due to the addition of $\mathcal{L}_{\text{rec}}$.

9.2 Results

Although the described framework is suitable for any global colour mappings, I only present results for global inverse tone-mapping (iTM) since this was my primary contribution to the project. Please see the paper [121] for further experiments related to forward tone mapping and interactive colour grading.

9.2.1 Dataset

3 Blu-ray movies, namely "BBC Planet Earth II Episode 3 - Jungles", "BBC Blue Planet II Episode 5 - Green Seas" and "The Lego Batman Movie" were decoded and colour graded separately for SDR and HDR displays. The frames from SDR and HDR streams were time-synchronised by finding the offset that maximised cross-correlation. Finally, the frames were manually inspected to ensure close correspondence. For good diversity, a sequence is constructed from each video by collecting every 120^{th} frame. For each such video sequence, the first 80% of frames are used for training and the rest 20% for testing. On average, the dataset contains 500-800 frames per movie.

9.2.2 Other dimensionality reduction techniques

A simplified view of the learned INN is that it approximates a linear relationship between the conditioned input image $\mathcal{C}(\mathbf{y})$ and output image, $\mathbf{x} = \mathcal{C}(\mathbf{y}) \cdot \mathbf{M}$. For input $\mathbf{y}$, the INN simultaneously predicts and distills this matrix into a compact latent *style* vector

[‡]This is implemented as a single forward pass with a batch of K.

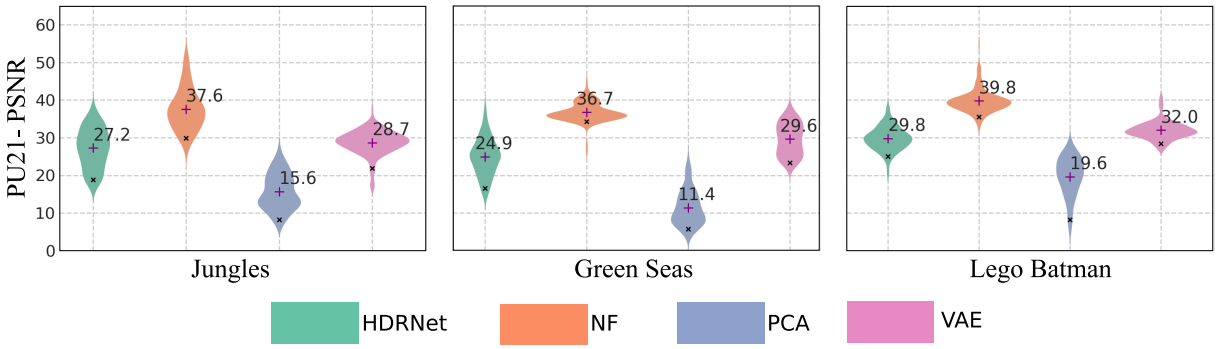


Figure 9.3: Comparison of results on different datasets for the task of inverse tone-mapping on the movies dataset. Note that the normalizing flow based method (orange) achieves a substantial improvement in performance compared to other dimensionality reduction methods across datasets. The perceptually uniform PU21 transform [104] adapts PSNR to HDR images. The purple '+' in the plots show the mean and the black '×' show the lowest 5^{th} percentiles.

$\mathbf{z}$. Here, I compare the INN with the alternate approach of condensing $\mathbf{M}$ by popular dimensionality reduction methods.

Principal component analysis: PCA was run on training pairs to reduce the flattened *style* matrix $\mathbf{M}_{\text{flat}}$ into the required number of latent dimensions. During test time, $\hat{\mathbf{M}}_{\text{flat}}$ is reconstructed from the principal components and used to map the colours.

Variational autoencoder: Another candidate deep generative model tested was a VAE [80], where the latent follows a normal distribution. The input to the VAE is the original matrix of polynomial coefficients $\mathbf{M}$. The goal is to train the VAE so that the matrix can be predicted from a low-dimensional latent vector. The training loss includes a reconstruction loss between predicted (after matrix multiplication using decoded matrix $\hat{\mathbf{M}}$) and ground truth pixels and the evidence lower-bound. For a fair comparison with the normalizing flow, a fully connected network with approximately the same number of parameters was chosen.

9.2.3 Quantitative comparison

Inverse tone-mapping often involves bit-depth expansion and hallucination of over- and under-exposed pixels (see Section 8.1). The focus of this chapter is on the problem of learning global SDR→HDR colour mapping. As a representative example of existing learning-based iTM, the normalizing flow (NF in Figure 9.3) is compared with HDRNet [47] retrained on the movies dataset. The numerical results and quality metric distributions, shown in Figure 9.3, demonstrate a dramatic improvement of over 10 dB relative to all other dimension reduction methods.

We observe that the strict linearity assumption of PCA results in poor performance.

Although the VAE attains higher quality scores than PCA and its results are comparable to HDRNet (which does not use a style vector), VAE still performs much worse than the iTM INN.

9.3 Summary

In this final chapter, I have emphasised the significance of modelling the one-to-many relationship inherent in the inverse tone-mapping problem. The proposed conditional INN effectively models this by extracting and encoding artistic choice from examples of image pairs into a low-dimensional style vector. I showed that our method significantly outperforms representative deep architectures that ignore style, as well as alternate dimensionality reduction methods that incorporate latent style.

Our primary focus has been on colour mapping relevant to video content, which exhibits a global nature (same mapping for all pixels in an image). Consequently, I have conducted experiments using exclusively video datasets. Nevertheless, the method can be readily expanded to facilitate manual, local mapping by dividing frames into multiple tiles [32]. However, when confronted with challenges posed by single-image HDR (refer to Chapter 8) such as recovering details in underexposed and saturated regions, this approach proves to be inadequate. The objective shifts from learning per-pixel distributions to learning the joint distribution of all pixels within an image. The problem also resembles image inpainting, a topic that has received considerable attention in the recent past due to the explosion of diffusion models [63]. Exploring this avenue constitutes an intriguing direction for future research.

Chapter 10

Conclusions

The primary objective of my PhD was to capture and recreate the radiance of the real world. Instead of reconstructing scenes that merely appear realistic, I strived to recover the physically accurate high dynamic range (HDR) information of complex scenes. Achieving this for still scenes involved skipping the typical camera-specific post-processing and, instead, merging a stack of exposure-modulated RAW images. Although the techniques of merging exposure stacks have been around for some time, I discovered that there was still potential for improvement. Finally, my research helped develop a new capture-render-display system that can generate highly detailed and realistic images of complex scenes, to the extent that human observers can be fooled into thinking they are real.

10.1 Key contributions

The following are the key contributions that summarise my work:

Noise-aware merging of exposure stacks

In Chapter 4, I describe several methods to estimate the radiance of a scene. Further, I showed the estimator that models the photon component of camera noise is near-optimal as it is unbiased and has minimal variance. A key advantage is that it does not rely on camera-specific noise parameters even though it is motivated by a statistical noise model. Therefore, the estimator can be used with any camera without a need for noise calibration.

Automatic exposure estimation

Then, in Chapter 5, I show that the camera metadata of an exposure stack may not always be reliable due to various reasons. To remedy this, I propose a robust algorithm to estimate exposure ratios directly from the images. The methods introduced in Chapters 4 and 5 were used to reconstruct accurate HDR scenes for all subsequent experiments.

Applications

The next two chapters, Chapters 6 and 7, highlight real-world applications of the improved reconstruction pipeline. The former focuses on reproducing realistic scenes for humans and also serves as an ideal validation of the pipeline. The latter shows that accurate HDR reconstruction is important for problems in machine vision as well.

Single-image HDR

Finally, I tackle the challenging problem of hallucinating HDR content from a single standard dynamic range (SDR) image in Chapters 8 and 9. A comprehensive subjective study demonstrated that state-of-the-art methods still need to improve. Interestingly, very few works correctly model the one-to-many SDR-to-HDR mapping. Chapter 9 shows an approach to model such an inverse mapping for global operations and serves as a first step in this promising direction.

10.2 Future work

Acquiring accurate HDR information for complex, dynamic scenes continues to pose a challenge. The process of merging an exposure stack, known as HDR deghosting in the literature, necessitates explicit frame alignment or the utilisation of deep neural networks for motion compensation. On the other hand, the alternative approach of estimating high dynamic range (HDR) information from a single image represents an even more formidable problem. As demonstrated in Chapter 8, this approach remains a work in progress, highlighting the need for a better metric in evaluating the HDR quality. Exploring and developing an improved metric in this context would be an interesting direction for future research. Additionally, the application of deep generative models could enhance single-image HDR methods, as these models have already shown excellent results in image inpainting, which bears similarity to this problem. Finally, it is worth noting that high dynamic range (HDR) video presents its own set of challenges, particularly concerning temporal artefacts.

Bibliography

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise Flow: Noise Modeling with Conditional Normalizing Flows. In *International Conference on Computer Vision (ICCV)*, 2019.

[3] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. Study of the digital camera acquisition process and statistical modeling of the sensor raw data. Technical report, Laboratoire Traitement et Communication de l'Information, Instituto de Ingeniería Eléctrica, August 2013. URL `https://hal.archives-ouvertes.fr/hal-00733538`.

[4] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. Best Algorithms for HDR Image Generation. A Study of Performance Bounds. *SIAM Journal on Imaging Sciences*, 7(1):1–34, jan 2014. ISSN 1936-4954.

[5] Kurt Akeley, Simon J. Watt, Ahna Reza Girshick, and Martin S. Banks. A stereo display prototype with multiple focal distances. *ACM Trans. Graph.*, 23(3):804–813, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015804. URL `https://doi.org/10.1145/1015706.1015804`.

[6] Ahmet Oğuz Akyüz and Erik Reinhard. Noise reduction in high dynamic range imaging. *Journal of Visual Communication and Image Representation*, 18(5):366–376, 2007.

[7] Hatem Alismail, Brett Browning, and Simon Lucey. Robust tracking in low light and sudden illumination changes. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 389–398, 2016. doi: 10.1109/3DV.2016.48.

[8] David Alleysson, Sabine Susstrunk, and Jeanny Hérault. Linear demosaicing inspired by the human visual system. *IEEE Transactions on Image Processing*, 14(4):439–449, 2005.

[9] Lynton Ardizzone, Jakob Kruse, Carsten Lüth, Niels Bracher, Carsten Rother, and Ullrich Köthe. Conditional invertible neural networks for diverse image-to-image translation. In *DAGM German Conference on Pattern Recognition*, pages 373–387. Springer, 2020.

[10] Yousef Atoum. *Detecting Objects under Challenging Illumination Conditions*. PhD thesis, Michigan State University, 2018.

[11] Tunç Ozan Aydın, Rafał Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full dynamic range images. In *Human Vision and Electronic Imaging XIII*, Proceedings of SPIE, pages 6806–10, San Jose, USA, January 2008.

[12] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011.

[13] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006.

[14] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 33–41, 2008.

[15] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. Nor-vdpnet: A no-reference high dynamic range quality metric trained on hdr-vdp 2. In *IEEE International Conference on Image Processing (ICIP)*, pages 126–130. IEEE, October 2020. doi: 10.1109/ICIP40778.2020.9191202. URL `http://vcg.isti.cnr.it/Publications/2020/BAMC20`.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 975–984, 2017.

[18] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[19] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020.

[20] Lukas Cerman and Vaclav Hlavac. Exposure time estimation for high dynamic range imaging with hand held camera. In *Proc. of Computer Vision Winter Workshop, Czech Republic*. Citeseer, 2006.

[21] Ke-Chi Chang, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang, and Hwann-Tzong Chen. Learning camera-aware noise models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[22] Jie Chen, Zaifeng Yang, Tsz Nam Chan, Hui Li, Junhui Hou, and Lap-Pui Chau. Attention-guided progressive neural texture fusion for high dynamic range image restoration. *IEEE Transactions on Image Processing*, 31:2661–2672, 2022. doi: 10.1109/TIP.2022.3160070.

[23] Yeyao Chen, Gangyi Jiang, Mei Yu, You Yang, and Yo-Sung Ho. Learning stereo high dynamic range imaging from a pair of cameras with different exposure parameters. *IEEE Transactions on Computational Imaging*, 6:1044–1058, 2020.

[24] Roberto Costantini and Sabine Susstrunk. Virtual sensor design. In *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications V*, volume 5301, pages 408–419. International Society for Optics and Photonics, 2004.

[25] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[26] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[27] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of SIGGRAPH '97*, pages 369–378, Los Angeles, CA. USA, 1997. ACM Press. ISBN 0897918967.

[28] P Didyk, R Mantiuk, M Hein, and H.P. Seidel. Enhancement of Bright Video Features for HDR Displays. *Computer Graphics Forum*, 27(4):1265–1274, jun 2008. ISSN 01677055. doi: 10.1111/j.1467-8659.2008.01265.x. URL `http://blackwell-synergy.com/doi/abs/10.1111/j.1467-8659.2008.01265.x`.

[29] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[30] Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer graphics forum*, volume 22, pages 419–426. Wiley Online Library, 2003.

[31] David Dunn, Cary Tippets, Kent Torell, Petr Kellnhofer, Kaan Akşit, Piotr Didyk, Karol Myszkowski, David Luebke, and Henry Fuchs. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1322–1331, 2017. doi: 10.1109/TVCG.2017.2657058.

[32] Gabriel Eilertsen, Rafał K. Mantiuk, and Jonas Unger. Real-time noise-aware tone mapping. *ACM Trans. Graph.*, 34(6), nov 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818092. URL `https://doi.org/10.1145/2816795.2818092`.

[33] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.

[34] Gabriel Eilertsen, Saghi Hajisharif, Param Hanji, Apostolia Tsirikoglou, Rafal K Mantiuk, and Jonas Unger. How to cheat with metrics in single-image hdr reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3998–4007, 2021.

[35] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (TOG)*, 36(6):1–10, 2017.

[36] David Eppstein. Finding the k smallest spanning trees. In *Scandinavian Workshop on Algorithm Theory*, pages 38–47. Springer, 1990.

[37] Mark D Fairchild. The hdr photographic survey. In *Color and imaging conference*, volume 2007, pages 233–238. Society for Imaging Science and Technology, 2007.

[38] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45103-7.

[39] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, page 249–256, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135211. doi: 10.1145/566570.566573. URL `https://doi.org/10.1145/566570.566573`.

[40] Graham D Finlayson and Mark S Drew. Constrained least-squares regression in color spaces. *Journal of electronic imaging*, 6(4):484–493, 1997.

[41] Graham D Finlayson, Michal Mackiewicz, and Anya Hurlbert. Color correction using root-polynomial regression. *IEEE Transactions on Image Processing*, 24(5): 1460–1470, 2015.

[42] Alessandro Foi. Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing*, 89(12):2609–2629, 2009. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2009.04.035. URL `https://www.sciencedirect.com/science/article/pii/S0165168409001996`. Special Section: Visual Information Analysis for Security.

[43] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. doi: 10.1109/TIP.2008.2001399.

[44] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[45] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[46] O. Gallo and P. Sen. Stack-Based Algorithms for HDR Capture and Reconstruction. In *High Dynamic Range Video*, pages 85–119. Elsevier, 2016.

[47] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):118, 2017.

[48] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proc. of SIGGRAPH '96*, pages 43–54. ACM Press, 1996. ISBN 0897917464. doi: 10.1145/237170.237200. URL `http://portal.acm.org/citation.cfm?doid=237170.237200`.

[49] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal hdr reconstruction with linear digital cameras. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 215–222. IEEE, 2010.

[50] Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free hdr reconstruction. *ACM Transactions on Graphics (TOG)*, 32(6):201, 2013.

[51] Michael D. Grossberg and Shree K. Nayar. What can be known about the radiometric response from images? In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision — ECCV 2002*, pages 189–205, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[52] Michael D Grossberg and Shree K Nayar. What is the space of camera response functions? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, volume 2, pages II–602. IEEE, 2003.

[53] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualiso hdr reconstruction. *EURASIP Journal on Image and Video Processing*, 2015(1):41, 2015.

[54] Johannes Hanika. Profiling sensor and photon noise, Dec 2012. URL `https://www.darktable.org/2012/12/profiling-sensor-and-photon-noise`. https://www.darktable.org/2012/12/profiling-sensor-and-photon-noise.

[55] Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. Noise-aware merging of high dynamic range image stacks without camera calibration. In *Advances in Image Manipulation (ECCV workshop)*, pages 376–391. Springer, 2020. URL `https://doi.org/10.1007/978-3-030-67070-2_23`.

[56] Param Hanji, Rafał K. Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. Comparison of single image hdr reconstruction methods — the caveats of quality assessment. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings)*, 2022. doi: 10.1145/3528233.3530729. URL `https://www.cl.cam.ac.uk/research/rainbow/projects/sihdr_benchmark/`.

[57] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 35(6):192, 2016.

[58] S.W. Hasinoff, F. Durand, and W.T. Freeman. Noise-optimal capture for high dynamic range photography. In *CVPR*, pages 553–560. IEEE, 2010.

[59] G.E. Healey and R Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16 (3):267–276, mar 1994. ISSN 01628828.

[60] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014.

[61] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, Mahmoud Afifi, Michael S Brown, Kele Xu, Hengxing Cai, Yuzhong Liu, Li-Wen Wang, et al. Aim 2020: Scene relighting and illumination estimation challenge. *arXiv preprint arXiv:2009.12798*, 2020.

[62] Keigo Hirakawa and Thomas W Parks. Adaptive homogeneity-directed demosaicing algorithm. *IEEE Transactions on Image Processing*, 14(3):360–369, 2005.

[63] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`.

[64] Jay M. Ver Hoef. Who invented the delta method? *The American Statistician*, 66 (2):124–127, 2012. doi: 10.1080/00031305.2012.687494. URL `https://doi.org/10.1080/00031305.2012.687494`.

[65] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1170, 2013.

[66] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Trans. Graph.*, 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766922. URL `https://doi.org/10.1145/2766922`.

[67] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[68] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008.

[69] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In *Advances in neural information processing systems*, pages 769–776, 2009.

[70] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[71] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[72] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003.

[73] K Karaduzovic-Hadziabdic, J Hasic Telalovic, and R Mantiuk. Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging. In *Second International Conference and SME Workshop on HDR imaging*, pages 1–4. Citeseer, 2014.

[74] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[75] Naoki Katoh, Toshihide Ibaraki, and Hisashi Mine. An algorithm for finding k minimum spanning trees. *SIAM Journal on Computing*, 10(2):247–255, 1981.

[76] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. FHDR: HDR image reconstruction from a single ldr image using feedback network. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019.

[77] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3116–3125, 2019.

[78] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. JSI-GAN: GAN-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for UHD HDR video. In *AAAI*, pages 11287–11295, 2020.

[79] Davis E King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015.

[80] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[81] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[82] Kristian Kirk and Hans Jørgen Andersen. Noise characterization of weighting schemes for combination of multiple exposures. In *BMVC*, volume 3, pages 1129–1138. Citeseer, 2006.

[83] Mikhail Konnik and James Welsh. High-level numerical simulations of noise in CCD and CMOS photosensors: review and tutorial. *arXiv preprint arXiv:1412.4031*, 2014.

[84] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, Anders Ynnerman, and Jonas Unger. A unified framework for multi-sensor hdr video reconstruction. *Signal Processing: Image Communication*, 29(2):203–215, 2014.

[85] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

[86] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive HDRi: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018.

[87] Jinghui Li and Peiyu Fang. HDRNET: Single-image-based HDR reconstruction using channel attention CNN. In *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, pages 119–124, 2019.

[88] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In *Visual Communications and Image Processing 2008*, volume 6822, page 68221J. International Society for Optics and Photonics, 2008.

[89] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[90] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[91] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[92] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*, pages 1651–1660, 2020.

[93] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, December 2015.

[94] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3238. IEEE, jun 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00335. URL `https://ieeexplore.ieee.org/document/8953888/`.

[95] Wenmiao Lu and Yap-Peng Tan. Color filter array demosaicking: new method and performance measures. *IEEE transactions on image processing*, 12(10):1194–1210, 2003.

[96] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European conference on computer vision*, pages 715–732. Springer, 2020.

[97] Kede Ma, Zhengfang Duanmu, Hojatollah Yeganeh, and Zhou Wang. Multi-exposure image fusion by optimizing a structural similarity index. *IEEE Transactions on Computational Imaging*, 4(1):60–72, 2018. doi: 10.1109/TCI.2017.2786138.

[98] Kevin J MacKenzie, David M Hoffman, and Simon J Watt. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. *Journal of vision*, 10(8):22, jan 2010. ISSN 1534-7362. doi: 10.1167/10.8.22. URL `http://www.ncbi.nlm.nih.gov/pubmed/20884597`.

[99] Kevin J. MacKenzie, Ruth A. Dickson, and Simon J. Watt. Vergence and accommodation to multiple-image-plane stereoscopic displays: "real world" responses with practical image-plane separations? *Journal of Electronic Imaging*, 21(1):011002, feb 2012. ISSN 1017-9909. doi: 10.1117/1.JEI.21.1.011002. URL `http://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JEI.21.1.011002`.

[100] Markku Makitalo and Alessandro Foi. Optimal inversion of the generalized anscombe transformation for poisson-gaussian noise. *IEEE transactions on image processing*, 22(1):91–103, 2012.

[101] S Mann and R Picard. Being "undigital" with digital cameras: extending dynamic range by combining differently exposed pictures. In *Proc. IS&T 48th Annual Conference*, pages 442–428, 1995.

[102] Rafał Mantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, New York, NY, USA, 2008.

Association for Computing Machinery. ISBN 9781450301121. doi: 10.1145/1399504.1360667. URL https://doi.org/10.1145/1399504.1360667.

[103] Rafal K. Mantiuk. Practicalities of predicting quality of high dynamic range images and video. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 904–908. IEEE, sep 2016. ISBN 978-1-4673-9961-6. doi: 10.1109/ICIP.2016.7532488. URL http://ieeexplore.ieee.org/document/7532488/.

[104] Rafał K. Mantiuk and Maryam Azimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium*, 2021.

[105] Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Trans. Graph.*, 40(4), jul 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459831. URL https://doi.org/10.1145/3450626.3459831.

[106] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[107] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

[108] Sujitha Martin, Kevan Yuen, and Mohan M Trivedi. Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1010–1014. IEEE, 2016.

[109] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE Transactions on Image Processing*, 16(1):132–141, 2006.

[110] Laurence Meylan, Scott Daly, and Sabine Süsstrunk. The reproduction of specular highlights on high dynamic range displays. In *Color and Imaging Conference*, volume 2006, pages 333–338. Society for Imaging Science and Technology, 2006.

[111] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K. Mantiuk. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2559–2566, 2021. doi: 10.1109/ICPR48806.2021.9412676.

[112] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24.

[113] Scott Miller, Mahdi Nezamabadi, and Scott Daly. Perceptual signal coding for more efficient usage of bit codes. *SMPTE Motion Imaging Journal*, 122(4):52–59, 2013.

[114] Tomoo Mitsunaga and Shree K Nayar. Radiometric self calibration. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 374–380. IEEE, 1999.

[115] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

[116] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

[117] Ankit Mohan, Reynold Bailey, Jonathan Waite, Jack Tumblin, Cindy Grimm, and Bobby Bodenheimer. Tabletop computed lighting for practical digital photography. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):652–662, 2007.

[118] Rebecca Moody. Screen time statistics: Average screen time in us vs. the rest of the world, March 2022. URL `https://www.comparitech.com/tv-streaming/screen-time-statistics/`.

[119] Christopher Z. Mooney and Robert D. Duval. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage, 1993. ISBN 9780803953819.

[120] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4080–4089, 2019.

[121] Aamir Mustafa, Param Hanji, and Rafal Mantiuk. Distilling style from image pairs for global forward and inverse tone mapping. In *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450399395. doi: 10.1145/3565516.3565520. URL `https://doi.org/10.1145/3565516.3565520`.

[122] Rang M. H. Nguyen and Michael S. Brown. Why you should forget luminance conversion and do something better. In *CVPR*, pages 5920–5928. IEEE, jul 2017.

ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.627. URL http://ieeexplore.ieee.org/document/8100110/.

[123] Gary W. Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992. doi: 10.1080/00031305.1992.10475842. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475842.

[124] Ingram Olkin and John W. Pratt. Unbiased Estimation of Certain Correlation Coefficients. *The Annals of Mathematical Statistics*, 29(1):201 – 211, 1958. doi: 10.1214/aoms/1177706717. URL https://doi.org/10.1214/aoms/1177706717.

[125] Sumanta N Pattanaik, James A Ferwerda, Mark D Fairchild, and Donald P Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 287–298. Citeseer, 1998.

[126] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal K. Mantiuk. From Pairwise Comparisons and Rating to a Unified Quality Scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2020. ISSN 1057-7149. doi: 10.1109/TIP.2019.2936103. URL https://ieeexplore.ieee.org/document/8818316/.

[127] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 691–700, 2021.

[128] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, jan 2015. ISSN 09235965. doi: 10.1016/j.image.2014.10.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0923596514001490.

[129] Javier Portilla, Vasily Strela, Martin J Wainwright, and Eero P Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11), 2003.

[130] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, 2019. URL http://rpg.ifi.uzh.ch/docs/TPAMI19_Rebecq.pdf.

[131] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[132] Erik Reinhard and Kate Devlin. Dynamic range reduction inspired by photoreceptor physiology. *IEEE transactions on visualization and computer graphics*, 11(1):13–24, 2005.

[133] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, jul 2002. ISSN 0730-0301. doi: 10.1145/566654.566575. URL `https://doi.org/10.1145/566654.566575`.

[134] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3):39–es, 2007.

[135] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[136] Mark A Robertson, Sean Borman, and Robert L Stevenson. Estimation-theoretic approach to dynamic range enhancement using multiple exposures. *Journal of Electronic Imaging*, 12(2):219–229, 2003.

[137] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[138] Ehsan Samei, Michael J Flynn, and David A Reimann. A method for measuring the presampled mtf of digital radiographic systems using an edge test device. *Medical physics*, 25(1):102–113, 1998.

[139] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020.

[140] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[141] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.

[142] Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. High dynamic range display systems. *ACM Trans. Graph.*, 23(3):760–768, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015797. URL https://doi.org/10.1145/1015706.1015797.

[143] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012.

[144] Zhenfeng Shao, Wenjing Wu, Zhongyuan Wang, Wan Du, and Chengyuan Li. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Transactions on Multimedia*, 20(10):2593–2604, oct 2018. ISSN 1520-9210. doi: 10.1109/TMM.2018.2865686. URL https://ieeexplore.ieee.org/document/8438999/.

[145] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. doi: 10.1109/TIP.2005.859378.

[146] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323008. URL https://doi.org/10.1145/3306346.3323008.

[147] Nai-Sheng Syu, Yu-Sheng Chen, and Yung-Yu Chuang. Learning deep convolutional networks for demosaicing. *arXiv preprint arXiv:1802.03769*, 2018.

[148] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

[149] Daniel Stanley Tan, Wei-Yang Chen, and Kai-Lung Hua. Deepdemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. *IEEE Transactions on Image Processing*, 27(5):2408–2419, 2018.

[150] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[151] Tadaaki Tani. *Photographic sensitivity: theory and mechanisms*. Oxford University Press on Demand, 1995.

[152] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[153] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. In *ACM Transactions on Graphics (TOG)*, volume 30, page 41. ACM, 2011.

[154] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. In *The 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, pages 49–56. Václav Skala-UNION Agency, 2007.

[155] Yanghai Tsin, Visvanathan Ramesh, and Takeo Kanade. Statistical calibration of CCD imaging process. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 480–487. IEEE, 2001.

[156] A. Tsirikoglou, G. Eilertsen, and J. Unger. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum*, 39(6):426–451, sep 2020. ISSN 0167-7055. doi: 10.1111/cgf.14047. URL `https://onlinelibrary.wiley.com/doi/10.1111/cgf.14047`.

[157] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: a survey and evaluation. In *Computer Graphics Forum*, volume 34, pages 683–707. Wiley Online Library, 2015.

[158] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971. IEEE, jul 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.316. URL `http://ieeexplore.ieee.org/document/8099799/`.

[159] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, 2015. doi: 10.1109/NCC.2015.7084843.

[160] Tu V.V. High dynamic range image synthesis via attention non-local network. 2020. URL `github.com/tuvovan/ANL-HDRI`.

[161] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019.

[162] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

[163] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[164] Kevin Westcott, Jana Arbanas, Chris Arkenberg, Brooke Auxier, Jeff Loucks, and Kevin Downs. 2022 digital media trends, 16th edition: Toward the metaverse, March 2022. URL `https://www2.deloitte.com/us/en/insights/industry/technology/digital-media-trends-consumption-habits-survey/summary.html`.

[165] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018.

[166] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics*, 37 (4):1–13, 2018. ISSN 15577368. doi: 10.1145/3197517.3201313.

[167] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019.

[168] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. doi: 10.1109/TIP.2020.2971346.

[169] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating hdr transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2018.

[170] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, pages 4519–4528, 2017.

[171] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[172] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[173] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8): 2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

[174] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23 (10):4270–4281, 2014. doi: 10.1109/TIP.2014.2346028.

[175] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

[176] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E. Gonzalez, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, and Kurt Keutzer. A Review of Single-Source Deep Unsupervised Visual Domain Adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020. ISSN 2162-237X. doi: 10.1109/TNNLS.2020.3028503. URL `https://ieeexplore.ieee.org/document/9238468/`.

[177] Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, Param Hanji, Simon J. Watt, and Rafał K. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Trans. Graph.*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480513. URL `https://doi.org/10.1145/3478513.3480513`.

[178] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pages 479–486. IEEE, 2011.

Appendix A

Proof of MVUE

The following proof was provided by my collaborator, Fangcheng Zhong, who worked with me on this project. I am grateful for his assistance.

Here I use the Lehmann-Scheffe theorem to show that with no static noise, the Poisson Estimator $\hat{\phi}_{\text{ppne}}$ is the unique minimum variance unbiased estimator (MVUE). I will state the theorem and a few related definitions first. In this section, the per-pixel notation is dropped and unless explicitly stated otherwise, every summation and product is indexed by i going from 1 to N, which identifies each image in a stack of size N.

Definition A.1 (Sufficient statistic)**.** Let $\vec{U}$ denote a random vector sampled from a distribution that has joint pdf or pmf $f(\vec{u};\theta)$ parameterized by $\theta \in \Omega$. Let $S = g(\vec{U})$ be a statistic whose PDF or PMF is $f_S(s;\theta)$. S is a sufficient statistic for θ if $\frac{f(\vec{u};\theta)}{f_S(s;\theta)} = h(\vec{u})$ where $h(\vec{u})$ does not depend on θ.

Definition A.2 (Complete family)**.** Let the random variable S have a PDF or PMF that is a member of the family $\{f_S(s;\theta)|\theta \in \Omega\}$. Let $w(S)$ be a function of S. If the expectation $E[w(S)]$ being zero for every $\theta \in \Omega$ requires that w(S) be zero for every s in the support of $f_S(s;\theta)$, then the family $\{f_S(s;\theta)|\theta \in \Omega\}$ is called a complete family.

Theorem 1 (Lehmann and Scheffe)**.** *Let $\vec{U}$ denote a random vector sampled from a distribution that has joint PDF or PMF $f(\vec{u};\theta)$ parameterized by $\theta \in \Omega$. Let $S = g(\vec{U})$ be a sufficient statistic for θ, and let the family $\{f_S(s;\theta)|\theta \in \Omega\}$ be complete. If there is a function of S that is an unbiased estimator of θ, then this function of S is its unique MVUE.*

Proof. Let $x_i t_i$ be denoted as u_i, then $\hat{\phi}_{\text{ppne}} = \frac{\sum u_i}{\sum t_i}$. Let $\vec{u}$ denote $(u_1, u_2, ...u_N)$, a random vector with joint PMF $f(\vec{u};\phi)$. Let $s = \sum u_i$. We first show that s is a sufficient statistic for ϕ.

According to Equation Eq. (2.1), $u_i \sim \text{Pois}(\phi\, t_i)$ if we assume no static noise. Since

each u_i is independent,

$$f(\vec{u};\phi) = \prod \frac{(\phi t_i)^{u_i} e^{-\phi t_i}}{u_i!} = e^{-\sum \phi t_i} \phi^{\sum u_i} \prod (t_i)^{u_i} \prod \frac{1}{u_i!}$$

$s \sim \text{Pois}(\sum \phi t_i)$ is a compound Poisson distribution with PMF

$$f_S(s;\phi) = \frac{(\sum \phi t_i)^s e^{-\sum \phi t_i}}{s!} = \frac{\phi^s (\sum t_i)^s e^{-\sum \phi t_i}}{s!}$$

Then,

$$\frac{f(\vec{u};\phi)}{f_S(s;\phi)} = \frac{(\sum t_i)^s \prod (t_i)^{u_i}}{s! \prod u_i!},$$

which does not depend on ϕ. Therefore, $s = \sum u_i$ is a sufficient statistic for ϕ by definition. Next, we show that the family $\{f_S(s;\phi)|\phi > 0\}$ is complete.

Suppose a function $w(s)$ of s is such that $E[w(s)] = 0$ for every $\phi > 0$. We want to show that this requires $w(s) = 0$ for every $s >= 0$. First, we know that for all $\phi > 0$,

$$0 = E[w(s)] = \sum_{s=0}^{\infty} w(s) f_S(s;\phi) = \sum_{s=0}^{\infty} w(s) \frac{(\sum \phi t_i)^s e^{-\sum \phi t_i}}{s!}$$

$$= e^{-\sum \phi t_i} \sum_{s=0}^{\infty} w(s) \frac{\phi^s (\sum t_i)^s}{s!}$$

Since $e^{-\sum \phi t_i}$ is nonzero, $\sum_{s=0}^{\infty} w(s) \frac{(\sum t_i)^s}{s!} \phi^s$ is a power series of ϕ which converges to zero for all $\phi > 0$. Therefore its coefficients must be zero. Since $\frac{(\sum t_i)^s}{s!}$ is also non-zero, $w(s)$ must be zero for every $s >= 0$, which completes the proof of $\{f_S(s;\phi)|\phi > 0\}$ being a complete family.

At last, note that $\hat{\phi}_{\text{ppne}}$ is an unbiased estimator of ϕ (this is also true for nonzero static noise):

$$E[\hat{\phi}_{\text{ppne}}] = \frac{\sum E[u_i]}{\sum t_i} = \frac{\sum \phi t_i}{\sum t_i} = \phi$$

And since $\hat{\phi}_{\text{ppne}}$ is a function of a sufficient statistic for ϕ, $\hat{\phi}_{\text{ppne}}$ is the unique MVUE of ϕ by the *Lehmann-Scheffe* theorem.

□

Appendix B

Moments of a transformed random variable

I rely on the Taylor expansion to derive expressions for the first and second moments (expected value and variance respectively) of a random variable under an arbitrary invertible function. In Section 5.4.2 I apply these results to obtain expressions for the log of a normally distributed random variable.

To get an approximate variance for a nonlinear function of a random variable, the Delta method [64, 123] uses a Taylor expansion. Let Z be an asymptotically normal random variable with a known mean μ_Z and variance σ_Z^2, and let f be an invertible and differentiable function. To get expressions for the expected value and variance of $f(Z)$ we use its first order Taylor expansion about the mean μ_Z,

$$\begin{aligned} f(Z) = f(\mu_Z) + f'(\mu_Z)(Z - \mu_Z) \\ + \text{higher-order terms}\,. \end{aligned} \tag{B.1}$$

Ignoring the diminishing contributions of higher-order terms, the expected value is,

$$\begin{aligned} \mathbb{E}[f(Z)] &\approx \mathbb{E}[f(\mu_Z)] + \mathbb{E}[f'(\mu_Z)(Z - \mu_Z)] \\ &= \mathbb{E}[f(\mu_Z)] + f'(\mu_Z)(\mathbb{E}[Z] - \mathbb{E}[\mu_Z])] \\ &= f(\mu_Z) + f'(\mu_Z)(\mu_Z - \mu_Z) \\ &= f(\mu_Z)\,, \end{aligned} \tag{B.2}$$

and the variance is

$$\begin{aligned}
\mathbb{V}[f(Z)] &\approx \mathbb{V}[f(\mu_Z)] + \mathbb{V}[f'(\mu_Z)(Z - \mu_Z)] \\
&= \mathbb{V}[f(\mu_Z)] + f'(\mu_Z)^2 \, \mathbb{V}[Z - \mu_Z] \\
&= \mathbb{V}[f(\mu_Z)] + f'(\mu_Z)^2 \, (\mathbb{V}[Z] + \mathbb{V}[\mu_Z]) \\
&= 0 + f'(\mu_Z)^2 \, (\sigma_Z^2 + 0) \\
&= f'(\mu_Z)^2 \, \sigma_Z^2 \, .
\end{aligned} \tag{B.3}$$