HDR Image Reconstruction Using Segmented Image Learning

BYEONG DAE LEE AND MYUNG HOON SUNWOO (Fellow, IEEE)

School of Electrical Engineering, Ajou University, Suwon 16499, South Korea

Corresponding author: Myung Hoon Sunwoo (sunwoo@ajou.ac.kr).

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01461) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and by the R&D program of MOTIE/KEIT. [No. 20010582, Developing low power deep learning HW IP design technology for image processing in CMOS Image sensors] and National R&D Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2021R1A2C2010228)

ABSTRACT

Converting a low dynamic range (LDR) image into a high dynamic range (HDR) image produces an image that closely depicts the real world without requiring expensive devices. Recent deep learning developments can produce highly realistic and sophisticated HDR images. This paper proposes a deep learning method to segment the bright and dark regions from an input LDR image and reconstruct the corresponding HDR image with similar dynamic ranges in the real world. The proposed multi-stage deep learning network brightens bright regions and darkens dark regions, and features with extended brightness range are combined to form the HDR image. Dividing the LDR image into the bright and dark regions effectively implements information on lost over-exposed and under-exposed areas, reconstructing a natural HDR image with color and appearance that is similar to reality. Experimental results confirm that the proposed method achieves an 8.52% higher HDR visual difference predictor (HDR-VDP) and a 41.2% higher log exposure range than current methods. Qualitative evaluation also verifies that the proposed method generates images that are close in quality to the ground truth.

INDEX TERMS

High dynamic range imaging, inverse tone mapping, deep learning, single exposure image

I. INTRODUCTION

Substantial image processing advances over the past few years have increased the need for technology to produce images similar to real world resolution. High dynamic range (HDR) is the most widely used image quality enhancement technology. Fig. 1 shows that HDR generates images with similar brightness to the range of human perception compared with low dynamic range (LDR) images with low brightness. However, HDR images should be obtained with expensive devices that can operate beyond the standard camera sensor limitations to acquire a wide dynamic range [1], and hence, most people have limited access to HDR images.

To overcome these limitations, multiple exposure fusion (MEF) methods [2], [3] have emerged, which generate multi-exposure LDR images. The images are then merged to reconstruct an HDR image. However, since multiple images cannot be taken simultaneously, the ghost artifact that projects the flow of time occurs when merging the bracketed images [4]. The areas with missing information also occur in ranges outside a given exposure.

Therefore, inverse tone mapping (ITM) methods [5] – [11] have been proposed to generate HDR images using single LDR images. However, previous ITM methods are insufficient for finding a function similar to the inverse camera response function (CRF), and obtaining parameters that satisfy all of the photo environments is difficult.

Deep learning methods have been proposed to overcome this problem, including convolutional neural network (CNN)-based MEF methods [12] – [18] and CNN-based ITM methods [19] – [23]. CNN-based MEF methods [12] – [18] are able to uniformly restore the saturated pixels by generating and merging multi-exposure images. However, they [12] – [18] have difficulty restoring the saturated pixels which do not exist in the brightness range of multi-exposure images. CNN-based ITM methods [19] – [23] have emerged to reconstruct HDR images from a single LDR image. These
methods map the LDR image to the HDR image through a CNN that acts as an inverse CRF. Nevertheless, despite the efforts to obtain near-original HDR images, effective restoration for saturated pixels is still limited. As shown in Fig. 2, the number of saturated pixels in the HDR image is minimal compared to the total number of pixels. Since image restoration typically uses loss functions based on L1, L2, or a mean square error (MSE), weights are updated to reduce the global pixel value differences between the ground truth images and the inferred images. As a result, the network weights are not efficiently updated to restore the saturated pixels. Moreover, since over-exposed pixels have very high values and under-exposed pixels have very low values, it is difficult to restore all of the saturated pixels uniformly. Therefore, this paper proposes a deep learning ITM method to generate HDR images by learning from segmented LDR images.

The main contributions of this paper are as follows:

1) The proposed method divides the input image into bright and dark regions and trains the two regions individually. Bright regions are learned to effectively restore over-exposed areas, and dark regions are learned to restore under-exposed areas. Therefore, the proposed approach can effectively restore all saturated pixels.

2) The proposed mask generation method for image segmentation can avoid artifacts. The proposed Mask allows smooth and clear image segmentation, avoiding boundary plane artifacts that occur when bright and dark regions are combined.

3) The proposed multi-stage network structure increases the dynamic range and uniformly restores the saturated pixels. To extend the dynamic range, bright regions are brightened, and dark regions are darkened. Then, the features are blended to generate the final HDR image.

The remainder of this paper is structured as follows: Section II discusses how the proposed method differs from the previous ITM method. Section III describes the algorithm, network structure, and loss function of the proposed method. Section IV describes the datasets and parameters used for training and testing. Section V discusses the qualitative and quantitative comparisons between the proposed ITM method and the existing ITM methods. Finally, Section VI summarizes and concludes the paper and discusses future study directions.

II. RELATED WORKS
The restoration of HDR images from the clipped LDR images is a complex problem. To solve this problem, various approaches have been studied in computer graphics, which can be classified into MEF and ITM methods.

A. MULTIPLE EXPOSURE FUSION METHODS
Debevec et al. [2] generated an HDR image by combining bracket images. Mertens et al. [3] used a Laplacian pyramid and a Gaussian pyramid to fuse the contrasting features in multi-exposure images. [2] and [3] proposed guidelines for the MEF method. However, the values of the bracket images were not adequately mixed during the image merging process, resulting in a halo artifact.

The CNN-based MEF methods have been proposed to optimally extract the weight values of bracket images and generate HDR images with evenly merged brightness values. Endo et al. [12] used U-Net [24] to generate HDR images with wide brightness ranges by bracketing images with various exposure values. Similarly, Lee et al. generated HDR images using a chain network to create multiple exposure stacks [13] and a generative adversarial network (GAN) [14], [15]. Since MEF methods generally use more than three LDR images, they require extensive processing time, power consumption, and storage capacity. To overcome these problems, Ma et al. [16] proposed a fast MEF method that generated weight maps for down sampled bracketed images using a multiscale context aggregation network (CAN) [25]. However, this method [16] also requires a large amount of computation since a large number of images were fed into the network.

To minimize the shortcomings of the CNN-based MEF methods, two exposure fusion (TEF) methods that use only two images have been proposed. Prabhakar et al. [17] proposed a network that fuses the luminance information of two exposure LDR images. Yin et al. [18] combined two exposure images through a prior aware-GAN, which consists of a content prior guided (CPG) encoder and a detail prior guided (DPG) decoder.

However, since TEF methods [17], [18] require two exposure images, generating images has a substantial impact on the results. Moreover, similar to the CNN-based MEF method, it is difficult to restore information that can be observed at exposure values other than preset exposure values.
B. INVERSE TONE MAPPING METHODS

Akyuz et al. [5] implemented HDR images using linear expansion based on gamma correction. However, efficient range expansion is difficult to achieve unless the input parameters are properly set. Banterle et al. generated HDR images using an expanded map and the median cut algorithm [6] along with a power function and an expansion map [7]. However, these approaches require parameters for density estimation, and the HDR image inferences have limitations for restoring over-exposed regions.

Huo et al. [8] generated HDR images with fewer parameters using a physiological ITM algorithm based on the human visual system (HVS), and Masia et al. [9] proposed an automatic global ITM based on gamma expansion. Nevertheless, the functions employed were insufficient to reconstruct saturated pixels.

In contrast, some proposed methods process a single image by segmenting them into specific ranges. Didyk et al. [10] classified images into diffuse, reflection, and light regions and subsequently applied different enhancements for each area to generate HDR images with a high luminance range. Rempel et al. [11] used Gaussian filtering and an image pyramid to extract pixels with high luminance and assigned information to saturated pixels. These approaches could effectively restore the saturated pixels but were unable to obtain functions satisfying the ground truth level color and texture.

The traditional ITM methods generate HDR images by applying a function to LDR images. However, finding a function similar to the inverse CRF is very difficult. Several deep learning methods have recently emerged to solve this tricky problem. Eilertsen et al. [19] generated HDR images using an autoencoder network with a loss function that considered illuminance and reflectance. They generated HDR images with an increased dynamic range by blending images to convert network outputs and LDR images to linear ranges. However, in contrast to the bright regions generated by the CNN, the dark regions were generated using virtual camera curves. Therefore, they could not effectively restore all saturated regions.

Cai et al. [20] decomposed the input LDR image into a low-frequency luminance region and a high-frequency detail region with weighted least squares (WLS) [26]. Then, the features of each region were extracted through a two-stage network. Cai et al. [20] generated HDR images with increased contrast, even in low-exposure images. However, they did not effectively increase the dynamic range since they focused on increasing the contrast rather than restoring the saturated pixels.

Marnerides et al. [21] utilized regional and local information from LDR images over an end-to-end network and divided it into global, dilation, and local branches. All reconstructions were performed within the network without human intervention. Although the proposed approach could obtain natural HDR images, there were limitations to infer the clipped pixel values of the under, over-exposed regions to have a dynamic range similar to the real world.

Liu et al. [22] proposed a method for generating HDR images using multiple networks based on a conventional camera pipeline structure that generated LDR images. They implemented dequantization, linearization, hallucination, and refinement networks to reverse the process of acquiring LDR images using deep learning. They obtained the final output by sequentially passing the LDR image to these networks. However, there is a significant risk that inference will significantly change if even one network is not properly learned, since this method utilizes various networks.

Ye et al. [23] showed that the bright and dark region characteristics differed, suggesting that different restoration methods should be employed for the different regions. Although they achieved effective restoration for both under- and over-exposed regions, it was difficult to increase the dynamic range efficiently due to the lack of direct brightness region enhancement.

Most previous CNN-based ITM methods reconstruct HDR images by inputting LDR images to the network without
significant pre-processing. Cai et al. [20] segmented the image through pre-processing, but based on only the frequency component. Eilertsen et al. [19] and Ye et al. [23] divided the image based on the brightness value, but [19] divided the image after learning, and [23] segmented it just before the final HDR image was generated. Therefore, it is difficult to claim that [19] and [23] can effectively restore the saturated pixels since CNN weights only consider creating an image similar to the ground truth image. Efficient reconstruction of saturated pixels becomes difficult using raw LDR image inputs. Therefore, this paper proposes separating LDR images into bright and dark regions before input and then learning them in different ways.

III. PROPOSED METHOD

Fig. 3 shows the overall flow of the proposed method. The image threshold is calculated during pre-processing, and the Mask is generated. The LDR image is then segmented into bright and dark regions (ImgBright and ImgDark, respectively) using the Mask, and ImgBright and ImgDark become the inputs of the proposed multi-stage network. ImgBright is learned to produce the increased brightness value in the Brightening block, and ImgDark is learned by the Darkening block to reduce the lower brightness value. Finally, the two feature maps with the enhanced dynamic ranges are combined to obtain the final HDR image. The original image ImgOriginal is also input into the network to help the calculations in the Brightening and Darkening blocks and feature blending.

A. PRE-PROCESSING: LDR IMAGE SEGMENTATION

Bright regions have information for only the bright pixels, and dark regions have information for only the dark pixels. As a result, the proposed network can effectively focus on over- and under-exposed pixels when bright regions and dark regions are fed into the network, respectively. Therefore, the input image is segmented into bright and dark regions during pre-processing to enhance the dynamic range.

Otsu [27] proposed the representative image segmentation method that separates the image into two classes using the threshold value that provides a maximum pixel variance over the image. In most cases, the Otsu [27] method classifies the image very clearly into two classes ImgBright and ImgDark. However, when the pixel values of the image are similar, the class boundaries are not clear because many pixel values are close to the threshold.

To clarify the difference between the two classes (ImgBright and ImgDark), it is necessary to find the threshold that maximizes the sum of the variances between the two classes, and the threshold should be located at the valley point of the image histogram [28]. Therefore, this paper proposes a method to find the optimal valley point in the histogram. Before applying the proposed method, the brightness value \( V \) is measured by converting the RGB image to the HSV color space, and a histogram is generated.

| Algorithm 1 Finding Threshold |
|--------------------------------|
| 1: Input: \( H \) (Histogram), \( O \) (Otsu Threshold) |
| 2: Output: \( \text{thr} \) (Threshold) |
| 3: \( A = [] \# \) Valley point set |
| 4: \( T = \text{sum}(H) \) |
| 5: for \( j \) in range(\( 0, 252 \)) |
| 6: \( k1 = H(j) - H(j-1) < 0 \) and \( H(j+1) - H(j) > 0 \) |
| 7: \( k2 = H(j) - H(j-2) < 0 \) and \( H(j+2) - H(j) > 0 \) |
| 8: \( k3 = H(j) - H(j-3) < 0 \) and \( H(j+3) - H(j) > 0 \) |
| 9: if \( k1 \) and \( k2 \) and \( k3 \): \# Find valley point |
| 10: \( V1 = \text{var}(H(0:j)) \) |
| 11: \( V2 = \text{var}(H(j+1:255)) \) |
| 12: \( W1 = \text{sum}(H(0:j))/T \) |
| 13: \( W2 = \text{sum}(H(j+1:255))/T \) |
| 14: \( s = V1 \times W1 + V2 \times W2 \) |
| 15: \( A.append(s, j)) \) |
| 16: end if |
| 17: end for |
| 18: if \( A \) is not empty: |
| 19: \( \text{thr} = \text{max}(A)[1] \) |
| 20: else: |
| 21: \( \text{thr} = 0 \) |
| 22: end if |

![Image](Image)

**FIGURE 4.** The flow of the proposed segmentation algorithm. (a) Find the Otsu threshold value, (b), (c) find the valley points, and (d) find the optimal valley point.
Alg. 1 describes the algorithm of the proposed method, and Fig. 4 shows an example of finding the optimal valley point. As described in Alg. 1, the first step is to increase the pixel value from the Otsu threshold value to 252 (Fig. 4 (a)). The second step is to find the valley points by checking the slope between the histogram value at the current pixel and the histogram values at the points \( k \in \{ \pm 1, \pm 2, \pm 3 \} \), which represent the distance from the current pixel (Alg. 1 lines 6 – 8). Third, the two classes are classified based on the valley point, and the sum of the variance values of the classes, i.e., \( s \), is calculated (Alg. 1 lines 10 – 14). Then, we include \( s \) and valley points in the valley set \( A \) (Fig. 4 (b), (c)). Finally, in the valley set, we can find the valley point with the largest \( s \) and assign it as a threshold, i.e., \( thr \). (Fig. 4 (d)).

The proposed Mask generated based on the \( thr \) can be expressed as

\[
Mask(n,m) = \left( \frac{\max(0, V(n,m) - thr)}{(1 - thr)} \right)^2,
\]

where \( V(n,m) \) is image brightness; \( n, m \) are horizontal and vertical pixels, respectively. Since the square function is used for \( Mask \) (1), the pixel with a brightness value slightly larger than \( thr \) has a very small mask value, which makes the boundary surface smooth when segmenting the image.

Fig. 5 and Fig. 6 show the images segmented by the proposed method and Otsu’s [27] method. As shown in Fig. 5, the \( thr \) values of some images are the same as the Otsu threshold values. However, as shown in Fig. 6, if most image pixels have similar values, the Otsu threshold may not be located at the optimal valley point. As a result, the dividing boundary is created at any false valley point, as shown in Fig. 6 (d). In contrast, the proposed method can find a new optimal valley point that clearly separates the bright and dark regions, as shown in Fig. 6 (c).

The generated Mask is then multiplied by the original LDR image and segmented into \( Img_{Bright} \) and \( Img_{Dark} \):

\[
Img_{Bright} = Img \times (Mask)
\]

and

\[
Img_{Dark} = Img \times (1 - Mask).
\]

The proposed Mask is similar to those previously used in [15], [19], and [23], but the proposed method enables more accurate image segmentation by setting the \( thr \). The generated Mask is also non-linear, i.e., the square of the difference, producing smoother segmentation boundaries. Consequently, the input LDR image can be segmented into \( Img_{Bright} \) and \( Img_{Dark} \), as shown in Figs. 5 and 6 (c).

B. PROPOSED NETWORK ARCHITECTURE

Fig. 7 shows the proposed multi-stage CNN structure. Network inputs include \( Img_{Bright} \) and \( Img_{Dark} \), generated in pre-processing, and the original LDR image \( Img_{Original} \). \( Img_{Bright} \) expands the dynamic range using the Brightening block (Fig. 8 (a)), increasing brightness, and \( Img_{Dark} \) expands the dynamic range using the Darkening block (Fig. 8 (c)), reducing brightness. The two feature maps from distinct learning processes are subsequently combined properly to prevent artifacts.

The proposed network structure includes three stages.

1) The **dynamic range extension stage** increases the bright range brightness and decreases the dark range brightness;
2) The **global feature concatenation stage** prevents outliers; and
3) The **blending and generation stage** combines the learned feature maps.

Details for each stage are explained in the following sections.

1) **DYNAMIC RANGE EXTENSION**

Fig. 8 shows that the Brightening (Fig. 8 (a)) and Darkening (Fig. 8 (c)) blocks are based on Zhang *et al.* [29] proposed residual channel attention block (RCAB), with squeeze
parameters obtained from the original LDR image. For example, brightness should be adjusted more strongly for images taken outdoors with strong sunlight because most pixels have fairly bright values. In contrast, image brightness should be slightly weakened for images taken indoors because they have appropriate brightness. Thus, feature brightness changes depending on the input image (Fig. 8 (b)). The feature brightness values can be obtained as

\[ Z_B = P_{\text{Avg}}(P_{\text{max}}(F_{\text{Original}})) \]

and

\[ Z_D = P_{\text{Avg}}(P_{\text{min}}(F_{\text{Original}})), \]

where \( F_{\text{Original}} \) is a feature map created after the original image passed through the convolution layer. We use the ReLU activation function; hence, \( P_{\text{max}}, P_{\text{min}}, \) and \( P_{\text{Avg}} \) denote the maximum, minimum, and average pooling, respectively. The squeeze parameter \( Z_B \) for the Brightening block is obtained by maximum pooling with \( 4 \times 4 \) pixels in \( F_{\text{Original}} \) and the average pooling with \( 1 \times 1 \) pixels to properly represent brightness (4).

Akyuz et al. [5] showed that HVS is more sensitive to dark than bright regions, suggesting that dynamic ranges should be scaled to different sizes based on the range. In particular, Ye et al. [23] argued that the dark regions texture and color information are ambiguous, in contrast to the bright regions; hence, it may not be able to effectively restore under-exposed ranges if brightness changes significantly. Reibel et al. [31] showed that dark regions filmed with a camera contain more noise than bright regions; hence, unintended artifacts can appear if the change is severe when darkening the dark regions. Thus, the degree of change in \( \text{Img}_{\text{Dark}} \) should be less than that for \( \text{Img}_{\text{Bright}} \). Hence, the squeeze parameter \( Z_D \) used in the Darkening block was obtained using the minimum rather than maximum pooling (5).

The final channel statistics \( C_B \) and \( C_D \) were determined after calculating \( Z_B \) and \( Z_D \),

\[ C_B = F_s(F_r(Z_B)) \]

and

\[ C_D = F_s(F_r(Z_D)), \]

where \( F_r \) is a convolutional layer employing the ReLU activation function, and \( F_s \) is a convolutional layer employing the sigmoid activation function. As a gating function for rescaling the size of feature maps, we used the sigmoid as in [28]. The channel size was downscaled to \( 64/r \) at \( F_r \) and then upscaled to \( r \) again at \( F_s \). The reduction ratio \( r \) for reducing the parameter overhead was designated as 16 from Zhang et al. [29]. Finally, feature maps \( F_{\text{Bright}} \) and \( F_{\text{Dark}} \) with adjusted brightness values can be obtained as

\[ F_{\text{Bright}} = X_{\text{Bright1}} + C_B \cdot X_{\text{Bright2}} \]

and

\[ F_{\text{Dark}} = X_{\text{Dark1}} - C_D \cdot X_{\text{Dark2}}, \]

where \( X_{\text{Bright1}} \) and \( X_{\text{Dark1}} \) represent the convolutional layers for \( \text{Img}_{\text{Bright}} \) and \( \text{Img}_{\text{Dark}} \), respectively, and \( X_{\text{Bright2}} \) and \( X_{\text{Dark2}} \) represent the...
calculation parameter extraction from the convolutional layers for \(X_{\text{Bright1}}\) and \(X_{\text{Dark1}}\), respectively. 

\(F_{\text{Bright}}\) generated using (8) has increased brightness compared to \(X_{\text{Bright1}}\), and \(F_{\text{Dark}}\) generated using (9) has reduced brightness compared to \(X_{\text{Dark1}}\). In particular, the degree of change in \(F_{\text{Dark}}\) is reduced in comparison to \(F_{\text{Bright}}\), preventing possible problems when restoring underexposed ranges.

When PReLU [32] is used as the activation function, negative parameters may remain in the feature map, which can cause \(I_{\text{mgBright}}\) to be darker rather than brighter and \(I_{\text{mgDark}}\) to be brighter. To prevent negative parameters, ReLU [30] was used as an activation function of the brightening and darkening blocks. Consequently, the bright image is brighter and the dark image is darker through the Brightening and Darkening blocks, respectively. Naturally inferred feature maps can be obtained since the dynamic range expansion was achieved within a single network.

2) GLOBAL FEATURE CONCATENATION

Unwanted outliers can be generated from the final HDR image because the level at which brightness values change is determined automatically within the proposed deep learning network. Although the image is smoothly segmented during pre-processing, artifacts can appear around the segmented image boundaries when combining the two feature maps, since \(I_{\text{mgBright}}\) and \(I_{\text{mgDark}}\) learning were conducted using different parameters.

Therefore, we used the global branch structure proposed by Marnerides et al. [21] to generate a feature map with abstract features from the original image, which limits significant deformation for the image learned through this feature map. Fig. 7 shows the generated global image feature \(F_{\text{Global}}\) with size 64 that is subsequently used as

\[
F_{BG} = F_P([F_{Bright}, F_{Global}]),
\]

\[
F_{DG} = F_P([F_{Dark}, F_{Global}]),
\]

and

\[
F_{Sum} = F_P([(F_{BG} + F_{DG}), F_{Global}]),
\]

where \(F_P\) is a convolution layer that uses the PReLU [32] activation function.

Thus, \(F_{Global}\) was used to combine \(F_{BG}\) and \(F_{DG}\) segmented feature maps. Since \(F_{BG}\), \(F_{DG}\), and \(F_{Sum}\) contain abstract information from the original LDR image, using these maps rather than \(F_{Bright}\), \(F_{Dark}\), and \(F_{BG} + F_{DG}\) prevented excessive outlier generation when increasing the dynamic range.

3) BLENDING AND GENERATION

Although \(F_{Global}\) was used to concatenate \(F_{Bright}\) and \(F_{Dark}\) (see (12)), a deep learning network structure was required to properly blend the feature maps that were learned under different directions. Therefore, we merged the feature maps using DenseNet [33], which could effectively extract input image features using only a few parameters, as shown in Fig. 4. Thus, \(F_{Sum}\) passed through a 3-Layer DenseNet,

\[
F_{D-Layer1} = F_P(F_{Sum}),
\]

\[
F_{D-Layer2} = F_P([F_{Sum}, F_{D-Layer1}]),
\]

and

\[
F_{D-Layer3} = F_P([F_{Sum}, F_{D-Layer1}, F_{D-Layer2}]),
\]

where \(F_{D-Layer1}\) and \(F_{D-Layer2}\) are vectors with \(G\) features and \(F_{D-Layer3}\) is a vector with 64 features. DenseNet’s growth rate \(G = 32\), half that for the input feature channel. Finally, we created the HDR image as

\[
\text{HDR image} = F_5(F_P(F_{D-Layer3})).
\]

A sigmoid function was used in the final layer to map the generated HDR image to the \([0, 1]\) range.

C. LOSS FUNCTION

The Weber–Fechner law [34] states that stimuli should be transformed into a constant ratio rather than a constant quantity for human senses to detect intensity changes because luminance variation is logarithmic; hence, loss functions should be handled in the logarithmic domain to create the HDR image optimized for human senses. Therefore, we used the logarithmic loss function

\[
L(h, h) = L_{RGB}(h, h) + \beta L_{\text{lumiance}}(h, h),
\]

where \(h\) is the HDR image estimated by the network, \(h\) is the ground truth HDR image, and \(\beta\) is the loss function weight.

We set \(\beta = 0.3\) to generate the highest experimental result and implemented the loss function as a sum of \(L_{RGB}\), comparing
pixel values in RGB and $l_{\text{Luminance}}$ in luminance to reconstruct the HDR image as close to real life as possible,

$$L_{\text{RGB}}(h, h) = |\log(Y(h + \epsilon)) - \log(Y(h + \epsilon))|$$

and

$$L_{\text{Luminance}} = |\log(Y(h + \epsilon)) - \log(Y(h + \epsilon))|.$$  

$L_{\text{RGB}}$ (18) and $L_{\text{Luminance}}$ (19) were based on the L1 distance and use a small constant $\epsilon$ to eliminate specificity at a zero-pixel value. $L_{\text{RGB}}$ (18) computes the difference between the ground truth and the predicted values in the RGB range, whereas $L_{\text{Luminance}}$ (19) computes in the luminance range. According to [1], the pixel value of the HDR image is an approximation of photometric quantities, in contrast with the LDR image, and is linearly related to luminance in general. Therefore, it is reasonable to evaluate the HDR image in the luminance range, extracting luminance values from the RGB range,

$$Y = 0.2126R + 0.7152G + 0.0722B,$$

where $R$, $G$, and $B$ indicate red, green, and blue channel pixel values, respectively. Finally, we obtain the absolute difference between the estimated and ground truth after converting HDR pixel values to the luminance range using (20).

IV. EXPERIMENTS

A. Dataset

Datasets are important factors in determining deep learning networks. Moreover, objectivity is maintained by learning in an environment with similar objects. Therefore, we used the most common HDR image reconstruction datasets: Funt [35], [36], Stanford [37], Ward [38], [39], and Fairchild [40]. Since these datasets include only HDR images rather than LDR–HDR pairs, we employed tone mapping [41] – [43] to synthesize corresponding LDR images, obtaining 4,935 total LDR images by randomly cropping 512×512 regions five times for each tone mapping.

We also used the HDR-Real dataset organized by Liu et al. [22], which included 480 HDR and 4,893 LDR images. LDR images were taken with 42 different cameras and converted into HDR images using Photomatix software [44]. Therefore, there was no need to generate additional LDR images, unlike the other datasets. Of the 9,828 data images, we used 8,118 images for training and the remaining 1,710 for testing. The training images were resized to 256×256 pixels, and HDR images were normalized onto $[0, 1]$ for uniform learning. For testing, we used the 512×512 pixel images.

B. Training

Training and testing were implemented using Python v.3.6 on an Intel i7-7820X CPU with an Nvidia GTX 1080 Ti GPU. We conducted 200 repeated training sessions on the 8,118 images, and the results were checked every 20 epochs to prevent overfitting. The initial learning rate was 0.00007, and we use the Adam optimizer [45]. Weights were reset using Xavier initialization [46], mini batch size = 8, and 1014 iterations were run for each epoch.

V. RESULTS

This section compares the results from the proposed and the existing methods using the peak signal noise to rate (PSNR) [47], structural similarity index measure (SSIM) [48], and multiscale SSIM (MS-SSIM) [49] metrics, which are the most common evaluation indicators. PSNR is calculated based on the MSE, which is the Euclidean distance between pixels, with a higher value implying better similarity to the original. SSIM measures the similarity to the original image using pixel variance and the mean from the two images. MS-SSIM is calculated by adding scale-space concepts to SSIM. Reconstructed images are structurally similar to the original images where SSIM and MS-SSIM are close to 1. PSNR, SSIM, and MS-SSIM metrics express mathematical pixel differences between the original and the comparison image. However, HDR images should be measured based on human visual perception as well as mathematical differences. Therefore, in this paper, we also use the quality Q score for the HDR visual difference predictor (HDR-VDP) [50] as an evaluation metric, which is based on the human visual model for all luminance environments and has the highest level of objectivity to assess HDR image quality. The log exposure range is used to evaluate the brightness range, which is the most critical element for the HDR image.

$$\text{Log Exposure Range} = \log_2(l_{\text{high}}) - \log_2(l_{\text{low}})$$

where $l_{\text{high}}$ and $l_{\text{low}}$ are the highest and smallest luminance, respectively. The dynamic range of the HDR image is evaluated by comparing the log exposure range with reference methods, including the traditional and previous deep learning methods. The traditional ITM methods include Akyuz et al. [5], Huo et al. [8], and Masia et al. [9], referred to as AEO, HOU, and MEO, respectively, and the deep learning methods include Endo et al. [12], Eilertsen et al. [19], Cai et al. [20], and Marnerides et al. [21], referred to as DrTMO, HDRCNN, SICE, and EXP, respectively.

A. QUANTITATIVE COMPARISON

Table 1 compares PSNR, SSIM, MS-SSIM, and HDR-VDP for the various methods. The proposed method achieves the highest score on all four evaluation metrics. In particular, the structural detail is increased, returning higher SSIM and MS-SSIM scores than the other methods due to the increased contrast with a wider brightness range. The VDP-Q score is also the highest for the proposed method, i.e., the best HDR image is reconstructed with respect to human cognitive aspects.

The deep learning methods [12], [19] – [21] use map-ranged ground truth images in the range $[0, 1]$. In contrast, the traditional ITM methods [5], [8], [9] multiple the LDR image by specific functions without a specified range to create the
HDR image. Consequently, when mapping the HDR image of [5], [8], [9] generated for objective comparison to a range of [0, 1], the minimum and maximum values of pixels are set to 0 and 1. Therefore, the size of the log exposure range becomes infinite, which makes it impossible to objectively compare with the deep learning methods. Therefore, Fig. 13 compares the AEO, HUO, and MEO luminance ranges with the proposed method by adjusting the exposure values.

Table 2 compares the log exposure range between the proposed method and the existing deep learning methods [12], [19] – [21], where the entire test dataset (1,710 images) includes 307 night or dark room images with low brightness and 322 daytime or bright room images with high brightness. The mean log exposure range for the ground truth images = 16.15 Stops, whereas HDR images generated by the proposed method achieve a mean log exposure range = 15.07 Stops. The mean log exposure range of the proposed method is the closest result to the ground truth, with a subsequent decreasing order of HDRCNN, SICE, EXP, and DrTMO.

The proposed method also achieves the best results for test images with high brightness, a ground truth error = 2.4% and a mean log exposure range = 12.13 Stops, which means that the saturated pixels are effectively restored, even in images with many over-exposed regions. In addition, for test images with low brightness, the proposed method shows a mean log exposure range = 15.28 Stops, which is 38.49% higher on average than those of the other deep learning methods [12], [19] – [21].

**B. QUALITATIVE COMPARISON**

In this paper, we use Reinhard [43] tone mapping with the 8-bit LDR image for the HDR image to compare with the proposed method. Qualitative image comparisons are evaluated for less saturated and highly saturated LDR image inputs. The dynamic brightness range is compared by increasing HDR image exposure.

**TABLE 1. Comparisons of image qualification among the reference methods and the proposed method. Bold indicates the best score.**

|                      | AEO [5] | HUO [8] | MEO [9] | DrTMO [12] | HDRCNN [19] | SICE [20] | EXP [21] | Proposed |
|----------------------|---------|---------|---------|------------|-------------|-----------|----------|----------|
| PSNR (dB)            | 19.08   | 20.82   | 18.55   | 18.80      | 24.15       | 23.28     | 22.18    | 27.08    |
| SSIM                 | 0.5106  | 0.5726  | 0.4880  | 0.3545     | 0.5810      | 0.6085    | 0.5493   | 0.7265   |
| MS-SSIM              | 0.4788  | 0.5457  | 0.4534  | 0.3739     | 0.5811      | 0.5996    | 0.5392   | 0.7124   |
| HDR-VDP 2.2 (Q score)| 44.52   | 44.44   | 40.37   | 45.51      | 48.52       | 47.12     | 46.34    | 48.97    |

**TABLE 2. Comparisons of log exposure range among the reference methods and the proposed method. Bold indicates the best score.**

| Ground truth | All images | Bright images | Dark images |
|--------------|------------|---------------|-------------|
|              | 16.15      | 12.43         | 18.34       |
| DrTMO [12]   | 9.06       | 8.43          | 9.1         |
| HDRCNN [19]  | 11.57      | 10.8          | 11.93       |
| SICE [20]    | 11.53      | 10.38         | 11.66       |
| EXP [21]     | 10.96      | 10.43         | 12.04       |
| Proposed     | 15.07      | 12.13         | 15.28       |

**1) LESS SATURATED LDR IMAGES**

To evaluate images in various environments, photographs taken in bright and dark places were compared. Fig. 9 compares the LDR image for a park in the daytime with the reconstructed HDR image. Sun shape, nearby tree details (red box), and bush color information (blue box) were damaged due to the bright light sources. The HDR image generated by the proposed method (j) was closest to the ground truth (b). In particular, the image captured the sun shape (red box) due to high brightness most naturally and had the most similar color to the ground truth (blue box).

The traditional ITM methods (c) – (e) generated HDR images that were similar to the input LDR image (a), but the deep learning methods (f) – (j) reconstructed HDR images that had better quality with SICE (h), generating the second best HDR image after the proposed method. However, SICE (h) did not accurately reconstruct the sun shape (red box), HDRCNN (g) and EXP(i) had lower overall contrast than the proposed method, and DrTMO (f) generated an overall blurred HDR image.

Fig. 10 compares HDR images inferred from the LDR image of the park at night. Information regarding streetlights (red box) and grass (blue box) reflected on the streetlights was lost due to over-exposed ranges having a higher brightness than under-exposed ranges. Similar to the case for daytime park images, the traditional ITM methods (c) – (e) generated HDR images that did not adequately restore the saturated pixel values for lights (red box) and grass (blue box) and differed little from the input LDR image (a). DrTMO (f) generated the unnatural HDR image compared to the ground truth (b) in the process of merging the bracketed images. HDRCNN (g), SICE (h) and EXP (i) detected saturated pixels for streetlights (red box) and grass (blue box) better than the other methods, but they exhibited significant sharpness loss compared to the ground truth (b).

In contrast, the proposed method (j) restored the saturated streetlight information (red box) closest to the ground truth (b) and was also able to realize similar grass color and tone (blue box). Thus, the proposed method could restore the clearest HDR image.

**2) HIGHLY SATURATED LDR IMAGES**

Fig. 11 compares the outcomes for the LDR image that was severely clipped and therefore lost most of its information before being restored. The original LDR image (a) was taken.
inside a dark building, with considerable information loss outside the relatively bright building (red box) compared to the interior.

As shown in Fig. 11, the traditional ITM methods (c) – (e) that did not use deep learning did not properly restore the information outside of the building (red box). On the other hand, the methods using deep learning effectively inferred the
saturated pixels based on the limited information, and in particular, the proposed method (j) restored the most out-of-building information (red box) corrupted by over-exposed. DrTMO (f) is the HDR image synthesized stacks of various exposed images, effectively realizing the damaged region in the bright range, and hence, restoring more shapes than the other comparison methods. However, DrTMO (f) was quite blurred, with the overall color tone and contrast differing greatly from the ground truth (b). EXP (i) could not restore as many saturated pixels (red box) as DrTMO (f), but overall reconstructed the good HDR image, since EXP (i) efficiently used local and global information from the input LDR image. SICE (h) extracted the feature of the low- and high-frequency information of the image and generated the HDR image with color and contrast close to the ground truth (b), but it could not restore the detailed information of over-exposed pixels.

Fig. 12 shows the example regions where under-exposed information was damaged, except for very bright pixels, since the exposure time was very short. Most LDR image information was damaged, aside from the tree bulbs, making it difficult to visibly identify image shapes (red box). Generating the HDR image from such a damaged LDR image is a painful task.

Not all comparison methods could generate images close to the ground truth (b). HUO (d) could not restore most information, and although MEO (e), DrTMO (f), SICE (h), and EXP (i) restored the statue (red box) to some extent, the restored images contain substantial noise and are severely blurred. AEO (c) and HDRCNN (g) were clearer than those by MEO (e), DrTMO (f), SICE (h), and EXP (i), but the colors deviated considerably from the ground truth (b). In contrast, the image reconstructed by the proposed method (j) achieved higher clarity and a similar color tone to the ground truth.

3) COMPARISON FOR DYNAMIC RANGE

The proposed method generates images with a wider dynamic range than the existing approaches (see Table 2). For qualitative analysis of the dynamic range, we observed the HDR image up to 4 exposure value (EV) units. We did not use the Reinhard [43] tone mapping method to compare methods, rather, we multiplied the HDR image by 255 and converted the values into integers.

Fig. 13 shows that the proposed method gives the most information for all of the exposure values. All considered methods showed similar information when the exposure was increased by 4 EVs. However, DrTMO (d) lost the most pixel value information when EV was increased by 8, and AEO (a), MEO (c), HDRCNN (e), SICE (f), and EXP (g) lost almost all pixel value information when EV was increased by 12, except for very blurry shapes. In contrast, HUO (b) and the proposed method (h) retained pixel information in most ranges, except in over-exposed regions, and the proposed method retained more detailed information than HUO (b). HUO (b) and the proposed method (h) lost most of the pixel information when EV was increased by 16, but the proposed method retained the most information compared with the original image.

Thus, the HDR image generated by the proposed method had the widest luminance range and retained more information, even for high-exposure regions. In particular, not only was the brightness range extended, but it was also very low, while high
To examine whether the saturated pixels could be effectively restored, we compared the results between the methods with and without segmentation. Table 3 shows that the saturated pixels were effectively restored using the method with segmentation by showing higher scores in all metrics, especially with 2.65 dB PSNR higher than the method without segmentation.

Nevertheless, the method without segmentation had an average log exposure range that was 33% higher than the other deep learning methods [12], [19] – [21] (see Table 2). Even if only the global image $\mathbf{img}_{\text{original}}$ was used, the proposed network could restore the saturated pixels more precisely. However, by dividing the image into two regions, the proposed network could focus more on the saturated pixels in each region. Therefore, segmenting the image is much more effective for restoring the saturated pixels.

Fig. 14 shows HDR images reconstructed by the method with segmentation and the method without segmentation. In both (a) and (b), when the segmented image is used as input, the color and shape information of the saturated pixels are restored in more detail, and the contrast of the image increases as the dynamic range improves.
We compared the results of cases with and without $F_{Global}$ in the proposed network that uses the segmented images as input to verify whether $F_{Global}$ reduces the number of artifacts. Table 4 compares the quantitative evaluation with and without $F_{Global}$ in the proposed network. As shown in Table 4, when $F_{Global}$ was used in the proposed network, the PSNR was 4.1% higher than that without $F_{Global}$. In other words, using $F_{Global}$ could improve the overall HDR image quality since the number of artifacts could be reduced on the segmentation boundary.

Fig. 15 shows the HDR images reconstructed by the proposed network with $F_{Global}$ and the proposed network without $F_{Global}$. As shown in Fig. 15, when $F_{Global}$ was not used, artifacts appeared at the division boundary near the reflected light area of the ceiling. However, when $F_{Global}$ was used, the boundary of the reflected light area of the ceiling showed a natural shape without any artifacts.

**TABLE 4.** Comparisons of image qualification among the proposed network without the global feature and the proposed network with the global feature. Segmented images are fed into each network. Bold indicates the best score.

|                        | Without Global feature | With Global feature |
|------------------------|------------------------|---------------------|
| PSNR (dB)              | 26.01                  | 27.08               |
| SSIM                   | 0.6980                 | 0.7265              |
| MS-SSIM                | 0.6845                 | 0.7124              |
| HDR-VDP 2.2 (Q score)  | 48.47                  | 48.97               |
| Log Exposure ratio (STOPS) | 14.95                | 15.07               |

**FIGURE 15.** Comparisons of HDR images among the proposed network without the global feature and the proposed network with the global feature.

**VII. CONCLUSIONS**

Recent developments in deep learning technology have substantially advanced ITM approaches, with many studies employing deep learning to reconstruct saturated pixels and regenerate high-quality HDR images [12] – [23]. However, the LDR image should be separated and learned according to brightness values to accurately reflect reality. Reconstructed HDR images using the proposed method achieved average PSNR, SSIM, MS-SSIM, and HDR-VDP2 metrics of 30.4%, 42.8%, 42.9%, and 8.52% higher than the previous methods, respectively. In particular, the proposed method provided a high mean log exposure range = 41.2%, confirming that the dynamic range was the most similar to the real world. Furthermore, qualitative evaluations enabled the generation of an HDR image similar to the original image with the highest resilience, even though the input LDR image had lost considerable pixel information.

Generating the image close to the real world using limited information from the clipped LDR image remains difficult. The proposed method could not reconstruct irradiance perfectly, and LDR images with severely damaged exposure values remain unable to generate HDR images close to the ground truth. We think this issue can be resolved by classifying images into more classes for training. In addition to the bright and dark ranges, the LDR image could be further classified by brightness, helping the clipped ranges to become effectively generated.

**REFERENCES**

[1] R. K. Mantik, K. Myszowski, and H. P. Seidel, “High dynamic range imaging,” Wiley encyclopedia of electrical and electronics engineering, pp. 1-42, 1999.

[2] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in Proc. ACM SIGGRAPH Classes, 2008, p. 31.

[3] T. Mertens, J. Kautz, and F. Van Reeth, “Exposure fusion,” in Proc. 15th Pacific Conf. Comput. Graph. Appl. (PG), Maui, HI, USA, Oct./Nov. 2007, pp. 382-390.

[4] S. Srikanta and D. Siddh, “Ghost detection and removal for high dynamic range images: Recent advances,” Signal Process., Image Commun., vol. 27, no. 6, pp. 650–662, 2012.

[5] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, “Do HDR displays support LDR content? A psychophysical evaluation,” ACM Trans. Graphics, vol. 26, no. 3, pp. 38, 2007.

[6] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, “Inverse tone mapping,” in Proc. 4th Int. Conf. Comput. Graph. Interact. Techn. Australasia Southeast Asia, New York, NY, USA, 2006, pp. 349–356.

[7] F. Banterle et al., “High dynamic range imaging and low dynamic range expansion for generating HDR content,” in Comput. Graph. Forum, vol. 28, no. 8, pp. 2343-2367, 2009.

[8] Y. Huo, F. Yang, L. Dong, and V. Brost, “Physiological inverse tone mapping based on retina response,” The Vis. Comput., vol. 30, no. 5, pp. 507-517, 2014.

[9] B. Masia, A. Serrano, and D. Gutierrez, “Dynamic range expansion based on image statistics,” Multimedia Tools Appl., vol. 76, no. 1, pp. 631–648, 2017.

[10] P. Didyk, R. Mantik, M. Hein, and H.-P. Seidel, “Enhancement of bright video features for HDR displays,” in Comput. Graph. Forum, vol. 27, no. 4, pp. 1265-1274, 2008.

[11] A. G. Rempel et al., “LDr2hdr: on-the-fly reverse tone mapping of legacy video and photographs,” ACM Trans. Graph., vol. 26, no. 3, 2007, Art. no. 39.

[12] Y. Endo, Y. Kanaori, and J. Mitani, “Deep reverse tone mapping,” ACM Trans. Graph., vol. 36, no. 6, p. 177, 2017.

[13] S. Lee, G. H. An, and S.-J. Kang, “Deep chain hdr: Reconstructing a high dynamic range image from a single low dynamic range image,” IEEE Access, vol. 6, pp. 49913-49924, 2018.

[14] S. Lee, G. H. An, and S.-J. Kang, “Deep recursive HDR: Inverse tone mapping using generative adversarial networks,” in Proc. Eur. Confr. Comput. Vis. (ECCV), Sep. 2018, pp. 596–611.

[15] S. Lee, S. Y. Jo, G. H. An, and S.-J. Kang, “Learning to Generate Multi-Exposure Stacks with Cycle Consistency for High Dynamic Range Imaging,” IEEE Trans. Multimed., vol. 23, pp. 2561–2574, Aug. 2020.

[16] K. Ma, Z. Duannu, H. Zhu, Y. Fang, and Z. Wang, “Deep guided learning for fast multi-exposure image fusion,” IEEE Trans. Image Process., vol. 29, pp. 2808–2819, Jun. 2020.

[17] K. R. Prabhabak, V. S. Srikar, and R. V. Babu, “DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs,” in Proc. ICCV, Oct. 2017, pp. 4724–4732.

[18] J. Yin, B.H. Chen, and Y.T. Peng, “Two Exposure Fusion Using Prior-Aware Generative Adversarial Network,” IEEE Trans. Multimed., 2021.

[19] 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, p. 178, 2017.

[20] 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.

[21] D. Memontides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, “Expendnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content,” in Comput. Graph. Forum, vol. 37, no. 2, pp. 37-49, May 2018.

[22] J.-L. Yin, B.-H. Chen, and Y.-T. Peng, “Two-exposure fusion using multi-exposure images,” IEEE Trans. Image Process., vol. 28, pp. 2049–2062, Apr. 2019.
