RawHDR: High Dynamic Range Image Reconstruction from a Single Raw Image

Yunhao Zou[1] Chenggang Yan[2] Ying Fu[1*]
[1]Beijing Institute of Technology [2]Hangzhou Dianzi University

Abstract

High dynamic range (HDR) images capture much more intensity levels than standard ones. Current methods predominantly generate HDR images from 8-bit low dynamic range (LDR) sRGB images that have been degraded by the camera processing pipeline. However, it becomes a formidable task to retrieve extremely high dynamic range scenes from such limited bit-depth data. Unlike existing methods, the core idea of this work is to incorporate more informative Raw sensor data to generate HDR images, aiming to recover scene information in hard regions (the darkest and brightest areas of an HDR scene). To this end, we propose a model tailor-made for Raw images, harnessing the unique features of Raw data to facilitate the Raw-to-HDR mapping. Specifically, we learn exposure masks to separate the hard and easy regions of a high dynamic scene. Then, we introduce two important guidances, dual intensity guidance, which guides less informative channels with more informative ones, and global spatial guidance, which extrapolates scene specifics over an extended spatial domain. To verify our Raw-to-HDR approach, we collect a large Raw/HDR paired dataset for both training and testing. Our empirical evaluations validate the superiority of the proposed Raw-to-HDR reconstruction model, as well as our newly captured dataset in the experiments.

1. Introduction

The dynamic range of real-world scenes often surpasses the recording capability of standard consumer camera sensors, leading images to lose details in both over- and under-exposed regions [6]. To endow today's digital photos with the capacity to contain more scene information, the technique called high dynamic range (HDR) that records data with a wide range of intensity levels has been extensively explored in the computational imaging community [45]. Compared with conventional low dynamic range (LDR) images, HDR retains more details in over- and under-exposure regions. Thus, HDR benefits downstream vision tasks including segmentation [33], object detection [36], and also provide more aesthetically appealing pictures [19, 25], which computer vision researchers have longly pursued.

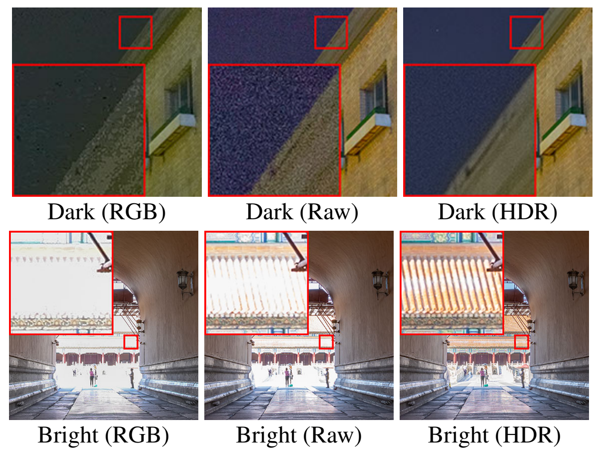


Figure 1: The RGB/Raw/HDR images of the darkest (first row) and brightest (second row) regions in a high dynamic scene. This paper is motivated by two observations: (1) HDR scenes contain both extremely dark and bright regions, which are very challenging to reconstruct from a single image; (2) Raw images contain much more information in these hard regions, compared to low-bit RGB images.

Methods to obtain HDR data broadly fall into three categories, *i.e.*, reconstruction from multi-exposure, single-exposure images, and novel camera sensors. Among new sensors, some noteworthy examples include HDR cameras [34, 41], event cameras [18, 40, 44, 58], and infrared sensors [29]. Since these sensors are all specialized devices, more works focus on the reconstruction from multi-/single-exposures captured by commercial cameras, which are more practical and commercially friendly.

Historically, considering that single exposure cannot record the intensity information of a scene that covers a large dynamic range, researchers often combined multi-exposure images to generate HDR content [10, 14, 19, 23, 52, 53]. The quality of these images greatly depend on align-

*Corresponding Author: fuying@bit.edu.cn

ing different exposures, and may suffer from ghosting effect caused by imperfect alignment.

To avoid the potential risk of alignment failure, more recent works incorporate only a single LDR image to reconstruct an HDR image [9, 13, 30, 50]. Nonetheless, single-image HDR reconstruction is more challenging due to the physical limitation of consumer camera's dynamic range. Consequently, under-exposed regions are often noisy [7, 8, 16, 54], while over-exposed regions are difficult to recover [2, 22]. Most previous works utilize low-bit sRGB images for HDR reconstruction. However, sRGB images have been degraded by lossy in-camera operations, which is not enough to record details in an HDR scene. Even as innovative and interpretable deep models emerge [13, 30], the HDR reconstructions from single LDR images often falter in highly dynamic scenes due to the inherent paucity of low-bit sRGB input.

In this work, we aim to relieve the information limitation for single-exposure setting with a specialized reconstruction model and high-quality dataset. Using a single image, the challenge comes down to the recovery of the darkest and brightest regions (*hard* regions) in high dynamic scenes. The visualization in Figure 1 shows that commonly used sRGB images contain limited information in hard regions. Raw images retain more details than sRGB, but are still far from HDR. Therefore, we propose using unprocessed Raw sensor data, which has higher available bit-depth and better intensity tolerance, thus can circumvent the long-standing drawback of insufficient scene information. To perform specific operations on hard regions, we learn an exposure mask to adaptively separate the over-/under- and well-illuminated regions for each scene. We also devise a deep neural network specially designed for Raw input images to exploit the information in hard regions. Crucially, we propose a dual intensity guidance based on the channel-variant attribute of Raw images to guide less informative image channels and global spatial guidance to well exploit longer-range information. Finally, we collect a high-quality Raw/HDR paired dataset for both training and testing. The quality of our Raw-to-HDR reconstruction methods and dataset are verified in the experiments.

Our main contributions can be summarized as follows:

1. We focus on the essential issue of HDR imaging — the challenge in recovering the dark and bright regions, for which we propose to learn an exposure mask to separate the image into hard and easy regions.

2. We propose a deep network to deal with the hard regions, including a dual intensity guidance built on the channel-variant attribute of Raw images and a global spatial guidance built on transformer with spatial attention that exploits information from a longer range.

3. We directly reconstruct HDR from a single Raw image, which is endowed with higher bit-depth to handle high dynamic scenes and can be potentially integrated into modern camera processing pipelines. In addition, we collect a high-quality paired Raw/HDR dataset for training and evaluation.

2. Related Work

In this section, we review the most relevant works, including multi-image HDR reconstruction and single-image HDR reconstruction.

Multi-image HDR reconstruction. In early years, a number of researchers reconstruct HDR images by merging a series of bracketed exposure LDR images [11]. For multi-exposure HDR imaging, the step to align multiple images is crucial yet challenging. Numerous methods [23, 38, 53] align bracket multi-exposures before fusing. Given a sequence of LDR images with different exposures, Kalantari *et al*. [23] regarded the medium exposure as reference and align low- and high-exposure images to the reference by flow warping. Then they learned the mapping from aligned LDR to HDR through a deep neural network. Peng *et al*. [38] investigated the potential of advanced optical flow estimation techniques, such as FlowNet [12], to refine alignment before HDR reconstruction. In addition to fusing different exposures, HDR+ [19] fused burst images under the same exposure and claimed that it was easier to align images with the same exposure. However, all these methods would encounter difficulties in alignment in very high dynamic scenes. Therefore, multi-exposure HDR reconstruction is mainly used in situations without significant motion.

Single-image HDR reconstruction. Given the inherent challenges with multi-exposure HDR, especially the difficulty in LDR alignment, some researchers have gravitated towards deriving HDR from a single image. However, the recording capacity of a single image makes it even more ill-posed. Traditional single-exposure HDR works proposed to expand the dynamic range by estimating the density of light sources [4, 5]. As recent deep learning-based tools, especially CNNs, have made a great breakthrough in computer vision, CNN-based methods [21, 30] were presented to directly learn to reconstruct HDR from a single LDR image. HDRCNN [13] and ExpandNet [32] proposed to directly learn the mapping from LDR-to-HDR in an end-to-end manner. HDRUNet [9] and SingHDR [30] designed the network architecture following the physics formation model of LDR images. Other methods [27] first synthesize pseudo-multi-exposure from a single exposure, then fuse these generated multi-exposures. All of these single-exposure HDR methods [4, 5, 13, 21, 27, 30] tried to improve HDR quality from the perspective of designing a better reconstruction algorithm. However, using sRGB LDR images as input prevents them from obtaining higher-quality HDR. In fact, during the intricate processing pipeline to generate sRGB, there are lossy and invertible in-camera operations including nonlinearization, clipping, compression and quantization, which degrade the original Raw images (generally 14-bit) to lower bit-depth images (generally 8-bit). Therefore, we believe that the linearity and high bit-depth attributes of untouched Raw images are ideal for single-image HDR reconstruction. In this paper, we reconstruct HDR images directly from a single Raw image, with customized network design and specially captured paired dataset.

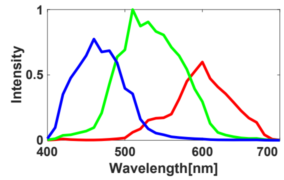
3. Method

This section demonstrate our primary motivations, problem formulation for Raw-to-HDR reconstruction, and the network architecture of our proposed RawHDR model. The overarching methodology is depicted in Figure 3.

3.1. Motivation

Existing sing-exposure HDR reconstruction methods [9, 13, 30] mainly focus on the reconstruction from low-bit sRGB images. Despite their ability to improve the quality beyond the original sRGB images, they cannot handle hard regions in extremely high dynamic scenes. This limitation lies in the irreversible and lossy operations in the in-camera signal processing workflow, including nonlinearization, clipping, compression and quantization [24, 49, 57]. Furthermore, these methods target at enhancing previously captured LDR images rather than optimizing the image capture process itself. Motivated by these shortcomings, we plan to reconstruct HDR images in higher dynamic scenes with a novel reconstruction method and new data setting. To resolve the problem of information loss and insufficient details in commonly used sRGB images, we directly utilize the unprocessed Raw data to reconstruct HDR images, as Raw images have higher bit-depth to maintain a large dynamic range of a scene. Moreover, since we directly perform HDR mapping on Raw images that are unprocessed, the operation can be potentially integrated into today's camera processing pipeline to facilitate the imaging process.

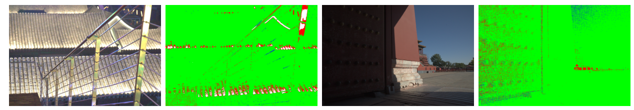
Then, we carefully analyze the special attribute of Raw images to present a deep model tailored for Raw-to-HDR mapping. Notably, the intensity values vary with image channels in Raw space [28]. We provide the camera response function (CRF) of a typical consumer camera (Canon 5D II) in Figure 2(a). It illustrates that the integration of green spectral curve is considerably larger than red and blue, which indicates higher sensitivity of green patterns. Consequently, green channels have larger values in Raw images. This conclusion is also verified in Figure 2(b), where the channel-wise mean values for our dataset (described in Section 4) is presented. Actually, apart from regions with a strong inclination towards red or blue, the pixel values in most areas tend to be higher for green. We give



(a) CRF

Channel	Mean value
Red	704.93
Green	1273.61
Blue	942.00

(b) Channel Mean



Scene 1 DCR 1 Scene 2 DCR 2

(c) Examples for dominant color channel (denoted as DCR).

Figure 2: Analysis for the channel-wise attributes of Raw images. (a) A typical camera response function; (b) The channel-wise mean values for the captured dataset; (c) Examples of two scenes, with the original captured images and the corresponding pixel-wise dominant color channels.

example scenes in Figure 2(c), which shows the predominant channel for each pixel, we can see that even the *blue sky* and *red wall* have larger green value in RAW space.

Therefore, green channels of Raw images are more likely to lose information due to the intensity upper bound. Red and blue channels face similar situations in poorly lit areas. Utilizing this attribute, we lean on the more informative channels to guide the others. Moreover, considering the severe information loss in hard areas (Figure 1), we introduce longer range feature exploitation models for the reconstruction of hard regions to compensate for insufficient information in these regions. In addition, since we plan the handle hard and easy regions separately, we propose to learn exposure masks to achieve this task.

3.2. Model Architecture

Given a Raw image R and a deep neural network model $f(\cdot, \theta)$, the process to reconstruct an HDR image H can be formulated as

$$H = f(R, \theta), \quad (1)$$

where θ is the network parameters. The overview of our RawHDR model is depicted in Figure 3. We first learn exposure masks through a mask estimation module, which separate the over-/under- and well-exposed areas. Then, on the basis of the intensity attribute of Raw images, we present dual intensity guidance to guide less informative channels with more informative ones. Finally, we use a transformer-based model to harness longer-range spatial features, which we call global spatial guidance.

Mask estimation. As we know, HDR images contain information details in both brightest and darkest areas, which cannot be well captured by ordinary LDR images. To employ customized operations on these regions, we need

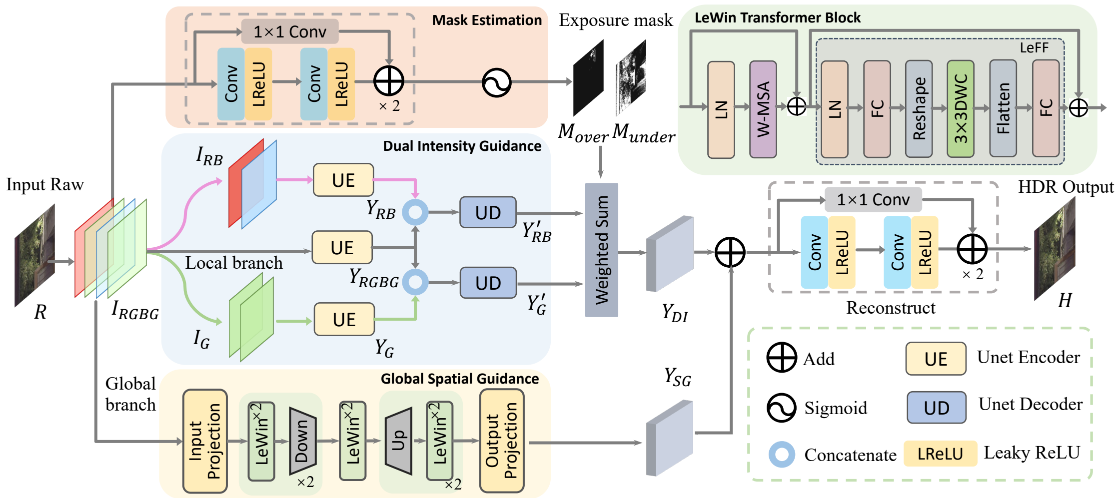


Figure 3: The overview of our HDR reconstruction method with dual intensity and global spatial guidance.

to find out the mask to separate the over-exposed, under-exposed, and well-illuminated regions of a scene. In this work, we use learnable masks to separate different regions, then perform different operations for these regions.

Previous works prefer to use manual-set thresholds to obtain the masks. For example, [30] classifies intensities surpassing 0.95 as over-exposed and those below 0.05 as under-exposed, subsequently generating a blending mask from these thresholds. Such masks, acquired through rigid thresholding, are denoted as M^h_{over} and M^h_{under}. However, such mask is too hard to separate the regions, potentially leading to undesirable edge artifacts in smooth areas. Therefore, we propose a gentler and more adaptive thresholding strategy. Here, we use deep neural networks to predict the exposure mask. Our mask estimation model incorporates local feature extraction, and the pixel-wise intensity value is not the only factor in deciding the exposure masks. The process to estimate the masks for over-, under-, and well-exposed regions (denoted as M_{over}, M_{under} and M_{well}) can be formulated as

$$\begin{aligned} M_{over} &= \mathcal{S}(P_o(R)), \\ M_{under} &= \mathcal{S}(P_u(R)), \\ M_{well} &= \max\{1 - M_{over} - M_{under}, 0\}, \end{aligned} \tag{2}$$

where $\mathcal{S}$ is the sigmoid function and P_o and P_u are networks to predict the over and under-exposure masks. In our experiment, we simply use two residual blocks [21] to implement P_o and P_u. In order to guarantee that M_{over} is close to 1 at high intensity values while M_{under} is close to 1 at low intensity values, we design a constraint loss $\mathcal{L}_{mask}$ which lead the mask to for mask learning

$$\mathcal{L}_{mask} = \|M_{over} - M^h_{over}\|_1 + \|M_{under} - M^h_{under}\|_1. \tag{3}$$

Dual intensity guidance. Modern cameras typically employ the Bayer pattern for image capture, where green sensors on the CMOS receive a greater luminous flux relative to other color sensors [56]. As a result, Raw images have higher intensity values in green channels compared to red and blue. Another notable observation is that different channels share similar edges [28]. Based on the two observations, existing works have already utilized this attribute of sensors by guiding red and blue channels with green channels in applications including denoising and demosaicing [17, 28, 56]. In this work, we present a *dual intensity guidance* module to recover missing details in hard regions of a high dynamic scene. Since green channels are more sensitive to light (Figure 2), it has higher signal-to-noise ratio (SNR) in under-exposed regions. With higher SNR, green channel can guide other channels during reconstruction since all channels share the same edges. In addition, in our HDR reconstruction case, over-exposure regions also need guidance, as these regions lose much information due to intensity clipping. In these regions, less sensitive channels, *i.e.*, red and blue, have smaller intensity values and are less likely to overexpose. So we use red and blue channels as guidance in over-exposed regions.

The architecture of our dual intensity guidance module is illustrated in Figure 3. Our method first packs Raw images with size $H \times W$ into four channels representing RGBG pixels in Bayer pattern, and obtains a $H/2 \times W/2 \times 4$ input image I_{RGBG}. Then, we extract the two green channels as under-exposure guiding image I_G, and the red/blue channels served as over-exposure guiding image, denoted as I_{RB}. Together with the packed input image I_{RGBG}, these three tensors are fed into U-net [42] encoders (UE) to separately obtain hidden features Y_G, Y_{RB}, and Y_{RGBG}. According to the analysis in Section 3.1, I_G and I_{RB} are less likely to saturate in under- and over-exposed regions. So we regard the features extracted by these two branches as guidance and concatenate them with Y_{RGBG} to incorporate more details in the darkest and brightest regions. Later, the concatenated tensors are decoded by a UNet decoder (UD), and the output of dual intensity guidance module is obtained by weighted summation of over and under-exposure guided features through an intensity mask. More concretely, we operate under-exposure mask M_{under} on the green channel guided feature and mask the red/blue channel guided feature with over-eposure mask M_{over}. The process can be formulated as

$$\begin{aligned} Y'_G &= D(\text{Concat}(Y_G, Y_{RGBG})), \\ Y'_{RB} &= D(\text{Concat}(Y_{RB}, Y_{RGBG})), \\ Y_{DI} &= M_{under} \odot Y'_G + M_{over} \odot Y'_{RB}, \end{aligned} \tag{4}$$

where $\odot$ denotes the Hadamard product, Y'_G and Y'_{RB} are the intensity guided features.

Global spatial guidance. Our dual intensity guidance module is designed from the perspective of channel-wise guidance. In other words, we guide some channels with more informative channels pixel-wisely. However, in more extreme regions, the most informative channel may still be far from sufficient details. Under these conditions, longer range features can be utilized since there may exist similar patches that can help the recovery of hard regions [51]. Therefore, besides channel-wise local guidance, we further present a spatial guidance branch that exploits long range features. Specifically, we use a series of transformer blocks to extract features, which we call global spatial guidance.

We incorporate a U-net like transformer structure [47] to build the global spatial guidance. As shown in Figure 3, the global spatial guidance consists of K stages, and each stage contains 2 Locally-enhanced Window (LeWin) transformer blocks [47]. Among stages, there are down- and up-sampling operations to obtain a large receptive field. The LeWin transformer block is built with attention operations to exploit the relationship between pixels. Based on the U-shape structure and attention mechanism, the global spatial guidance would find similar patches for hard regions from a global receptive field. These additional spatial features introduce more available details and facilitate the reconstruction of hard regions. In addition, the LeWin transformer blocks divide the feature maps into non-overlapping windows [31] and employ attention mechanism within each window. Though the attention for each block only focuses on a single window, the module still reaches an almost global receptive field due to the U-shape structure.

Figure 3 illustrates the structure of each LeWin block. For the i-th block, given the input feature F_{i-1}, the output feature F_i can be obtained by

$$\begin{aligned} F'_i &= \text{W-MSA}(\text{LN}(F_{i-1})) + F_{i-1}, \\ F_i &= \text{LeFF}(\text{LN}(F'_i)) + F'_i, \end{aligned} \tag{5}$$

where W-MSA denotes the window-based multi-head attention module [47], LeFF is the locally-enhanced feed-forward network, and LN is the layer normalization [3]. In LeFF, tokens are first processed by a linear projection, then reshaped to 2D feature maps. Next, a 3×3 depth-wise convolution layer is applied to extract spatial features, and the results are flattened back to tokens.

3.3. Learning Details

Considering that HDR output images have high bit-depth, normal L_1 or L_2 loss tends to be dominated by bright areas with extensive intensity values. To counteract this, we compute loss functions in the $\log$ space. Given the reconstructed HDR image H_t and the corresponding ground truth $\hat{H}_t$, we employ L_2 loss in the $\log$ space [30] to evaluate the reconstruction fidelity

$$\mathcal{L}_{rec} = \sum_{i=1}^{T} \| \log H_i - \log \hat{H}_i \|_2^2, \tag{6}$$

where T is the total number of training samples.

In addition to pixel-level metrics, we also adopt the Learned Perceptual Image Patch Similarity (LPIPS) loss [55] to ensure high-level structural similarity, denoted as $\mathcal{L}_{LPIPS}$. Incorporating this loss with the mask constraint, the complete loss function for becomes:

$$\mathcal{L} = \mathcal{L}_{rec} + \tau_1 \mathcal{L}_{LPIPS} + \tau_2 \mathcal{L}_{mask}. \tag{7}$$

For our experiments, we empirically set the weights τ_1, τ_2 to 0.5 each. During the training phase, we initialize our network using Kaiming initialization [20], and minimize the loss using the adaptive moment estimation method [26], with a momentum parameter to 0.9. The initial learning rate is 10^{-4}, which is subsequently divided by 10 at 1000-th epoch. Training is conducted with a batch size of 1 across 2000 epochs, using the PyTorch deep learning framework [37] on an NVIDIA Geforce 3090 GPU.

4. The Proposed Raw-to-HDR Dataset

Dataset is significant for current data-driven HDR imaging methods. To train the Raw-to-HDR model, we need a high-quality paired dataset to feed the deep neural networks. Existing datasets mainly contain HDR images or paired sRGB/HDR images [13,15,23,30,35,39]. Table 1 illustrates some recent and representative HDR datasets [45]. To the best of our knowledge, there is no high-quality paired Raw-to-HDR dataset that is designed for supervising single-exposure Raw-to-HDR mapping. Therefore, we build a new Raw-to-HDR paired dataset. We first carefully choose HDR scenes. Then, we use the bracket exposure mode to capture images. We fix the camera on a tripod to ensure that there is no vibration during the capture. Then, three exposure values are used to capture, including -3EV, 0EV, and +3EV. As a result, for each scene, we have three Raw images with varied exposures. Then, the Raw images at 0EV

Table 1: Summary of existing HDR dataset and the proposed Raw-to-HDR dataset.

Dataset	Year	Size	Resolution	Real/Syn.	Input	Output	Application
Froehlich *et al.* [15]	2014	15	1920×1080	Real	None	12-bit video	HDR video sequences
HDREye [35]	2015	46	1920 × 1080	Syn.	8-bit sRGB	16-bit HDR	Static indoor and outdoor scenes
Kalantari *et al.* [23]	2017	74	1500×1000	Real	14-bit Raw	16-bit HDR	Multi-exposure HDR
Liu *et al.* [30]	2020	N/A	1536×1024	Real+Syn.	8-bit sRGB	12-16-bit HDR	Single-exposure sRGB-to-HDR
NITIRE 2021 [39]	2021	1761	1920×1080	Syn.	8-bit sRGB	12-bit HDR	Curated from video sequences
Ours	2023	324	6720×4480	Real	14-bit Raw	20-bit HDR	Single-exposure Raw-to-HDR

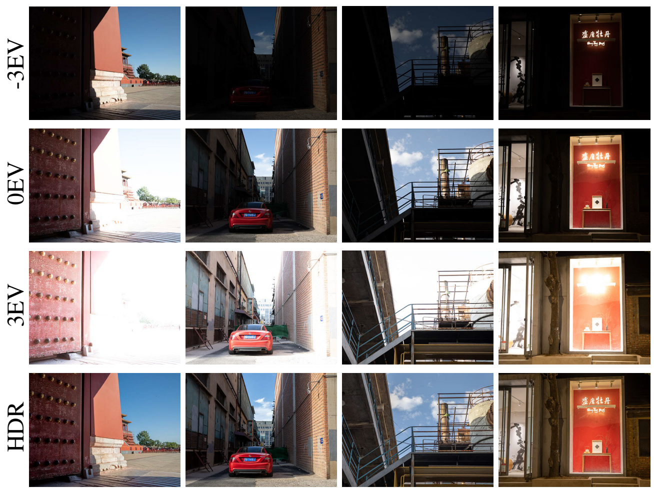


Figure 4: Representative scenes of our captured Raw-to-HDR dataset.

are served as input images, and we follow the well-known HDR merging method [11] to fuse HDR from raw image series, which are served as ground truth. In total, we collect 324 pairs of Raw/HDR images using Canon 5D Mark IV camera. For each scene, images are with a high resolution of 4480×6720, and the final dataset is carefully checked and filtered to exclude misaligned pairs. The input Raw images of our dataset are recorded in 14-bit Raw format, and the corresponding HDR images are 20-bit, with additional image profiles (white balance, color correction matrix) recorded in the file. The dataset is available at https://github.com/jackzou233/RawHDR.

5. Experiments

In this section, we first introduce the experimental settings, including the compared methods and the evaluation metrics. Then, we conduct experiments on our Raw-to-HDR dataset. Next, we perform ablation studies, to validate the superiority of the proposed model. Finally, we perform extended experiments for further evaluation.

5.1. Experimental Settings

Compared methods. To assess the performance of our RawHDR model, we compare with leading single-exposure HDR reconstruction methods, including HDRCNN [13], ExpandNet [32], DeepHDR [43], and HDRUNet [9]. In addition, we also compare with methods that are designed to deal with under-exposed regions (SID [7]) and over-exposed regions (EC [2]). Besides our original model, we incorporate a small version (denoted as Ours-S) by simply reducing the number of layers and channel numbers.

Evaluation metrics. Peak signal-to-noise ratio (PSNR) is a widely used metric to evaluate pixel-level fidelity. We compute PSNR directly on HDR images in linear space. Since HDR images generally need to be tone-mapped before displaying on modern displays, we also test the PSNR on the tone-mapped HDR images. The results are denoted as PSNR-μ, where μ is the tone mapping parameter. Compared with PSNR, which may stress much importance on the brightest regions of the image, PSNR-μ is a more balanced metric that can represent the reconstruction fidelity in both dark and bright areas. In other words, PSNR-μ reveals more information from the HDR display perspective. In addition to pixel-level evaluations, we also employ Structural Similarity [46] (SSIM) and multi-scale SSIM [48] (MS-SSIM) to evaluate the structural similarity of the tone-mapped images. Larger PSNR, PSNR-μ, SSIM, and MS-SSIM show better performances. For each method, we also report the computational cost (GMACs) and number of parameters to evaluate the mode size.

5.2. Results on Our Raw-to-HDR Dataset

To verify the effectiveness of our method, we compare our RawHDR model against prevailing methods on our Raw-to-HDR dataset. Notably, the original architectures of EC [2], HDRCNN [13], and HDRUNet [9] are designed to accommodate 3-channel RGB images. We adapted these models by modifying their input to 4 channels, ensuring compatibility with the Raw-to-HDR framework. Table 2 summarizes the numerical results according to the averaged values of all evaluation metrics. We can see that our method outperforms competing methods in PSNR, PSNR-μ, MS-SSIM, and presents a competitive performance in SSIM. One observation is that the advantage of our RawHDR is greater in PSNR-μ, compared with PSNR. This is benefited from our specially designed deep architecture for hard region recovery. The qualitative results are shown in Figures 5-6. It can be seen that the images recovered from our method approximate ground truth well and is significantly

Table 2: The comparisons of HDR reconstruction performances on our dataset. The best results are highlighted in **bold**.

Metrics	HDRCNN [13]	ExpandNet [32]	SID [7]	DeepHDR [43]	EC [2]	HDRUNet [9]	Ours-S	Ours
MACs(G)	54.46	13.75	13.73	19.13	**10.57**	23.61	17.85	42.07
Params(M)	155.45	**0.485**	7.76	51.55	7.02	1.65	4.47	10.49
PSNR	28.13	36.18	36.76	36.77	35.16	36.67	37.03	**37.24**
PSNR-μ	18.04	41.15	37.86	41.18	36.15	41.16	41.48	**41.95**
SSIM	0.4360	0.9701	0.9689	**0.9739**	0.9423	0.9232	0.9716	0.9714
MS-SSIM	0.7446	0.9919	0.9925	0.9910	0.9806	0.9851	0.9921	**0.9934**

Input ExpandNet [32] SID [7] DeepHDR [43] EC [2] HDRUNet [9] Ours GT

Figure 5: The result on our dataset. The visualization of typical *under-exposed* regions of Input/ExpandNet/SID/ DeepHDR/EC/HDRUNet/Ours/GT are presented. All images except the input are visualized through tone mapping.

better than other reconstruction methods, especially for extremely dark and bright regions. This phenomenon is benefited from our dual intensity and global spatial guidance, which better exploits features from Raw images.

5.3. Ablation Studies

This section performs ablation studies to verify the effectiveness of our model components and the proposed dataset.

Mask estimation. First, we evaluate the effectiveness of the proposed mask estimation module. In this module, we use deep neural networks to learn a soft mask, which well separates the over- and under-exposed areas. Here, we also conduct ablation studies that uses hard masks M^h_{over} and M^h_{under}. From the results in Figure 7(b), we see that using hard masks suffers from color artifacts, which do not appear in the results of our soft mask estimation.

Dual intensity guidance. Here, we evaluate the promotion brought by the dual intensity guidance. As shown in Table 3, by replacing each or both channel intensity guidance (denoted as w/o DIG), we notice apparent performance degradation. From the visual results illustrated in Figure 7(c), we see more details are recovered in hard regions with the help of the dual intensity guidance.

Global spatial guidance. As shown in Table 3 and Figures 7(d)-(f), we also evaluate our global spatial guidance by removing this module (denoted as w/o GSG). By comparing both quantitative and qualitative results, we can infer that global spatial guidance indeed plays an important role in extremely dark and bright regions.

Table 3: The ablation study results on network design.

Setting	PSNR	PSNR-μ	SSIM	MS-SSIM
Hard mask	35.70	39.37	0.9617	0.9904
w/o DIG	35.41	38.27	0.9206	0.9826
G guid.	35.85	40.30	0.9643	0.9913
RB guid.	36.59	38.72	0.9561	0.9889
w/o GSG	36.45	41.04	0.9687	0.9933
Ours	**37.04**	**41.50**	**0.9707**	**0.9935**

5.4. Further Evaluation for Raw-to-HDR

In this section, we further conduct extended experiments to verify the effectiveness of the setting to reconstruction HDR from Raw images.

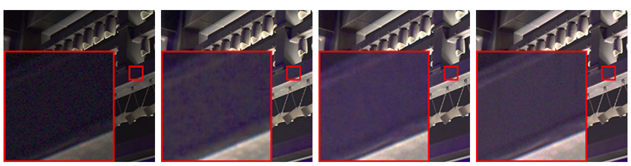
Evaluation of the Raw-to-HDR mapping. To validate the superiority of Raw-to-HDR reconstruction compared with commonly used sRGB-to-HDR, we design experiments to compare sRGB and Raw data for HDR reconstruction. Specifically, we synthesize sRGB images from the Raw data of the proposed dataset through a simple camera ISP (following [1]), and then form paired sRGB/HDR dataset. We train our RawHDR model on both the newly curated sRGB/HDR dataset and the proposed dataset in the same setting. For the results of Raw-to-HDR setting, we use the same processing pipeline to guarantee that both settings are in the same color space. Results in Table 4 and Figure 8 validate that by introducing linear high-bit Raw images, the reconstruction of HDR suffers less from information loss. Therefore, in order to obtain high-quality HDR, it is necessary to feed HDR reconstruction model with Raw images.

Input ExpandNet [32] SID [7] DeepHDR [43] EC [2] HDRUNet [9] Ours GT

Figure 6: The result on our dataset. The visualizations of typical *over-exposed* regions of Input/ExpandNet/SID/ DeepHDR/EC/HDRUNet/Ours/GT are presented. All images except the input are visualized through tone mapping.

(a) Input (b) Hard mask (c) w/o GSG (d) w/o DIG (e) G guid. (f) RB guid. (g) Ours (h) GT

Figure 7: The visualization results for ablation studies.



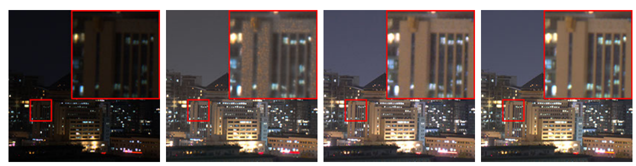
Input Ours (RGB) Ours (Raw) GT

Figure 8: The quantitative results for our RawHDR model trained on sRGB data and Raw data.

Table 4: The comparison between sRGB and Raw.

Setting	PSNR	PSNR-μ	SSIM	MS-SSIM
sRGB	37.25	37.82	0.9385	0.9798
Raw	**44.15**	**39.76**	**0.9509**	**0.9848**

Cross-camera HDR reconstruction. Besides the advantage that Raw images have higher bit-depth than sRGB (14-bit *vs*. 8-bit), another crucial attribute is the linearity. When attempting HDR reconstruction from a singular sRGB image, the learned models tend to be camera-specific, since the network is overfitted to a specific camera processing pipeline. However, Raw images are in linear space and unprocessed by any in-camera operation. As a result, discrepancies between Raw images from various sources are much less noticeable than those in sRGB format. To validate this, we capture Raw images from a different camera, *i.e*., Sony A7R4, and directly evaluate the HDR reconstruction results using the model trained on Canon 5D IV. As shown in Figure 9, when transferring to another camera, the model trained in sRGB setting fails to produce high-quality HDR, while the results are robust using Raw data.



Input Ours (RGB) Ours (Raw) GT

Figure 9: The results for cross-camera setting. The model is trained on Canon 5D IV and evaluated on Sony A7R4.

6. Conclusion

In this work, we aim to solve the most challenging problem in single exposure HDR reconstruction, which is the insufficient information in the darkest and brightest regions of a high dynamic scene. Our work solves this problem from two aspects. First, we propose to reconstruct HDR from a single Raw image instead of sRGB, since Raw images have higher bit-depth and intensity tolerance. Then, we propose a novel deep model, *i.e*., RawHDR, to learn the Raw-to-HDR mapping. Specifically, the proposed mask estimation, dual intensity guidance, and global spatial guidance help us to separate hard and easy regions and fully exploit local and long range features. We capture high-quality Raw-to-HDR datasets for data-driven methods, and our dataset is evaluated to be very necessary for HDR reconstruction. Experimental results verify that both our dataset and RawHDR model are of high quality. The Raw-to-HDR mapping can be potentially integrated in real camera ISP once the model is simplified, and we remain this as our future work.

Acknowledgments This work was supported by the National Natural Science Foundation of China (62171038, 61931008, 62088101, and U21B2024), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), and the Fundamental Research Funds for the Central Universities.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1692–1700, 2018. 7

[2] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9157–9167, 2021. 2, 6, 7, 8

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Francesco Banterle, Kurt Debattista, Alessandro Artusi, Sumanta Pattanaik, Karol Myszkowski, Patrick Ledda, and Alan Chalmers. High dynamic range imaging and low dynamic range expansion for generating hdr content. In *Computer graphics forum*, volume 28, pages 2343–2367. Wiley Online Library, 2009. 2

[5] Francesco Banterle, Patrick Ledda, Kurt Debattista, Alan Chalmers, and Marina Bloj. A framework for inverse tone mapping. *The Visual Computer*, 23(7):467–478, 2007. 2

[6] Sebastiano Battiato, Alfio Castorina, and Massimo Mancuso. High dynamic range imaging for digital still camera: an overview. *Journal of electronic Imaging*, 12(3):459–469, 2003. 1

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, 2018. 2, 6, 7, 8

[8] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, pages 1–21, 2023. 2

[9] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proc. of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 354–363, 2021. 2, 3, 6, 7, 8

[10] Sungil Choi, Jaehoon Cho, Wonil Song, Jihwan Choe, Jisung Yoo, and Kwanghoon Sohn. Pyramid inter-attention for high dynamic range imaging. *Sensors*, 20(18):5102, 2020. 1

[11] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, pages 1–10. 2008. 2, 6

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[13] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Trans. on Graphics*, 36(6):1–15, 2017. 2, 3, 5, 6, 7

[14] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. on Graphics*, 36(6):177–1, 2017. 1

[15] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, volume 9023, pages 279–288. SPIE, 2014. 5, 6

[16] Ying Fu, Yang Hong, Linwei Chen, and Shaodi You. Le-gan: Unsupervised low-light image enhancement network using attention module and identity invariant loss. *Knowledge-Based Systems*, 240:108010, 2022. 2

[17] Shi Guo, Zhetong Liang, and Lei Zhang. Joint denoising and demosaicking with green channel prior for real-world burst images. *IEEE Trans. Image Processing*, 30:6930–6942, 2021. 4

[18] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1730–1739, 2020. 1

[19] 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 Trans. on Graphics*, 35(6):1–12, 2016. 1, 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 4

[22] Jie Huang, Man Zhou, Yajing Liu, Mingde Yao, Feng Zhao, and Zhiwei Xiong. Exposure-consistency representation learning for exposure correction. In *Proc. of International Conference on Multimedia (MM)*, pages 6309–6317, 2022. 2

[23] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. on Graphics*, 36(4):144–1, 2017. 1, 2, 5, 6

[24] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 429–444. Springer, 2016. 3

[25] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 3116–3125, 2019. 1

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of International Conference on Learning representations (ICLR)*, 2015. 5

[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 (ECCV)*, pages 596–611, 2018. 2

[28] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint demosaicing and denoising with self guidance. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2240–2249, 2020. 3, 4

[29] Shuo Liu, Mingliang Gao, Vijay John, Zheng Liu, and Erik Blasch. Deep learning thermal image translation for night vision perception. *ACM Trans. on Intelligent Systems and Technology*, 12(1):1–18, 2020. 1

[30] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1651–1660, 2020. 2, 3, 4, 5, 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 5

[32] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49, 2018. 2, 6, 7, 8

[33] Miguel Ángel Martínez-Domingo, Eva M Valero, Javier Hernández-Andrés, Shoji Tominaga, Takahiko Horiuchi, and Keita Hirai. Image processing pipeline for segmentation and material classification based on multispectral high dynamic range polarimetric images. *Optics Express*, 25(24):30073–30090, 2017. 1

[34] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1375–1385, 2020. 1

[35] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in ldr and hdr images. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015. 5, 6

[36] Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7710–7720, 2021. 1

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of Conference on Neural Information Processing Systems (NeurIPS)*, pages 8026–8037, 2019. 5

[38] Feiyue Peng, Maojun Zhang, Shiming Lai, Hanlin Tan, and Shen Yan. Deep hdr reconstruction of dynamic scenes. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 347–351. IEEE, 2018. 2

[39] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proc. of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 691–700, 2021. 5, 6

[40] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Analysis and Machine Intelligence* , 2019. 1

[41] Nicolas Robidoux, Luis E Garcia Capel, Dong-eun Seo, Avinash Sharma, Federico Ariza, and Felix Heide. End-to-end high dynamic range camera pipeline optimization. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*. 1

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[43] Marcel Santana Santos, Ren Tsang, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Trans. on Graphics*, 39(4), 7 2020. 6, 7, 8

[44] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10081–10090, 2019. 1

[45] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Trans. Pattern Analysis and Machine Intelligence* , 2021. 1, 5

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 6

[47] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, 2022. 5

[48] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 6

[49] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Trans. Pattern Analysis and Machine Intelligence* , 2021. 3

[50] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6287–6296, 2021. 2

[51] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17714–17724, 2022. 5

[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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2019. 1

[53] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50, 2019. 1, 2

[54] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4967–4976, 2021. 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5

[56] Tao Zhang, Ying Fu, and Cheng Li. Deep spatial adaptive network for real image demosaicing. *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 36(3):3326–3334, Jun. 2022. 4

[57] Yunhao Zou and Ying Fu. Estimating fine-grained noise model via contrastive learning. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12682–12691, 2022. 3

[58] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2024–2033, 2021. 1