IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 23, 2019, accepted May 19, 2019, date of publication May 27, 2019, date of current version June 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919296

iTM-Net: Deep Inverse Tone Mapping Using Novel Loss Function Considering Tone Mapping Operator

YUMA KINOSHITA, (Student Member, IEEE), AND HITOSHI KIYA, (Fellow, IEEE)
Department of Computer Science, Tokyo Metropolitan University, Tokyo 191-0065, Japan

Corresponding author: Hitoshi Kiya (kiya@ tmu.ac.jp)

This work was supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for JSPS Fellows under JSPS KAKENHI Grant Number JP18J20326.

ABSTRACT In this paper, we propose a novel inverse tone mapping network, called "iTM-Net." For training iTM-Net, we also propose a novel loss function that considers the non-linear relation between low dynamic range (LDR) and high dynamic range (HDR) images. For inverse tone mapping with convolutional neural networks (CNNs), we first point out that training CNNs with a standard loss function causes a problem due to the non-linear relation between the LDR and HDR images. To overcome the problem, the novel loss function non-linearly tone-maps target HDR images into LDR ones on the basis of a tone mapping operator, and the distance between the tone-mapped images and predicted ones are then calculated. The proposed loss function enables us not only to normalize the HDR images but also to reduce the non-linear relation between LDR and HDR ones. The experimental results show that the HDR images predicted by the proposed iTM-Net have higher-quality than the HDR ones predicted by conventional inverse tone mapping methods, including the state of the art, in terms of both HDR-VDP-2.2 and PU encoding + MS-SSIM. In addition, compared with loss functions that do not consider the non-linear relation, the proposed loss function is shown to improve the performance of CNNs.

INDEX TERMS Convolutional neural networks, deep learning, high dynamic range imaging, inverse tone mapping, loss function.

I. INTRODUCTION

The low dynamic range (LDR) of modern digital cameras is a major factor preventing cameras from capturing images as well as human vision. This is due to the limited dynamic range that imaging sensors have. For this reason, interest in high dynamic range (HDR) imaging has been increasing.

The goal of HDR imaging is to obtain HDR images whose pixel values describe absolute or relative luminance of scenes, which is proportional to scene radiance. Since LDR images are distorted by sensor saturation and a non-linear camera response function (CRF), the objective of HDR imaging can be separated into two goals: saturation recovery and linearization. The most common approach to tackle the problems is stack-based methods that use a stack of differently exposed images, called "multi-exposure images," for both saturation recovery and linearization [1]–[8]. Although these stack-based methods work very well, they still have two limitations due to the use of multi-exposure images: ghost-like artifacts appear due to the motion of objects in a scene and a camera during taking multi-exposure images, and they are inapplicable to existing single images. For this reason, HDR imaging methods without multi-exposure images are expected to be developed.

With the aim of generating an HDR image from a single LDR image, various research works on inverse tone mapping have so far been reported [9]–[18]. Traditional ways of inverse tone mapping are based on expanding the dynamic range of input LDR images by using a fixed function or a specific parameterized function [9]–[15]. However, inverse tone mapping without prior knowledge is generally an ill-posed problem for two reasons: pixel values might be lost by sensor saturation, and the CRF used for photographing is unknown. Hence, HDR images produced by these methods

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

have limited quality. To obtain high-quality HDR images, inverse tone mapping methods based on deep learning have recently attracted attention.

Several convolutional neural network (CNN) based inverse tone mapping methods have so far been proposed [16]–[18]. These CNN-based methods significantly improve the performance of inverse tone mapping. In [16] and [17], CNNs are utilized for saturation recovery, but these methods do not employ deep learning for linearization. Although Marnerides *et al.* [18] tackled the linearization problem by training a CNN with simply normalized HDR images under the use of the min-max normalization, the performance was still limited because most pixel values of the normalized images were distributed in a narrow range. This is due to the non-linear relation between LDR and HDR images.

Thus, in this paper, we propose a novel inverse tone mapping network, called "iTM-Net." Similarly to [18], we aim to obtain relative luminance by using iTM-Net for linearization. To realize this, we also propose a novel loss function that considers the non-linear relation between LDR and HDR images. In the novel loss function, target HDR images are tone-mapped into LDR ones by using an invertible tone mapping operator, and the distance between the tone-mapped images and predicted ones is then calculated. The proposed loss function enables us not only to normalize HDR images but also to widely distribute the pixel values of HDR images like LDR ones. Our iTM-Net is implemented without using generative adversarial networks in order to make iTM-Net applicable to images having various resolutions.

In an experiment, the proposed method was compared with state-of-the-art inverse tone mapping methods. Experimental results illustrate that the proposed method outperforms the conventional methods in terms of two objective quality metrics: HDR-VDP-2.2 and PU encoding + MS-SSIM. In addition, the proposed loss function is shown to improve the performance of CNNs compared with standard loss functions that do not consider the non-linear relation.

II. RELATED WORK

Here, we summarize typical stack-based HDR imaging methods and inverse tone mapping methods. The term "HDR imaging" has been used with two meanings depending on contexts:

- recovering scene radiance (or intensity that is proportional to scene radiance) [1]–[8],
- capturing wide-dynamic-range information of real scene [19]–[30].

We use the term "HDR imaging" as the former meaning throughout this paper.

A. HDR IMAGING

As mentioned, the goal of HDR imaging including inverse tone mapping is to restore the absolute or relative luminance of a scene. This objective can be separated into two goals: saturation recovery and linearization. The most common approach of HDR imaging is stack-based one [1]–[8] that uses a stack of differently exposed images, called "multi-exposure images," for both saturation recovery and linearization. In the stack-based methods, a non-linear CRF is estimated by using multi-exposure images, and linearization is then done by applying the inverse CRF to the input multi-exposure images. After that, these images are fused into a single HDR image, in order to recover saturation.

The stack-based HDR imaging methods work very well when a scene is static and the camera is tripod-mounted. However, when scenes are dynamic or the camera moves while multi-exposure images are being captured, the multi-exposure images will not line up properly with one another. This misalignment results in ghost-like artifacts in the final HDR image. To deal with motion, Sen *et al.* [6] proposed a method that aligns multi-exposure images with patch-based optimization. Oh *et al.* [8] also proposed a robust HDR imaging method on the basis of rank minimization. By these research works, the problem of the ghost-like artifacts is being solved. However, those stack-based methods cannot be applied to existing single LDR images.

B. TRADITIONAL INVERSE TONE MAPPING

For generating an HDR image from a single LDR image, many inverse tone mapping methods have already been studied. Traditional ways of inverse tone mapping are based on expanding the dynamic range of input LDR images by using a fixed function or a specific parameterized function [9]–[15]. Banterle *et al.* [9] employed the inverse function of Reinhard's global operator [31] for expanding the dynamic range. Similarly, Youngquing *et al.* [12] used an S-shaped curve for the purpose. However, a fixed function or a specific parameterized function cannot correctly linearize an input LDR image because each camera has a different CRF and changes in temperature additionally alter the CRF. Moreover, saturation recovery without prior knowledge is also impossible since all saturated pixel values might be lost. Hence, HDR images produced by traditional inverse tone mapping have limited quality in terms of both linearization and saturation recovery.

C. DEEP-LEARNING-BASED INVERSE TONE MAPPING

CNN-based inverse tone mapping methods [16]–[18] have recently attracted attention because of their effectiveness. Eilertsen *et al.* [17] aim to reconstruct saturated areas in input LDR images via a CNN. Predicted pixel values are combined with an input LDR image, which is linearized by using a fixed function that does not consider CRFs, to produce an HDR image. Endo *et al.* [16] proposed a CNN based method that produces a stack of differently exposed images from a single LDR image. The generated images are linearized and fused by using an existing stack-based method such as Debevec's method [3]. These two methods enable us to recover saturated regions in images, but the linearization problem still remains.

In a work by Marnerides *et al.* [18], they tackled the linearization problem and sought to directly produce HDR images by using a CNN. To calculate prediction loss in train-

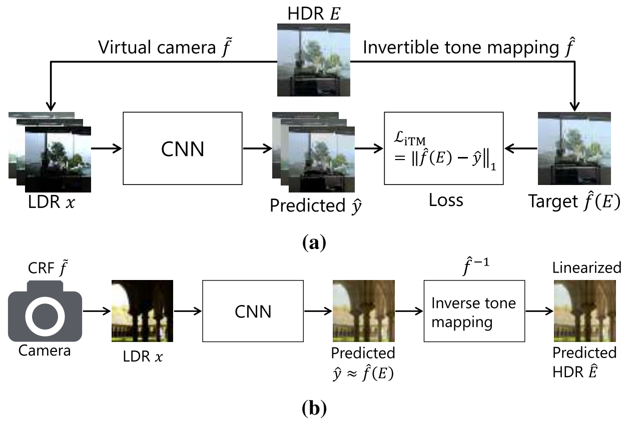

FIGURE 1. **Proposed inverse tone mapping. (a) Training. (b) Predicting.**

ing a CNN, all HDR images are simply normalized into the range of [0, 1] by using min-max normalization. However, using normalized HDR images to calculate prediction loss causes a problem in that most pixel values of the normalized images are distributed in a narrow range. This is due to the non-linear relation between LDR and HDR images, so the image statistics of LDR and HDR images differ considerably as pointed out in [17]. Therefore, we aim to improve the performance of CNN-based inverse tone mapping, by using a novel loss function that considers the non-linear relation for learning HDR images.

III. PROPOSED INVERSE TONE MAPPING

A. NOTATION

The following notations are utilized throughout this paper.

- i and j are used to denote pixel indexes.
- E denotes an HDR image or scene irradiance which is proportional to scene radiance.
- X denotes the integrated irradiance over the time the shutter is open, commonly referred to as "exposure".
- Luminance of an image is denoted by L, where L is the same as the Y-component in the CIE XYZ color space. For example, Luminance of an HDR image E is written as L_E.
- x and y denote input and output images of a CNN, respectively.

B. OVERVIEW

Figure 1 shows an overview of our training procedure and prediction procedure. In the training, all input LDR images x are generated from target HDR images E by using various virtual cameras $\tilde{f}$ [17]. To calculate loss between a predicted image, $\hat{y}$, and a target HDR one, E, a tone mapping function, $\hat{f}$, which is generally a non-linear one, is applied to E.

After the training, various LDR images are applied to the proposed CNN as input images, where the CNN then predicts tone-mapped versions of HDR images. The linearization is done by mapping the predicted images $\hat{y}$ by using an inverse tone mapping function, $\hat{f}^{-1}$. Detailed training conditions are described in Section III-F.

C. LOSS FUNCTION

For training a CNN, an error between target images and predicted images is calculated by using a loss function, and parameters in the CNN are optimized so that the error will be minimized.

In [18], a loss function for training a CNN is defined by using the L_1-distance $\mathcal{L}_1$ and the cosine similarity $\mathcal{L}_{\cos}$. $\mathcal{L}_1$ and $\mathcal{L}_{\cos}$ are calculated as

$$\mathcal{L}_1(\hat{y}, E) = \frac{1}{P}\sum_{i,j}\|E_{i,j} - \hat{y}_{i,j}\|_1, \tag{1}$$

$$\mathcal{L}_{\cos}(\hat{y}, E) = 1 - \frac{1}{P}\sum_{i,j}\frac{E_{i,j} \cdot \hat{y}_{i,j}}{\|E_{i,j}\|_2\|\hat{y}_{i,j}\|_2}, \tag{2}$$

where $E_{i,j}$ and $\hat{y}_{i,j}$ denote an RGB pixel vector at pixel (i, j) in HDR image E and predicted image $\hat{y}$, respectively, and P is the total number of pixels. By using eqs. (1) and (2), the loss function utilized for ExpandNet [18] is given by

$$\mathcal{L}_{\mathrm{Expand}}(\hat{y}, E) = \mathcal{L}_1(\hat{y}, m(E)) + \lambda\mathcal{L}_{\cos}(\hat{y}, m(E)), \tag{3}$$

where λ is a constant factor that adjusts the contribution of the cosine similarity and $m(E)$ denotes min-max normalization that simply normalizes E into the range of [0, 1] by

$$m(E) = \frac{E - \min E}{\max E - \min E}. \tag{4}$$

However, min-max normalization is unsuitable for learning HDR images because pixel values of HDR images are non-uniformly distributed in an extremely wide range [17] unlike LDR ones.

For this reason, we utilize an invertible tone mapping operator, $\hat{f}(\cdot)$, which is designed to transform HDR images into LDR ones, instead of min-max normalization $m(\cdot)$. For example, the L_1-distance with $\hat{f}(\cdot)$ is calculated by

$$\mathcal{L}_{\mathrm{iTM}}(\hat{y}, E) = \mathcal{L}_1(\hat{y}, \hat{f}(E)). \tag{5}$$

In this paper, Reinhard's global operator [31] is utilized as $\hat{f}(\cdot)$, where the operator transforms HDR images into high-quality LDR ones, and it has an inverse function. By using the luminance matrix L_E of E, the operator is given by the equations

$$\hat{f}(E) = (\hat{g}(L_E) \oslash L_E) \odot E, \tag{6}$$

$$\hat{g}(L_E) = L_X \oslash (1 + L_X), \tag{7}$$

$$L_X = \frac{a}{G(L_E)}L_E, \tag{8}$$

where $\odot$ and $\oslash$ mean pixel-wise multiplication and division, respectively. The parameter $a \in [0, 1]$ determines the brightness of an output image $\hat{f}(E)$, and $G(L_E)$ is the geometric mean of L_E given by

$$G(L_E) = \exp\left(\frac{1}{P}\sum_{i,j}\log\left(\max\left(L_{E\,i,j}, \epsilon\right)\right)\right), \tag{9}$$

where ϵ is a small value for avoiding singularities at $L_{E\,i,j} = 0$. Eq. (8) enables us to calibrate HDR images by

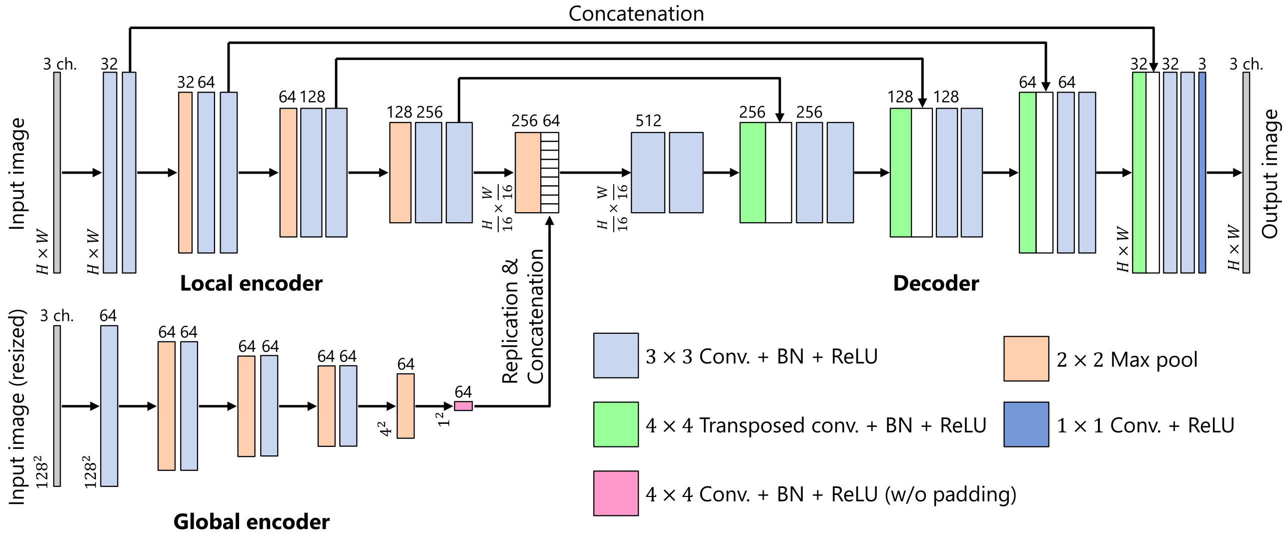

FIGURE 2. Network architecture. Architecture consists of local encoder, global encoder, and decoder. Each box denotes multi-channel feature map produced by each layer. Number of channels is denoted above each box. Feature map resolutions are denoted to left of boxes.

adjusting the geometric mean of each HDR image to a, and eq. (7) allows us to distribute pixel values of HDR images like those of LDR ones. Since $\hat{f}$ is invertible, HDR images can be predicted by using the inverse tone mapping operator $\hat{f}^{-1}$, as shown in the next Section.

D. PREDICTION

The proposed CNN generates tone-mapped versions of HDR images E because the CNN is trained by using the loss function shown in eq. (5). Hence, HDR images are predicted by applying the inverse tone mapping function $\hat{f}^{-1}$ to image $\hat{y}$ as

$$\hat{E} = \hat{f}^{-1}(\hat{y}) = (\hat{g}^{-1}(L_{\hat{y}}) \oslash L_{\hat{y}}) \odot \hat{y}, \tag{10}$$

$$\hat{g}^{-1}(L_{\hat{y}}) = L_{\hat{y}} \oslash (1 - L_{\hat{y}}), \tag{11}$$

where $L_{\hat{y}}$ is the luminance of $\hat{y}$. Note that eq. (8) can be ignored in the inverse tone mapping since our goal is to obtain relative luminance.

E. ITM-NET ARCHITECTURE

Figure 2 shows the overall network architecture of iTM-Net. The architecture consists of three networks: a local encoder, a global encoder, and a decoder. The input for the local encoder is a $P = H \times W$ pixels, 24-bit color LDR image. For the global encoder, the input image is resized to a fixed size (128 × 128). iTM-Net has five types of layers as shown in Fig. 2:

3 × 3 **Conv. + BN + ReLU** which calculates a 3 × 3 convolution with a stride of 1 and a padding of 1. After convolution, batch normalization [32] and the rectified linear unit activation function [33] (ReLU) are applied. In the local encoder and the decoder, two adjacent 3 × 3 Conv. + BN + ReLU layers will have the same number K of filters. From the first two layers to the last ones, the numbers of filters are $K = 32, 64, 128, 256, 512, 256, 128, 64$, and 32, respectively. In the global encoder, all layers have 64 filters.

2 × 2 **Max pool** which downsamples feature maps by max pooling with a kernel size of 2 × 2 and a stride of 2.

4 × 4 **Transposed Conv. + BN + ReLU** which calculates a 4 × 4 convolution with a stride of 1/2 and a padding of 1. After convolution, BN and ReLU are applied. From the first layer to the last one, the numbers of filters are $K = 256, 128, 64$, and 32, respectively.

1 × 1 **Conv. + ReLU** which calculates a 1 × 1 convolution with a stride of 1 and a padding of 1. After convolution, ReLU is applied. The number of filters in the layer is 3.

4 × 4 **Conv. + BN + ReLU (w/o padding)** which calculates a 4 × 4 convolution without padding. The number of filters in the layer is 64.

The local encoder and the decoder in the proposed method are almost the same as those used in U-Net [34]. Concatenated skip connections between the local encoder and the decoder are also utilized like in U-Net.

The main difference between iTM-Net and U-Net is that iTM-Net has an additional encoder, i.e., the global encoder, for extracting global image information. In the most recent work [18], Marnerides et al. claimed that U-Net causes unwanted blocking artifacts in predicted HDR images. Our preliminary experimental results showed that the blocking artifacts are attributed to its network architecture that cannot handle global image information. For this reason, we utilize the global encoder and combine features extracted by both encoders to prevent the distortions.

In addition to the novel network architecture, the use of the novel loss function $\mathcal{L}_{\mathrm{iTM}}$ enables us to improve the performance of inverse tone mapping.

F. TRAINING

Numerous LDR images taken under various conditions, x, and corresponding HDR images, E, are needed to train iTM-Net. To prepare a sufficient amount of training data, we utilize various virtual cameras to generate x from HDR images E [17]. For training, 336 HDR images were collected from databases online available [35]–[40].

The training procedure of our CNN is shown as follows.

(i) Select 16 HDR images from the 336 HDR images at random.

(ii) Generate 16 pairs of an input image and its target one $(x, \tilde{E})$ from each HDR image. Each pair is generated in accordance with the following steps.

(a) Crop HDR image E to an image patch $\tilde{E}$ at $N \times N$ pixels. The size N is given as a product of a uniform random number in the range $[0.2, 0.6]$ and the length of the short side of E. In addition, the position of the patch in E is also determined at random.

(b) Resize $\tilde{E}$ to 256×256 pixels.

(c) Flip $\tilde{E}$ horizontally or vertically with a probability of 0.5.

(d) Calculate exposure X from $\tilde{E}$ with $X = \Delta t(v) \cdot \tilde{E}$, where pixel values larger than 1 are clipped. Shutter speed Δt is calculated as $\Delta t(v) = 0.18 \cdot 2^v / G(L_{\tilde{E}})$ as in [31] by using a uniform random number, v, in the range $[-2, 2]$. $G(L_{\tilde{E}})$ is the geometric mean of the luminance of $\tilde{E}$.

(e) Generate an input LDR image x from X by using virtual camera $\tilde{f}$, as

$$x = \tilde{f}(X) = (\tilde{g}(L_X) \oslash L_X) \odot X, \tag{12}$$

$$\tilde{g}(L_X) = (1 + \eta)(L_X^\gamma \oslash (L_X^\gamma + \eta)), \tag{13}$$

where η and γ are random numbers that follow normal distributions with a mean of 0.6 and a variance of 0.1 and with a mean of 0.9 and a variance of 0.1, respectively. L_X is the luminance of X, and exponentiation L_X^γ is calculated as a pixel-wise operation.

iii Predict 16 LDR images $\hat{y}$ from 16 input LDR images x by using iTM-Net.

iv Evaluate the loss between predicted images $\hat{y}$ and target images $\tilde{E}$ by using eq. (5) with Reinhard's global operator $\hat{f}$. Here, $a = 0.18$ is used in (8).

v Update filter weights ω and biases b in the CNN by backpropagation.

In our experiments, iTM-Net was trained with 1000 epochs, where the above procedure was repeated 42 times in each epoch. In addition, each HDR image had only one chance to be selected in Step i in each epoch. He's method [41] was used for initializing iTM-Net. In addition, the Adam optimizer [42] was utilized for optimization, where parameters in Adam were set as $\alpha = 0.002$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We implemented iTM-Net with Tensorflow and Keras, and the training was run on a single NVIDIA GeForce 1080Ti GPU.

IV. SIMULATION

We evaluated the effectiveness of the proposed method by using two objective quality metrics in addition to visual evaluation.

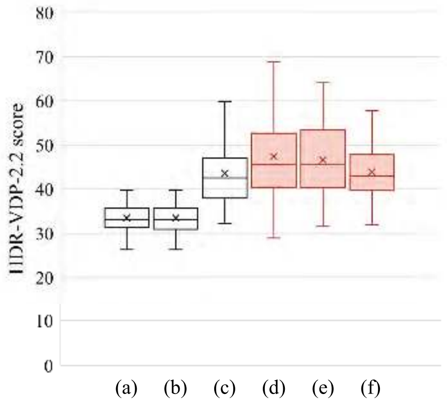

FIGURE 3. HDR-VDP-2.2 scores. (a) ITMO [15], (b) PMET [13], (c) ExpandNet [18], (d) iTM-Net with $\mathcal{L}_{\text{iTM}}$ (Proposed), (e) iTM-Net with $\mathcal{L}_1$, and (f) iTM-Net with $\mathcal{L}_{\text{Expand}}$. Boxes span from first to third quartile, referred to as Q_1 and Q_3, and whiskers show maximum and minimum values in range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. Band and cross inside boxes indicate median and average value, respectively.

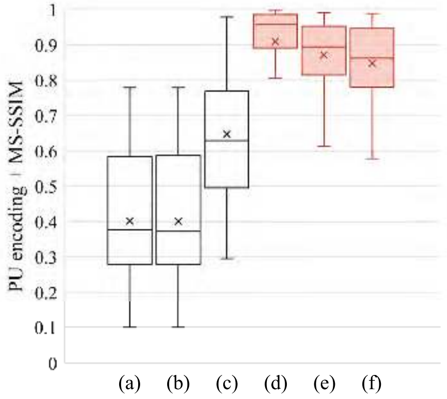

FIGURE 4. PU-encoding + MS-SSIM scores (a) ITMO [15], (b) PMET [13], (c) ExpandNet [18], (d) iTM-Net with $\mathcal{L}_{\text{iTM}}$ (Proposed), (e) iTM-Net with $\mathcal{L}_1$, and (f) iTM-Net with $\mathcal{L}_{\text{Expand}}$. Boxes span from first to third quartile, referred to as Q_1 and Q_3, and whiskers show maximum and minimum values in range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. Band and cross inside boxes indicate median and average value, respectively.

A. SIMULATION CONDITIONS

The quality of HDR images $\hat{E}$ generated by using iTM-Net was evaluated by using two objective quality metrics: HDR-VDP-2.2 [43] and PU encoding [44] with MS-SSIM [45], which utilize an original HDR image, E, as a reference. In [46], it was shown that these metrics are suitable for evaluating the quality of HDR images.

The two metrics are designed for evaluating the difference between two HDR images that have the absolute luminance of a scene. Such HDR images, namely HDR images having absolute luminance, are only in the dataset [37]. Hence, 44 HDR images randomly selected from the dataset [37] were used for the experiment. Note that they were not used for training. Input LDR images x were generated in accordance with Steps ii(d) and ii(e) in Section III-F. In addition, predicted HDR images

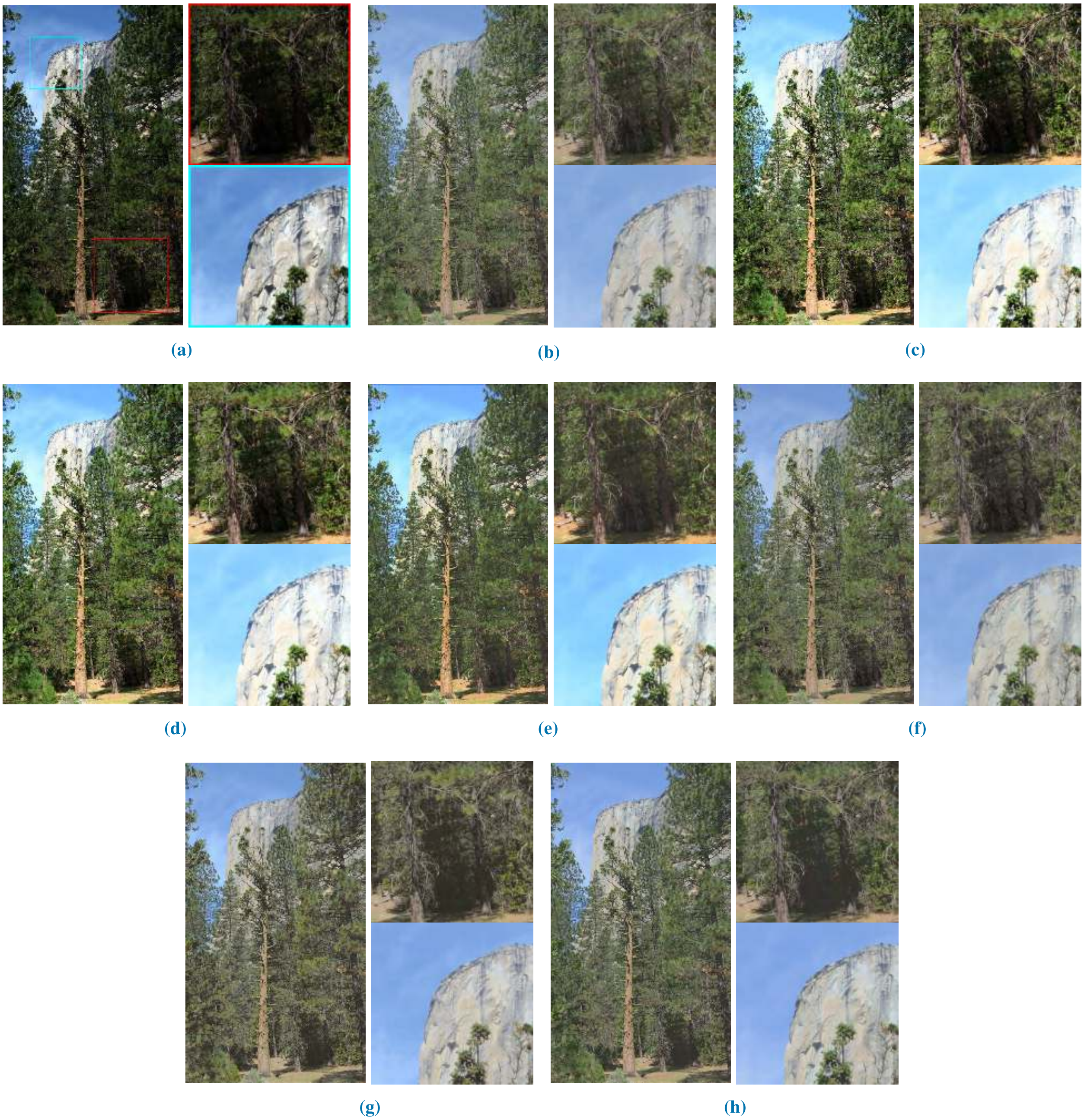

FIGURE 5. Experimental results [for image “ElCapitan”]. Zoom-ins of boxed regions are shown in right of each HDR image. HDR images (b)–(h) were tone-mapped for visualization, where scaling to match range of predicted HDR images with that of original HDR one was not performed. Proposed iTM-Net (f) provided most similar image to original one (b), in six methods. (a) Input x. (b) Ground truth $\tilde{E}$ (c) Direct ITMO [15]. HDR-VDP: 32.27, MS-SSIM: 0.0990. (d) PMET [13]. HDR-VDP: 32.27, MS-SSIM: 0.0990. (e) ExpandNet [18]. HDR-VDP: 40.12, MS-SSIM: 0.7534. (f) iTM-Net with $\mathcal{L}_{\text{iTM}}$ (Proposed). HDR-VDP: 71.77, MS-SSIM: 0.9966. (g) iTM-Net with $\mathcal{L}_1$. HDR-VDP: 49.99, MS-SSIM: 0.9664. (h) iTM-Net with $\mathcal{L}_{\text{Expand}}$. HDR-VDP: 43.39, MS-SSIM: 0.9660.

$\hat{E}$ were scaled to match the range of $\hat{E}$ with that of the original HDR image E because inverse tone mapping methods can predict only HDR images having relative luminance.

The proposed method was compared with three existing methods including state-of-the-art ones: direct inverse tone mapping operator (Direct ITMO) [15], pseudo-multi-exposure-based tone fusion (PMET) [13], ExpandNet [18]. The third method is CNN-based one, but the other methods are not based on deep learning. For ExpandNet, the predictions from this method were generated by using the trained network which was made available online by the authors. Furthermore, to clarify the effectiveness of the proposed loss function, iTM-Net trained by three different loss functions, i.e., the proposed loss $\mathcal{L}_{\text{iTM}}$ (iTM-Net with $\mathcal{L}_{\text{iTM}}$), the standard L_1-loss $\mathcal{L}_1(\hat{y}, m(E))$ without tone mapping (iTM-Net with $\mathcal{L}_1$), and ExpandNet's loss $\mathcal{L}_{\text{Expand}}(\hat{y}, E)$ without tone mapping (iTM-Net with $\mathcal{L}_{\text{Expand}}$).

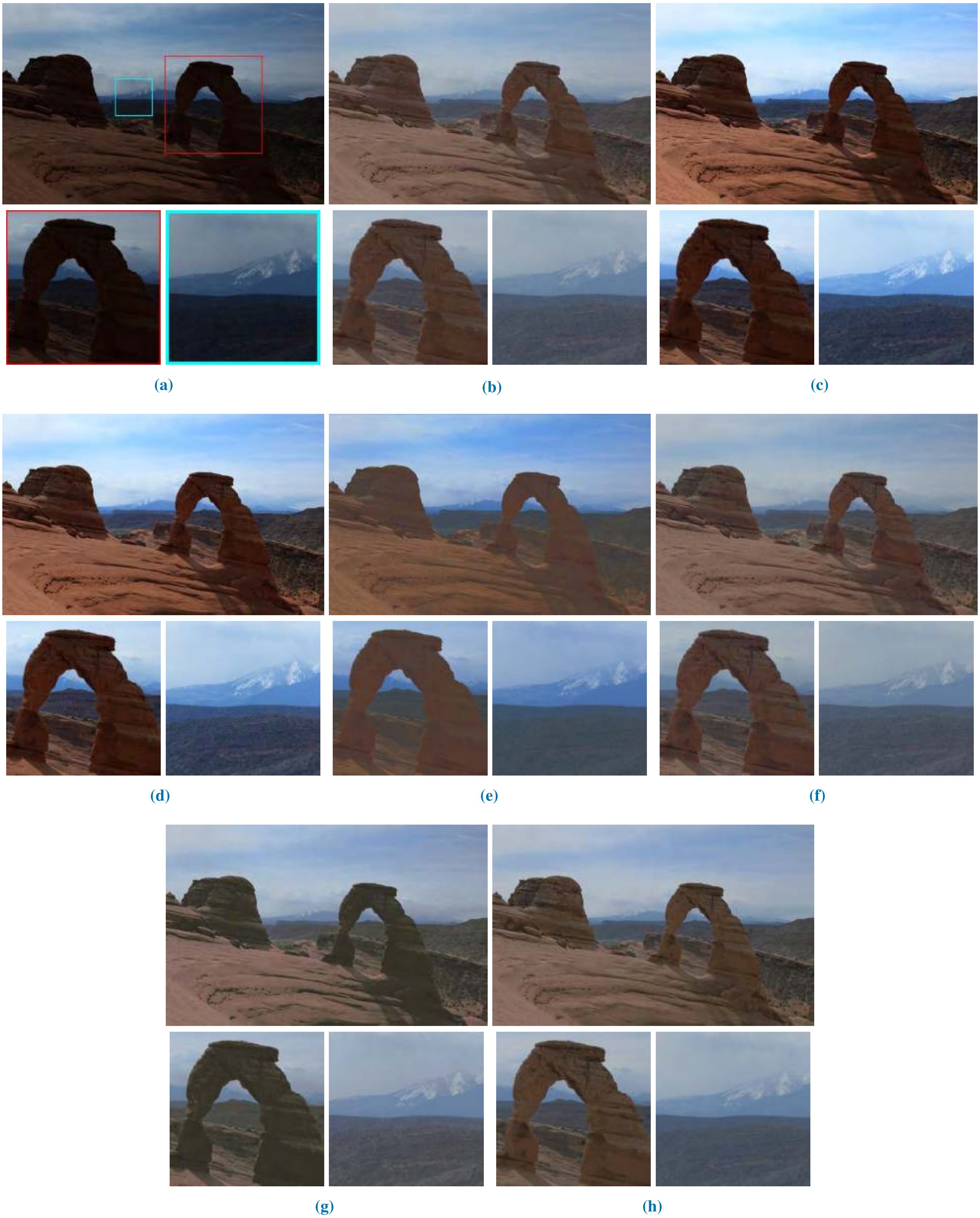

FIGURE 6. Experimental results [for image "DelicateArch"]. Zoom-ins of boxed regions are shown in bottom of each predicted HDR image. HDR images (b)–(h) were tone-mapped for visualization, where scaling to match range of predicted HDR images with that of original HDR one was not performed. Proposed iTM-Net (f) provided most similar image to original one (b), in six methods. (a) Input x. (b) Ground truth $\tilde{E}$ (c) Direct ITMO [15].HDR-VDP: 38.73, MS-SSIM: 0.6927. (d) PMET [13]. HDR-VDP: 38.71, MS-SSIM: 0.7183. (e) ExpandNet [18]. HDR-VDP: 57.72, MS-SSIM: 0.9442. (f) iTM-Net with $\mathcal{L}_{\text{iTM}}$ (Proposed). HDR-VDP: 68,87, MS-SSIM: 0.9933. (g) iTM-Net with $\mathcal{L}_1$. HDR-VDP: 58.73, MS-SSIM: 0.9692. (h) iTM-Net with $\mathcal{L}_{\text{Expand}}$. HDR-VDP: 57.73, MS-SSIM: 0.9522.

B. RESULTS

Figures 3 and 4 summarize quantitative evaluation results as box plots for the 44 images in terms of HDR-VDP and PU encoding + MS-SSIM, respectively. The boxes span from the first to the third quartile, referred to as Q_1 and Q_3, and the whiskers show the maximum and the minimum values in the range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. The band inside boxes indicates the median, i.e., the second quartile Q_2, and the cross inside boxes denotes the average value. A larger value for both metrics means higher similarity between a predicted HDR image and its original HDR image.

As shown in Figures 3 and 4, all iTM-Nets provided higher median and average scores in terms of the two metrics, than the three conventional methods including ExpandNet. These results indicate that HDR images predicted by iTM-Nets were more similar to the original HDR images than those predicted by the conventional methods. Since all of the predicted HDR images were scaled to match the original HDR images, the results illustrate that the proposed method can linearize images with high quality. Hence, it is confirmed that the proposed architecture can predict better HDR images than ExpandNet's one.

By comparing with iTM-Net with $\mathcal{L}_1$ and iTM-Net with $\mathcal{L}_{\text{Expand}}$, iTM-Net with the proposed loss $\mathcal{L}_{\text{iTM}}$ produced higher scores of both metrics. Hence, the proposed loss function is effective at training CNNs for inverse tone mapping.

Figures 5 and 6 show examples of HDR images generated by the six methods. Here, these images were tone-mapped from predicted HDR images because HDR images cannot be displayed in commonly used LDR devices, where scaling to match the range of predicted HDR images with that of corresponding original HDR ones was not performed. From Figs. 5 and 6, it is confirmed that the proposed method produced higher-quality HDR images, which are similar to corresponding original HDR ones $\tilde{E}$, than the other methods.

For these reasons, it is shown that the proposed method is effective at generating high-quality HDR images from single LDR images. In particular, the use of the proposed loss function enables us to improve the performance of CNNs for inverse tone mapping.

V. CONCLUSION

In this paper, a novel inverse tone mapping network, called "iTM-Net", was proposed. For training iTM-Net, a novel loss function that considers the non-linear relation between HDR and LDR images was also proposed. In the proposed loss function, target HDR images are tone-mapped into LDR images by an invertible tone mapping operator. The use of the proposed loss function enables us not only to normalize HDR images, but also to distribute the pixel values of HDR images like those of LDR ones. As a result, the performance of CNNs for inverse tone mapping can be improved. Experimental results showed that HDR images predicted by iTM-Net trained with the proposed loss function have higher quality than HDR ones predicted by conventional methods including the state-of-the-art in terms of HDR-VDP-2.2 and PU encoding + MS-SSIM. In addition, it was also confirmed that the proposed loss function improves the performance of CNNs compared with loss functions that do not consider the non-linear relation.

REFERENCES

[1] B. C. Madden, "Extended intensity range imaging," Dep. Comput. Inf. Sci., Univ. Pennsylvania, Philadelphia, PA USA, Tech. Rep. MS-CIS-93-96, 1993.

[2] S. Mann and R. W. Picard, "On being undigital with digital cameras: Extending dynamic range by combining exposed pictures," in *Proc. IST*, May 1995, pp. 422–428.

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

[4] M. G. Grossberg and S. K. Nayar, "Determining the camera response from images: What is knowable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1455–1467, Nov. 2003.

[5] M. Granados, B. Ajdin, M. Wand, C. Theobalt, H.-P. Seidel, and H. P. A. Lensch, "Optimal HDR reconstruction with linear digital cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 215–222.

[6] 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. Graph.*, vol. 31, no. 6, pp. 203:1–203:11, Nov. 2012.

[7] A. Badki, N. K. Kalantari, and P. Sen, "Robust radiometric calibration for dynamic scenes in the wild," in *Proc. IEEE Int. Conf. Comput. Photogr.*, Apr. 2015, pp. 1–10.

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

[9] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Inverse tone mapping," in *Proc. Int. Conf. Comput. Graph. Interact. Tech. Australas. Southeast Asia*, Nov./Dec. 2006, pp. 349–356.

[10] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, and G. Ward, "Ldr2Hdr: On-the-fly reverse tone mapping of legacy video and photographs," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007, Art. no. 39.

[11] P.-H. Kuo, C.-S. Tang, and S.-Y. Chien, "Content-adaptive inverse tone mapping," in *Proc. Vis. Commun. Image Process.*, Nov. 2012, pp. 1–6.

[12] H. Youngquing, Y. Fan, and V. Brost, "Dodging and burning inspired inverse tone mapping algorithm," *J. Comput. Inf. Syst.*, vol. 9, no. 9, pp. 3461–3468, May 2013.

[13] T.-H. Wang, C.-W. Chiu, W.-C. Wu, J.-W. Wang, C.-Y. Lin, C.-T. Chiu, and J.-J. Liou, "Pseudo-multiple-exposure-based tone fusion with local region adjustment," *IEEE Trans. Multimed.*, vol. 17, no. 4, pp. 470–484, Apr. 2015.

[14] Y. Kinoshita, S. Shiota, and H. Kiya, "Fast inverse tone mapping with Reinhard's global operator," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2017, pp. 1972–1976.

[15] Y. Kinoshita, S. Shiota, and H. Kiya, "Fast inverse tone mapping based on Reinhard's global operator with Estimated Parameters," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 100, no. 11, pp. 2248–2255, Nov. 2017.

[16] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. Graph.*, vol. 36, no. 6, p. 177, Nov. 2017.

[17] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, Nov. 2017, Art. no. 178.

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

[19] A. A. Goshtasby, "Fusion of multi-exposure images," *Image Vis. Comput.*, vol. 23, no. 6, pp. 611–618, Jun. 2005.

[20] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," *Comput. Graph. Forum*, vol. 28, no. 1, pp. 161–171, Mar. 2009.

[21] A. Saleem, A. Beghdadi, and B. Boashash, "Image fusion-based contrast enhancement," *EURASIP J. Image Video Process.*, vol. 2012, no. 1, Art. no. 10.

[22] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 626–632, May 2012.

[23] J. Wang, G. Xu, and H. Lou, "Exposure fusion based on sparse coding in pyramid transform domain," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service*, Aug. 2015, pp. 1–4.

[24] Z. Li, J. Zheng, Z. Zhu, and S. Wu, "Selectively detail-enhanced fusion of differently exposed images with moving objects," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4372–4382, Oct. 2014.

[25] T. Sakai, D. Kimura, T. Yoshida, and M. Iwahashi, "Hybrid method for multi-exposure image fusion based on weighted mean and sparse representation," in *Proc. Eur. Signal Process. Conf.*, Aug./Sep. 2015, pp. 809–813.

[26] M. Nejati, M. Karimi, S. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Fast exposure fusion using exposedness function," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 2234–2238.

[27] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4724–4732.

[28] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[29] Y. Kinoshita and H. Kiya, "Automatic exposure compensation using an image segmentation method for single-image-based multi-exposure fusion," *APSIPA Trans. Signal Inf. Process.*, vol. 7, pp. 1–10, Jan. 2018.

[30] Y. Kinoshita and H. Kiya, "Scene segmentation-based luminance adjustment for multi-exposure image fusion," *IEEE Trans. Image Process.*, to be published. doi: 10.1109/TIP.2019.2906501.

[31] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Feb. 2015, pp. 1–11, *arXiv: 1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, Jul. 2011, pp. 315–323.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Nov. 2015, pp. 234–241.

[35] *GitHub-Openexr*. Accessed: Mar. 2019. [Online]. Available: https://github.com/openexr/

[36] *High Dynamic Range Image Examples*. Accessed: Mar. 2019. [Online]. Available: http://www.anyhere.com/gward/hdrenc/pages/originals.html

[37] *The HDR Photographic Survey*. Accessed: Mar. 2019. [Online]. Available: http://rit-mcsl.org/fairchild/HDRPS/HDRthumbs.html

[38] *Max Planck Institut Informatik*. Accessed: Mar. 2019. [Online]. Available: http://resources.mpi-inf.mpg.de/hdr/gallery.html

[39] P. Zolliker and Z. Bara czuk, D. Küpper, I. Sprow, and T. Stamm, "Creating HDR video content for visual quality assessment using stop-motion," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

[40] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in LDR and HDR images," in *Proc. 9th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, Feb. 2015, pp. 1–6.

[41] 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.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, pp. 1–15, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[43] M. Narwaria, R. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images," *Proc. SPIE*, vol. 24, no. 1, Jan. 2015, Art. no. 010501.

[44] T. O. Aydin, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," *Proc. SPIE*, vol. 6806, Feb. 2008, Art. no. 68060B.

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[46] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, Art. no. 39.

YUMA KINOSHITA received the B.Eng. and M.Eng. degrees from Tokyo Metropolitan University, Japan, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. He was a recipient of the IEEE ISPACS Best Paper Award, in 2016, the IEEE Signal Processing Society Japan Student Conference Paper Award, in 2018, and the IEEE Signal Processing Society Tokyo Joint Chapter Student Award, in 2018. His research interest includes image processing. He is a Student Member of IEICE.

HITOSHI KIYA received the B.E. and M.E. degrees from the Nagaoka University of Technology, in 1980 and 1982, respectively, and the Dr.Eng. degree from Tokyo Metropolitan University, in 1987.

In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor, in 2000. From 1995 to 1996, he was a Visiting Fellow of the University of Sydney, Australia. From 2009 to 2013, he served as the Inaugural Vice President (Technical Activities) of APSIPA. He was the President of IEICE Engineering Sciences Society, from 2011 to 2012, and he served as a Vice President and the Editor-in-Chief for *IEICE Society Magazine and Society Publications*. From 2016 to 2017, he was a Regional Director-at-Large for Region ten of the IEEE Signal Processing Society. He serves as the President of APSIPA. He was an Editorial Board Member of eight journals, including the IEEE Transactions on Signal Processing, Image Processing, and Information Forensics And Security, the Chair of two technical committees and a member of nine technical committees including APSIPA Image, Video, and Multimedia Technical Committee (TC), and the IEEE Information Forensics and Security TC. He has organized a lot of international conferences, such as the TPC Chair of the IEEE ICASSP 2012 and as a General Co-Chair of the IEEE ISCAS 2019. He was a recipient of numerous awards, including six best paper awards. He is a Fellow of IEICE and ITE.