Polarization Guided HDR Reconstruction via Pixel-Wise Depolarization

Chu Zhou, Yufei Han, Minggui Teng, Jin Han, Si Li, Chao Xu, and Boxin Shi*, *Senior Member, IEEE*

***Abstract*—Taking photos with digital cameras often accompanies saturated pixels due to their limited dynamic range, and it is far too ill-posed to restore them. Capturing multiple low dynamic range images with bracketed exposures can make the problem less ill-posed, however, it is prone to ghosting artifacts caused by spatial misalignment among images. A polarization camera can capture four spatially-aligned and temporally-synchronized polarized images with different polarizer angles in a single shot, which can be used for ghost-free high dynamic range (HDR) reconstruction. However, real-world scenarios are still challenging since existing polarization-based HDR reconstruction methods treat all pixels in the same manner and only utilize the spatially-variant exposures of the polarized images (without fully exploiting the degree of polarization (DoP) and the angle of polarization (AoP) of the incoming light to the sensor, which encode abundant structural and contextual information of the scene) to handle the problem still in an ill-posed manner. In this paper, we propose a pixel-wise depolarization strategy to solve the polarization guided HDR reconstruction problem, by classifying the pixels based on their levels of ill-posedness in HDR reconstruction procedure and applying different solutions to different classes. To utilize the strategy with better generalization ability and higher robustness, we propose a network-physics-hybrid polarization-based HDR reconstruction pipeline along with a neural network tailored to it, fully exploiting the DoP and AoP. Experimental results show that our approach achieves state-of-the-art performance on both synthetic and real-world images.**

***Index Terms*—High dynamic range imaging, polarization, deep learning.**

I. Introduction

REAL-WORLD scenes usually have a much higher dynamic range than a digital camera (image) can record, so that the recorded images would be corrupted by disturbing artifacts such as saturation, leading to poor visual experience for human and degenerated performance for computer vision algorithms. To handle this problem, a kind of techniques termed as high dynamic range (HDR) reconstruction has been proposed, aiming to conquer the bottleneck of digital cameras for acquiring HDR images. A straightforward way would be directly restoring an HDR image from its low dynamic range (LDR) counterpart. However, recovering the truncated dynamic range from a single observation is highly ill-posed due to the lack of information in badly-exposed areas. By adopting numerical optimization [3], [30], [38] or deep neural networks [8], [9], [28], [40], priors handcrafted from natural image statistics or features extracted from training data could be used to achieve this goal. However, their generalization ability is limited since they rely strongly on the priors or features they could obtain.

For better generalization ability, several methods propose to merge multiple images (*e.g.*, multiple LDR images with bracketed exposures [6], or a burst of frames with constant exposure [14]) to reconstruct an HDR image in a less ill-posed manner. Despite their effectiveness has been shown in a large variety of scenes, these methods require multiple shots with specific exposures, which makes the capturing process inconvenient, leading to poor photographic experience. Besides, ghosting artifacts would occur when there is spatial misalignment among images caused by camera shake or object motion during the exposure time, leading to degenerated reconstruction results.

With the development of polarization-based vision, polarization cameras have been introduced to achieve ghost-free single-shot HDR reconstruction [45], [51]. A polarization camera (*e.g.*, Lucid Vision Phoenix polarization camera[1]) can capture four spatially-aligned and temporally-synchronized polarized LDR images with different polarizer angles ($0°$, $45°$, $90°$, and $135°$) in a single shot, by virtue of four-directional, on-chip micro-polarizers. Unlike conventional LDR images which only provide bracketed exposures for the whole image, the captured four polarized LDR images not only offer spatially-variant exposures[2], but also contain abundant structural and contextual information of the scene (encoded in the degree of polarization (DoP) and the angle of polarization (AoP) of the incoming light to the sensor, which can be computed from the polarized images). Such information can guide HDR reconstruction if properly used. Besides, polarizers can suppress specular highlights (which often have strong radiance values causing saturation) because the light reflected by specular surfaces is often significantly polarized [4], [49], while multi-image methods without using polarization need to capture additional images with low exposures to complete HDR

Chu Zhou and Chao Xu are with the Key Laboratory of Machine Perception, School of Intelligence Science and Technology, Peking University, Beijing 100080, China (e-mail: zhou_chu@pku.edu.cn; xuchao@cis.pku.edu.cn).

Minggui Teng and Boxin Shi are with the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100080, China. (e-mail: minggui_teng@pku.edu.cn; shiboxin@pku.edu.cn).

Yufei Han and Si Li are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: hanyufei@bupt.edu.cn; lisi@bupt.edu.cn).

Jin Han is with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: jinhan@nii.ac.jp).

* Corresponding author.

[1]https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/

[2]A polarizer attenuates scene radiance, and the extent of attenuation varies not only with the polarizer angles, but also with the polarization properties of light (depending on the properties of scene objects (*e.g.*, surface normal and texture), which are often spatially-variant [10], [26], [57]).

reconstruction. However, important properties of polarized LDR image formation model have not been considered by existing polarization-based HDR reconstruction methods [45], [51]:

- Different pixels are treated in the same manner without considering their variant levels of ill-posedness.
- Depolarization is directly conducted on captured images by assuming all pixels are unsaturated and unquantized (optionally a post-processing approach is applied to refine the inaccurate output).
- Spatially-variant exposures of the polarized images are considered, but structural and contextual information of the scene (encoded in the DoP and AoP of the incoming light to the sensor) is ignored.

The above issues motivate us to carefully revisit the formation model of polarized LDR images. To enable the polarization-based HDR reconstruction methods to handle images captured in the wild with better generalization ability and higher robustness, this paper proposes a pixel-wise depolarization strategy, by classifying the pixels based on their levels of ill-posedness in HDR reconstruction procedure and applying different solutions to different classes. To handle both ill-posed and well-posed computations in pixel-wise depolarization, we propose a network-physics-hybrid polarization-based HDR (Pol-HDR) reconstruction pipeline by complementarily adopting network modules and physics modules, taking the pixel saturation and quantization into account. Tailored to our pipeline, we design a neural network to make full use of the structural and contextual information of the scene encoded in the DoP and AoP of the incoming light to the sensor to achieve high-quality HDR reconstruction. To summarize, this paper makes contributions by demonstrating:

- A pixel-wise depolarization strategy for HDR.
- A network-physics-hybrid Pol-HDR pipeline.
- A neural network fully exploiting the DoP and AoP.

Experimental results show our method achieves state-of-the-art performance over existing polarization-based and single-image solutions on both synthetic and real-world images.

The rest of this paper is organized as follows: Section II briefly summaries the related works; Section III demonstrates the polarized LDR image formation model, our pixel-wise depolarization strategy, and our Pol-HDR pipeline; Section IV introduces the architecture of our neural network; Section V presents our synthetic dataset generation pipeline and implementation details; Section VI shows the experimental results; The whole paper is concluded in Section VII.

II. Related work

Generally, HDR reconstruction methods could be divided into three categories according to their properties: multi-image methods, single-image methods, and unconventional camera-based methods. We will have an overview on them respectively in the following.

A. Multi-image HDR reconstruction

Debevec and Malik proposed the classic multi-image HDR reconstruction method [6] by merging multiple LDR images captured with bracketed exposures. However, it is prone to ghosting artifacts when there is spatial misalignment among images caused by camera shake or object motion during the exposure time. This problem provokes a series of studies on ghosting removal. Khan *et al*. [21] supposed that the HDR image could be directly represented as the weighted sum of those bracketed exposed LDR images, and proposed to compute the weights iteratively to determine the contribution of each pixel to the final image. Sen *et al*. [42] proposed a patch-based energy minimization approach to jointly optimize the image alignment and HDR reconstruction processes. Oh *et al*. [35] formulated the HDR reconstruction problem into a rank minimization problem where misalignment errors could be considered as sparse outliers. Recently, neural networks have shown great potential for boosting the performance of HDR deghosting [7], [20], [34], [36], [37], [50], [52], [53].

Although these bracketed exposure-based approaches are successful in reconstructing plausible HDR contents in a large variety of scenes, their applicability is still limited since they require multiple shots with specific exposures, which makes the capturing process inconvenient, leading to poor photographic experience. To reduce the difficulty of choosing exposures, Hasinoff *et al*. [14] proposed to fuse a burst of frames with constant exposure to rebuild an HDR image. However, it still requires multiple shots so that deghosting is still needed.

B. Single-image HDR reconstruction

Single-image HDR reconstruction, also known as inverse tone-mapping [3], aims to reconstruct the HDR image from a single LDR image in a post-processing manner. Since it only requires a single shot, it is immune to ghosting artifacts. However, this problem is far more ill-posed than its multi-image counterpart due to the lack of information in badly-exposed areas. Several methods attempted to solve this challenging problem by adopting numerical optimization [3], [30], [38] based on handcrafted priors from natural image statistics. Recently, neural networks have also been introduced to solve it by directly hallucinating HDR contents based on extracted features from an amount of training data. Eilertsen *et al*. [8] labeled the saturated areas in LDR images with masks and used a network to restore them. Endo *et al*. [9], Lee *et al*. [27], and Kim *et al*. [22] predicted the LDR images with bracketed exposures from a single LDR image in a data-driven manner and merged them. Yang *et al*. [54] proposed a learning-based approach to perform HDR reconstruction and tone-mapping in an end-to-end manner. Marnerides *et al*. [29] concatenated and fused different levels of features extracted by a network to recover HDR details. Liu *et al*. [28] modeled the HDR-to-LDR image formation pipeline and adopted several network modules to simulate the reverse pipeline for mapping LDR images back to their HDR counterparts. Santos *et al*. [40] proposed to use feature masking mechanism along with the perceptual loss to improve the HDR reconstruction quality. Zheng *et al*. [56] proposed a dual-path network by collaboratively learning textural and chromatic features in the bilateral space to increase the process speed.

However, the generalization ability of these post-processing-based methods is limited, because image priors are not always

observed in the input and the image features extracted from synthetic training data often have a large domain gap with real-world ones. To reduce the dependence on priors or features, Metzler *et al*. [31] proposed a single-shot HDR imaging system by jointly optimizing a diffractive optical element-based encoder and a network-based decoder. Nevertheless, it requires a specially-fabricated optical element to build the system, which increases the difficulty of implementation.

C. Unconventional camera-based HDR reconstruction

With the development of unconventional cameras, single-shot ghost-free HDR reconstruction with high performance becomes possible. Unlike conventional cameras which only capture a single LDR image in a single shot, unconventional ones can obtain additional information of the scene to alleviate the ill-posedness. Some methods designed concept cameras with heterogeneous imaging pipelines (*e.g*., by recording the gradients [46] or modulo images [55], [58] that will never get saturated) to conquer the bottleneck of digital cameras for acquiring HDR images. Recently, neuromorphic cameras (*e.g*., event cameras [5] and spike cameras [59]) have also shown effectiveness in guiding HDR reconstruction [11], [12], [48] by virtue of the HDR observations encoded in captured data.

Another practical solution is coded exposure, which aims to modify the pixel architecture of conventional cameras with optical masks for spatially-variant exposures. Some methods focus on the design of optical masks. Nayar *et al*. [33] placed an optical mask with a spatially-variant regular attenuation pattern adjacent to a conventional image detector array. Hirakawa *et al*. [18] adopted a combination of photographic filter placed over the lens and the color filter array on image sensor to induce differences in red, green, and blue channel sensitivities. Schöberl *et al*. [41] proposed a nonregular arrangement of the attenuation pattern to alleviate the aliasing problems of regular sampling patterns adopted by Nayar *et al*. [33]. Sony proposed a per-pixel exposure camera [19] by setting different exposure times for two group of pixels. Fujifilm, on the other hand, proposed Super CCD [24] that has paired pixels with different effective pixel areas. Some methods focus on the performance of reconstruction algorithms. Aguerrebere *et al*. [1] proposed to use piecewise linear estimators to reconstruct the irradiance information of a scene from a single shot acquired with spatially-variant pixel exposures following a random pattern. Serrano *et al*. [43] introduced convolutional sparse coding (CSC) to recover high-quality HDR images from a single, coded exposure. Alghamdi *et al*. [2] not only proposed to place the mask at a small standoff distance in front of the sensor to make the camera reconfigurable, but also proposed a learning-based method to perform reconstruction.

With the development of polarization-based vision, polarization cameras have also been introduced to achieve ghost-free single-shot HDR reconstruction. Wu *et al*. [51] proposed to conduct depolarization on the captured four spatially-aligned and temporally-synchronized polarized LDR images with different polarizer angles (0°, 45°, 90°, and 135°) by a polarization camera in a single shot. Ting *et al*. [45] adopted a single-image HDR reconstruction network [9] to refine the result of Wu *et al*. [51]. However, these polarization-based methods simply treat the polarized images as LDR images with different exposures, while ignoring to make full use of the structural and contextual information of the scene encoded in the DoP and AoP of the incoming light to the sensor.

III. Method

In this section, we introduce the formation model of polarized LDR images captured by a polarization camera in Section III-A, our pixel-wise depolarization strategy in Section III-B, and our network-physics-hybrid polarization-based HDR reconstruction pipeline in Section III-C.

A. Polarized LDR image formation model

While the real-world scenes often have a high dynamic range, the digital sensor in cameras can only capture and store a limited extent, usually with 8 bits. Given the scene radiance $\mathbf{E}$ and sensor exposure time t, an HDR image can be expressed as

$$\mathbf{H} = \mathbf{E} \cdot t. \tag{1}$$

Inside a polarization camera, as shown in Figure 1 (a), the process of converting the HDR image $\mathbf{H}$ into four spatially-aligned and temporally-synchronized 8-bit polarized LDR images $\mathbf{L}_i$ $(i = 1, 2, 3, 4)$ with different polarizer angles α_i ($i = 1, 2, 3, 4$ and $\alpha_{1,2,3,4} = 0°, 45°, 90°, 135°$ respectively) in a single shot could be modeled by the following three major steps:

(1) **Polarization filtering.** The four-directional on-chip micro-polarizers first filter the HDR image $\mathbf{H}$ to get four polarized HDR images $\mathbf{H}_i$ $(i = 1, 2, 3, 4)$ by

$$\mathbf{H}_{1,2,3,4} = \mathcal{P}(\mathbf{H}), \tag{2}$$

where $\mathcal{P}$ is the polarization filtering operation. According to Malus' law [16], the relationship between $\mathbf{H}_i$ and $\mathbf{H}$ can be expressed as the following equation:

$$\mathbf{H}_i = \mathcal{P}_i(\mathbf{H}) = \frac{1}{2}\mathbf{H} \cdot (1 - \mathbf{p} \cdot \cos(2(\alpha_i - \boldsymbol{\theta}))), \tag{3}$$

where $\mathcal{P}_i$ $(i = 1, 2, 3, 4)$ stands for the polarization filtering operation performed by the on-chip micro-polarizer at α_i, $\mathbf{p} \in [0, 1]^3$ and $\boldsymbol{\theta} \in [0, 180°]$ denote the DoP and AoP of the incoming light to the sensor respectively. Note that the DoP and AoP are not properties of specific scene objects, which means that even if the incoming light to the sensor has a complex polarization distribution (*i.e*., it consists of multiple components with different polarized states), we can always calculate its unique DoP and AoP. Defining the Stokes parameters [25] of the incoming light to the sensor as

$$\begin{cases} \mathbf{S}_0 = \frac{1}{2}(\mathbf{H}_1 + \mathbf{H}_2 + \mathbf{H}_3 + \mathbf{H}_4) \\ \mathbf{S}_1 = \mathbf{H}_3 - \mathbf{H}_1 \\ \mathbf{S}_2 = \mathbf{H}_4 - \mathbf{H}_2 \end{cases}, \tag{4}$$

[3] $\mathbf{p} = 0$ means the incoming light to the sensor is unpolarized, and $0 < \mathbf{p} \leq 1$ means it is polarized ($0 < \mathbf{p} < 1$: partially polarized; $\mathbf{p} = 1$: totally polarized).

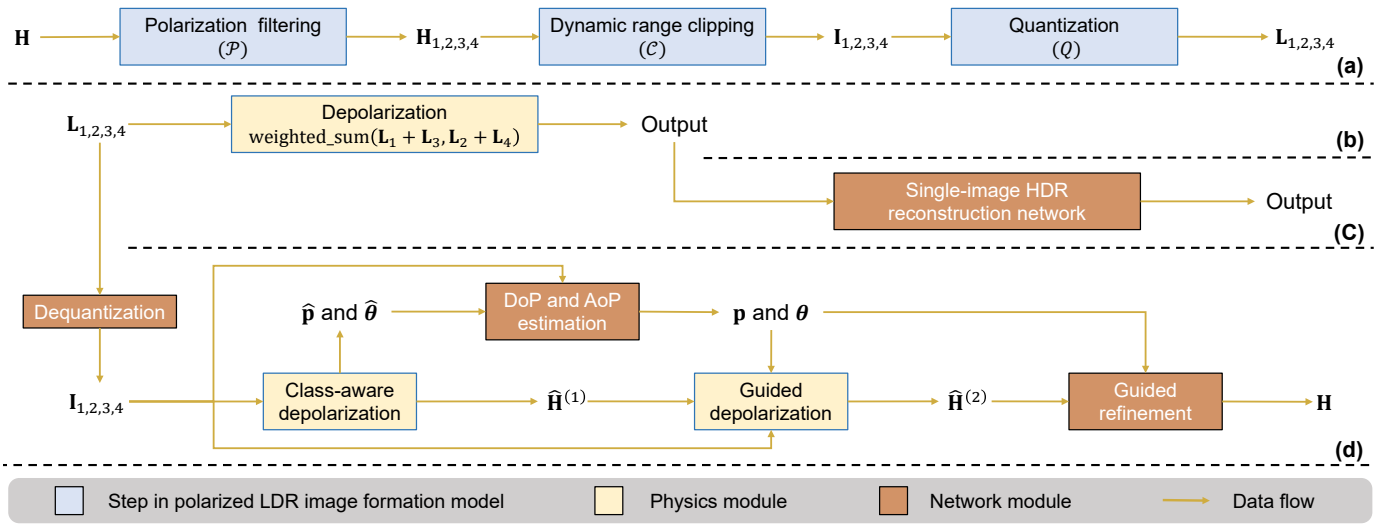


Fig. 1. (a) Polarized LDR image formation model (Section III-A). (b) Directly conducting uniform depolarization on $\mathbf{L}_{1,2,3,4}$ [51]. (c) Adopting a single-image HDR reconstruction network [9] to refine the output of (b) [45]. (d) We propose a pixel-wise depolarization strategy (Section III-B) and a network-physics-hybrid polarization-based HDR reconstruction pipeline (Section III-C), complementarily adopting network modules and physics modules to handle the ill-posed and well-posed computations respectively.

the DoP $\mathbf{p}$ and AoP $\boldsymbol{\theta}$ can be calculated by

$$\mathbf{p}=\frac{\sqrt{\mathbf{S}_1^2+\mathbf{S}_2^2}}{\mathbf{S}_0} \quad \text{and} \quad \boldsymbol{\theta}=\frac{1}{2}\arctan\left(\frac{\mathbf{S}_2}{\mathbf{S}_1}\right). \tag{5}$$

Note that we only consider linear polarization since the polarization camera contains only linear polarizers. Plugging the values of α_i $(i=1,2,3,4)$ into Equation (3), we get

$$\begin{cases}\mathbf{H}_{3,1}=\frac{1}{2}\mathbf{H}\cdot(1\pm\mathbf{p}\cdot\cos(2\boldsymbol{\theta}))\\ \mathbf{H}_{4,2}=\frac{1}{2}\mathbf{H}\cdot(1\pm\mathbf{p}\cdot\sin(2\boldsymbol{\theta}))\end{cases}, \tag{6}$$

and we call it the *internal constraints of polarization* for brevity.

(2) **Dynamic range clipping.** The polarization camera then clips the pixel values[4] of $\mathbf{H}_i$ $(i=1,2,3,4)$ to a limited range to get the unquantized polarized LDR images $\mathbf{I}_i$ $(i=1,2,3,4)$ by

$$\mathbf{I}_i=\mathcal{C}(\mathbf{H}_i)=\min(\mathbf{H}_i,1), \tag{7}$$

where $\mathcal{C}$ is the dynamic range clipping operation.

(3) **Quantization.** The last step is quantizing $\mathbf{I}_i$ $(i=1,2,3,4)$ to 8-bit polarized LDR images $\mathbf{L}_i$ $(i=1,2,3,4)$ by

$$\mathbf{L}_i=\mathcal{Q}(\mathbf{I}_i)=\lfloor 255(\mathbf{I}_i+\boldsymbol{\epsilon})\rfloor/255, \tag{8}$$

where $\mathcal{Q}$ is the quantization operation and $\boldsymbol{\epsilon}$ denotes noise[5].

Note that in the above-mentioned steps we do not include non-linear mapping like the image formation model introduced in [28], this is because a polarization camera can output images with a linear radiometric response function, *i.e.*, the pixel values linearly relate to scene radiance.

[4]All pixel values are normalized to $[0,1]$ in this paper.

[5]Discussions about the noise model can be found in the supplementary material.

In summary, the polarized LDR image formation model $\mathcal{M}$ can be expressed as

$$\mathbf{L}_{1,2,3,4}=\mathcal{M}(\mathbf{H}), \tag{9}$$

in which

$$\mathbf{L}_i=\mathcal{Q}(\mathcal{C}(\mathcal{P}_i(\mathbf{H}))) \quad (i=1,2,3,4). \tag{10}$$

Our goal is finding an inverse mapping $\mathcal{M}^{-1}$ to reconstruct the HDR image $\mathbf{H}$ from four 8-bit polarized LDR images $\mathbf{L}_{1,2,3,4}$ captured by a polarization camera in a single shot, based on an assumption that all pixels tend to have non-zero DoPs in our daily photography (*i.e.*, all pixels are polarized)[6].

Analysis of the potential dynamic range gain. In general, the dynamic range of a conventional image sensor is often expressed as

$$R=20\log_{10}\frac{L_{max}}{L_{min}}, \tag{11}$$

where L_{max} and L_{min} denote the maximum gray level (corresponding to the full-well capacity) and the minimum gray level (corresponding to the read-noise) respectively [33]. Therefore, an 8-bit sensor with a linear response function results in a dynamic range of $20\log_{10}255\approx 48.13$ dB. By recording with multiple exposures, the sensed dynamic range can be expanded [6], and it is related to the ratio between the maximum exposure e_{max} and the minimum exposure e_{min} [33], which can be written as

$$R_{me}=20\log_{10}\frac{L_{max}e_{max}}{L_{min}e_{min}}. \tag{12}$$

In the case of a polarization camera, the parameters of polarization filtering could be regarded as different exposures [51]. Similar to Equation (12), defining $f_{3,1}=\frac{1}{2}(1\pm p\cdot\cos(2\theta))$ and $f_{4,2}=\frac{1}{2}(1\pm p\cdot\sin(2\theta))$ (where $p\in[0,1]$ and $\theta\in[0,180^\circ]$

[6]We experimentally verify that in our dataset, the proportion of pixels with small DoPs ($\mathbf{p}\leq 0.03$) is around 4.7%, so that this assumption is reasonable.

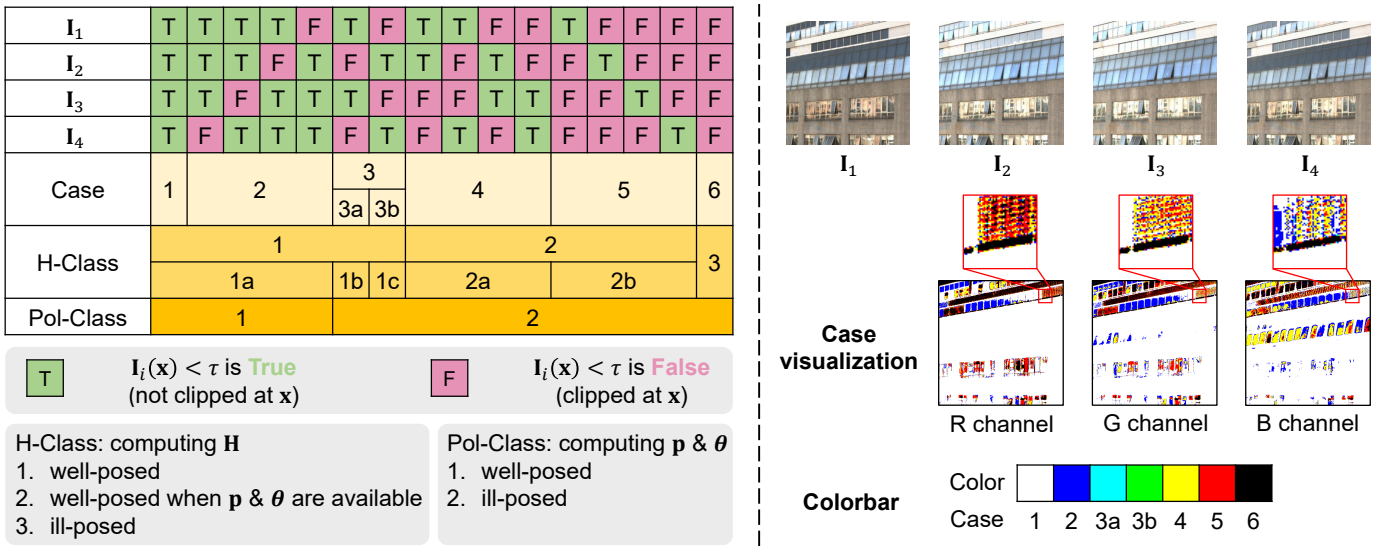


Fig. 2. Left: We label the pixels $\mathbb{X} = \{\mathbf{x} \mid \mathbf{x} = (x, y, c)\}$ into several cases based on whether the images in $\mathbb{I} = \{\mathbf{I}_i\}_{i=1}^4$ have not been clipped at $\mathbf{x}$, and classify these cases based on the level of ill-posedness in computing $\mathbf{H}$ (H-Class), $\mathbf{p}$ and $\boldsymbol{\theta}$ (Pol-Class). Right: An example to visualize each case at each color channel.

denote the DoP and AoP of a certain pixel), the dynamic range of an ideal polarization camera (*i.e.*, the micro-polarizers inside it are perfect) at a certain pixel can be written as

$$R_{pc} = 20 \log_{10} \frac{L_{max} \max(f_1, f_2, f_3, f_4)}{L_{min} \min(f_1, f_2, f_3, f_4)}, \tag{13}$$

Subtracting Equation (11) from Equation (13), we could obtain the potential dynamic range gain of an ideal polarization camera at a certain pixel as

$$G_{pc} = 20 \log_{10} \frac{\max(f_1, f_2, f_3, f_4)}{\min(f_1, f_2, f_3, f_4)}. \tag{14}$$

From Equation (14) we can see that G_{pc} depends on both p and θ, and it increases monotonically with p when θ is fixed, which means that the larger amount of polarization in the incoming light to the sensor, the higher the potential dynamic range gain of a polarization camera; besides, the maximum value of G_{pc} could be very large (*e.g.*, when $p \to 1$ and $\theta \to 0$, $G_{pc} \to \infty$). However, since the real micro-polarizers inside a polarization camera are not perfect, their extinction ratio e_r (the ratio between the maximum and minimum signal detected by a particular polarization pixel channel for a linearly polarized input, which depends on the wavelength) is not very large (often about several hundred, see the link in Footnote 1 for the extinction ratio of the Lucid Vision Phoenix polarization camera we used in this paper), so that the actual dynamic range gain could be smaller than the one computed by Equation (14) (*i.e.*, the maximum value of G_{pc} can only reach $20 \log_{10} e_r$). Practically, since p can be spatially-variant, different regions could obtain different dynamic range gains. These properties are very attractive when handling highly polarized scenes with strong radiance values (*e.g.*, specular highlights and scattered sunlight). By assuming a uniform distribution of θ, Wu *et al.* [51] quantify G_{pc} under three different values of p: at $p = 0.2$, $p = 0.5$, and $p = 0.8$, the dynamic range gains are expected to be 3.2, 8.5, and 16.4 respectively.

B. Pixel-wise depolarization

According to Equation (6), a straightforward way to achieve our goal is to estimate $\mathbf{H}_{1,2,3,4}$ first and then compute $\mathbf{H}$ by conducting depolarization on them:

$$\begin{aligned} \mathbf{H} &= \texttt{weighted\_sum}(\mathbf{H}_1 + \mathbf{H}_3, \mathbf{H}_2 + \mathbf{H}_4) \\ &= \frac{\sum_{i=1}^{2} (\mathcal{W}(\mathbf{H}_i + \mathbf{H}_{i+2}) \cdot (\mathbf{H}_i + \mathbf{H}_{i+2}))}{\sum_{i=1}^{2} \mathcal{W}(\mathbf{H}_i + \mathbf{H}_{i+2})}, \end{aligned} \tag{15}$$

where $\mathcal{W}$ is a weight function (*e.g.*, in [51] and [45], $\mathcal{W}$ is selected to be the Gaussian function with a standard deviation of $\sigma = 0.2$). However, estimating $\mathbf{H}_{1,2,3,4}$ requires handling multiple independent single-image HDR reconstruction problems, which is far too ill-posed. Wu *et al.* [51] directly conducted uniform[7] depolarization on $\mathbf{L}_{1,2,3,4}$ (Figure 1 (b)), which is not robust since the values of $\mathbf{L}_i$ have been clipped and quantized during the formation (*i.e.*, $\mathbf{L}_i \neq \mathbf{H}_i$). Ting *et al.* [45] supposed that the pixels in the output of Wu *et al.* [51] are still LDR, and adopted a single-image HDR reconstruction network [9] to refine them (Figure 1 (c)), which still tried to solve the problem in an ill-posed manner. In contrast, we find that the depolarization process could be optimized to handle the problem if the internal constraints of polarization are properly used. So, we propose a pixel-wise depolarization strategy, which aims to classify the pixels based on their levels of ill-posedness in computing $\mathbf{H}$ and apply different solutions to different classes, as detailed below.

Since $\mathbf{L}_{1,2,3,4}$ are noisy and often suffer from quantization artifacts in smooth regions, directly using them to compute $\mathbf{H}$ may not be suitable. We therefore propose to map them back to $\mathbf{I}_{1,2,3,4}$ by denoising and dequantization first. Then, to make the problem more tractable, we label the pixels $\mathbb{X} = \{\mathbf{x} \mid \mathbf{x} = (x, y, c)\}$ (x and y are pixel coordinates, and c denotes the color channel index) into several cases based on whether

[7] "Uniform" means treating all pixels in the same manner.

the images in $\mathbb{I} = \{\mathbf{I}_i\}_{i=1}^4$ satisfy $\mathbf{I}_i(\mathbf{x}) < \tau$[8] (*i.e.*, $\mathbf{I}_i$ has not been clipped at $\mathbf{x}$ so that $\mathbf{I}_i(\mathbf{x}) = \mathbf{H}_i(\mathbf{x})$), as shown in the left column of Figure 2 (*Case*):

Case1: $\mathbb{X}_1 = \{\mathbf{x} \mid \sum_{\mathbf{I}_i \in \mathbb{I}} \texttt{b2i}(\mathbf{I}_i(\mathbf{x}) < \tau) = 4\}$;
Case2: $\mathbb{X}_2 = \{\mathbf{x} \mid \sum_{\mathbf{I}_i \in \mathbb{I}} \texttt{b2i}(\mathbf{I}_i(\mathbf{x}) < \tau) = 3\}$;
Case3: $\mathbb{X}_3 = \mathbb{X}_{3a} \cup \mathbb{X}_{3b}$, where
$\mathbb{X}_{3a} = \{\mathbf{x} \mid \mathbf{I}_{1,3}(\mathbf{x}) < \tau, \mathbf{I}_{2,4}(\mathbf{x}) \geq \tau\}$,
$\mathbb{X}_{3b} = \{\mathbf{x} \mid \mathbf{I}_{2,4}(\mathbf{x}) < \tau, \mathbf{I}_{1,3}(\mathbf{x}) \geq \tau\}$;
Case4: $\mathbb{X}_4 = \{\mathbf{x} \mid \sum_{\mathbf{I}_i \in \mathbb{I}} \texttt{b2i}(\mathbf{I}_i(\mathbf{x}) < \tau) = 2\} - \mathbb{X}_3$;
Case5: $\mathbb{X}_5 = \{\mathbf{x} \mid \sum_{\mathbf{I}_i \in \mathbb{I}} \texttt{b2i}(\mathbf{I}_i(\mathbf{x}) < \tau) = 1\}$;
Case6: $\mathbb{X}_6 = \{\mathbf{x} \mid \sum_{\mathbf{I}_i \in \mathbb{I}} \texttt{b2i}(\mathbf{I}_i(\mathbf{x}) < \tau) = 0\}$,

where `b2i` is a function converting bool values into integer values, which can be written as

$$\texttt{b2i}(\textit{cond.}) = \begin{cases} 1 & \text{if } \textit{cond.} \text{ is True} \\ 0 & \text{if } \textit{cond.} \text{ is False} \end{cases}. \tag{16}$$

An example to visualize each case at each color channel could be found in the right column of Figure 2.

We further classify these cases based on the level of ill-posedness in computing $\mathbf{H}$, as shown in the left column of Figure 2 (*H-Class*):

H-Class1: $\mathbb{X}_1^H = \mathbb{X}_{1a}^H \cup \mathbb{X}_{1b}^H \cup \mathbb{X}_{1c}^H$, where
$\mathbb{X}_{1a}^H = \mathbb{X}_1 \cup \mathbb{X}_2$, $\mathbb{X}_{1b}^H = \mathbb{X}_{3a}$, $\mathbb{X}_{1c}^H = \mathbb{X}_{3b}$;
H-Class2: $\mathbb{X}_2^H = \mathbb{X}_{2a}^H \cup \mathbb{X}_{2b}^H$, where
$\mathbb{X}_{2a}^H = \mathbb{X}_4$, $\mathbb{X}_{2b}^H = \mathbb{X}_5$;
H-Class3: $\mathbb{X}_3^H = \mathbb{X}_6$,

and analyze them class-by-class:

H1. For H-Class1 ($\mathbb{X}_1^H$), $\mathbf{H}$ can be directly computed from $\mathbf{I}_{1,2,3,4}$ in a well-posed manner, but different approaches should be taken to handle three different subclasses ($\mathbb{X}_{1a}^H$, $\mathbb{X}_{1b}^H$, and $\mathbb{X}_{1c}^H$):

H1a. For $\mathbb{X}_{1a}^H$, $\mathbf{H}$ can be obtained by solving a linear system derived from Equation (3) with the images which have not been clipped as the input[9]:

$$\mathbf{I}_i = \begin{bmatrix} \frac{1}{2} & \frac{-\cos(2\alpha_i)}{2} & \frac{-\sin(2\alpha_i)}{2} \end{bmatrix} \begin{bmatrix} \mathbf{H} \\ \mathbf{D}_c \\ \mathbf{D}_s \end{bmatrix}, \tag{17}$$

where

$$\begin{cases} \mathbf{D}_c = \mathbf{H} \cdot \mathbf{p} \cdot \cos(2\boldsymbol{\theta}) \\ \mathbf{D}_s = \mathbf{H} \cdot \mathbf{p} \cdot \sin(2\boldsymbol{\theta}) \end{cases}. \tag{18}$$

$\mathbf{H}$, $\mathbf{D}_c$, and $\mathbf{D}_s$ are three unknowns to be solved from the linear system;

H1b. For $\mathbb{X}_{1b}^H$, $\mathbf{H}$ can be obtained by

$$\mathbf{H} = \mathbf{I}_1 + \mathbf{I}_3; \tag{19}$$

H1c. For $\mathbb{X}_{1c}^H$, $\mathbf{H}$ can be obtained by

$$\mathbf{H} = \mathbf{I}_2 + \mathbf{I}_4. \tag{20}$$

H2. For H-Class2 ($\mathbb{X}_2^H$), computing $\mathbf{H}$ is ill-posed with only $\mathbf{I}_{1,2,3,4}$. However, from Equation (6) we notice that once the polarization-related physical parameters ($\mathbf{p}$ and $\boldsymbol{\theta}$) are available, we could compute $\mathbf{H}$ in a well-posed manner again by using an arbitrary image $\mathbf{I}_i$ which has not been clipped:

$$\mathbf{H} = \begin{cases} \frac{2\mathbf{I}_i}{(1 \pm \mathbf{p} \cdot \cos(2\boldsymbol{\theta}))} & \text{if } i = 3, 1 \\ \frac{2\mathbf{I}_i}{(1 \pm \mathbf{p} \cdot \sin(2\boldsymbol{\theta}))} & \text{if } i = 4, 2 \end{cases}. \tag{21}$$

The only difference between subclasses $\mathbb{X}_{2a}^H$ and $\mathbb{X}_{2b}^H$ is the number of images which have not been clipped:

H2a. For $\mathbb{X}_{2a}^H$, the problem is overdetermined since we have two images which have not been clipped and both of them can be used to compute $\mathbf{H}$ by Equation (21). We could just average their computed results as the final $\mathbf{H}$;

H2b. For $\mathbb{X}_{2b}^H$, only one image has not been clipped, and we could use it to compute $\mathbf{H}$ by Equation (21).

H3. For H-Class3 ($\mathbb{X}_3^H$), computing $\mathbf{H}$ is ill-posed since all images have been clipped.

Note that the words "well-posed" and "ill-posed" denote the problems to which there is a unique solution (determined or overdetermined problem, note that for overdetermined problem we apply the least-squares solution to ensure the uniqueness) and where there are multiple solutions (underdetermined problem) respectively.

Since computing $\mathbf{H}$ in $\mathbb{X}_2^H$ requires estimating $\mathbf{p}$ and $\boldsymbol{\theta}$ first, we perform an additional classification based on the level of ill-posedness in computing $\mathbf{p}$ and $\boldsymbol{\theta}$, as shown in the left column of Figure 2 (*Pol-Class*):

Pol-Class1: $\mathbb{X}_1^{\text{Pol}} = \mathbb{X}_1 \cup \mathbb{X}_2$;
Pol-Class2: $\mathbb{X}_2^{\text{Pol}} = \mathbb{X}_3 \cup \mathbb{X}_4 \cup \mathbb{X}_5 \cup \mathbb{X}_6$,

and also analyze them class-by-class:

Pol1. For Pol-Class1 ($\mathbb{X}_1^{\text{Pol}}$), $\mathbf{p}$ and $\boldsymbol{\theta}$ can be directly computed from $\mathbf{I}_{1,2,3,4}$ in a well-posed manner by solving Equation (17):

$$\mathbf{p} = \frac{\sqrt{\mathbf{D}_c^2 + \mathbf{D}_s^2}}{\mathbf{H}} \text{ and } \boldsymbol{\theta} = \frac{1}{2}\arctan(\frac{\mathbf{D}_s}{\mathbf{D}_c}); \tag{22}$$

Pol2. For Pol-Class2 ($\mathbb{X}_2^{\text{Pol}}$), computing $\mathbf{p}$ and $\boldsymbol{\theta}$ is ill-posed.

Therefore, estimating $\mathbf{p}$ and $\boldsymbol{\theta}$ is equivalent to an "interpolation-like" task in $\mathbb{X}_2^{\text{Pol}}$ using the priors of computed values in $\mathbb{X}_1^{\text{Pol}}$, which can be handled by optimization-based or learning-based approaches with the help of semantic information provided by $\mathbf{I}_{1,2,3,4}$. The reason why we propose to estimate $\mathbf{p}$ and $\boldsymbol{\theta}$ first and then use them to compute $\mathbf{H}$ instead of hallucinating $\mathbf{H}$ directly is that the ranges of $\mathbf{p}$ and $\boldsymbol{\theta}$ are bounded, making them much easier to estimate.

In the end, only the pixel values of $\mathbf{H}$ in $\mathbb{X}_3^H$ have not been computed yet, and we propose to hallucinate their values under the guidance of $\mathbf{p}$ and $\boldsymbol{\theta}$ since they encode abundant structural and contextual information of the scene, which can offer useful priors to guide the hallucination.

[8] τ denotes the threshold used for checking if the pixel is saturated or not. In our implementation, we set τ to 0.95 as other works [8], [28] do.

[9] If all images in $\mathbb{I}$ have not been clipped, the linear system becomes overdetermined, and we apply the least-squares solution to it.

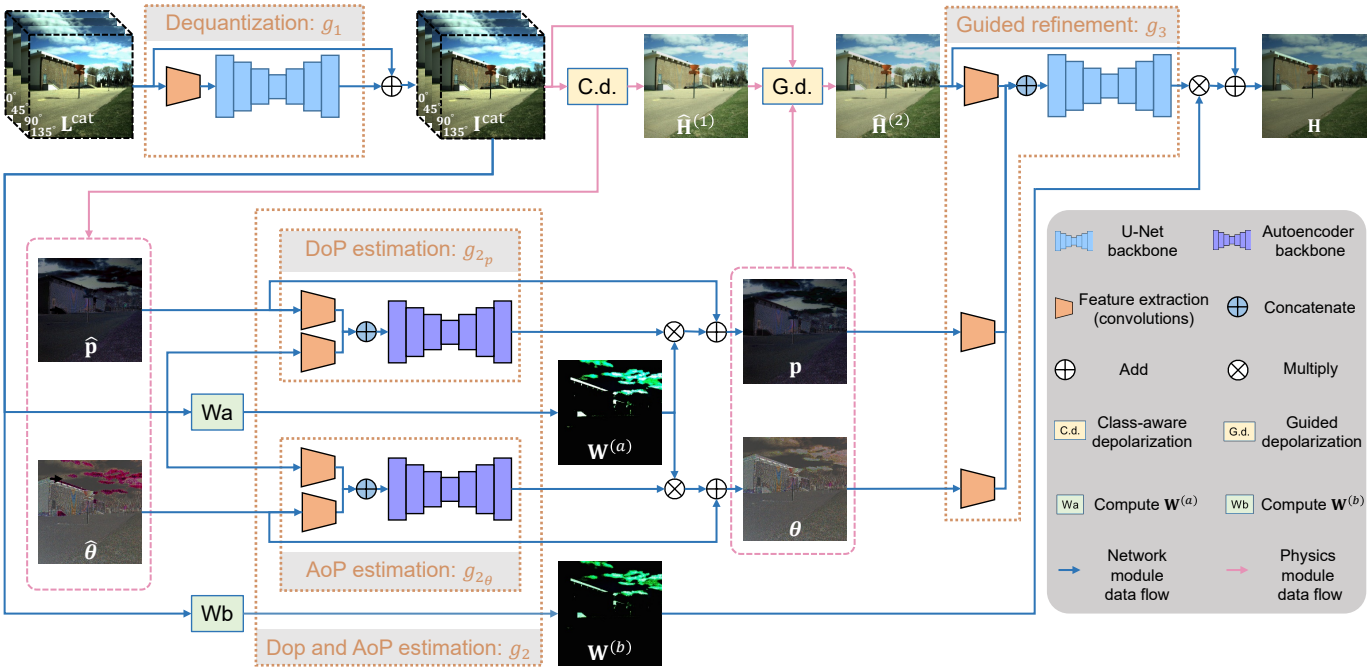


Fig. 3. Architecture of our Pol-HDR network, which consists of three subnetworks (g_1, g_2, and g_3) to implement three network modules (Dequantization, DoP and AoP estimation, and Guided refinement) in our hybrid Pol-HDR pipeline (Figure 1 (d)) respectively.

C. Network-physics-hybrid Pol-HDR pipeline

From Section III-B, we can see that not all computations in our pixel-wise depolarization strategy can be handled in a well-posed manner. So, we propose a network-physics-hybrid polarization-based HDR (Pol-HDR) reconstruction pipeline, complementarily adopting *network modules* and *physics modules* to handle the *ill-posed* and *well-posed* computations respectively. As shown in Figure 1 (d), it consists of five stages:

(1) **Dequantization** (network module). It estimates $\mathbf{I}_{1,2,3,4}$ from $\mathbf{L}_{1,2,3,4}$ by performing denoising and dequantization on them.

(2) **Class-aware depolarization** (physics module). It computes $\mathbf{H}$ in H-Class1 ($\mathbb{X}_1^H$) using Equation (17), Equation (19), and Equation (20), and computes $\mathbf{p}$ and $\boldsymbol{\theta}$ in Pol-Class1 ($\mathbb{X}_1^{\text{Pol}}$) using Equation (22) from $\mathbf{I}_{1,2,3,4}$. Besides, for other pixels, it temporarily assumes all images have not been clipped and uses Equation (17) and Equation (22) to fill them with inaccurate values. We denote its output as $\hat{\mathbf{H}}^{(1)}$, $\hat{\mathbf{p}}$, and $\hat{\boldsymbol{\theta}}$ (the coarse values of $\mathbf{H}$, $\mathbf{p}$, and $\boldsymbol{\theta}$ respectively).

(3) **DoP and AoP estimation** (network module). It estimates $\mathbf{p}$ and $\boldsymbol{\theta}$ in Pol-Class2 ($\mathbb{X}_2^{\text{Pol}}$) by updating $\hat{\mathbf{p}}$ and $\hat{\boldsymbol{\theta}}$ with the help of semantic information provided by $\mathbf{I}_{1,2,3,4}$.

(4) **Guided depolarization** (physics module). It computes $\mathbf{H}$ in H-Class2 ($\mathbb{X}_2^H$) using Equation (21) from $\mathbf{I}_{1,2,3,4}$ by updating $\hat{\mathbf{H}}^{(1)}$ under the guidance of $\mathbf{p}$ and $\boldsymbol{\theta}$. We denote its output as $\hat{\mathbf{H}}^{(2)}$ (the updated values of $\hat{\mathbf{H}}^{(1)}$).

(5) **Guided refinement** (network module). It estimates $\mathbf{H}$ in H-Class3 ($\mathbb{X}_3^H$) by refining $\hat{\mathbf{H}}^{(2)}$ under the guidance of structural and contextual information of the scene encoded in $\mathbf{p}$ and $\boldsymbol{\theta}$.

IV. Pol-HDR network

According to Section III-C, three network modules (Dequantization, DoP and AoP estimation, and Guided refinement) are required in our hybrid Pol-HDR pipeline. So, we design a neural network named Pol-HDR network which consists of three subnetworks (g_1, g_2, and g_3) to implement them respectively, as shown in Figure 3.

A. Dequantization (g_1)

As shown in Figure 3 (g_1), we choose the U-Net architecture [39] as the backbone of this subnetwork. By adopting the `Tanh` activation function to normalize the output of the last layer to $[-1, 1]$, it learns the residual between $\mathbf{L}^{\text{cat}}$ and $\mathbf{I}^{\text{cat}}$ (the concatenated $\mathbf{L}_{1,2,3,4}$ and $\mathbf{I}_{1,2,3,4}$), which can be described as

$$\mathbf{I}^{\text{cat}} = g_1(\mathbf{L}^{\text{cat}}) + \mathbf{L}^{\text{cat}}. \tag{23}$$

Since g_1 is only designed to perform denoising and dequantization on $\mathbf{L}_{1,2,3,4}$ simultaneously with little consideration about their relationship, it would break the internal constraints of polarization in the results, bringing errors to other stages. To handle this issue, we propose a polarization-based regularization term $\mathcal{L}_{p_1}$ in the loss function to enforce

$$\mathbf{I}_1 + \mathbf{I}_3 \equiv \mathbf{I}_2 + \mathbf{I}_4 \tag{24}$$

in $\mathbb{X}_1$, which is defined as

$$\mathcal{L}_{p_1}(\mathbf{I}^{\text{cat}}) = \mathcal{L}_2\left(\mathbf{m}_1 \cdot (\mathbf{I}_1 + \mathbf{I}_3), \mathbf{m}_1 \cdot (\mathbf{I}_2 + \mathbf{I}_4)\right), \tag{25}$$

where $\mathcal{L}_2$ denotes the L_2 loss, and $\mathbf{m}_1$ denotes a binary mask which equals to 1 (0) in $\mathbb{X}_1$ ($\mathbb{X}-\mathbb{X}_1$). The loss function of g_1 is defined as

$$\mathcal{L}_{g_1} = \lambda_1 \cdot \mathcal{L}_2(\mathbf{I}^{\text{cat}}, \mathbf{I}^{\text{cat}}_{\text{gt}}) + \lambda_2 \cdot \mathcal{L}_{p_1}(\mathbf{I}^{\text{cat}}), \tag{26}$$

where $\lambda_{1,2}$ are empirically set to be 100.0 and 10.0 respectively, and gt labels the ground truth across this paper.

B. DoP and AoP estimation (g_2)

As shown in Figure 3 (g_2), we choose the autoencoder [17] architecture as the backbone of this subnetwork. Since this subnetwork aims to learn two different parameters ($\mathbf{p}$ and $\boldsymbol{\theta}$), we design it as a two-branch architecture. The two branches g_{2_p} and g_{2_θ} learn the residual between $\hat{\mathbf{p}}$ and $\mathbf{p}$, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ respectively with the help of semantic information provided by $\mathbf{I}^{\text{cat}}$. Similar to g_1, g_{2_p} and g_{2_θ} can be described as

$$\begin{cases} \mathbf{p} = g_{2_p}(\hat{\mathbf{p}}, \mathbf{I}^{\text{cat}}) \cdot \mathbf{W}^{(a)} + \hat{\mathbf{p}} \\ \boldsymbol{\theta} = g_{2_\theta}(\hat{\boldsymbol{\theta}}, \mathbf{I}^{\text{cat}}) \cdot \mathbf{W}^{(a)} + \hat{\boldsymbol{\theta}} \end{cases}, \tag{27}$$

where $\mathbf{W}^{(a)}$ is a weight map to reweight the output of g_{2_p} and g_{2_θ}, which can be computed from $\mathbf{I}^{\text{cat}}$.

Now we explain why we need the weight map $\mathbf{W}^{(a)}$ and how to compute it. Since the pixel values of $\mathbf{p}$ and $\boldsymbol{\theta}$ in $\mathbb{X}_1^{\text{Pol}}$ can be computed almost accurately (*i.e.*, $\mathbf{p} \approx \hat{\mathbf{p}}$ and $\boldsymbol{\theta} \approx \hat{\boldsymbol{\theta}}$ in $\mathbb{X}_1^{\text{Pol}}$) in Class-aware depolarization stage, we should not alter them too much. Instead, we are supposed to focus on repairing the inaccurate values in $\mathbb{X}_2^{\text{Pol}}$. Therefore, we propose to use a weight map to reweight the output of g_{2_p} and g_{2_θ}:

$$\mathbf{W}^{(a)} = \sum_{k=1}^{6} (\mathbf{m}_k \cdot w_k^{(a)}), \tag{28}$$

where $\mathbf{m}_k$ denotes a binary mask which equals to 1 (0) in $\mathbb{X}_k$ ($\mathbb{X}-\mathbb{X}_k$), and $w_k^{(a)}$ is a confidence coefficient which equals to 0.0 (1.0) if the computed values in Class-aware depolarization stage are accurate (inaccurate). We empirically set $w_{1,2,3,4,5,6}^{(a)}$ to 0.01, 0.05, 1.0, 1.0, 1.0, 1.0 respectively since the values in $\mathbb{X}_1^{\text{Pol}} = \mathbb{X}_1 \cup \mathbb{X}_2$ are not that accurate[10], while the values in $\mathbb{X}_2^{\text{Pol}} = \mathbb{X}_3 \cup \mathbb{X}_4 \cup \mathbb{X}_5 \cup \mathbb{X}_6$ are inaccurate. The reason why we set $w_1^{(a)} < w_2^{(a)}$ is that the values in $\mathbb{X}_1$ are computed by four images (overdetermined), while the values in $\mathbb{X}_2$ are computed by three images (determined).

Directly using the ground truth to supervise the estimation of $\mathbf{p}$ and $\boldsymbol{\theta}$ is not robust enough since it would also break the internal constraints of polarization in the results. To address this problem, we propose another regularization term $\mathcal{L}_{p_2}$ derived from Equation (3) to enforce

$$\frac{\mathbf{I}_i}{\mathbf{I}_j} \equiv \frac{1 - \mathbf{p} \cdot \cos(2(\alpha_i - \boldsymbol{\theta}))}{1 - \mathbf{p} \cdot \cos(2(\alpha_j - \boldsymbol{\theta}))} \tag{29}$$

for pixels which have not been clipped in both $\mathbf{I}_i$ and $\mathbf{I}_j$, which is defined as

$$\mathcal{L}_{p_2}(\mathbf{I}^{\text{cat}}, \mathbf{p}, \boldsymbol{\theta}) = \sum_{\mathbf{I}_i \in \mathbb{I}, \mathbf{I}_j \in \mathbb{I}}^{i \neq j} \mathcal{L}_2\left(\mathbf{m}_{ij} \cdot \mathcal{P}_j(\mathbf{I}_i), \mathbf{m}_{ij} \cdot \mathcal{P}_i(\mathbf{I}_j)\right), \tag{30}$$

[10]Since the output of Dequantization stage may contain errors, the values in $\mathbb{X}_1^{\text{Pol}}$ computed by Class-aware depolarization stage may be inaccurate. We attempt to refine them in DoP and AoP estimation stage.

where $\mathbf{m}_{ij}$ denotes a binary mask which equals to 1 (0) for pixels which have not been clipped in both $\mathbf{I}_i$ and $\mathbf{I}_j$ (otherwise), and the definition of $\mathcal{P}_i$ can be found in Equation (3). The loss function of g_2 is defined as

$$\begin{aligned} \mathcal{L}_{g_2} = {} & \lambda_3 \cdot \mathcal{L}_1(\mathbf{p}, \mathbf{p}_{\text{gt}}) + \lambda_4 \cdot \mathcal{L}_2(\mathbf{p}, \mathbf{p}_{\text{gt}}) + \lambda_5 \cdot \mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{gt}}) \\ & + \lambda_6 \cdot \mathcal{L}_2(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{gt}}) + \lambda_7 \cdot \mathcal{L}_{p_2}(\mathbf{I}^{\text{cat}}, \mathbf{p}, \boldsymbol{\theta}), \end{aligned} \tag{31}$$

where $\mathcal{L}_1$ denotes the L_1 loss, and $\lambda_{3,4,5,6,7}$ are empirically set to be 10.0, 100.0, 10.0, 100.0, 1.0 respectively.

C. Guided refinement (g_3)

As shown in Figure 3 (g_3), we also choose the U-Net architecture [39] as the backbone of this subnetwork. Since the values of $\mathbf{H}$ should be no smaller than $\hat{\mathbf{H}}^{(2)}$, we aim to learn the positive residual between $\hat{\mathbf{H}}^{(2)}$ and $\mathbf{H}$ (by adopting the `ReLU` activation function to normalize the output of the last layer to $[0, \infty]$) with the help of structural and contextual information encoded in $\mathbf{p}$ and $\boldsymbol{\theta}$. Similar to g_1, g_3 can be described as

$$\mathbf{H} = g_3(\hat{\mathbf{H}}^{(2)}, \mathbf{p}, \boldsymbol{\theta}) \cdot \mathbf{W}^{(b)} + \hat{\mathbf{H}}^{(2)}, \tag{32}$$

where $\mathbf{W}^{(b)}$ is another weight map to reweight the output of g_3. Similar to Equation (28), $\mathbf{W}^{(b)}$ is defined as

$$\mathbf{W}^{(b)} = \sum_{k=1}^{6} (\mathbf{m}_k \cdot w_k^{(b)}), \tag{33}$$

where $w_k^{(b)}$ is another confidence coefficient which equals to 0.0 (1.0) if the computed values in Guided depolarization stage are accurate (inaccurate). We empirically set $w_{1,2,3,4,5,6}^{(b)}$ to be 0.01, 0.05, 0.1, 0.25, 0.5, 1.0 respectively. The loss function of g_3 is defined as

$$\mathcal{L}_{g_3} = \lambda_8 \cdot \mathcal{L}_1(\mathbf{H}, \mathbf{H}_{\text{gt}}) + \lambda_9 \cdot \mathcal{L}_2(\Gamma(\mathbf{H}), \Gamma(\mathbf{H}_{\text{gt}})), \tag{34}$$

where $\lambda_{8,9}$ are empirically set to be 1.0 and 10.0 respectively, and Γ is a μ-law function to compress the data range, which is defined as

$$\Gamma(\mathbf{H}) = \frac{\log(1 + \mu \cdot \mathbf{H})}{\log(1 + \mu)}, \tag{35}$$

where $\mu = 5000$ is a parameter deciding the extent of compression. Note that in this situation we should not directly compute the L_2 loss in the linear domain (*i.e.*, use $\mathcal{L}_2(\mathbf{H}, \mathbf{H}_{\text{gt}})$ instead of $\mathcal{L}_2(\Gamma(\mathbf{H}), \Gamma(\mathbf{H}_{\text{gt}}))$ in Equation (34)). This is because in an HDR image, the highlight regions (*e.g.*, sun and light sources) typically have values with orders of magnitude larger than those of other regions, which would cause the loss function being dominated by the large values, while the effect of small values tends to be ignored [20], [28], [53].

V. Data preparation and implementation details

In this section, we first detail our synthetic dataset generation pipeline in Section V-A, then show our implementation details in Section V-B.

A. Synthetic dataset generation pipeline

It is difficult to obtain pairwise HDR image and the corresponding 8-bit polarized LDR images at four different polarizer angles $(0°, 45°, 90°, 135°)$ at a large scale. Therefore, we propose to generate a synthetic dataset for training our Pol-HDR network. Since simulating the formation model of polarized LDR images (Section III-A) requires the ground truth values of the DoP and AoP, we cannot directly generate the polarized LDR images from the HDR images in existing HDR reconstruction benchmark datasets [20], [28]. However, we find that EdPolCommunity Dataset[11] [45] provides 103 different scenes, and each scene is captured with 17 bracketed exposures at four different polarizer angles $(0°, 45°, 90°, 135°)$ by a FLIR BFSU3-51S5p polarization camera[12], making it the desired data source for generating our dataset. In short, for each scene, our synthetic dataset generation pipeline[13] could be described as

(1) Reconstruct polarized HDR images at each polarizer angle from the bracketed exposed polarized LDR images using the classic multi-image HDR reconstruction method [6][14];
(2) calculate the ground truth DoP $\mathbf{p}_{\text{gt}}$ and AoP $\boldsymbol{\theta}_{\text{gt}}$ using Equation (4) and Equation (5);
(3) compute the unpolarized HDR image using Equation (15), normalize it to $[0, 1]$, and treat it as the ground truth scene radiance $\mathbf{E}_{\text{gt}}$;
(4) randomly generate an exposure time t to compute $\mathbf{H}_{\text{gt}}$ using Equation (1), and compute $\mathbf{I}_{\text{gt}}^{\text{cat}}$ and $\mathbf{L}^{\text{cat}}$ using the polarized image formation model (Section III-A);
(5) compute $\hat{\mathbf{p}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{H}}^{(2)}$ using the ways in our hybrid Pol-HDR pipeline (Section III-C);
(6) compute other parameters (*e.g.*, the weight maps $\mathbf{W}^{(a)}$ and $\mathbf{W}^{(b)}$) according to their definitions.

Since EdPolCommunity Dataset [45] does not provide a ready-made split file for splitting the training and testing sets, we randomly select 80 scenes for training, and the rest 23 scenes for testing. Due to the fact that the training set only contains 80 scenes, it is too small to train a neural network if we feed the original full-size images to the network. Therefore, for each scene, we randomly resize and crop the images (including both the input and output images of the three subnetworks) to 256×256 patches for training, and to 512×512 patches for testing. Besides, to avoid overfitting, we also perform data augmentation on those image patches (*e.g.*, flipping and rotating $90°$).

B. Implementation details

The total loss function is the sum of loss functions of each subnetwork:

$$\mathcal{L}_{total} = \mathcal{L}_{g_1} + \mathcal{L}_{g_2} + \mathcal{L}_{g_3}. \tag{36}$$

[11]https://github.com/jtuoa/Deep-Polarized-HDRreconstruction

[12]https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-51S5PC-C

[13]More details can be found in the supplementary material.

[14]Since multi-image HDR reconstruction [6] can be nearly regarded as a well-posed procedure, the errors between the reconstructed polarized HDR images and their corresponding ground truth values could be ignored. We could assume that the reconstructed polarized HDR images still satisfy the internal constraints of polarization as [51] and [45] do.

The feature extraction blocks in Figure 3 consist of three convolution layers with a stride of 1, and each of the convolution layers is followed by instance normalization [47] and `LeakyReLU` (negative slope is set to be 0.1) activation function. Their kernel sizes are set to be 1×1, 3×3, and 1×1 respectively. To keep the size of features, we employ zero padding for the second convolution layer (the padding size is set to be 1). For the U-Net [39] backbones in Figure 3, we set the number of downsampling blocks to 4, and add instance normalization [47], `ReLU` activation function, and Dropout [44] to the output of their downsampling blocks. For the autoencoder [17] backbones in Figure 3, we set the number of downsampling blocks to 3, and embed 5 residual bottleneck blocks [15] (instance normalization [47], `ReLU` activation function, and Dropout [44] are also added to their output) in the coarsest layer.

We implement our method using PyTorch[15] on a PC with an Intel Core i7-8700K CPU and an NVIDIA 2080Ti GPU, and apply a two-phase training strategy. First, to ensure a stable initialization, we train three subnetworks independently for 400 epochs. Then, we finetune the entire network in an end-to-end manner for another 300 epochs. For optimization, we use Adam optimizer [23] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate is set to be 5×10^{-4} during the training process without change. Dropout [44] and instance normalization [47] are also added during training.

VI. Experiments

In this section, we first evaluate the performance of our Pol-HDR network by conducting quantitative and qualitative comparisons on our synthetic dataset in Section VI-A and on real-world images in Section VI-B, then conduct ablation study in Section VI-C.

A. Evaluation on synthetic data

Since our Pol-HDR network takes four 8-bit polarized LDR images ($\mathbf{L}_{1,2,3,4}$) with different polarizer angles $(0°, 45°, 90°, 135°)$ captured by a polarization camera in a single shot (without adjusting bracketed exposures) as input, we compare our results with two existing polarization-based HDR reconstruction methods (Wu *et al*. [51] and Ting *et al*. [45]) which take the same type of input data as ours as input, and four state-of-the-art learning-based single-image HDR reconstruction methods (Eilertsen *et al*. [8], Endo *et al*. [9], Liu *et al*. [28], and Santos *et al*. [40]) which take a single conventional LDR image captured by a digital camera in a single shot as input. Note that comparing with learning-based single-image HDR reconstruction methods [8], [9], [28], [40] might be a bit unfair because of the difference in types of input data (polarized images *vs*. conventional images), and we conduct such comparisons to demonstrate the advantages of using polarized images w.r.t. state-of-the-art single-image approaches.

For conducting comparisons with single-image HDR reconstruction methods [8], [9], [28], [40], we adopt such a pipeline to get their results (taking $\mathbf{L}_{1,2,3,4}$ as input):

[15]https://pytorch.org/

Fig. 4. Qualitative comparisons on synthetic data among our Pol-HDR network, two existing polarization-based HDR reconstruction methods (Wu *et al*. [51] and Ting *et al*. [45]), and four state-of-the-art learning-based single-image HDR reconstruction methods (Eilertsen *et al*. [8], Endo *et al*. [9], Liu *et al*. [28], and Santos *et al*. [40]). Please zoom-in for better details.

TABLE I
QUANTITATIVE EVALUATION RESULTS ON SYNTHETIC DATA AMONG OUR POL-HDR NETWORK, TWO EXISTING POLARIZATION-BASED HDR RECONSTRUCTION METHODS (WU *et al*. [51] AND TING *et al*. [45]), AND FOUR STATE-OF-THE-ART LEARNING-BASED SINGLE-IMAGE HDR RECONSTRUCTION METHODS (EILERTSEN *et al*. [8], ENDO *et al*. [9], LIU *et al*. [28], AND SANTOS *et al*. [40]). THE HIGHER THE METRICS, THE BETTER THE RESULTS. **BOLD** FONT INDICATES THE BEST PERFORMANCE.

	Wu *et al*. [51]	Ting *et al*. [45]	Eilertsen *et al*. [8]	Endo *et al*. [9]	Liu *et al*. [28]	Santos *et al*. [40]	Ours
SSIM	0.985	0.917	0.989	0.971	0.986	0.981	**0.991**
MS-SSIM	0.990	0.935	0.993	0.972	0.993	0.992	**0.997**
PSNR	39.29	28.64	41.61	37.79	40.39	39.24	**43.12**
Q-Score [32]	60.30	58.20	64.77	60.07	61.79	62.66	**66.30**

(1) Estimate $\mathbf{H}_{1,2,3,4}$ from $\mathbf{L}_{1,2,3,4}$ using those methods;
(2) compute $\mathbf{H}$ by conducting depolarization on $\mathbf{H}_{1,2,3,4}$ using Equation (15).

Note that the reason why we do not adopt the opposite pipeline (conduct depolarization on $\mathbf{L}_{1,2,3,4}$ using Equation (15) first, and then use those methods to estimate $\mathbf{H}$) is that the opposite pipeline has already been adopted by one of the existing polarization-based HDR reconstruction methods (Ting *et al*. [45], see Figure 1 (c)), and it have already been included in our comparison group.

Visual quality comparisons are shown in Figure 4[16]. All HDR images are visualized using a gamma-exposure function. Compared to polarization-based HDR reconstruction methods [45], [51], our Pol-HDR network performs much better in recovering HDR contents, with recovered color appearance resembling the ground truth more closely; compared to single-image HDR reconstruction methods [8], [9], [28], [40], our Pol-HDR network generates fewer artifacts but finer image details, especially in areas where saturation occurs in the input polarized images. Taking the green box in the bottom right group of Figure 4 as an example, we could see that polarization-based HDR reconstruction methods [45], [51] produce color distortion artifacts, and single-image HDR reconstruction methods [8], [9], [28], [40] fail to recover texture details of the clouds. This is because polarization-based HDR reconstruction methods [45], [51] only utilize the spatially-variant exposures of the input polarized images and conduct uniform depolarization on them, and single-image HDR reconstruction methods [8], [9], [28], [40] need to handle multiple independent single-image HDR reconstruction problems, which is far too ill-posed.

To evaluate the results quantitatively, we adopt four image quality metrics including SSIM, MS-SSIM (multi-scale SSIM), PSNR, and Q-Score (produced by HDR-VDP-2.2 [32]). We measure the SSIM, MS-SSIM, and PSNR scores using the method proposed by Hanji *et al*. [13] (encoding the HDR images using PU21[17] with the `CRF_correction` function enabled before measurement), and only measure the Q-Score in the linear domain (as required by HDR-VDP-2.2 [32]). Results are shown in Table I. Our Pol-HDR network consistently outperforms the compared ones on all metrics.

B. Evaluation on real-world images

To demonstrate that our Pol-HDR network has a good generalization ability, we use the Lucid Vision Phoenix polarization camera (RGB) to capture real-world images[18] for further evaluation. For better evaluation, we also capture each scene with bracketed exposures at four different polarizer angles $(0°, 45°, 90°, 135°)$ similar to EdPolCommunity Dataset [45], and adopt the synthetic dataset generation pipeline proposed in Section V-A to obtain the ground truth HDR images.

Visual quality comparisons are shown in Figure 5[19]. Our Pol-HDR network can reconstruct visually more plausible HDR images with finer details than other methods. Taking the green box in the left group of Figure 5 as an example, we could see that the borders of holes in the building in both $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_4$ are invisible due to saturation, and those saturated pixels cannot be correctly reconstructed by all the other compared methods [8], [9], [28], [40], [45], [51]. This is because Wu *et al*. [51] do not consider the dynamic range clipping step in polarized LDR image formation model, Ting *et al*. [45] adopt a single-image HDR reconstruction network [9] to process all pixels in the same manner without considering their variant levels of ill-posedness, and single-image HDR reconstruction methods [8], [9], [28], [40] cannot make full use of the polarization information.

C. Ablation study

To verify the validity of each model design choice, we conduct a series of ablation studies, which can be expressed as

(1) End2end: training the networks in an end-to-end manner (to demonstrate the effectiveness of our network-physics-hybrid Pol-HDR pipeline).
(2) W/o g_1: removing g_1 from our Pol-HDR network (to verify the necessity of g_1).
(3) W/o g_2: removing g_2 from our Pol-HDR network (to verify the necessity of g_2).
(4) W/o g_3: removing g_3 from our Pol-HDR network (to verify the necessity of g_3).
(5) W/o $\mathbf{p}$ & $\boldsymbol{\theta}$: removing $\mathbf{p}$ and $\boldsymbol{\theta}$ from the input of g_3 (to validate the usefulness of structural and contextual information encoded in $\mathbf{p}$ and $\boldsymbol{\theta}$).
(6) W/o $\mathbf{W}^{(a)}$ & $\mathbf{W}^{(b)}$: removing the weight maps (to show the advantage of using weight maps over treating all pixels in the same manner).
(7) W/o $\mathcal{L}_{p_1}$ & $\mathcal{L}_{p_2}$: removing the regularization terms (to show the significance of enforcing the satisfaction of internal constraints of polarization).

[16]Additional synthetic results can be found in the supplementary material.
[17]https://github.com/gfxdisp/pu21
[18]We directly obtain the raw images from the camera, without demosaicing and color correction.
[19]Additional real-world results can be found in the supplementary material.

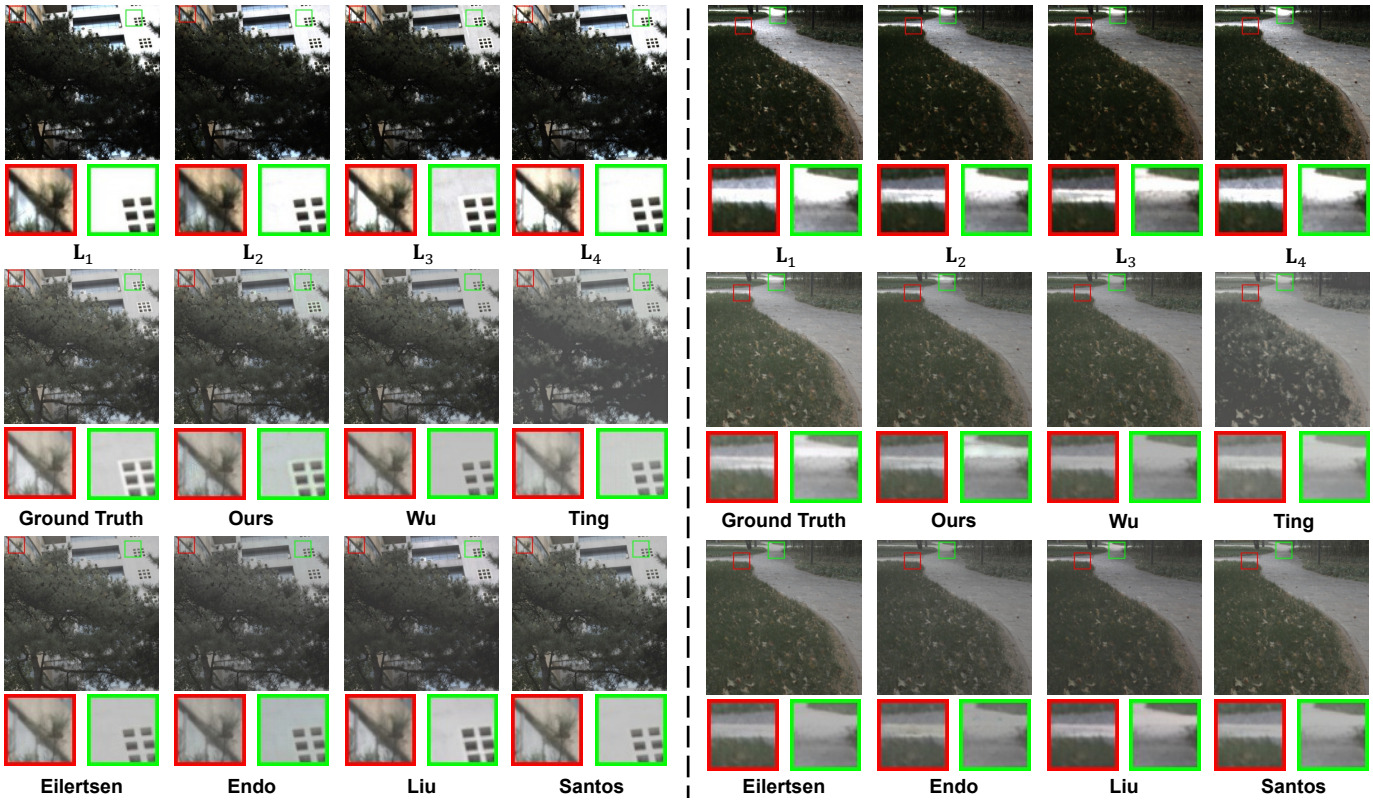


Fig. 5. Qualitative comparisons on real data. See the caption of Figure 4 for explanation.

TABLE II
QUANTITATIVE EVALUATION RESULTS OF ABLATION STUDY. SEE THE CAPTION OF TABLE I FOR EXPLANATION.

	End2end	W/o g_1	W/o g_2	W/o g_3	W/o $\mathbf{p}$ & $\boldsymbol{\theta}$	W/o $\mathbf{W}^{(a)}$ & $\mathbf{W}^{(b)}$	W/o $\mathcal{L}_{p_1}$ & $\mathcal{L}_{p_2}$	W/o classification	Our complete model
SSIM	0.989	0.977	0.988	0.984	0.989	0.990	0.986	0.985	**0.991**
MS-SSIM	0.995	0.993	0.995	0.990	0.996	0.995	0.996	0.990	**0.997**
PSNR	40.85	40.18	42.09	39.32	42.39	43.04	42.77	38.39	**43.12**
Q-Score [32]	65.24	62.73	64.87	61.41	65.78	65.92	65.08	60.57	**66.30**

(8) W/o classification: conducting uniform depolarization without classifying the pixels (to show the importance of our pixel-wise depolarization strategy).

Quantitative comparisons are shown in Table II. These results demonstrate that our complete model achieves the optimal performance with these specific design choices.

VII. CONCLUSION

We presented a learning-based method that leverages the properties of polarized light for HDR reconstruction. By using a polarization camera, our method can achieve single-shot HDR reconstruction without ghosting artifacts. To solve the polarization guided HDR reconstruction problem, we proposed a pixel-wise depolarization strategy, a network-physics-hybrid Pol-HDR pipeline, and a neural network fully exploiting the DoP and AoP, showing state-of-the-art performance. Our method enabled polarization-based HDR reconstruction methods to handle images captured in the wild with better generalization ability and higher robustness.

ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China under Grant 2021ZD0109803, National Natural Science Foundation of China under Grant No. 62136001 and 62088102.

REFERENCES

[1] Cecilia Aguerrebere, Andrés Almansa, Yann Gousseau, Julie Delon, and Pablo Muse. Single shot high dynamic range imaging using piecewise linear estimators. In *Proc. of International Conference on Computational Photography*, pages 1–10, 2014.
[2] Masheal M Alghamdi, Qiang Fu, Ali Kassem Thabet, and Wolfgang Heidrich. Reconfigurable snapshot hdr imaging using coded masks and inception network. In *Vision, Modeling and Visualization*, 2019.
[3] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proc. of International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, 2006.
[4] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.
[5] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 dB 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.
[6] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, 1997.

[7] Xuan Dong, Xiaoyan Hu, Weixin Li, Xiaojie Wang, and Yunhong Wang. MIEHDR CNN: Main image enhancement based ghost-free high dynamic range imaging using dual-lens systems. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 1264–1272, 2021.

[8] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 36(6):178:1–178:15, 2017.

[9] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 36(6):177:1–177:10, 2017.

[10] Shuai Fang, XiuShan Xia, Xing Huo, and ChangWen Chen. Image dehazing using polarization effects of objects and airlight. *Optics Express*, 22(16):19523–19537, 2014.

[11] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proc. of Computer Vision and Pattern Recognition*, pages 1730–1739, 2020.

[13] Param Hanji, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. Comparison of single image hdr reconstruction methods—the caveats of quality assessment. In *Proc. of ACM SIGGRAPH*, pages 1–8, 2022.

[14] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 35(6):1–12, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Eugene Hecht et al. *Optics*, volume 5. Addison Wesley San Francisco, 2002.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[18] Keigo Hirakawa and Paul M Simon. Single-shot high dynamic range imaging with conventional camera hardware. In *Proc. of International Conference on Computer Vision*, pages 1339–1346, 2011.

[19] Atsushi Ito, Toshio Yamazaki, and Seiji Kobayashi. Image processing apparatus, image processing method and program. *US Patent US8866884B2*, 2014.

[20] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 36(4):144–1, 2017.

[21] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *Proc. of International Conference on Computational Photography*, 2006.

[22] Junghee Kim, Siyeong Lee, and Suk-Ju Kang. End-to-end differentiable learning to HDR image synthesis for multi-exposure images. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 1780–1788, 2021.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Makoto Kobayashi, Seiji Tanaka, Kazuya Oda, Katsumi Ikeda, Kenkichi Hayashi, and Toru Nishimura. Development of Super CCD EXR. *ITE Technical Report*, 33.18:1–4, 2009.

[25] GP Können. *Polarized light in nature*. CUP Archive, 1985.

[26] Meredith K Kupinski, Christine L Bradley, David J Diner, Feng Xu, and Russell A Chipman. Angle of linear polarization images of outdoor scenes. *Optical Engineering*, 58(8):082419, 2019.

[27] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive HDRI: Inverse tone mapping using generative adversarial networks. In *Proc. of European Conference on Computer Vision*, pages 596–611, 2018.

[28] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proc. of Computer Vision and Pattern Recognition*, pages 1651–1660, 2020.

[29] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum*, 37:37–49, 2018.

[30] Belen Masia, Sandra Agustin, Roland W. Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 28(5):160:1–160:8, 2009.

[31] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. of Computer Vision and Pattern Recognition*, pages 1375–1385, 2020.

[32] Manish Narwaria, Rafal Mantiuk, Mattheiu P Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501, 2015.

[33] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, pages 472–479, 2000.

[34] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

[35] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2014.

[36] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution HDR deghosting with CNN. In *Proc. of European Conference on Computer Vision*, pages 497–513, 2020.

[37] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep HDR deghosting. In *Proc. of Computer Vision and Pattern Recognition*, pages 4875–4885, 2021.

[38] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. LDR2HDR: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 26(3), 2007.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

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

[41] Michael Schöberl, Alexander Belz, Jürgen Seiler, Siegfried Foessel, and André Kaup. High dynamic range video by spatially non-regular optical filtering. In *Proc. of International Conference on Image Processing*, pages 2757–2760, 2012.

[42] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 31(6):203:1–203:11, 2012.

[43] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum*, pages 153–163, 2016.

[44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[45] Juiwen Ting, Xuesong Wu, Kangkang Hu, and Hong Zhang. Deep snapshot HDR reconstruction based on the polarization camera. In *Proc. of International Conference on Image Processing*, 2021.

[46] Jack Tumblin, Amit Agrawal, and Ramesh Raskar. Why i want a gradient camera. In *Proc. of Computer Vision and Pattern Recognition*, pages 103–110, 2005.

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[48] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous Kalman filter for hybrid event cameras. In *Proc. of International Conference on Computer Vision*, pages 448–457, 2021.

[49] Lawrence B Wolff and Terrance E Boult. Constraining object features using a polarization reflectance model. *Physics-Based Vision: Principles and Practice: Radiometry*, 1:167, 1993.

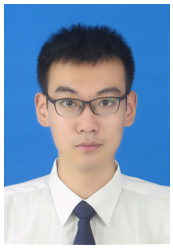
[50] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proc. of European Conference on Computer Vision*, pages 117–132, 2018.

[51] Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, and Juiwen Ting. HDR reconstruction based on the polarization camera. *IEEE Robotics and Automation Letters*, 5(4):5113–5119, 2020.

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

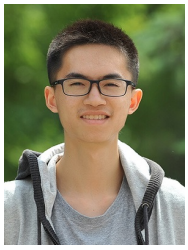
[53] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020.

[54] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating HDR transformation. In *Proc. of Computer Vision and Pattern Recognition*, pages 1798–1807, 2018.
[55] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. of International Conference on Computational Photography*, pages 1–10, 2015.
[56] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, and Xiuyi Jia. Ultra-high-definition image HDR reconstruction via collaborative bilateral learning. In *Proc. of International Conference on Computer Vision*, pages 4449–4458, 2021.
[57] Chu Zhou, Minggui Teng, Yufei Han, Chao Xu, and Boxin Shi. Learning to dehaze with polarization. In *Proc. of Advances in Neural Information Processing Systems*, 2021.
[58] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. UnModNet: Learning to unwrap a modulo image for high dynamic range imaging. In *Proc. of Advances in Neural Information Processing Systems*, 2020.
[59] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proc. of International Conference on Multimedia and Expo*, pages 1432–1437, 2019.



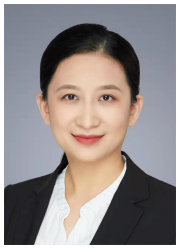
Chu Zhou is a PhD student in the School of Intelligence Science and Technology, Peking University. His research interests span event-based vision, polarization-based vision, and HDR imaging. He has served as a reviewer for CVPR, ICCV, ECCV, *etc*.

Yufei Han received a B.S. degree in Communication Engineering from the Beijing University of Posts and Telecommunications, Beijing, China in 2022. He is currently working toward an M.S. degree in Artificial Intelligence with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include polarization-based vision and HDR imaging.



Minggui Teng received the B.S. degree from Peking University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. His research interests are focused on neuromorphic camera and image enhancement.

Jin Han is currently a PhD candidate at the Graduate School of Information Science and Technology, the University of Tokyo. He received the B.Sc. degree in computer science from Sichuan University in 2018, and the M.Sc. degree in machine intelligence from Peking University in 2021. His research interests lie in neuromorphic cameras, event-based vision, and image restoration. He has served as a reviewer for CVPR, ICCV, ECCV, IJCV, TMM, *etc*.



Si Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2012. She is currently an associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her current research interests include multimodal artificial intelligence and machine learning.

Chao Xu received the B.E. degree from Tsinghua University, Beijing, China, in 1988, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 1991, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 1997. From 1991 to 1994, he was an Assistant Professor with the University of Science and Technology of China. Since 1997, he has been with the School of Electronics Engineering and Computer Science (EECS), Peking University, Beijing, where he is currently a Professor. His research interests are in image and video coding, processing, and understanding. He has authored or coauthored more than publications and five patents in these fields.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He has served as an editorial board member of IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.

Supplementary Material: Polarization Guided HDR Reconstruction via Pixel-Wise Depolarization

Chu Zhou, Yufei Han, Minggui Teng, Jin Han, Si Li, Chao Xu, and Boxin Shi*, *Senior Member, IEEE*

VIII. Discussions about the noise model

In this section, we provide discussions about the noise model, corresponding to Footnote 5 in Section III-A of the paper.

In our work, we follow the way of Ting *et al*. [16] (the only existing polarization-based HDR reconstruction method using deep learning) to adopt the Gaussian noise model to generate the dataset, which is also adopted in many other works (*e.g*., single-image HDR reconstruction methods [4], [5] and unconventional camera-based HDR reconstruction methods [6], [21]). To evaluate how the noise model can affect the results, we further consider the advanced noise model proposed by Konnik *et al*. [8] and Aguerrebere *et al*. [1]. We will have an brief introduction on it in the following.

Briefly, the electronic imaging pipeline could be written as three stages:

(1) photons to electrons;
(2) electrons to voltage;
(3) voltage to digital numbers.

These stages could produce different kinds of noise patterns, which can be simulated by the method proposed by Wei *et al*. [17]:

(1) the noise produced in the first stage can be described as the photon shot noise N_p, which can be simulated using a Poisson distribution $(I + N_p) \sim \mathcal{P}(I)$ (where I is the number of photoelectrons that is proportional to the scene irradiation);
(2) the noise produced in the second stage can be described as two terms including the read noise N_{read} and row noise N_r, which can be simulated using a Gaussian distribution;
(3) the noise produced in the third stage can be described as the quantization noise N_q, which can be simulated using a uniform distribution $N_q \sim U(-1/2q, 1/2q)$ (where q is the quantization step).

Chu Zhou and Chao Xu are with the Key Laboratory of Machine Perception, School of Intelligence Science and Technology, Peking University, Beijing 100080, China (e-mail: zhou_chu@pku.edu.cn; xuchao@cis.pku.edu.cn).

Minggui Teng and Boxin Shi are with the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100080, China. (e-mail: minggui_teng@pku.edu.cn; shiboxin@pku.edu.cn).

Yufei Han and Si Li are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: hanyufei@bupt.edu.cn; lisi@bupt.edu.cn).

Jin Han is with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: jinhan@nii.ac.jp).

* Corresponding author.

To summarize, the noise formation model consists of four major noise components: $N = KN_p + N_{read} + N_r + N_q$ (where K denotes the overall system gain), and the resulting noisy digital sensor raw image D can be written as $D = KI + N$.

Based on the analysis above, we regenerate the dataset adopting the above mentioned noise model and retrain our network. The quantitative evaluation results become: 0.990 (SSIM), 0.996 (MS-SSIM), 42.37 (PSNR), 65.45 (Q-Score). Compared to our original results (0.991 (SSIM), 0.997 (MS-SSIM), 43.12 (PSNR), 66.30 (Q-Score)), we can see that the results do not change too much. This is because the images are not captured in low-light conditions (*i.e*., the signal-to-noise ratio (SNR) of images is relatively high) so that the influence of noise model is not that significant. Therefore, adopting the Gaussian noise model could be practicable in our work, as many other HDR reconstruction works [4]–[6], [16], [21] do.

IX. Details about dataset generation

In this section, we explain how to generate our synthetic dataset in detail, corresponding to Footnote 13 in Section V-A of the paper.

With the ground truth scene radiance, DoP, and AoP (denoted as $\mathbf{E}_{\text{gt}} \in [0, 1]$, $\mathbf{p}_{\text{gt}} \in [0, 1]$, and $\boldsymbol{\theta}_{\text{gt}} \in [0, 180^\circ]$ respectively) calculated from EdPolCommunity Dataset [16], we first generate an appropriate exposure time t to re-expose $\mathbf{E}_{\text{gt}}$ to get the ground truth HDR image $\mathbf{H}_{\text{gt}}$ by

$$\mathbf{H}_{\text{gt}} = \mathbf{E}_{\text{gt}} \cdot t, \tag{37}$$

and use

$$\begin{cases} \mathbf{H}_{3_{\text{gt}},1_{\text{gt}}} = \dfrac{1}{2}\mathbf{H}_{\text{gt}} \cdot (1 \pm \mathbf{p}_{\text{gt}} \cdot \cos(2\boldsymbol{\theta}_{\text{gt}})) \\ \mathbf{H}_{4_{\text{gt}},2_{\text{gt}}} = \dfrac{1}{2}\mathbf{H}_{\text{gt}} \cdot (1 \pm \mathbf{p}_{\text{gt}} \cdot \sin(2\boldsymbol{\theta}_{\text{gt}})) \end{cases} \tag{38}$$

to compute four ground truth polarized HDR images $\mathbf{H}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ at polarizer angles $\alpha_{1,2,3,4} = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ respectively. Then, we clip $\mathbf{H}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ to obtain four ground truth unquantized polarized LDR images $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ by

$$\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}} = \min(\mathbf{H}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}, 1), \tag{39}$$

and quantize $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ to acquire four 8-bit polarized LDR images $\mathbf{L}_{1,2,3,4}$ by

$$\mathbf{L}_{1,2,3,4} = \left\lfloor 255(\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}) + \boldsymbol{\epsilon}) \right\rfloor / 255, \tag{40}$$

where $\boldsymbol{\epsilon}$ is a noise term which is set to be 2% Gaussian noise. And we could obtain the input ($\mathbf{L}^{\text{cat}}$) and target ($\mathbf{I}^{\text{cat}}_{\text{gt}}$)

of subnetwork g_1 by concatenating $\mathbf{L}_{1,2,3,4}$ and $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ along their channel index respectively. As $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ become available, the input and target of subnetworks g_2 and g_3 could be computed easily from them as follows:

- For subnetwork g_2:
  a. The input of subnetwork g_2 includes $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$, $\hat{\mathbf{p}}$, and $\hat{\boldsymbol{\theta}}$. Since $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ have already been computed, we only need to compute $\hat{\mathbf{p}}$ and $\hat{\boldsymbol{\theta}}$. We first solve a linear system:

$$\mathbf{I}_{i_{\text{gt}}} = \begin{bmatrix} \frac{1}{2} & \frac{-\cos(2\alpha_i)}{2} & \frac{-\sin(2\alpha_i)}{2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}} \\ \hat{\mathbf{D}}_c \\ \hat{\mathbf{D}}_s \end{bmatrix}, \quad (41)$$

  where $\hat{\mathbf{H}}$, $\hat{\mathbf{D}}_c$, and $\hat{\mathbf{D}}_s$ are three unknowns to be solved. Note that for pixels which have not been clipped in three of $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$, we use the three unclipped images to solve the linear system, while for other pixels we use all of $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ to solve it (by applying the least-squares solution). Then, we compute $\hat{\mathbf{p}}$ and $\hat{\boldsymbol{\theta}}$ using

$$\hat{\mathbf{p}} = \frac{\sqrt{\hat{\mathbf{D}}_c^2 + \hat{\mathbf{D}}_s^2}}{\hat{\mathbf{H}}} \quad \text{and} \quad \hat{\boldsymbol{\theta}} = \frac{1}{2}\arctan(\frac{\hat{\mathbf{D}}_s}{\hat{\mathbf{D}}_c}). \quad (42)$$

  b. The target of subnetwork g_2 includes $\mathbf{p}_{\text{gt}}$ and $\boldsymbol{\theta}_{\text{gt}}$, which have already been calculated.
- For subnetwork g_3:
  a. The input of subnetwork g_3 includes $\mathbf{p}_{\text{gt}}$, $\boldsymbol{\theta}_{\text{gt}}$, and $\hat{\mathbf{H}}^{(2)}$. Since $\mathbf{p}_{\text{gt}}$, and $\boldsymbol{\theta}_{\text{gt}}$ have already been calculated, we only need to compute $\hat{\mathbf{H}}^{(2)}$ by

$$\hat{\mathbf{H}}^{(2)} = \begin{cases} \hat{\mathbf{H}} & \text{for pixels which have been clipped} \\ & \text{in all of } \mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}} \\ \mathbf{H}_{\text{gt}} & \text{for other pixels} \end{cases}, \quad (43)$$

  where $\hat{\mathbf{H}}$ is obtained by solving Equation (41) using all of $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$.
  b. The target of subnetwork g_3 is $\mathbf{H}_{\text{gt}}$, which could be obtained from Equation (37).

We explain how to find an appropriate exposure time t. Defining the ill-posedness rate r as the proportion of pixels where $\mathbf{H}_{\text{gt}}$ could be compute in a well-posed manner using $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$, *i.e.*, the proportion of pixels whose values satisfy

$$[(\mathbf{H}_{1_{\text{gt}}} \geq \tau) \vee (\mathbf{H}_{3_{\text{gt}}} \geq \tau)] \wedge [(\mathbf{H}_{2_{\text{gt}}} \geq \tau) \vee (\mathbf{H}_{4_{\text{gt}}} \geq \tau)]. \quad (44)$$

Our goal is to control the ill-posedness rate to $[2.5\%, 15\%]$ using t, by assuming people usually do not take or save photos that contain too many saturated pixels. To this end, we apply a binary search to find an appropriate t, as shown in Algorithm 1. The upper limit m of the number of iterations is set to 15 in our work.

Note that the computed input and target values from $\mathbf{I}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}}$ are only for separate training (training phase 1). As for finetuning in an end-to-end manner (training phase 2), we use the output of former subnetwork to compute the input of later subnetwork.

Algorithm 1: The algorithm of finding an appropriate exposure time t

Input:
The ground truth HDR image $\mathbf{H}_{\text{gt}}$, DoP $\mathbf{p}_{\text{gt}}$, and AoP $\boldsymbol{\theta}_{\text{gt}}$;
The given range $[a, b]$ of the ill-posedness rate;
The upper limit m of the number of iterations.
Output:
An appropriate exposure time t.

$n \leftarrow 0$;
$f_m \leftarrow \max(\frac{1}{2}(1 \pm \mathbf{p}_{\text{gt}} \cdot \cos(2\boldsymbol{\theta}_{\text{gt}})), \frac{1}{2}(1 \pm \mathbf{p}_{\text{gt}} \cdot \sin(2\boldsymbol{\theta}_{\text{gt}})))$;
$l \leftarrow \frac{1}{f_m}$;
$u \leftarrow \frac{100}{f_m}$;
while $n < m$ **do**
  if $n = 0$ **then**
    $t \leftarrow \texttt{rand}([l, u])$;
  end
  else
    $t \leftarrow \frac{l+u}{2}$;
  end
  $\mathbf{H}_{3_{\text{gt}},1_{\text{gt}}} = \frac{1}{2}\mathbf{H}_{\text{gt}} \cdot (1 \pm \mathbf{p}_{\text{gt}} \cdot \cos(2\boldsymbol{\theta}_{\text{gt}}))$;
  $\mathbf{H}_{4_{\text{gt}},2_{\text{gt}}} = \frac{1}{2}\mathbf{H}_{\text{gt}} \cdot (1 \pm \mathbf{p}_{\text{gt}} \cdot \sin(2\boldsymbol{\theta}_{\text{gt}}))$;
  $r \leftarrow \texttt{ill\_posedness\_rate}(\mathbf{H}_{1_{\text{gt}},2_{\text{gt}},3_{\text{gt}},4_{\text{gt}}})$;
  if $r > b$ **then**
    $u \leftarrow t$;
  end
  else if $r < a$ **then**
    $l \leftarrow t$;
  end
  else
    break;
  end
  $n \leftarrow n + 1$;
end
if $n = m$ **then**
  return;
end
else
  return t;
end

X. The variances of quantitative evaluation results

In this section, we provide the variances (measured using the standard deviation σ) of the quantitative evaluation results, as shown in Table III.

XI. Special conditions

In this section, we analyze three special conditions, including:

(1) dark objects;
(2) completely unpolarized scenes;
(3) highly unpolarized scenes.

TABLE III
THE VARIANCES (MEASURED USING THE STANDARD DEVIATION σ) OF THE QUANTITATIVE EVALUATION RESULTS ON SYNTHETIC DATA AMONG OUR POL-HDR NETWORK, TWO EXISTING POLARIZATION-BASED HDR RECONSTRUCTION METHODS (WU *et al*. [19] AND TING *et al*. [16]), AND FOUR STATE-OF-THE-ART LEARNING-BASED SINGLE-IMAGE HDR RECONSTRUCTION METHODS (EILERTSEN *et al*. [4], ENDO *et al*. [5], LIU *et al*. [9], AND SANTOS *et al*. [14]). THE LOWER THE METRICS, THE BETTER THE RESULTS. **BOLD** FONT INDICATES THE BEST PERFORMANCE.

	Wu *et al*. [19]	Ting *et al*. [16]	Eilertsen *et al*. [4]	Endo *et al*. [5]	Liu *et al*. [9]	Santos *et al*. [14]	Ours
σ_{SSIM}	0.013	0.061	0.013	0.072	0.018	0.015	**0.007**
$\sigma_{\text{MS-SSIM}}$	0.013	0.072	0.014	0.094	0.016	0.010	**0.004**
σ_{PSNR}	5.69	5.01	6.52	5.27	5.17	4.98	**4.19**
$\sigma_{\text{Q-Score [11]}}$	7.70	**4.86**	8.48	7.87	6.39	7.97	7.36

As for dark objects, since the DoP and AoP may be falsely calculated due to the noise [15], it may indeed bring negative effects to the physics modules. However, inaccurate DoP and AoP in dark regions may not have much impact on the HDR reconstruction process. This is because dark objects usually have small pixel values so that they are often unsaturated in the captured polarized images (*i.e*., they often belong to H-Class1 ($\mathbb{X}_1^H$)). In this situation, we can directly obtain the corresponding HDR pixels from the captured polarized images without computing the DoP and AoP (see the analysis of H-Class1 in Section III-B of the paper for details). Even if there are still some dark objects whose pixels do not belong to H-Class1, our network modules (*e.g*., g_1 can reduce the noise level in the image domain, and g_2 can repair the inaccurate values of the calculated DoP and AoP) could alleviate this problem to some extent. In our experiments (Figure 4 and Figure 5 of the paper) we can see that the visual artifacts caused by this problem are not obvious.

As for completely unpolarized scenes, the captured polarized images would have almost the same pixel values. In such a situation, the DoP tends to be zero, which degenerates the original polarization guided HDR reconstruction problem into a single-image one. The physics modules cannot provide any help, and the dynamic range gain completely depends on the network modules. However, since the proportion of pixels with small DoPs is not large in our dataset (see Footnote 6 of the paper), we adopt the assumption that all pixels tend to have non-zero DoPs in our daily photography, as Wu *et al*. [19] do.

As for highly unpolarized scenes, the DoP could be close to 1. In such a situation, the potential dynamic range gain could be very large (see "Analysis of the potential dynamic range gain" in Section III-A of the paper), making the HDR reconstruction process easier, not harder.

XII. RECONSTRUCTION RESULTS OF THE SCENES CONTAINING VERY BRIGHT HIGHLIGHTS

In this section, we provide some reconstruction results of the scenes containing very bright highlights, as shown in Figure 6. We can see that our method still obtains acceptable reconstruction results, because these highlights are caused by specular reflection so that they are significantly polarized, which could be suppressed by polarizers at certain angles.

XIII. COMPARISON WITH CODED EXPOSURE

Since the structure of four-directional on-chip micro-polarizers inside the polarization camera is similar to an optical mask with 2×2 patterns, we make a comparison

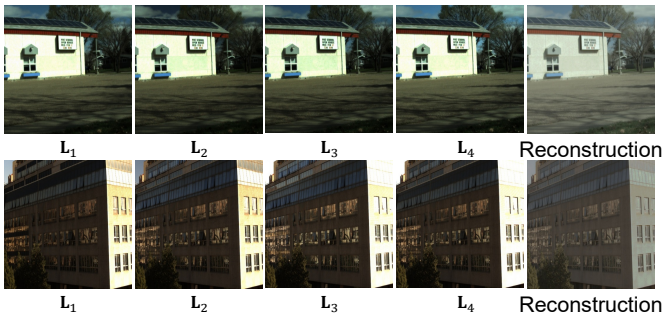


Fig. 6. Reconstruction results of the scenes containing very bright highlights.

with a coded exposure method SVE [12] that adopts an optical mask with repeated 2×2 patterns specially designed for HDR reconstruction. The output of SVE [12] can be regarded as four images with different exposures $e_3 = 4e_2 = 16e_1 = 64e_0$ generated by four neighboring cells with different optical transparencies in each pattern, so that the HDR images can be easily obtained using a multi-image HDR reconstruction method [3]. Theoretically, according to Equation (12) in the main paper the dynamic range of SVE [12] can reach 84.25 dB in the whole image plane, while according to Equation (13) in the main paper the dynamic range of a polarization camera could be spatially-variant since it depends on the DoP and AoP (which are spatially-variant). Assuming a uniform distribution of the AoP, the theoretical dynamic range of a polarization camera could outperform SVE [12] only when the average DoP is very high (*i.e*., large than 0.994). According to the above analysis, SVE [12] seems to be more suitable for HDR reconstruction than using a polarization camera, considering that the scenes could not always be highly polarized. However, since the optical mask used in SVE [12] would result in a significantly lower SNR for the sensor, the actual dynamic range of SVE [12] could be much lower than the theoretical value. Besides, SVE [12] is a prototype specially designed for HDR reconstruction and it has not been manufactured (only a patent [13] is available), while the polarization camera is already available in the market (*e.g*., Lucid and FLIR polarization cameras, as mentioned in Footnote 1 and Footnote 12 of the paper) and it has a broad range of applications in the field of computational photography (*e.g*., shape from polarization [2], reflection removal [10], image dehazing [20], *etc*.) and factory automation (*e.g*., transparent object segmentation [7]). Therefore, it is interesting to explore the extra ability of polarization cameras to achieve ghost-free single-shot HDR reconstruction.

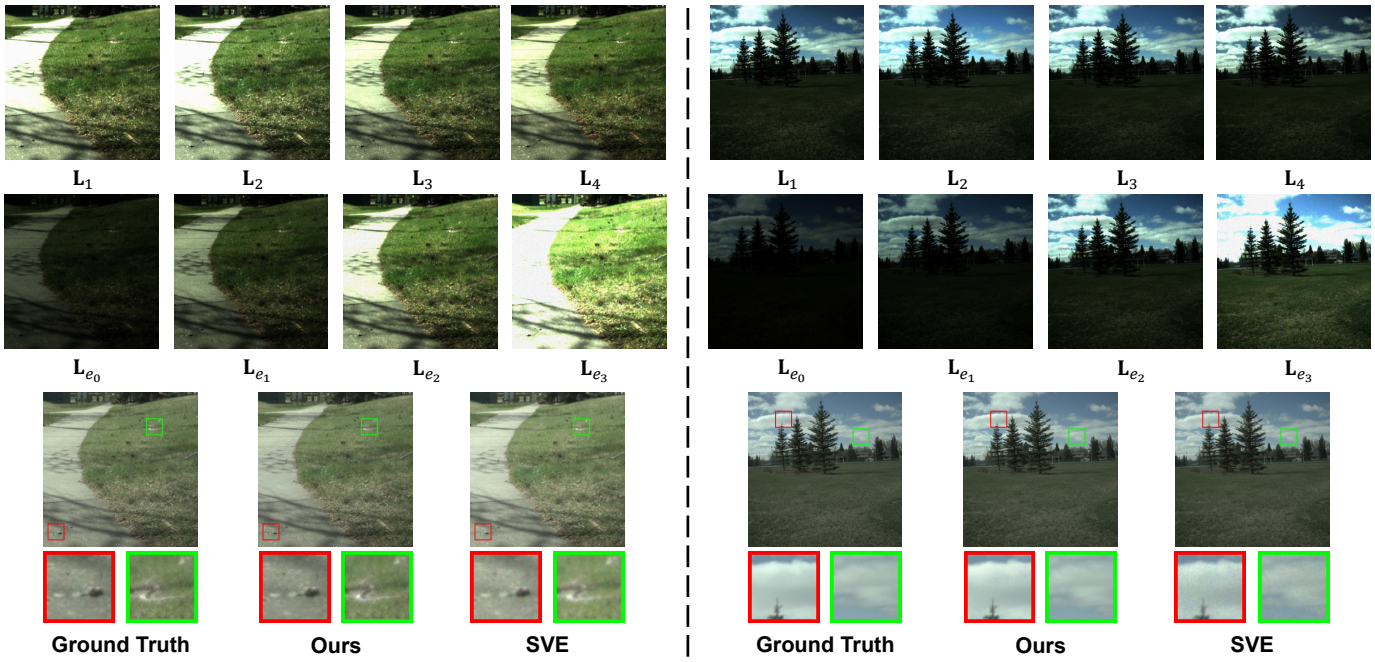


Fig. 7. Qualitative comparisons with SVE [12] (a coded exposure method adopting an optical mask with repeated 2×2 patterns to obtain four different exposures). $\mathbf{L}_{e_{3,2,1,0}}$ denote the images with four different exposures $e_3 = 4e_2 = 16e_1 = 64e_0$ generated by SVE [12].

TABLE IV
COMPARISON WITH A CODED EXPOSURE METHOD SVE [12] THAT ADOPTS AN OPTICAL MASK WITH REPEATED 2×2 PATTERNS TO OBTAIN FOUR DIFFERENT EXPOSURES $e_3 = 4e_2 = 16e_1 = 64e_0$. Δ DENOTES THE GAP BETWEEN THEM. THE HIGHER THE METRICS, THE BETTER THE RESULTS. **BOLD** FONT INDICATES THE BEST.

	Ours	SVE [12]
SSIM	**0.991**	0.931
MS-SSIM	**0.997**	0.989
PSNR	**43.12**	35.83
Q-Score [11]	**66.30**	65.56

We choose to perform simulations to reproduce the concept of SVE [12] since there is no off-the-shelf device available in the market. Quantitative results are shown in Table IV. We can see that our method outperforms SVE [12] on all metrics. This is because SVE [12] suffers from the noise caused by the use of optical mask, while our method can improve the performance of polarization cameras in HDR reconstruction, by virtue of the proposed network-physics-hybrid Pol-HDR pipeline that adopts both physical constraints and learned features to fully exploit the polarization information. Visual quality comparisons are shown in Figure 7. We can see that our method can recover finer details than SVE [12], especially for scenes that contain highly polarized objects such as the specular highlights on the ground (in the left group) and the sky regions (in the right group).

XIV. ADDITIONAL SYNTHETIC RESULTS

In this section, we provide additional comparisons on synthetic data among our Pol-HDR network, two existing polarization-based HDR reconstruction methods (Wu *et al*. [19] and Ting *et al*. [16]), and four state-of-the-art learning-based single-image HDR reconstruction methods (Eilertsen *et al*. [4], Endo *et al*. [5], Liu *et al*. [9], and Santos *et al*. [14]), as shown in Figure 8, corresponding to Footnote 16 in Section VI-A of the paper.

XV. ADDITIONAL REAL RESULTS

In this section, we provide additional comparisons on real data among our Pol-HDR network, two existing polarization-based HDR reconstruction methods (Wu *et al*. [19] and Ting *et al*. [16]), and four state-of-the-art learning-based single-image HDR reconstruction methods (Eilertsen *et al*. [4], Endo *et al*. [5], Liu *et al*. [9], and Santos *et al*. [14]), as shown in Figure 9, corresponding to Footnote 19 in Section VI-B of the paper.

XVI. LIMITATIONS

Since our method relies on polarization-related physical parameters (DoP $\mathbf{p}$ and AoP $\boldsymbol{\theta}$), once the proportion of pixels in Pol-Class2 ($\mathbb{X}_2^{\text{Pol}}$) accounts for the vast majority, estimating them becomes more ill-posed, leading to inaccurate HDR reconstruction. As our future work, we plan to propose a more sophisticated subnetwork for DoP and AoP estimation to handle such a situation more adequately. Furthermore, if all pixels have zero DoPs, the captured polarized LDR images would tend to be the same, which degenerates the original polarization guided HDR reconstruction problem into a single-image one. And this problem could be tackled by inverse tone-mapping approaches [4], [5], [9], [14]. Moreover, since RGB polarization cameras require joint chromatic and polarimetric demosaicing [18], the spatial resolution they could achieve is often lower than that of digital cameras; besides, since a polarizer attenuates scene radiance, the signal-to-noise ratio of a polarization camera is often lower than that of digital cameras; in addition, since the real micro-polarizers inside a polarization camera are not

Fig. 8. Additional comparisons on synthetic data.

perfect, their extinction ratio is not very large (often about several hundred), which would limit the maximum value of the dynamic range gain. How to dealing with those defects caused by the currently available design of polarization cameras is beyond the scope of this paper. Another limitation is that our method can only achieve HDR image reconstruction now, and

Fig. 9. Additional comparisons on real data.

we aim to extend it for HDR video reconstruction in the future.

REFERENCES

[1] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. Study of the digital camera acquisition process and statistical modeling of the sensor raw data. 2013.

[2] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proc. of European Conference on Computer Vision*, 2020.

[3] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, 1997.

[4] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 36(6):178:1–178:15, 2017.

[5] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 36(6):177:1–177:10, 2017.

[6] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proc. of Computer Vision and Pattern Recognition*, pages 1730–1739, 2020.

[7] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proc. of Computer Vision and Pattern Recognition*, 2020.

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

[9] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proc. of Computer Vision and Pattern Recognition*, pages 1651–1660, 2020.

[10] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *Proc. of Advances in Neural Information Processing Systems*, 2019.

[11] Manish Narwaria, Rafal Mantiuk, Mattheiu P Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501, 2015.

[12] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, pages 472–479, 2000.

[13] Shree K Nayar and Tomoo Mitsunaga. Apparatus and method for high dynamic range imaging using spatially varying exposures. *US Patent 7,924,321*, 2011.

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

[15] AB Tibbs, IM Daly, DR Bull, and NW Roberts. Noise creates polarization artefacts. *Bioinspiration & biomimetics*, 13(1):015005, 2017.

[16] Juiwen Ting, Xuesong Wu, Kangkang Hu, and Hong Zhang. Deep snapshot HDR reconstruction based on the polarization camera. In *Proc. of International Conference on Image Processing*, 2021.

[17] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proc. of Computer Vision and Pattern Recognition*, pages 2758–2767, 2020.

[18] Sijia Wen, Yinqiang Zheng, and Feng Lu. A sparse representation based joint demosaicing method for single-chip polarized color sensor. *IEEE Transactions on Image Processing*, 30:4171–4182, 2021.

[19] Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, and Juiwen Ting. HDR reconstruction based on the polarization camera. *IEEE Robotics and Automation Letters*, 5(4):5113–5119, 2020.

[20] Chu Zhou, Minggui Teng, Yufei Han, Chao Xu, and Boxin Shi. Learning to dehaze with polarization. In *Proc. of Advances in Neural Information Processing Systems*, 2021.

[21] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. UnModNet: Learning to unwrap a modulo image for high dynamic range imaging. In *Proc. of Advances in Neural Information Processing Systems*, 2020.