IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Single-Shot High Dynamic Range Imaging via Multiscale Convolutional Neural Network

AN GIA VIEN[1], (Student, IEEE), AND CHUL LEE[1], (Member, IEEE)
[1]Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Chul Lee (e-mail: chullee@dongguk.edu).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. NRF-2019R1A2C4069806).

ABSTRACT We propose a single-shot high dynamic range (HDR) imaging algorithm with row-wise varying exposures in a single raw image based on a deep convolutional neural network (CNN). We first convert a raw Bayer input image into a radiance map by calibrating rows with different exposures, and then we design a new CNN model to restore missing information at the under- and over-exposed pixels and reconstruct color information from the raw radiance map. The proposed CNN model consists of three branch networks to obtain multiscale feature maps for an image. To effectively estimate the high-quality HDR images, we develop a robust loss function that considers the human visual system (HVS) model, color perception model, and multiscale contrast. Experimental results on both synthetic and captured real images demonstrate that the proposed algorithm can achieve synthesis results of significantly higher quality than conventional algorithms in terms of structure, color, and visual artifacts.

INDEX TERMS Spatially varying exposure (SVE) image, high dynamic range (HDR) imaging, convolutional neural network (CNN), and human visual system (HVS).

I. INTRODUCTION

DESPITE significant recent advances in digital imaging technology, conventional cameras can capture only a limited range of intensity levels perceptible by the human eye. For example, images captured by conventional cameras cannot represent bright light sources or scenes with a bright exterior and a dark interior. To overcome the limitations of these conventional imaging systems, high dynamic range (HDR) imaging was developed to represent, store, and reproduce the full visible luminance range of real-world scenes [1], [2]. Because of the advantages of HDR imaging over conventional imaging systems, extensive research efforts have been made to acquire high-quality HDR images. One approach is to design specialized camera systems to extend the dynamic range of conventional cameras [3]–[5]. For example, a beam splitter to reflect the light on multiple sensors [3], a modulo sensor to keep only the least significant bits [4], and programmable sensors to optimize per-pixel shutter functions [5] have been developed. However, the devices designed to acquire HDR images directly are too complex and expensive to be used in practical applications. Instead, most researches have focused on the effective synthesis of HDR images using conventional low dynamic range (LDR) imaging devices.

The most common approach to HDR image acquisition with conventional cameras is to combine multiple LDR images captured with different exposure times [6]–[9]. The main challenge of this approach is the misalignment of the images due to movements of the camera or objects in the scene, which results in ghosting artifacts in the synthesized HDR image. To address this problem, many algorithms for the removal of such ghosting artifacts in HDR synthesis have been proposed [10]–[19]. Conventional ghost-free HDR imaging algorithms can be categorized into three groups differentiated by how they handle motion. The first category of algorithms attempts to estimate the correspondences between the input LDR images in the stack and then merge the aligned images [10]–[12]. However, when the reference image contains poorly exposed regions, these algorithms may fail to find accurate correspondences, thus degrading the quality of the results. The second category of algorithms alleviates the contributions of regions that contain object movements by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

identifying the ghost regions [13]–[15]. These algorithms achieve high-quality results but may fail when the scene contains objects with complex motions. The algorithms of the third category attempt to detect ghost regions and estimate correspondences simultaneously [16]–[19]. These algorithms formulate and solve joint optimization problems and achieve high-quality synthesis results. However, a common drawback of these ghost-free HDR imaging algorithms is that they generally require high computational costs for correspondence estimation and numerical optimization.

Recently, convolutional neural networks (CNNs) have been applied to HDR imaging. CNN-based algorithms reconstruct an HDR image from a stack of LDR images captured with different exposure times by using encoder–decoder structures that learn to handle misaligned pixels and merge LDR images into the final HDR images [20]–[25]. These models have the advantage of exploiting information learned from training data and compensating for missing details in the HDR synthesis. Although each algorithm addresses an important issue, no algorithm has enough robustness to completely handle the misalignment caused by object motion in images. Concurrently, another approach attempts to infer an HDR image directly from a single LDR image using CNNs [26]–[31], which is also known as inverse tone-mapping (ITM). Although the networks in this approach can recover missing details in under- and over-exposed regions, one limitation of this approach is its high dependence on a single input LDR image, thereby lacking underlying information. For example, if an image has a large area of under- and over-exposure, the ITM algorithms may fail to restore those regions faithfully because no information is available in their neighboring regions.

Another effective approach to avoid ghosting artifacts in synthesizing an HDR image is to employ spatially varying exposure (SVE) images [32]. In SVE imaging, a scene is captured with pixel-wise varying exposures in a single image, and then multiple sub-images corresponding to each exposure are merged to synthesize the HDR image. The SVE-based algorithm, which is called single-shot HDR imaging, benefits from multiple exposures in a single image and thereby exploits more information than the ITM approaches, where a single exposure is used. Because of these merits, several algorithms have recently been developed, which improve the HDR synthesis performance [33]–[39]. In single-shot HDR imaging, poorly exposed pixels in the SVE image are recovered using differently exposed neighboring pixels. For example, Gu *et al.* [33] applied cubic interpolation with optical flow. Cho *et al.* [37] employed the bilateral filter with edge-directional weights and obtained HDR images using a demosaicing algorithm. However, interpolation makes these algorithms susceptible to causing blurring artifacts in the synthesized HDR images when they fail to faithfully recover the missing pixels in poorly exposed regions. Choi *et al.* [39] developed sparse representation with dual dictionary learning to construct HDR videos via SVE frames. Although their algorithm provides high-quality HDR frames, it relies on a dictionary trained from a set of LDR images to synthesize the HDR frames.

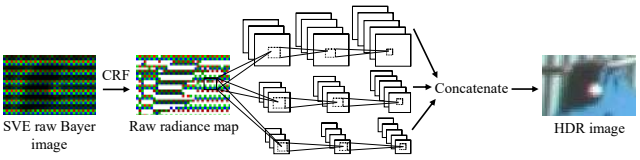


FIGURE 1. Overview of the proposed end-to-end CNN architecture for directly synthesizing full-color HDR images from raw SVE Bayer images. The proposed CNN is composed of multiple branches to mine different aspects of an input image. Missing values in the raw radiance map, which are caused by poor exposures, are illustrated in white.

Recently, deep learning-based approaches to single-shot HDR imaging were developed [40], [41], which can reconstruct over- and under-exposed pixels in raw SVE Bayer images. While these algorithms synthesize high-quality HDR images, two persistent issues must be addressed. First, these algorithms use conventional demosaicing techniques to obtain a full-color HDR image after recovering missing pixels, which does not confer any additional information to the final output. Moreover, the demosaicing increases the computational complexity and may yield color artifacts. Second, both networks in [40], [41] are optimized with an L_2 loss, which is sensitive to large and small errors and does not correlate well to human perception [42]–[44]. Therefore, these algorithms often provide over-smoothed HDR images that are inconsistent with human visual perception, as we will discuss in Section IV.

In this work, to address the aforementioned issues, we propose a novel single-shot HDR imaging algorithm using a deep CNN that takes raw SVE Bayer images as input and directly synthesizes full-color HDR images in an end-to-end manner, as shown in Figure 1. Specifically, we make the following contributions:

- We develop an end-to-end CNN to synthesize full-color HDR images directly from raw SVE Bayer images, as shown in Figure 1, by learning semantic information of an input image. Specifically, we design a multibranch network architecture, each branch of which mines the features of a particular aspect of the input image to reconstruct the HDR image progressively in a coarse-to-fine manner.
- We develop a robust loss function, which incorporates the human visual system (HVS) model, color perception model, and multiscale contrast, to recover under- and over-exposed regions better and minimize perceptual distortions.
- We experimentally show, with both synthetic and real image datasets, that the proposed CNN model trained with the robust loss function produces clear and natural-looking HDR images of significantly higher quality than the conventional algorithms [33], [37], [39]–[41].

The remainder of this paper is organized as follows: Section II describes the proposed single-shot HDR synthesis

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

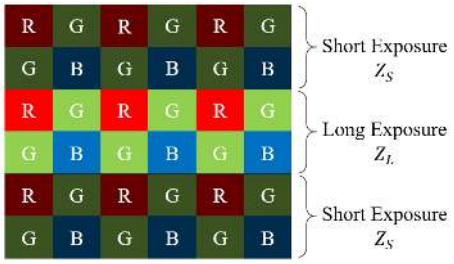


FIGURE 2. Illustration of the raw SVE Bayer image with row-wise varying exposures.

algorithm. Section III presents the robust loss function, and Section IV discusses the experimental results. Finally, Section V concludes this paper.

II. PROPOSED SINGLE-SHOT HDR IMAGING

Figure 1 shows an overview of the proposed single-shot HDR imaging algorithm, which takes an SVE raw Bayer image as input and reconstructs a full-color HDR image. First, we convert the input SVE image into a raw radiance map. Then, the proposed multibranch network synthesizes the HDR image in a coarse-to-fine manner by learning semantic information of the input image. In this section, we first present an overview of the raw SVE Bayer images and the conversion of these images into radiance maps. Then, we provide the details of our network architecture for HDR imaging.

A. SPATIALLY VARYING EXPOSURE (SVE) IMAGE

In this work, we consider the SVE image with row-wise varying exposures in a single raw Bayer image using two different exposure times: a short exposure time Δt_S and a long exposure time Δt_L, as in [37]–[40]. We also consider the 2×2 RGGB color filter array, as shown in Figure 2. The input image Z with a resolution of $H \times W$ is modeled as

$$Z = \begin{cases} Z_S, & \text{on } 4n+1 \text{ and } 4n+2\text{-th rows}, \\ Z_L, & \text{on } 4n+3 \text{ and } 4n+4\text{-th rows}, \end{cases} \tag{1}$$

where Z_S and Z_L denote the short- and long-exposure sub-images, respectively, and $n = 0, 1, \ldots, \frac{H}{4}$. In this work, we assume that Z is an 8-bit image.

B. RADIANCE MAP RECONSTRUCTION

The input SVE raw Bayer image Z is first converted into the radiance map E for HDR imaging, assuming that the camera response function (CRF) [7] is known *a priori*. In this work, the CRF is calibrated with actual luminance values. Specifically, let f denote the CRF, then the image acquisition can be modeled as

$$Z_j = f(E\Delta t_j), \tag{2}$$

where $j \in \{S, L\}$ indexes the exposure time. As f is invertible [7], we can normalize the model in (2) using the logarithm as

$$g(Z_j) = \ln(E) + \ln(\Delta t_j), \tag{3}$$

where $g = \ln(f^{-1})$. We then obtain radiance map E for the pixel value in Z_j with exposure time Δt_j by

$$\ln(E) = g(Z_j) - \ln(\Delta t_j). \tag{4}$$

As the input raw SVE Bayer image Z contains poorly exposed pixels, radiance map E obtained by (4) includes unreliable radiance values at the corresponding pixel locations, which are represented by white in Figures. 1 and 3. We define the well-exposed pixels as $Z_{\text{th}} \leq Z \leq 255 - Z_{\text{th}}$, where Z_{th} is the threshold value. In this work, we fix $Z_{\text{th}} = 15$. Additionally, two color values in each pixel should be estimated to obtain a full-color radiance map from E.

C. MULTIPLE BRANCH NETWORK FOR HDR IMAGING

Multibranch networks have been shown to achieve better results than single networks in image enhancement [45], [46] and HDR imaging [28], [47]. Inspired by these recent works, in this work, we develop a multiscale CNN to reconstruct missing information effectively and synthesize high-quality HDR images. Note that the missing information is caused by both poor exposures and the Bayer filter of the image sensor. The architecture of the proposed network, which takes the raw radiance map E as input and reconstructs the full-color HDR image, is shown in Figure 3. Specifically, the proposed network is designed to synthesize a full-color HDR image directly with three different network branches: global, medium, and local. Each branch is a convolutional encoder–decoder network that accepts a radiance map as input with a different scale of initial feature maps and decodes them to feature maps at the final resolution. In other words, each branch is responsible for a particular aspect of an input image: global branch for high-level features, medium branch for medium-level features, the local branch for fine details.

Although the proposed CNN builds on recent works, it has the following novelties over conventional algorithms. First, each branch of the proposed CNN is an encoder–decoder structure, whereas conventional multibranch networks, *e.g.*, [28], are constructed without downsampling and upsampling. By progressively downsampling the input in the encoder, each branch of the network can fully exploit well-exposed information with a larger receptive field, and thus can learn semantic information with high-level understanding of the images more faithfully than conventional networks. Second, initial feature maps of each branch are learned at different resolutions using different kernel sizes, unlike in conventional multibranch networks. This enables the proposed network to extract more important information with larger receptive fields. Finally, the proposed network is trained with a robust loss function, which considers human perception on luminance, chrominance, and contrast. The network is described in detail below.

The four types of blocks used in the proposed network are shown in Figure 3. The convolutional block, which is composed of a convolutional layer (Conv), batch normalization (BN) [48], and parametric rectified linear unit (PReLU) [49], extracts feature maps from the input. Note that both positive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

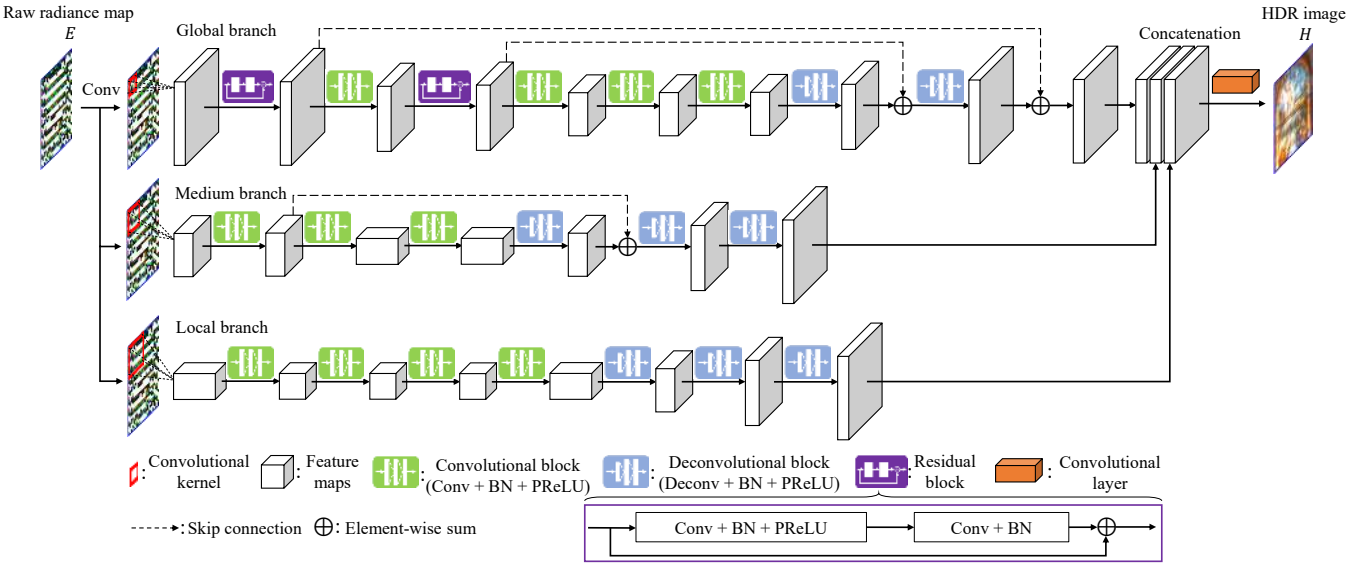


FIGURE 3. Proposed CNN architecture, which is composed of three branches: global, medium, and local. The output features of the branches are concatenated, and then fused using 3×3 convolutions to form a full-color HDR image.

TABLE 1. Architecture of the global branch network, which contains five convolutional blocks, two residual blocks, two deconvolutional blocks, and two skip connections.

Layer	Activation size
Input	$32 \times 32 \times 1$
$3 \times 3 \times 128$ conv, stride 1, pad 1, BN	$32 \times 32 \times 128$
Resdiual block	$32 \times 32 \times 128$
$2 \times 2 \times 256$ conv, stride 2, BN	$16 \times 16 \times 256$
Resdiual block	$16 \times 16 \times 256$
$2 \times 2 \times 512$ conv, stride 2, BN	$8 \times 8 \times 512$
$3 \times 3 \times 512$ conv, stride 1, pad 1, BN	$8 \times 8 \times 512$
$3 \times 3 \times 512$ conv, stride 1, pad 1, BN	$8 \times 8 \times 512$
$4 \times 4 \times 256$ deconv, stride 2, pad 1, BN	$16 \times 16 \times 256$
Skip connection	$16 \times 16 \times 256$
$4 \times 4 \times 128$ deconv, stride 2, pad 1, BN	$32 \times 32 \times 128$
Skip connection	$32 \times 32 \times 128$

TABLE 2. Architecture of the medium branch network, which contains a convolutional block, a residual block, and two deconvolutional blocks.

Layer	Activation size
Input	$32 \times 32 \times 1$
$4 \times 4 \times 512$ conv, stride 4, BN	$8 \times 8 \times 512$
$3 \times 3 \times 512$ conv, stride 1, BN	$8 \times 8 \times 512$
$2 \times 2 \times 1024$ conv, stride 2, BN	$4 \times 4 \times 1024$
$3 \times 3 \times 1024$ conv, stride 1, BN	$4 \times 4 \times 1024$
$4 \times 4 \times 512$ deconv, stride 2, pad 1, BN	$8 \times 8 \times 512$
Skip connection	$8 \times 8 \times 512$
$4 \times 4 \times 256$ deconv, stride 2, pad 1, BN	$16 \times 16 \times 256$
$4 \times 4 \times 128$ deconv, stride 2, pad 1, BN	$32 \times 32 \times 128$

and negative values contain important information for image restoration [46]. We employ the residual block [50] to learn larger receptive fields. The deconvolutional block that consists of the deconvolutional layer (Deconv), BN, and PReLU upsamples the feature maps. The skip connection connects the feature maps of two layers using element-wise sum to construct the new feature maps. Finally, the last convolutional layer in the proposed network includes three filters of size 3×3, stride 1, and zero padding to synthesize the HDR image from multiscale feature maps.

Global branch: The global branch learns large receptive fields that represent high-level features of the input. It first extracts initial feature maps with the same resolution as the input. As the global branch represents the coarse feature maps, we use the residual block [50] after convolutional layers in the encoder to recover unreliable information and missing color information of each pixel. Further, we use two additional convolution layers after a third convolution layer to enlarge the receptive field for learning high-level features. Further, to prevent information loss caused by the convolution operations, we use skip connections from residual blocks in the encoder to deconvolutional blocks in the decoder. Table 1 summarizes the details of the global branch network.

Medium branch: The medium branch in the network learns to represent mid-level feature maps and extracts useful neighboring features. Because the input raw radiance map E has information missing in under- and over-exposed regions, as well as missing color values to be estimated, the input is first downsampled by a factor of four using the convolutional block to obtain the initial feature maps. Three convolutional blocks are then used to ensure that the receptive field is large enough to extract the mid-level features. At the end of the encoder structure, one convolutional layer is used as a nonlinear mapping to connect the encoder features to the decoder. Table 2 lists the details of the medium branch

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

TABLE 3. Architecture of the local branch network, which contains two convolutional blocks and three deconvolutional blocks.

Layer	Activation size
Input	$32 \times 32 \times 1$
$8 \times 8 \times 1024$ conv, stride 8, BN	$4 \times 4 \times 1024$
$1 \times 1 \times 512$ conv, stride 1, BN	$4 \times 4 \times 512$
$3 \times 3 \times 512$ conv, stride 1, pad 1, BN	$4 \times 4 \times 512$
$3 \times 3 \times 512$ conv, stride 1, pad 1, BN	$4 \times 4 \times 512$
$1 \times 1 \times 1024$ conv, stride 1, BN	$4 \times 4 \times 1024$
$4 \times 4 \times 512$ deconv, stride 2, pad 1, BN	$8 \times 8 \times 512$
$4 \times 4 \times 256$ deconv, stride 2, pad 1, BN	$16 \times 16 \times 256$
$4 \times 4 \times 128$ deconv, stride 2, pad 1, BN	$32 \times 32 \times 128$

network.

Local branch: The local branch of the proposed network extracts the feature maps of the input at the finest resolution. To this end, the input raw radiance map is downsampled by a factor of eight by the convolutional block to provide the finest initial feature maps. The small receptive field of the finest features provides learning at the pixel-level to preserve high-frequency details in images. Convolutional layers with a kernel size of 1×1 are used to reduce the number of parameters. Table 3 summarizes the details of the local branch.

Fusion: The outputs of the global, medium, and local branches are concatenated to construct multiscale features. Then, a convolutional layer with a kernel of size 3×3, stride 1, and padding 1 is applied to fuse the coarse-to-fine feature maps extracted from the three branches, providing the full-color HDR image.

III. LOSS FUNCTIONS

The L_2 loss was used in [40]. However, the L_2 loss penalizes larger errors and is tolerant of smaller errors, regardless of the underlying structures in an image [51]. Furthermore, the L_2 loss does not correlate well with the human perception of image quality [43]. Therefore, An and Lee's algorithm [40] results in visible artifacts in the synthesized results, especially in highly textured regions. To overcome these limitations, we develop a new and robust loss function for HDR imaging that considers the HVS model, color perception, and multiscale contrast.

Given a reconstructed HDR image H and the ground-truth HDR image Y, we define the robust loss L_{Robust} as

$$L_{\text{Robust}} = L_{\text{HVS}}(H,Y) + \alpha L_{\text{C}}(H,Y) + \beta L_{\text{MC}}(H,Y), \quad (5)$$

where L_{HVS}, L_{C}, and L_{MC} are the HVS, chromatic, and multiscale contrast losses, respectively. The hyper-parameters α and β in (5) control the balance between the three losses. In this work, unless otherwise specified, α and β are fixed to 0.5 and 0.75, respectively. We now describe the different losses in (5).

A. HUMAN VISUAL SYSTEM (HVS) LOSS

For the HVS loss, we employ the just-noticeable difference (JND) to consider the human perception model [1], [44]. Specifically, the JND $l(\cdot)$ determines the perceptually uniform integer values via conversion from absolute luminance values y in units of cd/m^2. In this work, we use the model in [52], [53], which is given by

$$l(y) = \begin{cases} ay, & \text{if } y < y_l, \\ by^c + d, & \text{if } y_l \le y < y_h, \\ e\ln(y) + f, & \text{if } y_h \le y, \end{cases} \quad (6)$$

where the coefficients a, b, c, d, e, f, y_l, and y_h are 17.554, 826.81, 0.10013, -884.17, 209.16, -731.28, 5.6046, and 10.469, respectively [52], [53].

Then, we define the HVS loss as

$$L_{\text{HVS}}(H,Y) = \frac{1}{N}\sum_{i=1}^{N} \|l(H_i) - l(Y_i)\|_1, \quad (7)$$

where N is the number of training samples. The L_1 loss function is less sensitive to outliers than the L_2 loss function in [40], which yields visual artifacts [43]. In the backpropagation, the backward function computes the partial derivative of the loss in (7) with respect to each luminance value in H as

$$\frac{\partial L_{\text{HVS}}}{\partial H} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial l(H_i)}{\partial H_i} \text{sign}\big(l(H_i) - l(Y_i)\big) \quad (8)$$

with

$$\frac{\partial l(y)}{\partial y} = \begin{cases} a, & \text{if } y < y_l, \\ bcy^{c-1}, & \text{if } y_l \le y < y_h, \\ \frac{e}{y}, & \text{if } y_h \le y, \end{cases} \quad (9)$$

and

$$\text{sign}(x) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ 1, & \text{if } x > 0. \end{cases} \quad (10)$$

In (8), since the L_1 norm is not differentiable, we use its subgradient instead of the gradient.

B. CHROMATIC LOSS

As the proposed network incorporates demosaicing, we also define the chromatic loss to minimize the visible color differences, based on the color perception model. To this end, we first compute the chrominance values u and v [53], given by

$$u = 410 \cdot \frac{4X}{X + 15Y + 3Z}, \quad (11)$$

$$v = 410 \cdot \frac{9Y}{X + 15Y + 3Z}, \quad (12)$$

where X, Y, and Z are the tristimulus values in the XYZ color space.

Then, we define the chromatic loss as

$$L_{\text{C}}(H,Y) = \frac{1}{2N}\sum_{i=1}^{N} \Big(\|u(H_i) - u(Y_i)\|_1 + \|v(H_i) - v(Y_i)\|_1 \Big), \quad (13)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

where the parameter $\frac{1}{2}$ is used to compute the average of the two chrominance differences. We also use the L_1 loss to reduce visual artifacts. The derivative of L_{C} with respect to each luminance value in H for the backpropagation is given by

$$\frac{\partial L_{\mathrm{C}}}{\partial H} = \frac{1}{2N}\sum_{i=1}^{N}\Bigg\{\frac{\partial u(H_i)}{\partial H_i}\,\mathrm{sign}\big(u(H_i)-u(Y_i)\big) + \frac{\partial v(H_i)}{\partial H_i}\,\mathrm{sign}\big(v(H_i)-v(Y_i)\big)\Bigg\} \quad (14)$$

with

$$\frac{\partial u(H_i^c)}{\partial H_i^c} = \begin{cases} 410\cdot\left(\frac{4t_{11}}{A} - \frac{4(t_{11}+15t_{21}+3t_{31})X}{A^2}\right), & c=R,\\ 410\cdot\left(\frac{4t_{12}}{A} - \frac{4(t_{12}+15t_{22}+3t_{32})X}{A^2}\right), & c=G,\\ 410\cdot\left(\frac{4t_{13}}{A} - \frac{4(t_{13}+15t_{23}+3t_{33})X}{A^2}\right), & c=B, \end{cases} \quad (15)$$

and

$$\frac{\partial v(H_i^c)}{\partial H_i^c} = \begin{cases} 410\cdot\left(\frac{9t_{21}}{A} - \frac{9(t_{11}+15t_{21}+3t_{31})Y}{A^2}\right), & c=R,\\ 410\cdot\left(\frac{9t_{22}}{A} - \frac{9(t_{12}+15t_{22}+3t_{32})Y}{A^2}\right), & c=G,\\ 410\cdot\left(\frac{9t_{23}}{A} - \frac{9(t_{13}+15t_{23}+3t_{33})Y}{A^2}\right), & c=B, \end{cases} \quad (16)$$

where $c \in \{R,G,B\}$ indexes the RGB color channel and $t_{ij} \in \mathbb{R}^{3\times 3}$ is the value of the transformation matrix that converts the RGB color space to the XYZ color space [1]. Further, we denote $A = X+15Y+3Z$ for simpler notations.

C. MULTISCALE CONTRAST LOSS

The contrast of the images can be considered in CNNs using the structural similarity index (SSIM) in the loss function [43], [46]. The SSIM is computed using three terms: luminance, contrast, and structure. However, as the luminance component is already used in the HVS loss L_{HVS} in (7), we only consider contrast and structure as

$$cs(x,y) = \frac{2\sigma_{xy}+C_2}{\sigma_x^2+\sigma_y^2}, \quad (17)$$

where σ_x and σ_y are the standard deviations for images x and y, respectively, and σ_{xy} is the cross-covariance.

In this work, to consider the contrast and structure of the image more faithfully at different scales, we employ multiscale SSIM (MS-SSIM) to compute $cs(\cdot)$ in (17). Additionally, we compute $cs(\cdot)$ in the perceptually uniform domain via (6). Specifically, the multiscale contrast loss L_{MC} is defined as

$$L_{\mathrm{MC}}(H,Y) = 1 - \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{M} cs_j\big(l(H_i),l(Y_i)\big), \quad (18)$$

where M is the number of scales and cs_j denotes the contrast and structure term in (17) at the jth scale. Then, as was similarly done in [43], the derivative of the multiscale contrast loss is given by

$$\frac{\partial L_{\mathrm{MC}}}{\partial H} = -\frac{1}{N}\sum_{i=1}^{N}\Bigg\{\Bigg(\sum_{j=1}^{M}\frac{1}{cs_j(l(H_i),l(Y_i))} \times \frac{\partial cs_j(l(H_i),l(Y_i))}{\partial H_i}\Bigg)\prod_{k=1}^{M} cs_k(l(H_i),l(Y_i))\Bigg\} \quad (19)$$

with

$$\frac{\partial cs_j(l(H_i),l(Y_i))}{\partial H_i} = \frac{2}{\sigma^2_{l(H_i)}+\sigma^2_{l(Y_i)}+C_2}\circ\left(G_\sigma * \frac{\partial l(H_i)}{\partial H_i}\right) \circ\Big\{\big(l(Y_i)-\mu_{l(Y_i)}\big) - cs_j(l(H_i),l(Y_i))\circ\big(l(H_i)-\mu_{l(H_i)}\big)\Big\}, \quad (20)$$

where $\circ$ and $*$ denote the element-wise product and convolution operator, respectively; μ_{Y_i} and μ_{H_i} are the local means for images $l(Y_i)$ and $l(H_i)$, respectively; and σ_{Y_i} and σ_{H_i} are the standard deviations for images $l(Y_i)$ and $l(H_i)$, respectively.

IV. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed algorithm against four conventional single-shot HDR imaging algorithms: Gu *et al.*'s algorithm [33] and Cho *et al.*'s algorithm [37] are interpolation-based algorithms, Choi *et al.*'s algorithm [39] is a sparse representation model-based algorithm [39], and An and Lee's algorithm [40] and Çoğalan and Akyüz's algorithm [41] are learning-based algorithms. The performance of these algorithms is evaluated using synthetic noncalibrated and calibrated images, and captured real SVE images. To print the synthesized HDR images, we used the tone-mapping algorithm [54] in all the experiments.

We set the parameters α_s and α_d in Cho *et al.*'s algorithm [37] to 2 and 0.1, respectively, in all experiments. For Choi *et al.*'s algorithm [39], we used the patch size of 6×6 and fixed the sparsity regularization parameter λ to 0.15. For An and Lee's algorithm [40] and Çoğalan and Akyüz's algorithm [41], we retrained their networks with our dataset, which will be described subsequently. For reproducibility, we provide the source codes and pretrained models on our project website.[1]

A. EXPERIMENTAL SETTINGS

We collected 70 calibrated HDR images for training from the Fairchild's HDR database,[2] HDR-Eye dataset,[3] and calibrated HDR video sequences.[4] To avoid overfitting, we performed data augmentation to increase the size of training data as done in [55], [56]. Specifically, we used geometric transformations of 90, 180, and 270° rotations and horizontal flipping, thereby producing six additional augmented versions of

[1] https://github.com/viengiaan/Single-Shot-HDR-Imaging-with-MSCNN
[2] http://rit-mcsl.org/fairchild/HDRPS/HDRthumbs.html
[3] https://mmspg.epfl.ch/hdr-eye
[4] http://www.hdrv.org/Resources.php

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

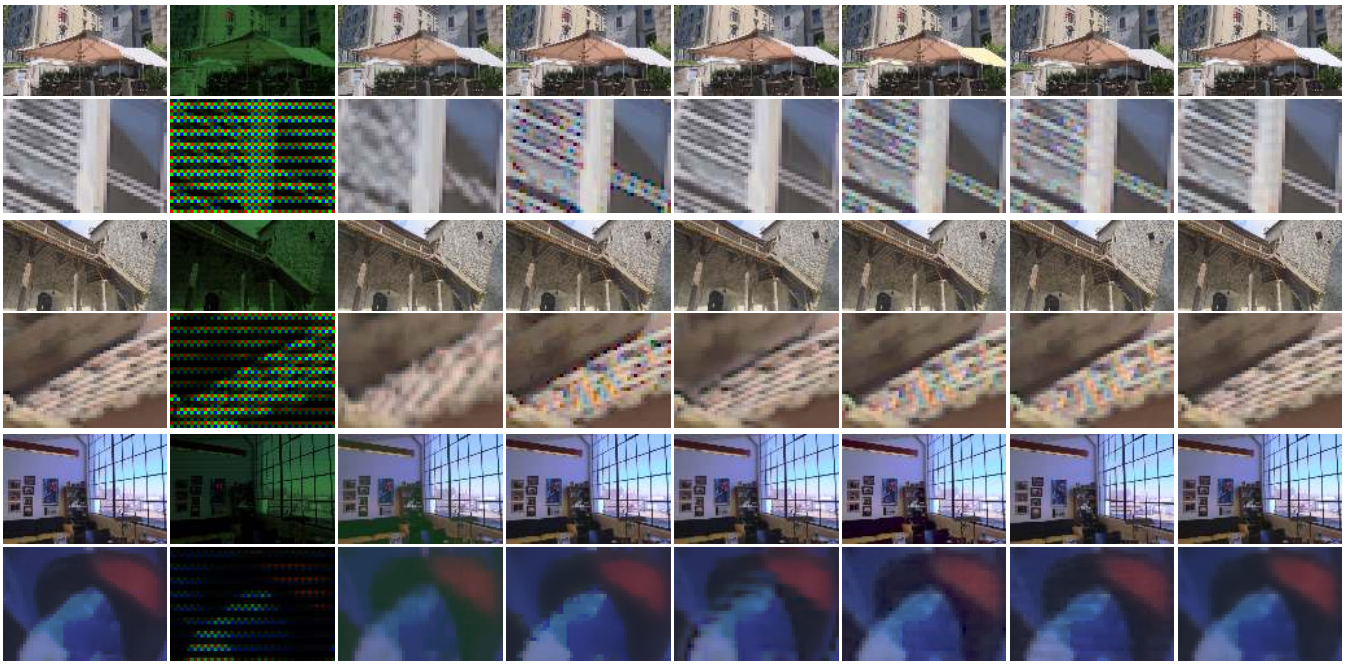


FIGURE 4. Synthesized results of the noncalibrated images. (a) Ground-truth, (b) synthetic raw Bayer images, (c) Gu *et al.*'s algorithm [33], (d) Cho *et al.*'s algorithm [37], (e) Choi *et al.*'s algorithm [39], (f) An and Lee's algorithm [40], (g) Çoğalan and Akyüz's algorithm [41] , and (h) the proposed algorithm. (Top) Full-resolution images. (Bottom) Magnified parts.

each image. Next, to synthesize a row-wise SVE image, we set the short and long exposure values as $EV = \{-1, +1\}$, and then split them into a raw image with the RGGB pattern as shown in Figure 2.

Further, we randomly selected 70% and 10% of the images for training and validation, respectively, and used the remaining images for testing. We extracted 200,000 patches of size 32×32 from the training images with a stride of 80. We used the recent stochastic gradient descent solver, Adam [57], with a batch size of 32 patches and learning rate of 10^{-4} with momentums $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All the experiments were performed using the Caffe library [58]. The training took approximately two days with an Nvidia Titan V GPU using a PC with a 3.30 GHz Intel® Core™ i9-10900 CPU and 32 GB memory.

B. SYNTHETIC IMAGES

We evaluate the performance of the proposed algorithm on two synthetic image datasets: noncalibrated and calibrated HDR images. The noncalibrated dataset contains 12 indoor and outdoor scene images, and the calibrated dataset consists of nine natural scene images. Because a noncalibrated HDR image is not calibrated in units of cd/m^2, we multiplied all pixel values by a single constant to convert them to approximate luminance values, as done in [52], [53]. Specifically, we assumed that the typical maximum luminance values of some objects in specific environments, *e.g.*, sky on a sunny day, are known *a priori*.

Figure 4 compares the synthesized results obtained by each algorithm from the raw SVE images in Figure 4(b), for the noncalibrated images. Gu *et al.*'s algorithm [33] in Figure 4(c) results in blurring and aliasing artifacts in the highly textured regions, *e.g.*, window, roof, and object texture. This is because their algorithm uses bicubic interpolation to upsample the subimages Z_S and Z_L to synthesize the HDR image. Cho *et al.*'s algorithm [37] in Figure 4(d) yields better results than Gu *et al.*'s algorithm, but still produces artifacts. For example, the high-frequency details in the window and roof are lost because the algorithm uses bilateral interpolation to recover missing values. Choi *et al.*'s algorithm [39] in Figure 4(e) recovers the shape of the texture but yields blurring artifacts around the edges due to the failure of sparse reconstruction, whose performance depends on dual dictionary learning. As shown in Figures 4(f) and (g), An and Lee's algorithm [40] and Çoğalan and Akyüz's algorithm [41] produce higher-quality images and preserve high-frequency details. However, these algorithms still yield false-color artifacts because of demosaicing. On the contrary, the proposed algorithm in Figure 4(h) achieves the highest-quality synthesized images, preserving high-frequency details faithfully and avoiding visual artifacts. This is because the proposed algorithm explicitly considers the HVS, chromatic, and multiscale contrast losses in (7), (13), and (18), respectively, during training.

Figure 5 shows the synthesized HDR images from the calibrated dataset. Gu *et al.*'s algorithm in Figure 5(c) results in blurring artifacts due to interpolation. Cho *et al.*'s algorithm in Figure 5(d) yields artifacts around textures when the expo-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

(a) (b) (c) (d) (e) (f) (g) (h)

FIGURE 5. Synthesized results of the noncalibrated images. (a) Ground-truth, (b) synthetic raw Bayer images, (c) Gu *et al.*'s algorithm [33], (d) Cho *et al.*'s algorithm [37], (e) Choi *et al.*'s algorithm [39], (f) An and Lee's algorithm [40], (g) Çoğalan and Akyüz's algorithm [41] , and (h) the proposed algorithm. (Top) Full-resolution images. (Bottom) Magnified parts.

TABLE 4. Quantitative comparison of the synthesis performance using three objective quality metrics: pu-PSNR, log-PSNR, and HDR-VDP. Boldface numbers denote the best scores, thus indicating the best quality.

	Noncalibrated dataset				Calibrated dataset			
	pu-PSNR	log-PSNR	HDR-VDP (P)	HDR-VDP (Q)	pu-PSNR	log-PSNR	HDR-VDP (P)	HDR-VDP (Q)
Gu *et al.* [33]	32.80	32.13	0.80	69.00	31.77	26.48	0.80	68.93
Cho *et al.* [37]	40.02	30.40	0.62	70.80	39.37	28.00	0.60	69.64
Choi *et al.* [39]	36.17	35.10	0.48	69.23	36.30	28.51	0.57	70.14
An and Lee [40]	39.92	35.49	**0.40**	69.37	37.32	28.20	0.52	66.88
Çoğalan and Akyüz [41]	42.00	39.86	0.48	70.23	40.33	31.08	**0.45**	71.08
Proposed	**44.20**	**40.18**	0.48	**73.66**	**42.55**	**31.34**	0.46	**73.76**

sure difference between rows is large since no information is available in neighboring pixels to recover missing regions. Choi *et al.*'s algorithm in Figure 5(e) achieves better performance than the interpolation-based algorithms, especially in well-exposed regions, *e.g.*, the leaf region in the second-row image. However, when a region has poorly exposed values, *e.g.*, the characters in the fourth and sixth rows, blurring artifacts around edges are brought in. An and Lee's algorithm and Çoğalan and Akyüz's algorithm in Figures 5(f) and (g), respectively, synthesize high-quality HDR images, alleviating the artifacts via the learning-based approaches. However, this is at the cost of splotchy artifacts appearing around the strong edges because their networks are trained with L_2 loss functions. It is also observed that Cho *et al.*'s, An and Lee's, and Çoğalan and Akyüz's algorithms produce false color artifacts caused by demosaicing. In contrast, the proposed algorithm provides the highest quality HDR images without noticeable artifacts. This is because our model is trained with the HVS, chromatic, and multiscale contrast losses, and it effectively alleviates the effects of the outliers via the L_1 loss.

To complement the subjective assessment, we compare the proposed algorithm with conventional algorithms using three objective quality metrics: perceptually uniform extension to PSNR (pu-PSNR) [59], log-PSNR [59], and high dynamic range visible difference predictor (HDR-VDP) [60], [61]. Table 4 quantitatively compares the synthesis performance of the proposed algorithm against those of the conventional algorithms for the noncalibrated and calibrated test images. For each metric on each dataset, the highest scores, which indicate the best results in terms of the lowest differences, are boldfaced. The pu-PSNR and log-PSNR metrics, which are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

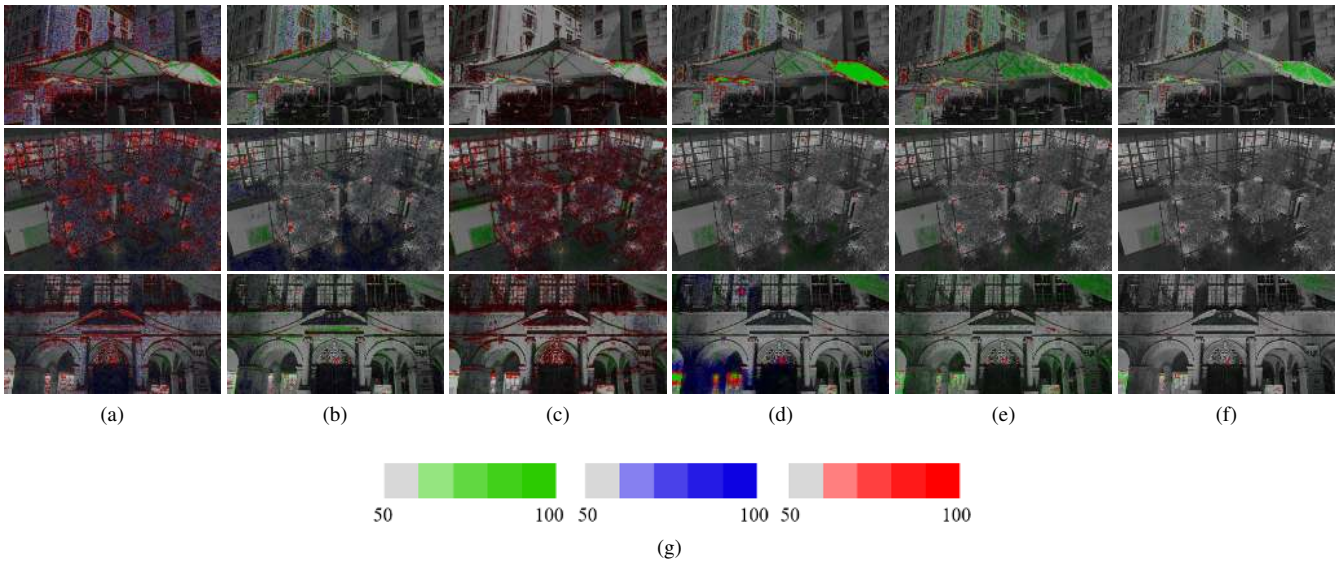


FIGURE 6. DRIM assessment on a test image. (a) Gu *et al.*'s algorithm [33], (b) Cho *et al.*'s algorithm [37], (c) Choi *et al.*'s algorithm [39], (d) An and Lee's algorithm [40], (e) Çoğalan and Akyüz's algorithm [41], and (f) the proposed algorithm. The colormaps in (g) depict the predicted visible differences.

extensions of the PSNR metric, measure the quality of pixel value reconstruction by considering the HVS model in absolute luminance. As the proposed algorithm effectively reconstructs missing information via multiscale CNN with the robust loss function, it provides significantly higher scores than the conventional algorithms in these metrics. HDR-VDP measures perceptual differences between the reference and test images by predicting the visibility and quality differences at which an average human observer would detect. In this work, we use the quality index score Q and the probability score P [61]. Lower P and higher Q values imply that the query image provides higher image quality with less difference compared to the reference image. Because the proposed algorithm effectively synthesizes high-quality HDR images via a multiscale CNN model with the HVS model-based loss function, the proposed algorithm also achieves the best performance in terms of the overall HDR-VDP scores.

Figure 6 shows the distortion map for the test images in Figures 4 and 5 estimated by another quality metric: dynamic range independent image quality metric (DRIM) [62]. DRIM estimates the probability that the differences between two images in structural changes in each local region would be noticed by a viewer. The DRIM detects three types of structural change, namely, loss of visible contrast (green), amplification of invisible contrast (blue), and reversal of visible contrast (red), as shown in Figure 6(g). Gu *et al.*'s algorithm [33] in Figure 6(a) results in severe visible differences with significant losses in visible contrast because their algorithm produces blurring artifacts that change the structures of the image. In Figure 6(b), Cho *et al.*'s algorithm [37] produces less visible differences than Gu *et al.*'s algorithm, but still results in visible differences with the amplification of invisible contrast. Although Choi *et al.*'s algorithm [39] in Figure 6(c) produces a smaller loss of visible contrast and amplification of invisible contrast than the interpolation-based algorithms, it contains a higher reversal of visible contrast (red color). Additionally, most of the differences appear in under-exposed regions. An and Lee's algorithm [40] in Figure 6(d) synthesizes HDR images with a smaller amount of artifacts, but the results still exhibit differences in the over-exposed regions with the loss of visible contrast. Çoğalan and Akyüz's algorithm [41] in Figure 6(e) provides less artifacts than An and Lee's algorithm but still loses visible contrast in the over-exposed regions. On the contrary, the proposed algorithm in Figure 6(f) achieves high-quality results with significantly less visible differences in terms of all the structural changes than the conventional algorithms.

C. CAPTURED IMAGES

Next, we evaluate the HDR synthesis performance on captured SVE images. We controlled the exposure times of a FLIR Grasshopper®3 camera to capture two raw Bayer images of a static scene. Then, we merged the images to synthesize the row-wise SVE images, as shown in Figure 7(a). The exposure times were set such that the longer exposure was two stops more than the shorter exposure. In this test, the synthesized images are only compared qualitatively, because the ground-truth HDR images are unavailable.

The synthesis results in Figure 7 exhibit similar tendencies to the results for the synthetic images in Figures 4 and 5. Gu *et al.*'s algorithm in Figure 7(b) smears details and causes blurry textures. Cho *et al.*'s algorithm in Figure 7(c) yields severe artifacts around textures and edges. Choi *et al.*'s algorithm in Figure 7(d) achieves better results but still produces blurring around the edges. An and Lee's and Çoğalan and Akyüz's algorithms in Figures 7(e) and (f), respectively, provide low performance since the image contains highly over-exposed regions. Moreover, Cho *et al.*'s, An and Lee's, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access

(a) (b) (c) (d) (e) (f) (g)

FIGURE 7. Synthesis results of the captured images. (a) SVE raw Bayer input image, (b) Gu *et al.*'s algorithm [33], (c) Cho *et al.*'s algorithm [37], (d) Choi *et al.*'s algorithm [39], (e) An and Lee's algorithm [40], (f) Çoğalan and Akyüz's algorithm [41], and (g) the proposed algorithm.

TABLE 5. Ablation study of the proposed multiple-branch network against different combinations of network architectures at the final epoch. Boldface numbers denote the highest scores for each metric.

	pu-PSNR	log-PSNR	HDR-VDP (P)	HDR-VDP (Q)
U-Net [63]	42.04	34.89	0.35	70.64
G	45.48	34.83	**0.31**	73.04
G+M	42.43	33.33	0.32	73.20
G+L	43.73	34.72	**0.31**	73.11
G+M+L	**45.80**	**36.60**	**0.31**	**73.73**

Çoğalan and Akyüz's algorithms elicit false color artifacts. The proposed algorithm in Figure 7(g) provides more faithful results with less visible artifacts.

D. ABLATION STUDIES FOR NETWORK ARCHITECTURES

We conduct ablation studies to analyze the contributions of three different branch networks in the proposed algorithm described in Section II-C: global, medium, and local. Specifically, we analyze the effectiveness of each branch network simultaneously by training the proposed networks to synthesize HDR images using the following combinations:

- G: A single global branch network is used.
- G+M: Multiscale features, obtained using global and medium branch networks, are combined.
- G+L: Multiscale features, obtained using global branch and local branch networks, are combined.
- G+M+L: In addition to G+M, feature maps obtained from local branch network are used as well (proposed model). Thus, three-branch networks are used in total.

In addition, we analyze the effectiveness of the proposed networks by comparing them with U-Net [63], which is commonly used as a backbone for image restoration.

Table 5 shows a quantitative comparison of the synthesis performance of the proposed multiple-branch networks and U-Net in terms of pu-PSNR, log-PSNR, and HDR-VDP. Note that all the models are trained with the robust loss function in (5). First, it can be observed that U-Net provides the worst performance in all HDR image quality metrics, which indicates that the proposed networks can synthesize HDR images more effectively than general-purpose restoration networks. Except for U-Net, G yields the worst performance in Q-value. Second, a higher performance is achieved in Q-value with multiscale features, which are obtained through medium branch (M) or local branch (L) networks. Finally, by combining the three branch networks (G+M+L), the proposed network achieves the best score in all HDR image quality metrics.

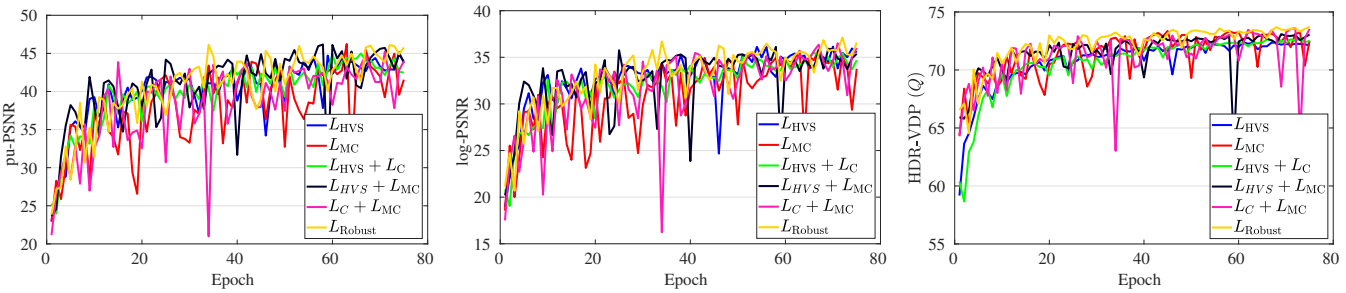
E. ABLATION STUDIES FOR LOSS FUNCTIONS

Finally, we conduct ablation studies to analyze the contribution of each loss function described in Section III to the synthesis performance. To this end, we trained the proposed network using six different combinations of loss functions, namely L_{HVS}, L_{MC}, $L_{\text{HVS}}+L_{\text{C}}$, $L_{\text{HVS}}+L_{\text{MC}}$, $L_{\text{C}}+L_{\text{MC}}$, and L_{Robust}. In addition, we compare the proposed loss functions with the tone-mapped loss L_{TM} developed in [20] as

$$L_{\text{TM}}(H,Y) = \|\mathcal{T}(H) - \mathcal{T}(Y)\|_2^2, \tag{21}$$

where $\mathcal{T}(H) = \frac{\log(1+\mu H)}{\log(1+\mu)}$ denotes the tone-mapping function, and the HVS loss L_{HVS,L_2} using the L_2 norm. After training, we compute three objective quality metrics: pu-PSNR, log-PSNR, and HDR-VDP.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3078457, IEEE Access



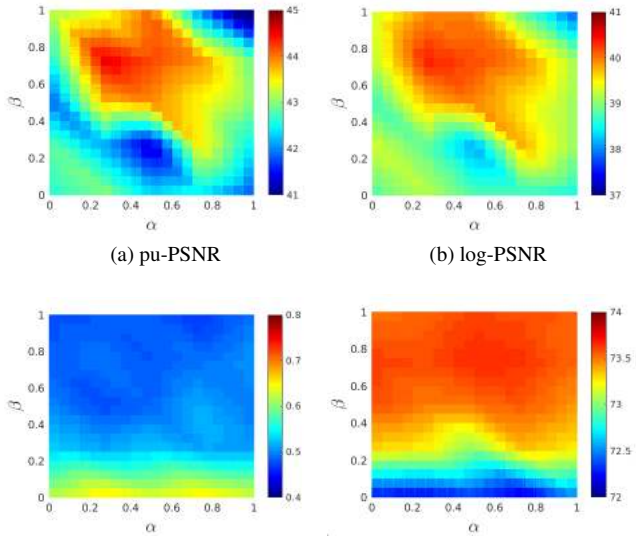
(a) pu-PSNR (b) log-PSNR (c) HDR-VDP

FIGURE 8. Convergence behavior for the different combinations of loss functions, measured via (a) pu-PSNR, (b) log-PSNR, and (c) HDR-VDP.

TABLE 6. Ablation study of the proposed robust loss function against different combinations of losses at the final epoch. Boldface numbers denote the highest scores for each metric.

	pu-PSNR	log-PSNR	HDR-VDP (P)	HDR-VDP (Q)
L_{TM} [20]	43.41	**37.44**	0.39	72.00
L_{HVS,L_2}	43.25	34.71	0.40	72.05
L_{HVS}	43.85	35.55	0.39	72.14
L_{MC}	41.76	33.91	**0.30**	72.82
$L_{\text{HVS}} + L_{\text{C}}$	42.70	34.73	0.41	72.51
$L_{\text{MC}} + L_{\text{C}}$	44.20	36.03	**0.30**	73.60
$L_{\text{HVS}} + L_{\text{MC}}$	45.05	35.34	**0.30**	73.00
L_{Robust}	**45.80**	36.60	0.31	**73.73**

Table 6 shows a quantitative comparison of the synthesis performance of the different loss functions at the final epoch. First, since L_{TM} considers human perception using the logarithm function that approximates the HVS's contrast sensitivity [52], [53], it provides the highest log-PSNR but the lowest Q-value, thereby producing lower perceptual quality. Second, since L_{HVS} measures perceptual differences by considering the HVS model, L_{HVS} yields better performance than L_{TM} in terms of pu-PSNR and Q-value. Third, L_{HVS,L_2} yields inferior performance than L_{HVS} in terms of all metrics, since the L_2 norm is more sensitive to outliers than the L_1 norm [43]. Fourth, as L_{MC} measures the structural and contrast similarity of the images in a multiscale approach, it achieves higher perceptual qualities corresponding to lower P-value and higher Q-value than L_{HVS}. Fifth, as L_{C} measures color information with chroma functions to avoid color artifacts, the combinations $L_{\text{HVS}} + L_{\text{C}}$ and $L_{\text{MC}} + L_{\text{C}}$ increase Q-values of L_{HVS} and L_{MC}, respectively. Sixth, $L_{\text{HVS}} + L_{\text{MC}}$ takes advantage of pixel-wise quality estimation from L_{HVS} and multiscale structural and contrast similarity from L_{MC}. Therefore, it provides the second best pu-PSNR and the lowest P-value. Finally, by combining L_{HVS}, L_{C}, and L_{MC}, the robust loss function L_{Robust} achieves the highest pu-PSNR and Q-value and the second best scores in terms of log-PSNR and P-score, producing high-quality HDR images with less visual artifacts than other loss functions. Therefore, this analysis confirms the effectiveness of the proposed robust loss function.



(a) pu-PSNR (b) log-PSNR

(c) HDR-VDP (P) (d) HDR-VDP (Q)

FIGURE 9. Synthesis performance evaluation for various combinations of α and β in terms of (a) pu-PSNR, (b) log-PSNR, (c) HDR-VDP (P), and (d) HDR-VDP (Q).

We also examine the convergence behavior and stability of different combinations of the proposed loss functions. Figure 8 shows the plots of three image quality metrics, *i.e.*, pu-PSNR, log-PSNR, and HDR-VDP, at all epoches in training. The robust loss function L_{Robust} achieves the best performance with significantly less fluctuation than other combinations on all quality metrics. This result also confirms the stability as well as the effectiveness of the proposed robust loss function in training.

F. EFFECTS OF PARAMETERS α AND β ON SYNTHESIS PERFORMANCE

As discussed in Section III, α and β in (5) control the trade-off between the HVS loss L_{HVS}, chromatic loss L_{C}, and multiscale contrast loss L_{MC}. We evaluate how those parameters affect the synthesis performance. Figure 9 shows the average performance for various combinations of α and β on the

test dataset. More specifically, 25 combinations with $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ and $\beta \in \{0, 0.25, 0.5, 0.75, 1.0\}$ were evaluated, and then the results were interpolated for visualization.

As shown in Figure 9, the proposed algorithm provides varying performance for different values of α and β in terms all the objective quality metrics. More specifically, in Figures 9(a) and (b), the pu-PSNR and log-PSNR scores are considerably affected by both α and β. On the contrary, the effects of α on the HDR-VDP scores is smaller than those of β in Figures 9(c) and (d). This is in line with the fact that the HVS is less sensitive to color than to luminance. An improper combination of α and β may yield undesirable artifacts in the synthesized image. Therefore, to achieve overall high synthesis performance in terms of the objective quality metrics, we fixed α and β to 0.5 and 0.75, respectively, in this work.

V. CONCLUSIONS

In this work, we proposed an end-to-end CNN-based single-shot HDR image synthesis algorithm in this work. We first designed a multiscale CNN that consists of multiple encoder–decoder networks to obtain multiscale feature maps for an SVE image. Then, we developed the robust loss function composed of the HVS, chromatic, and multiscale contrast losses, to effectively measure the differences in HDR images. The experimental results showed that the proposed algorithm outperforms the conventional algorithms in terms of both subjective and objective quality metrics. In addition, the ablation studies of network architecture and loss function showed that the proposed network and robust loss function achieve reliable and high-quality synthesis results.

REFERENCES

[1] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, 2nd ed. Morgan Kaufmann Publishers, 2010.

[2] P. Sen and C. Aguerrebere, "Practical high dynamic range imaging of everyday scenes: Photographing the world as we see it with our own eyes," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 36–44, Sep. 2016.

[3] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," *ACM Trans. Graphics*, vol. 30, no. 4, pp. 41:1–10, Jul. 2011.

[4] H. Zhao, B. Shi, C. Fernandez-Cull, S. Yeung, and R. Raskar, "Unbounded high dynamic range photography using a modulo camera," in *Proc. IEEE Int. Conf. Comput. Photography*, Jul. 2015, pp. 1–10.

[5] J. N. P. Martel, L. K. Müller, S. J. Carey, P. Dudek, and G. Wetzstein, "Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 42, no. 7, pp. 1642–1653, Jul. 2020.

[6] S. Mann and R. W. Picard, "On being undigital with digital cameras: Extending dynamic range by combining differently exposed pictures," in *Proc. IS&T's Annual Conf.*, 1995, pp. 422–428.

[7] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. ACM SIGGRAPH*, Aug. 1997, pp. 369–378.

[8] T. Mitsunaga and S. K. Nayar, "Radiometric self calibration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 1999, pp. 374–380.

[9] M. A. Robertson, S. Borman, and R. L. Stevenson, "Estimation-theoretical approach to dynamic range enhancement using multiple exposures," *J. Electron. Imaging*, vol. 12, no. 2, pp. 219–228, Apr. 2003.

[10] L. Bogoni, "Extending dynamic range of monochrome and color images through fusion," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Sep. 2000, pp. 7–12.

[11] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition," in *Proc. Int. Conf. Central Europ. Comput. Graph. Vis. Comput. Vis*, Jan./Feb. 2007, pp. 49–56.

[12] M. Gupta, D. Iso, and S. K. Nayar, "Fibonacci exposure bracketing for high dynamic range imaging," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1473–1480.

[13] C. Lee and E. Y. Lam, "Computationally efficient truncated nuclear norm minimization for high dynamic range imaging," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4145–4157, Sep. 2016.

[14] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand HDR imaging of moving scenes with simultaneous resolution enhancement," *Comput. Graph. Forum*, vol. 30, no. 2, pp. 405–414, Apr. 2011.

[15] C. Lee, Y. Li, and V. Monga, "Ghost-free high dynamic range imaging via rank minimization," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1045–1049, Sep. 2014.

[16] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graphics*, vol. 31, no. 6, pp. 203:1–203:11, Nov. 2012.

[17] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1163–1170.

[18] C. Aguerrebere, J. Delon, Y. Gousseau, and P. Musé, "Simultaneous HDR image reconstruction and denoising for dynamic scenes," in *Proc. IEEE Int. Conf. Comput. Photography*, Apr. 2013, pp. 1–11.

[19] T.-H. Oh, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Robust high dynamic range imaging by rank minimization," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 37, no. 6, pp. 1219–1232, Jun. 2015.

[20] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graphics*, vol. 36, no. 4, pp. 144:1–144:12, Jul. 2017.

[21] S. Wu, J. Xu, Y. W. Tai, and C. K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. European Conf. Comput. Vis.*, Sep. 2018, pp. 120–135.

[22] Q. Yan, D. Gong, Q. Shi, A. V. D. Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1751–1760.

[23] K. R. Prabhakar, R. Arora, A. Swaminathan, K. P. Singh, and R. V. Babu, "A fast, scalable, and reliable deghosting method for extreme exposure fusion," in *Proc. IEEE Int. Conf. Comput. Photography*, May 2019, pp. 1–8.

[24] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang, "Deep HDR imaging via a non-local network," *IEEE Trans. Image Process.*, vol. 29, pp. 4308–4322, 2020.

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

[26] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graphics*, vol. 36, no. 6, pp. 178:1–178:15, Nov. 2017.

[27] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. Graphics*, vol. 36, no. 6, pp. 177:1–177:10, Nov. 2017.

[28] D. Marnerides, T. B.-Rogers, J. Hatchett, and K. Debattista, "ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 37–49, May 2018.

[29] S. Lee, G. H. An, and S.-J. Kang, "Deep recursive HDRI: Inverse tone mapping using generative adversarial networks," in *Proc. European Conf. Comput. Vis.*, Sep. 2018, pp. 613–628.

[30] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image HDR reconstruction by learning to reverse the camera pipeline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1651–1660.

[31] M. S. Santos, T. I. Ren, and N. K. Kalantari, "Single image HDR reconstruction using a CNN with masked features and perceptual loss," *ACM Trans. Graphics*, vol. 39, no. 4, pp. 80:1–80:10, Jul. 2020.

[32] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2000, pp. 472–479.

[33] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling," in *Proc. IEEE Int. Conf. Comput. Photography*, Mar. 2010, pp. 1–8.

[34] K. Hirakawa and P. M. Simon, "Single-shot high dynamic range imaging with conventional camera hardware," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1339–1346.
[35] C. Aguerrebere, A. Almansa, Y. Gousseau, J. Delon, and P. Musé, "Single shot high dynamic range imaging using piecewise linear estimators," in *Proc. IEEE Int. Conf. Comput. Photography*, May 2014, pp. 1–10.
[36] J. Li, C. Bai, Z. Lin, and J. Yu, "Penrose high-dynamic-range imaging," *J. Electron. Imaging*, vol. 25, no. 3, pp. 033 024:1–13, Jun. 2016.
[37] H. Cho, S. J. Kim, and S. Lee, "Single-shot high dynamic range imaging using coded electronic shutter," *Comput. Graph. Forum*, vol. 33, no. 7, pp. 329–338, Oct. 2014.
[38] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 153–163, May 2016.
[39] I. Choi, S. Baek, and M. H. Kim, "Reconstructing interlaced high-dynamic-range video using joint learning," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5353–5366, Nov. 2017.
[40] V. G. An and C. Lee, "Single-shot high dynamic range imaging via deep convolutional neural network," in *Proc. APSIPA ASC*, Dec. 2017, pp. 1768–1772.
[41] U. Çoğalan and A. O. Akyüz, "Deep joint deinterlacing and denoising for single shot dual-ISO HDR reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 7511–7524, Jun. 2020.
[42] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.
[43] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Compt. Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[44] O. M. Blackwell and H. R. Blackwell, "IERI: Visual performance data for 156 normal observers of various ages," *J. Illuminating Eng. Soc.*, vol. 1, no. 1, pp. 3–13, Oct. 1971.
[45] Y. Sun, Y. Yu, and W. Wang, "Moiré photo restoration using multiresolution convolutional neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4160–4172, Aug. 2018.
[46] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.
[47] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep tone mapping operator for high dynamic range images," *IEEE Trans. Image Process.*, vol. 29, pp. 1285–1298, 2020.
[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Machine Learning*, Jul. 2015, pp. 448–456.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.
[50] ——, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[51] S. Winkler and S. Süsstrunk, "Visibility of noise in natural images," in *Proc. SPIE, Human Vision and Electronic Imaging IX*, vol. 5292, 2004, pp. 121–129.
[52] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, "Backward compatible high dynamic range MPEG video compression," *ACM Trans. Graphics*, vol. 25, no. 3, pp. 713–723, Jul. 2006.
[53] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Lossy compression of high dynamic range images and video," in *Proc. SPIE, Human Vision and Electronic Imaging*, Feb. 2006, pp. 6057–6057–10.
[54] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graphics*, vol. 21, no. 3, pp. 267–276, Jul. 2002.
[55] R. Timofte, R. Rothe, and L. V. Gool, "Seven ways to improve example-based single image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1865–1873.
[56] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2790–2798.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Represent.*, May 2015.
[58] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.
[59] O. T. Aydın, R. Mantiuk, and H. P. Seidel, "Extending quality metrics to full dynamic range images," in *Proc. SPIE, Human Vision and Electronic Imaging XIII*, Jan. 2008, pp. 6806–10.
[60] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graphics*, vol. 30, no. 4, Jul. 2011.
[61] M. Narwaria, R. K. Mantiuk, M. P. D. Silva, and P. L. Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high dynamic range and standard images," *J. Electron. Imaging*, vol. 24, no. 1, Jan. 2015.
[62] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graphics*, vol. 27, no. 3, pp. 69:1–10, Aug. 2008.
[63] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Imag. Comput. Computer-Assiated Intervention*, Nov. 2015, pp. 234–241.

AN GIA VIEN (S'19) received the B.S. degree in mathematics and computer science from University of Science, Vietnam National University, Ho Chi Minh City, Vietnam, in 2015, and the M.S. degree in computer engineering from Pukyong National University, Busan, South Korea, in 2019. He is currently working toward the Ph.D. degree in the Department of Multimedia Engineering at Dongguk University, Seoul, South Korea.

His current research interests include image restoration, image enhancement, and high dynamic range imaging.

CHUL LEE (S'06–M'13) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2003, 2008, and 2013, respectively.

He was with Biospace Inc., Seoul, South Korea, from 2002 to 2006, where he was involved in the development of medical equipment. From 2013 to 2014, he was a Postdoctoral Scholar with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA. From 2014 to 2015, he was a Research Scientist with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. From 2015 to 2019, he was an Assistant Professor with the Department of Computer Engineering, Pukyong National University, Busan, South Korea. In March 2019, he joined the Department of Multimedia Engineering, Dongguk University, Seoul, South Korea, where he is currently an Assistant Professor. His current research interests include image processing and computational imaging with an emphasis on restoration and high dynamic range imaging.

He received the Best Paper Award from the *Journal of Visual Communication and Image Representation* in 2014. He is currently an Editorial Board Member of the *Journal of Visual Communication and Image Representation*.