Deep Snapshot HDR Reconstruction Based on the Polarization Camera

by

Jui Wen Ting

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

© Jui Wen Ting, 2021

Abstract

In this thesis, we propose the use of the polarization camera for high-dynamic-range (HDR) imaging. Specifically, observing that natural light can be attenuated differently by varying the orientation of the polarization filter, we treat the multiple images captured by the polarization camera as a set captured under different exposure times, to support the development of solutions for the HDR reconstruction problem. Most existing methods are developed for conventional camera images. However, polarization cameras capture images differently than conventional cameras. In this thesis, we propose two deep snapshot HDR reconstruction frameworks, that uses polarimetric cues available from the polarization camera. With our deep-learning based methods, the obtained polarimetric information enables us to regress the missing pixels in polarization images more effectively. We train and validate the methods on our collected polarization dataset. We demonstrate through experimental results that our approach can reconstruct visually pleasing HDR results, and performs favorably than state-of-the-art HDR reconstruction algorithms. The source code is publicly available on Github[1].

[1] Link to source code: https://github.com/jtuoa/Deep-Polarized-HDRreconstruction

Preface

The following papers are part of this manuscript:

1. Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, Juiwen Ting "HDR Reconstruction Based on the Polarization Camera", IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 7203-7216, 2020.

2. Juiwen Ting, Xuesong Wu, Kangkang Hu, Hong Zhang, "Deep Snapshot HDR Reconstruction Based on the Polarization Camera", under submission to IEEE International Conference on Image Processing (ICIP), Alaska, USA, 2021.

3. Juiwen Ting, Moein Shakeri, Hong Zhang, "Deep Polarimetric HDR Reconstruction", under submission to ACM Transactions on Graphics (ACM TOG), 2021.

In the list mentioned above, paper 1 is described in Chapter 3.2, paper 2 is described in Chapter 3.3, and paper 3 is described in Chapter 3.4.

In the paper 1 submission I helped to build the datasets used in the experiments. In the paper 2 and 3 submissions I was the primary contributor, and discussions with the co-authors helped in completing the experiments and writing.

Dr. Hong Zhang provided critical inputs and was the primary instructor and co-author for all the publications.

To my family

For their constant support and motivation.

I could either watch it happen or be a part of it.

– Elon Reeve Musk, 2019.

Acknowledgements

Foremost, I would like to thank my supervisor, Dr. Hong Zhang. His valuable guidance and encouragement throughout the period of my program made this work possible. His valuable suggestions and time spent in discussions has added greatly to my knowledge.

I would like to thank the examination committee Dr. Nilanjan Ray and Dr. Li Cheng for reviewing my thesis and providing insightful feedback. I would like to thank Dr. Moein Shakeri for the valuable guidance and feedback in the collaboration on the topic of “Deep Polarimetric HDR Reconstruction”. I would also like to thank Sara Elkerdawy for improving my work through in-depth discussions.

Last but not least, I would like to thank my family for entrusting me and providing me with their unfailing support and continuous encouragement all the time. It has been a long journey up until finishing the thesis, and this is just a beginning for more adventure ahead.

Thank you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter first introduces the problem by defining the concept of high dynamic range, followed by the motivation for the research. Then background information to lay the necessary foundation for the rest of the thesis is presented, where we describe the HDR reconstruction pipeline and HDR image applications. Next, the thesis contributions are described, and finally the structure of the thesis is outlined.

1.1 Problem definition

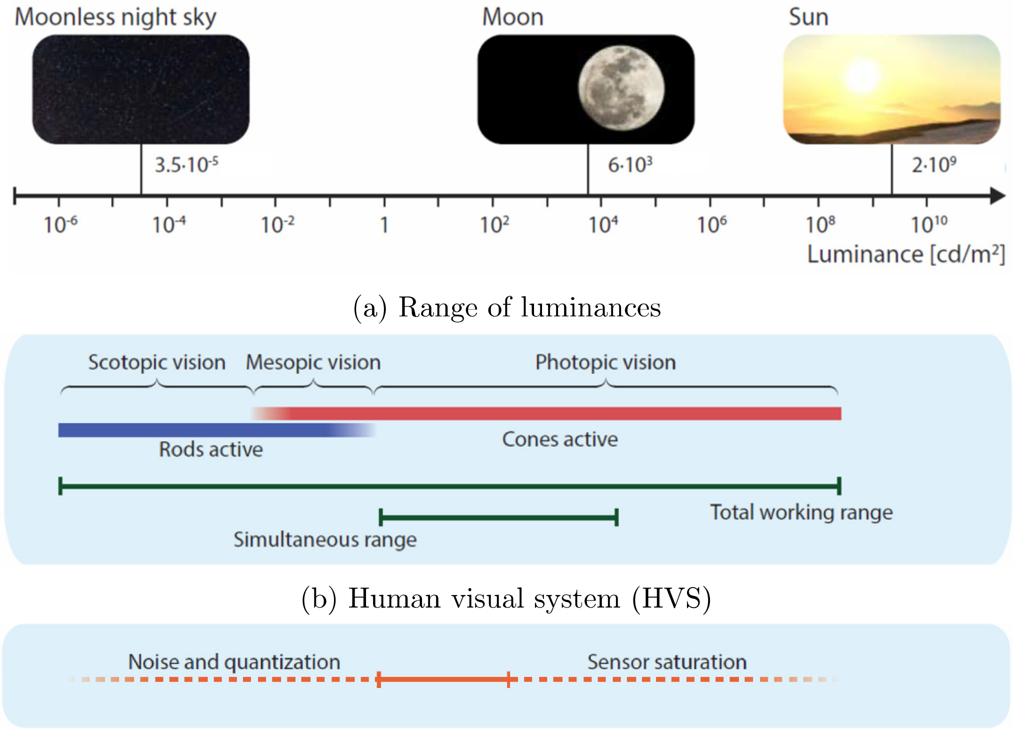
Dynamic range is the ratio between the brightest and darkest value registered by an imaging device [14]. In photography and imaging, dynamic range represents the ratio of two luminance values. Luminance is the integrated outgoing light over a surface area in a certain direction, and it is what we measure when registering the light as it falls on the area of a pixel in a camera sensor [23]. This is distinct from illuminance, which is the incident light from the surrounding environment onto a specific point on a surface. The SI unit for measuring the luminance in a scene or on a display is candela per square meter (cd/m^2). In the display manufacturing industry, the same unit is also commonly referred to as nit (1 nit = 1 cd/m^2). In real-world scenes, the luminance varies over several orders of magnitude. For instance, Figure 1.1a shows the luminance of a moonless night sky can be 3.5×10^{-5} cd/m^2, while a sunlit sky can be 2×10^9 cd/m^2. An imaging device that can simultaneously render such a range of luminances has a dynamic range of 5.714×10^{13}:1 or 45.70 bits or

45.70 stops/f-stops. The stops/f-stops unit is commonly seen in photography, and can be computed by $log_2(dynamic\,range)/log_2$.

From the literature in high dynamic range (HDR) imaging, it is not exactly clear what the definition of low and high dynamic range is, and it may vary depending on the application. The term low dynamic range (LDR) generally refers to conventional cameras and displays with a dynamic range of 256:1 or 8 bits or 8 stops/f-stops. Furthermore, some literature use the terms LDR and standard dynamic range (SDR) interchangeably to denote images that are not HDR. The term HDR generally refers to anything that has a wider dynamic range than the conventional cameras and displays.

Figure 1.1b depicts the range of luminances the human visual system (HVS) is capable of capturing. The HVS can observe a very wide range of luminances from around 10×10^{-6} cd/m^2 up to 10×10^{8} cd/m^2, for a total dynamic range of 1×10^{14}:1 or 46.51 bits or 46.51 stops/f-stops. However, in order to do so the eye needs to adapt to the different lighting conditions. This is achieved partly by changing the pupil size, but mostly from bleaching and regeneration processes in the photo-receptors [52]. This transition for our eyes to adapt from a bright to a dark environment can take considerable time before details can be discerned. The simultaneous dynamic range of the eye, which is also depicted in Figure 1.1b, is around 10×10^{-1} cd/m^2 up to 10×10^{4} cd/m^2, for a total dynamic range of 1×10^{5}:1 or 16.61 bits or 16.61 stops/f-stops.

A camera is designed for a similar task as the HVS, which is to capture the surrounding environment in order to attain information for subsequent processing. Given this similarity, we would expect that a physical scene captured by a camera and viewed on a display should invoke the same response as observing the scene directly. However, this is very seldom the case. Though the dynamic range of a camera sensor can vary greatly, from 8 bits LDR cameras up to 14 bits high-end HDR cameras, most consumer grade cameras are only 8 bits. Figure 1.1c illustrates the dynamic range for a conventional consumer level camera sensor. Luminances above the highest detectable value are lost from a hard cutoff at some peak intensity, since the pixels in the sensors that act as potential wells are saturated when the well capacity is reached. In-



(a) Range of luminances

(b) Human visual system (HVS)

(c) Typical camera sensor

Figure 1.1: Dynamic ranges of different capturing techniques.

formation below the lowest detectable value is lost due to sensor noise and quantization. This dynamic range is much narrower than that of the HVS. Therefore, this mismatch in dynamic range results in output images with over- or under-saturated areas where details in the bright and dark regions can not be captured, but can be detected by the HVS.

1.2 Motivation

Since the production of Charge-Coupled Devices (CCDs), the image restoration problem has been extensively studied for commercial applications, for decades. The CCD converts optical perception to digital signals, but due to the semiconductors used in the CCDs, there is an unknown non-linearity mapping between the scene radiance (luminance) and the pixel values in the image. This non-linearity is usually modelled by gamma correction, which has resulted in a series of image restoration methods [5], [21], [69]. However, these methods tend to focus on image pixel balance, and do not restore image details lost from the under- or over-exposure due to the limitations of conventional camera sensors.

The limitations of the camera as compared to the HVS is evident. Conventional camera sensors are unable to capture the wide range of luminances that the HVS can detect simultaneously, which means that there is more visual information available in the scene than what can be captured and processed. This problem falls within the class of image restoration problems, and is known as the HDR reconstruction problem. HDR reconstruction is the task to recover the broad range of luminances that the HVS can detect. Although different image pixel operators have been proposed for HDR reconstruction, the results are still unsatisfactory, due to the ill-posed nature of the problem.

Some studies have tackled the HDR reconstruction problem using hardware and software approaches. For example, different camera modalities are used in combination with image restoration algorithms to realize HDR images. The motivation is that the camera sensor can help to first narrow the mismatch between the dynamic range in natural scenes and the capturing device, then the image restoration algorithm can more effectively infer the HDR image since it now has more input information. Several camera sensors have been proposed to realize such designs, such as a spatially varying exposure (SVE) sensor [61], a color coded filter [22], a non-regular coded exposure mask [78], an optically coded mask [73], a dual-ISO sensor [20], and a multi-exposure color filter array (ME-CFA) sensor [81].

An interesting recent development in imaging technology is a polarization image sensor (Polarsens by Sony) with four directional, on-chip micropolarizers [26]. The pixels in a polarization sensor are organized into unit of four sensing elements, each arranged as a 2x2 matrix with polarizing filters at 0°, 45°, 90° and 135°, to capture four spatially and temporally synchronized images of a scene with the same ease of operation as a conventional camera. Since polarization conveys information about the surface normal, a common application of this image sensor is 3D reconstruction [3], [10]. As a less obvious application, since a polarizing filter attenuates irradiance, and the extent of attenuation varies with the direction of the polarizer, in a way similar to changing the exposure time setting, this sensor should also provide us with the possibility to reconstruct HDR images. We note irradiance refers to a portion of the luminance that actually falls on the area of a pixel in a camera sensor. In fact, [76] has made such an attempt based on the idea of light attenuation by a polarizer for the purpose of eliminating saturated points and enhancing contrast to achieve HDR; however, [76], did not examine the full HDR image reconstruction.

Inspired by the approach to combine hardware and software components to reconstruct better HDR images, we raise two inter-related questions:

1. Is it possible to perform HDR image reconstruction using the polarization camera?

2. If question 1 is possible, then given the polarization information, how do we effectively recover the missing details using a deep-learning based approach?

This thesis aims to answer the above two inter-related questions, using the polarization camera - Polarsens developed by Sony [26]. This will be accomplished with the development of an HDR polarization mathematical model, a deep HDR reconstruction framework and an image acquisition setup.

1.3 Background

1.3.1 Image restoration

Image restoration methods aim to restore an original image using prior knowledge about the degradation phenomena. Quantitatively, the image restoration task can be expressed as:

$$y = H(x) + d \tag{1.1}$$

where x is the corrupted image, H is the degradation function, d is the additive noise, and y is the restored latent clean image. A variety of state-of-the-art image restoration methods have been proposed. Some methods propose to obtain y by reducing the noise via deep network designs [35], [97], low-rank sparse representation learning [41] or soft-rounding regularization [54]. Other methods propose to strengthen details by edge aware filtering [5], [21] or histogram equalization [69], in order to obtain y. There are also methods that focus on pixel manipulation, such as color enhancement [92] to obtain a y that adapts to various user preferences. Although these methods can help to improve the image quality, they can hardly recover the missing details. The HDR reconstruction problem is a subclass of the image restoration problem, and aims to recover the broad range of luminances that the HVS can detect. Thus, visual details of the scene initially missing due to sensor under- or over-saturation can be restored.

Another subclass of the image restoration problem is image inpainting. Image inpainting requires generating plausible pixels for corrupted content according to uncorrupted contents. Several methods have demonstrated content generation [68], [96]. However, it is inadequate to apply inpainting methods to restore details in saturated regions. There are two distinct differences between the image inpainting and HDR reconstruction tasks. First, the missing regions of the images in inpainting are due to random masks, while the saturated regions are correlated. Second, to generate reasonable results, the existing contents in the saturated image are required to be adjusted for the full HDR reconstruction task.

Another subclass of the image restoration problem is the illumination invariant task. Illumination invariant aims to remove image variations caused by illumination changes. This is mainly achieved by reprojecting the colour space to obtain an illumination invariant representation of the image [4], [80]. Thus, unlike HDR reconstruction, the resulting images do not restore the natural color nor expand the dynamic range to restore the luminances of the scene.

1.3.2 HDR reconstruction pipeline

The HDR imaging pipeline, from capturing to display, can be illustrated by Figure 1.2, which highlights the four major components: capture, reconstruction, distribution and tone-mapping. The pipeline can be described as follows:

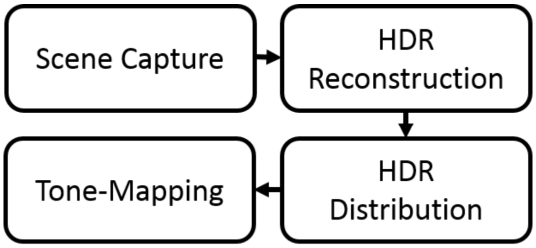


Figure 1.2: The HDR imaging pipeline

First, the physical scene is captured by a camera, followed by processing the captured information using techniques for HDR reconstruction. Next, the HDR image is stored using an HDR capable format. Finally, the HDR image is prepared for display, using a tone-mapping algorithm, which compresses the dynamic range of the HDR image in order to adapt to the range of the display while retaining visual image details.

The *first stage* of the HDR reconstruction pipeline is image capture. Various cameras can be used to capture the physical scene, and can be generally categorized as: HDR or LDR cameras. HDR cameras enable one-shot HDR capturing. The one-shot approach allows high speed capturing of the target, and avoids ghosting s that can result from scene dynamics or camera shakes during capture. Several HDR cameras can infer an HDR image directly from a single exposure value (EV) image, where EV represents the amount of light that reaches the camera sensor. For example, Figure 1.3a shows the Alexa model released by the camera manufacturer ARRI, which reported a dynamic range of 14 stops. As another example, Figure 1.3b shows the CineAlta Venice model released by Sony, which achieves a dynamic range of 15 stops. In another example, Figure 1.3c shows the Epic Dragon model released by the camera manufacturer RED, which further enhanced the dynamic range and reported a dynamic range of 16.5 stops. The increase in dynamic range by these cameras can be partly attributed to the hardware developments via large size and manufacturing quality of the sensor, which forms the reduction in the noise floor of the captured image. However, these high-end cameras have a large form factor and high cost. Thus, in general, unsuitable for consumer usage.

Instead, a compact and more affordable option is to capture images with

(a) ARRI Alexa (b) Sony CineAlta Venice (c) Red Epic Dragon

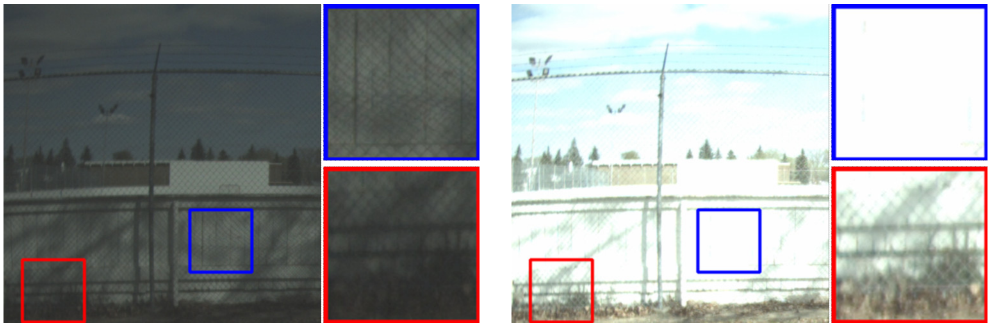
Figure 1.3: HDR cameras

conventional one-shot LDR cameras, and then apply image processing to achieve HDR reconstruction. The image processing operations can be performed via interactively through a post photo editing software, or through a post or onboard algorithm. However, photo editing can be an intensive manual process that requires skilled technicians and color artists to restore the missing details. Furthermore, prior knowledge about the contents in the scene can be required during photo editing, especially for large saturated regions, which may not always be available. To alleviate these problems, algorithms can be applied to automatically hallucinate the missing details in the saturated regions. However, when the input image for the algorithm is an one-shot single EV image, the algorithms report hallucination s for large saturated areas, due to the lack of information in the saturated areas [81].

In a LDR camera, the EV is controlled by the aperture, shutter speed, and sensor sensitivity, and can be expressed as:

$$EV = 2\log_2 F - \log_2 S + \log_2\left(\frac{ISO}{100}\right) \tag{1.2}$$

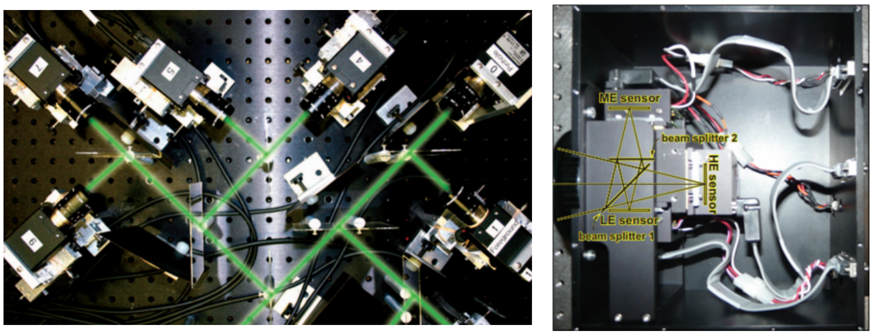
where F is the relative aperture (F-number), S is the exposure time (=1/shutter speed) and ISO is the sensor sensitivity. In a LDR camera, an EV image depicts a certain range of the scene luminance. Figure 1.4a shows that low EV images can capture details in bright image areas but information in dark image areas are lost due to sensor under-saturation. Figure 1.4b shows that large EV images can capture details in dark image areas but information in bright image areas are lost due to sensor over-saturation. Thus, combining multiple EV images can extract both bright and dark image features that represent the



(a) EV: 45 us.

(b) EV: 1153 us.

Figure 1.4: EV images



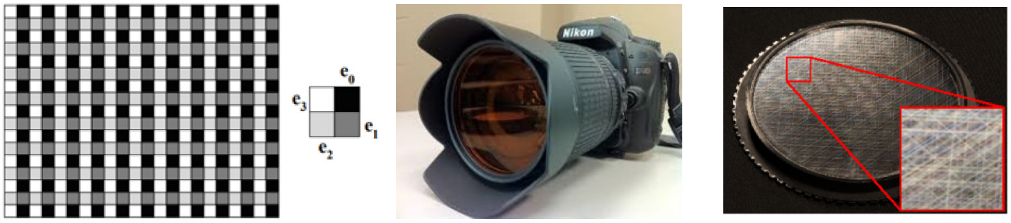
(a) [53]

(b) [83]

Figure 1.5: Multiple LDR camera setup for one-shot multi-EV image capturing

luminances of the physical world. Furthermore, the multiple EV images can be used to provide a more informative input that reduces the hallucination artifacts for large saturated areas encountered with a single EV input. Therefore, as another one-shot approach, multiple LDR cameras can be positioned, to form an acquisition rig, to acquire multiple EV images in one acquisition. Figure 1.5 illustrates two proposed beam splitter configurations [53], [83], where the incoming light is partitioned onto multiple sensors to achieve simultaneous capture of multiple EV images.

However, such sensor setup has high complexity, cost, footprint and requires rigorous calibration. To alleviate the aforementioned issues, there are LDR cameras based on a snapshot sensor, which acquires multiple EV images in one acquisition with the ease of using only one camera. Several snapshot sensors for LDR cameras have been proposed to realize HDR images, such as a SVE sensor [61], a color coded filter [22], a non-regular coded exposure mask



(a) [61] (b) [22] (c) [73]

Figure 1.6: LDR snapshot cameras

[78], an optically coded mask [73], a dual-ISO sensor [20], and a ME-CFA sensor [81]. Figure 1.6 illustrates some of these snapshot sensors.

Alternatively, LDR cameras can also perform multiple-shots to capture multiple EV images. Many modern cameras are equipped with specific multi-exposure capturing modes. For example, a burst of images with short EV duration can be captured. As another example, a sequence of images with both shorter and longer EVs can also be captured. While the multiple-shots approach is initially developed to handle only static scenes, since motion between input images due to scene dynamics or camera shaking during capture can present ghosting artifacts, the development of state-of-the-art techniques in image registration, deghosting and machine learning have shown to alleviate this image misalignment issue and achieve good results. However, the additional capturing and processing time can be prohibitive for applications that demand real-time feeds, such as robotics and autonomous driving.

In our study, we use the polarization camera as our capturing device. This is a LDR camera with a snapshot sensor. The sensor has four directional, on-chip micropolarizers to allow capturing of four spatially aligned and temporarily synchronized images of a scene. More details about the polarization camera are described in Section 3.

The *second stage* of the HDR reconstruction pipeline is image reconstruction. Different algorithms have been proposed to reconstruct HDR images, but it generally consists of two distinct steps: First estimate the camera response function (CRF). Second, reconstruct the HDR image. This can be expressed as:

$$H = R(F(L)) \tag{1.3}$$

where L is the LDR image, F is the CRF estimation, R is the HDR reconstruction, and H is the degradation function or the reconstructed HDR image. The CRF describes a mapping between pixel values the camera acquires and luminances of the scene. Since most cameras have a nonlinear CRF, it results in a pixel variation across the image that is nonlinear in luminance. To correct for this variation, the response needs to be estimated and inverted in order to derive pixel values that are linearly dependent on the captured luminances. Several methods have been proposed for CRF estimation. For example, [95] proposed to obtain the CRF by capturing an image with an uniformly lit calibration chart, such as a Macbeth chart, placed in the scene. Since the chart has patches with known reflectances, the known radiances of the patches and the corresponding measurements can be interpolated to estimate the CRF. However, this estimation method can be inconvenient or not feasible in the field. For example, when images are taken with a camera attached to a remote mobile device. Furthermore, changes in temperature alter the response function requiring frequent recalibration. Other chartless methods have also been proposed. For example, [12], [47], [58] proposed to obtain the CRF using the radiometric response function estimated using images of arbitrary scenes taken under different known exposures. While the measured brightness values change with exposure, scene radiance values remain constant. This observation allows the estimation of the CRF without prior knowledge of scene radiance.

The HDR reconstruction is the task to recover the broad range of luminances that the HVS can detect. Various algorithms can be used to reconstruct HDR images, and can be generally categorized as: one-shot image or multiple-shot image HDR reconstruction. One-shot image HDR reconstruction requires only a single acquisition to reconstruct an HDR image. Traditional techniques are non-learning based, and perform brightness enhancement through expansion operators or light-source detection. For example, [1], [51] proposed global expansion operators to expand the content equally across all pixels. As another example, [66], [72] proposed local expansion operators to expand the content locally by considering neighboring pixels. Multiple-shot image HDR reconstruction requires multiple acquisitions to reconstruct an HDR image. Traditional techniques are non-learning based, and perform image fusion through pixel selection or pixel weighting to fuse the bracketed images, to generate HDR content. For example, [46], [91] proposed pixel selection criterions, to select a single exposure per pixel, to fuse the bracketed images. As another example, [12], [99] proposed pixel weighting functions that assign different weights to pixels at each EV, to fuse the bracketed images. However, these non-learning based approaches may not accurately estimate the luminances of the physical scene due to the lack of knowledge about real HDR images; thus limits the quality of the reconstructed HDR images.

In our study, we use the Mitusnaga and Nayar [58] method to calculate the CRF. To overcome the limitations of non-learning based approaches, we investigate a deep-learning based approach for HDR reconstruction. More details are described in Section 3.

The *third stage* of the HDR reconstruction pipeline is image distribution. The reconstructed HDR image contains rich visual information that needs a higher bit-depth format to store and distribute the wide dynamic range data. A natural goal for an HDR image format is to store the linear pixel values with floating point precision. However, assuming 32 bit floating numbers, this means that 96 bits per pixel have to be used in order to encode colors. For a 10 megapixel image, this amounts to a file size of 120 MB with no compression applied, which in many situations is impractical. Therefore, floating point HDR image formats use reduced pixel descriptions. Several HDR distributions formats have been proposed to store HDR information. The two most widely used formats are Radiance RGBE and OpenEXR. The HDR pixel format used by the Radiance renderer uses the RGBE pixel description introduced by Ward [85]. It stores RGB values with 32 bits per pixel; 8 bits mantissa for each color channel, and an 8 bit common exponent. Thus, by using a shared exponent, the RGBE format is able to provide a better compress representation of the floating point numbers. The OpenEXR (Extended Range) HDR pixel format

[30] typically stores the pixels with half floats, which uses 16 bits for each color channel. The bits are allocated for 1 sign bit, 5 exponent bits, and 10 mantissa bits. There are also options for 32 bit float and 32 bit integers. The pixels can be encoded both by lossy and lossless compression schemes, and thus provides a better compress representation of the floating point numbers.

In our study, we use the OpenEXR method to store the HDR information. More details are described in Section 4.

The *fourth stage* of the HDR reconstruction pipeline is tone-mapping. The HDR image contains rich visual information that is stored with a higher bit-depth by an HDR encoding format. However, most displays, such as monitor, TV, smartphone and printed paper can only display LDR content. Tone-mapping aims to reduce the dynamic range of HDR images in order to map scene-referred HDR tones to display-referred LDR pixels, while preserving the perceptual content as much as possible. Different tone-mapping operators (TMOs) have been proposed, and can be generally categorized into global and local operators. For example, [84] proposed to apply a single transformation globally for all pixels. As another example, [15] proposed to apply local transformations for different parts of the image. Typically, local TMOs have two distinct steps: First, decompose the image into a base layer that is often smoothed but still maintains the original global dynamic range, and a detail layer that contains only local edges or detail information. Second, apply a tone mapping curve to compress the base layer, and add the result with the detail layer to output the final tone-mapped image. This is analogous to separating the image into a product of illumination and reflectance components [6], which is similar to how the HVS processes a scene. The illumination component describes global variations within the scene, and the reflectance component describes the image details and textures. Furthermore, compared to the global TMOs, the local TMOs have demonstrated better performance in preserving local contrast and detail information.

In our study, we use the Durand TMO [15] and TMOs in Photomatix to tone-map the HDR images for visualization on LDR displays. More details are described in Section 4.

1.3.3 Applications

In the field of computer vision, accurate image details is a necessary prerequisite for solving many vision related tasks in autonomous driving and robotics. For example, in robotics, visual odometry is the process of estimating the motion of a camera in real-time using sequential images, and is often used for mapping and navigation. In general, this is accomplished by accurate detection of image features to generate correspondences between image pairs for feature matching and tracking. Several common feature detectors include SIFT [44], SURF [7] and ORB [74] rely on detecting features, such as lines, edges and corners. However, due to the wide range of luminances in a physical scene, which can be beyond the acquisition range of the camera, it results in under- or over-saturated pixels that outputs featureless pixels. Thus, washes out the lines, edges and corners required as input for feature detection algorithms to detect accurately. As another example, in autonomous driving, accurate object detection and object segmentation allows for safe autonomous navigation. These tasks also rely on accurate feature representation to detect and segment objects in the physical scene. However, it can also suffer from pixel saturation that results in featureless content. To alleviate the aforementioned issues, HDR reconstruction can be a beneficial image restoration pre-processing step. HDR images seek to recover the broad range of luminances that the HVS can detect, to output a feature-rich image that matches the contents in the physical scene.

1.4 Thesis contribution

To tackle the hallucination artifacts observed in one-shot single EV image HDR reconstruction, we bring in a LDR camera with a snapshot sensor - polarization camera. We identify that it is possible to perform HDR image reconstruction using the polarization camera, and provide the mathematical equations to model the HDR polarization image formation. Furthermore, we demonstrate that this has the added benefit of combining with software approaches to first narrow the mismatch between the dynamic range in natural scenes and the capturing device, then HDR image can be reconstructed more accurately as it now has a more informative input image.

Next, we tackle the problem of using the polarization images for HDR reconstruction based on a deep-learning approach. We posit that the polarization images along with the polarization information deduced from the images can help guide the network to achieve a better reconstructed HDR image. The proposed training targets to the networks are polarized images of a scene captured at different EV to create a traditional exposure fusion HDR image using the technique in [56]. We intend to train our network to faithfully reconstruct HDR images from LDR acquisitions by integrating polarization information into our model.

Finally, as we are interested in studying a deep-learning based approach for HDR reconstruction, a dataset for training and testing the network is required. However, there is a lack of such public available dataset for training and testing HDR techniques with polarized images. We address this issue by collecting a dataset. The polarization image dataset will be a collection of high resolution polarization data, with four polarization components, degree of polarization (DoP), angle of polarization (AoP), and color, for each scene. The dataset will contain a myriad of scenes with indoor and outdoor illuminations, stationary targets, and diffuse/specular/hybrid objects. The collection of large amounts of data is a necessary prerequisite for a data-driven machine learning problem.

This thesis makes three primary contributions:

1. To our knowledge, there is no prior work on reconstructing HDR from polarized images with deep-learning. We introduce the theory and perform extensive experiments supporting the case of using a polarization camera for HDR reconstruction.

2. This thesis also presents novel deep-learning frameworks for deep snapshot HDR reconstruction based on the polarization camera. We leverage on the prior knowledge about the polarization data and integrate it into the design of our framework to achieve a more robust model.

3. To support the deep-learning based approach, our work introduces a new

dataset, available publicly, for training and testing HDR techniques with polarized images. The dataset is created to support the development of data-driven approaches for HDR imaging. We demonstrate promising results on this new dataset.

1.5 Organization of the Thesis

- **Chapter 1.** *Introduction*

  We introduce the problem, the motivation behind it, the background information, and the potential gaps in current approaches to solving it that can be addressed.

- **Chapter 2.** *Related Works*

  We provide an overview of the different families of approaches that were used to tackle the HDR reconstruction problem.

- **Chapter 3.** *Proposed Deep Snapshot HDR Reconstruction*

  Here we describe the individual ideas introduced in our thesis and how they all come together for our proposed algorithm.

- **Chapter 4.** *Experiments*

  We discuss our experiments and how we use them to evaluate our methodology against standard approaches.

- **Chapter 5.** *Conclusion*

  We present the conclusions of this thesis and summarize the ideas newly introduced here.

- **Chapter 6.** *Appendix*

  This section lists the hyperparameters for the different algorithms that were used.

Chapter 2

Related Works

Given that HDR reconstruction is a fundamental task in image processing pipelines, whether it be for object detection, object segmentation or visual odometry, it is an age old problem which has been approached from many angles. Traditional techniques are non-learning based that rely on heuristic strategies to expand the dynamic range. In recent years, deep-learning based approaches have been proposed, and demonstrated improved performances. Based on the approaches used to solve the HDR reconstruction problem, they can be broadly divided into two categories. In this chapter, we discuss these two categories which are:

1. One-shot image HDR reconstruction

2. Multiple-shot image HDR reconstruction

Furthermore, within each category, we introduce the traditional works followed by the learning based works, to provide a comprehensive overview of the different techniques within the category.

2.1 One-shot image HDR reconstruction

This approach uses a single image acquisition to reconstruct an HDR image. Several methods based on this approach have been proposed, and can be grouped into single image output and multiple image output methods.

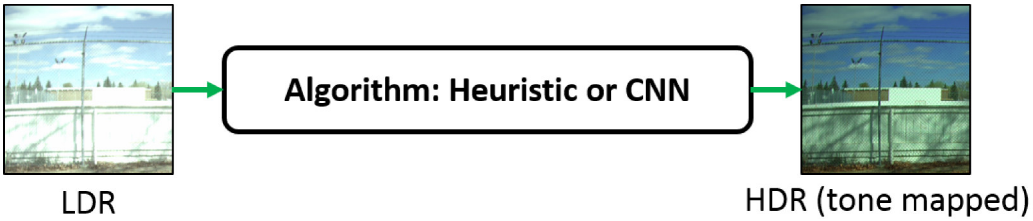


Figure 2.1: The one-shot single image output method HDR image formation pipeline

2.1.1 Single image output methods

Given a single input LDR image, the single image output methods design a framework to output a single HDR image. Figure 2.1 illustrates the pipeline of single image output methods, which aims to recover an HDR image (32 bits/pixel) from a given single LDR image (8 bits/pixel).

Several traditional works are based on heuristic approaches. Particularly, these traditional works explore different expansion operators heuristically to transform a single LDR image to an HDR image. Some works proposed global expansion operators to expand the content equally across all pixels. For example, in [1], a linear global expansion operator is proposed based on their psychophysical study on human visual perception. In [51], a gamma curve global expansion operator is proposed, in which the gamma value is determined automatically via regression. Other works proposed local expansion operators to expand the content locally by considering neighboring pixels. For example, in [72], the local expand map is proposed by selecting a constellation of bright points through Gaussian filtering, and then expanding them through density estimation. In [66], the local expand map is generated through cross bilateral filtering. While all the aforementioned methods are automatic, with the exception of parameter tuning, there are also interactive methods. For example, in [13], user markups determine the saturated regions as light, reflections or diffuse surfaces, in which different expansion functions are then applied for each of the classified surfaces. In [50], another interactive method is presented where the user adjusts regional tonal balance of the final HDR image by using a piecewise linear function. However, all the above techniques can only expand

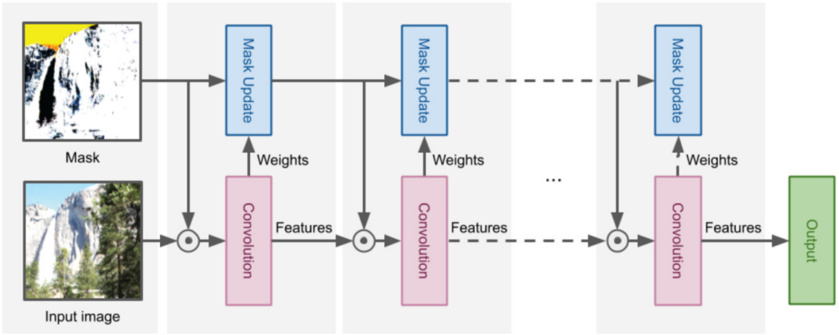


Figure 2.2: Feature mechanism as proposed in [77]

the contrast range but cannot reproduce missing details in saturated regions. Furthermore, most existing expansion operators derived from heuristics have difficulty handling significant under- or over-exposed LDR contents.

Recently, deep-learning has been extensively used in image recovery applications. Such data driven approaches have demonstrated improved performances compared to heuristic approaches in recovering missing details from under- or over-saturated LDR contents. Some works presented novel network architectures. For example, in [49], a three-branch convolutional neural network (CNN) is proposed to extract global, semi-local and local features to recover missing details in the saturated regions. In [16], a U-Net like architecture is implemented to predict values in the over-exposed regions with a fixed mask, and later blend the prediction with the input LDR image for the unsaturated regions. In [77], an autoencoder structure with a learnable feature masking mechanism is proposed to predict values in the over-exposed regions, and later blend the prediction with the input LDR image for the unsaturated regions. Figure 2.2 shows their proposed feature mechanism. The features at each layer are multiplied with the corresponding mask, computed based on the well-exposedness of the pixel, before going through the convolution process. The masks at each layer are obtained by updating the masks using the weights from the previous layer. In [60], [62], [90], Generative Adversarial Networks (GANs) are presented to introduce a hybrid loss that combines reconstruction loss and adversarial loss to recover realistic HDR content.

Other deep-learning works developed an end-to-end HDR framework that takes the image display or image formation pipeline into consideration. For example, in [94], an HDR-to-LDR framework is proposed to generate an HDR image, and a LDR image for visualization on conventional displays. They train a network for HDR reconstruction to restore the missing details from the input LDR image, and then a second network to transfer these details back to the LDR domain. In [43], an HDR framework that incorporates the domain knowledge of the LDR image formation pipeline is presented. They trained three specialized networks to reverse the image formation steps of dynamic range clipping, non-linear mapping from a CRF, and quantization, to reconstruct HDR images. Some works proposed to solve HDR and HDR-related tasks in a joint optimization framework. For example, in [33], a residual based network is implemented to learn the direct mapping from low resolution LDR video to their high resolution HDR version for displays on high-end TVs. They decomposed the image into base and detail components, then trained the network through separate feature extraction, and finally fused the images to obtain the high resolution HDR content. In a similar vein, [34] implements a GAN to convert low resolution LDR videos to high resolution HDR videos. As another example, in [29], a two-stage cascade network is designed to learn HDR image generation and HDR image color refinement to output accurate color representation of the physical scene. Alternatively, some works presented an end-to-end joint optimization for optics and HDR reconstruction. For example, in [57], [82], an optical based encoder and a CNN decoder are jointly trained to hallucinate the HDR content from a single LDR image.

As another one-shot single output method, several works investigated the use of snapshot sensors to acquire multiple EV LDR images in one acquisition, and output a single HDR image. In general, a snapshot sensor refers to an imaging device capable of acquiring multiple images in a single image capture (one-shot), as demonstrated in [2], [76], [81], [88]. This is beneficial for HDR imaging as more information about the scene can be captured in one-shot to help recover the scene's dynamic range. Some traditional works are implemented with heuristics strategies. For example, in [76], a snapshot sensor based on the polarization camera can acquire multiple polarization im-

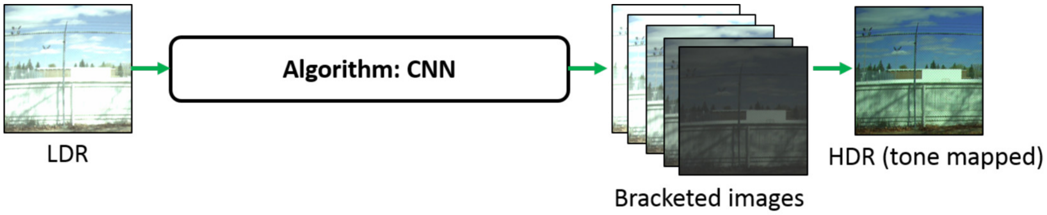


Figure 2.3: The one-shot multiple image output method HDR image formation pipeline

ages in one-shot. These polarization images are used to identify and eliminate the saturated pixels, and thus enhance contrast to achieve HDR. Later, [88] extended the work to examine the full HDR image reconstruction process by including the camera response calibration step.

There are also some deep-learning based approaches developed using snapshot sensors. For example, in cite [2], a snapshot sensor capable of acquiring multiple exposures in one-shot is co-designed with the HDR reconstruction network. They train an inception network to jointly optimize for demosaicking, HDR reconstruction, as well as the spatially varying modulation mask in the hardware. In [81], a deep snapshot HDR imaging framework is presented to reconstruct HDR values from the raw data captured using a ME-CFA sensor, which consists of a mosaic pattern of RGB filters with different exposure levels. They pre-train a luminance estimation network, then attach it to the main framework to reconstruct the final HDR values.

2.1.2 Multiple image output methods

Given a single input LDR image, the multiple image output methods design a framework to output bracketed LDR images at different exposures, which are then post-processed to generate the final single HDR image. Figure 2.3 illustrates the pipeline of multiple image output methods, which aim to recover an HDR image (32 bits/pixel) from the bracketed LDR images (8 bits/pixel), where each EV image infers a part of the luminance range. We note that the post-process fusion step is a separate task, which we do not tackle in our thesis; it focuses on generating a visually pleasing LDR image from the bracketed LDR images [45], [70], [89]. To the best of our knowledge, all multiple

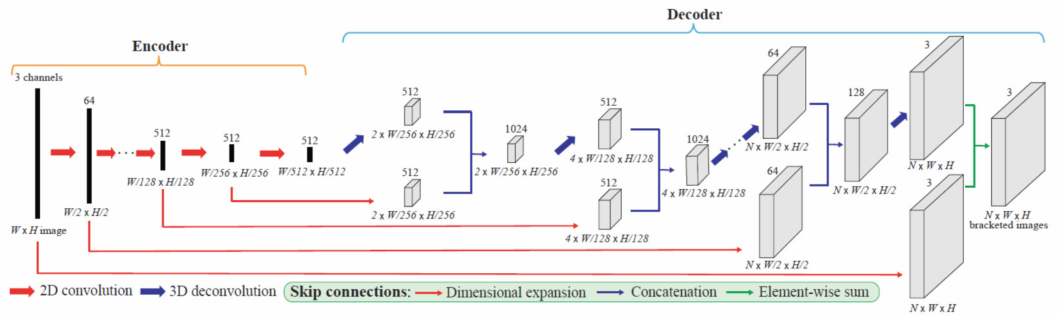


Figure 2.4: Network architecture as proposed in [17]

output methods are implemented with deep neural networks, and there are no heuristic works that transform a single LDR image to multiple LDR images at different EVs and then merge them into an HDR image.

Using deep-learning, some works presented novel network architectures for generating bracketed LDR images. For example, in [17], a modified U-Net architecture is used to generate the multiple exposure images from a single exposure image. These images are then merged to construct the final HDR image. Figure 2.4 shows their proposed network. The encoder consists of 2D convolutions and the decoder consists of 3D deconvolutions to generate consistent images with different exposures. In [40], a chained CNN structure is proposed to sequentially generate the bracketed LDR images. Later, [38], [39], proposed to handle this application through a recursive conditional GAN and a cycle GAN, respectively. Another work in [32] presented an end-to-end HDR framework that takes the CRF into consideration. They trained a recurrent network to generate the bracketed LDR images, and then added a differentiable HDR synthesis layer to learn the appropriate CRF, to reconstruct the final HDR image.

Here we summarize the pros and cons of the multiple image output method when compared to the single image output method. A pro of the multiple image output method over the single image output method is that it can alleviate the dataset quantity problem, as the focus is on transferring exposures to accurately generate the bracketed LDR stack. Another pro of the multiple image output method is that it can generate more accurate results, as each EV image infers a part of the luminance range rather than inferring the entire

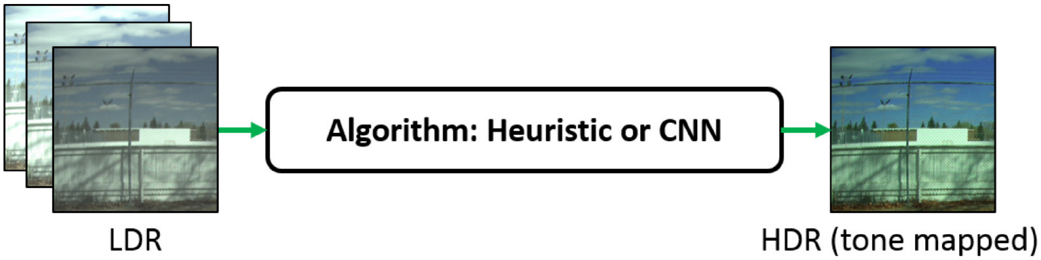


Figure 2.5: The multiple-shot method HDR image formation pipeline

range at once.

On the other hand, a con of the multiple image output method over the single image output method is that it requires a dataset of multiple EV LDR images per HDR image. Another con of the multiple image output method is that it can require a longer training time as two models: a down-exposure and an up-exposure model needs to be trained separately, to infer the full dim-to-bright set of LDR images.

In our study, we investigate both single image output [77] and multiple image output methods [17] for HDR reconstruction based on the snapshot sensor - polarization camera. For the post-process fusion step in the multiple image output method, we used the technique in [56] to fuse the bracketed LDR images. More details about our proposed methods are described in Section 3.

2.2 Multiple-shot image HDR reconstruction

This approach uses multiple acquisitions to reconstruct an HDR image. Figure 2.5 illustrates the pipeline of the multiple-shot image HDR reconstruction method. There is a series of works that focuses on removing ghosting artifacts caused by moving objects or misalignment in the images shot from multiple acquisitions for HDR reconstruction. Several traditional works perform alignment and HDR reconstruction in a unified optimization system. For example, in [79], a patch-based optimization system is proposed to fill in the missing details due to the under- or over-saturated regions in the reference image using other images within the stack. In [25], a similar patch-based system is proposed, but includes camera calibration as part of the optimization. Other works use rank minimization, in which image misalignment are considered as sparse outliers, to align and reconstruct a HDR image. For example in [65], a rank-1 matrix is proposed to reject ghost-artifacts and reconstruct an HDR image. In [36], a similar rank-1 matrix is proposed, but includes camera calibration as part of the unified optimization setup. Another work in [42] uses a content adaptive filtering scheme to simultaneously correct image misalignment and reconstruct an HDR image. However, the above techniques are not data driven and produces unsatisfactory results in challenging cases where the reference has significant under- or over-saturated areas.

Learning based approaches can help to alleviate the aforementioned issues and output better results. Some works proposed a pre-process step to align the images before feeding it to a CNN. For example, in [31], a pre-process step is proposed to first use optical flow to align the input images to the reference image, and then employ a CNN to obtain the HDR image. In [87], a pre-process step is introduced to first use homography transformation to align the background of the input images, and then use an autoencoder structure to translate multiple LDR images into a ghost-free HDR image. Other works implemented an end-to-end HDR framework. For example, in [93], an attention-merging network is presented to generate an HDR image with less ghosting artifacts, and restore details in the saturated regions. Their attention network detects useful regions and misaligned regions, while their merging network, based on a series of dilated residual dense blocks, merges the input images to reconstruct an HDR image. In [37], an alignment-merging network is proposed for ghost-free HDR imaging. Their alignment network aligns the input LDR images to the reference image, and the merging network, based on a residual dense block, merges the input images to restore an HDR image. In a similar vein, in [71], a pyramidal alignment and masked merging network is proposed to synthesize HDR images from the input LDR images. Their alignment network extracts image features at different scales and aligns them to the reference view, while their merging network, based on residual dense blocks, merges the input images to restore an HDR image. In [9], a non-local network is introduced to explicitly address the feature alignment problem to obtain a ghost-free HDR image. In [63], a GAN framework is proposed to handle this application. Alternatively, another work in [8] proposed to focus on reconstruction the details from the input images, and assumes the input images are aligned.

2.3 Summary

In this chapter, we discussed the two categories of approaches that can be used to group the existing HDR reconstruction methods. These two categories are: one-shot image HDR reconstruction, and the multiple-shot image HDR reconstruction. For each category, we first reviewed the traditional works, which are non-learning based that rely on heuristics to expand the dynamic range. Then, we reviewed the deep-learning based works and described their strategies to improve HDR imaging. In the one-shot image HDR reconstruction category, we further identify two classes of methods which are: single image output and multiple image output methods. Our work focuses on developing a deep-learning based approach for HDR reconstruction using the one-shot image HDR reconstruction approach. We study and adopt a framework from each class of this category, and present them in Chapter 3 Proposed Deep Snapshot HDR Reconstruction.

Chapter 3

Proposed Deep Snapshot HDR Reconstruction

Our work tackles the HDR reconstruction problem by using a combination of hardware and software approaches. In this section, we present the hardware and software approaches used in our method. First, for hardware, we use the polarization camera as the capture device, and then show how this new camera sensor can help to narrow the mismatch between the dynamic range in natural scenes and the capturing device, to reconstruct a more informative image. Second, for software, we propose two deep snapshot HDR imaging frameworks: Deep Snapshot Multiple image output HDR (DSMHDR), and Deep Snapshot Single image output HDR (DSSHDR). We use a snapshot sensor - polarization camera, with a deep neural network to design a deep-learning based approach for HDR reconstruction. The first, DSMHDR, brings in the ideas of using multiple polarization images as input for the network, to reconstruct HDR images. The second, DSSHDR, builds on these ideas and proposes a new framework that further integrates polarimetric cues available from the polarization camera, to reconstruct HDR images.

3.1 Polarization camera for scene capture

The polarization camera captures the polarization information of a scene. Polarization is one of the three fundamental properties of light, along with color and intensity. Specifically, polarization describes the direction in which light

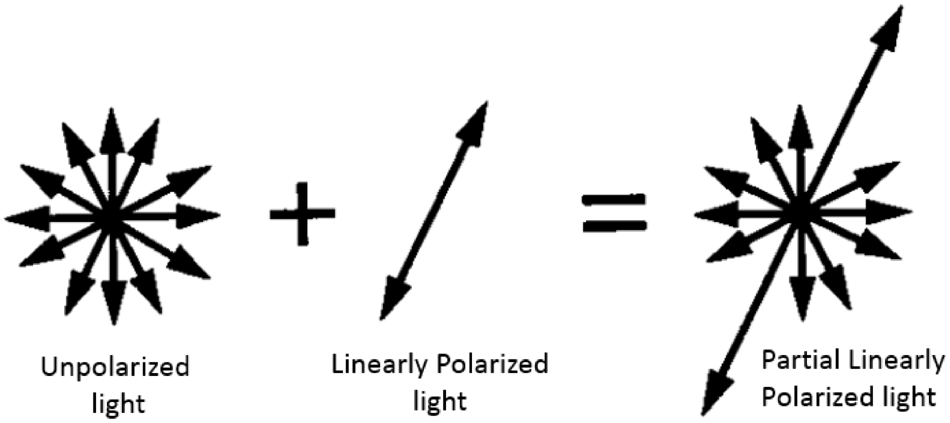


Figure 3.1: Polarization states of light

as an electromagnetic wave oscillates. Figure 3.1 illustrates three polarization states of light, in which the light is unpolarized when it oscillates at more than one angle, linearly polarized when it oscillates at a single angle, and partially linearly polarized when it oscillates at more than one angle but stronger in a particular angle. Such polarized light is everywhere [86]. Light can be polarized when scattered by particles in the atmosphere. This scattering occurs in the atmosphere due to the gas molecules and dust particles, where the light can be as much as 70% linearly polarized when observed at an angle of 90° to the sun. Light can also be polarized by refraction underwater, where the transmission of the light at the air/water interface causes significant polarization. Finally, light can also be polarized when reflected from a surface, and the strength of polarization can depend upon the reflecting material. Polarization is a key property of light, and has been utilized by many creatures, such as insects, birds and marine animals to achieve their visually guided behaviours. For example, bees and birds use the polarization pattern in the sky to aid navigation [24]. Several marine creatures also use similar patterns found underwater for the same purpose [11]. Despite the HVS's ability to identify the color and intensity of light, the HVS is blind to this light polarization.

With recent advances in camera technology, polarization effects can now be captured by imaging devices such as polarizers, and more recently, polarization cameras. The polarizers serve to attenuate irradiance and alleviate over-exposure by potentially boosting the contrast of dark regions of the en-

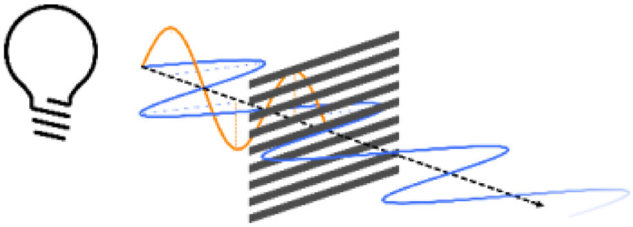


Figure 3.2: The polarizer filter passes the blue beam that is aligned parallel to its polarizer axis, and blocks the orange beam that is aligned perpendicular to its polarizer axis [27]

vironment to enable HDR imaging. Like filters, polarizers select a specific polarization of light while blocking the rest, as shown in Figure 3.2. The different angles of polarization can be achieved by mechanically rotating the polarizer placed in front of the camera lens. However, a polarization camera integrates the polarizers onto the surface of the sensors, which replaces the need for mechanically rotating the polarizers. The polarized image is thus captured by forcing light through a polarizer. Recently, Sony has introduced two CMOS sensors: IMX250MZR (mono) [28] and IMX250MYR (color) [26] where both sensors use nano-scale fabrication techniques to create a division-of-focal plane imaging sensor. Figure 3.3 shows the polarization cameras and their physical layouts. The camera integrates four on-chip directional polarizers, at 0°, 45°, 90°, and 135° onto the camera sensor (one calculation unit), to capture four spatially and temporally aligned high-resolution polarized images of a scene in real-time. The difference between the two sensors is that the IMX250MZR (mono) captures grayscale images while the IMX250MYR (color) captures color images. In our study, we use the IMX250MYR (color) to collect a colored dataset to train the HDR polarization model.

3.2 Polarization image formation (HDR)

When light reflects off a non-metallic object, it becomes partially polarized. This polarized light can be captured by a polarization camera, as shown in Figure 3.4. The image irradiance I_0, an attenuated version of the scene radiance, is first filtered by the on-chip directional micro-polarizers. Then the

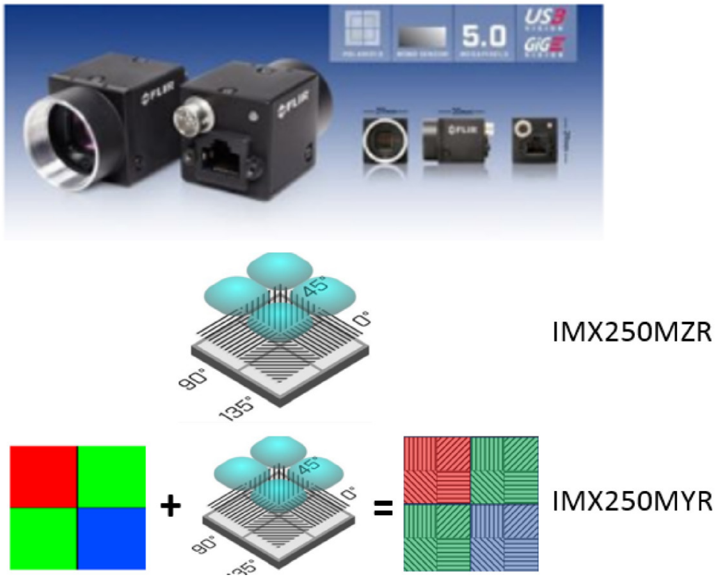


Figure 3.3: Polarization cameras and their physical layouts

camera's photosensitive elements convert the light signal, using the CRF, into four digital images L_1 to L_4. During this process, the camera has an exposure time t_0 (not shown), which can be varied to adjust the polarization images. The effect of a polarizer on image irradiance can be written as [19]:

$$I_i = 0.5 \times I_0\Big(1 + \rho\cos(2\theta - 2\alpha_i)\Big) \tag{3.1}$$

where ρ denotes DoP, θ denotes AoP, $\alpha_i = 0°$, $45°$, $90°$, $135°$ denotes the angle of the polarizers, and $i = 1, 2, 3, 4$ denotes the index of the four polarizers. Substituting the polarizer angles into Equation 3.1, we get the filtered images:

$$\begin{aligned} I_1, I_3 &= 0.5 \times I_0(1 \pm \rho\cos 2\theta) \\ I_2, I_4 &= 0.5 \times I_0(1 \pm \rho\sin 2\theta) \end{aligned} \tag{3.2}$$

The DoP (ρ) which measures the portion of light that is polarized for a given pixel is computed from the four polarization images, and is in the range of [0,1]. If $\rho = 0$, the light is unpolarized; if $\rho = 1$, the light is completely polarized; when $0 \leq \rho \leq 1$, the light is partially polarized. ρ can be computed by:

$$\rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \tag{3.3}$$

where S_0, S_1 and S_2 are the intermediate variables called Stokes and represent the total power, the power of the 0° linear polarization, and the power of the

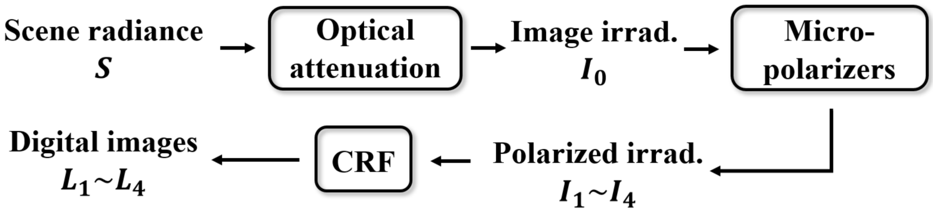


Figure 3.4: Polarization camera image acquisition pipeline

45° linear polarization, respectively. S_0, S_1 and S_2 can be computed by [19]:

$$\begin{aligned} S_0 &= 0.5 \times (L_1 + L_2 + L_3 + L_4) \\ S_1 &= L_1 - L_3 \\ S_2 &= L_2 - L_4 \end{aligned} \tag{3.4}$$

where L_1, L_2, L_3 and L_4 are the captured images with the polarizer filters at 0°, 45°, 90°, and 135°, respectively.

The AoP (θ) which measures the direction of light in which the polarized light oscillates at a given pixel is computed from the four polarization images, and is in the range of [0°, 180°]. If $\theta = 0$, the light is oscillating horizontally; if $\theta = 90°$, the light is oscillating vertically. θ can be computed by:

$$\theta = 0.5 \times \tan^{-1}\left(\frac{S_2}{S_1}\right) \tag{3.5}$$

In general, the relation between image irradiance I_i and pixel value L_i at exposure time t_0 can be written as [12]:

$$L_i = f(I_i t_0) \tag{3.6}$$

where f is the CRF. In particular, we use the method in [58] to calculate the CRF. While there are other methods that can be used to compute the response, we selected this method because it is formulated in such a way as is applicable to the polarization camera, as will be explained shortly. We briefly summarize the method below for the completeness of the presentation.

The method [58], rather than identifying f, estimates the inverse CRF (ICRF), g, which is related to Equation 3.6 as:

$$I_i = \frac{g(L_i)}{t_0} \tag{3.7}$$

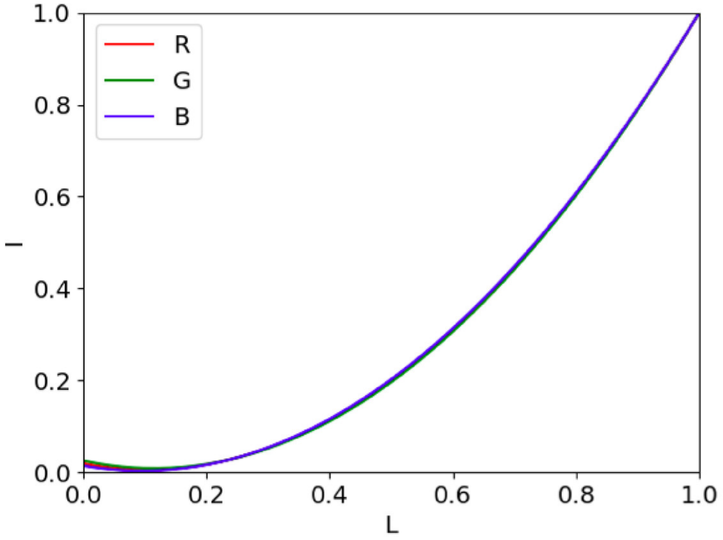


Figure 3.5: ICRF of the polarization camera

which is of direct use in HDR reconstruction. g can be modeled using a polynomial of appropriate order N as:

$$g(L_i) = \sum_{n=0}^{N} c_N L_i^N \tag{3.8}$$

With this formulation, the calibration process is viewed as one of determining the order N as well as the coefficients c_n. When multiple images are taken at different EVs, the calibration algorithm makes use of the observation that the ratio between exposures is the same as that between the scene luminance, as is defined by Equation 3.8. Using the multiple available ratios between known exposures, this observation is used to set up constraints on the coefficients c_n of Equation 3.8 whose pixel value L_i is captured and therefore known. This is applicable for the polarization camera because the basic constraint used in [58] takes the form of a ratio between scene luminance, the filtering effect of the polarizer appears on both numerator and the denominator of the ratio and cancels each other. The algorithm can therefore be used as is to calibrate the polarization camera.

We have used the above calibration method to perform radiometric calibration of our polarization camera, and the ICRF is plotted in Figure 3.5. Note that both the pixel intensity measurement and the scene luminance are on a normalized scale. In our experiment, we used a total of 17 different exposures.

We found $N = 2$ provided the optimal result as the goodness of fit in terms of R-squared for the RGB channels are 0.928, 0.925, 0.931, respectively. Thus, the coefficients of each color channels are:

$$\begin{array}{cc} & \begin{matrix} \quad c_0 & \quad c_1 & \quad c_2 \end{matrix} \\ \begin{matrix} R \\ G \\ B \end{matrix} & \begin{bmatrix} 0.0196 & -0.2596 & 1.2400 \\ 0.0256 & -0.2918 & 1.2662 \\ 0.0144 & -0.2332 & 1.2188 \end{bmatrix} \end{array}$$

This ICRF is then used to create HDR images from polarization images as will be described in Section 4.1.

Finally, by substituting Equation 3.2 into Equation 3.6, and applying the reciprocity relation in [12], we obtain:

$$\begin{aligned} t_1, t_3 &= 0.5 \times t_0(1 \pm \rho \cos 2\theta) \\ t_2, t_4 &= 0.5 \times t_0(1 \pm \rho \sin 2\theta) \end{aligned} \tag{3.9}$$

Equation 3.9 provides the polarization image formation model in the case of HDR reconstruction. When the incoming light is not entirely unpolarized ($\rho \neq 0$), the four pixels within one calculation unit of the polarization camera experience different exposure times, effectively creating the condition for multiple exposures. In the extreme case, when $\rho \approx 1$ one can expect a large difference in exposure between the four pixels.

Different from a conventional camera capturing multiple exposures, the variation in exposure time in a polarization camera is pixel specific as both ρ and θ vary from pixel to pixel. This is because the pixels in different calculation units of a polarization camera correspond to different points in space, which in general differ in terms of their light polarization, just as in color or in intensity.

To quantify the potential gain in dynamic range of a polarization camera versus a conventional camera, we compare the dynamic range of the two cameras by computing Equation 3.10:

$$DR = 20 \log \left(\frac{L_{max}}{L_{min}} \frac{t_{max}}{t_{min}} \right) \tag{3.10}$$

where L_{max} corresponds to the full-well capacity of the sensor, and L_{min} corresponds to the minimum signal detectable by the sensor. t_{max} and t_{min} are the

Table 3.1: Change in dynamic range as a function of DoP (ρ)

ρ		0.2	0.5	0.8
$\triangle$	DR (dB)	3.2	8.5	16.4
	# of bits	0.53±0.05	1.40±0.16	2.66±0.41

maximum and minimum exposure times, respectively. For a conventional ideal 8 bit camera, its dynamic range $= 20\log(255) = 48.13$ dB, and is fixed. For a polarization camera, the dynamic range can vary with the scene with a minimum of 48.13 dB and a maximum that depends on the ratio of the exposure times; where $t_{max} = \max(t_1, t_2, t_3, t_4)$ and $t_{min} = \min(t_1, t_2, t_3, t_4)$ within one calculation unit. The dynamic range of the entire camera can be computed using the mean of the dynamic ranges of all its calculation units.

Table 3.1 summarizes the results of the dynamic range based on the evaluation of Equation 3.10 under three different values of ρ. In the calculation of t_1, t_2, t_3, t_4, using Equation 3.9, we assume a uniform distribution for the AoP. We also calculate the change in dynamic range in terms of the number of additional bits in the pixel depth, i.e., $log_2\big(\frac{t_{max}}{t_{min}}\big)$. At $\rho = 0.2$ or a moderate amount of polarization in the environment light, we can expect to increase the dynamic range by 0.53 bits or 3.2 dB. On the other hand, for space points with significant light polarization with $\rho = 0.8$, the dynamic range of the images can be increased by 2.66 bits, from 8 bits to 10.66 bits, or by 16.4 dB, from 48.13 dB to 64.53 dB. In Table 3.1, the standard deviation of the change in dynamic range ($\triangle$) is due to the variation in the AoP (θ). Practically, ρ can vary spatially and, as a result, so can the different regions of an image in terms of their gain in dynamic range from this polarization camera.

3.3 Deep Snapshot Multiple output HDR framework (DSMHDR)

Here, we study and propose a multiple output deep-learning based method within the category of one-shot image HDR reconstruction to construct our DSMHDR framework. There are several deep-learning works proposed to output multiple exposure LDR images to combine and form an HDR output from a single LDR input [17], [32], [38]–[40], and have been described in Section 2.1.2. In particular, the work by Endo et al. [17] uses a modified U-Net architecture to predict multiple exposure images from a single exposure image, which are then merged to output an HDR image. This method in [17] is able to predict any number of bracketed images by simply changing the size of the training input image list. Additionally, [17] reconstructs HDR through supervised training, where the CNN learns the correlation between saturated and unsaturated regions to fill in the missing pixel values in the saturated regions. On the other hand, other methods require modification to the architecture, such as extending the sub-networks in [38]–[40] to increase the network depth, so the method can be used to infer a different size of bracketed set than what the network was originally trained on. However, such modification increases network complexity which is more likely to overfit to the training data [75]. Nevertheless, these methods have demonstrated their feasibility to reconstruct HDR images. However, they are all developed for conventional cameras, and thus is not suitable to be applied directly on polarization images.

Unlike [17], our proposed DSMHDR framework is developed to directly handle polarization images. It uses a more informative input derived from the polarization images to improve HDR reconstruction. Namely, we propose a pre-processing step to fuse the four LDR polarization images before feeding it to the network. Based on Equation 3.9, which relates the polarization images to different exposure times, the fusion step can be implemented by a pixel-weighting function. It assigns weights to the four pixels in one calculation unit depending on how well-exposed the pixels are as was proposed by Debevec in [12], and adopted in [88]. The resulting image I_{Deb} can be computed by:

$$I_{Deb} = \frac{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)\Big(g\big(L_i\big) + g\big(L_{i+2}\big)\Big)}{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)t_0} \tag{3.11}$$

where W is the Gaussian weighted function ($\sigma = 0.2$ in our study). Intuitively, Equation 3.25 transforms the input to the HDR space, which conveys more information as I_{Deb} is in floating point and spans a wider dynamic range than

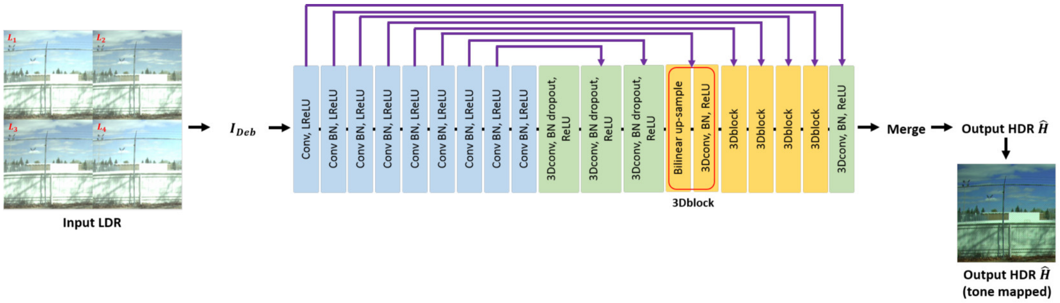


Figure 3.6: DSMHDR Framework

I_i.

Then we feed the I_{Deb} image into the network. Figure 3.6 illustrates the DSMHDR network. We first briefly summarize the architecture for the completeness of the presentation, and then discuss the improvements we introduced to the network to improve HDR reconstruction. The network presented in [17] and adopted for DSMHDR is an 18 level autoencoder architecture. The architecture consists of 9 levels for the encoder and 9 levels for the decoder. In the encoder, the first level consists of a 2D convolutional layer and a LeakyReLU (negative slope = 0.2) layer. The subsequent levels consist of a 2D convolutional layer, batch normalization layer followed by a LeakyReLU layer. In the decoder, the first two levels consist of a 3D deconvolutional layer, batch normalization layer, dropout layer followed by a ReLU layer. The subsequent five levels consist of a bilinear up-sample layer, convolutional layer, batch normalization layer followed by a ReLU layer. The final two levels consist of a 3D deconvolutional layer, batch normalization layer followed by a ReLU layer. We use skip connections between all the encoder layers and their corresponding decoder layers. The architecture generate 17 different exposure images, and then merge the images using the technique in [56] to output an HDR image. Details on the network architecture is summarized in Table 3.2.

Unlike [17], our proposed DSMHDR framework introduces the following improvements: 1) We replace the deconvolution operation with a bilinear interpolation operation. This reduces the visible tiling artifacts when the input image has large saturated regions [64]. We apply this operation to five of the deconvolutional layers due to memory restrictions. We observe that applying

Table 3.2: Overview of DSMHDR network architecture

Layer	Stage	# filters	Filter size	Conv. stride	Spatial pad	Activation
1	conv+act	64	(3,3)	(2,2)	(1,1)	LeakyReLU
2	conv+bn+act	64	(3,3)	(2,2)	(1,1)	LeakyReLU
3	conv+bn+act	128	(3,3)	(2,2)	(1,1)	LeakyReLU
4	conv+bn+act	256	(3,3)	(2,2)	(1,1)	LeakyReLU
5	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
6	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
7	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
8	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
9	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
10	conv+bn+dropout[1]+act	512	(4,4,4)	(2,2,2)	(1,1,1)	ReLU
11	conv+bn+dropout[1]+act	512	(4,4,4)	(2,2,2)	(1,1,1)	ReLU
12	conv+bn+dropout[1]+act	512	(4,4,4)	(2,2,2)	(1,1,1)	ReLU
13	upsample+conv+bn+act	512	(3,4,4)	(1,2,2)	(1,1,1)	ReLU
14	upsample+conv+bn+act	512	(3,4,4)	(1,2,2)	(1,1,1)	ReLU
15	upsample+conv+bn+act	256	(3,4,4)	(1,2,2)	(1,1,1)	ReLU
16	upsample+conv+bn+act	128	(3,4,4)	(1,2,2)	(1,1,1)	ReLU
17	upsample+conv+bn+act	64	(3,4,4)	(1,2,2)	(1,1,1)	ReLU
18	conv+bn+act	128	(3,4,4)	(1,2,2)	(1,1,1)	ReLU

[1] 50% dropout rate

the up-sampling operation to the five deconvolutional layers demonstrates the removal of tiling artifacts. 2) We add the SSIM loss to the loss function. This additional loss term improves the perceptual quality of the synthesized content [100]. Our overall loss function is:

$$\mathcal{L}(W,b) = \alpha \times \mathbb{E}[|H - \hat{H}|] + \beta \times SSIM(H, \hat{H}) \tag{3.12}$$

where H and $\hat{H}$ denotes the ground truth and predicted image, respectively. The ground truth image is constructed from the dataset, as discussed in Section 4.1.1. The first term in Equation 3.12 corresponds to the mean absolute error (MAE) loss which measures the pixel-level loss between the predicted and ground truth images. By minimizing the MAE loss, details can be better recovered. The second term in Equation 3.12 corresponds to the SSIM loss, which measures the perceptual loss between the predicted and ground truth images. By maximizing the SSIM loss, edges and contrast details can be better preserved to synthesize visually pleasing textures. We empirically set $\alpha = 1.0$, and $\beta = 0.1$ in our experiments. Additionally, during inference, the final HDR image is obtained directly from the predicted image $\hat{H}$.

3.4 Deep Snapshot Single-output HDR framework (DSSHDR)

Unlike DSMHDR which outputs multiple images, here we study and propose a method which outputs a single image, that also falls within the category of one-shot image HDR reconstruction. While the DSMHDR method has demonstrated its feasibility to estimate missing pixels in saturated regions, training time remains a key challenge in investigating potential framework solutions as it is a multiple output based one-shot image HDR reconstruction approach. Therefore, after the development of the DSMHDR framework, we studied and proposed a single output based one-shot image HDR reconstruction approach, to construct our DSSHDR framework.

Also unlike DSMHDR that uses only the polarization images for HDR reconstruction, DSSHDR integrates additional polarization information, specifically from the DoP image, to improve HDR reconstruction. As mentioned in Section 3.2, the polarization state of light can be quantified by DoP, which determines the portion of light that is polarized for a given pixel. Namely, DoP is a ratio of the power of the polarized light versus the total power, which can be computed by Equation 3.3, and is in the range of [0,1]. Large DoP values correspond to a strong measure of polarization. Equations 3.3 and 3.4 show that a large DoP values are obtained where there are large pixel variations among the polarization images (L_1, L_2, L_3, L_4). In other words, regions with large DoP values indicate where the polarization camera can provide rich and reliable polarimetric cues for HDR recovery. Equipped with this domain knowledge about DoP, we can employ the DoP image as a strong accurate prior for the DSSHDR framework.

There are several one-shot deep-learning works presented to output a single HDR image from a single LDR input image [16], [43], [49], [77], [90], and have been described in Section 2.1.1. In particular, the work by Santos et al. [77]

uses an autoencoder structure with a learnable feature mechanism to regress details in the over-saturated regions, and then reconstruct the HDR image through blending with the input's unsaturated regions. Namely, the feature masking mechanism in [77] is based on a soft mask computed from the pixel intensity of the input image. The soft mask maps the unsaturated pixels to one, and the saturated pixels to zero. Thus, dark pixels in the soft mask correspond to saturated pixels where the CNN focuses it learning upon, and the soft mask is continuously updated by the feature weights of each convolutional layer. Furthermore, [77] uses an input image and soft mask to reconstruct the HDR image through supervised training, where the CNN learns the correlation between saturated and unsaturated regions to fill in the missing pixel values in the saturated regions. Additionally, it also utilizes the weak textures in the saturated regions, obtained by the soft mask, to fill in the missing details. [77] claims that such unfixed masking strategy can reconstruct better HDR images that are free from visible artifacts as compared to using a fixed mask. While these methods have demonstrated their feasibility, robustness remains a key challenge. For example, in [16], the method ignores HDR reconstruction for under-saturated regions, and implements a fixed feature masking mechanism which generates visible artifacts in the results. As another example, in [77], the method neglects HDR reconstruction for under-saturated regions, is unable to improve HDR estimation for unsaturated regions, and has unreliable reconstruction in saturated regions due to the lack of prior to enforce consistency. Furthermore all of the aforementioned methods are developed for conventional cameras which captures images differently than the polarization camera, and thus is not suitable to be applied directly on polarization images.

Figure 3.7 illustrates the DSSHDR framework. Unlike [77], our proposed DSSHDR framework can reconstruct HDR for all unsaturated and saturated pixels with the help of DoP as a strong accurate prior. In our HDR reconstruction method, the first step is to compute the image L_c for input to the network. L_c computes the mean of the four LDR polarization images, thereby accounting for the contributions from each polarization filter, and is defined

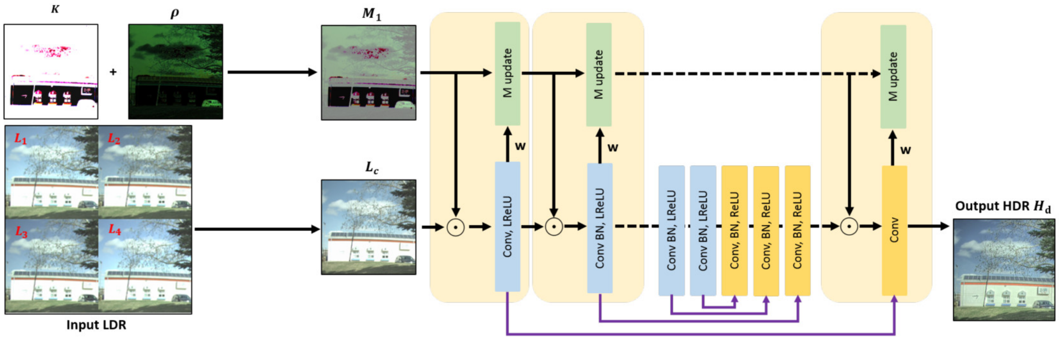


Figure 3.7: Overview of the deep neural network used in constructing H_d in our proposed DSSHDR framework. The LDR input image L_c is propagated through the network, while udpated by the corresponding mask M_l before going through the convolutional layers. The mask at each layer is obtained by updating the mask from the previous layer.

as follows:

$$L_c = \frac{L_1 + L_2 + L_3 + L_4}{2} \tag{3.13}$$

where L_1, L_2, L_3, L_4 are the normalized LDR polarization images in the range of [0,1], for filters oriented at 0°, 45°, 90°, and 135°, respectively.

Next, we compute the feature mask M_1 as input to the feature masking mechanism. The feature mask is constructed where if the pixel value of L_c is properly exposed and DoP is low, then the reconstruction is dominated by the predicted image H_d. However, if L_c is poorly exposed and DoP is high, then the reconstruction is dominated by the traditional model based image H_t. Then for L_c and DoP values that lie somewhere in between will share the reconstructions results of H_d and H_t. M_1 is in the range of [0,1], and is defined as follows:

$$M_1 = \frac{\rho + K}{max(\rho + K)} \tag{3.14}$$

where the subscript 1 indicates the first layer in the CNN ($l = 1$). ρ is the DoP for the input image which indicates the measure of polarization for a pixel, and is in the range of [0,1]. K is also in the range of [0,1], and indicates the proper exposedness of each input pixel based on the pixel intensity as shown in Figure 3.8. $K = 0$ indicates the input pixel is completely over-exposed. Therefore, M_1 computes the well exposedness of each input pixel based on the combination of polarization and intensity information. $M_1 = 0$

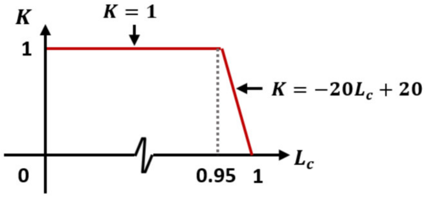


Figure 3.8: We use this function [77] to measure how properly exposed a pixel is. THe value 1 indicates the pixel as properly exposed, while 0 indicates the pixel as completely over-exposed. In our implementation we set the threshold at 0.95.

indicates the feature are computed from low polarization and poorly exposed pixels, and thus are invalid content that require reconstruction by the CNN. Then the feature mask is updated at each convolutional layer l to obtain M_l. As a result, the reconstruction is achieved by using the mask M_l to reduce the magnitude of the feature generated from the invalid content. Specifically updating the feature maps X_l extracted at each convolutional layer as follows:

$$Z_l = X_l \cdot M_l \tag{3.15}$$

where with the abuse of notation the multiplication sign $\cdot$ means the pixel-wise multiplication where appropriate throughout the thesis. In addition, the mask at each layer is computed by applying the convolutional filter to the mask at the previous layer. Since the masks are in the range of [0,1] and weights the contributions of the features, the magnitude of the filters is irrelevant. Therefore, we normalize the filter weights before convolving them with the mask as follows:

$$M_{l+1} = \left(\frac{|W_l|}{\|W_l\|_1 + \epsilon} \right) * M_l \tag{3.16}$$

where $\|\cdot\|_1$ is the l_1 function, $|\cdot|$ is the absolute operator, and $\epsilon = 10^{-6}$ is a small constant added to avoid division by 0.

Our loss function is a combination of an HDR reconstruction loss $\mathcal{L}_r$ and a perceptual loss $\mathcal{L}_p$ as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_p \tag{3.17}$$

where $\lambda_1 = 6.0$ and $\lambda_2 = 1.0$ in our implementation.

The reconstruction loss computes the pixel-wise l_1 distance between the ground truth image H_g and predicted image H_d for the invalid content defined as follows:

$$\mathcal{L}_r = \left\| (1 - M_1) \cdot (H_g - H_d) \right\|_1 \tag{3.18}$$

The perceptual loss has been demonstrated useful to improve visual quality in [77], [82], and its key idea is to enhance the similarity in feature space between the ground truth and predicted images. Our perceptual loss is a combination of the VGG loss $\mathcal{L}_v$ and style loss $\mathcal{L}_s$ as follows:

$$\mathcal{L}_p = \lambda_3 \mathcal{L}_v + \lambda_4 \mathcal{L}_s \tag{3.19}$$

where $\lambda_3 = 1.0$ and $\lambda_4 = 120.0$ in our implementation.

The VGG loss is defined as follows:

$$\mathcal{L}_v = \sum_l \left\| \phi_l(H_g) - \phi_l(H_d) \right\|_1 \tag{3.20}$$

where ϕ_l is the feature map from the l-th layer of the pretrained VGG-19. Then to recover more vivid textures, we add the style loss to Equation 3.19, and is defined as follows:

$$\mathcal{L}_s = \sum_l \left\| G_l(H_g) - G_l(H_d) \right\|_1 \tag{3.21}$$

where G_l is the Gram matrix [18] applied on the feature map at the l-th layer of the pretrained VGG-19, and is defined as follows:

$$G_l(X) = \frac{1}{K_l} \phi_l(X)^T \phi_l(X) \tag{3.22}$$

where K_l is a normalization factor computed as $H_l W_l C_l$. The feature ϕ_l is a matrix of shape $(H_l W_l) \times C_l$, and thus the Gram matrix has a size of $C_l \times C_l$. For the perceptual loss, we extract feature maps from pool1, pool2 and pool3 layers of the VGG-19 network.

Figure 3.9 shows the mask M_l at different layers l of the network. We can observe from the input image that the building and the fence regions are over-exposed. We can also observe from the input mask that these regions have a low measure of polarization. The input mask informs the network where the

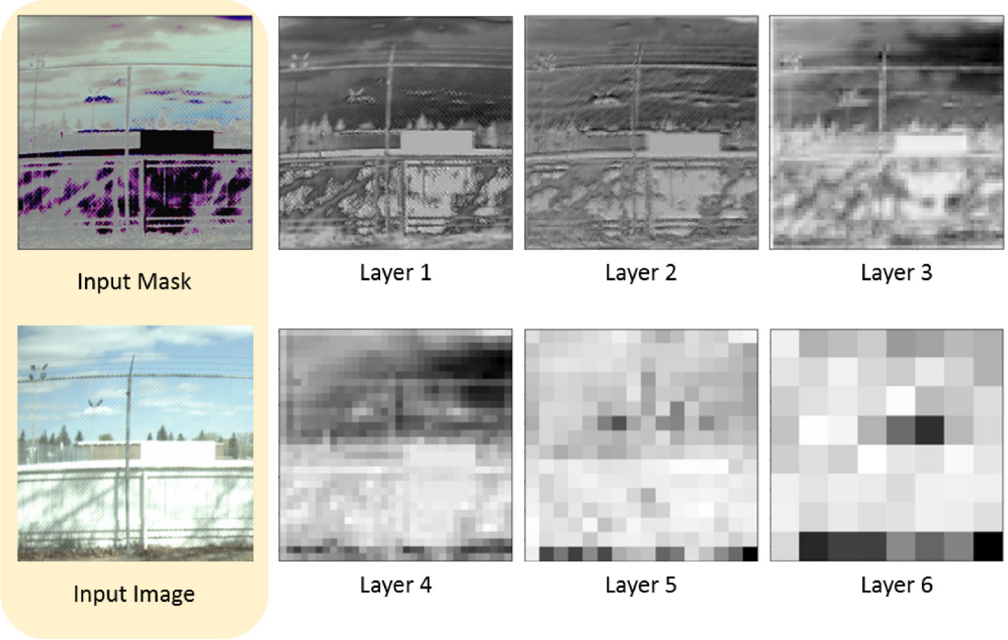


Figure 3.9: Visualization of the masks M_l at different layers of the network, and identifies regions that require CNN to direct its learning upon with higher feature weights.

invalid features are for the CNN to direct its learning upon. Therefore, the building and the fence regions consistently have a higher feature weight (i.e., brighter pixels) compared to other regions in the masks at each convolutional layer. As we move deeper through the network, the masks become blurrier and more uniform. This is expected since the receptive field of the features become larger in the deeper layers.

With the CNN predicted image H_d, we formulate the final HDR image H through HDR integration as follows:

$$H = \alpha \cdot H_t + (1 - \alpha) \cdot H_d \tag{3.23}$$

where α is as follows:

$$\alpha = \frac{\rho}{\rho + (1 - M_1)} \tag{3.24}$$

In Equation 3.23, H_t is computed as follows [88]:

$$H_t = \frac{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)\Big(g(L_i) + g(L_{i+2})\Big)}{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)t_0} \tag{3.25}$$

Table 3.3: Overview of DSSHDR network architecture

Layer	Stage	# filters	Filter size	Conv. stride	Spatial pad	Activation
1	conv+act	64	(7,7)	(2,2)	(3,3)	ReLU
2	conv+bn+act	128	(5,5)	(2,2)	(2,3)	LeakyReLU
3	conv+bn+act	256	(5,5)	(2,2)	(2,3)	LeakyReLU
4	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
5	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
6	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
7	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
8	conv+bn+act	512	(3,3)	(2,2)	(1,1)	LeakyReLU
9	conv+bn+act	512	(3,3)	(1,1)	(1,1)	LeakyReLU
10	conv+bn+act	512	(3,3)	(1,1)	(1,1)	LeakyReLU
11	conv+bn+act	512	(3,3)	(1,1)	(1,1)	LeakyReLU
12	conv+bn+act	256	(3,3)	(1,1)	(1,1)	LeakyReLU
13	conv+bn+act	128	(3,3)	(1,1)	(1,1)	LeakyReLU
14	conv+bn+act	64	(3,3)	(1,1)	(1,1)	LeakyReLU
15	conv+act	3	(3,3)	(1,1)	(1,1)	LeakyReLU

where W is the Gaussian weighted function ($\sigma = 0.2$ in our study), and g is the inverse camera response function.

The HDR formulation by Equation 3.23 uses a combination of H_t which is a traditional model based method, and H_d which is a deep-learning based method, to estimate HDR in all areas. H_t can estimate HDR well for all areas with high DoP [88]. Then we let H_d to estimate HDR for all areas with low DoP, using the predicted pixels from the CNN. Additionally, normalization is performed where necessary.

The DSSHDR network configuration adopts the 15 level autoencoder arhictecture in [77]. The architecture consists of 8 levels for the encoder and 7 levels for the decoder. In the encoder, the first level consists of a 2D convolutional layer and a ReLU layer. The subsequent levels consist of a 2D convolutional layer, batch normalization layer followed by a ReLU layer. In the decoder, the first six levels consists of a 2D convolutional layer, batch normalization layer followed by a LeakyReLU (negative slope = 0.2) layer. The final layer consists of a 2D convolutional layer. We use skip connections between all the encoder layers and their corresponding decoder layers. We use the feature masking strategy in all the convolutional layers and up-sample the features in the decoder using nearest neighbor. The architecture outputs an HDR image. Details on the network architecture is summarized in Table 3.3

3.5 Summary

In this chapter, we discussed our proposed methods to tackle the HDR reconstruction problem. Our methods are a combination of hardware and software approaches. We presented the polarization camera, and showed that the extent of irradiance attenuation by a polarizer filter is similar to multiple exposures, and thus aids in HDR reconstruction. Then we proposed two frameworks to achieve one-shot image HDR reconstruction. The results of our proposed methods are presented in Chapter 4 Experiments.

Chapter 4

Experiments

This chapter first introduces the dataset used in our experiments; here we describe the image processing pipeline used to create the images, then the dataset statistics to provide an overview of our dataset. Next, the experimental setups are discussed including the implementation details and the evaluation metrics. Then, we present ablation studies to quantify the effectiveness of the different components within our methods. Finally, we compare our method against competing algorithms.

4.1 Dataset

4.1.1 Image processing pipeline

Since there is a lack of a public dataset for training and testing HDR techniques with polarized images, we collected a dataset. The dataset is collected by the IMX250MYR (color) polarization camera, as described in Section 3.1. During data acquisition, the polarization camera is mounted on a sturdy tripod, and the Spinnaker SDK acquisition provided by the camera manufacturer, is launched with our custom script and settings summarized by Table 4.1. For each scene, a high-resolution 2448x2048 colored image is captured at 17 exposure times. The exposure times are:

$$\begin{aligned} t_0 = [&0.03, 0.045, 0.068, 0.101, 0.152, 0.228, 0.342, 0.513, 0.769, \\ &1.153, 1.73, 2.595, 3.592, 5.839, 8.758, 13.137, 19.705]\, ms \end{aligned} \tag{4.1}$$

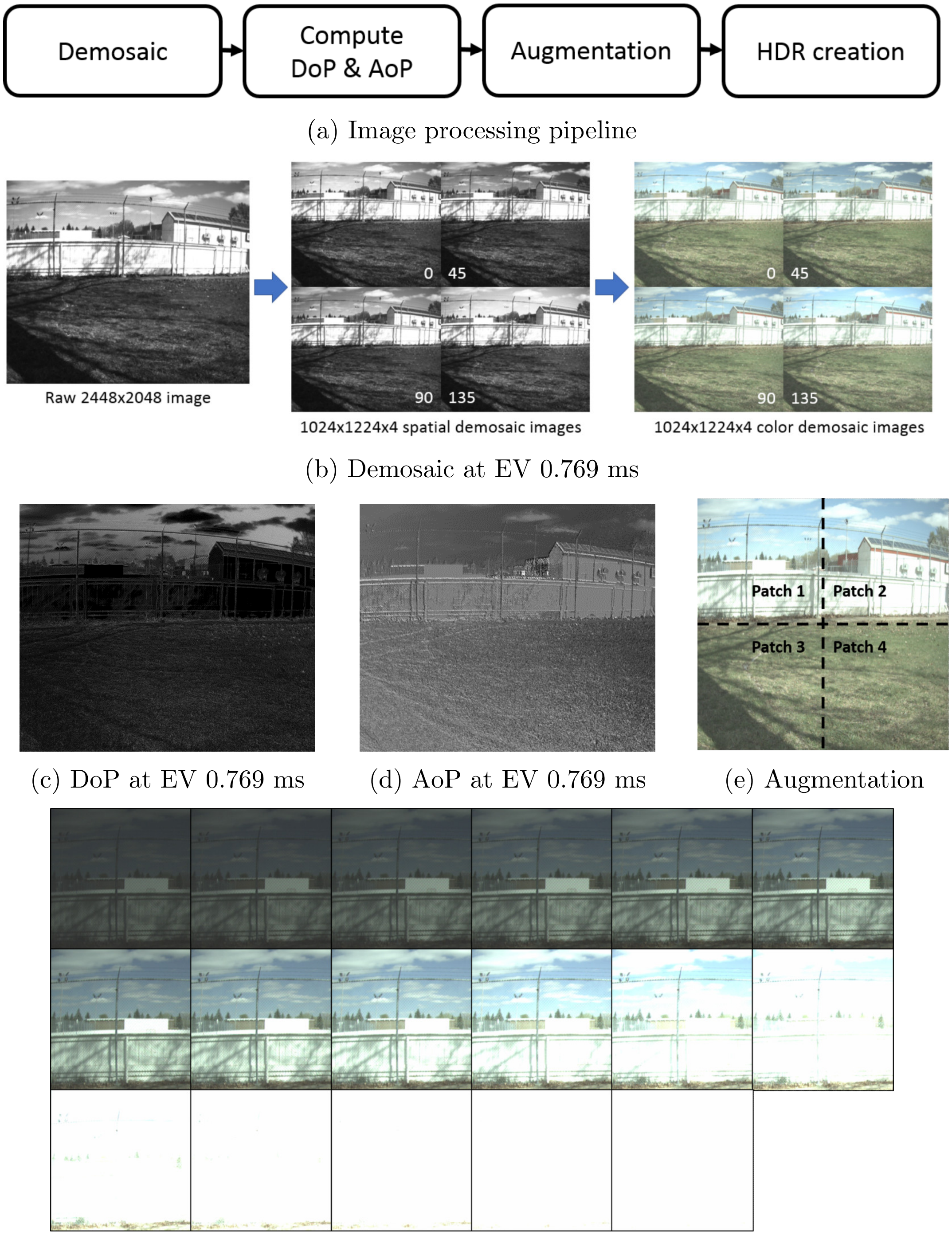
The lowest EV is selected at 0.03 ms, and each subsequent image is captured by increasing the exposure in multiples of 1.5. In total, each scene takes

Table 4.1: Polarized capture settings

Aperture	16	**FOV**	58°x49°x73°
Focal length	8	**Pixel format**	Polarized 8
Gain	0	**Pixel range**	0-255
Black level	2.22	**ADC bit depth**	12
Sensor size	2464x2056	**Frame rate**	73
Image size	2448x2048	**# Exposures**	17

about 1 second to capture, where the overhead is mainly due to saving the images into corresponding folders.

After scene capture, the raw data needs to be processed to reconstruct HDR images, and to extract polarization information. This image processing pipeline is shown in Figure 4.1a. The pipeline can be described as follows. First, the raw image needs to be demosaiced in order to extract the four colored polarization images (0°, 45°, 90°, and 135°). Since the physical layout of the polarized sensor is composed of an array of filters arranged as a 2x2 matrix, the raw 2448x2048 colored image is demosaiced in the spatial domain to extract four 1024x1224 polarization images. Then, an existing Bayer demosaicing method [55] is applied to each polarization image to interpolate the missing RGB values, and output colored polarization images. Figure 4.1b shows the process of applying demosaicing on a 0.769 ms EV image. Second, the DoP (ρ) and AoP (θ) images which characterize the polarization state of the light are computed from the four polarization images. The equations applied to compute ρ and θ have been described in Section 3.1. Figure 4.1c and Figure 4.1d show a DoP and AoP image computed from the 0.769 ms EV image, respectively. Third, the dataset is augmented by cropping four 512x512 patches from each polarization image, for improved readability. Figure 4.1e illustrates the data augmentation step. Additionally, Figure 4.1f shows the demosaiced and augmented multiple exposure images with the polarizer at 45°. Finally, ground truth HDR images are created by first applying the ICRF to linearize the relation between pixels and luminance. Next, we use the method in [56] to merge the bracketed images for each polarization channel (0°, 45°, 90°, and 135°), and then apply a pixel-weighting function to fuse the four



(a) Image processing pipeline

(b) Demosaic at EV 0.769 ms

(c) DoP at EV 0.769 ms

(d) AoP at EV 0.769 ms

(e) Augmentation

(f) Multiple-EV images at 45°

Figure 4.1: Image processing pipeline and results

images, which simultaneously map the pixels to the HDR domain [88]:

$$I_{gt}^{HDR} = \frac{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)\Big(g\big(L_i\big) + g\big(L_{i+2}\big)\Big)}{\sum_{i=1}^{2} W\Big(L_i + L_{i+2}\Big)t_0} \tag{4.2}$$

where W is the Gaussian weighted function ($\sigma = 0.2$), g is the ICRF, and I_{gt}^{HDR} is a RGB three-channel ground truth HDR image with 32 bit float per channel.

4.1.2 HDR distribution

To distribute I_{gt}^{HDR}, the OpenEXR method [30] is applied to store the HDR information. We used Python's OpenEXR library [67] to store the HDR content with 32 bit float, to obtain image files with .exr extensions that can be read with mainstream photo editing software.

4.1.3 Tone-mapping

To display the HDR content on an LDR display, we use the Durand [15] and Photomatix TMOs to map the scene-referred HDR tones to the display-referred LDR pixels. In our implementation, the Durand TMO parameters are Gamma = 1.0, contrast = 2.0, saturation = 0.5, sigma-space = 2.0, and sigma-color=2.0. The Photomatix TMOs selected are Enhanced and Detailed. Figure 4.2 illustrates some tone-mapped ground truth images for visualization.

4.1.4 Dataset statistics

We build two datasets, namely the EdPolCommunityOutdoor dataset and the UAPolCampusIndoor dataset. The datasets are collected using the Flir-BFS-U3-51S4p polarization camera with Sony CMOS sensor IMX250MYR (color).

The EdPolCommunityOutdoor dataset is collected outdoors during daytime under sunlight conditions. In total, it contains 50,048 LDR images and 736 HDR images. Specifically, 736 pairs of ground truth HDR and LDR images are available to train and test both DSMHDR and DSSHDR networks. The UAPolCampusIndoor dataset is collected indoors during the evening under night time conditions. In total, it contains 13,056 LDR images and 192

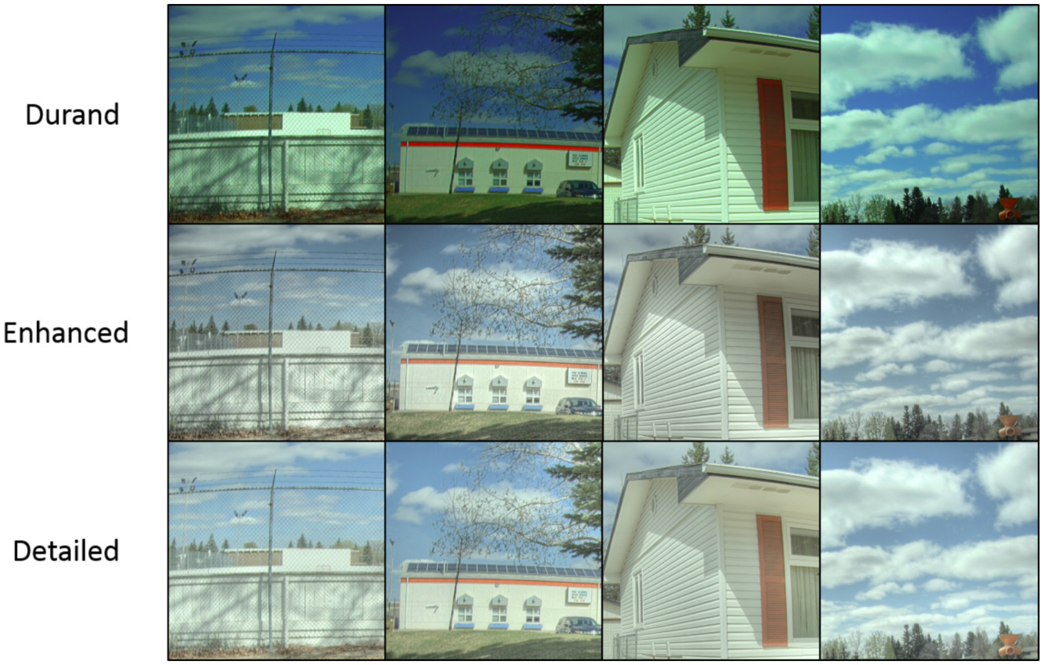
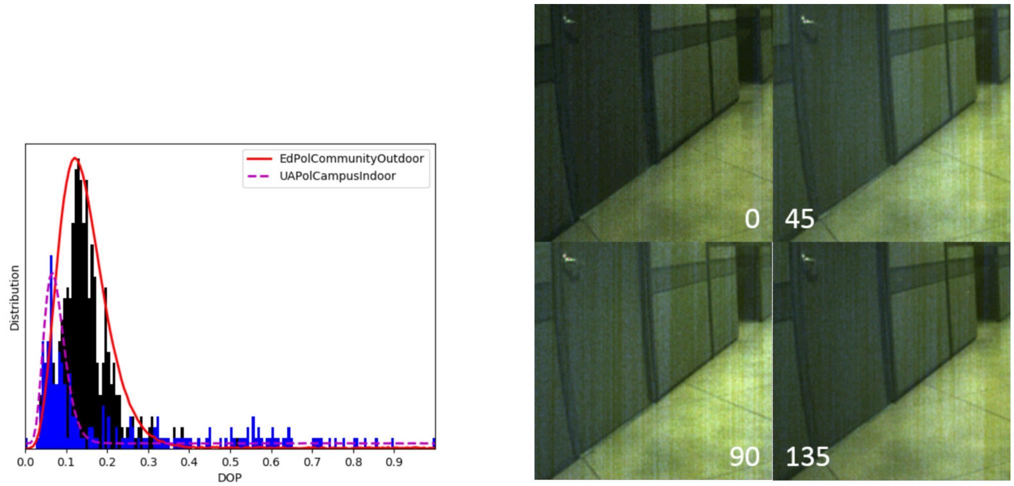


Figure 4.2: Tone-mapped ground truth images

HDR images. Specifically, 192 pairs of ground truth HDR and LDR images are also available to train and test deep-learning based approaches. The DoP (ρ) distributions and the fitted curves for the datasets are plotted in Figure 4.3a. A mixture model and the expectation maximization algorithm [59] is used to fit the DoP distributions. Shown in Table 4.2, the mixture model provided the largest negative log likelihood than the single model, which indicates a better model fit. We note that the mixture model fitted the EdPolCommunityOutdoor dataset better than the UAPolCampusIndoor dataset, and a more complex mixture model for the UAPolCampusIndoor dataset can improve the model fitting accuracy. The mixture model for the DoP distribution is given by:

$$f = w \cdot \gamma + (1 - w) \cdot U \tag{4.3}$$

where w is the weight of the distributions, γ is the Gamma distribution, and U is the uniform distribution. In our implementation for the EdPolCommunityOutdoor dataset: w = 0.934, γ_α = 6.264, γ_β = 0.023, U_{start} = 0.0 and U_{end} = 1.0. We note that Equation 4.3 is a general case that considers the presence of shot noise, as will be explained shortly. For the EdPolCommunityOutdoor



(a) DoP distribution

(b) UAPolCmapusIndoor image showing noise streaks in right regions of the image

Figure 4.3: Data statistics

Table 4.2: Negative log likelihood of different curve fitting model

Distribution	Model	neg-log likelihood
EdPolCommunityOutdoor	Gamma+Uniform	-548.895
UAPolCampusIndoor	Gamma+Uniform	-118.458

dataset, f in Equation 4.3 is predominantly weighted by the Gamma distribution since its weight w is 0.934, and thus the data can be explained well with only the Gamma distribution. However, in our implementation for the UAPolCampusIndoor dataset: $w = 0.491$, $\gamma_\alpha = 8.278$, $\gamma_\beta = 0.009$, $U_{start} = 0.0$ and $U_{end} = 1.0$. Due to the presence of shot noise that will be explained shortly, the general case represented by Equation 4.3 is more applicable for fitting the UAPolCampusIndoor data as it weights the Gamma distribution by $w = 0.491$ and the Uniform distribution by $(1 - w) = 0.509$.

Interestingly, the indoor poorly-lit scenes have a higher weight for the uniform distribution than the outdoor well-lit scenes. A closer inspection of the indoor images in the UAPolCampusIndoor dataset reveals there are visible noise streaks present, as shown in Figure 4.3b. This is the effect of shot noise as it is predominant when insufficient light reaches the sensor due to an exposure that is too brief for the lighting condition. The shot noise results in

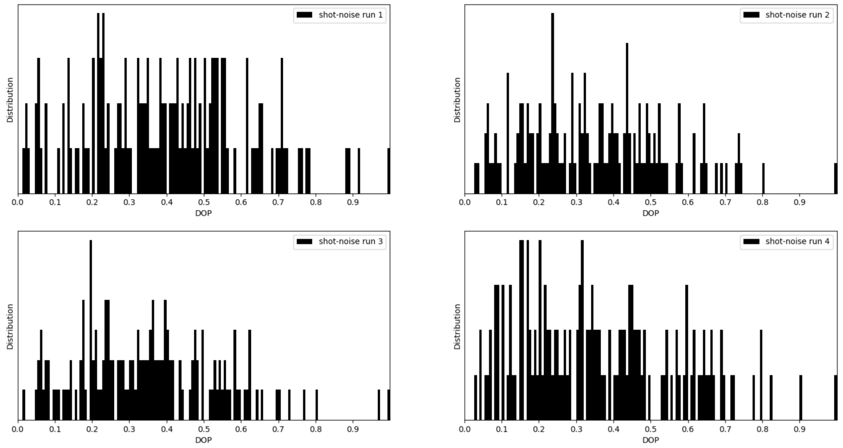


Figure 4.4: Study on effect of noise on DoP distribution over four runs

random pixel variations that contribute to a higher DoP, but do not indicate that the scene truly has a higher measure of polarization. In particular, we verified through an experiment that the shot noise can be modelled by an uniform distribution. In the experiment, the polarization images L_1, L_2, L_3, L_4 are generated by random numbers sampled from a uniform distribution over [0,1], and is used to compute the DoP using Equation 3.3. Figure 4.4 plots the DoP distribution over four runs where in each run, the DoP is computed from random generated pixel values to represent shot noise. Visually, we can observe that the resultant DoP distribution generated from shot noise is generally uniformly distributed. Finally, we take note of the observation that additional processing is required to account for shot noise when using the Flir-BFS-U3-51S4p polarization camera to capture poorly-lit scenes, such as the ones in our UAPolCampusIndoor dataset.

In our study, we focus on HDR reconstruction for outdoor scenes. We trained and evaluated our network using the EdPolCommunityOutdoor dataset.

4.2 Experiment setups

4.2.1 DSMHDR implementation details

The framework is implemented using Chainer, where the up- and down-exposure models are trained on NVIDIA GeForce GTX 1080 Ti, AMD Ryzen Threadripper 1950X 16-Core Processor 32 GB RAM. Our implementation is an extension of the existing code originally implemented with Chainer, and is publicly available for download[1].

4.2.2 DSSHDR implementation details

The framework is implemented using PyTorch and trained on NVIDIA GeForce GTX 1080 Ti, AMD Ryzen Threadripper 1950X 16-Core Processor 32 GB RAM. Our implementation is an extension of the existing code originally implemented with PyTorch, and is publicly available on Github[2].

4.2.3 Evaluation metrics

We evaluate the performance using different metrics. To evaluate the quality of the HDR images, we used the popular HDR-VDP2 [48] metric. We normalize the predicted and reference ground truth HDR images [49]. From the HDR-VDP2 metric, the Q score is reported, and the visibility probability map can be plotted. This map describes how likely it is for a difference to be noticed by the average observer, at each pixel. Warm values such as red in the map indicate high probability differences (undesired), while cold values such as blue indicate low probability differences (desired). Additionally, we evaluate the performance of the HDR images using the mean squared error (MSE) metric, as adapted from other works [43], [77]. Then to evaluate the accuracy of the HDR tone-mapped LDR images, we used the popular PSNR, SSIM [100] and FSIM [98] metrics.

[1] Link to source code: http://www.cgg.cs.tsukuba.ac.jp/ endo/projects/DrTMO

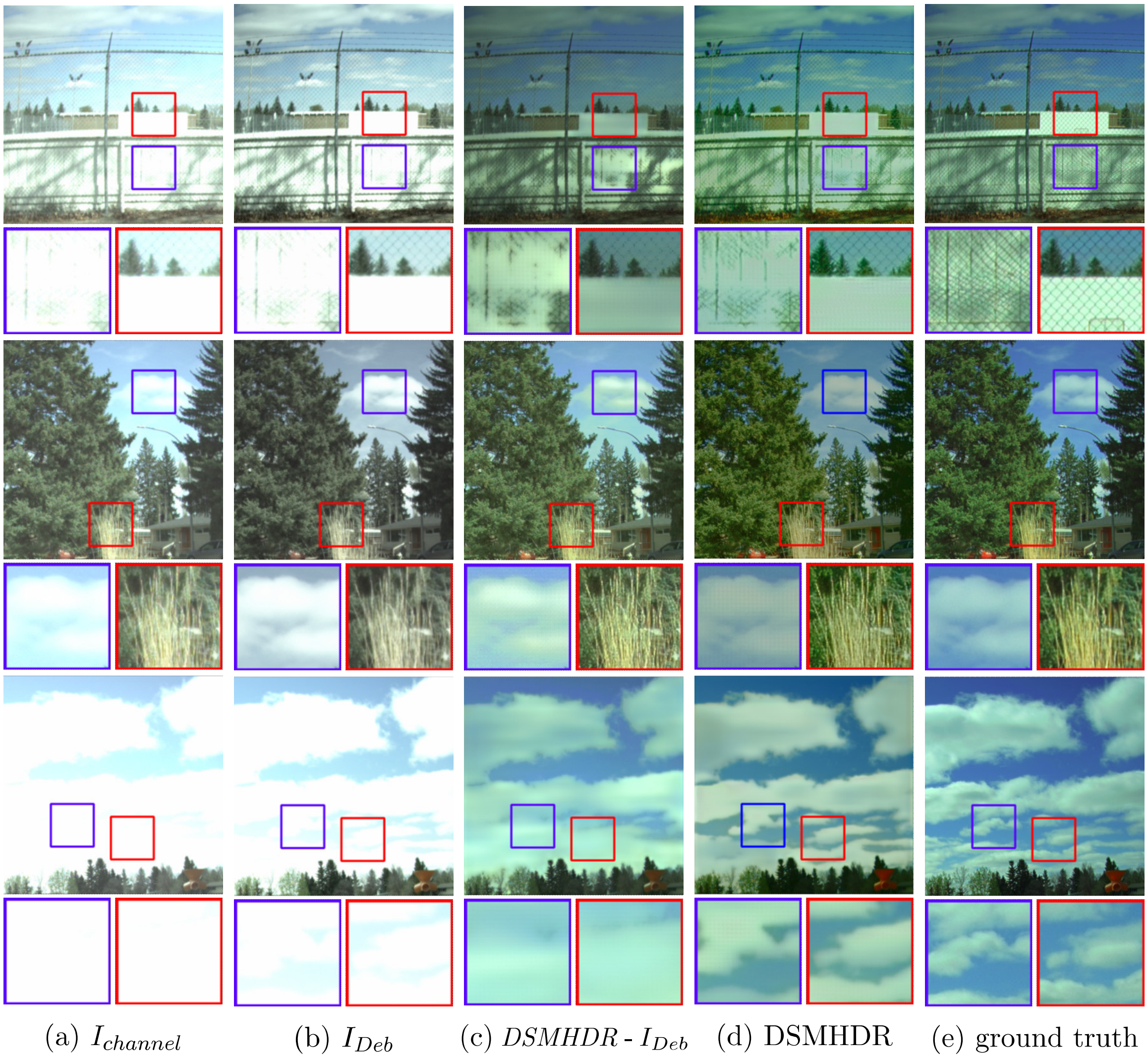
[2] Link to source code: https://github.com/marcelsan/Deep-HdrReconstruction

4.3 Ablation study

4.3.1 DSMHDR: study on framework

In this experiment, we validate the effect of the pre-processing step implemented by Equation 3.25, to fuse the four polarizer orientation LDR images before feeding it to the network. The ablation study is performed by considering three different cases. First, in the case of *DSMHDR* - I_{Deb}, we removed the pre-processing step, and trained the network to estimate an HDR image from an LDR image at a single polarizer orientation. Second, in the case of I_{Deb}, we removed the network and estimated an HDR image with only the pre-processing step. Third, in the case of $I_{channel}$, we removed the pre-processing step and the network, and showed a single polarizer orientation image.

Figure 4.5 illustrates the qualitative result for the ablation study. We can observe that the case of *DSMHDR* - I_{Deb} shown in Figure 4.5c reconstructs fewer details compared to DSMHDR, which is reconstructed with I_{Deb} as the input image (shown in Figure 4.5d). It is because *DSMHDR* - I_{Deb} uses only a single LDR image as input which conveys less information, and thus tends to neglect textures and local contrasts. On the other hand, with DSMHDR, it constructs an intermediate HDR image from the four LDR images with four different polarizer angles, which represent an information-rich input. These four images effectively correspond to four different exposure times on a per-pixel basis, and can thus better reveal scene details and color contrast. We can also observe that the case of I_{Deb} shown in Figure 4.5b remains saturated with limited details present compared to DSMHDR. Therefore, CNN prediction by training the network is required for a better HDR recovery. Next, we can observe that the case of $I_{channel}$ shown in Figure 4.5a reveals few details compared to DSMHDR, and has overall the worst performance. This is expected as $I_{channel}$ is simply a single LDR polarizer orientation image directly captured by the camera without any processing. Also, when comparing $I_{channel}$ and I_{Deb}, it reveals that I_{Deb} is a more informative and better input with more details. The quantitative results shown in Table 4.3 are aligned with our observations. It indicates that DSMHDR (last row) achieves the best results.



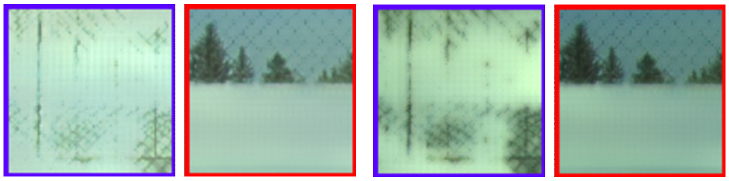
(a) $I_{channel}$ (b) I_{Deb} (c) *DSMHDR* - I_{Deb} (d) DSMHDR (e) ground truth

Figure 4.5: Qualitative results of DSMHDR and its variants. Example image from the test dataset where the image details are better restored from left to right. For comparison, the ground truth image is in the last column. The Durand TMO is used.

Table 4.3: Quantitative results of DSMHDR model variants.

	PSNR	SSIM	FSIM	HDR-VDP2
$I_{channel}$	11.76	0.59	0.87	44.85
I_{Deb} [88]	15.42	0.67	0.89	45.72
DSMHDR - I_{Deb}	20.61	0.80	0.90	49.56
DSMHDR	**25.35**	**0.91**	**0.94**	**55.26**

Additionally, the deconvolution interpolation operation produces clearer and smoother images.



(a) ℓ_1 loss (b) $\ell_1 + SSIM$ loss

Figure 4.6: Qualitative results of different loss functions, where our choice of $\ell_1 + SSIM$ loss for $DSMHDR$ - I_{Deb} restores more details.

Table 4.4: Quantitative results of different loss functions of $DSMHDR$ - I_{Deb}

	PSNR	SSIM	FSIM	HDR-VDP2
ℓ_1	17.64	0.78	0.91	48.48
$\boldsymbol{\ell_1}$ + SSIM	**20.61**	**0.80**	**0.90**	**49.56**

4.3.2 DSMHDR: study on loss function

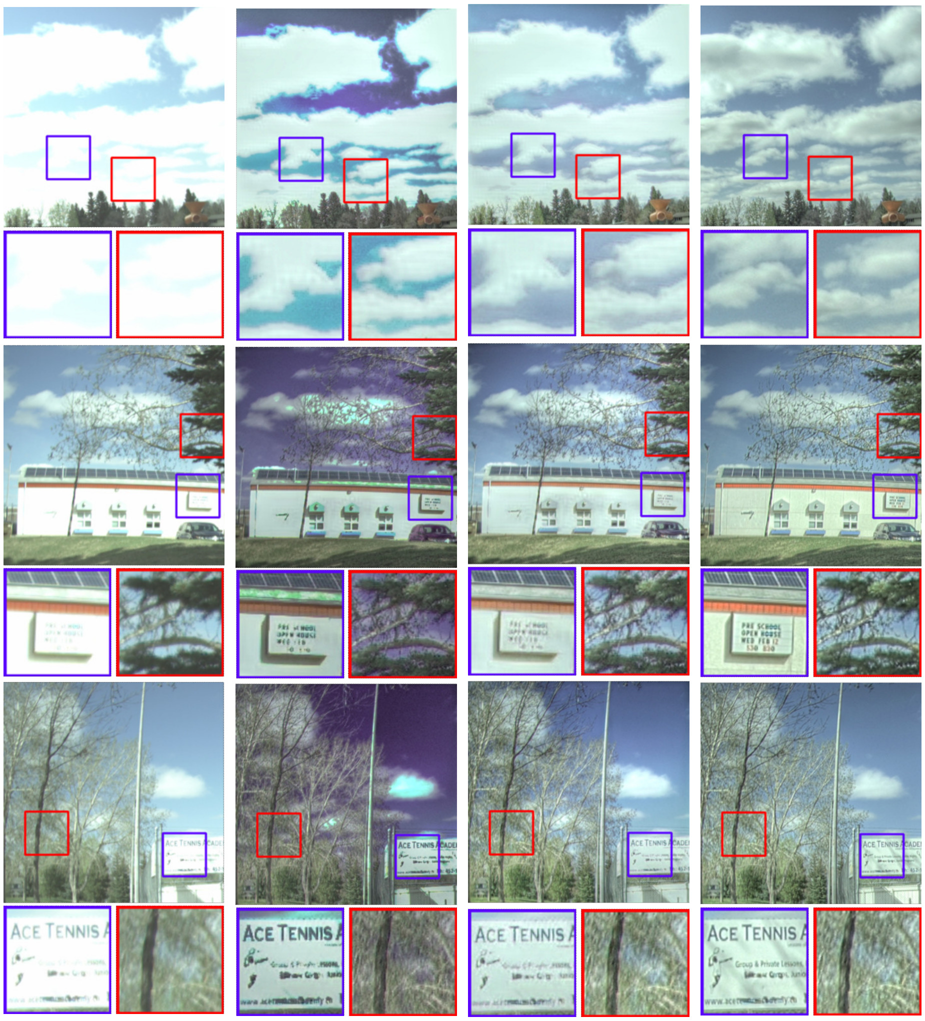
In this experiment, we compare the performances of our method with different loss functions. To solely analyze the effect of each loss function, we use the model $DSMHDR$ - I_{Deb} for analysis. Qualitative results are illustrated in Figure 4.6 where $\ell_1 + SSIM$ loss is better at preserving details. This is also reflected by the quantitative results shown in Table 4.4. Therefore, we train the DSMHDR model using $\ell_1 + SSIM$ loss.

4.3.3 DSSHDR: study on hyperparameter selection

In this experiment, we validate the model design choice for the HDR formulation, which is performed by comparing the following variants of the HDR formulation:

(a) **DSSHDR** (*i.e.*, $H = \alpha \cdot H_t + (1 - \alpha) \cdot H_d$). The full HDR formulation, which is the combination of traditional model based and deep-learning based HDR reconstruction results.

(b) **DSSHDR w/o $\boldsymbol{H_t}$** (*i.e.*, $H = H_d$). We remove the traditional model based result in this variant, and obtain the final HDR image H using our deep-learning based result to show the effect of polarimetric information in comparison with the regular images in traditional deep-learning



(a) L_c (b) DSSHDR w/o H_t (c) DSSHDR (d) ground truth

Figure 4.7: Qualitative results of DSSHDR hyperparameter selection. The DSSHDR model formulates HDR using a combination of the traditional model based method H_t, and the deep-learning based method H_d to estimate HDR in all areas. In contrast to L_c and DSSHDR w/o H_t, DSSHDR obtains results with better details that is more closely matched to the ground truth image. The Photomatix Enhanced tone-mapping operator is used.

methods.

The HDR formulation with (a) is effective for estimating HDR values. As shown in Figure 4.7, compared with the input L_c, (b) can restore more details as our CNN uses the polarimetric information as a prior. However, (b) suffers color distortions, and is an overall darker image with limited recovery of details in the poorly exposed regions. On the other hand, (a) can recover better

Table 4.5: Quantitative results of DSSHDR hyperparameter selection (the higher the better, except for MSE)

	PSNR	SSIM	FSIM	HDR-VDP2	HDR-MSE
DSSHDR w/o H_t	19.14	0.77	0.87	51.19	0.0341
DSSHDR	**22.66**	**0.89**	**0.94**	**56.16**	**0.0071**

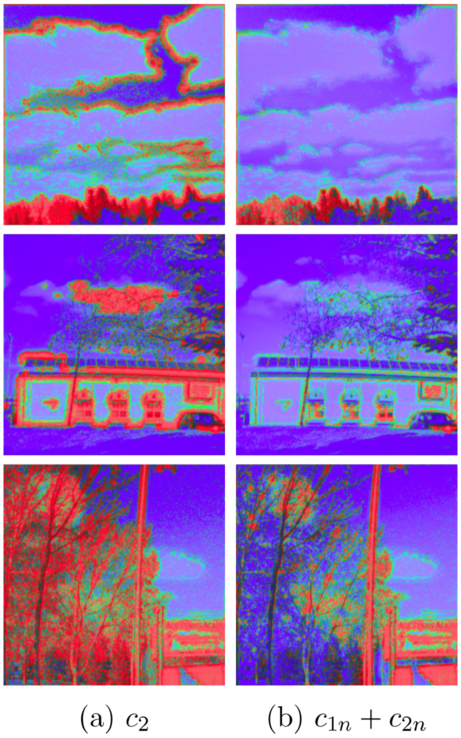

(a) c_2 (b) $c_{1n} + c_{2n}$

Figure 4.8: HDR-VDP2 probability maps of DSSHDR hyperparameter selection. Cold values such as blue indicate imperceptible differences to be noticed by the average observer (desired). Warm values such as red indicate perceptible differences to be noticed by the average observer (undesired). Overall, $c_{1n} + c_{2n}$ performs better than the other configurations with the most imperceptible difference (blue) show in the images.

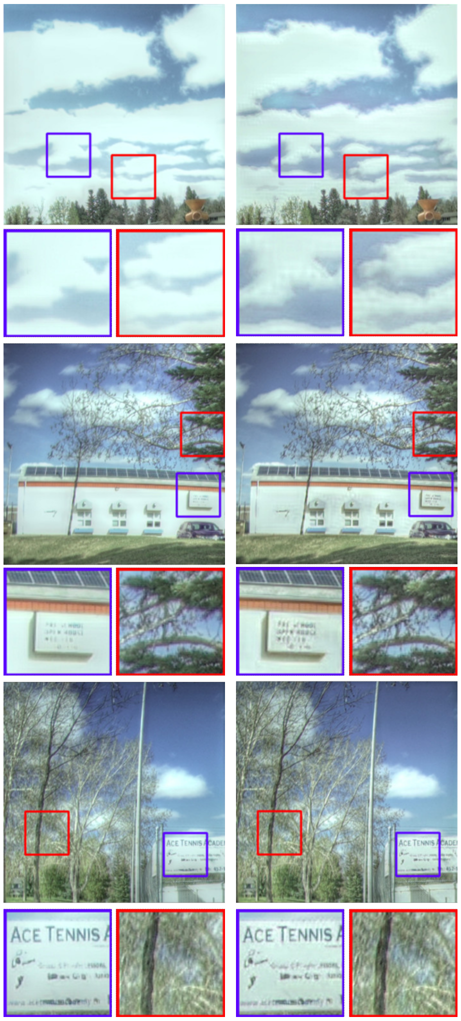
details in both under- and over-exposed regions, and alleviate color distortions and visible artifacts. An explanation for the removal of color distortions in (a) is that DoP is a mask shared among the colored channels, and thus can enforce consistency between images which helps to remove the colorization artifacts for a better HDR reconstruction. The quantitative results are shown in Table 4.5, and is aligned with the observation that (a) achieves a better performance. The key difference between these two variants is the effect of DoP. In the (a) variant, DoP is used as a strong accurate prior to recover HDR, where for regions with high DoP the traditional model based method recovers HDR details well compared to the deep-learning based method. Therefore, we achieve HDR formulation using the (a) variant. Additionally, we show the HDR-VDP2 probability maps for the different configurations in Figure 4.8.

4.3.4 DSSHDR: study on input mask

In this experiment, we evaluate the model design choice for the input mask. In particular, we compare the performances of our input mask K with a Gaussian input mask G defined by:

$$G = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{2\sigma})^2} \tag{4.4}$$

where $\sigma = 0.5$ and $\mu = 0.5$. We selected a Gaussian input mask for comparison because it can serve as a soft mask to identify pixels in the mid-range values as well-exposed, while identify pixels in both dark and bright ends as poorly-exposed and thus requires CNN prediction. The Gaussian mask is centered at $\mu = 0.5$ with a distribution of $\sigma = 0.5$. This is different than K which identifies all pixels below a threshold value as well-exposed, and only pixels in the bright end as poorly-exposed. Qualitative results are illustrated in Figure 4.9 where the results from using K reveals better textures and are more closely matched to the ground truth. On the other hand, using G results in a overall brighter image and reveals fewer textures. This is also reflected by the quantitative results shown in Table 4.6. One possible explanation for a limited texture recovery with the input mask G is that our dataset mainly consists of bright pixels due to the collection of images in outdoor scenes, and thus there is insufficient dark pixels in the dataset to train the under-exposed pixels. Additionally, there are significant noise present in the dark regions of an image that cannot be handled properly by the current setup. Thus, the network will try to predict based on incorrect pixel values, thus leading to poor results. Instead, the input mask K considers HDR recovery for only the bright regions, and predominantly address the recovery of the dark regions by the traditional model based method. Therefore, we train the DSSHDR model using K as our mask function, which we apply to the input image.



(a) G (b) K

Figure 4.9: Qualitative results of different input masks, where our choice of K mask for DSSHDR restores more details.

Table 4.6: Quantitative results of DSSHDR mask function selection (the higher the better, except for MSE)

	PSNR	SSIM	FSIM	HDR-VDP2	HDR-MSE
G	21.98	0.86	0.92	55.18	0.0115
$\boldsymbol{K}$	**22.66**	**0.89**	**0.94**	**56.16**	**0.0071**

4.4 Competing Algorithms

We evaluate our method against the following six popular state-of-the-art HDR reconstruction algorithms.

1. **ENet** [49]: This is a one-shot single output method that reconstructs an

HDR image from a single exposure LDR image using a novel three-branch CNN. The three branches of the network extract global, semi-local and local image features, respectively to recover missing details in the saturated regions. To reduce the effects of blocking artifacts that may arise from deconvolutions, and banding artifacts that may arise from nearest-neighbour upsampling, the architecture avoids the use of upsampling layers to reduce the aforementioned artifacts. Publicly available code[3] and author provided weights are used in the comparison study.

2. **HDRCNN** [16]: This is a one-shot single output method that maps the HDR image from a single exposure LDR image using a U-Net like architecture. The network predicts values for the over-saturated regions using a fixed mask, and later blends the prediction with the input LDR image for the unsaturated regions. Publicly available code[4] and author provided weights are used in the comparison study.

3. **HDRCNN-Mask** [77]: This is a one-shot single output method that recovers the HDR image from a single exposure LDR image with an autoencoder structure as well, but this approach predicts values in the saturated regions with an adaptable mask. The adaptable mask is generated by a feature masking mechanism that updates the mask at each convolutional layer to better recover the missing details in the saturated areas and to reduce visible artifacts. Publicly available code[5] and author provided weights are used in the comparison study. Additionally, fine-tuning of the models are also performed for the study.

4. **DRCP** [43]: This is a one-shot single output method that reconstructs an HDR image from a single exposure LDR image by incorporating domain knowledge about the image pipeline to train three specialized networks that reverses the image formation steps. Namely, a dequantization network to reduce the quantization artifacts in the input LDR image, a

[3] Link to source code: https://github.com/dmarnerides/hdr-expandnet

[4] Link to source code: https://github.com/gabrieleilertsen/hdrcnn

[5] Link to source code: https://github.com/marcelsan/Deep-HdrReconstruction

linearization network to estimate the CRF, and a hallucination network to reconstruct the saturated contents due to dynamic range clipping. Publicly available code[6] and author provided weights are used in the comparison study.

5. **DrTMO** [17]: This is a one-shot multiple output method that restores an HDR image from a single exposure LDR image using a novel encoder-decoder architecture. The encoder consists of 2D convolutions, and the decoder consists of 3D deconvolutions to generate consistent images with different exposures. The network synthesizes multiple LDR images with different exposures from a single exposure image, which are then post-processed to generate an HDR image using standard merging algorithms. Publicly available code[7] and author provided weights are used in the comparison study.

6. **PHDR** [88]: This is a one-shot single output method that directly maps to an HDR image from four LDR polarization images acquired at a single exposure value. To the best of our knowledge, this is the only existing work to reconstruct HDR images directly from polarization images. This is a non-learning approach that computes an HDR image heuristically based on the observation that images taken at different polarization angles are similar to images taken at different exposures.

4.5 Results

4.5.1 DSMHDR vs. State-of-the-art methods

In this experiment, we compare our DSMHDR results with the state-of-the-art algorithms described in Section 4.4. DSMHDR is a deep-learning based method that combines the LDR images to form an intermediate HDR for network input. Then, the network outputs multiple LDR images, each covering

[6] Link to source code: https://github.com/alex04072000/SingleHDR

[7] Link to source code: http://www.cgg.cs.tsukuba.ac.jp/ endo/projects/DrTMO

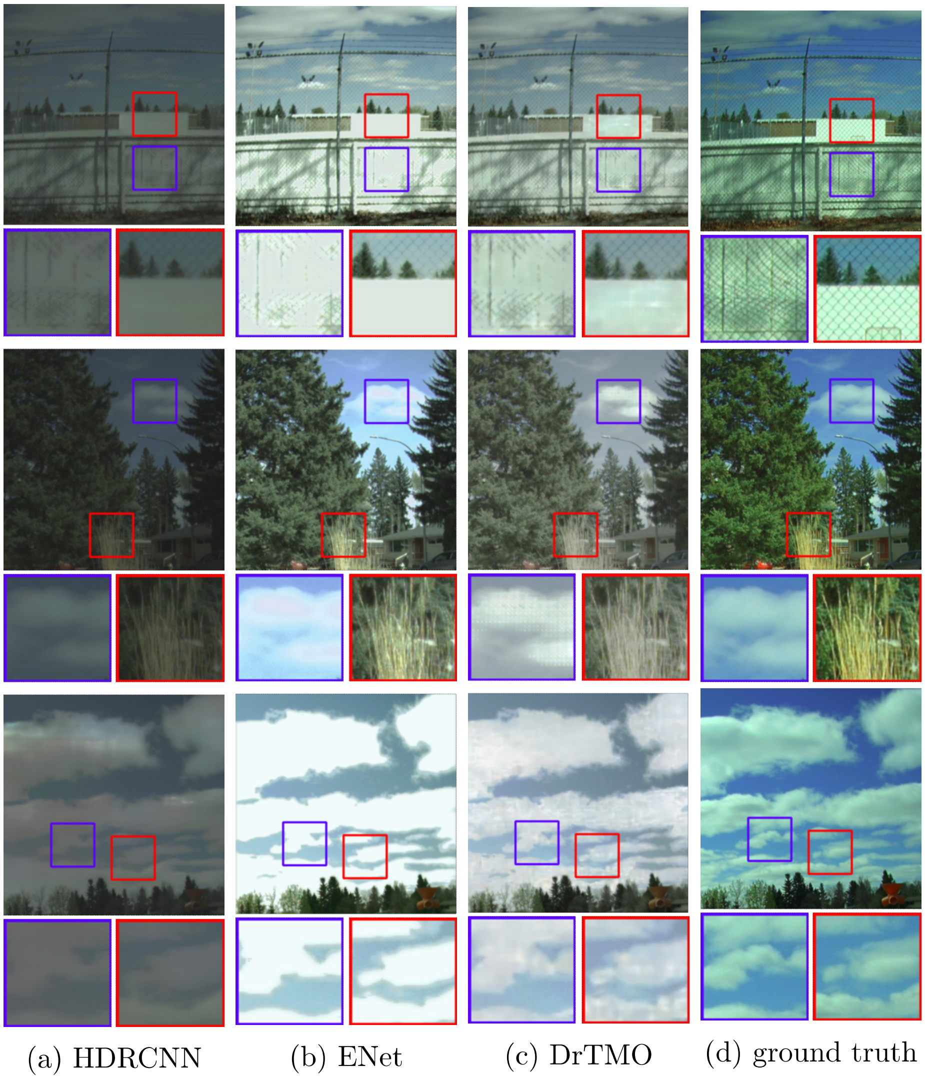


Figure 4.10: Qualitative comparative results with state-of-the-art methods. DSMHDR (ours) in Figure 4.11 recovers the most details. The Durand TMO is used.

a different luminance range, which are then merged to output an HDR image. Figure 4.10 and 4.11 illustrates the qualitative result for the comparative study. We can observe that HDRCNN [16] results tend to be dim, and the network is unable to restore details in the saturated regions. The ENet [49] generates overly-bright and smooth results, as it over-enhances the extracted illumination features. It also fails to recover details which reside in the over-exposed regions. The results of DrTMO [17] suffer from blocking artifacts and can not preserve details in the saturated areas. The DRCP [43] shares

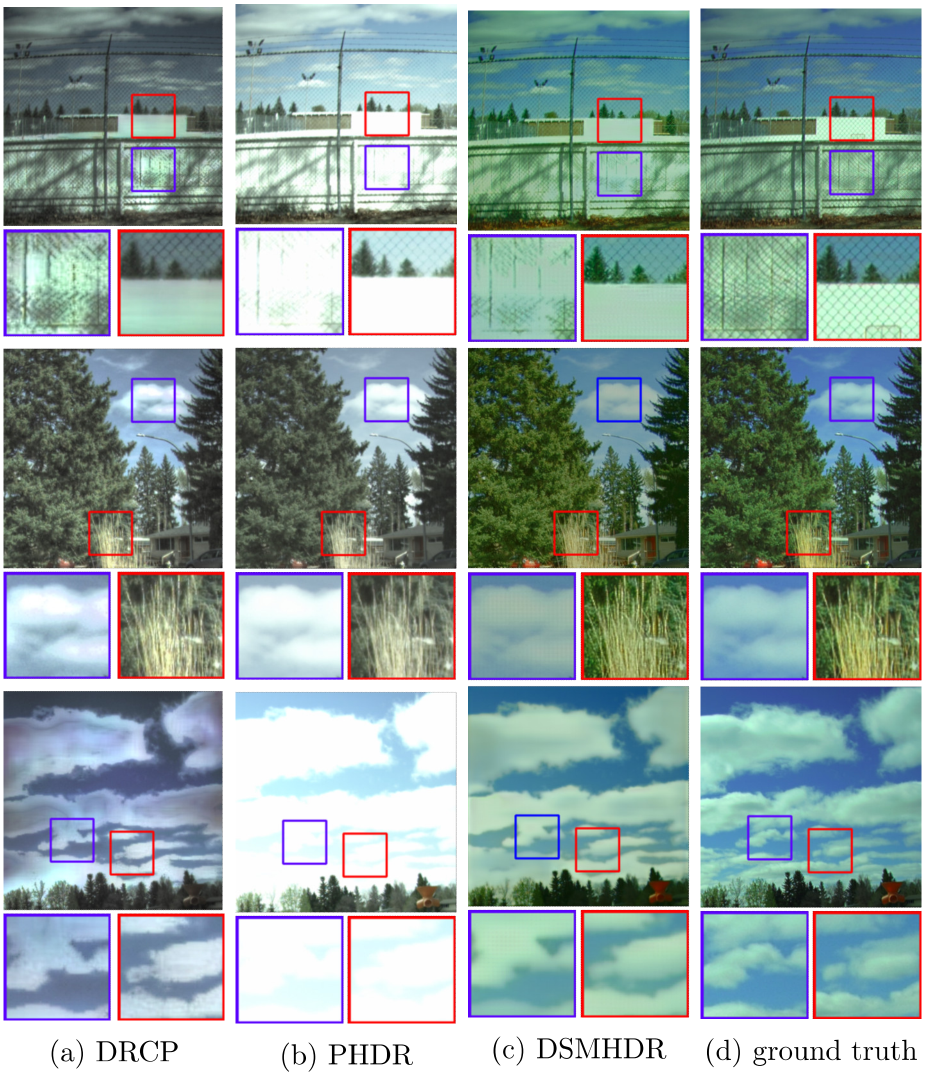


Figure 4.11: Qualitative comparative results with state-of-the-art methods. DSMHDR (ours) recovers the most details. The Durand TMO is used.

similar limitations and the results lack color consistency, as in some cases the generated colors are unnatural with artifacts. The PHDR [88] results tend to be bright, and the method cannot recover the information in the saturated regions. Since our method fuses four polarization images captured in a snapshot, where the images also correspond to those captured under four different exposure times, the given input is able to utilize information in the unsaturated pixels from one or more of the images to reveal details. As a result, the polarized input images collectively convey richer details compared to an image taken by a conventional camera. This fusion mechanism helps to reconstruct

Table 4.7: Quantitative comparative results with state-of-the-art methods. Underline indicates the best performing state-of-the-art.

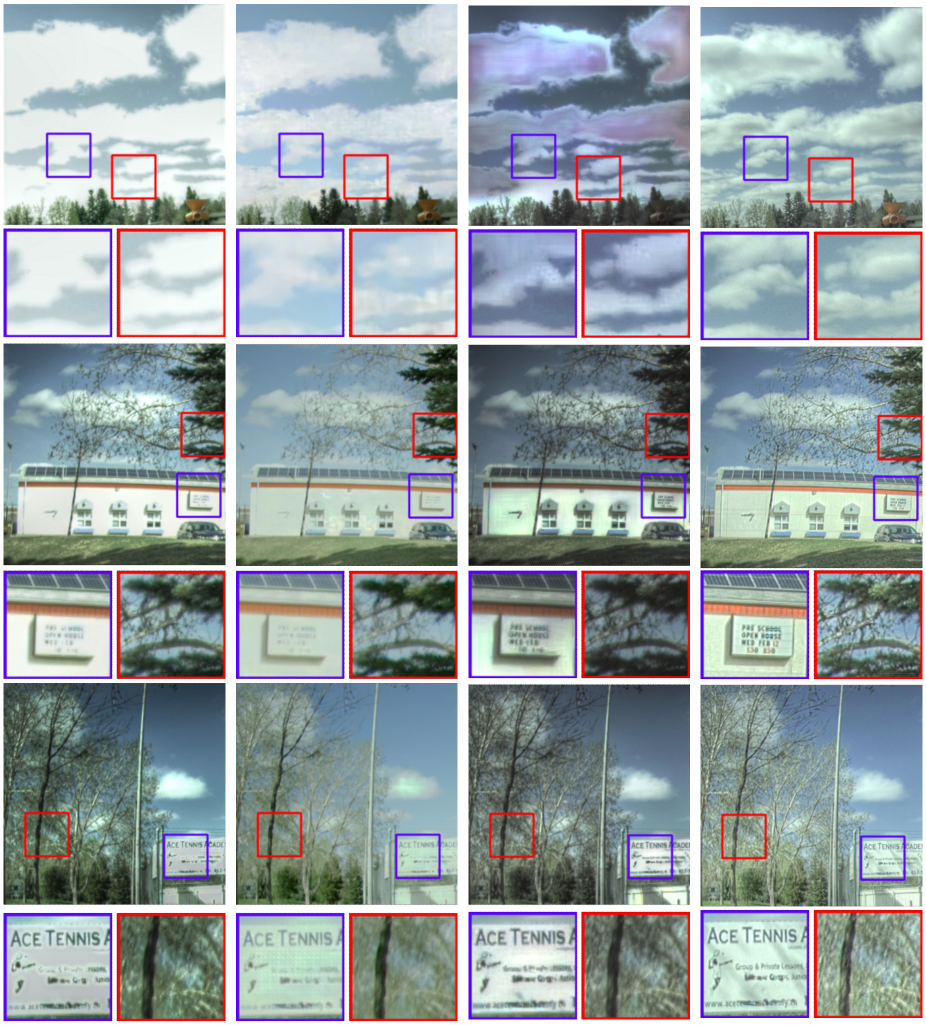
Methods	PSNR	SSIM	FSIM	HDR-VDP2
HDRCNN [16]	14.18	0.40	0.71	47.49
ENet [49]	15.37	0.67	0.88	47.24
DrTMO [17]	16.16	0.64	0.89	47.70
DRCP [43]	17.24	0.70	0.90	51.46
PHDR [88]	15.42	0.67	0.89	45.72
DSMHDR (ours)	**25.35**	**0.91**	**0.94**	**55.26**

details, and outputs visually pleasing textures.

In addition to visual evaluation, the quantitative results are summarized in Table 4.7. It shows that our method, in the last row, performs favorably compared to state-of-the-art methods under various evaluation metrics.

4.5.2 DSSHDR vs. State-of-the-art methods

In this experiment, we compare DSSHDR results with the state-of-the-art algorithms described in Section 4.4. The DSSHDR is a deep-learning based method that compute the mean of the LDR images and feed it to the network, then the network directly outputs a single HDR image. This is different than DSMHDR which provides a different input image, and outputs multiple LDR images that are then post-processed to output an HDR image. Figure 4.12, Figure 4.13 and Figure 4.14 illustrates the qualitative result for the comparative study. We can observe that images reconstructed with DSSHDR can overall recover better details in the saturated regions, and are free from color distortions and visible artifacts. The ENet [49] result tends to be overly-bright and smooth, as it over-enhances the extracted illumination features. Additionally, it's unable to improve HDR estimation for properly exposed regions. The DrTMO [17] method suffers from blocking artifacts, and has difficulty recovering details in all properly exposed and poorly exposed regions. The DRCP [43] method can restore details in the poorly exposed regions, but exhibits color distortions. Furthermore, the result is generally darker, thus contents lost in the under-exposed regions can not be restored. The PHDR [88] result

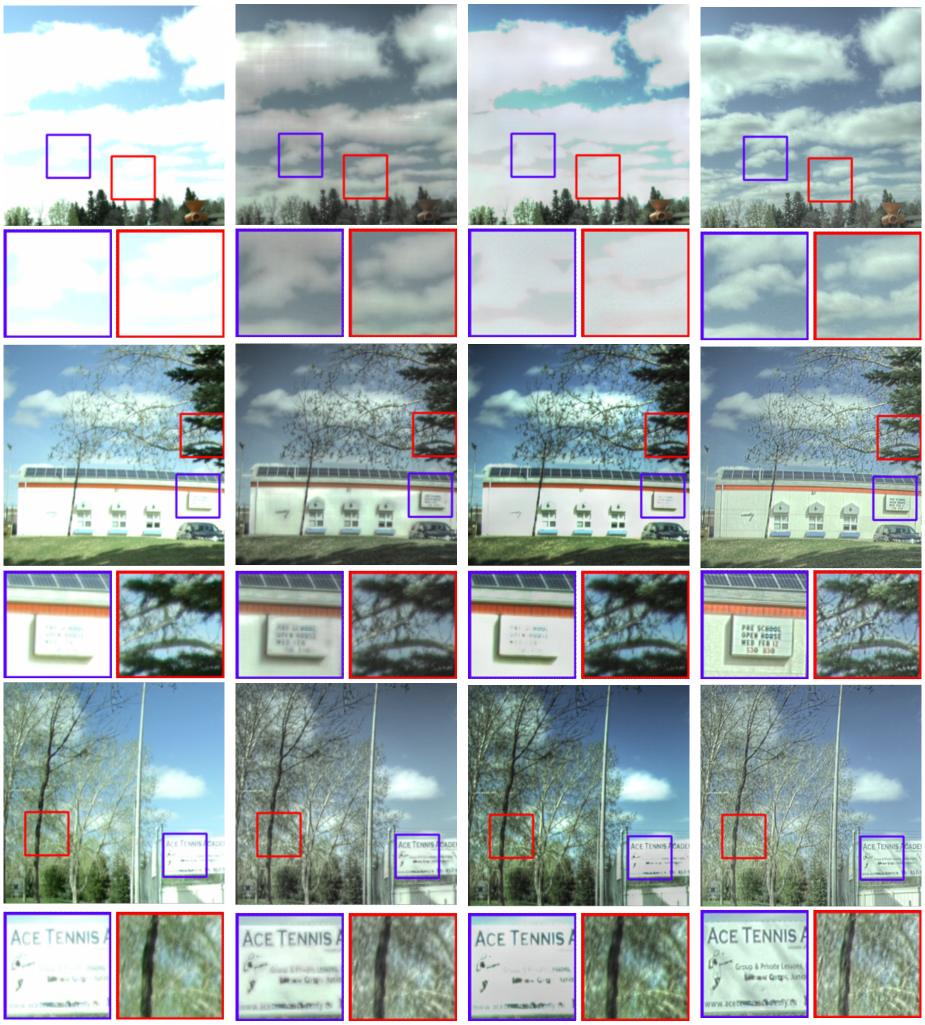


(a) ENet (b) DrTMO (c) DRCP (d) ground truth

Figure 4.12: Qualitative comparative results with state-of-the-art methods (also in Fig. 4.13 and 4.14). DSSHDR (ours) in Fig. 4.14 recovers the most details. The Photomatix Enhanced TMO is used.

is overall bright where pixels remain lost in the poorly exposed regions.

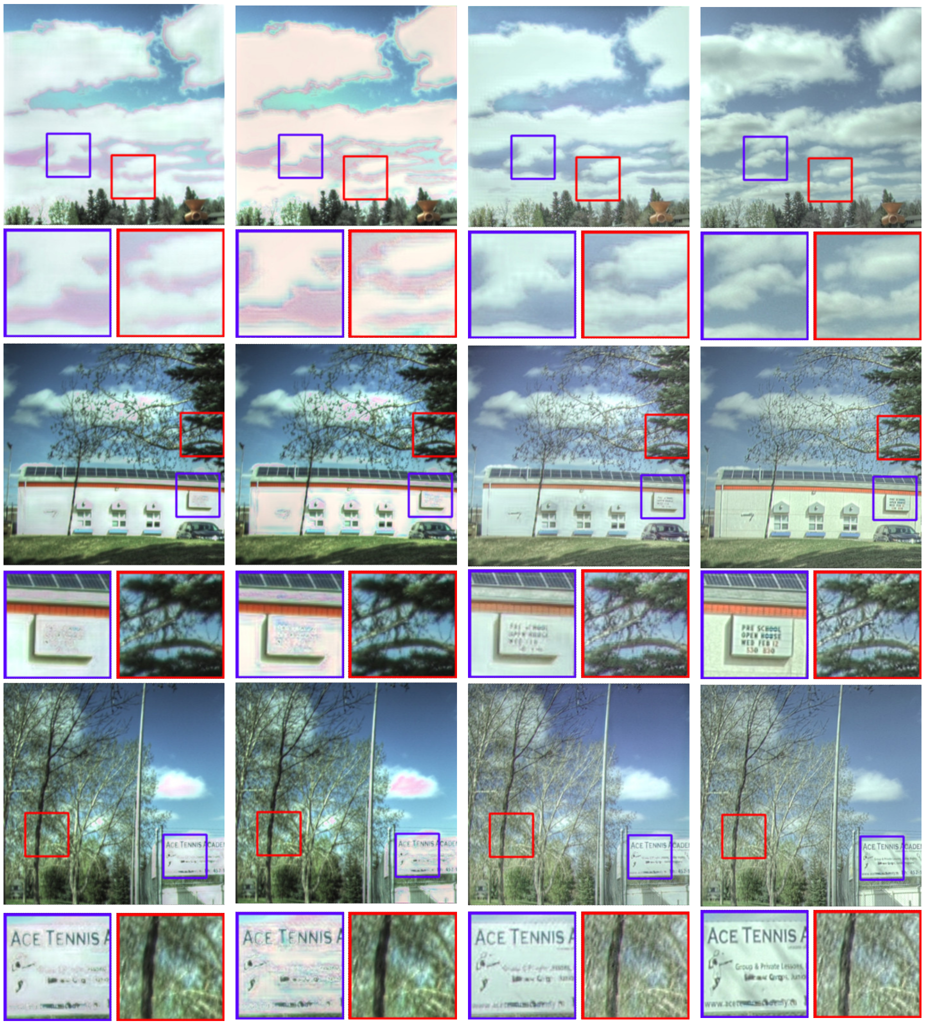
The HDRCNN [16] method can recover some contents in the over-exposed sky region, but the result is overall dim and presents visual artifacts. The HDRCNN-Mask [77] method follows the approach by HDRCNN to estimate details in over-exposed regions, and then reconstruct the final HDR image by combining with the input. However, unlike HDRCNN, HDRCNN-Mask proposes a feature masking mechanism to propagate valid features for properly exposed pixels. The HDRCNN-Mask-pretrain [77] uses the pretrain weights,



(a) PHDR (b) HDRCNN (c) HDRCNN-M^1 (d) ground truth

Figure 4.13: Qualitative comparative results with state-of-the-art methods (also in Fig. 4.12 and 4.14). DSSHDR (ours) in Fig. 4.14 recovers the most details. The Photomatix Enhanced TMO is used. HDRCNN-M^1 is HDRCNN-Mask-pretrain. The Photomatix Enhanced TMO is used.

provided by the author, to perform inference to generate the output image. The result is overly-smooth with color artifacts, and has difficulty reconstructing the boundaries of the cloud in the over-exposed sky area. Also, the content lost in the under-exposed tree areas can not be reconstructed. The HDRCNN-Mask-finetune [77] initializes the network with the pretrain weights, freezes the batch normalization parameters, and fine-tunes on our polarization dataset. The result is less smooth where the boundaries of the cloud in the over-exposed sky area can be better reconstructed. However, it also suffers from color ar-



(a) HDRCNN-M^2 (b) HDRCNN-M^3 (c) DSSHDR (d) ground truth

Figure 4.14: Qualitative comparative results with state-of-the-art methods (also in Fig. 4.12 and 4.13). DSSHDR (ours) recovers the most details. HDRCNN-M^2 is HDRCNN-Mask-finetune; HDRCNN-M^3 is HDRCNN-Mask-retrain. The Photomatix Enhanced TMO is used.

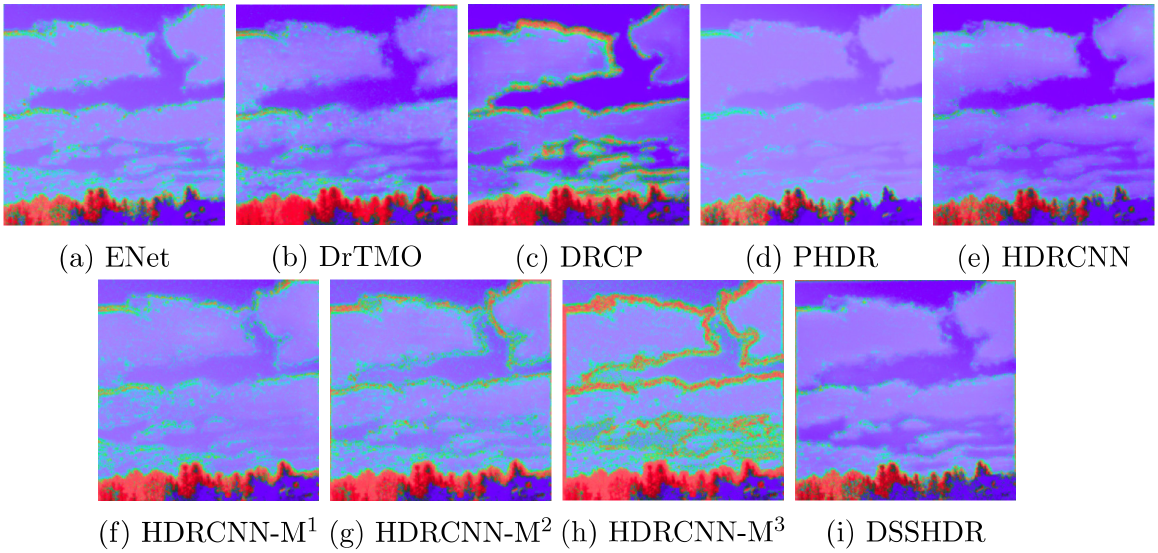
tifacts, and remains unable to reconstruct content lost in the under-exposed tree areas. The HDRCNN-Mask-retrain [77] trains the network from scratch on our polarization dataset. Similarly, the result exhibits color artifacts and difficulty in reconstructing details in the saturated regions of the image. On the other hand, the proposed DSSHDR method presents color consistency, free of visible artifacts, and overall able to recover richer textures in both under- and over-exposed regions. This is because the proposed DSSHDR method is designed to handle polarization images, as we have integrated the polariza-

Table 4.8: Quantitative comparative results with state-of-the-art methods (the higher the better, except for MSE). Underline indicates the best performing state-of-the-art.

Methods	PSNR	SSIM	FSIM	HDR-VDP2	HDR-MSE
ENet [49]	18.24	0.84	0.91	49.45	0.0429
DrTMO [17]	20.42	0.82	0.88	53.28	0.0591
DRCP [43]	17.99	0.80	0.92	49.25	0.0784
PHDR [88]	15.48	0.67	0.89	45.92	0.0639
HDRCNN [16]	22.32	0.86	0.92	53.66	0.0391
HDRCNN-Mask-pretrain [77]	16.92	0.79	0.90	47.72	0.0470
HDRCNN-Mask-finetune [77]	17.06	0.79	0.90	47.95	0.0449
HDRCNN-Mask-retrain [77]	17.03	0.71	0.89	47.70	0.0440
DSSHDR (ours)	**22.66**	**0.89**	**0.94**	**56.16**	**0.0071**

tion information (i.e., DoP) into the network during training and inference to formulate the final HDR image. In addition to visual evaluation, the quantitative results are summarized in Table 4.8. Since the proposed DSSHDR method uses DoP as a strong accurate prior to help HDR recovery for polarization images, it performs favorably compared to state-of-the-art methods under various evaluation metrics, including both HDR (HDR-VDP2, MSE) and HDR tone-mapped LDR evaluation metrics (PSNR, SSIM, FSIM).

Note that the results for the competing algorithms in Table 4.8 differ than those in Table 4.7. This is because of the input image used to evaluate the algorithms in the two tables are different. In Table 4.8, the input image to the algorithms is the mean of the four polarization filter images computed by Equation 3.13, as the study focus is on the effectiveness of the proposed feature masking mechanism. On the other hand, in Table 4.7, the input image to the algorithms is a single polarization filter image, as the study focus is on the effectiveness of the proposed pre-processing step to generate an informative input image. Furthermore, the input image used in Table 4.7 can be more closely matched to the input image used in DRCP [43], while the input image used in Table 4.8 can be more closely matched to the input image used in HDRCNN [16]. Thus, the difference in the best performing state-of-the-art. Additionally, we show the HDR-VDP2 probability maps for the different configurations in Figure 4.15. Results further confirms that our method performs better than



(a) ENet (b) DrTMO (c) DRCP (d) PHDR (e) HDRCNN

(f) HDRCNN-M^1 (g) HDRCNN-M^2 (h) HDRCNN-M^3 (i) DSSHDR

Figure 4.15: HDR-VDP2 probability maps of DSSHDR and the state-of-the-art methods. Cold values such as blue indicate imperceptible differences to be noticed by the average observer (desired). Warm values such as red indicate perceptible differences to be noticed by the average observer (undesired). Overall, DSSHDR (ours) performs better than the other configurations with the most imperceptible difference (blue) and details shown in the images. HDRCNN-M^1 is HDRCNN-Mask-pretrain; HDRCNN-M^2 is HDRCNN-Mask-finetune; HDRCNN-M^3 is HDRCNN-Mask-retrain. The Photomatix Enhanced TMO is used.

the competing algorithms with the most imperceptible difference (blue) and details illustrated in the image. We note that Figure 4.15d is mainly blue, but it reveals little boundaries of the clouds in the sky.

4.6 Summary

In this chapter, we introduced a polarization HDR dataset called EdPolCommunityOutdoor dataset, which we constructed and used to evaluate our two proposed methods: DSMHDR and DSSHDR. Then we presented the results of the ablation studies to quantify the effectiveness of the different components within our methods. In particular, for the DSMHDR method, we investigated the effect of the pre-processing step and the loss function terms. For the DSSHDR, we investigated the effect of the design choice for the HDR formulation equation, and for the input mask function. Finally, we presented the results of the comparative study against state-of-the-art algorithms, and showed that our method can outperform the competing algorithms.

In addition, the experiments performed correspond to testing our two important hypothesis: First, we showed that it is possible to perform HDR image reconstruction using the polarization camera. Second, given the polarization information from the DoP image, we are able to utilize a deep-learning based approach and showed that this information helps with HDR reconstruction.

Chapter 5

Conclusion

5.1 Overview

In this thesis, we have proposed two novel methods based on polarimetric information for HDR reconstruction. Our methods explores polarimetric cues obtained by a polarization camera to enhance the reconstruction in both under- and over-saturated regions, which is a long-standing challenge in computer vision and robotics applications. We first studied the feasibility of creating HDR images from a polarization camera that has on-chip multi-directional polarization filters. We exploited the fact that in environments with polarized lighting, the effect of the polarizers is analogous to that of imaging with multiple exposures. This gives rise to the possibility of reconstructing an HDR image from polarization images. In particular, we presented a radiometric model of the polarization camera, with which we can estimate the expected increase in dynamic range as a function of the polarimetric cue available from the polarization camera.

Then we proposed two deep-learning based methods to achieve deep snapshot HDR reconstruction. The first method, DSMHDR, leverages the prior knowledge that different polarization images are similar to different exposure images, which allows us to combine the multiple polarization images at the input, and feed it to the network for HDR reconstruction. Intuitively, this method transforms the input to the HDR space, to convey more information that can help with reconstructing HDR images. However, the first method only utilizes the polarization image information for reconstruction, while other polarimetric cues are also available from the polarization camera, such as the DoP. Therefore, in our second method, DSSHDR, we used the polarization images and the DoP image to reconstruct HDR images. In particular, we use the DoP image as a mask for the feature masking mechanism, to identify valid content to propagate through the CNN. Then we blend the predicted HDR image with the HDR image generated by a traditional model based approach to formulate the final HDR result.

Due to the lack of polarization images for HDR creation, we collected polarization images of outdoor scenes. We then processed the images to create the pairs of ground truth HDR and HDR tone-mapped LDR images. To our knowledge, this is the first polarization image dataset available for HDR reconstruction. Our experimental results on the polarization image dataset showed that our method demonstrated better quantitative and qualitative performances over state-of-the-art methods.

5.2 Limitation and Future Work

Despite the aforementioned success, currently the proposed methods are not sufficiently capable of handling scenes with extremely high dynamic range. This is mainly because such scenes are rare in our training dataset. In the future, we will augment our training dataset to incorporate such extreme cases to improve the performance. In addition, it will be interesting to explore the use of the AoP image, as an additional polarimetric cue, to further enhance the image dynamic range and HDR recovery.

References

[1] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do hdr displays support ldr content? a psychophysical evaluation," *ACM Trans. Graph.*, vol. 26, no. 3, 38–es, Jul. 2007, ISSN: 0730-0301. DOI: 10.1145/1276377.1276425. [Online]. Available: https://doi.org/10.1145/1276377.1276425.

[2] M. Alghamdi, Q. Fu, A. Thabet, and W. Heidrich, "Reconfigurable Snapshot HDR Imaging Using Coded Masks and Inception Network," in *Vision, Modeling and Visualization*, H.-J. Schulz, M. Teschner, and M. Wimmer, Eds., The Eurographics Association, 2019, ISBN: 978-3-03868-098-7. DOI: 10.2312/vmv.20191316.

[3] Y. Ba, A. R. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, *Deep shape from polarization*, 2020. arXiv: 1903.10210 [cs.CV].

[4] F. Bahri, M. Shakeri, and N. Ray, *Online illumination invariant moving object detection by generative neural network*, 2018. arXiv: 1808.01066 [cs.CV].

[5] L. Bao, Y. Song, Q. Yang, and N. Ahuja, "An edge-preserving filtering framework for visibility restoration," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 384–387.

[6] H. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images," 1978.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008, Similarity Matching in Computer Vision and Multimedia, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2007.09.014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314207001555.

[8] Y. Cao, Y. Ren, T. H. Li, and G. Li, "Over-exposure correction via exposure and scene information disentanglement," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, Nov. 2020.

[9] S.-Y. Chen and Y.-Y. Chuang, "Deep exposure fusion with deghosting via homography estimation and attention learning," 2020. arXiv: 2004.09089 [eess.IV].

[10] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, "Polarimetric multi-view stereo," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 369–378. DOI: 10.1109/CVPR.2017.47.

[11] I. Daly, M. How, J. Partridge, S. Temple, N. Marshall, T. Cronin, and N. Roberts, "Dynamic polarization vision in mantis shrimps," *Nature Communications*, vol. 7, p. 12 140, Jul. 2016. DOI: 10.1038/ncomms12140.

[12] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378, ISBN: 0897918967. DOI: 10.1145/258734.258884. [Online]. Available: https://doi.org/10.1145/258734.258884.

[13] P. Didyk, R. Mantiuk, M. Hein, and H.-P. Seidel, "Enhancement of bright video features for hdr displays," *Comput. Graph. Forum*, vol. 27, pp. 1265–1274, Jun. 2008. DOI: 10.1111/j.1467-8659.2008.01265.x.

[14] F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, *High Dynamic Range Video – From Acquisition to Display and Applications.* Apr. 2016.

[15] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, Jul. 2002, ISSN: 0730-0301. DOI: 10.1145/566654.566574. [Online]. Available: https://doi.org/10.1145/566654.566574.

[16] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–15, Nov. 2017, ISSN: 1557-7368. DOI: 10.1145/3130800.3130816. [Online]. Available: http://dx.doi.org/10.1145/3130800.3130816.

[17] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," 6, vol. 36, Nov. 2017.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[19] D. Goldstein, *Polarized light.* 2017.

[20] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualiso hdr reconstruction," *EURASIP Journal on Image and Video Processing*, vol. 2015, Dec. 2015. DOI: 10.1186/s13640-015-0095-0.

[21] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013. DOI: `10.1109/TPAMI.2012.213`.

[22] K. Hirakawa and P. M. Simon, "Single-shot high dynamic range imaging with conventional camera hardware," in *2011 International Conference on Computer Vision*, 2011, pp. 1339–1346. DOI: `10.1109/ICCV.2011.6126387`.

[23] P. Hiscocks and P. E. Syscomp, "Measuring luminance with a digital camera," 2011.

[24] G. Horváth and D. Varjú, "Underwater refraction-polarization patterns of skylight perceived by aquatic animals through snell's window of the flat water surface," *Vision Research*, vol. 35, no. 12, pp. 1651–1666, 1995, ISSN: 0042-6989. DOI: `https://doi.org/10.1016/0042-6989(94)00254-J`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/004269899400254J`.

[25] J. Hu, O. Gallo, K. Pulli, and X. Sun, "Hdr deghosting: How to deal with saturation?" In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1163–1170. DOI: `10.1109/CVPR.2013.154`.

[26] *Imx250myr (color) cmos sensor*, `https://https://www.flir.ca/products/blackfly-s-usb3/?model=BFS-U3-51S5PC-C`.

[27] *Imx250mzr (mono) cmos sensor*, `https://www.flir.ca/discover/iis/machine-vision/imaging-reflective-surfaces-sonys-first-polarized-sensor/`.

[28] *Imx250mzr (mono) cmos sensor*, `https://https://www.flir.ca/products/blackfly-s-usb3/?model=BFS-U3-51S5P-C`.

[29] H. Jang, K. Bang, J. Jang, and D. Hwang, "Inverse tone mapping operator using sequential deep neural networks based on the human visual system," *IEEE Access*, vol. 6, pp. 52 058–52 072, 2018. DOI: `10.1109/ACCESS.2018.2870295`.

[30] F. Kainz, R. Bogart, and P. Stanczyk, "Technical introduction to openexr.," in *Industrial light and magic*, 2009, p. 28.

[31] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, vol. 36, no. 4, 2017.

[32] J. H. Kim, S. Lee, and S.-J. Kang, "End-to-end differentiable learning to hdr image synthesis for multi-exposure images," 2020. arXiv: `2006.15833 [eess.IV]`.

[33] S. Y. Kim, J. Oh, and M. Kim, "Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications," 2019. arXiv: `1904.11176 [eess.IV]`.

[34] ——, "Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video," 2019. arXiv: 1909.04391 [eess.IV].

[35] Y. Kim, H. Jung, D. Min, and K. Sohn, *Deeply aggregated alternating minimization for image restoration*, 2016. arXiv: 1612.06508 [cs.CV].

[36] C. Lee, Y. Li, and V. Monga, "Ghost-free high dynamic range imaging via rank minimization," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1045–1049, 2014. DOI: 10.1109/LSP.2014.2323404.

[37] S. Lee, H. Chung, and N. I. Cho, "Exposure-structure blending network for high dynamic range imaging of dynamic scenes," *IEEE Access*, vol. 8, pp. 117 428–117 438, 2020. DOI: 10.1109/ACCESS.2020.3005022.

[38] S. Lee, S. Y. Jo, G. H. An, and S. Kang, "Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging," 2020, pp. 1–1. DOI: 10.1109/TMM.2020.3013378.

[39] S. Lee, G. An, and S.-J. Kang, "Deep recursive hdri: Inverse tone mapping using generative adversarial networks: 15th european conference, munich, germany, september 8-14, 2018, proceedings, part ii," Sep. 2018, pp. 613–628, ISBN: 978-3-030-01215-1. DOI: 10.1007/978-3-030-01216-8_37.

[40] S. Lee, G. H. An, and S.-J. Kang, "Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image," vol. 6, Institute of Electrical and Electronics Engineers (IEEE), 2018, pp. 49 913–49 924. DOI: 10.1109/access.2018.2868246. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2018.2868246.

[41] J. Li, X. Chen, D. Zou, B. Gao, and W. Teng, "Conformal and low-rank sparse representation for image restoration," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 235–243. DOI: 10.1109/ICCV.2015.35.

[42] Z. Li, J. Zheng, Z. Zhu, and S. Wu, "Selectively detail-enhanced fusion of differently exposed images with moving objects," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4372–4382, 2014. DOI: 10.1109/TIP.2014.2349432.

[43] Y. .-L. Liu, W. .-S. Lai, Y. .-S. Chen, Y. .-L. Kao, M. .-H. Yang, Y. .-Y. Chuang, and J. .-B. Huang, "Single-image hdr reconstruction by learning to reverse the camera pipeline," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1648–1657.

[44] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–, Nov. 2004. DOI: 10.1023/B:VISI.0000029664.99615.94.

[45] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," vol. 29, 2020, pp. 2808–2819. DOI: 10.1109/TIP.2019.2952716.

[46] B. C. Madden, "Extended intensity range imaging," 1993.

[47] S. Mann and R. W. Picard, "Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures," 1994.

[48] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011, ISSN: 0730-0301.

[49] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, "Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content," 2019. arXiv: 1803.02266 [cs.CV].

[50] B. Masia and D. Gutiérrez, "Content-aware reverse tone mapping," Jan. 2016. DOI: 10.2991/icaita-16.2016.58.

[51] ——, "Dynamic range expansion based on image statistics," *Multimedia Tools and Applications*, vol. 76, Jan. 2017. DOI: 10.1007/s11042-015-3036-0.

[52] S. McCarthy, "How independent are hdr, wcg, and hfr in human visual perception and the creative process?," Oct. 2015, pp. 1–18. DOI: 10.5594/M001630.

[53] M. Mcguire, W. Matusik, H. Pfister, B. Chen, J. Hughes, and S. Nayar, "Optical splitting trees for high-precision monocular imaging," *IEEE computer graphics and applications*, vol. 27, pp. 32–42, Apr. 2007. DOI: 10.1109/MCG.2007.45.

[54] X. Mei, H. Qi, B.-G. Hu, and S. Lyu, *Improving image restoration with soft-rounding*, 2015. arXiv: 1508.05046 [cs.CV].

[55] D. Menon, S. Andriani, and G. Calvagno, "Demosaicing with directional filtering and a posteriori decision," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 132–141, 2007. DOI: 10.1109/TIP.2006.884928.

[56] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," Oct. 2007, pp. 382–390, ISBN: 978-0-7695-3009-3. DOI: 10.1109/PG.2007.17.

[57] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," 2019. arXiv: 1908.00620 [eess.IV].

[58] T. Mitsunaga and S. K. Nayar, “Radiometric self calibration,” in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1, 1999, 374–380 Vol. 1. DOI: `10.1109/CVPR.1999.786966`.

[59] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996. DOI: `10.1109/79.543975`.

[60] K. Moriwaki, R. Yoshihashi, R. Kawakami, S. You, and T. Naemura, *Hybrid loss for learning single-image-based hdr reconstruction*, 2018. arXiv: `1812.07134 [cs.CV]`.

[61] S. K. Nayar and T. Mitsunaga, “High dynamic range imaging: Spatially varying pixel exposures,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 1, 2000, 472–479 vol.1. DOI: `10.1109/CVPR.2000.855857`.

[62] S. Ning, H. Xu, L. Song, R. Xie, and W. Zhang, *Learning an inverse tone mapping network with a generative adversarial regularizer*, 2018. arXiv: `1804.07677 [eess.IV]`.

[63] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. H. Lau, “Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions,” 2020. arXiv: `2007.01628 [eess.IV]`.

[64] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, 2016. [Online]. Available: `http://distill.pub/2016/deconv-checkerboard/`.

[65] T. Oh, J. Lee, Y. Tai, and I. S. Kweon, “Robust high dynamic range imaging by rank minimization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1219–1232, 2015. DOI: `10.1109/TPAMI.2014.2361338`.

[66] M. Oliveira and R. Kovaleski, “High-quality reverse tone mapping for a wide range of exposures,” Aug. 2014. DOI: `10.1109/SIBGRAPI.2014.29`.

[67] *Openexr documentation*, `https://excamera.com/files/OpenEXR.pdf`.

[68] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, *Context encoders: Feature learning by inpainting*, 2016. arXiv: `1604.07379 [cs.CV]`.

[69] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987, ISSN: 0734-189X.

[70] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, “Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs,” 2017. arXiv: 1712.07384 [cs.CV].

[71] Z. Pu, P. Guo, M. S. Asif, and Z. Ma, “Robust high dynamic range (hdr) imaging with complex motion and parallax,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, Nov. 2020.

[72] A. Rempel, M. Trentacoste, H. Seetzen, H. Young, W. Heidrich, L. Whitehead, and G. Ward, “Ldr2hdr: On-the-fly reverse tone mapping of legacy video and photographs,” *ACM Transactions on Graphics (TOG)*, vol. 26, p. 39, Aug. 2007. DOI: 10.1145/1275808.1276426.

[73] M. Rouf, R. Mantiuk, W. Heidrich, M. Trentacoste, and C. Lau, “Glare encoding of high dynamic range images,” in *CVPR 2011*, 2011, pp. 289–296. DOI: 10.1109/CVPR.2011.5995335.

[74] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” Nov. 2011, pp. 2564–2571. DOI: 10.1109/ICCV.2011.6126544.

[75] I. Safran and O. Shamir, *Depth-width tradeoffs in approximating natural functions with neural networks*, 2020. arXiv: 1610.09887 [cs.LG].

[76] B. Salahieh, Z. Chen, J. Rodriguez, and R. Liang, “Multi-polarization fringe projection imaging for high dynamic range objects,” *Optics express*, vol. 22, pp. 10 064–10 071, Apr. 2014. DOI: 10.1364/OE.22.010064.

[77] M. S. Santos, T. I. Ren, and N. K. Kalantari, “Single image hdr reconstruction using a cnn with masked features and perceptual loss,” *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020, ISSN: 0730-0301. DOI: 10.1145/3386569.3392403. [Online]. Available: https://doi.org/10.1145/3386569.3392403.

[78] M. Schöberl, A. Belz, J. Seiler, S. Foessel, and A. Kaup, “High dynamic range video by spatially non-regular optical filtering,” in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 2757–2760. DOI: 10.1109/ICIP.2012.6467470.

[79] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, “Robust Patch-Based HDR Reconstruction of Dynamic Scenes,” *ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH Asia 2012)*, vol. 31, no. 6, 203:1–203:11, 2012.

[80] M. Shakeri and H. Zhang, “Illumination invariant representation of natural images for visual place recognition,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 466–472. DOI: 10.1109/IROS.2016.7759095.

[81] T. Suda, M. Tanaka, Y. Monno, and M. Okutomi, *Deep snapshot hdr imaging using multi-exposure color filter array*, 2020. arXiv: 2011.10232 [cs.CV].

[82] Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, "Learning rank-1 diffractive optics for single-shot high dynamic range imaging," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[83] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile hdr video production system," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011, ISSN: 0730-0301. DOI: 10.1145/2010324.1964936. [Online]. Available: https://doi.org/10.1145/2010324.1964936.

[84] G. Ward, "A contrast-based scalefactor for luminance display," in *Graphics Gems IV*. USA: Academic Press Professional, Inc., 1994, pp. 415–421, ISBN: 0123361559.

[85] G. J. Ward, "The radiance lighting simulation and rendering system," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '94, New York, NY, USA: Association for Computing Machinery, 1994, pp. 459–472, ISBN: 0897916670. DOI: 10.1145/192161.192286. [Online]. Available: https://doi.org/10.1145/192161.192286.

[86] L. B. Wolff, "Polarization vision: A new sensory approach to image understanding," *Image and Vision Computing*, vol. 15, no. 2, pp. 81–93, 1997, ISSN: 0262-8856. DOI: https://doi.org/10.1016/S0262-8856(96)01123-7. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885696011237.

[87] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," 2018. arXiv: 1711.08937 [cs.CV].

[88] X. Wu, H. Zhang, X. Hu, M. Shakeri, C. Fan, and J. Ting, "Hdr reconstruction based on the polarization camera," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5113–5119, 2020.

[89] H. Xu, J. Ma, and X. Zhang, "Mef-gan: Multi-exposure image fusion via generative adversarial networks," vol. 29, 2020, pp. 7203–7216. DOI: 10.1109/TIP.2020.2999855.

[90] Y. Xu, S. Ning, R. Xie, and L. Song, "Gan based multi-exposure inverse tone mapping," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1–5. DOI: 10.1109/ICIP.2019.8803540.

[91] K. Yamada, T. Nakano, and S. Yamamoto, "Effectiveness of video camera dynamic range expansion for lane mark detection," in *Proceedings of Conference on Intelligent Transportation Systems*, 1997, pp. 584–588. DOI: 10.1109/ITSC.1997.660539.

[92] J. Yan, S. Lin, S. B. Kang, and X. Tang, “A learning-to-rank approach for image color enhancement,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2987–2994. DOI: 10.1109/CVPR.2014.382.

[93] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. Zhang, “Attention-guided network for ghost-free high dynamic range imaging,” 2019. arXiv: 1904.10293 [cs.CV].

[94] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. Lau, “Image correction via deep reciprocating hdr transformation,” Jun. 2018, pp. 1798–1807. DOI: 10.1109/CVPR.2018.00193.

[95] Young-Chang Chang and J. F. Reid, “Rgb calibration for color image analysis in machine vision,” *IEEE Transactions on Image Processing*, vol. 5, no. 10, pp. 1414–1422, 1996. DOI: 10.1109/83.536890.

[96] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, *Generative image inpainting with contextual attention*, 2018. arXiv: 1801.07892 [cs.CV].

[97] K. Zhang, W. Zuo, S. Gu, and L. Zhang, *Learning deep cnn denoiser prior for image restoration*, 2017. arXiv: 1704.03264 [cs.CV].

[98] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[99] W. Zhang and W. Cham, “Gradient-directed composition of multi-exposure images,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 530–536. DOI: 10.1109/CVPR.2010.5540168.

[100] 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*, vol. 13, no. 4, pp. 600–612, 2004.

Appendix A

A.1 Algorithm Hyperparameters

All architectures and hyper-parameters for our experiments are listed here (in case a hyperparameter isn't mentioned, its default value provided by the framework was used):

A.1.1 DSBDR

:

1. learningRate: 2e-4
2. numberOfIterations: 100
3. Optimizer: ADAM with $\beta_1 = 0.5$
4. batchSize: 1
5. weightInitialize: zero mean Gaussian noise with $\sigma = 0.02$

A.1.2 DSSHDR

:

1. learningRate: 2e-4
2. numberOfIterations: 150
3. Optimizer: ADAM with learning rate factor of 2.0
4. batchSize: 4
5. weightInitialize: Xavier

A.1.3 HDRCNN-Mask-finetune

:

1. learningRate: 2e-4
2. numberOfIterations: 60
3. Optimizer: ADAM with learning rate factor of 2.0
4. batchSize: 4
5. weightInitialize: HDRCNN-Mask weights with batch normalization parameters freeze

A.1.4 HDRCNN-Mask-retrain

:

1. learningRate: 2e-4
2. numberOfIterations: 100
3. Optimizer: ADAM with learning rate factor of 2.0
4. batchSize: 4
5. weightInitialize: Xavier