Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline

Yu-Lun Liu[1,2*] Wei-Sheng Lai[3*] Yu-Sheng Chen[1] Yi-Lung Kao[1]
Ming-Hsuan Yang[3,4] Yung-Yu Chuang[1] Jia-Bin Huang[5]
[1]National Taiwan University [2]MediaTek Inc. [3]Google [4]UC Merced [5]Virginia Tech

https://www.cmlab.csie.ntu.edu.tw/~yulunliu/SingleHDR

Figure 1: **HDR reconstruction from a single LDR image.** Our method recovers missing details for both backlit and over-exposed regions of real-world images by learning to reverse the camera pipeline. Note that the input LDR images are captured by different real cameras, and all reconstructed HDR images have been tone-mapped by [32] for display.

Abstract

Recovering a high dynamic range (HDR) image from a single low dynamic range (LDR) input image is challenging due to missing details in under-/over-exposed regions caused by quantization and saturation of camera sensors. In contrast to existing learning-based methods, our core idea is to incorporate the domain knowledge of the LDR image formation pipeline into our model. We model the HDR-to-LDR image formation pipeline as the (1) dynamic range clipping, (2) non-linear mapping from a camera response function, and (3) quantization. We then propose to learn three specialized CNNs to reverse these steps. By decomposing the problem into specific sub-tasks, we impose effective physical constraints to facilitate the training of individual sub-networks. Finally, we jointly fine-tune the entire model end-to-end to reduce error accumulation. With extensive quantitative and qualitative experiments on diverse image datasets, we demonstrate that the proposed method performs favorably against state-of-the-art single-image HDR reconstruction algorithms.

*Indicates equal contribution.

1. Introduction

HDR images are capable of capturing rich real-world scene appearances including lighting, contrast, and details. Consumer-grade digital cameras, however, can only capture images within a limited dynamic range due to sensor constraints. The most common approach to generate HDR images is to merge multiple LDR images captured with different exposures [12]. Such a technique performs well on static scenes but often suffers from ghosting artifacts on dynamic scenes or hand-held cameras. Furthermore, capturing multiple images of the same scene may not always be feasible (e.g., existing LDR images on the Internet).

Single-image HDR reconstruction aims to recover an HDR image from a single LDR input. The problem is challenging due to the missing information in under-/over-exposed regions. Recently, several methods [14, 15, 40, 53, 56] have been developed to reconstruct an HDR image from a given LDR input using deep convolutional neural networks (CNNs). However, learning a direct LDR-to-HDR mapping is difficult as the variation of HDR pixels (32-bit) is significantly higher than that of LDR pixels (8-bit). Recent methods address this challenge either by focusing on

recovering the over-exposed regions [14] or synthesizing several up-/down-exposed LDR images and fusing them to produce an HDR image [15]. The artifacts induced by quantization and inaccurate camera response functions (CRFs) are, however, only *implicitly* addressed through learning.

In this work, we incorporate the domain knowledge of the LDR image formation pipeline to design our model. We model the image formation with the following steps [12]: (1) dynamic range clipping, (2) non-linear mapping with a CRF, and (3) quantization. Instead of learning a direct LDR-to-HDR mapping using a generic network, our core idea is to decompose the single-image HDR reconstruction problem into three sub-tasks: i) dequantization, ii) linearization, and iii) hallucination, and develop three deep networks to specifically tackle each of the tasks. First, given an input LDR image, we apply a Dequantization-Net to restore the missing details caused by quantization and reduce the visual artifacts in the under-exposed regions (e.g., banding artifacts). Second, we estimate an inverse CRF with a Linearization-Net and convert the non-linear LDR image to a *linear* image (i.e., scene irradiance). Building upon the empirical model of CRFs [16], our Linearization-Net leverages the additional cues from edges, the intensity histogram and a monotonically increasing constraint to estimate more accurate CRFs. Third, we predict the missing content in the over-exposed regions with a Hallucination-Net. To handle other complicated operations (e.g., lens shading correction, sharpening) in modern camera pipelines that we do not model, we use a Refinement-Net and jointly fine-tune the whole model end-to-end to reduce error accumulation and improve the generalization ability to real input images.

By explicitly modeling the *inverse* functions of the LDR image formation pipeline, we significantly reduce the difficulty of training one single network for reconstructing HDR images. We evaluate the effectiveness of our method on four datasets and real-world LDR images. Extensive quantitative and qualitative evaluations, as well as the user study, demonstrate that our model performs favorably against the state-of-the-art single-image HDR reconstruction methods. Figure 1 shows our method recovers visually pleasing results with faithful details. Our contributions are three-fold:

- We tackle the single-image HDR reconstruction problem by reversing image formation pipeline, including the dequantization, linearization, and hallucination.
- We introduce specific physical constraints, features, and loss functions for training each individual network.
- We collect two HDR image datasets, one with synthetic LDR images and the other with real LDR images, for training and evaluation. We show that our method performs favorably against the state-of-the-art methods in terms of the HDR-VDP-2 scores and visual quality on the collected and existing datasets.

2. Related Work

Multi-image HDR reconstruction. The most common technique for creating HDR images is to fuse a stack of bracketed exposure LDR images [12, 38]. To handle dynamic scenes, image alignment and post-processing are required to minimize artifacts [25, 37, 50]. Recent methods apply CNNs to fuse multiple flow-aligned LDR images [23] or unaligned LDR images [52]. In contrast, we focus on reconstructing an HDR image from a *single* LDR image.

Single-image HDR reconstruction. Single-image HDR reconstruction does not suffer from ghosting artifacts but is significantly more challenging than the multi-exposure counterpart. Early approaches estimate the density of light sources to expand the dynamic range [1, 2, 3, 4, 5] or apply the cross-bilateral filter to enhance the input LDR images [20, 27]. With the advances of deep CNNs [17, 48], several methods have been developed to learn a direct LDR-to-HDR mapping [40, 53, 56]. Eilertsen et al. [14] propose the HDRCNN method that focuses on recovering missing details in the over-exposed regions while ignoring the quantization artifacts in the under-exposed areas. In addition, a fixed inverse CRF is applied, which may not be applicable to images captured from different cameras. Instead of learning a direct LDR-to-HDR mapping, some recent methods [15, 30] learn to synthesize multiple LDR images with different exposures and reconstruct the HDR image using the conventional multi-image technique [12]. However, predicting LDR images with different exposures from a single LDR input itself is challenging as it involves the non-linear CRF mapping, dequantization, and hallucination.

Unlike [15, 30], our method directly reconstructs an HDR image by modeling the inverse process of the image formation pipeline. Figure 2 illustrates the LDR image formation pipeline, state-of-the-art single-image HDR reconstruction approaches [14, 15, 40], and the proposed method.

Dequantization and decontouring. When converting real-valued HDR images to 8-bit LDR images, quantization errors inevitably occurs. They often cause scattered noise or introduce false edges (known as contouring or banding artifacts) particularly in regions with smooth gradient changes. While these errors may not be visible in the non-linear LDR image, the tone mapping operation (for visualizing an HDR image) often aggravates them, resulting in noticeable artifacts. Existing decontouring methods smooth images by applying the adaptive spatial filter [9] or selective average filter [49]. However, these methods often involve meticulously tuned parameters and often produce undesirable artifacts in textured regions. CNN-based methods have also been proposed [18, 35, 58]. Their focus is on restoring an 8-bit image from lower bit-depth input (e.g., 2-bit or 4-bit). In contrast, we aim at recovering a 32-bit floating-point image from an 8-bit LDR input image.

Radiometric calibration. As the goal of HDR reconstruction is to measure the full scene irradiance from an input LDR image, it is necessary to estimate the CRF. Recovering the CRF from a single image requires certain assumptions of statistical priors, e.g., color mixtures at edges [33, 34, 43] or noise distribution [41, 51]. Nevertheless, these priors may not be applicable to a wide variety of images in the wild. A CRF can be empirically modeled by the basis vectors extracted from a set of real-world CRFs [16] via the principal component analysis (PCA). Li and Peers [31] train a CRF-Net to estimate the weights of the basis vectors from a single input image and then use the principal components to reconstruct the CRF. Our work improves upon [31] by introducing new features and monotonically increasing constraint. We show that an accurate CRF is crucial to the quality of the reconstructed HDR image. After obtaining an accurate HDR image, users can adopt advanced tone-mapping methods (e.g., [32, 46]) to render a more visually pleasing LDR image. Several other applications (e.g., image-based lighting [11] and motion blur synthesis [12]) also require linear HDR images for further editing or mapping.

Image completion. Recovering the missing contents in saturated regions can be posed as an image completion problem. Early image completion approaches synthesize the missing contents via patch-based synthesis [6, 13, 19]. Recently, several learning-based methods have been proposed to synthesize the missing pixels using CNNs [21, 36, 45, 55, 54]. Different from the generic image completion task, the missing pixels in the over-exposed regions always have equal or larger values than other pixels in an image. We incorporate this constraint in our Hallucination-Net to reflect the physical formation in over-exposed regions.

Camera pipeline. We follow the forward LDR image formation pipeline in HDR reconstruction [12] and radiometric calibration [8] algorithms. While the HDRCNN method [14] also models a similar LDR image formation, this model does not learn to estimate accurate CRFs and reduce quantization artifacts. There exist more advanced and complex camera pipelines to model the demosaicing, white balancing, gamut mapping, noise reduction steps for image formation [7, 24, 26]. In this work, we focus on the components of great importance for HDR image reconstruction and model the rest of the pipeline by a refinement network.

3. Learning to Reverse the Camera Pipeline

In this section, we first introduce the image formation pipeline that renders an LDR image from an HDR image (the scene irradiance) as shown in Figure 2(a). We then describe our design methodology and training procedures for single-image HDR reconstruction by reversing the image formation pipeline as shown in Figure 2(e).

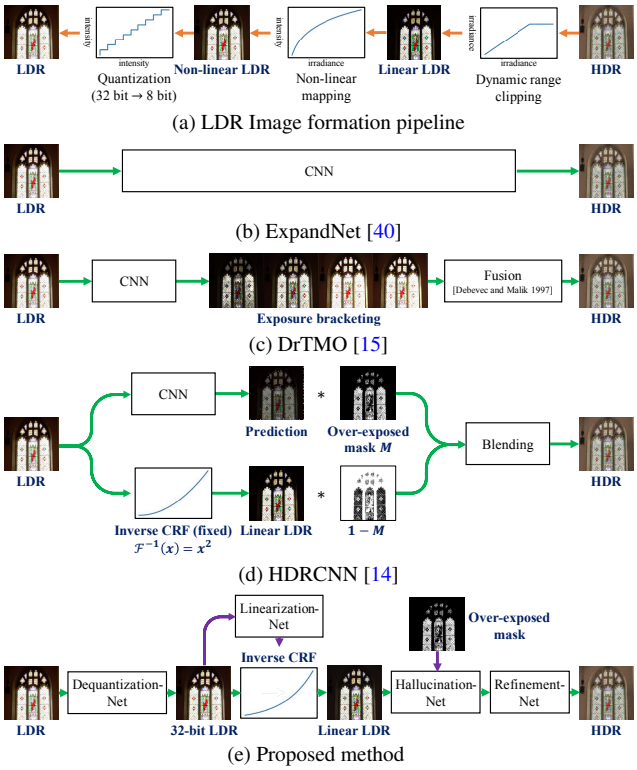


Figure 2: **The LDR Image formation pipeline and overview of single-image HDR reconstruction methods.** (a) We model the LDR image formation by (from right to left) dynamic range clipping, non-linear mapping, and quantization [12]. (b) ExpandNet [40] learns a direct mapping from LDR to HDR images. (c) DrTMO [15] synthesizes multiple LDR images with different exposures and fuses them into an HDR image. (d) HDRCNN [14] predicts details in over-exposed regions while ignoring the quantization errors in the under-exposed regions. (e) The proposed method explicitly learns to "reverse" each step of the LDR image formation pipeline.

3.1. LDR image formation

While the real scene irradiance has a high dynamic range, the digital sensor in cameras can only capture and store a limited extent, usually with 8 bits. Given the irradiance E and sensor exposure time t, an HDR image is recorded by $H = E \times t$. The process of converting one HDR image to one LDR image can be modeled by the following major steps:

(1) **Dynamic range clipping**. The camera first clips the pixel values of an HDR image H to a limited range, which can be formulated by $I_c = \mathcal{C}(H) = \min(H, 1)$. Due to the clipping operation, there is information loss for pixels in the over-exposed regions.

(2) **Non-linear mapping**. To match the human perception of the scene, a camera typically applies a non-linear CRF

mapping to adjust the contrast of the captured image: $I_n = \mathcal{F}(I_c)$. A CRF is unique to the camera model and unknown in our problem setting.

(3) **Quantization**. After the non-linear mapping, the recorded pixel values are quantized to 8 bits by $\mathcal{Q}(I_n) = \lfloor 255 \times I_n + 0.5 \rfloor / 255$. The quantization process leads to errors in the under-exposed and smooth gradient regions.

In summary, an LDR image L is formed by:

$$L = \Phi(H) = \mathcal{Q}(\mathcal{F}(\mathcal{C}(H))), \tag{1}$$

where Φ denotes the pipeline of dynamic range clipping, non-linear mapping, and quantization steps.

To learn the inverse mapping Φ^{-1}, we propose to decompose the HDR reconstruction task into three sub-tasks: dequantization, linearization, and hallucination, which model the inverse functions of the quantization, non-linear mapping, and dynamic range clipping, respectively. We train three CNNs for the three sub-tasks using the corresponding supervisory signal and specific physical constraints. We then integrate these three networks into an end-to-end model and jointly fine-tune to further reduce error accumulation and improve the performance.

3.2. Dequantization

Quantization often results in scattered noise or contouring artifacts in smooth regions. Therefore, we propose to learn a Dequantization-Net to reduce the quantization artifacts in the input LDR image.

Architecture. Our Dequantization-Net adopts a 6-level U-Net architecture. Each level consists of two convolutional layers followed by a leaky ReLU ($\alpha = 0.1$) layer. We use the `Tanh` layer to normalize the output of the last layer to $[-1.0, 1.0]$. Finally, we add the output of the Dequantization-Net to the input LDR image to generate the dequantized LDR image $\hat{I}_{\text{deq}}$.

Training. We minimize the ℓ_2 loss between the dequantized LDR image $\hat{I}_{\text{deq}}$ and corresponding ground-truth image I_n: $\mathcal{L}_{\text{deq}} = \|\hat{I}_{\text{deq}} - I_n\|_2^2$. Note that $I_n = \mathcal{F}(\mathcal{C}(H))$ is constructed from the ground-truth HDR image with dynamic range clipping and non-linear mapping.

3.3. Linearization

The goal of linearization (i.e., radiometric calibration) is to estimate a CRF and convert a non-linear LDR image to a linear irradiance. Although the CRF (denoted by $\mathcal{F}$) is distinct for each camera, all the CRFs must have the following properties. First, the function should be monotonically increasing. Second, the minimal and maximal input values should be respectively mapped to the minimal and maximal output values: $\mathcal{F}(0) = 0$ and $\mathcal{F}(1) = 1$ in our case. As the CRF is a one-to-one mapping function, the inverse CRF (denoted by $\mathcal{G} = \mathcal{F}^{-1}$) also has the above properties.

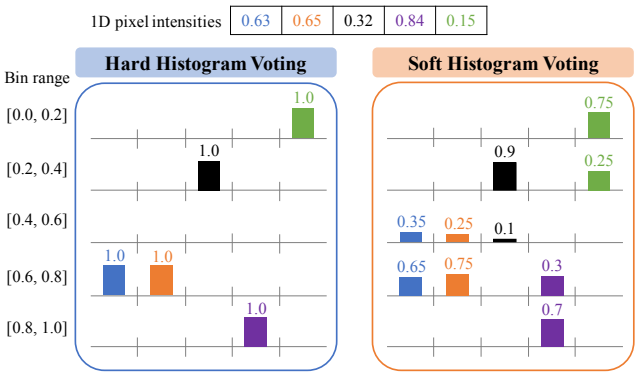


Figure 3: **Spatial-aware soft histogram layer.** We extract histogram features by soft counting on pixel intensities and preserving the spatial information.

To represent a CRF, we first discretize an inverse CRF by uniformly sampling 1024 points between $[0, 1]$. Therefore, an inverse CRF is represented as a 1024-dimensional vector $\mathbf{g} \in \mathbb{R}^{1024}$. We then adopt the Empirical Model of Response (EMoR) model [16], which assumes that each inverse CRF $\mathbf{g}$ can be approximated by a linear combination of K PCA basis vectors. In this work, we set $K = 11$ as it has been shown to capture the variations well in the CRF dataset [31]. To predict the inverse CRF, we train a Linearization-Net to estimate the weights from the input non-linear LDR image.

Input features. As the edge and color histogram have been shown effective to estimate an inverse CRF [33, 34], we first extract the edge and histogram features from the non-linear LDR image. We adopt the Sobel filter to obtain the edge responses, resulting in 6 feature maps (two directions $\times$ three color channels). To extract the histogram features, we propose a *spatial-aware soft-histogram layer*. Specifically, given the number of histogram bins B, we construct a *soft* counting of pixel intensities by:

$$h(i, j, c, b) = \begin{cases} 1 - d \cdot B\,, & \text{if } d < \frac{1}{B} \\ 0\,, & \text{otherwise} \end{cases} \tag{2}$$

where i, j indicate horizontal and vertical pixel positions, c denotes the index of color channels, $b \in \{1, \cdots, B\}$ is the index for the histogram bin, and $d = |I(i, j, c) - (2b - 1)/(2B)|$ is the intensity distance to the center of the b-th bin. Every pixel contributes to the two nearby bins according to the intensity distance to the center of each bin. Figure 3 shows a 1D example of our soft-histogram layer. Our histogram layer preserves the spatial information and is fully differentiable.

Architecture. We use the ResNet-18 [17] as the backbone of our Linearization-Net. To extract a global feature, we add a global average pooling layer after the last convolutional

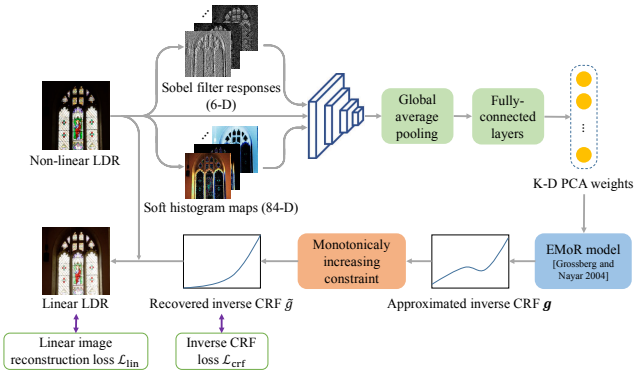


Figure 4: **Architecture of the Linearization-Net.** Our Linearization-Net takes as input the non-linear LDR image, edge maps, and histogram maps, and predicts the PCA coefficients for reconstructing an inverse CRF, followed by enforcing the monotonically increasing constraint.

layer. We then use two fully-connected layers to generate K PCA weights and reconstruct an inverse CRF.

Monotonically increasing constraint. To satisfy the constraint that a CRF/inverse CRF should be monotonically increasing, we adjust the estimated inverse CRF by enforcing all the first-order derivatives to be non-negative. Specifically, we calculate the first-order derivatives by $g'_1 = 0$ and $g'_d = g_d - g_{d-1}$ for $d \in [2, \cdots, 1024]$ and find the smallest negative derivative $g'_m = \min(\min_d(g'_d), 0)$. We then shift the derivatives by $\tilde{g}'_d = g'_d - g'_m$. The inverse CRF $\tilde{\mathbf{g}} = [\tilde{g}_1, \cdots, \tilde{g}_{1024}]$ is then reconstructed by integration and normalization:

$$\tilde{g}_d = \frac{1}{\sum_{i=1}^{1024} \tilde{g}'_i} \sum_{i=1}^{d} \tilde{g}'_i \,. \tag{3}$$

We normalize $\tilde{g}_d$ to ensure the inverse CRF satisfies the constraint that $\mathcal{G}(0) = 0$ and $\mathcal{G}(1) = 1$. Figure 4 depicts the pipeline of our Linearization-Net. With the normalized inverse CRF $\tilde{\mathbf{g}}$, we then map the non-linear LDR image $\hat{I}_{\text{deq}}$ to a linear LDR image $\hat{I}_{\text{lin}}$.

Training. We define the linear LDR image reconstruction loss by: $\mathcal{L}_{\text{lin}} = \|\hat{I}_{\text{lin}} - I_c\|_2^2$, where $I_c = \mathcal{C}(H)$ is constructed from the ground-truth HDR image with the dynamic range clipping process. In addition, we formulate the inverse CRF reconstruction loss by: $\mathcal{L}_{\text{crf}} = \|\tilde{\mathbf{g}} - \mathbf{g}\|_2^2$, where $\mathbf{g}$ is the ground-truth inverse CRF. We train the Linearization-Net by optimizing $\mathcal{L}_{\text{lin}} + \lambda_{\text{crf}}\mathcal{L}_{\text{crf}}$. We empirically set $\lambda_{\text{crf}} = 0.1$ in all our experiments.

3.4. Hallucination

After dequantization and linearization, we aim to recover the missing contents due to dynamic range clipping. To this end, we train a Hallucination-Net (denoted by $\mathcal{C}^{-1}(\cdot)$) to predict the missing details within the over-exposed regions.

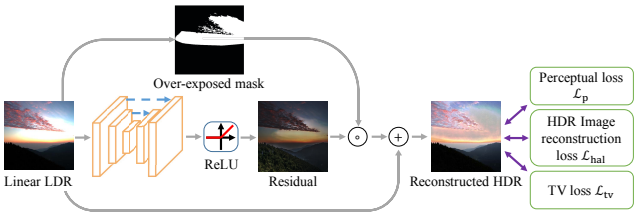


Figure 5: **Architecture of the Hallucination-Net.** We train the Hallucination-Net to predict *positive* residuals and recover missing content in the over-exposed regions.

Architecture. We adopt an encoder-decoder architecture with skip connections [14] as our Hallucination-Net. The reconstructed HDR image is modeled by $\hat{H} = \hat{I}_{\text{lin}} + \alpha \cdot \mathcal{C}^{-1}(\hat{I}_{\text{lin}})$, where $\hat{I}_{\text{lin}}$ is the image generated from the Linearization-Net and $\alpha = \max(0, \hat{I}_{\text{lin}} - \gamma)/(1 - \gamma)$ is the over-exposed mask with $\gamma = 0.95$. Since the missing values in the over-exposed regions should always be greater than the existing pixel values, we constrain the Hallucination-Net to predict *positive residuals* by adding a `ReLU` layer at the end of the network. We note that our over-exposed mask is a *soft* mask where $\alpha \in [0, 1]$. The soft mask allows the network to smoothly blend the residuals with the existing pixel values around the over-exposed regions. Figure 5 shows the design of our Hallucination-Net.

We find that the architecture of [14] may generate visible checkerboard artifacts in large over-exposed regions. In light of this, we replace the transposed convolutional layers in the decoder with the resize-convolution layers [44].

Training. We train our Hallucination-Net by minimizing the $\log - \ell_2$ loss: $\mathcal{L}_{\text{hal}} = \|\log(\hat{H}) - \log(H)\|_2^2$, where H is the ground-truth HDR image. We empirically find that training is more stable and achieves better performance when minimizing the loss in the log domain. As the highlight regions (e.g., sun and light sources) in an HDR image typically have values with orders of magnitude larger than those of other regions, the loss is easily dominated by the errors in the highlight regions when measured in the linear domain. Computing the loss in the log domain reduces the influence of these extremely large errors and encourages the network to restore more details in other regions.

To generate more realistic details, we further include the perceptual loss $\mathcal{L}_{\text{p}}$ [22]: As the VGG-Net (used in $\mathcal{L}_{\text{p}}$) is trained on *non-linear RGB images*, directly feeding an linear HDR image to the VGG-Net may not obtain meaningful features. Therefore, we first apply a differentiable global tone-mapping operator [52] to map the HDR images to a non-linear RGB space. We can then compute the perceptual loss on the tone-mapped HDR images. To improve the spatial smoothness of the predicted contents, we also minimize the total variation (TV) loss $\mathcal{L}_{\text{tv}}$ on the recovered HDR

image. Our total loss for training the Hallucination-Net is $\mathcal{L}_{\text{hal}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}$. We empirically set $\lambda_{\text{p}} = 0.001$ and $\lambda_{\text{tv}} = 0.1$ in our experiments.

3.5. Joint training

We first train the Dequantization-Net, Linearization-Net, and Hallucination-Net with the corresponding input and ground-truth data. After the three networks converge, we jointly fine-tune the entire pipeline by minimizing the combination of loss functions $\mathcal{L}_{\text{total}}$:

$$\lambda_{\text{deq}}\mathcal{L}_{\text{deq}} + \lambda_{\text{lin}}\mathcal{L}_{\text{lin}} + \lambda_{\text{crf}}\mathcal{L}_{\text{crf}} + \lambda_{\text{hal}}\mathcal{L}_{\text{hal}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} \quad (4)$$

where we set the weights to $\lambda_{\text{deq}} = 1$, $\lambda_{\text{lin}} = 10$, $\lambda_{\text{crf}} = 1$, $\lambda_{\text{hal}} = 1$, $\lambda_{\text{p}} = 0.001$, and $\lambda_{\text{tv}} = 0.1$. The joint training reduces error accumulation between the sub-networks and further improves the reconstruction performance.

3.6. Refinement

Modern camera pipeline contains significant amounts of *spatially-varying* operations, e.g. local tone-mapping, sharpening, chroma denoising, lens shading correction, and white balancing. To handle these effects that are not captured by our image formation pipeline, we introduce an optional Refinement-Net.

Architecture. Our Refinement-Net adopts the same U-Net architecture as the Dequantization-Net, which learns to refine the output of the Hallucination-Net by a residual learning. The output of the Refinement-net is denoted by $\hat{H}_{\text{ref}}$.

Training. To model the effects of real camera pipelines, we train the Refinement-Net using HDR images reconstructed from exposure stacks captured by various cameras (more details in the supplementary material). We minimize the same $\mathcal{L}_{\text{total}}$ for end-to-end fine-tuning (with λ_{deq}, λ_{lin}, λ_{crf}, and λ_{hal} set to 0 as there are no stage-wise supervisions), and replace the output of Hallucination-Net $\hat{H}$ with refined HDR image $\hat{H}_{\text{ref}}$.

4. Experimental Results

We first describe our experimental settings and evaluation metrics. Next, we present quantitative and qualitative comparisons with the state-of-the-art single-image HDR reconstruction algorithms. We then analyze the contributions of individual modules to justify our design choices.

4.1. Experiment setups

Datasets. For training and evaluating single-image HDR reconstruction algorithms, we construct two HDR image datasets: HDR-SYNTH and HDR-REAL. We also evaluate our method on two publicly available datasets: RAISE (RAW-jpeg pairs) [10] and HDR-EYE [42].

Evaluation metrics. We adopt the HDR-VDP-2 [39] to evaluate the accuracy of HDR reconstruction. We normalize both the predicted HDR and reference ground-truth HDR images with the processing steps in [40]. We also evaluate the PSNR, SSIM, and perceptual score with the LPIPS metric [57] on the tone-mapped HDR images in the supplementary material.

4.2. Comparisons with state-of-the-art methods

We compare the proposed method with five recent CNN-based approaches: HDRCNN [14], DrTMO [15], ExpandNet [40], Deep chain HDRI [29], and Deep recursive HDRI [30]. As the ExpandNet does not provide the code for training, we only compare with their released pre-trained model. Both the Deep chain HDRI and Deep recursive HDRI methods do not provide their pre-trained models, so we compare with the results on the HDR-EYE dataset reported in their papers.

We first train our model on the training set of the HDR-SYNTH dataset (denoted by Ours) and the fine-tune on the training set of the HDR-REAL dataset (denoted by Ours+). For fair comparisons, we also re-train the HDRCNN and DrTMO models with both the HDR-SYNTH and HDR-REAL datasets (denoted by HDRCNN+ and DrTMO+). We provide more comparisons with the pre-trained models of HDRCNN and DrTMO and the our results from each training stage in the supplementary material.

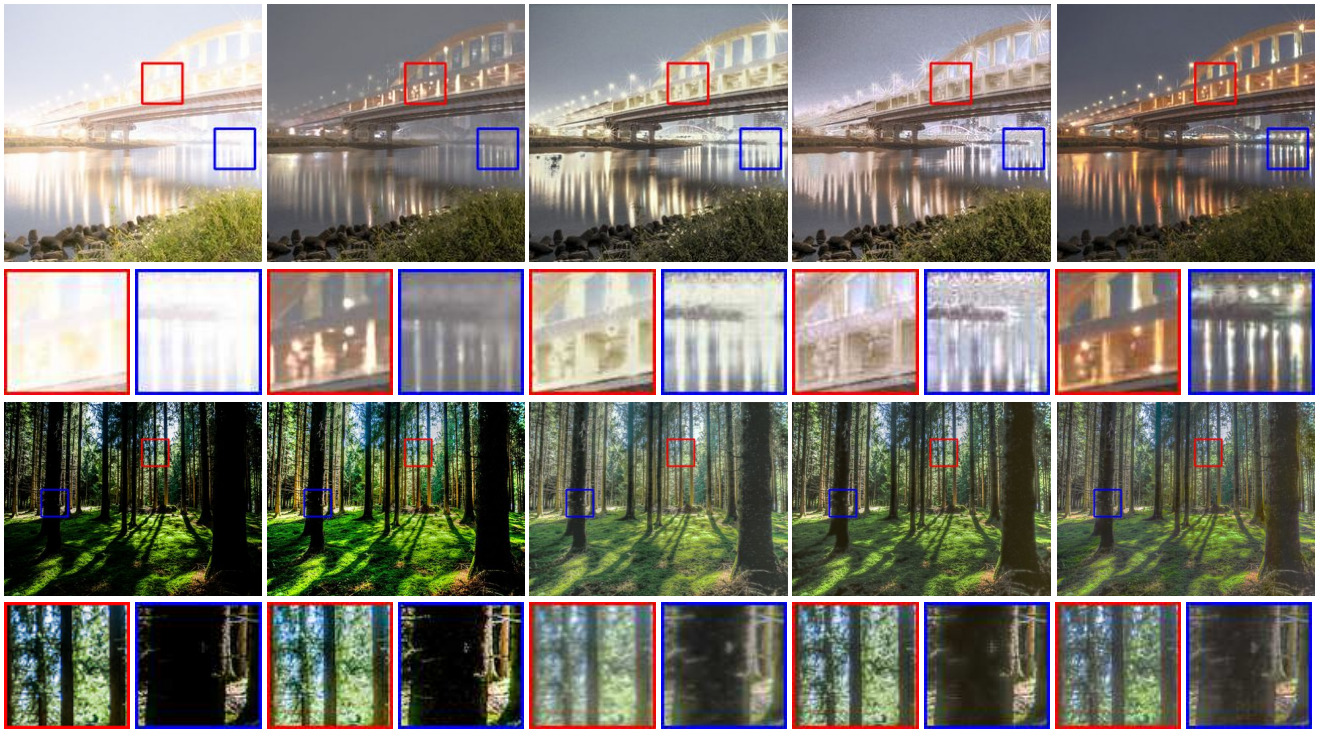
Quantitative comparisons. Table 1 shows the average HDR-VDP-2 scores on the HDR-SYNTH, HDR-REAL, RAISE, and HDR-EYE datasets. The proposed method performs favorably against the state-of-the-art methods on all four datasets. After fine-tuning on the HDR-REAL training set, the performance of our model (Ours+) is further improved by 1.57 on HDR-REAL, 0.41 on the RAISE, and 0.5 on HDR-EYE datasets, respectively.

Visual comparisons. Figure 6 compares the proposed model with existing methods on a real image captured using NIKON D90 provided by HDR-REAL and an example provided in [15]. We note that both two examples in Figure 6 come from unknown camera pipeline, and there are no ground-truth HDRs. In general, the HDRCNN [14] often generates overly-bright results and suffers from noise in the under-exposed regions as an aggressive and fixed inverse CRF x^2 is used. The results of the DrTMO [15] often looks blurry or washed-out. The ExpandNet [40] cannot restore the details well in the under-exposed regions and generates visual artifacts in the over-exposed regions, such as sky. Due to the space limit, we provide more visual comparisons in the supplementary material.

User study. We conduct a user study to evaluate the human preference on HDR images. We adopt the paired comparison [28, 47], where users are asked to select a preferred image from a pair of images in each comparison. We design the user study with the following two settings: (1) *With-*

Table 1: **Quantitative comparison on HDR images with existing methods.** * represents that the model is re-trained on our synthetic training data and + is fine-tuned on both synthetic and real training data. **Red** text indicates the best and blue text indicates the best performing state-of-the-art method.

Method	Training dataset	HDR-SYNTH	HDR-REAL	RAISE [10]	HDR-EYE [42]
HDRCNN+ [14]	HDR-SYNTH + HDR-REAL	55.51 ± 6.64	51.38 ± 7.17	56.51 ± 4.33	51.08 ± 5.84
DrTMO+ [15]	HDR-SYNTH + HDR-REAL	56.41 ± 7.20	50.77 ± 7.78	57.92 ± 3.69	51.26 ± 5.94
ExpandNet [40]	Pre-trained model of [40]	53.55 ± 4.98	48.67 ± 6.46	54.62 ± 1.99	50.43 ± 5.49
Deep chain HDRI [29]	Pre-trained model of [29]	-	-	-	49.80 ± 5.97
Deep recursive HDRI [30]	Pre-trained model of [30]	-	-	-	48.85 ± 4.91
Ours*	HDR-SYNTH	$\mathbf{60.11 \pm 6.10}$	51.59 ± 7.42	58.80 ± 3.91	52.66 ± 5.64
Ours+	HDR-SYNTH + HDR-REAL	59.52 ± 6.02	$\mathbf{53.16 \pm 7.19}$	$\mathbf{59.21 \pm 3.68}$	$\mathbf{53.16 \pm 5.92}$



(a) Input LDR (b) HDRCNN+ (c) DrTMO+ (d) ExpandNet (e) Ours+

Figure 6: **Visual comparison on real input image.** The example on the top is captured by NIKON D90 from HDR-REAL, and the bottom one is from DrTMO [15]. The HDRCNN [14] often suffers from noise, banding artifacts or over-saturated colors in the under-exposed regions. The DrTMO [15] cannot handle over-exposed regions well and leads to blurry and low-contrast results. The ExpandNet [40] generates artifacts in the over-exposed regions. In contrast, our method restores fine details in both the under-exposed and over-exposed regions and renders visually pleasing results.

reference test: We show both the input LDR and the ground-truth HDR images as reference. This test evaluates the *faithfulness* of the reconstructed HDR image to the ground-truth. (2) *No-reference test*: The input LDR and ground-truth HDR images are not provided. This test mainly compares the *visual quality* of two reconstructed HDR images.

We evaluate all 70 HDR images in the HDR-REAL test set. We compare the proposed method with the HDR-CNN [14], DrTMO [15], and ExpandNet [40]. We ask each participant to compare 30 pairs of images and collect the results from a total of 200 unique participants. Figure 7 reports the percentages of the head-to-head comparisons in which users prefer our method over the HDRCNN, DrTMO, and ExpandNet. Overall, there are 70% and 69% of users prefer our results in the with-reference and no-reference tests, respectively. Both user studies show that the proposed method performs well to human subjective perception.

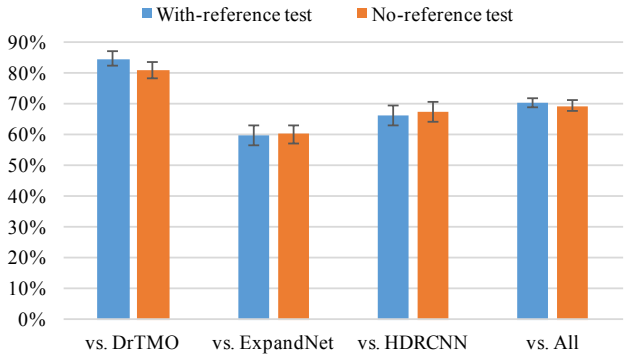


Figure 7: **Results of user study.** Our results are preferred by users in both with-reference and no-reference tests.

Table 2: **Comparisons on Dequantization-Net.** Our Dequantization-Net restores the missing details due to quantization and outperforms existing methods.

Method	PSNR (↑)	SSIM (↑)
w/o dequantization	33.86 ± 6.96	0.9946 ± 0.0109
Hou et al. [18]	33.79 ± 6.72	0.9936 ± 0.0110
Liu et al. [35]	34.83 ± 6.04	0.9954 ± 0.0073
Dequantization-Net (Ours)	$\mathbf{35.87 \pm 6.11}$	$\mathbf{0.9955 \pm 0.0070}$

4.3. Ablation studies

In this section, we evaluate the contributions of individual components using the HDR-SYNTH test set.

Dequantization. We consider the LDR images as the input and the image $I_n = \mathcal{F}(\mathcal{C}(H))$ synthesized from the HDR images as the ground-truth of the dequantization procedure. We compare our Dequantization-Net with CNN-based models [18, 35]. Table 2 shows the quantitative comparisons of dequantized images, where our method performs better than other approaches.

Linearization. Our Linearization-Net takes as input the non-linear LDR image, Sobel filter responses, and histogram features to estimate an inverse CRF. To validate the effectiveness of these factors, we train our Linearization-Net with different combinations of the edge and histogram features. Table 3 shows the reconstruction error of the inverse CRF and the PSNR between the output of our Linearization-Net and the corresponding ground-truth image $I_c = \mathcal{C}(H)$. The edge and histogram features help predict more accurate inverse CRFs. The monotonically increasing constraint further boosts the reconstruction performance on both the inverse CRFs and the linear images.

Hallucination. We start with the architecture of Eilertsen et al. [14], which does not enforce the predicted residuals being positive. As shown in Table 4, our model design (predicting positive residuals) can improve the performance by 1.19 HDR-VDP-2 scores. By replacing the transposed convolution with the resize convolution in the decoder, our model effectively reduces the checkerboard artifacts. Furthermore, introducing the perceptual loss for training not only improves the HDR-VDP-2 scores but also helps the model to predict more realistic details. We provide visual comparisons in the supplementary material.

Table 3: **Analysis on alternatives of Linearization-Net.** We demonstrate the edge and histogram features and monotonically increasing constraint are effective to improve the performance of our Linearization-Net.

Image	Edge	Histogram	Monotonically increasing	L2 error (↓) of inverse CRF	PSNR (↑) of linear image
✓	-	-	-	2.00 ± 3.15	33.43 ± 7.03
✓	✓	-	-	1.66 ± 2.93	34.31 ± 6.94
✓	-	✓	-	1.61 ± 3.03	34.51 ± 7.14
✓	✓	✓	-	1.58 ± 2.73	34.53 ± 6.83
✓	✓	✓	✓	$\mathbf{1.56 \pm 2.52}$	$\mathbf{34.64 \pm 6.73}$

Table 4: **Analysis on alternatives of Hallucination-Net.** With the positive residual learning, the model predicts physically accurate values within the over-exposed regions. The resize convolution reduces the checkerboard artifacts, while the perceptual loss helps generate realistic details.

Positive residual	Resize convolution	Perceptual loss	HDR-VDP-2 (↑)
-	-	-	63.60 ± 15.32
✓	-	-	64.79 ± 15.89
✓	✓	-	64.52 ± 16.05
✓	✓	✓	$\mathbf{66.31 \pm 15.82}$

End-to-end training from scratch. To demonstrate the effectiveness of explicitly reversing the camera pipeline, we train our entire model (including all sub-networks) from scratch without any intermediate supervisions. Compared to the proposed model shown in Table 1, the performance of such a model drops significantly (-4.43 and -3.48 HDR-VDP-2 scores in the HDR-SYNTH and HDR-REAL datasets, respectively). It shows that our stage-wise training is effective, and the performance improvement does not come from the increase of network capacity.

5. Conclusions

We have presented a novel method for single-image HDR reconstruction. Our key insight is to leverage the domain knowledge of the LDR image formation pipeline for designing network modules and learning to *reverse* the imaging process. Explicitly modeling the camera pipeline allows us to impose physical constraints for network training and therefore leads to improved generalization to unseen scenes. Extensive experiments and comparisons validate the effectiveness of our approach to restore visually pleasing details for a wide variety of challenging scenes.

Acknowledgments. This work is supported in part by NSF CAREER (#1149783), NSF CRII (#1755785), MOST 109-2634-F-002-032, MediaTek Inc. and gifts from Adobe, Toyota, Panasonic, Samsung, NEC, Verisk, and Nvidia.

References

[1] Ahmet Oğuz Akyüz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bülthoff. Do hdr displays support ldr content?: A psychophysical evaluation. *ACM TOG*, 2007. 2

[2] 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. *Computer Graphics Forum*, 2009. 2

[3] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *International conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, 2006. 2

[4] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Expanding low dynamic range videos for high dynamic range applications. In *Spring Conference on Computer Graphics*, 2008. 2

[5] Francesco Banterle, Patrick Ledda, Kurt Debattista, Alan Chalmers, and Marina Bloj. A framework for inverse tone mapping. *The Visual Computer*, 2007. 2

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 2009. 3

[7] MS Brown and SJ Kim. Understanding the in-camera image processing pipeline for computer vision. 2015. 3

[8] Ayan Chakrabarti, Ying Xiong, Baochen Sun, Trevor Darrell, Daniel Scharstein, Todd Zickler, and Kate Saenko. Modeling radiometric uncertainty for vision with tone-mapped color images. *TPAMI*, 2014. 3

[9] Scott J Daly and Xiaofan Feng. Decontouring: Prevention and removal of false contour artifacts. In *Human Vision and Electronic Imaging IX*, 2004. 2

[10] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *ACM MM*, 2015. 6, 7

[11] Paul Debevec. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*. 2006. 3

[12] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. *ACM TOG*, 1997. 1, 2, 3

[13] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *ACM TOG*, 2001. 3

[14] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM TOG*, 2017. 1, 2, 3, 5, 6, 7, 8

[15] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM TOG*, 2017. 1, 2, 3, 6, 7

[16] Michael D. Grossberg and Shree K. Nayar. What is the space of camera response functions? In *CVPR*, 2003. 2, 3, 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4

[18] Xianxu Hou and Guoping Qiu. Image companding and inverse halftoning using deep convolutional neural networks. *arXiv*, 2017. 2, 8

[19] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM TOG*, 2014. 3

[20] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 2014. 2

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 2017. 3

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[23] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM TOG*, 2017. 2

[24] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, 2016. 3

[25] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *ICIP*, 2006. 2

[26] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *TPAMI*, 2012. 3

[27] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, 2014. 2

[28] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 6

[29] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 2018. 6, 7

[30] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *ECCV*, 2018. 2, 6, 7

[31] Han Li and Pieter Peers. Crf-net: Single image radiometric calibration using CNNs. In *European Conference on Visual Media Production*, 2017. 3, 4

[32] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *CVPR*, 2018. 1, 3

[33] Stephen Lin, Jinwei Gu, Shuntaro Yamazaki, and Heung-Yeung Shum. Radiometric calibration from a single image. In *CVPR*, 2004. 3, 4

[34] Stephen Lin and Lei Zhang. Determining the radiometric response function from a single grayscale image. In *CVPR*, 2005. 3, 4

[35] Chang Liu, Xiaolin Wu, and Xiao Shu. Learning-based dequantization for image restoration against extremely poor illumination. *arXiv*, 2018. 2, 8

[36] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[37] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*, 2010. 2

[38] Steve Mann and Rosalind W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combin-

ing differently exposed pictures. In *Proceedings of IS&T*, 1995. 2

[39] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM TOG*, 2011. 6

[40] 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 *EG*, 2018. 1, 2, 3, 6, 7

[41] Yasuyuki Matsushita and Stephen Lin. Radiometric calibration from noise distributions. In *CVPR*, 2007. 3

[42] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in ldr and hdr images. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015. 6, 7

[43] Tian-Tsong Ng, Shih-Fu Chang, and Mao-Pei Tsui. Using geometry invariants for camera response function estimation. In *CVPR*, 2007. 3

[44] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 5

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[46] Photomatix. https://www.hdrsoft.com/. 3

[47] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. *ACM TOG*, 2010. 6

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[49] Qing Song, Guan-Ming Su, and Pamela C Cosman. Hardware-efficient debanding and visual enhancement filter for inverse tone mapped high dynamic range images and videos. In *ICIP*, 2016. 2

[50] Abhilash Srikantha and Désiré Sidibé. Ghost detection and removal for high dynamic range images: Recent advances. *Signal Processing: Image Communication*, 2012. 2

[51] Jun Takamatsu, Yasuyuki Matsushita, and Katsushi Ikeuchi. Estimating radiometric response functions from image noise variance. In *ECCV*, 2008. 3

[52] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, 2018. 2, 5

[53] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Lau Rynson. Image correction via deep reciprocating hdr transformation. In *CVPR*, 2018. 1, 2

[54] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 3

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *ICCV*, 2019. 3

[56] Jinsong Zhang and Jean-Francois Lalonde. Learning high dynamic range from outdoor panoramas. In *ICCV*, 2017. 1, 2

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[58] Yang Zhao, Ronggang Wang, Wei Jia, Wangmeng Zuo, Xiaoping Liu, and Wen Gao. Deep reconstruction of least significant bits for bit-depth expansion. *TIP*, 2019. 2