IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 8, 2021, accepted October 20, 2021, date of publication October 25, 2021, date of current version October 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122540

Deep Gradual Multi-Exposure Fusion via Recurrent Convolutional Network

JE-HO RYU, JONG-HAN KIM, AND JONG-OK KIM, (Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Jong-Ok Kim (jokim@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF), Ministry of Science and ICT (MSIT), South Korea, by the Korean Government, under Grant 2019R1A2C1005834; and in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2021-2018-0-01421.

ABSTRACT The performance of multi-exposure image fusion (MEF) has been recently improved with deep learning techniques but there are still a couple of problems to be overcome. In this paper, we propose a novel MEF network based on recurrent neural network (RNN). Multi-exposure images have different useful information depending on their exposure levels, and in order to fuse them complementarily, we first extract the local detail and global context features of input source images, and both features are separately combined. A weight map is learned from the local features for effectively fusing according to the importance of each source image. Adopting RNN as a backbone network enables gradual fusion, where more inputs result in further improvement of the fusion gradually. Also, information can be transferred to the deeper level of the network. Experimental results show that the proposed method achieves the reduction of fusion artifacts and improves detail restoration performance, compared to conventional methods.

INDEX TERMS Multi-exposure image fusion, recurrent convolutional network, dilated convolution filter, gradual fusion.

I. INTRODUCTION

Multi-exposure fusion (MEF) has been a popular approach for HDR image generation. It is a method to fuse a couple of low dynamic range images obtained by taking the same scene at different exposure levels. Due to the limited dynamic range of digital camera sensors, some regions of the scene may be under-exposed or over-exposed (even leading to saturation), depending on exposure time. Thus, multi-exposure images can be complementarily combined for generating a single HDR image, which contains the whole dynamic range of the scene. In general, over/under-exposure images lose detailed information, which leads to low contrast and quality. Therefore, MEF aims to fuse a high-quality image with better brightness and detailed information restoration.

Since MEF was proposed by Mertens *et al.* [1], various researches have been conducted in literature. The conventional MEF approach can be divided into spatial and transform domain based methods. The spatial domain based methods fuse multi-exposure images on the spatial domain. The weights for their fusion are calculated by analyzing MEF images from various perspectives such as contrast ratio, saturation [1], image block [2], and gradient [3]. In contrast, in the case of the transform domain approach, source images are first transformed into another domain and fusion is proceeded. A variety of the relevant methods have been proposed, including Wavelet [4], [16], multi-scale decomposition [6], [15], and sparse representation [7], [8]. However, the performances of both MEF approaches are fundamentally limited in that they mainly rely on hand-crafted features for image fusion. For further performance improvement, we need well-designed feature extraction and fusion rules, which are a challenging task.

Recently, convolutional neural networks (CNNs) have been popularly used for image fusion [9], [12], [23]. While this CNN-based fusion approach achieves better performance than non-deep learning, there are still some challenges to be overcome. First, a deep learning framework is used only in the limited part of MEF such as feature extraction, and conventional fusion strategies such as weighted sum are used identically. In addition, the image quality of the fused image is deteriorated by using features extracted only limited information (*e.g.,* Y channel) from the source image. As a result, detail restoration performance is degraded in the over/

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

under-exposure region. Finally, it was observed through experiments that artifacts such as local dark region often occur in fused images in particular on the regions whose brightness between multi-exposure images is significantly different.

In this paper, we propose a novel CNN-based MEF architecture, which is called Deep Gradual Multi-Exposure Fusion via Recurrent Convolutional Network (DGMEF-RNN). In general, deep learning based MEF methods go through the processes such as feature extraction, fusion, and reconstruction. In the proposed network, features are extracted in both global context and local detail using a dilated convolution filter. Fusion and reconstruction are done through RNN and a residual network, respectively. Unlike the conventional methods where auto-encoder is mainly adopted as a backbone, we propose a novel RNN-based fusion model. RNN builds a connection between the output and the next input of the network. As shown in Fig. 1, the stepwise fusion process of the proposed RNN-based network allows long-range dependency so that each source image information is transmitted to a deeper level in the network to generate a high-quality fusion image. The fusion module consists of two blocks: The global context fusion block naturally fuses global components such as color and style of source images. And the local detail fusion block preserves the detail components of source images by learning appropriate weights for fusion. The experimental result of the proposed method is richer in color and contains better illumination for all regions, thus more fully revealing the details of source images with higher saturation. The contributions of the paper are summarized as the followings:

(1) We propose a novel RNN-based MEF architecture, which sequentially transmits global context information to the entire network. As far as we know, this is the first work to implement the sequential fusion of multiple exposed images with RNN.

(2) The proposed method strengthens the detail feature fusion of source images through the learned weight map and effectively restores the local detail components of the fused image.

(3) From the experimental results, it could be confirmed that the proposed method reduces the local dark region artifact by global context fusion and RNN-based architecture.

The rest of the paper is organized as follows. Section II introduces related works on CNN-based multi-exposure image fusion. Section III describes our RNN-based multi-exposure image fusion architecture. Section IV verifies the effectiveness of our proposed MEF method visually and quantitatively. Finally, Section V provides the concluding statements.

II. RELATED WORKS

Deep neural networks have been recently applied to various image fusion problems. Liu *et al.* [9] studied convolutional sparse representation (CSR) for image fusion, where a fusion weight map is learned to distinguish the focus and unfocus regions of the source image. Liu *et al.* [10] proposed a medical image fusion method based on CNN which is used to generate a weighted map to represent the extent of pixel activity in the source image. They also introduced a local similarity-based strategy to adaptively adjust the fusion rules through decomposed coefficients. In the above methods, CNN is only adopted to generate a weighted map that incorporates pixel activity information, and the entire fusion process is still performed in a traditional way of multi-scale image pyramids.

Li and Wu [11] proposed DenseFuse for the fusion of infrared and visible images. Dense blocks are used in the encoder and it proposes a new fusion strategy to fuse feature maps. The feature maps of source images in the fusion layer are combined into two manually designed fusion strategies (additional and 1-norm). And it uses a non-referential metric (structural similarity index measurement and Euclidean distance) as a loss function for unsupervised learning. Li *et al.* [12] decomposed source images into base parts and detail content. Then, the base parts are fused by weighted-averaging. For detail content, deep learning networks are used to extract multi-layered features. These features are used to generate multiple candidates of fused detail content using *l1-norm* and weighted average strategies. Finally, the two parts are combined for reconstruction. Due to availability and effectiveness of generative adversarial network (GAN), Ma *et al.* proposed FusionGAN [13] to fuse infrared and visible images. The fused image generated by the generator is forced to restore more details existing in the visible image by applying the discriminator to distinguish differences between them. Kalantari and Ramamoorthi [14] proposed to obtain tone-mapped and ghost-free fused images from multi-exposure images through CNN. They collected a static set of low dynamic range (LDR) images, and then fused them into a high dynamic range (HDR) image using a simple triangle weighting scheme. In this way, fusion research is being conducted in various fields. Recently, a study was proposed by Xu *et al.* [25] to solve the fusion problem of several cases including multi-modal, multi-exposure, and multi-focus at once.

Deep learning was first introduced into the field of multi-exposure image fusion by DeepFuse [23]. It was designed as an encoder-decoder based image fusion architecture. DeepFuse uses the MEF-SSIM [33] metric as a loss function and trains the network through unsupervised learning. In DeepFuse [23], CNN is used only to Y channels for feature extraction and reconstruction, and the fusion rules of the chrominance channels are still designed manually. However, this manual fusion of chrominance channels may fail to restore color information accurately. Recently, MEF-NET was proposed by Ma *et al.* [24]. It trains a high resolution weight map from source images using a context aggregation network based on dilated convolution filter and a guided filter. Although it exhibits a very high performance, artifacts are often observed on a region whose brightness is significantly

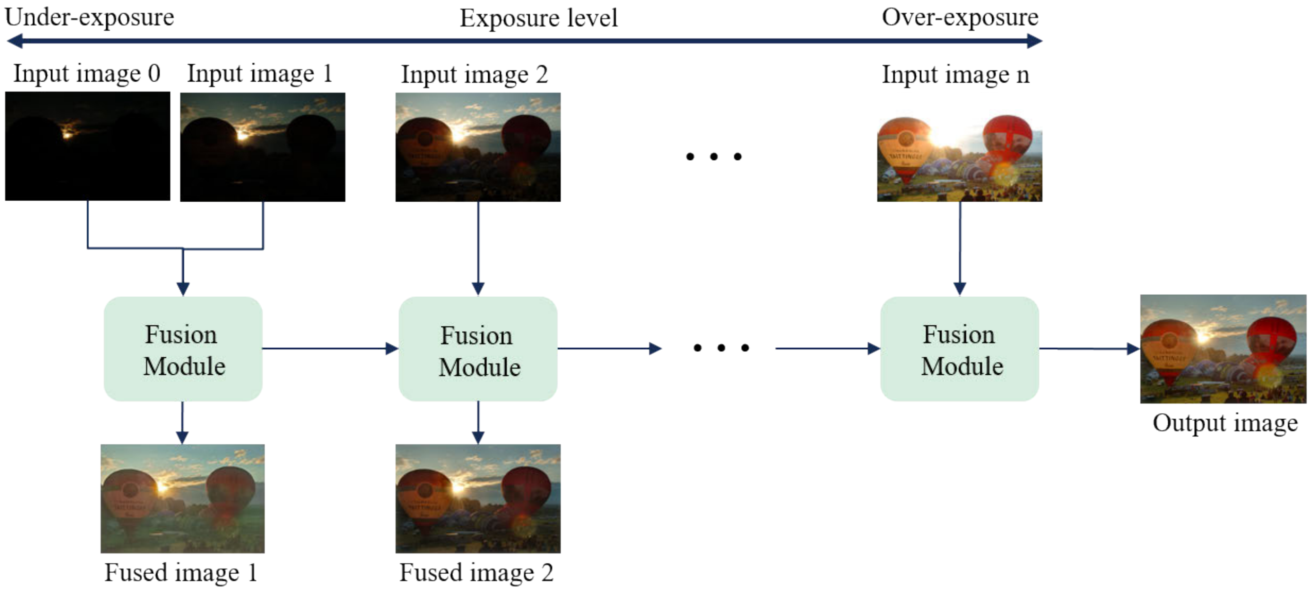


FIGURE 1. The concept of the proposed RNN-based progressive multi-exposure fusion.

distinct among source images. This performance degradation is highly improved in the proposed RNN-based MEF method as demonstrated in experimental results.

III. THE PROPOSED METHOD

This section describes the overall network architecture and fusion modules of the proposed method in detail.

A. NETWORK ARCHITECTURE

RNN has a characteristic that the connection between units has a recursive structure. It is distinguished by its hidden state (memory) as it takes information from prior inputs to influence the current input and output. This structure makes it possible to store a current state inside a neural network so that time-varying dynamic features can be modeled. In the proposed method, we design a multi-exposure fusion architecture to utilize the characteristics of RNN. The reason for using the RNN structure as a backbone network in the proposed method is that it is suitable for MEF, considering the characteristics of multi-exposure images. Multi-exposure images have different brightness for the same scene, and accordingly, each image contains different information (e.g., brightness, detail, and color). We observed that the behavior of the increasing brightness along multi-exposure images is similar to information changes over time in time series data. In addition, in order to generate a high-quality fused image, it is important to naturally fuse multi-exposure images. To this end, we propose an RNN-based step-by-step fusion structure. When multi-exposure source images with different brightness are fed as inputs, the fusion process is made step by step by leveraging the features of the fused result in the previous step. As a result, the current output can be used as the next input, and then it is continuously fused with the next new input, consequently leading to natural fusion and gradual enhancement. Fig. 2. illustrates the entire architecture of the proposed method. Four multi-exposure source images were used as the input of the RNN network. A set of four given source images is denoted by I_n $(n=0,\ 1,\ 2,\ 3)$, and they are arranged so that the brightness of the image increases sequentially as shown in Fig. 2.

TABLE 1. Specification of CAN in DGMEF-RNN for DEB.

	1	2	3	4	5	6
Channel	32	32	64	64	64	64
Conv Filter Size	3x3	3x3	3x3	3x3	3x3	3x3
Dilation	1	2	4	8	16	1
Receptive field	3x3	7x7	15x15	31x31	63x63	65x65

The proposed RNN fusion network generates a fused image through three processes; Initial feature extraction (Dilated encoding block: DEB), fusion module, and reconstruction. In DEB, the global context feature and local detail feature of the source image are extracted. The extracted features are fused through both the global context fusion block (GCFB) and the local detail fusion block (LDFB), and then, the fused feature is transferred to the next fusion module. Finally, the fused features generated in each fusion module are concatenated and reconstructed. Each process is described later in order.

B. DILATED ENCODING BLOCK

In order to generate a high-quality multi-exposure fusion image, it is important to acquire and restore detail information from each source image. In addition, a process of naturally fusing context information such as color and style

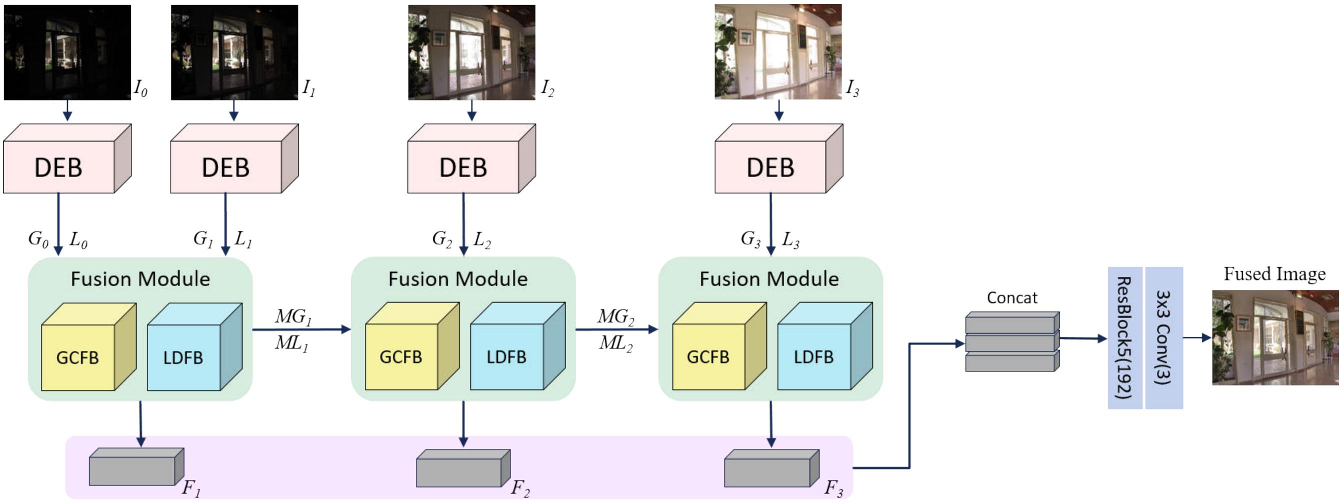


FIGURE 2. The overall network structure of the proposed DGMEF-RNN.

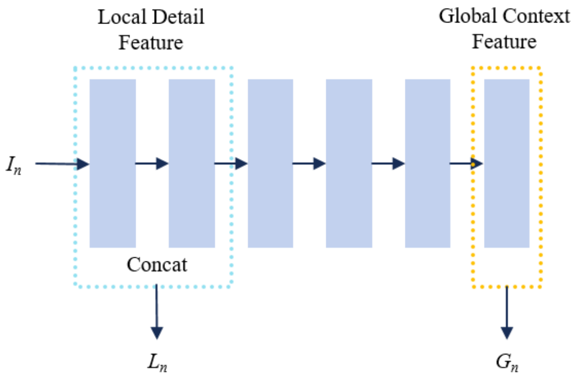


FIGURE 3. Dilated encoding block (DEB) structure.

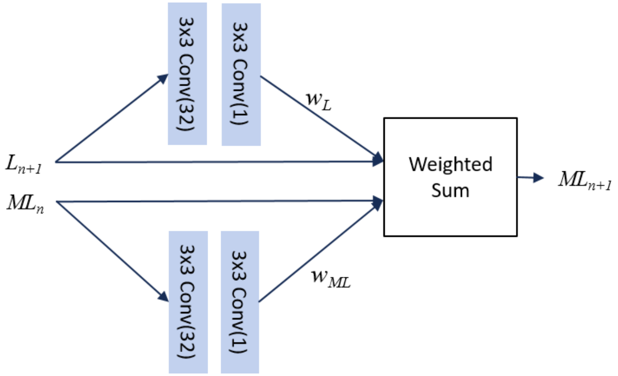


FIGURE 4. Local detail fusion block (LDFB) structure.

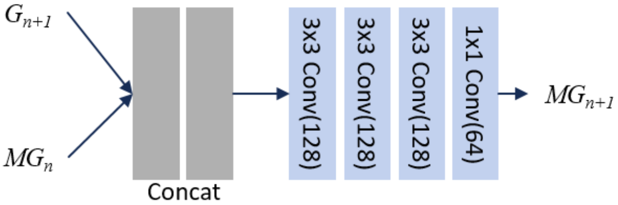


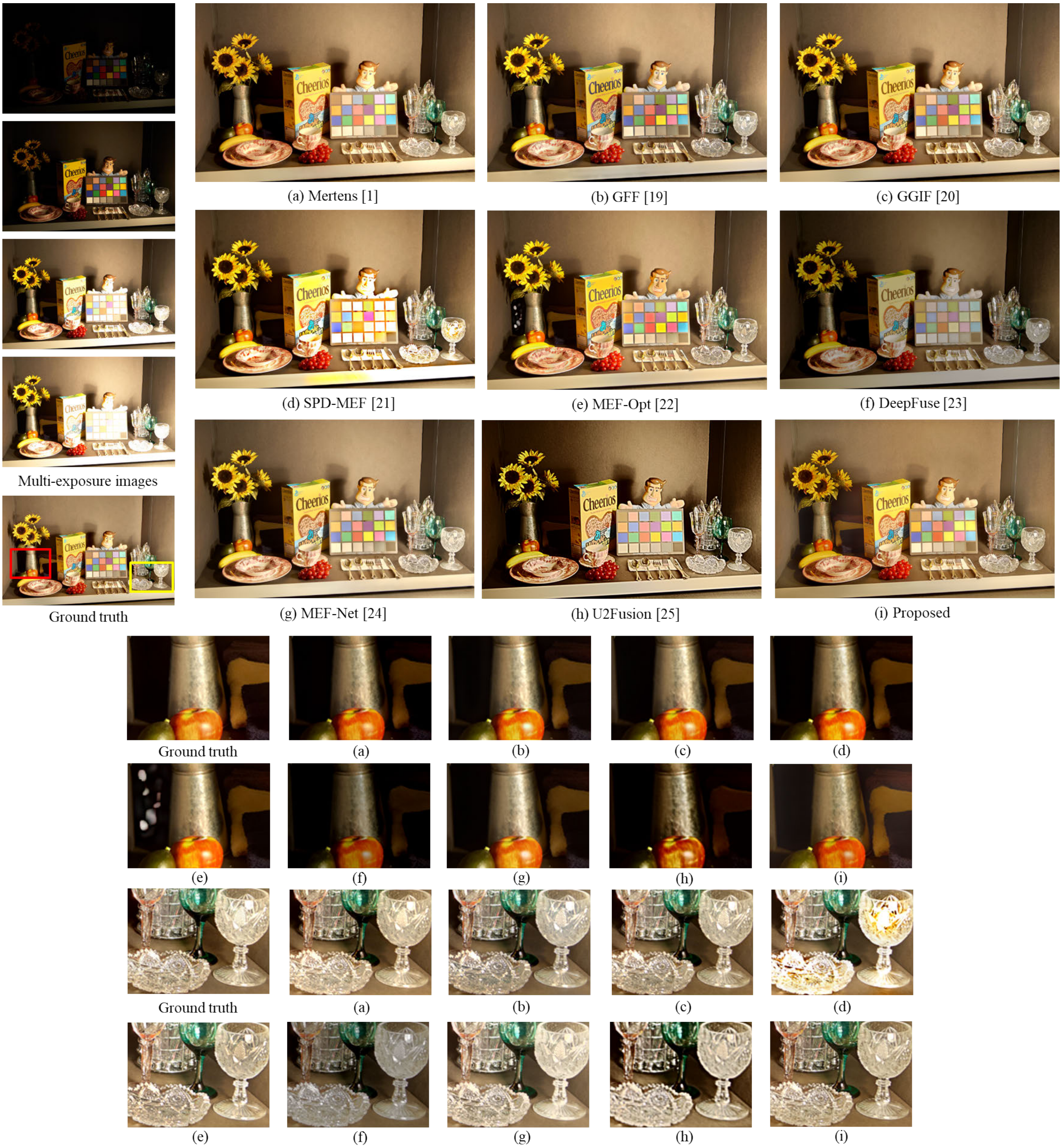
FIGURE 5. Global context fusion block (GCFB) structure.

of source images is also required. Therefore, in this paper, dilated encoding block (DEB) is adopted to extract the local detail feature and the global context feature of the source image. As shown in Fig. 3, DEB uses the context aggregation network (CAN) proposed by Yu and Koltun [37]. The CAN structure to utilize the characteristics of the dilated convolution filter gradually expands the receptive field of the convolution filter. In this paper, we construct a convolution filter for each step, as shown in Table 1. We employ adaptive normalization and leaky rectified linear unit (LReLu) right after convolution. The local detail feature (L_n) of the source image was generated by concatenation of the features in steps 1 and 2 extracted from the small receptive field. The feature extracted from the relatively wide receptive field in the last step is defined as a global context feature (G_n).

C. FUSION MODULE

The proposed fusion module consists of the local detail fusion block (LDFB) and the global context fusion block (GCFB). LDFB and GCFB fuse local detail features (L_n) and global context features (G_n) extracted from DEB, respectively. And the fused global context and local detail features are sequentially transferred to the next RNN fusion module.

The LDFB is structured as shown in Fig. 4. In the conventional method, it was difficult to reconstruct the detail of the fused image in too saturation and dark areas. In addition, the restoration performance of texture details can be further improved. In order to improve the detail reconstruction performance of the fused image, we generate a weight map to reinforce the local details of source images, and the local



Multi-exposure images

Ground truth

(a) Mertens [1]

(b) GFF [19]

(c) GGIF [20]

(d) SPD-MEF [21]

(e) MEF-Opt [22]

(f) DeepFuse [23]

(g) MEF-Net [24]

(h) U2Fusion [25]

(i) Proposed

Ground truth (a) (b) (c) (d)

(e) (f) (g) (h) (i)

Ground truth (a) (b) (c) (d)

(e) (f) (g) (h) (i)

FIGURE 6. Qualitative comparison on scene 1. Images on the far upper left column are source images and the corresponding ground truth. Fused results of the eight existing methods (Mertens [1], GFF [19], GGIF [20], SPD-MEF [21], MEF-Opt [22], DeepFuse [23], MEF-Net [24] and U2Fusion [25]) and the proposed DGMEF-RNN are shown on the upper right. The bottom rows are the highlighted regions corresponding to the marked boxes in the ground truth.

detail features are fused by a weighted sum. As shown in Fig. 4, each weight map is generated using the local detail feature (L_{n+1}) extracted from the newly source image at each stage of the RNN and the ML_n transferred from the fusion module at the previous stage.

The GCFB for global context fusion is designed as shown in Fig. 5. Like LDFB, GCFB concatenates the global context feature (G_{n+1}) extracted from a newly arrived source image at each stage of the RNN and MG_n transferred from the previous stage of the fusion module and then, they passed through the

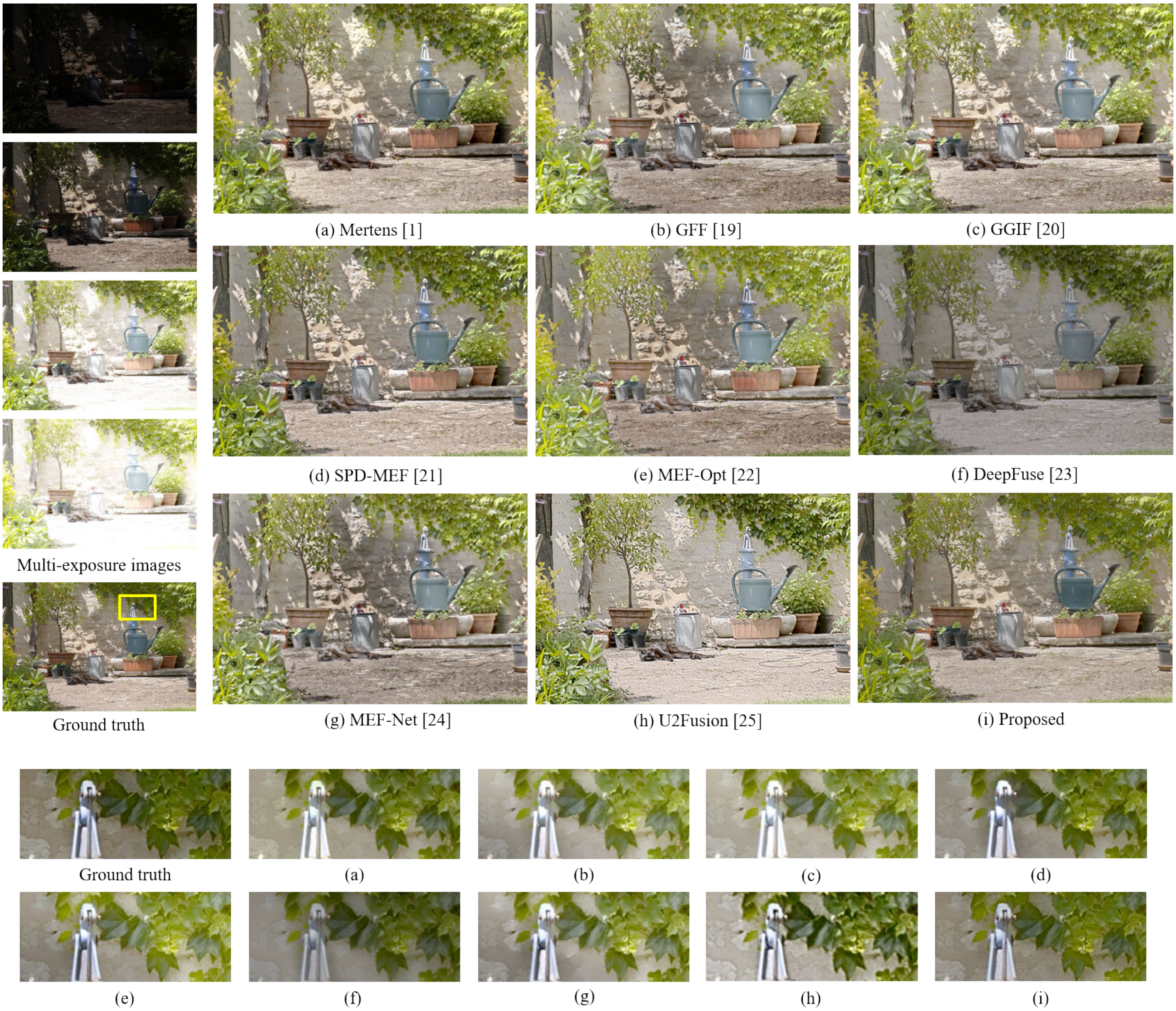


FIGURE 7. **Qualitative comparison on scene 2. For more information, refer to the caption in Fig. 6.**

four convolution filters to generate a fused global context feature.

D. RECONSTRUCTION

The reconstruction network for the restoration of a fusion image consists of five ResBlocks and one convolution filter. It is fed with the concatenation of the output features (F_n) of the RNN fusion module as shown in Fig. 2. Note that F_n is the fusion of MG_n and ML_n.

E. LOSS FUNCTION

The loss function L used in the proposed method is given by

$$L = L_{ssim} + \lambda L_{mef} + L_{L1} \tag{1}$$

which is a combination of *l1* loss L_{L1}, structural similarity index (SSIM) [35] loss L_{ssim} and MEF-SSIM [34] loss L_{mef} with a weight λ. L_{L1}, L_{ssim}, and L_{mef} are losses between the fused image and groundtruth. As described above, the above loss equation is used to reinforce the local detail information of the fusion image and for natural fusion of the global context information.

IV. EXPERIMENTAL RESULTS

In this section, we first compare the proposed DGMEF-RNN with conventional and recent MEF methods in qualitative and quantitative ways. We then conduct a series of ablation experiments diversely to demonstrate the usefulness of DGMEF-RNN.

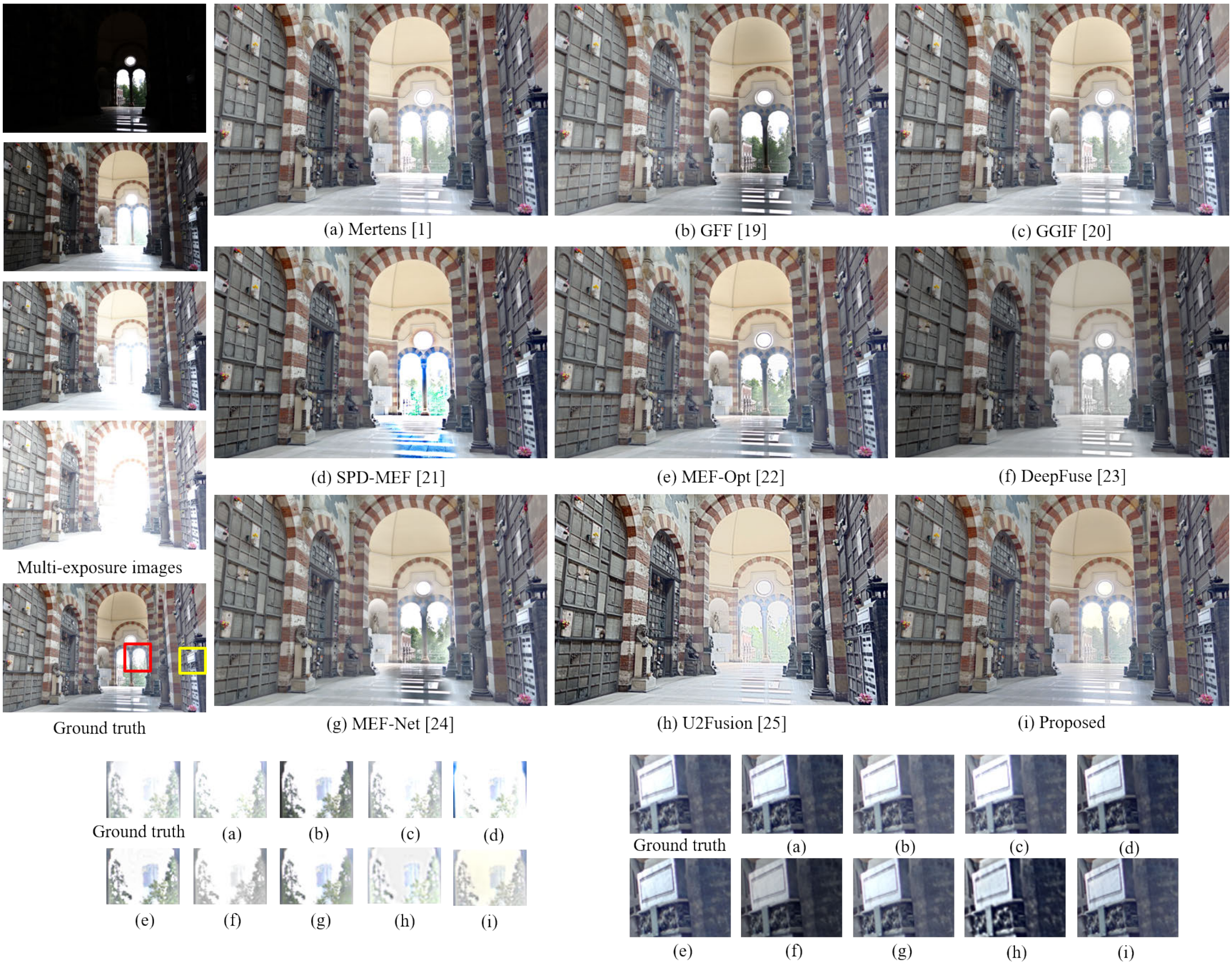


FIGURE 8. **Qualitative comparison on scene 3. For more information, refer to the caption in Fig. 6.**

A. TRAINING

1) DATASET

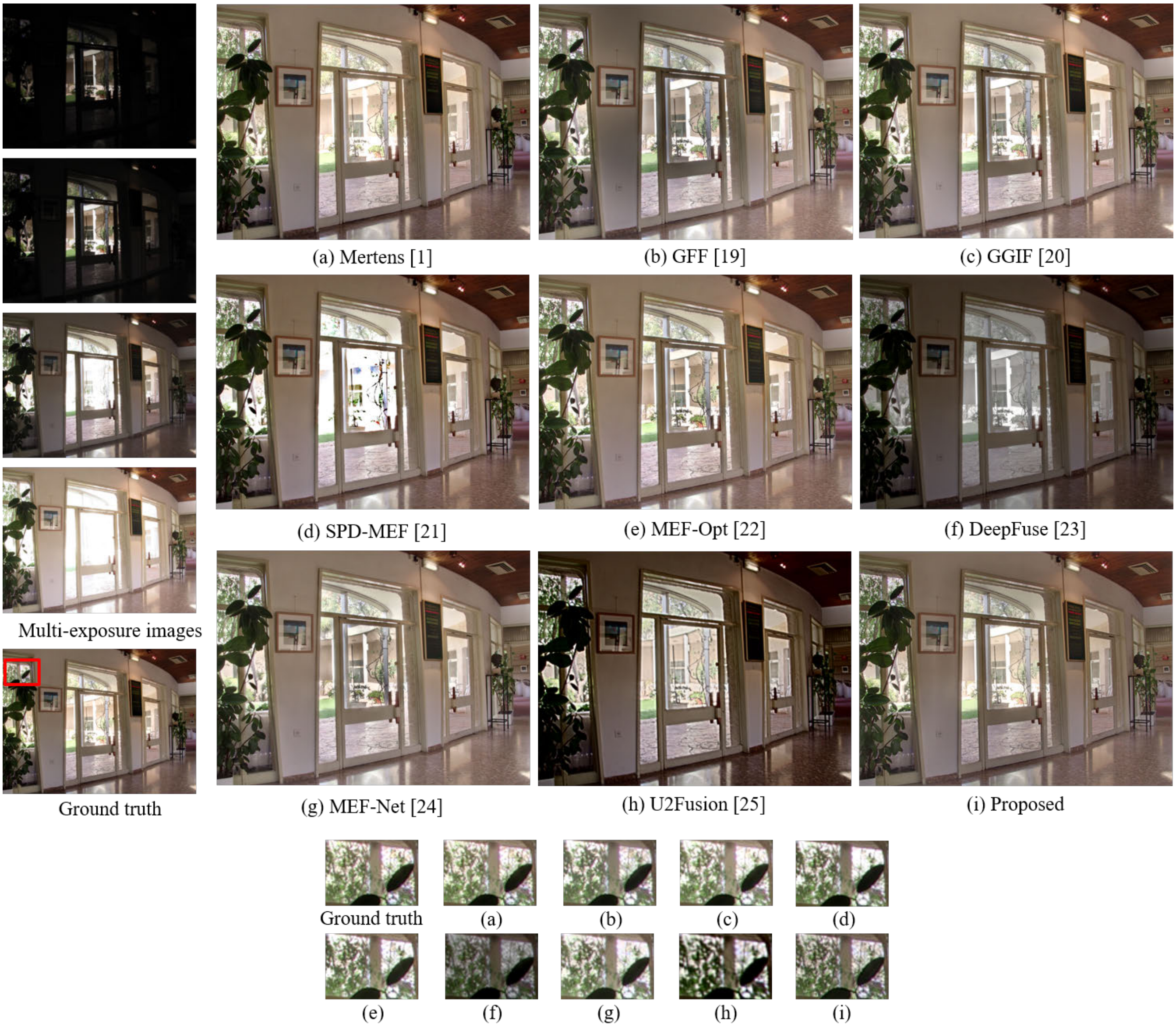
To validate the performance of the proposed DGMEF-RNN, we perform qualitative and quantitative experiments on the publicly available dataset provided by [53] and [54], with multi-exposure sequences including indoor and outdoor, human-life, day and night scenes and the corresponding high-quality reference images (ground truth). We use a sequence of multi-exposure images under different exposure settings which have been accurately aligned. In each scene, dataset was constructed by selecting four multi-exposure images with different brightness, and total of 270 scenes dataset was obtained. We train DGMEF-RNN on 227 scenes and use the remaining 43 scenes for testing. During training, the resolution of the training dataset was reduced to 1/5∼1/7 for the reduction of the GPU memory cost while maintaining the aspect ratio. The resolution of the image is kept from 500 to 700 pixels at least. 227 scenes of multi-exposure images, and corresponding ground truth images are cropped into 10000+ patches for the training data. All patches are of size 160 × 160.

2) IMPLEMENTATION

To train our network, we used a sequence of four multi-exposure source images and the corresponding groundtruth images with a batch size of 8. It is implemented using the PyTorch framework on a PC with 2 NVIDIA RTX 2080ti GPUs. For loss optimization, we adopted the Adam optimizer with a learning rate of 10^{-4} which is divided by 10 for every 500 iterations. Finally, DGMEF-RNN is evaluated at a full resolution during testing

B. QUALITATVE COMPARISON

The proposed DGMEF-RNN is compared with the eight state-of-the-art methods, including Metens09 [1], GFF [19], GGIF [20], SPD-MEF [21], MEF-Opt [22], DeepFuse [23], MEF-Net [24] and U2Fusion [25]. Mertens09 [1] is one of the representative methods for MEF. GFF [19] is a guided filter-based fusion method, and GGIF [20] extends it to



Multi-exposure images

Ground truth

(a) Mertens [1]

(b) GFF [19]

(c) GGIF [20]

(d) SPD-MEF [21]

(e) MEF-Opt [22]

(f) DeepFuse [23]

(g) MEF-Net [24]

(h) U2Fusion [25]

(i) Proposed

FIGURE 9. Qualitative comparison on scene 4. For more information, refer to the caption in Fig. 6.

the gradient domain. SPD-MEF [21], MEF-Opt [22], and MEF-Net [24] are inspired by the MEF-SSIM proposed by Ma *et al.* DeepFuse [23] is the first work to propose a deep learning-based MEF and has been used as a reference in so many papers. U2Fusion [25] is for solving three fusion tasks at once: multi-modal, multi-exposure, and multi-focus. For all the comparison methods, we used the code and setting provided by the original authors. However, in case of DeepFuse [23], it accepts only two images of under-over exposure, and for a fair comparison, its input is expanded to accept four source images. In addition, the MEF-Net [24] is also trained using four source images. U2Fusion [25] conducted an experiment using the author-provided code and model. For fair comparison, four multi-exposure images were used, and as described in the U2Fusion [25] paper, two multi-exposure images were fused step by step.

Subjective comparisons were made with 43 image scenes, some of which are shown in Figs. 6-9. We analyzed the visual quality factors of the fused image such as brightness, color and detail restoration. Through these analyses, we can confirm the usefulness of LDFB and GCFB modules in the proposed network, and the merit of the RNN structure.

In terms of subjective image quality, it can be confirmed from the fused images of Figs. 6-9, that the proposed method achieves high performances by restoring sufficient brightness and color. In particular, for DeepFuse [23], its fused image is generally grayish and suffers from insufficient color saturation. In U2Fusion [25], the color saturation of the fusion image is not sufficient, and as shown in Fig. 6, the restoration performance of dark areas is deteriorated. Also, for SPD-MEF [21], color artifact occurs in several experimental images. But the proposed method achieves

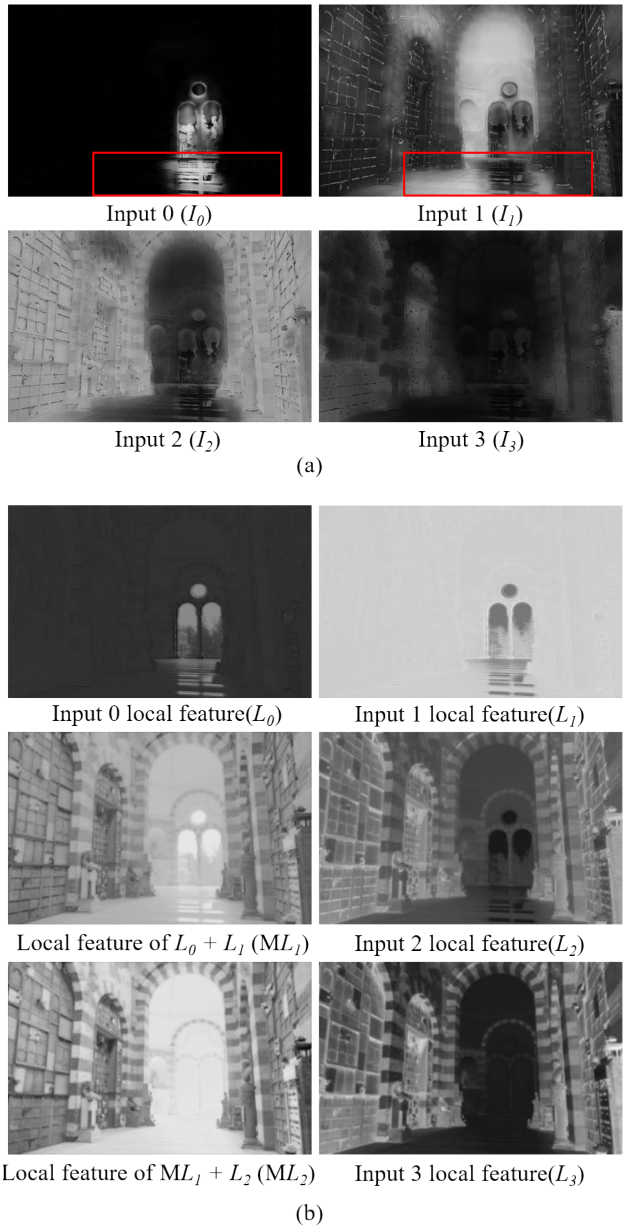

Input 0 (I_0)
Input 1 (I_1)
Input 2 (I_2)
Input 3 (I_3)
(a)
Input 0 local feature(L_0)
Input 1 local feature(L_1)
Local feature of L_0 + L_1 (ML_1)
Input 2 local feature(L_2)
Local feature of ML_1 + L_2 (ML_2)
Input 3 local feature(L_3)
(b)

FIGURE 10. Illustration of the trained weight map of MEF-Net and the proposed method. (a) MEF-Net weight map (b) the proposed LDFB weight map.

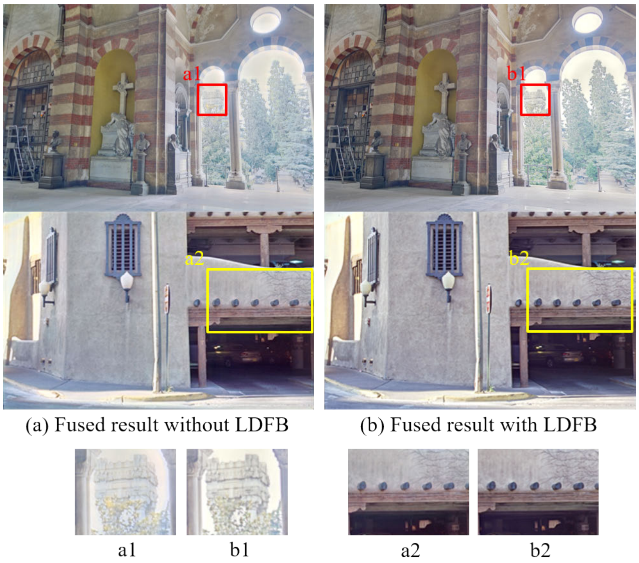


FIGURE 11. The fused results (a) without LDFB and (b) with LDFB.

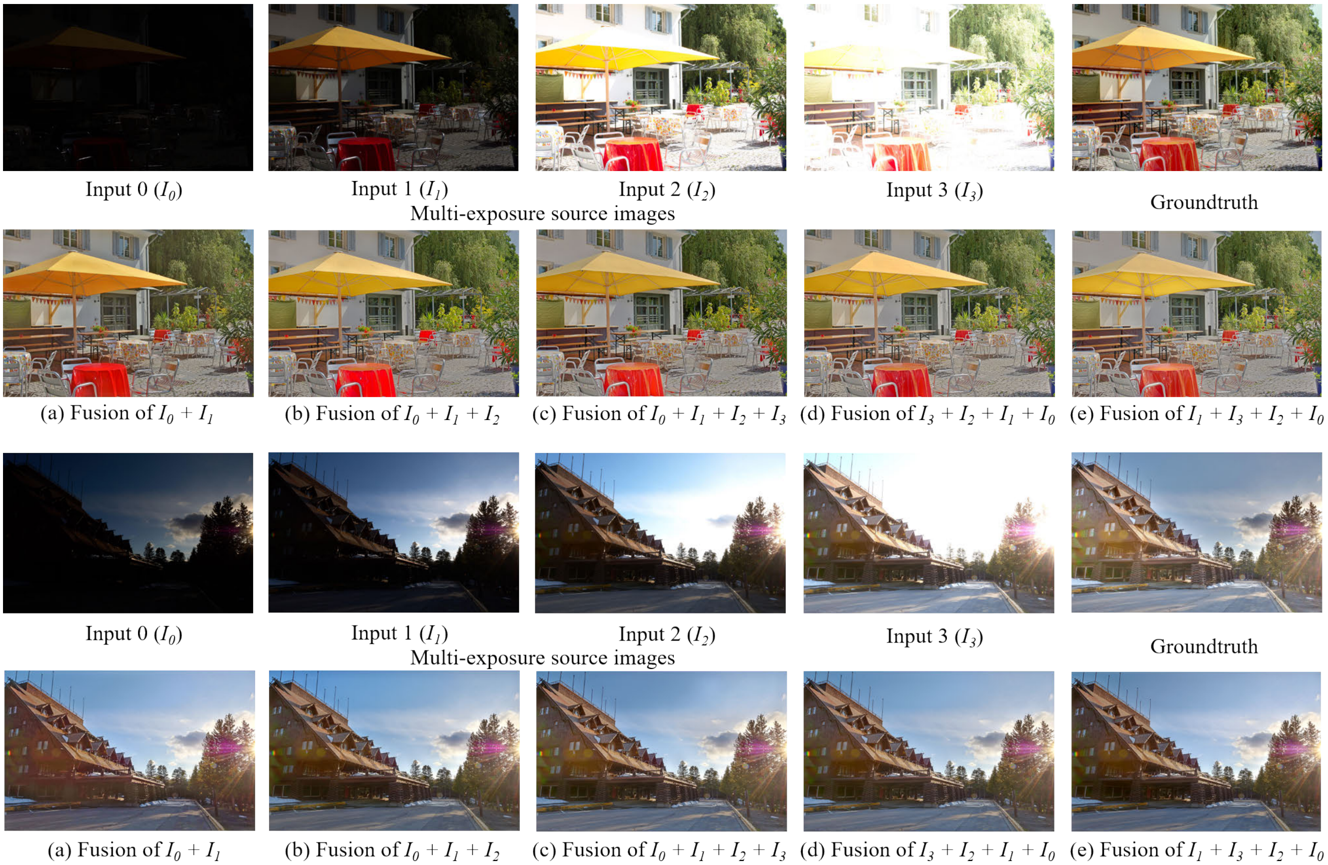
a superior performance in terms of color distortion. Furthermore, as shown in Fig. 8, the proposed method naturally restores the original luminance without local dark region artifact observed from GFF [19], MEF-Opt [22], and MEF-Net [24], leading to the improvement of texture detail. In GGIF [20], the detail reconstruction ability is good and there are few artifacts, but the global contrast decreases as shown in Fig. 7.

As shown in Figs. 8 and 9, GFF [19], MEF-Opt [22], and MEF-NET [24] suffer from local dark region artifact. Consequently, the fused image is deteriorated by artificial and uncomfortable appearances. This is caused by inaccurate weight determination for each input image when the brightness of the local region among source multi-exposure images is significantly different. For example, looking at Fig. 10, (a) and (b) are the weight maps trained by MEF-Net [24] and the proposed method for the experimental image in Fig. 8, respectively. MEF-Net [24] surely determines the weight of each source image according to its important information, but the weight is heavily placed on a specific source image for some local regions. As a result, it is not naturally fused in the boundary region. For this reason, local dark region or halo artifact occurs when multi-exposure images with a large brightness difference are fused. In the proposed method, this phenomenon is highly minimized by the global context fusion block GCFB and RNN-based fusion architecture.

Figs. 6-9 show the local detail restoration performance of the proposed method. It can be seen that the texture detail is preserved in the glass of Fig. 6, the walls and leaves of Fig. 7 and the windows of Fig. 9. The detail restoration performance of SPD-MEF [21] is also good, but it suffers from color distortion and halo artifacts. In addition, the proposed method has excellent restoration performances in very bright and dark regions such as the windows of Fig. 9 and the vases of Fig. 6.

C. QUANTITATIVE COMPARISON

We conducted quantitative evaluation using peak signal-to-noise ratio (PSNR), SSIM, and MEF-SSIM metrics, and the results are shown in Table 2. Red indicates the best performing MEF method, and blue represents the second one. We can see that the proposed method shows the highest performance on SSIM and PSNR scores. The MEF-SSIM score is slightly inferior to that of GGIF [20] and MEF-Net [24], but visual artifacts occur in both methods as described above. Through experimental results, it can be confirmed that the proposed method generally achieves a superior performance compared to the conventional methods.



Input 0 (I_0) Input 1 (I_1) Input 2 (I_2) Input 3 (I_3)

Multi-exposure source images

Groundtruth

(a) Fusion of $I_0 + I_1$ (b) Fusion of $I_0 + I_1 + I_2$ (c) Fusion of $I_0 + I_1 + I_2 + I_3$ (d) Fusion of $I_3 + I_2 + I_1 + I_0$ (e) Fusion of $I_1 + I_3 + I_2 + I_0$

Input 0 (I_0) Input 1 (I_1) Input 2 (I_2) Input 3 (I_3)

Multi-exposure source images

Groundtruth

(a) Fusion of $I_0 + I_1$ (b) Fusion of $I_0 + I_1 + I_2$ (c) Fusion of $I_0 + I_1 + I_2 + I_3$ (d) Fusion of $I_3 + I_2 + I_1 + I_0$ (e) Fusion of $I_1 + I_3 + I_2 + I_0$

FIGURE 12. The fused results obtained by changing the number of source images and the order of the source images in the proposed RNN network.

TABLE 2. Quantitative comparison. Red indicates the best result, and blue indicates the second best.

	PSNR	SSIM	MEF-SSIM
Mertens09	21.3302	0.8968	0.9289
GGF	19.1921	0.9032	0.9288
GGIF	22.5614	0.9255	0.9335
SPD-MEF	21.2736	0.9047	0.9242
MEF-opt	21.4410	0.9201	0.9287
DeepFuse	17.2099	0.7607	0.8617
MEF-Net	21.5181	0.9134	0.9350
U2Fusion	17.8389	0.7798	0.8612
Proposed	23.5052	0.9306	0.9311

Table 3 is the result of the running time comparison. We compared the running times of DeepFuse [23], MEF-Net [24], U2Fusion [25], and the proposed method. The experiments were conducted in an i7-8700 CPU and NVIDIA TITAN RTX pc environment. Running time was measured using 42 test images of various resolutions. The contents of Table 3 measure the average running time to process 1 million pixels. On average, the DeepFuse [23] method is the fastest. However, the DeepFuse [23] network structure is relatively simple compared to the other methods, and thus the fused image quality is deteriorated. The proposed method takes relatively less running time than the existing methods.

TABLE 3. Running time comparison.

	DeepFuse	MEF-Net	U2Fusion	Proposed
Running time (1M px/s)	0.03376	0.14738	3.24747	0.06992

Considering the fused image quality and running time, we believe that the proposed method surely has merits over the existing methods.

D. ABLATION STUDY

A number of ablation studies are conducted to verify the importance of each component in the proposed deep network. The fusion module of the proposed method is composed of GCFB and LDFB, and LDFB is a module to strengthen local details in the fused image. To verify the effect of LDFB, an experiment is performed without it. The results of this experiment are shown in Fig. 11 (a). Compared with the proposed method DGMEF-RNN in Fig. 11 (b), the fused image is blurred on the whole and the reconstruction performance is degraded in texture details. Through this experiment, it can be conformed that LDFB is effective in enhancing local details. The quantitative comparison is shown in Table 4. In addition, when the experimental results in Fig. 11 (a) and (b) are compared, local dark region artifact does not occur yet even if

TABLE 4. Quantitative comparison results on ablation study note that the summation indicates the order of fusion in the proposed RNN.

	PSNR	SSIM	MEF-SSIM
Fusion of $I_0 + I_1$	21.2569	0.8934	0.9042
Fusion of $I_0 + I_1 + I_2$	22.6902	0.9205	0.9244
Fusion of $I_3 + I_2 + I_1 + I_0$	22.9487	0.9241	0.9324
Fusion of $I_1 + I_3 + I_2 + I_0$	23.4515	0.9232	0.9299
Fused result without LDFB	23.3695	0.8976	0.9123
Fusion of $I_0 + I_1 + I_2 + I_3$	23.5052	0.9306	0.9311

LDFB is excluded. That is, we can see that the aforementioned gradual fusion using GCFB and RNN-based proposed network architecture is effective for natural image fusion.

Next, some experiments were conducted by changing the number of source images and the order of the source images in the RNN network input (note that the input images are originally fed with the network in the order of increasing exposure level). Fig. 12 illustrates the results of the ablation study. The proposed method basically accepts four source images. We evaluate the performance of the proposed network when the number of source images is two and three. As the number of source images decreases, the number of fusion modules configured in RNN should be reduced accordingly. Multi-exposure images have different information depending on the exposure level. As shown in Fig. 12 (a), the source images I_0 and I_1 (which are excessive dark) have little information, so their contributions to the image fusion is marginal. Even though the source image information is insufficient, the entire structure of the fused image is correctly restored in the proposed method, but the color and detail restoration performances can be further improved. Fig. 12 (b) and (c) are the fusion result by adding the source image one by one. As confirmed in Fig. 12 (a), (b) and (c), the color and detail restoration performances are gradually improved as the source image is added. This can be confirmed that the proposed network progressively enhances the fusion according to the number of input images. Fig. 12 (d) and (e) are the experimental results by changing the order of the input source images. In (d), the inputs are changed in the reverse order from (4) to (1). In (e), the order is taken randomly (I_1-I_3-I_2-I_0). From the results (c), (d) and (e), it can be seen that there is no significant difference in color and detail restoration performances of the fused image. It can be thought that the DEB, which is used for feature extraction in the proposed method, accurately extracts important information of each source image and transfers it to the deep level of the RNN-based network. Through such experiments, the RNN-based proposed network can be easily extended with inputs and the fusion can be improved accordingly.

V. CONCLUSION

We propose a novel multi-exposure image fusion method based on RNN so called DGMEF-RNN. The goal of DGMEF-RNN is to transfer source image information to the entire network to maintain long-range dependency, and to generate a fused image with artifact reduction. Moreover, we attempted to accomplish the good performance of detail restoration in the fused image. For this, we designed both GCFB and LDFB modules and demonstrated their superior performances through experimental results when compared with the conventional MEF methods. Our proposed network is trained with four multi-exposure images. Through comparisons with 7 state-of-the-art MEF methods, it was confirmed that the proposed method outperforms from both qualitative and quantitative perspectives.

REFERENCES

[1] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer Graphics Forum*, vol. 28, no. 1. Oxford, U.K.: Blackwell, 2009.

[2] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Inf. Fusion*, vol. 2, no. 3, pp. 169–176, Sep. 2001.

[3] 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.

[4] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.

[5] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, Apr. 2007.

[6] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, Jul. 2009.

[7] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[8] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, Sep. 2015.

[9] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36 pp. 191–207, Jul. 2017.

[10] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–7.

[11] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2018.

[12] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 2705–2710.

[13] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

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

[15] H. M. Hu, J. Wu, and B. Li, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2706–2719, Dec. 2017.

[16] H. Pang, M. Zhu, and L. Guo, "Multifocus color image fusion using quaternion wavelet transform," in *Proc. 5th Int. Congr. Image Signal Process.*, Chongqing, China, Oct. 2012, pp. 543–546.

[17] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Burlington, MA, USA: Morgan Kaufmann, 2010.

[18] R. Lai, Y. Li, and J. Guan, "Multi-scale visual attention deep convolutional neural network for multi-focus image fusion," *IEEE Access*, vol. 7, pp. 114385–114399, 2019.

[19] 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.

[20] F. Kou, Z. Li, C. Wen, and W. Chen, "Multi-scale exposure fusion via gradient domain guided image filtering," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, Jul. 2017, pp. 1105–1110.

[21] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2519–2532, May 2017.

[22] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index," *IEEE Trans. Comput. Imag.*, vol. 4, no. 1, pp. 60–72, Mar. 2018.

[23] 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. (ICCV)*, Oct. 2017, vol. 1, no. 2, pp. 4714–4722.

[24] K. Ma, Z. Duanmu, 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, 2020.

[25] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 28, 2020, doi: 10.1109/TPAMI.2020.3012548.

[26] J. Shen, Y. Zhao, S. Yan, and X. Li, "Exposure fusion using boosting Laplacian pyramid," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1579–1590, Sep. 2014.

[27] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 341–357, Jan. 2012.

[28] K. Ma, T. Zhao, K. Zeng, and Z. Wang, "Objective quality assessment for color-to-gray image conversion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Dec. 2015.

[29] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. C-31, no. 4, pp. 532–540, Apr. 1983.

[30] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Proc. 4th Int. Conf. Comput. Vis.*, Berlin, Germany, 1993, pp. 173–182.

[31] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process.*, vol. 29, pp. 3845–3858, 2020.

[32] H. Yan and Z. Li, "Infrared and visual image fusion based on multi-scale feature decomposition," *Optik*, vol. 203, Feb. 2020, Art. no. 163900.

[33] S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography," in *Proc. EUROGRAPHICS*, 2009, pp. 1–4.

[34] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[35] M. Hassan and C. Bhagvati, "Structural similarity measure for color images," *Int. J. Comput. Appl.*, vol. 43, no. 14, pp. 7–12, 2012.

[36] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jun. 2017.

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–13.

[38] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[39] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Inf. Fusion*, vol. 74, pp. 111–131, Oct. 2021.

[40] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Longbeach, CA, USA, Jun. 2019, pp. 3897–3906.

[41] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4107–4116.

[42] H. Xu, J. Ma, and X. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.

[43] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[44] J. Liang, J. Wang, Y. Quan, T. Chen, J. Liu, H. Ling, and Y. Xu, "Recurrent exposure generation for low-light face detection," 2020, *arXiv:2007.10963*. [Online]. Available: http://arxiv.org/abs/2007.10963

[45] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3106–3115.

[46] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 714–722.

[47] Y. Kinoshita and H. Kiya, "Scene segmentation-based luminance adjustment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4101–4116, Aug. 2019.

[48] J.-W. Kim, J.-H. Ryu, and J.-O. Kim, "Deep gradual flash fusion for low-light enhancement," *J. Vis. Commun. Image Represent.*, vol. 72, Oct. 2020, Art. no. 102903.

[49] 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.

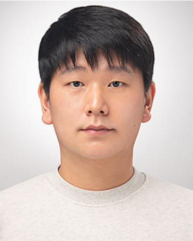
[50] H. Xu, F. Fan, H. Zhang, Z. Le, and J. Huang, "A deep model for multi-focus image fusion based on gradients and connected regions," *IEEE Access*, vol. 8, pp. 26316–26327, 2020.

[51] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12797–12804.

[52] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12484–12491.

[53] *DATASET*. Accessed: Aug. 2020. [Online]. Available: https://github.com/csjcai/SICE

[54] *DATASET*. Accessed: Aug. 2020. [Online]. Available: http://ivc.uwaterloo.ca/database/MEF.html



JE-HO RYU received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, where he is currently pursuing the integrated M.S. and Ph.D. degrees in electrical engineering. His research interests include machine learning and deep learning in particular with applications on computer vision tasks, such as visual recognition and image enhancement tasks, including high dynamic range imaging, image fusion, and super resolution.

JONG-HAN KIM received the B.S. degree in electronic engineering from Kwangwoon University, Seoul, South Korea, in 2020. He is currently pursuing the M.S. degree with Korea University, Seoul. His current research interests include deep learning-based various image processing and computer vision algorithms.

JONG-OK KIM (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Korea University, Seoul, Korea, in 1994 and 2000, respectively, and the Ph.D. degree in information networking from Osaka University, Osaka, Japan, in 2006. From 1995 to 1998, he served as an Officer for Korea Air Force. From 2000 to 2003, he was with the SK Telecom Research and Development Center and Mcubeworks Inc., South Korea, where he was involved in research and development on mobile multimedia systems. From 2006 to 2009, he was a Researcher with the Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan. He joined Korea University, Seoul, South Korea, in 2009, where he is currently a Professor. His current research interests include image processing, computer vision, and intelligent media systems. He was a recipient of the Japanese Government Scholarship, from 2003 to 2006.