GALFusion: Multi-Exposure Image Fusion via a Global–Local Aggregation Learning Network

Lei, J., Li, J., Liu, J., Zhou, S., Zhang, Q., & Kasabov, N. (2023). GALFusion: Multi-Exposure Image Fusion via a Global–Local Aggregation Learning Network. *IEEE Transactions on Instrumentation and Measurement*, *72*, 16. [5011915]. https://doi.org/10.1109/TIM.2023.3267525

Link to publication record in Ulster University Research Portal

Published in:
IEEE Transactions on Instrumentation and Measurement

Publication Status:
Published (in print/issue): 21/04/2023

DOI:
https://doi.org/10.1109/TIM.2023.3267525

Document Version
Author Accepted version

GALFusion: Multi-exposure Image Fusion via a Global-local Aggregation Learning Network

Jia Lei, Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, *Member, IEEE*
and Nikola K. Kasabov, *Life Fellow, IEEE*

***Abstract*—The goal of multi-exposure image fusion is to generate synthetic results with abundant details and balanced exposure from low dynamic range(LDR) images. The existing multi-exposure fusion methods often use convolution operations to extract features. However, these methods only consider the pixel values in local view field and ignore the long-range dependencies between pixels. To solve the aforementioned problem, we propose a global-local aggregation network for fusing extreme exposure images in an unsupervised way. Firstly, we design a collaborative aggregation module, composed of two sub-modules covering a non-local attention inference module and a local adaptive learning module, to mine the relevant features from source images. So that we successfully formulate a feature extraction mechanism with aggregating global and local information. Secondly, we provide a special fusion module to reconstruct fused images, which effectively avoids artifacts and suppresses information decay. Moreover, we further fine-tune the fusion results by a recursive refinement module to capture more textural details from source images. The results of both comparative and ablation analyses on two datasets demonstrate that our work is superior to ten existing state-of-the-art fusion methods.**

***Index Terms*—Image fusion, multi-exposure image, non-local attention**

I. Introduction

Natural scenes are rich in light and shadow information, which presents colorful images to human eyes. However, the well-exposure image is difficult to accurately record by digital cameras. A single image often has the phenomenon of being excessively bright or dark in some regions. The reason for these phenomena is that the dynamic range in realistic scenes is much higher than the output dynamic range of imaging or display devices. To deal with this limitations, researchers generally conduct highly dynamic imaging equipments [1] or common imaging equipments [2]. The highly dynamic imaging equipment can directly obtain high dynamic range (HDR) images of shooting scenes under professional hardware conditions. The common imaging equipment relies on multiple LDR image sequences under various exposure settings, and then uses fusion algorithms to synthesize an HDR image with clear details and faithful color. Considering the cost of actual execution, we mostly choose the fusion algorithm to achieve our purpose. Therefore, a stable HDR image is affected by two factors: the number of LDR images and the performance of fusion strategies. Depending on the quantity of LDR images required for fusion, the current fusion tasks can be divided into the non-extreme exposure fusion [3], [4], [5] and the extreme exposure fusion [6], [7], [8], [9]. However, excessive LDR images will undoubtedly increase storage burden and time cost. So the following work in this paper is aimed at extreme exposure image fusion.

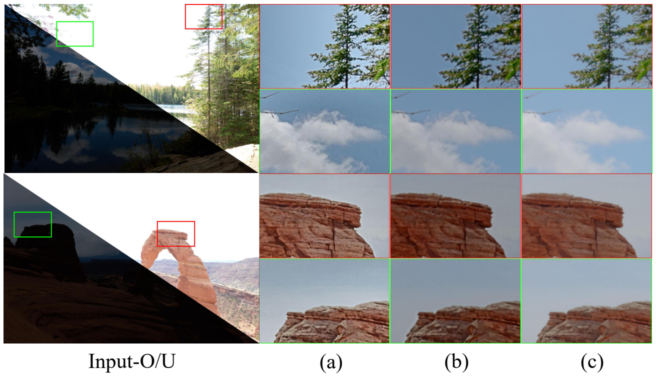

Fig. 1. (a), (b) and (c) refer to the detail patches of IFCNN, U2Fusion and our approaches. Comparing the detail of various approaches, it is clear that our work can achieve a more balanced exposure outcome.

Researchers have been working on the multi-exposure fusion(MEF) task and have produced plenty of excellent works in decades. These works may be broadly classified into two groups: *i.e.*, traditional-based approaches [10], [11], [12], [13] and deep learning-based approaches [14], [15], [16]. For the traditional group, the existing fusion methods are mainly based on transform domain [17] and spatial domain [18]. The former decompose image sequences into the transform domain, and then design fusion rules to reconstruct the fused image. The latter directly fuse on the pixel space of the image, whose pixel values are calculated linearly. However, the limitation of such methods is that it needs to manually design fusion rules,

This work is supported by the National Key Technology R&D Program of China (No. 2018YFC0910500), the National Natural Science Foundation of China (Nos. 61425002, 61751203, 61772100, 61972266, 61802040), LiaoNing Revitalization Talents Program (No. XLYC2008017), the Innovation and Entrepreneurship Team of Dalian University (No.XQN202008), Natural Science Foundation of Liaoning Province (Nos. 2021-MS-344, 2021-KF-11-03, 2022-KF-12-14). (Corresponding author: Shihua Zhou, Qiang Zhang)

Jia Lei, Jiawei Li and Shihua Zhou are with Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian, China (e-mail: leijia@s.dlu.edu.cn; lijiawei@s.dlu.edu.cn; zhoushihua@dlu.edu.cn).

Jinyuan Liu is with School of Software Technology, Dalian University of Technology, Dalian 116024, China (e-mail: atlantis918@hotmail.com).

Qiang Zhang is with School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China (e-mail: zhangq26@126.com).

Nikola K. Kasabov is with Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland 1010, New Zealand; Intelligent Systems Research Center, Ulster University, Londonderry BT52 1SA, UK and Auckland Bioengineering Institute (ABI), The University of Auckland, Auckland 1010, New Zealand (e-mail: nkasabov@aut.ac.nz).

which can easily lead to color deviation and detail blurriness of the fused image and lower time efficiency.

In addition, deep learning has become very popular recently, and meanwhile it has been effectively served in image enhancement [19], super-resolution [20], image restoration [21] and other low-level image tasks [22], [23], [24]. Researchers have also used deep learning in the MEF field owing to its excellent feature representation capabilities. Most existing MEF approaches utilize convolutional neural networks (CNN) for extracting features. Compared with traditional methods, convolutional neural networks can directly interact with image pixels to obtain required information, which is closer to the working mode of human visual system. Nevertheless, these MEF methods have some drawbacks: (i) some methods only capture the features of local neighborhoods through convolution operations, which neglect the overall dependencies and lost some essential global contexts. (ii) the fusion strategies used by some methods fail to integrate the features obtained from source images well. Moreover, these methods are prone to cause information degradation throughout fusion process, which lead that synthetic results produce structure defect and color deviation. (iii) Due to the large gaps between extremely exposed images, the detail information in the too bright or dark areas is difficult to be fully utilized. The fusion results may appear unclear targets and poor visual perception.

In this work, we solve the above limitations by proposing a global-local aggregation learning network in an unsupervised learning way. Concretely, the collaborative aggregation module is implemented to acquire the relevant information from input images, which not only pays attention to the pixel values of small receptive field, but also fully considers the long-range dependencies between pixels. To make full use of global-local features, we design a special fusion module to reconstruct fused images. Besides, we further refine the fusion results to recover more textural details by the recursive refinement module. A fused comparison is displayed in Fig. 1. Our work has more balanced exposure, clearer details and more vivid color. The following four aspects list the efforts made by our work.

- In order to fully mine the information of source images, we design a collaborative aggregation module to obtain the required features for fusion. It consists of two sub-modules. One is a non-local attention inference module, which is used to obtain global dependencies from feature maps. The other is a local adaptive learning module, which is regarded as a supplement to learn relevant local features. With the help of the two sub-modules, our collaborative aggregation module formulates a complete feature extraction mechanism to learn global-local features.
- According to the characteristics of the information extracted from source images, we provide a fusion module to reconstruct fused results, which effectively avoids artifacts and suppresses information decay.
- The recursive refinement module is introduced to fine-tune the fused results. It can use the loop mechanism to continuously excavate the information from the feature maps, improving the texture details of fused images.
- Extensive experiments on different datasets are undertaken to prove the effectiveness of our approach and highlight its superiority over ten other state-of-the-art approaches.

Our remaining work is arranged below. Section II mainly describes MEF methodology development and the background of non-local attention mechanisms. Section III illustrates the principle of our work in detail. Section IV. presents the comparative experiments and ablation studies. Finally, Section V makes a few conclusions about our findings.

II. Related Works

We particularly reviews both conventional and deep learning approaches to multi-exposure image fusion and explains the principle of standard non-local attention mechanisms.

A. Traditional-based MEF approaches

Traditional MEF approaches are generally classified into spatial domain-based approaches [18], [25] and transform domain-based approaches [26]. Spatial domain approaches, which are further divided into pixel-based [25], [27], patch-based [28], [29] and optimization-based [10] approaches, directly calculate the pixel values of input images. The primary technique of transform domain approaches is to decompose input images, convert it to various domains and then complete the subsequent fusion and construction operation. Commonly used transform methods include pyramid transform [30], Laplacian pyramid [31], wavelet transform [26], edge-preserving smoothing [32], *etc.*

As a representative, Lee *et al.* [18] defined two adaptive weights, which reflect the overall brightness and global gradient related to the pixel quality. The method only requires modest computational complexity to achieve visually pleasing results. Ma *et al.* [28] introduced a patch-wise method, which attaches importance to two independent parts in the image patch, pixel intensity and signal information. The input image is initially split into patches, two elements are deconstructed, processed, and integrated, and then a full-color fusion image is recombined. Song *et al.* [27] presented a maximum posterior probability architecture that guides the fusion process to obtain additional details using the ideal gradient information of input contrast images. However, as the perfect contrast image doesn't exist, the authors employ a luminance image to approximate it. Mertens *et al.* [31] proposed a framework for fusion that places a strong emphasis on multi-scale information fusion. Given input are decomposed by the Laplacian pyramid, which then calculates and normalizes the weight map from three factors: contrast, color saturation and exposure. The Gaussian pyramid is used to smooth the weight map. The Laplace pyramid of the multi-exposure images is multiplied by the Gaussian pyramid of the weighted maps to get the final fusion result. Zhang *et al.* [33] proposed a method based on the color space variation of input images, which converted the input image into YUV space. The different components of this YUV space involve luminance and color information, respectively. The luminance component is converted to the wavelet domain, and well-exposed weight and adjustable contrast weight are fused. The chromatic components are fused according to saturation weights. The fused image is next transformed into the standard RGB space. Kou *et al.* [34] successfully constructed a multi-scale method by introducing an edge-preserving smooth pyramid to smooth the weight maps. Information about the most complex areas of the image, including the brightest and darkest parts, is well-preserved thanks to the edge-preserving properties.

B. Deep Learning-based MEF approaches

Deep learning has seen notable progress in computer vision and image processing applications [35], [36]. Deep learning-based MEF approaches have proliferated, and they may be classified as supervised or unsupervised depending on whether or not ground truth is used. In supervised MEF algorithms, Kalantari [37] introduced the supervised CNN-based MEF framework for the first time. The ground truth image datasets are generated by combining three static images with various exposure settings. The convolutional neural network (CNN), which also provides the fusion weights, is employed to generate the final fused outputs. Zhang *et al.* [38] built a full convolution method that can fuse images with arbitrary resolution. The element average fusion rules are used to combine the convolution features that were derived from two branches. Additionally, perceptual loss improves the model. A general image fusion network that can solve multiple modal tasks was introduced in [42]. It discusses the commonalities and properties of various fusion works and analyzes the effect of different network structures on the performance of image fusion. Liu *et al.* [39] proposed a MEF framework based on attention guidcance, which uses two discriminators to preserve global-local features. Ground truth images are always missing in real scenes, and thus researchers have developed unsupervised fusion methods. Prabhakar *et al.* [7] built the first unsupervised architecture for MEF. With this technique, the image's color space is altered during the fusion process, the information from various channels is processed separately, and the color space is finally reversed in accordance with the actual requirement. The specific fusion strategy of this method is to use convolutional neural networks to extract features for Y in YCbCr space and manually fuse the CbCr channels using a predetermined formula. Xu *et al.* [40] developed an unsupervised fusion network, which was applied to build a unified image fusion model without ground truth or standard reference metrics and solve various fusion tasks. Deng *et al.* [41] used a coupling feedback mechanism to simultaneously fuse and super-resolve a pair of input overexposed and underexposed images.

C. Non-local Attention Mechanism

Non-Local Attention(NLA) was first proposed by Wang *et al.* [42]. As a new technology, it has been widely used in Computer Vision(*e.g.*, image restoration [43], image compression [44], image super-resolution [45]). Formally, standard non-local attention is defined as: $\sum_{j=1}^{M} \frac{\exp\left(Q_i^T K_j\right)}{\sum_{\hat{j}=1}^{M} \exp\left(Q_i^T K_{\hat{j}}\right)} V_j$. In detail, Q, K and V are respectively the results obtained from the feature transformation of input. i and j are representation of specific positions about the feature map. M represents the size of the input. c^i and c are the number of channels. Standard non-local attention gathers all features, importing unnecessary noises into results and causing quadratic computational cost. To solve the problems, Mei *et al.* [20] used Locality Sensitive Hashing to quickly assemble crucial information. However, the method ignores global correlation information. To alleviate the issue, Xia *et al.* [46] proposed a Efficient NLA to aggregate all features efficiently. Efficient NLA adopts the kernel method to close to exponential function $\exp\left(Q_i^\top K_j\right)$. The technique can not only compress the fusion of excessive irrelevant features, but also can reduce the linear computation cost. Therefore, this efficient NLA is introduced into our cooperative aggregation module to help our model obtain global information and formulate an effective feature extraction mechanism.

III. Proposed Method

A core part of our approach is a collaborative aggregation module for global and local information aggregation, a fusion module for generating fused images and a recursive refinement module for improving the detail texture of the final fusion images. Prior to accessing the bulk of our network, we first process source images to obtain initial feature maps F_o. In specific, two source LDR images I_1 and I_2 with extreme exposure settings are concatenated to go through a 3×3 convolution layer, and LeakyReLU as the activation layer. In Fig. 2, where we apply the proposed network to an unsupervised image fusion task, the detail architecture is displayed.

A. Collaborative Aggregation Module

The retention degree of the relevant information in feature maps affects the performance of fused images. We give more weight to the information that needs attention. Based on this consideration, we come up with a collaborative aggregation module(CAM) to acquire more relevant features. Different from previous methods, the CAM formulates a complete feature extraction mechanism, which acquires information from both global and local aspects. In detail, the CAM consists of a non-local attention inference module with Sparse Aggregation and a local adaptive learning module. The former is employed to acquire pertinent non-local features, and the latter is applied to be a supplement to preserve more local information. We utilize the collaborative aggregation module to refine the initial feature maps F_o, converted into two feature representation sequences with merging global-local information. So, the process can be defined as:

$$[F_{\mathcal{A}}, F_{\mathcal{C}}] = [\mathcal{A}(F_o), \mathcal{C}(F_o)] \tag{1}$$

where $\mathcal{A}(\cdot), \mathcal{C}(\cdot)$ refer to the non-local attention inference module with Sparse Aggregation and the local adaptive learning module respectively. $F_{\mathcal{A}}$ and $F_{\mathcal{C}}$ represent the output sequence of the two sub-modules.

1) Non-Local Attention Inference Module: The main goal of the non-local attention inference (NLAI) module, which takes the NLA mechanism as its foundation, is to leverage more relevant non-local features and acquire long-range visual

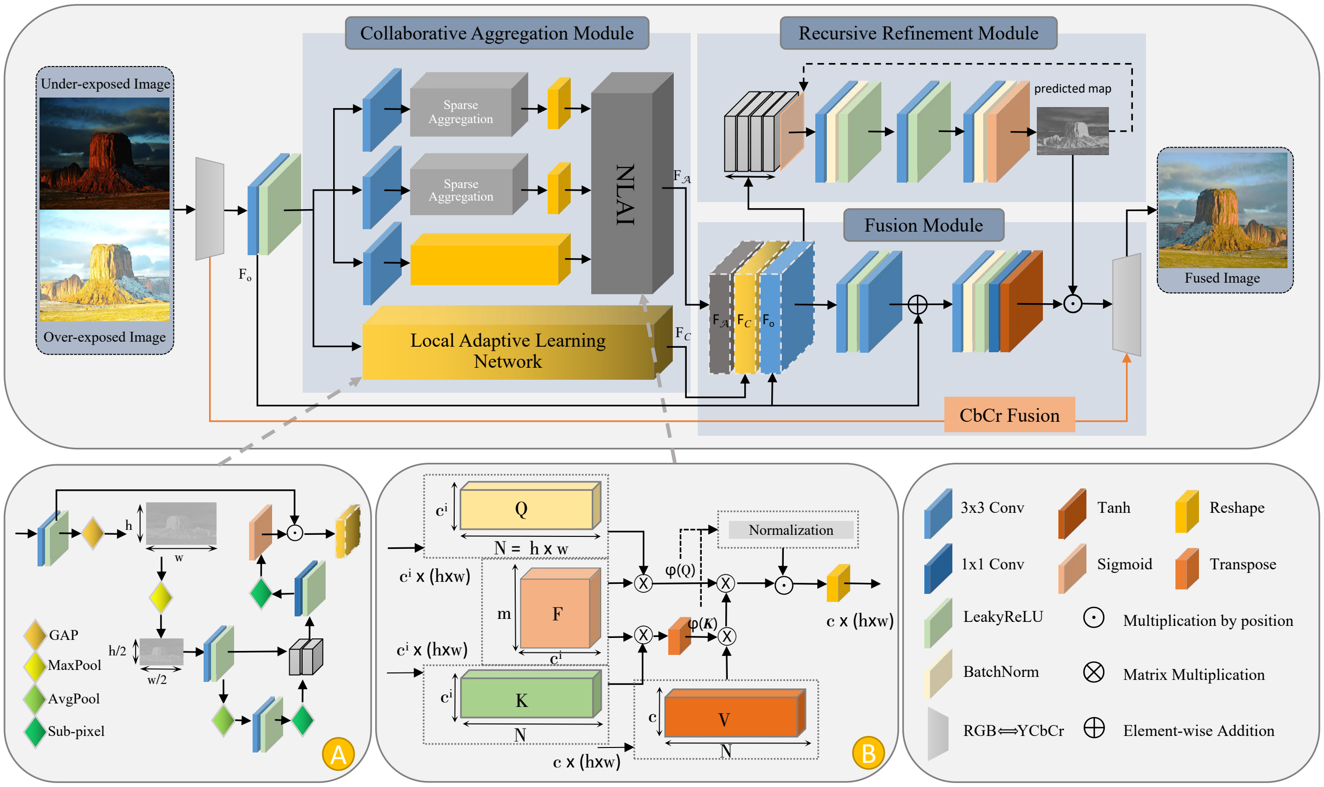

Fig. 2. Our image fusion framework. The network is composed of a collaborative aggregation module(CAM), a special fusion module(FM) and a recursive refinement module(RRM). The collaborative aggregation module has two sub-modules(*i.e.*, a non-local attention inference module(NLAI) with Sparse Aggregation and a local adaptive learning module) to learn global and local information from feature maps. The learned information is passed into the fusion module to reconstruct fusion results. The RRM is implemented to further fine-tune texture details of the fusion results.

dependencies. However, standard NLA gathers all features and propagates meaningless noise into fused results. Therefore, we adopt kernel method [46] to close to the exponential function in Standard NLA, which can suppress the fusion of excessive irrelevant features and obtain lower computation complexity.

The detail architecture of the NLAI is displayed in the middle part of Fig. 2 (Part B). The input F_o of dimensions $c \times h \times w$, where c stands for channel numbers, h and w are height and width of input, respectively. Firstly, We transferred the input F_o to three feature representations K, Q and V, respectively. Then we decompose the exponential function $\exp\left(Q_i^\top K_j\right)$ in Standard NLA. Concretely, we combine K, Q with Gaussian random matrix F to approximate the exponential function and modify the multiplication order to reduce the computational cost. Among them, Gaussian random matrix $F \in \mathbb{R}^{m \times c}$ consists of m different Gaussian random samples $f_1, ..., f_m$. The specific projection $\varphi(\cdot)$ is shown in Fig. 3, which displays the conversion process from Q, K to $\varphi(Q)$, $\varphi(K)$. Therefore, we use $W = \left(\varphi(Q)^\top \varphi(K)\right)$ close to exponential function for converging global information. And we define the NLAI as:

$$F_{\mathcal{A}} = D^{-1} \odot \left(WV^\top\right), \tag{2}$$

where $F_{\mathcal{A}}$ is the approximated standard NLA, D denotes the normalization in softmax operators for $\varphi(Q)$, $\varphi(K)$. $\odot$ is the point-wise multiplication.

To further improve the performance of the NLAI, Sparse Aggregation(SA) is used to filter out irrelevant features and expand the weight of relevant features. The essence is to enhance the sparsity of weight by applying a raising factor $k(k > 1)$ to the input and forcing NLAI to assign higher weights to relevant features. In Fig. 2, the Sparse Aggregation may be written:

$$Z = \sqrt{a}\frac{\theta(X)}{\|\theta(X)\|}, \tag{3}$$

where X is the input, and raising factor $a(a > 1)$ is employed to strengthen non-local sparseness. $\theta(\cdot)$ is feature transformation.

2) Local Adative Learning Module: The role of Sparse Aggregation is to filter out irrelevant parts and expand the weight of relevant parts in feature maps, which greatly affects the performance improvement of the NLAI. However, Sparse Aggregation doesn't have the function of increasing the gap between irrelevant and relevant features. So the irrelevant regions ignored by the SA may still retain a little bit of important information, which may impact the final fusion effect. To solve this problem, we introduce a U-Net-like local adaptive learning module to learn local feature weights from source images, which is also seen as a complement to the NLAI. The local adaptive learning module be separated into two stages: the down-sampling stage and the up-sampling stage, as shown in Fig. 2(Part A). In the down-sampling stage, the initial feature maps F_o first go through a convolutional layer, termed F_c.

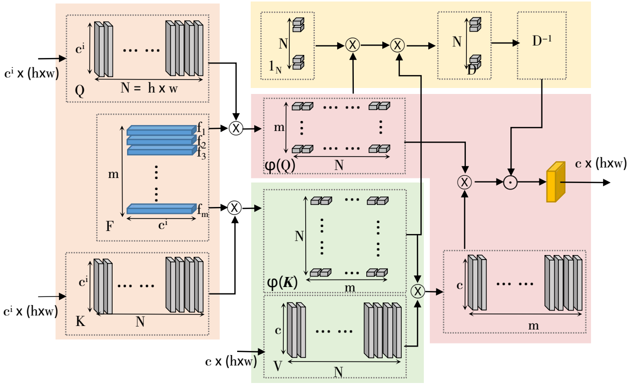

Fig. 3. The specific projection process of NLAI.

And then put it into Global Average Pooling(GAP). GAP can be used to aggregate information to generate $F_g \in \mathbb{R}^{c\times h\times w}$. The down-sampling stage is composed of two steps. In the first step, we use the max-pooling operation(MP) to keep the most meaningful information in receptive field while reducing the resolution of F_g to $1 \times \frac{1}{2}h \times \frac{1}{2}w$. After that, a convolutional operation is employed to extract information even more thoroughly. The first step is expressed as follows:

$$F_m = Conv_3(MP(GAP(F_c))), \tag{4}$$

where $Conv_3$ denotes the 3×3 convolution. The resolution of F_m is $8 \times \frac{1}{2}h \times \frac{1}{2}w$. In the second step, we scale-down the feature maps F_m using the average pooling(AP) operation. And the results after AP operation are transferred to a convolution layer to obtain the feature maps F_a of dimension $16 \times \frac{1}{4}h \times \frac{1}{4}w$. We calculate the feature maps F_a by

$$F_a = Conv_3(AP(F_m)). \tag{5}$$

In the up-sampling stage, we employ a sub-pixel convolution(Sp) operation to scale-up the feature maps F_a to $4 \times \frac{1}{2}h \times \frac{1}{2}w$. Concatenated the results from the Sp operation and F_m are pumped into a 1×1 convolutional layer. Finally, a additional Sp layer is employed to acquire weight maps $F_s \in \mathbb{R}^{1\times h\times w}$ containing relevant local information, and the sigmoid function is executed to normalize the value F_s to $[0, 1]$. Specifically, the up-sampling process can be defined as

$$F_s = Sp(Conv_1[Sp(F_a), F_m]), \tag{6}$$

$$F_{\mathcal{C}} = Sigmoid(F_s) \odot F_c, \tag{7}$$

B. Recursive Refinement Module

We propose the recursive refinement module(RRM) for improving the detail texture of fused images, which has a iterative refinement mechanism using the previously estimated distribution as a guidance of the current distribution. To realize this recursive module, we formulate a recursive refinement network $\mathcal{R}$ with three convolutional layers, whose detail shows in Fig. 2. The RRM is defined as

$$p^t = \mathcal{R}\left(F_o; p^{t-1}\right) \tag{8}$$

where p^t and p^{t-1} are the predicted maps at t^{th} and $(t-1)^{th}$ iteration, respectively. The initial predicted map is defined as zeros. For the number of iteration, we set t_{max} to 3. With the number of iterations rising, the data distribution of the predicted map is closer to initial feature maps F_o.

C. Fusion Module

1) Ours Fusion Strategy: We formulate a special fusion module, which can employ the global and local features of sources images and suppress information degradation, to ensure the reconstruction of a visually pleasant fusion image. By the collaborative aggregation module, two sequences of features $F_{\mathcal{A}}, F_{\mathcal{C}}$ are generated and retain global-local information from source images. We can utilize the feature representation sequences to reconstruct fused image. The layout of our fusion module(FM) is depicted in Fig. 2. In specific, sequences of features $F_{\mathcal{A}}, F_{\mathcal{C}}$ are first concatenated to balance the intensity distribution of pixels and restore information lost in localized areas. Then we perform four convolution operations on these sequences to carrying out a preliminary feature fusion. At the same time, in order to make up for the representation that was lost during the CAM process, we also introduce initial feature maps F_o during the concatenation and add a skip connection, which adds initial feature maps F_o and the output of the second convolution together. After the above process, we can get a preliminary friendly fusion image $\widetilde{I}$. To further improving the detail quality of the fused image, we introduce a recursive refinement module(RRM) to generate a predicted map p, which guides the $\widetilde{I}$ to create a final outcome I_f that is merged with complimentary and comprehensive information. As a result, the fusion process may be described:

$$I_f = \mathcal{F}\left(F_{\mathcal{A}}, F_{\mathcal{C}}, F_o\right) \odot p \tag{9}$$

2) Managing RGB Input: We apply common strategies to fuse RGB images, the key of which is to fuse specific information on different channels. First, the color channel of RGB images is transformed to the YCbCr. Using our proposed strategy, the Y channel is fused. Then the conventional technique [47] is utilized to fuse the information in the Cb and Cr channels, which is characterized as:

$$C_f = \frac{C_x\left(|C_x - \tau|\right) + C_y\left(|C_y - \tau|\right)}{|C_x - \tau| + |C_y - \tau|}, \tag{10}$$

where C_x and C_y refer to the Cb, Cr channel values from multi-exposure images. C_f denotes fused channel results. τ is set to 128. Finally, the two channels (*i.e.,* the fused Y channel and the C_f channel) are reversed into RGB space together.

D. Loss Function

To motivate our model to learn the pixel dependencies from input, we introduce three loss functions to restrain the resemblance between the original image I_i and the final fused image I_f. The total loss function is employed as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{2} \mathcal{L}_{\text{fused}}\left(I_i, I_f\right), \tag{11}$$

where $\mathcal{L}_{\text{fused}}$ is the fusion loss function, containing the structural similarity(SSIM) loss function $\mathcal{L}_{\text{SSIM}}$ [48], the mean square error(MSE) loss function $\mathcal{L}_{\text{MSE}}$ and the total variation loss function $\mathcal{L}_{\text{TV}}$ [49]. In addition, two hyperparameters α and β are introduced, $\mathcal{L}_{\text{fused}}$ is defined as:

$$\mathcal{L}_{\text{fused}} = \mathcal{L}_{\text{SSIM}} + \alpha\mathcal{L}_{\text{MSE}} + \beta\mathcal{L}_{\text{TV}}. \tag{12}$$

The SSIM loss improves the model's ability to extract structural elements from input, while the MSE loss constrains the pixel difference between two images. $\mathcal{L}_{\text{SSIM}}$ and L_{MSE} are defined as:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}\left(I_f, I_i\right), \tag{13}$$

$$\mathcal{L}_{\text{MSE}} = \left\|I_f - I_i\right\|_2. \tag{14}$$

The total variation loss $\mathcal{L}_{\text{TV}}$ introduced in [49] is employed to recover gradient information from input images and further suppress noise. It is described as follows:

$$R(m,n) = I_f(m,n) - I_i(m,n), \tag{15}$$

$$\begin{aligned}\mathcal{L}_{\text{TV}} = \sum_{m,n} (&\|R(m,n+1) - R(m,n)\|_2 \\ &+\|R(m+1,n) - R(m,n)\|_2),\end{aligned} \tag{16}$$

where $R(m,n)$ means the difference between the input image and the fused image, $\|\cdot\|_2$ is the l_2 norm, and m,n refer to coordinate relationship of the image's pixels, respectively.

IV. Experiments and Results

This section elaborates on our experimental setup, covering training details, optional datasets, evaluation metrics and comparison methods. In addition, we discuss the performance of our methods through comparative and ablation experiments. Finally, the efficiency comparison further proves the superiority of our work.

A. Experimental Settings

1) Training Details: For training our modules, initial resizing of the input images to 512×512, and then the downsized images are randomly split into 128×128 patches in each round of training. The Adam is used to optimize the network with default parameters and fixed learning rate 1e-4. The batch size is 17. Hyperparameters α and β are both set to 20. Five hundred epochs make up the entire training procedure, implemented in the Pytorch framework. The hardware support with a NVIDIA RTX 2080 Ti GPU, a 11GB RAM Memory, and an Intel Core i7-7700HQ CPU.

2) Optional Datasets: To train and evaluate our method, we introduce the SICE dataset [50], which provides all kinds of exposure image sequences(*e.g.*, church, lawn, railway, sunset). Each sequence contains a variable number of well-aligned images. To demonstrate the fusion capability of our approach for extreme scenarios, we only select two images with significant gaps in exposure levels from each sequence to train our model. We randomly select 374 image sequences with extreme exposure images as our training dataset. From the remaining sequences, 100 image sequences are chosen to serve as the test dataset. In addition, as a supplement to the test set, we introduce 18 image sequences from dataset [29] and select pairs of exposure images in the same ways.

3) Evaluation Metrics: Refer to previous work, we select two important fusion metrics, the structure similarity index measure (MEF-SSIM) [51] and peak signal-to-noise ratio(PSNR), to accomplish quantitative analysis for the subsequent experiments.

- MEF-SSIM evaluates the structural consistency of image patches by removing luminance components. The statistical values range of MEF-SSIM is constrained between 0 and 1. The image perception quality increases as the value gets closer to 1. Given the structural comparison element S:

  $$S\left(\{\mathbf{x}_k\}, \mathbf{y}\right) = \frac{2\sigma_{\hat{\mathbf{x}}\mathbf{y}} + C}{\sigma_{\hat{\mathbf{x}}}^2 + \sigma_{\mathbf{y}}^2 + C}, \tag{17}$$

  where $\{\mathbf{x}_k\}$ and $\mathbf{y}$ indicate the corresponding patches obtained from the source image and the fused image, respectively. $\sigma_{\hat{\mathbf{x}}}^2$ and $\sigma_{\mathbf{y}}^2$ denote the local variances of $\hat{\mathbf{x}}$ and $\mathbf{y}$, and $\sigma_{\hat{\mathbf{x}}\mathbf{y}}$ is the local covariance between $\hat{\mathbf{x}}$ and $\mathbf{y}$. C is a constant reflecting low contrast effects [48].
- PSNR is a metric used to represent the retio between peak power and noise power of the fused image. It makes it possible to evaluate aberrations in the fusion process, defined as follows:

  $$PSNR = 10\log_{10}\frac{r^2}{MSE}, \tag{18}$$

  where r is the peak of the fused result, and is set to 255. MSE is the mean squared error that measures the inressemblance between two images. It calculates as:

  $$MSE_{SF} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[S(i,j) - F(i,j)], \tag{19}$$

  where S and F denotes the input and fused image. The final calculation can be constructed as: $0.5 * MSE_{AF} + 0.5 * MSE_{BF}$, A and B are the two input with different exposure settings. In general, getting higher PSNR values means that the fused results retains more details with less deviation.

4) Comparison Methods: We compare our method with ten state-of-the-art methods, involving four traditional methods, *i.e.*, DEM [5], DSIFT [25], FMMEF [17] and GFF [3], and six deep learning methods, *i.e.*, U2Fusion [40], DeepFuse [7], IFCNN [38], MEF-Net [52], CF-Net [41] and AGAL [39]. In specific, DEM performs simple multi-scale operation in the YUV space. DSIFT can effectively extract local details to obtain fusion images without ghosting artifacts through scale-invariant feature transform.FMMEF utilizes a fast structural patch decomposition method to improve fused results. The highlight of GFF is information fusion based on guided filtering. U2Fusion can solve multiple fusion tasks, which limits the resemblance between fused results and input images. DeepFuse is the iconic MEF method that builds the unsupervised training structure for the first time. IFCNN can use full-convolutional architecture to fuse various types of images. MEF-Net develops a fast MEF method, which can

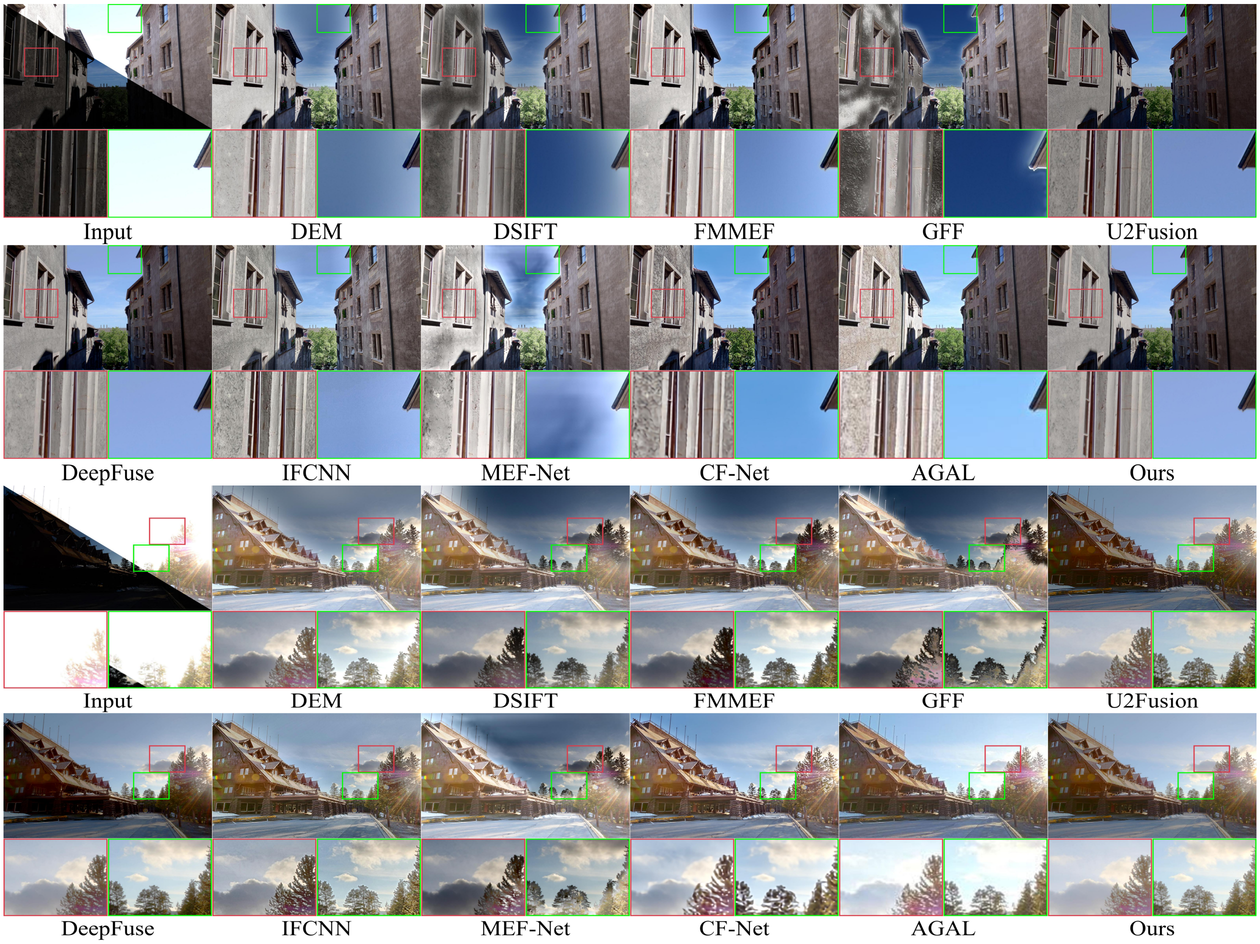

Fig. 4. Qualitative comparisons with eleven methods on the SICE dataset. The part of fused images is magnified by red and green boxes to observe the differences between these methods. For the above examples, our method generates fused images with more balanced exposure and abundant details(*e.g.*, sky, walls, trees and clouds).

fuse images of random resolution and exposure number. A coupled feedback network provided by CF-Net that allows both MEF tasks and super-resolution tasks to be carried out. AGAL is an adversarial netwok based on two discriminators for processing exposure images. Each approach is assessed in accordance with its official code.

B. Observational Analysis Using The SICE Dataset

1) Qualitative Comparisons: In this analysis, we compare our approach against ten other excellent methods using two ordinary image sequences from the SICE test set. Fig. 4 shows the results of our visual comparison. To evaluate these methods comprehensively and reasonably, we consider the overall image and local details(the red and green boxes are utilized to enlarge the local details below each image). Look at the whole image, areas of DEM, DSIFT, FMMEF, GFF and MEF-Net are clearly shaded with black, such as the sky and clouds above the building. In the second sequence of images, the house color in methods U2Fusion and DeepFuse is low-exposure, which means that these two methods can not recover the color information well. Moreover, IFCNN and AGAL have a seemingly good visual effect, but it can be found from the figure that their exposure is out of balance, the former is slightly dark and the latter is slightly light. The details of CF-Net are blurred and flawed, especially this wall and the edge of the tree. After comparison, we found that our method can produce an image with a balanced exposure, both its detail

TABLE I
THE VALUES OF TWO COMMON METRICS REPRESENT THE FINDINGS OF THE QUANTITATIVE COMPARISON OF ELEVEN APPROACHES ON SICE TEST DATASET. THE TOP TWO OUTCOMES ARE DENOTED IN RED AND BLUE.

	Methods	MEF-SSIM	PSNR
Traditional	DEM	0.9300	57.1168
	DSIFT	0.9242	56.8504
	FMMEF	0.9311	57.4353
	GFF	0.8547	56.3125
DL-based	U2Fusion	0.9447	**58.5469**
	DeepFuse	0.9019	58.2120
	IFCNN	0.8863	57.6739
	MEF-Net	**0.9486**	56.6181
	CF-Net	0.7561	57.8450
	AGAL	0.9019	57.9328
	Ours	**0.9512**	**58.6529**

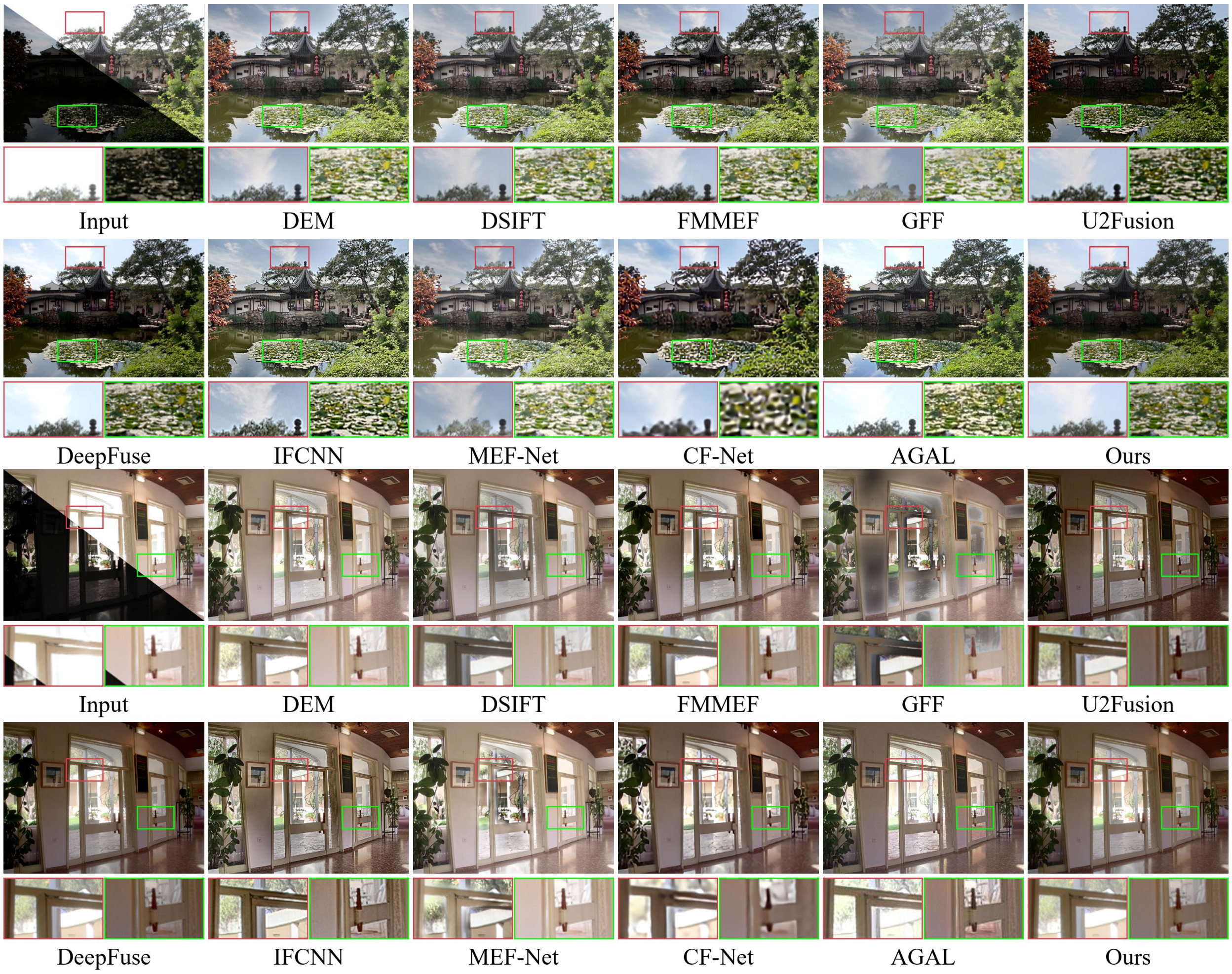

Fig. 5. Qualitative comparisons with eleven methods on dataset [29]. The part of fused images is magnified by red and green boxes to observe the differences between these methods. For these examples, our proposed method can achieve the desired results with balance pixel intensity and realistic details(*e.g.*, lotus leaves and doorhead).

and color can be recovered well.

2) Quantitative Comparisons: After the subjective assessment, we begin to analyze the image quality objectively. Based on previous MEF works, we select the most commonly used image metrics, which is described in the experimental settings section, to measure the fusion ability of different methods. The quantitative analyses are performed on the SICE dataset. On the basis of 100 test image pairs, we calculate the average scores of the two metrics of eleven methods separately. It should be noted that the training dataset and the test dataset do not cross over. Table. I reports the quantitative results, which shows that our method generates fused results with higher metric values than those of ten other approaches. Additionally, the statistical data in the table reflects that our approach can deliver a strong visual image with balanced exposure and accurate texture details.

C. Observational Analysis Using The Dataset [29]

1) Qualitative Comparisons: In addition to the SICE dataset, we also use another dataset as a supplement to ensure the rigor of the entire experimental analysis. Fig. 5 shows two image pairs on the Dataset [29], involving the visual comparison of two categories about 11 approaches. Among these traditional methods, DEM, DSIFT, FMMEF and GFF appear to exposure imbalances and obvious artifacts. For example, the lotus leaf area in the first image pair has noticeable brightness, and shades of black appear near the doorhead in the second image pair. Among the deep learning approaches, the fused images acquired by U2Fusion and DeepFuse have poor color, whose local areas are pretty gloomy and fail to recover proper luminance. For methods IFCNN, MEF-Net and CF-Net, observing the local parts of their fused images, it can be found that the details are either too sharp or too fuzzy. Moreover, AGAL can avoid shadow artifacts but produce color distortion. Although the exposure gap of the source images is quite large, our proposed method still achieves the desired results, which can balance pixel intensity and realistic details.

2) Quantitative Comparisons: We selected 18 image pairs from dataset [29] and obtained the corresponding fused images by using different fusion methods. Afterward, we calculate the metric values, which quantifies the ability of the fused images produced by each method to retain information. Fig. 6 shows the concrete quantification results and our method is indicated

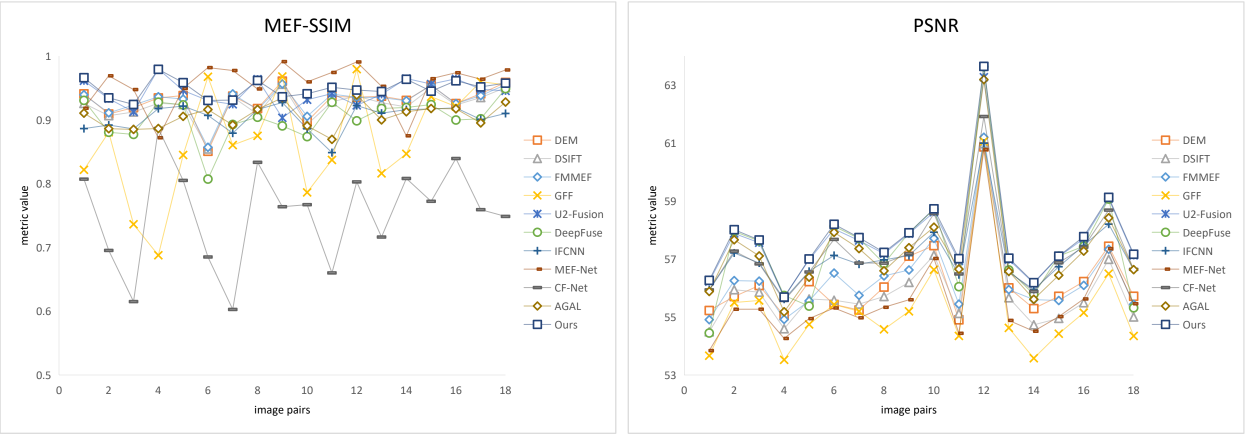

Fig. 6. Quantitative comparisons of two metrics with eleven methods on dataset [29], the horizontal and vertical axes refer to image pairs and metric values.

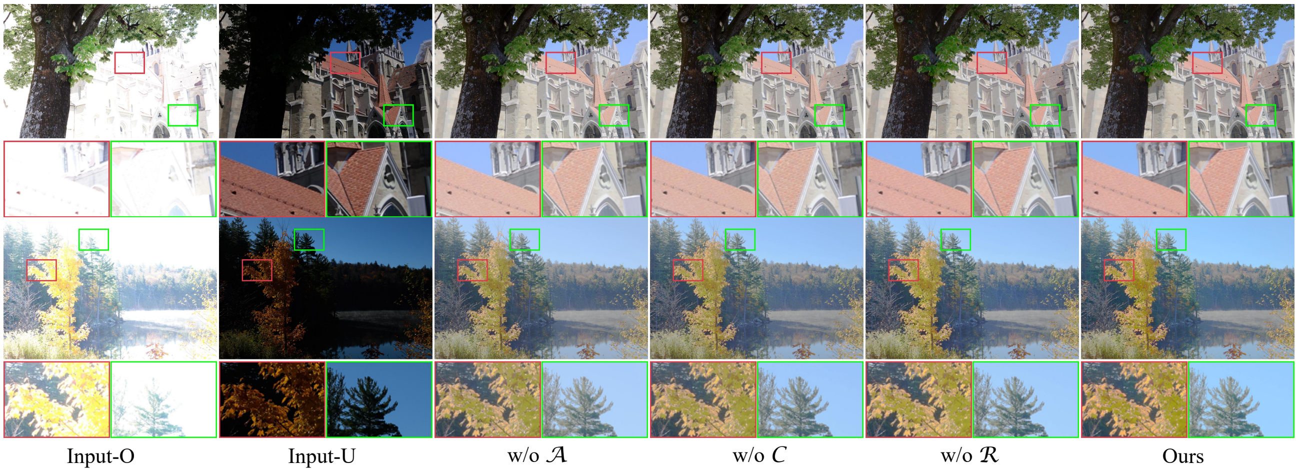

Fig. 7. Ablation analysis on the SICE dataset validates the effectiveness of three modules(*i.e.*, the non-local attention inference module $\mathcal{A}$, the local adaptive learning module $\mathcal{C}$ and the recursive refinement module $\mathcal{R}$) in our method.

by dark blue boxes. It is clear from the figure that our PSNR is the best, which means that the distortion is the lowest in our fusion process. For MEF-SSIM, our method is not always the best, and that's because MEF-SSIM is not particularly accurate in accordance with the human visual system since it concentrates on the structure consistency between the fused target and input target. In other words, a great MEF approach should balance the metric score with the visual results.

D. Ablation Experiments

1) Discussion of Sub-modules: We set up experiments to verify the work utility of three modules, including the non-local attention inference module, the local adaptive learning module and the recursive refinement module. In Fig. 7, we display the outcomes of our entire model with and without per module, and three modules are denoted by $\mathcal{A}$, $\mathcal{C}$ and $\mathcal{R}$, respectively. From the figure, it is obvious that each module is indispensable and affects the final fusion result. Compared to the visual output of the full model, the color of the fused results without $\mathcal{A}$ or $\mathcal{C}$ is obviously not recovered well, as well as the images have relatively decreased sharpness and fragile details, especially the roof and the leaves. These are because the global and local features are not fully utilized, which seen that only a complete cooperative aggregation module can ensure the final fusion effect. The textural expression of the fusion results can be further improved using the recursive refining module, which can correct color distortion and enrich details. In addition, we conduct quantitative analysis of different modules. Table. II reports whether or not each

TABLE II
QUANTITATIVE ANALYSIS OF TWO METRICS ON SICE DATASET VALIDATES THE EFFECTIVENESS OF THREE MODULES(*i.e.*, THE NON-LOCAL ATTENTION INFERENCE MODULE $\mathcal{A}$, THE LOCAL ADAPTIVE LEARNING MODULE $\mathcal{C}$ AND THE RECURSIVE REFINEMENT MODULE $\mathcal{R}$) IN OUR METHOD.

$\mathcal{A}$	$\mathcal{C}$	$\mathcal{R}$	MEF-SSIM	PSNR
✗	✓	✓	0.938	58.678
✓	✗	✓	0.947	**58.733**
✓	✓	✗	0.947	58.725
✓	✓	✓	**0.951**	58.652

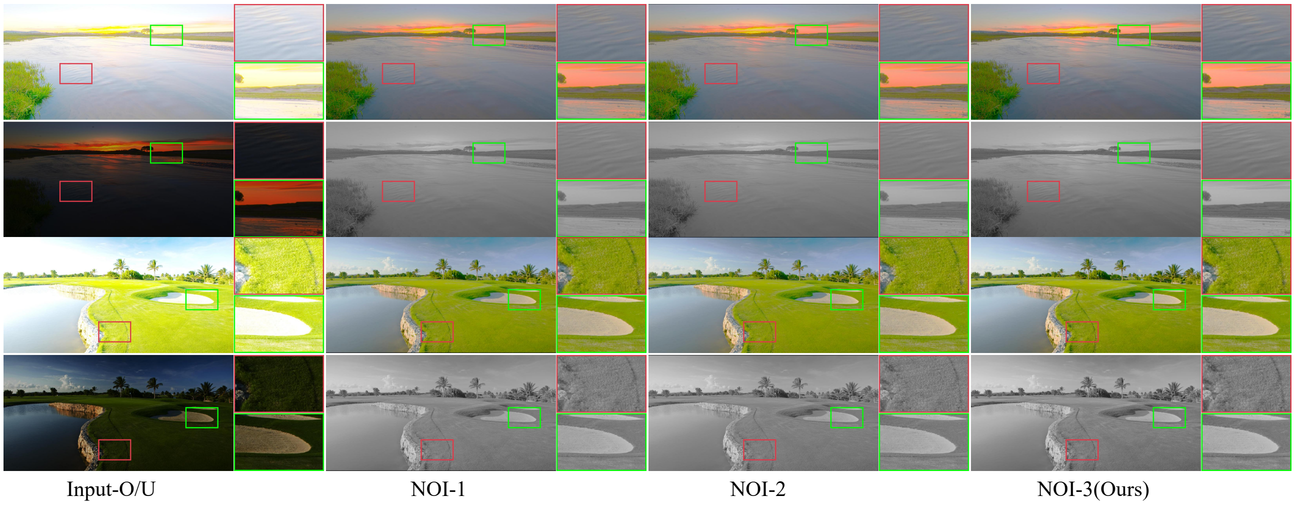

Fig. 8. Ablation analysis on SICE dataset tests the performance of iteration counts on fused outcomes in recursive refinement module. The grayscale image in the middle is the predicted map obtained by our RRM.

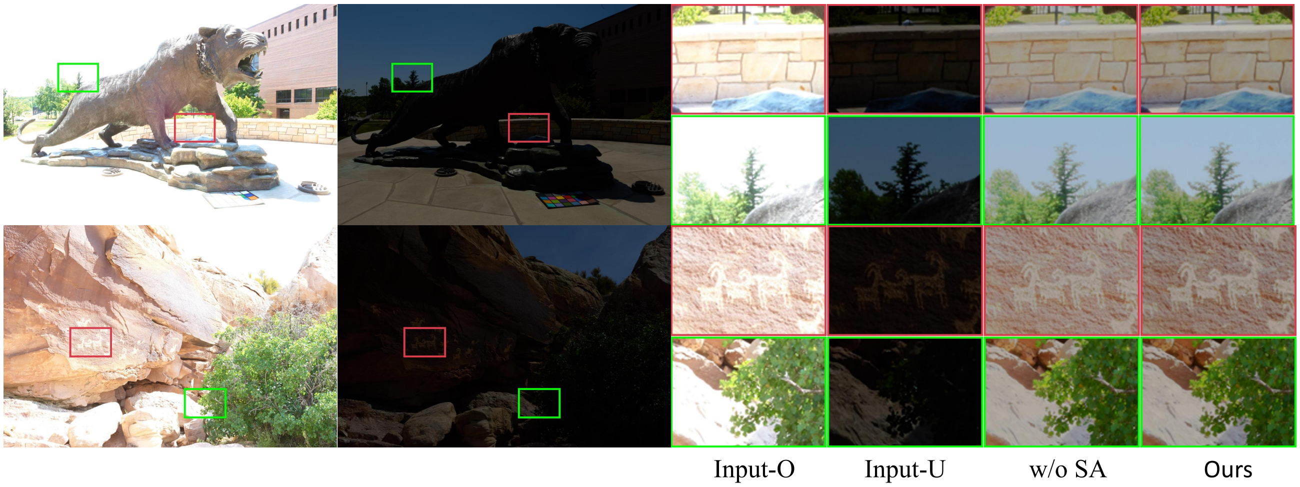

Fig. 9. Ablation analysis on SICE dataset validates the effectiveness of Sparse Aggregation.

module affects the final score. The PSNR score in the table fluctuates. The calculation of PSNR is related to MSE, which has a limitation that it may not completely conform to the visual law. During the experiment, MSE is very sensitive to pixel changes, leading to changes in PSNR scores.

2) Discussion of the Number of Iteration: The fusion performance of the results might be impacted by the iteration count of the recursive refinement module. The NOI-3 results, as shown in Fig. 8, offer significantly clearer details, such as ripples in the water and markings in the grass. With a higher iteration count, the detail quality and contrast of the reconstructed results are obviously enhanced. The essence of this iterative refinement procedure is to fit the numerical distribution of the initial feature map over and over again. By the mechanism, weight maps are obtained and finally produce a satisfactory result. The quantitative analysis in Table. III. The change of PSNR is affected by the pixel change in each iteration.

TABLE III
ABLATION ANALYSIS OF TWO METRICS ON THE SICE DATASET TESTS THE PERFORMANCE OF ITERATIONS COUNTS ON FUSED OUTCOMES IN RECURSIVE REFINEMENT MODULE.

	MEF-SSIM	PSNR
NOI-1	0.943	**58.669**
NOI-2	0.946	58.629
NOI-3(All)	**0.951**	58.652

3) Discussion of Sparse Aggreagation: We discuss the validity of Sparse Aggregation in the non-local attention inference module. Concretely, we add or remove SA blocks in the network to experiment with their effects on the fused image. Fig. 9 shows two pairs of exposed images zoomed in on local details, input-O/U. In addition, the corresponding fusion image patches with or without SA are presented successively. The fact that the engraving on the stone walls and the details of the plants are much clearer under the guidance of the SA block. The method with SA blocks can enlarge the weight of

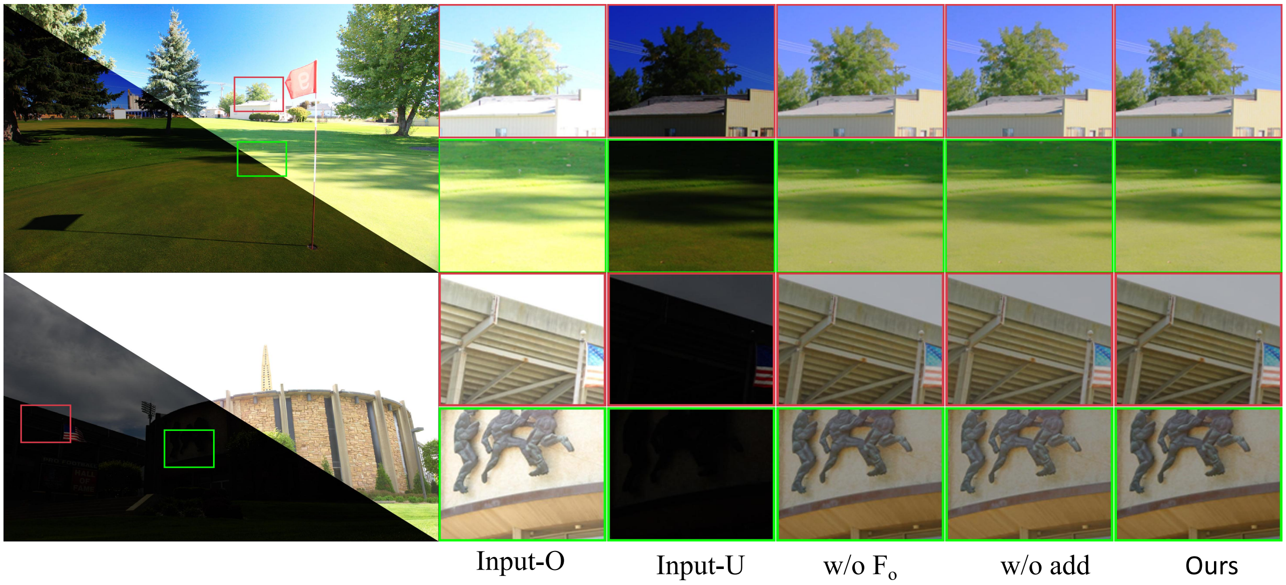

Fig. 10. Ablation analysis about fusion tradeoff on SICE dataset. From front to back: over-exposed image patch, under-exposed image patch, the concatenation without F_o, without addition operation, full module.

useful information and ensure the texture details and color of the fusion result. It may be seen from the quantitative analysis in Table. IV that metric values have a positive increase when Sparse Aggregation is used. Thus, we can prove that Sparse Aggregation is effective for our model.

4) Discussion of fusion tradeoff: To compensate for the lost information in the network, we take two measures, including introducing the initial feature map F_o during the concatenation and using the addition operation in the fusion process. For our fusion module, the essence of the initial feature maps and the addition operation is to reemploy previous features to prevent information decay. In Fig. 10, From front to back, we present the visual results of various operations, including the exposed image without any processing, unused F_o during the concatenation, removing the addition operation and the complete model. From the figure, the model with the full operation has better saturation and more detail. In addition, we also execute a quantitative analysis in Table. V.

5) Discussion of Other Conditions: Our approach can utilize two images with different exposure settings, generating a balanced exposure and visually pleasing result. To further test the reconstruction capability of our method, we randomly select an image sequence from SICE dataset, which contains five images with various exposure settings of the same scene. These images are arranged and combined, and the fusion outcomes are presented in Fig. 11. Data with the same exposure levels as (1) and (5) are used to train our model, so the result of this combination is optimal in terms of saturation and detail preservation.

TABLE IV
ABLATION ANALYSIS OF TWO METRICS ON SICE DATASET VALIDATES THE EFFECTIVENESS OF SPARSE AGGREGATION.

	MEF-SSIM	PSNR
w/o SA	0.938	58.605
All	**0.951**	**58.652**

TABLE V
ABLATION ANALYSIS ABOUT FUSION TRADEOFF.

	MEF-SSIM	PSNR
w/o F_o	0.946	58.676
w/o add	0.947	**58.690**
All	**0.951**	58.652

E. Efficiency Evaluation

The NVIDIA RTX 2080Ti GPU and the Intel i7-7700HQ CPU as our calculate power support are used to test traditional approaches and deep learning approaches, respectively. We can see that, among all approaches, our work is exceptionally effective (producing the third best result). We identify two causes for the given method's high time effectiveness. Operation efficiency of various approaches is offered in Table. VI. First, the backbone of our method is the collaborative aggregation module, which has two sub-modules: a non-local attention inference module and a local adaptive learning module. For the former, we use the Gaussian random matrix to reconstruct the non-local attention mechanism and use Sparse Aggregation to filter out irrelevant information, which all improve the efficiency. For the former, we also use a network structure similar to U-net, which requires a shorter amount of time than a single feedforward procedure. Second, our proposed method can perform unsupervised tasks, which further improves the time efficiency. However, the key to our method is to focus on extracting useful information from multiple feature maps,

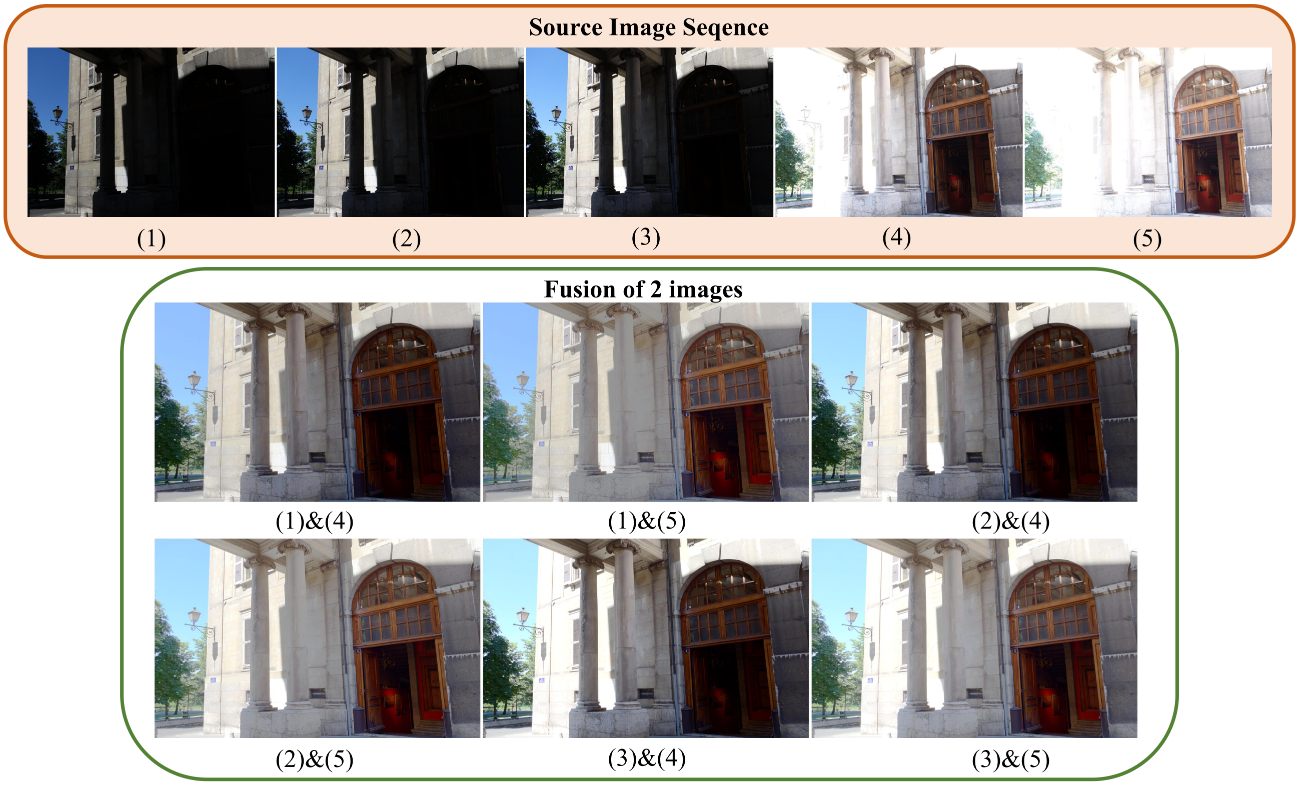

Fig. 11. Six visual results of our method generated by combing images with five exposure settings. Among them, (1)-(3) are under-exposed images, (4),(5) are over-exposed images.

TABLE VI
EFFICIENCY COMPARISON ON 100 TEST IMAGE PAIRS. THE RESULTS OF THREE MOST EFFICIENT METHODS ARE SHOW IN RED, BLUE AND BLACK(RED IS BEST, BLUE IS SECOND).

Methods	DEM	DSIFT	FMMEF	GFF	U2Fusion	DeepFuse	IFCNN	MEF-Net	CF-Net	AGAL	Ours
Platform	Matlab(CPU)	Matlab(CPU)	Matlab(CPU)	Matlab(CPU)	Tensorflow(GPU)	Tensorflow(GPU)	Pytorch(GPU)	Pytorch(GPU)	Pytorch(GPU)	Pytorch(GPU)	Pytorch(GPU)
Runtime	0.497	0.732	0.333	0.361	0.145	0.332	0.214	**0.023**	1.319	**0.066**	**0.118**
Parameters	-	-	-	-	0.659	0.018	0.083	0.026	2.897	1.591	1.597

causing a larger number of parameters than other approaches. MEF-Net keeps the efficiency by downsampling the images, and AGAL uses a relatively simple layer attention module to guarantee time efficiency. Our method combines both advantages, taking into account time efficiency and fusion quality.

V. CONCLUSIONS

We develop a global-local aggregation network for addressing multi-exposure image fusion tasks in an unsupervised way. Our work is composed of a collaborative aggregation module, a special fusion module and a recursive refinement module. The collaborative aggregation module formulates a complete feature extraction mechanism to learn global-local features from source images. The special fusion module is used to reconstruct fused results in terms of the characteristics of extracted information. The recursive refinement module further fine-tunes the fusion results by loop mechanism. Our module, whose each part is indispensable, significantly generates fused images with abundant details and balanced exposure. Extensive experiments on two common datasets are performed to prove the fusion capability of the proposed approach and highlight its competitiveness over ten other state-of-the-art methods.

REFERENCES

[1] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 267–276, 2002.

[2] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer graphics forum*, vol. 28, pp. 161–171, Wiley Online Library, 2009.

[3] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[4] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3634–3646, 2011.

[5] Q. Wang, W. Chen, X. Wu, and Z. Li, "Detail-enhanced multi-scale exposure fusion in yuv color space," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2418–2429, 2019.

[6] J. Li, J. Liu, S. Zhou, Q. Zhang, and N. K. Kasabov, "Learning a coordinated network for detail-refinement multi-exposure image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[7] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proceedings of the IEEE international conference on computer vision*, pp. 4714–4722, 2017.

[8] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3333–3348, 2020.

[9] S.-Y. Chen and Y.-Y. Chuang, "Deep exposure fusion with deghosting via homography estimation and attention learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1464–1468, IEEE, 2020.

[10] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index," *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 60–72, 2017.

[11] S. Hu and W. Zhang, "Exploiting patch-based correlation for ghost removal in exposure fusion," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1099–1104, IEEE, 2017.

[12] N. Hayat and M. Imran, "Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 295–308, 2019.

[13] J.-L. Yin, B.-H. Chen, Y.-T. Peng, and C.-C. Tsai, "Deep prior guided network for high-quality image fusion," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2020.

[14] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12484–12491, 2020.

[15] Z. Li, J. Liu, R. Liu, X. Fan, Z. Luo, and W. Gao, "Multiple task-oriented encoders for unified image fusion," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2021.

[16] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2761–2785, 2021.

[17] H. Li, K. Ma, H. Yong, and L. Zhang, "Fast multi-scale structural patch decomposition for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 5805–5816, 2020.

[18] S.-h. Lee, J. S. Park, and N. I. Cho, "A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1737–1741, IEEE, 2018.

[19] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5637–5646, 2022.

[20] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3517–3526, 2021.

[21] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16205–16215, 2021.

[22] Y. Yu, W. Yang, Y.-P. Tan, and A. C. Kot, "Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6013–6022, 2022.

[23] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16304–16313, 2022.

[24] Y. Cai, J. Lin, X. Hu, H. Wang, X. Yuan, Y. Zhang, R. Timofte, and L. Van Gool, "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17502–17511, 2022.

[25] Y. Liu and Z. Wang, "Dense sift for ghost-free multi-exposure fusion," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 208–224, 2015.

[26] H. Li, B. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graphical models and image processing*, vol. 57, no. 3, pp. 235–245, 1995.

[27] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 341–357, 2011.

[28] K. Ma and Z. Wang, "Multi-exposure image fusion: A patch-wise approach," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1717–1721, IEEE, 2015.

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

[30] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *1993 (4th) international Conference on Computer Vision*, pp. 173–182, IEEE, 1993.

[31] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pp. 382–390, IEEE, 2007.

[32] F. Kou, Z. Li, C. Wen, and W. Chen, "Multi-scale exposure fusion via gradient domain guided image filtering," in *2017 IEEE international conference on multimedia and expo (ICME)*, pp. 1105–1110, IEEE, 2017.

[33] W. Zhang, X. Liu, W. Wang, and Y. Zeng, "Multi-exposure image fusion based on wavelet transform," *International Journal of Advanced Robotic Systems*, vol. 15, no. 2, p. 1729881418768939, 2018.

[34] F. Kou, Z. Li, C. Wen, and W. Chen, "Edge-preserving smoothing pyramid based multi-scale exposure fusion," *Journal of Visual Communication and Image Representation*, vol. 53, pp. 235–244, 2018.

[35] Y. Zhang, T. Liu, M. Singh, E. Çetintaş, Y. Luo, Y. Rivenson, K. V. Larin, and A. Ozcan, "Neural network-based image reconstruction in swept-source optical coherence tomography using undersampled spectral data," *Light: Science & Applications*, vol. 10, no. 1, pp. 1–14, 2021.

[36] X. Li, G. Zhang, H. Qiao, F. Bao, Y. Deng, J. Wu, Y. He, J. Yun, X. Lin, H. Xie, *et al.*, "Unsupervised content-preserving transformation for optical microscopy," *Light: Science & Applications*, vol. 10, no. 1, pp. 1–11, 2021.

[37] N. K. Kalantari, R. Ramamoorthi, *et al.*, "Deep high dynamic range imaging of dynamic scenes.," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017.

[38] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[39] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[40] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2020.

[41] X. Deng, Y. Zhang, M. Xu, S. Gu, and Y. Duan, "Deep coupled feedback network for joint exposure fusion and image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 3098–3112, 2021.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

[43] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[44] H. Liu, T. Chen, Q. Shen, and Z. Ma, "Practical stacked non-local attention modules for image compression.," in *CVPR Workshops*, p. 0, 2019.

[45] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5690–5699, 2020.

[46] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," *arXiv preprint arXiv:2201.03794*, 2022.

[47] K. R. Prabhakar and R. V. Babu, "Ghosting-free multi-exposure image fusion in gradient domain," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1766–1770, IEEE, 2016.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[49] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "Vif-net: an unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[50] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

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

[52] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 2808–2819, 2019.