Focus Affinity Perception and Super-Resolution Embedding for Multifocus Image Fusion

Huafeng Li, Ming Yuan, Jinxing Li, *Member, IEEE*, Yu Liu, *Member, IEEE*, Guangming Lu, *Senior Member, IEEE*, Yong Xu, *Senior Member, IEEE*, Zhengtao Yu, and David Zhang, *Life Fellow, IEEE*

***Abstract*— Despite the fact that there is a remarkable achievement on multifocus image fusion, most of the existing methods only generate a low-resolution image if the given source images suffer from low resolution. Obviously, a naive strategy is to independently conduct image fusion and image super-resolution. However, this two-step approach would inevitably introduce and enlarge artifacts in the final result if the result from the first step meets artifacts. To address this problem, in this article, we propose a novel method to simultaneously achieve image fusion and super-resolution in one framework, avoiding step-by-step processing of fusion and super-resolution. Since a small receptive field can discriminate the focusing characteristics of pixels in detailed regions, while a large receptive field is more robust to pixels in smooth regions, a subnetwork is first proposed to compute the affinity of features under different types of receptive fields, efficiently increasing the discriminability of focused pixels. Simultaneously, in order to prevent from distortion, a gradient embedding-based super-resolution subnetwork is also proposed, in which the features from the shallow layer, the deep layer, and the gradient map are jointly taken into account, allowing us to get an upsampled image with high resolution. Compared with the existing methods, which implemented fusion and super-resolution independently, our proposed method directly achieves these two tasks in a parallel way, avoiding artifacts caused by the inferior output of image fusion or super-resolution. Experiments conducted on the real-world dataset substantiate the superiority of our proposed method compared with state of the arts.**

***Index Terms*— Affinity, image fusion, multifocus, receptive field, super-resolution.**

I. Introduction

DUE to the existence of depth of field (DOF) in imaging devices, it is intractable to enforce all pixels to fall in the DOF, subsequently generating partial-blur or partial-in-focus images. Fortunately, the multifocus image fusion technique [1], [2], [3] is proposed to detect and fuse complementary in-focus pixels from multiple source images, finally obtaining an all-in-focus image. Because of this fused image, more accurate image interpretation [4], [5] and recognition [6], [7] are allowed.

In recent years, various multifocus image fusion methods have been proposed, which can be roughly classified to three types: transform domain-based [8], [9], [10], spatial domain-based [1], [11], [12], [13], [14], and machine learning-based [15], [16], [17], [18], [19], [20], [21] methods. Transform domain-based methods first decompose the source images into multiple coefficients, which are then fused by following certain fusion rules. According to this fused coefficient, the all-in-focus image is finally obtained by exploiting the inverse transformation. Differently, spatial domain-based methods directly make focus-probability estimation in the pixel level or the block level, which then obtain fused image by combining focused pixels or blocks from source images. For the machine learning-based methods, a part of training samples is first selected to train a data-driven model, which aims to extract focus pixels-related features for image fusion. A typical strategy is the deep learning-based [22], [23], [24], [25] method [15], [19], which uses deep networks, e.g., convolution neural network (CNN) [26], to predict the probability of each pixel [15] or patch [19] that belongs to the focused or blur category. Alternatively, the fused result can also be generated by an end-to-end model, such as IFCNN [27], MFF-GAN [28], and ZMFF [29].

Despite the fact that a number of multifocus image fusion approaches have been studied, most of them are constraint on the fixed resolution, which is the same to the source images. It is true that if the source images enjoy the high resolution, the resolution of the fused image is fine. However, it would be far from our requirement if the source images only have low resolution. To tackle this problem, a straightforward way is to make super-resolution [30], [31], [32], [33], [34], [35] for the source images and then fuse them or conduct super-resolution for the fused but low-resolution image. However, if the result obtained from the first manner is corrupted from artifacts, these will then be enlarged or transmitted to the second step, subsequently making an inferior influence on the final result. By contrast, the study on a unit model for joint image fusion

Manuscript received 7 November 2021; revised 27 January 2023, 31 July 2023, and 9 December 2023; accepted 13 February 2024. This work was supported in part by the NSFC Fund under Grant 62276120 and Grant 62272133, in part by Shenzhen Colleges and Universities Stable Support Program under Grant GXWD20220811170100001, and in part by Yunnan Fundamental Research Projects under Grant 202301AV070004. *(Huafeng Li and Ming Yuan contributed equally to this work.) (Corresponding author: Jinxing Li.)*

Huafeng Li, Ming Yuan, and Zhengtao Yu are with Kunming University of Science and Technology, Kunming 650500, China (e-mail: hfchina99@163.com; 20192204225@stu.kust.edu.cn; ztyu@hotmail.com).

Jinxing Li, Guangming Lu, and Yong Xu are with Harbin Institute of Technology at Shenzhen, Shenzhen 518055, China, and also with Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen 518055, China (e-mail: lijinxing158@gmail.com; luguangm@hit.edu.cn; yongxu@ymail.com).

Yu Liu is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn).

David Zhang is with The Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China, and also with Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: davidzhang@cuhk.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2024.3367782, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2024.3367782

and super-resolution has attracted much attention in recent years [36], [37], [38]. For instance, Li et al. [38] introduced the meta learning for infrared and visible image fusion and super-resolution. Although these methods can be applied to multifocus image fusion, they are not specially designed for fusing multifocus images, ignoring the prior that the focus area of the multifocus image is clear, and the out-of-focus area is blurred, which leads to unsatisfied performance.

In this article, we construct a unit deep network, which is efficient of fusing image and enhancing spatial-resolution simultaneously. In particular, this network is composed of the focus affinity perception network (FAPN), super-resolution network (SRN), and fusion network (FN). FAPN is exploited for focus pixel detection, while SRN is designed to encourage the network to enjoy the capability of super-resolution. Jointly taking both outputs from FAPN and SRN into account, FN is finally introduced to achieve multifocus image fusion as well as spatial resolution enhancement. Note that, both FAPN and SRN take the source images as inputs, preventing from enlarging the artifacts of a semioutput as mentioned above.

The main contributions of our proposed method are concluded as follows.

1) A deep architecture is proposed for multifocus image fusion and super-resolution. To the best of our knowledge, this is the first work, which jointly takes these two tasks into an end-to-end network. This work contributes a novel way to researchers to conduct multifocus fusion if the source images have low resolution.
2) In FAPN, the affinity of features under different types of receptive fields is estimated, so that the focus pixels in both texture areas and smooth areas are more easily detected.
3) In SRN, the features from the shallow layer, the deep layer, and the gradient map are jointly considered, efficiently allowing our upsampled and fused image to follow our human visual perception.
4) Experimental results on the real-world dataset substantiate the effectiveness and superiority of our method in comparison with the existing state-of-the-art approaches on multifocus image fusion and super-resolution.

The rest of this article is organized as follows. In Section II, the related works, including multifocus image fusion and multisource image fusion, followed by super-resolution are briefly introduced. We then analyze the proposed method in detail in Section III. In Section IV, experiments are conducted to show the effectiveness of our proposed method, followed by the conclusion in Section V.

II. Related Works

It is a challenge task to jointly achieve multifocus image fusion and super-resolution in a unit model, which has not been studied in the existing works. In this section, we briefly describe some related works on multifocus image fusion, and joint image fusion and super-resolution, to make readers have a better understand on our proposed method.

A. Multifocus Image Fusion

In the last decades, both transform domain and spatial domain-based multifocus image fusion methods achieved much improvement. In 1983, Burt and Adelson [39] proposed a Laplacian pyramid-based approach, which is capable of preserving some texture details around the boundaries. Because of their contribution, a large number of transform domain-based strategies were subsequently studied, including wavelet transform [40], nonsubsampled contourlet transform (NSCT) [41], ridgelet transform [42], and curvelet transform [43]. Although these methods contributed to the presentation of details in edges or boundaries, they only adopted the fixed transformation, which is not adaptive for natural images due to their diversity. Thus, the dictionary learning-based methods [16], [44], [45] were then proposed, which can sparsely represent a natural image in an adaptive way. However, it is intractable to make sure which parts of the represented sparse coefficients are associated with in-focus pixels, resulting in that the reconstructed image may not be all-in-focus. By contrast, spatial domain-based methods [12], [13], [14] are more accessible of extracting focused pixels from the source images. However, this type of pixel-level or block-level strategies additionally introduces the image stitch tracer, which has an inferior influence on the fused image.

Because of the strong capability of representation, deep learning-based multifocus image fusion methods also have attracted much attention. Liu et al. [46] primarily proposed a CNN-based method to classify a patch from the source images to the focused or blur category. Instead of dividing images into overlapped patches, which costs much time, Li et al. [15] presented an end-to-end network, which takes the whole image as the input and efficiently reduces the time complexity. In addition, an unsupervised deep learning approach [47] was proposed to directly estimate the weight of different source images. Considering that all-in-focus images well contain gradient information, Ma et al. [48] embedded the gradient map into the network to exploit this prior. To solve the fusion issue near the focused/defocused boundary, Ma et al. [49] presented an α-matte boundary defocus model for multifocus image fusion. To reduce the number of parameters in the model, Nie et al. [50] proposed a lightweight multifocus image fusion method. Wang et al. [51] proposed a self-supervised learning model for multifocus image fusion, which frees the model from the dependence on the multifocus image training set. Liu et al. [3] developed a lightweight unified image fusion framework, achieving the fusion of multisource images, including multifocus images. Despite the fact that these methods achieved much better results compared with transform and spatial domain-based methods, they fail to take the low resolution, which does exist in natural images into account. Although an additional super-resolution processing followed by image fusion can tackle this problem, it would enlarge the artifacts, heavily resulting in the quality degradation of fused images.

B. Joint Image Fusion and Super-Resolution

To remove the constraint on low resolution of the source images, many researchers also tried to combine the image fusion and super-resolution into a joint model [36], [37], [38], [52]. According to the assumption that the fused image and the source inputs should have the same or similar structure tensor, Li et al. [36] proposed a fractional differential and variational algorithm in which the energy function for the image fusion and super-resolution is established by combining with the downsampling operator. Yin et al. [37] presented a sparse representation-based image fusion method, followed by the super-resolution. Specifically, it first interpolates and

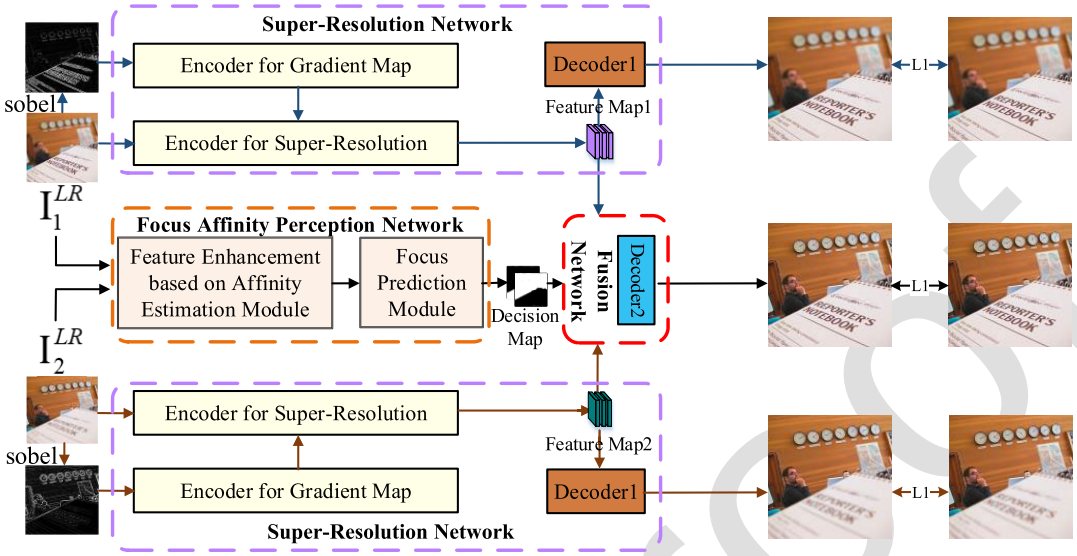


Fig. 1. Pipeline of the proposed method. Here, it is composed of SRN, FAPN, and FN, which is used for image super-resolution, in-focus pixel detection, and image fusion, respectively.

decomposes the low-resolution source images into high- and low-frequency components. These two kinds of components are then sparsely represented, fused, and reconstructed to get an image with high resolution. Besides, in [52], two pairs of dictionaries were introduced in which one pair is exploited to extract the low-rank and sparse components of low-resolution source images, and the other one is utilized for reconstructing the fused high-resolution image. Even though these methods achieve fusion and resolution improvement in a unit model, they were not typically designed for the multifocus image fusion task, ignoring the prior that the focus area of the multifocus image is clear, and the out-of-focus area is blurred. Images in multifocus image fusion enjoy their own specific characteristics, and naively applying aforementioned methods to this task is not always optimal. Thus, in this article, we present a novel method for multifocus image fusion by fully considering the characteristics of multifocus images. It is tailored for low-resolution multifocus image fusion, jointly achieving the focused pixel detection and super-resolution.

III. Proposed Method

The pipeline of our proposed method is shown in Fig. 1, which consists of FAPN, SRN, and FN. Without loss of the generality, two source images with low resolution are represented as $\mathbf{I}_1^{\mathrm{LR}} \in \mathbb{R}^{h\times w\times 3}$ and $\mathbf{I}_2^{\mathrm{LR}} \in \mathbb{R}^{h\times w\times 3}$, where h and w are their height and width, respectively. Because of FAPN, the affinity of features under different receptive fields is computed, which allows in-focus pixels that are more discriminatively classified. Simultaneously, in SRN, a source image and its gradient map are regarded as the input to upsample it to a high-resolution image. By combining these two blocks together, the high-resolution image, which is all-in-focus, is generated.

A. Focus Affinity Perception Network

1) Feature Enhancement Based on Affinity Estimation Module: It is obvious that the focused part contains more texture details compared with that of the defocused part. Thus, it is reasonable to classify which part is focused or defocused according to their comparison on texture and details. However, due to the diversity of natural images, it does exist the smooth but focused area in the source image, which does not have the characteristic mentioned above, increasing the difficulty of the in-focus pixel detection. In deep learning, for areas with rich edge details, a small receptive field is enough for judging the focusing characteristics of pixels. By contrast, since the convolutional kernel under a large size enjoys a larger receptive field, their associated features are more suitable for the identification of focus pixels from the smooth area located in the focused region. Because of these advantages of the feature maps generated by convolution kernels with different receptive fields, we jointly take them into account to make a focused feature enhancement.

Fig. 2 displays the insight of the feature enhancement structure, in which affinity estimation module (AEM), spatial attention module (SAM), and channel attention module (CAM) are included. As the receptive field is influenced by the sizes of convolutional kernels, here, two sizes of kernels, 3×3 and 5×5, are used for feature extraction. In other words, the 3×3 kernel is associated with features under the small receptive field, while the 5×5 kernel is associated with features under the large receptive field. Given a source image $\mathbf{I}_i^{\mathrm{LR}}, i\in\{1,2\}$, its feature maps associated with the 3×3 and 5×5 kernels can be represented as follows:

$$\mathbf{F}_{i,3}^{\mathrm{AEM}} = \mathrm{conv}(\mathbf{I}_i^{\mathrm{LR}}, k=3), \quad \mathbf{F}_{i,5}^{\mathrm{AEM}} = \mathrm{conv}(\mathbf{I}_i^{\mathrm{LR}}, k=5) \quad (1)$$

where "conv" denotes the convolutional processing and k means the size of the kernel. As shown in the left part of Fig. 2, the affinity map $\mathbf{W}_i^{\mathrm{AEM}}$ according to these two feature maps can be then computed by using element product and convolution blocks, followed by the active functions, including ReLU and sigmoid:

$$\mathbf{W}_i^{\mathrm{AEM}} = \mathrm{sigmoid}\big(\mathrm{conv}\big(\mathrm{ReLU}\big(\mathrm{conv}\big(\mathbf{F}_{i,3}^{\mathrm{AEM}} \odot \mathbf{F}_{i,5}^{\mathrm{AEM}}, k=3\big)\big), k=3\big)\big) \quad (2)$$

where $\odot$ denotes the Hadamard product. The affinity map $\mathbf{W}_i^{\mathrm{AEM}}$ implies the relationship between $\mathbf{F}_{i,3}^{\mathrm{AEM}}$ and $\mathbf{F}_{i,5}^{\mathrm{AEM}}$ at

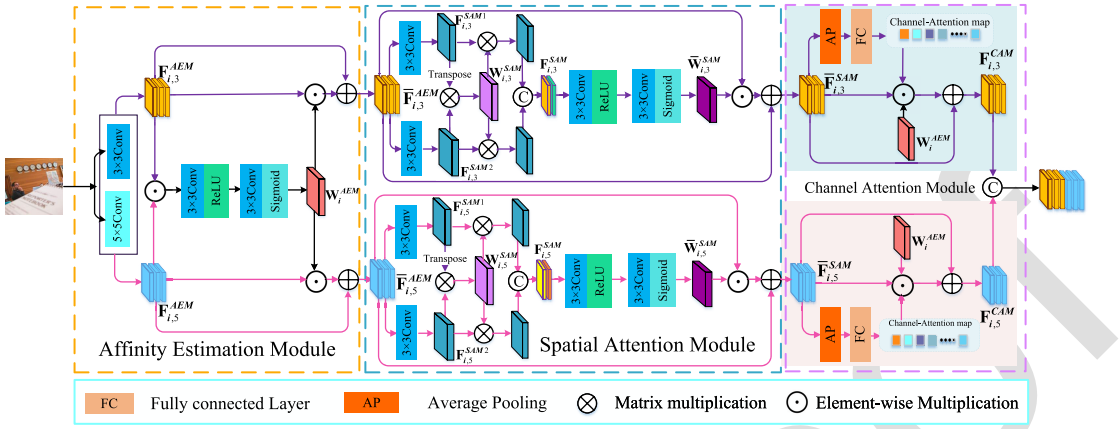


Fig. 2. Details of the feature enhancement based on affinity estimation, in which the source image is forwarded through AEM, SAM, and CAM. AEM is used for computing the affinity of feature maps obtained via different convolutional kernels; SAM is used for spatial correlation extraction; CAM is used for channel correlation extraction.

the same location. The larger the value of $\mathbf{W}_i^{\mathrm{AEM}}$, the higher the probability that the pixels at the same position in different feature maps come from the focused region and vice versa. From (2), we can also see that, if $\mathbf{F}_{i,3}^{\mathrm{AEM}}$ and $\mathbf{F}_{i,5}^{\mathrm{AEM}}$ do focus on the focused pixels, their corresponding components would gain large values, so that their discriminability is enhanced. Otherwise, if these components inversely encounter small values, they would be recorrected under the constraint of loss functions in (17). After this enhancement, features are transformed to

$$\begin{aligned}\bar{\mathbf{F}}_{i,3}^{\mathrm{AEM}} &= \mathbf{F}_{i,3}^{\mathrm{AEM}} + \mathbf{F}_{i,3}^{\mathrm{AEM}} \odot \mathbf{W}_i^{\mathrm{AEM}} \\ \bar{\mathbf{F}}_{i,5}^{\mathrm{AEM}} &= \mathbf{F}_{i,5}^{\mathrm{AEM}} + \mathbf{F}_{i,5}^{\mathrm{AEM}} \odot \mathbf{W}_i^{\mathrm{AEM}}. \end{aligned} \tag{3}$$

Since focused or blur pixels in a natural image also have the spatial correlation, we further propose a SAM for $\bar{\mathbf{F}}_{i,3}^{\mathrm{AEM}}$ and $\bar{\mathbf{F}}_{i,5}^{\mathrm{AEM}}$ to gain more discriminative features. As shown in Fig. 2, $\bar{\mathbf{F}}_{i,3}^{\mathrm{AEM}}$ and $\bar{\mathbf{F}}_{i,5}^{\mathrm{AEM}}$ are first forwarded into two different 3×3 convolution layers to, respectively, get ($\mathbf{F}_{i,3}^{\mathrm{SAM1}}$, $\mathbf{F}_{i,3}^{\mathrm{SAM2}}$) and ($\mathbf{F}_{i,5}^{\mathrm{SAM1}}$, $\mathbf{F}_{i,5}^{\mathrm{SAM2}}$). To achieve spatial improvement, $\mathbf{F}_{i,3}^{\mathrm{SAM1}}$/$\mathbf{F}_{i,5}^{\mathrm{SAM1}}$ are transposed to multiply $\mathbf{F}_{i,3}^{\mathrm{SAM2}}$/$\mathbf{F}_{i,5}^{\mathrm{SAM2}}$ to get the self-attention maps $\mathbf{W}_{i,3}^{\mathrm{SAM}} = (\mathbf{F}_{i,3}^{\mathrm{SAM1}})^T \mathbf{F}_{i,3}^{\mathrm{SAM2}}$ and $\mathbf{W}_{i,5}^{\mathrm{SAM}} = (\mathbf{F}_{i,5}^{\mathrm{SAM1}})^T \mathbf{F}_{i,5}^{\mathrm{SAM2}}$. Then, the improved features can be obtained through

$$\begin{aligned}\mathbf{F}_{i,3}^{\mathrm{SAM}} &= \mathrm{concat}\big(\mathbf{W}_{i,3}^{\mathrm{SAM}}\mathbf{F}_{i,3}^{\mathrm{SAM1}}, \mathbf{W}_{i,3}^{\mathrm{SAM}}\mathbf{F}_{i,3}^{\mathrm{SAM2}}\big) \\ \mathbf{F}_{i,5}^{\mathrm{SAM}} &= \mathrm{concat}\big(\mathbf{W}_{i,5}^{\mathrm{SAM}}\mathbf{F}_{i,5}^{\mathrm{SAM1}}, \mathbf{W}_{i,5}^{\mathrm{SAM}}\mathbf{F}_{i,5}^{\mathrm{SAM2}}\big). \end{aligned} \tag{4}$$

Followed by 3×3 Conv + ReLU + 3×3 Conv + sigmoid, the spatial attention maps $\bar{\mathbf{W}}_{i,3}^{\mathrm{SAM}}$ and $\bar{\mathbf{W}}_{i,5}^{\mathrm{SAM}}$ are computed, which are then used to get the final features from SAM:

$$\begin{aligned}\bar{\mathbf{F}}_{i,3}^{\mathrm{SAM}} &= \bar{\mathbf{W}}_{i,3}^{\mathrm{SAM}} \odot \bar{\mathbf{F}}_{i,3}^{\mathrm{AEM}} + \bar{\mathbf{F}}_{i,3}^{\mathrm{AEM}} \\ \bar{\mathbf{F}}_{i,5}^{\mathrm{SAM}} &= \bar{\mathbf{W}}_{i,5}^{\mathrm{SAM}} \odot \bar{\mathbf{F}}_{i,5}^{\mathrm{AEM}} + \bar{\mathbf{F}}_{i,5}^{\mathrm{AEM}}. \end{aligned} \tag{5}$$

Although SAM allows each feature to be enhanced in a spatial way, the channel relationship among different feature maps is ignored. Feature maps from different channels enjoy various discriminability. It is reasonable to make a channel-based enhancement, so that the defocused parts can be removed, while the focused parts get more discriminative characteristics. Therefore, as displayed in the right part of Fig. 2, a CAM is further proposed. This module contains the average pooling (AP) layer and the fully connected (FC) layer. By forwarding $\bar{\mathbf{F}}_{i,3}^{\mathrm{SAM}}$ or $\bar{\mathbf{F}}_{i,5}^{\mathrm{SAM}}$ through AP and FC, they are transformed via

$$\begin{aligned}\mathbf{F}_{i,3}^{\mathrm{CAM}} &= \bar{\mathbf{F}}_{i,3}^{\mathrm{SAM}} \odot \big(\mathrm{FC}\big(\mathrm{AP}\big(\bar{\mathbf{F}}_{i,3}^{\mathrm{SAM}}\big)\big)\big) \odot \mathbf{W}_i^{\mathrm{AEM}} + \bar{\mathbf{F}}_{i,3}^{\mathrm{SAM}} \\ \mathbf{F}_{i,5}^{\mathrm{CAM}} &= \bar{\mathbf{F}}_{i,5}^{\mathrm{SAM}} \odot \big(\mathrm{FC}\big(\mathrm{AP}\big(\bar{\mathbf{F}}_{i,5}^{\mathrm{SAM}}\big)\big)\big) \odot \mathbf{W}_i^{\mathrm{AEM}} + \bar{\mathbf{F}}_{i,5}^{\mathrm{SAM}}. \end{aligned} \tag{6}$$

Here, the attention map $\mathbf{W}_i^{\mathrm{AEM}}$ obtained from the AEM is also used to enforce each channel to enhance the focused pixels-related features.

By taking two source images into account, the final enhanced feature is obtained via the following equation:

$$\mathbf{F}^{\mathrm{EN}} = \mathrm{concat}\big(\mathbf{F}_{1,3}^{\mathrm{CAM}}, \mathbf{F}_{2,3}^{\mathrm{CAM}}, \mathbf{F}_{1,5}^{\mathrm{CAM}}, \mathbf{F}_{2,5}^{\mathrm{CAM}}\big). \tag{7}$$

2) Focus Prediction Module: After getting $\mathbf{F}^{\mathrm{EN}}$, which enhances the discriminability of focused pixels, the focus prediction module (FPM), as shown in Fig. 3, is then introduced to get the decision map for multifocus image fusion. Being like AEM, two types of convolution groups whose kernel sizes are 3×3 and 5×5, respectively, are used for feature extraction first. This local and global combination can further get an enhancement for the concatenated features from two source images. As shown in Fig. 3, denote these two convolution groups as conv_3^a and conv_5. $\mathbf{F}^{\mathrm{EN}}$ can be transformed to $\mathbf{F}_3^{\mathrm{FPM}}$ and $\mathbf{F}_5^{\mathrm{FPM}}$ via

$$\begin{aligned}\mathbf{F}_3^{\mathrm{FPM}} &= \mathrm{conv}_3(\mathbf{F}^{\mathrm{EN}}, k = 3) \\ \mathbf{F}_5^{\mathrm{FPM}} &= \mathrm{conv}_5(\mathbf{F}^{\mathrm{EN}}, k = 5). \end{aligned} \tag{8}$$

We then input the concatenation of $\mathbf{F}_3^{\mathrm{FPM}}$ and $\mathbf{F}_5^{\mathrm{FPM}}$ into a feature refine (FR) block to gradually refine the predictions. Specifically, this FR is composed of a 3×3 convolution layer, a ReLU function, a sigmoid function, and a batch norm (BN) layer. Here, the sigmoid function is introduced to encourage the output to fall in [0, 1], which is similar to the ground truth. According to this strategy, the learned feature from FR would be more adaptive for the final prediction. Followed by the BN layer, which allows the network to have a better convergence, the output from FR is added to $\mathbf{F}_3^{\mathrm{FPM}}$, being similar to the residual learning for more information preservation. Further taking the global feature $\mathbf{F}_5^{\mathrm{FPM}}$ into account, we can obtain the updated feature according to the following equation:

$$\bar{\mathbf{F}}_1^{\mathrm{FR}} = \mathrm{concat}\big(\mathbf{F}_1^{\mathrm{FR}} + \mathbf{F}_3^{\mathrm{FPM}}, \mathbf{F}_5^{\mathrm{FPM}}\big) \tag{9}$$

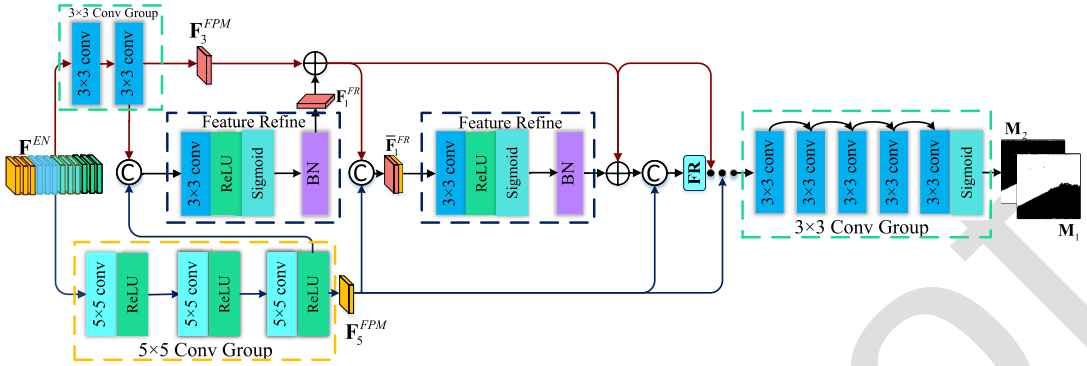


Fig. 3. Details of the FPM. Features are first input into two convolution groups with 3 × 3 and 5 × 5 kernels. Then, updated features are input into FR block to refine the features. Repeat FR for ten times, and the decision maps are finally obtained through another 3 × 3 convolution group followed by a sigmoid layer.

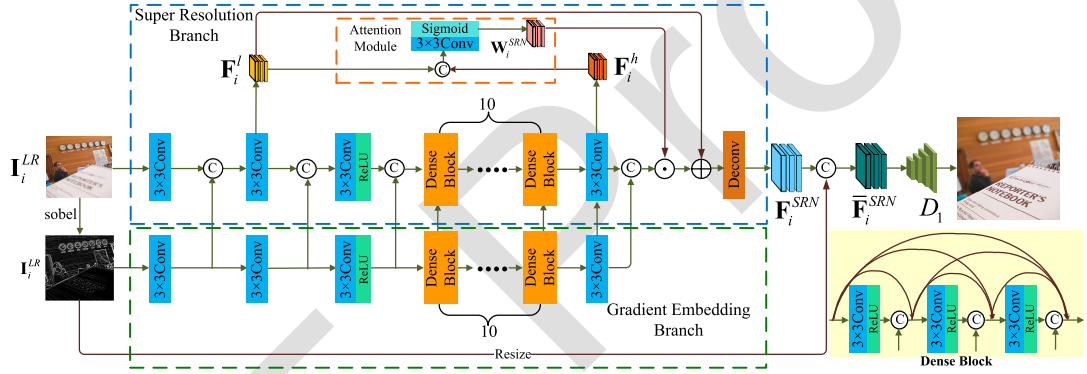


Fig. 4. Details of the SRN. It consist of the SRB and GEB, through which the low-level, high-level, and gradient features are jointly taken into account.

where $\mathbf{F}_1^{\mathrm{FR}}$ is the output from the BN in the first FR. As shown in Fig. 3, inputting $\bar{\mathbf{F}}_1^{\mathrm{FR}}$ into another FR, which is totally repeat for ten times, the binary predictions are finally computed by utilizing five 3 × 3 convolution layers and one sigmoid layer. In this article, we denote the predicted decision maps as $\mathbf{M}_1$ and $\mathbf{M}_2$.

B. Super-Resolution Network

From low resolution to high resolution, it is general that the reconstructed image would be smoother. In fact, our human visual perception is quite sensitive to the edge or texture details. The smooth recovery has a great inferior influence on performance. To preserve these details in the high-resolution images, the gradient map of the source image is additionally embedded into the network. Furthermore, a superiority of deep learning-based methods is to use the deep structure to learn the semantic or high-level information. Empirically, in the deep network, the top layers can enhance the global semantic contexts hidden in an in-focus object, while the shallow layers contribute to obtaining rich low-level features, which are quite adaptive for the sparse and texture details. Thus, here, a dense residual block is also proposed to jointly exploit high- and low-level information for super-resolution performance improvement.

The architecture of this SRN is shown in Fig. 4. In detail, output from the gradient embedding branch (GEB) is embedded into the outputs of layers in the super-resolution branch (SRB). Followed by ten dense residual blocks, the high-level feature $\mathbf{F}_i^h$ from the last dense residual block is combined with the low-level features $\mathbf{F}_i^l$ from the shallow layer to jointly learn an attention map $\mathbf{W}_i^{\mathrm{SRN}}$:

$$\mathbf{W}_i^{\mathrm{SRN}} = \mathrm{sigmoid}\big(\mathrm{conv}\big(\mathrm{concat}\big(\mathbf{F}_i^l, \mathbf{F}_i^h\big), k = 3\big)\big). \quad (10)$$

As shown in Fig. 4, by jointly taking the low-level feature $\mathbf{F}_i^l$, the high-level feature $\mathbf{F}_i^h$, the gradient feature $\nabla\mathbf{F}_i^h$, and the attention map $\mathbf{W}_i^{\mathrm{SRN}}$ into account, the feature associated with high-resolution images is recovered using a set of deconvolution layers

$$\mathbf{F}_i^{\mathrm{SRN}} = \mathrm{deconv}\big(\mathbf{W}_i^{\mathrm{SRN}} \odot \mathrm{concat}\big(\mathbf{F}_i^h, \nabla\mathbf{F}_i^h\big) + \mathbf{F}_i^l\big). \quad (11)$$

To make reconstructed high-resolution images gain more texture details, $\mathbf{F}_i^{\mathrm{SRN}}$ is further concatenated with the gradient map to be the final feature of the SRN

$$\bar{\mathbf{F}}_i^{\mathrm{SRN}} = \mathrm{concat}\big(\mathbf{F}_i^{\mathrm{SRN}}, \mathrm{Resize}\big(\nabla\mathbf{I}_i^{\mathrm{LR}}\big)\big) \quad (12)$$

where $\nabla\mathbf{I}_i^{\mathrm{LR}}$ is the gradient map of $\mathbf{I}_i^{\mathrm{LR}}$ and $\mathrm{Resize}(\nabla\mathbf{I}_i^{\mathrm{LR}})$ denotes that $\nabla\mathbf{I}_i^{\mathrm{LR}}$ is rescaled to the size, which is the same to the high-resolution image.

By using decoder D_1, the reconstructed images $\mathbf{I}_i^{\mathrm{HR}}$ with high resolution from the source image $\mathbf{I}_i^{\mathrm{LR}}$ are finally obtained

$$\mathbf{I}_i^{\mathrm{HR}} = D_1\big(\bar{\mathbf{F}}_i^{\mathrm{SRN}}\big). \quad (13)$$

C. Feature Fusion

Assume that the predicted binary masks from low-resolution source images as $\mathbf{M}_1$ and $\mathbf{M}_2$. Since it is inevitable to suffer

from some incorrect predictions, which are sparely located, the hole filling operation HF is exploited to finetune $\mathbf{M}_1$ and $\mathbf{M}_2$

$$\mathbf{M}_1^{\text{pred}} = \text{HF}(\mathbf{M}_1), \quad \mathbf{M}_2^{\text{pred}} = \text{HF}(\mathbf{M}_2). \tag{14}$$

Then, the fused feature map corresponding to the all-in-focus image is calculated via the following equation:

$$\mathbf{F}_{\text{fus}} = \text{Resize}\left(\mathbf{M}_1^{\text{pred}}\right) \odot \bar{\mathbf{F}}_1^{\text{SRN}} + \text{Resize}\left(\mathbf{M}_2^{\text{pred}}\right) \odot \bar{\mathbf{F}}_2^{\text{SRN}}. \tag{15}$$

Inputting $\mathbf{F}_{\text{fus}}$ into decoder D_2, the fused image with high resolution is finally obtained

$$\mathbf{I}_{\text{fus}}^{\text{HR}} = D_2(\mathbf{F}_{\text{fus}}). \tag{16}$$

D. Loss Functions

1) Loss Functions for FAPN: In this network, three types of loss functions, including the mask measurement L_m^{LR}, the gradient measurement $L_{\text{gd}}^{\text{LR}}$, and the structure similarity measurement $L_{\text{ssim}}^{\text{LR}}$, are jointly exploited to update the weights in the FAPN. In particular, denote the ground truth of decision maps from two low-resolution source images as $\mathbf{M}_1^{\text{GT}}$ and $\mathbf{M}_2^{\text{GT}}$, the ground truth of the all-in-focus source image with low resolution as $\mathbf{I}_{\text{GT}}^{\text{LR}}$, and the fused low-resolution image as $\mathbf{I}_{\text{fus}}^{\text{LR}} = \mathbf{M}_1 \odot \mathbf{I}_1^{\text{LR}} + \mathbf{M}_2 \odot \mathbf{I}_2^{\text{LR}}$. L_m^{LR}, $L_{\text{gd}}^{\text{LR}}$, and $L_{\text{ssim}}^{\text{LR}}$ can be, respectively, represented as follows:

$$\begin{aligned} L_m^{\text{LR}} &= \left\|\mathbf{M}_1 - \mathbf{M}_1^{\text{GT}}\right\|_1 + \left\|\mathbf{M}_2 - \mathbf{M}_2^{\text{GT}}\right\|_1 \\ L_{\text{gd}}^{\text{LR}} &= \left\|\nabla \mathbf{I}_{\text{fus}}^{\text{LR}} - \nabla \mathbf{I}_{\text{GT}}^{\text{LR}}\right\|_1 \\ L_{\text{ssim}}^{\text{LR}} &= 1 - \text{SSIM}\left(\mathbf{I}_{\text{fus}}^{\text{LR}}, \mathbf{I}_{\text{GT}}^{\text{LR}}\right) \end{aligned} \tag{17}$$

where SSIM means the structural similarity index measure (SSIM) [53] and $\nabla \mathbf{I}_{\text{fus}}^{\text{LR}}/\nabla \mathbf{I}_{\text{GT}}^{\text{LR}}$ is the gradient of $\mathbf{I}_{\text{fus}}^{\text{LR}}/\mathbf{I}_{\text{GT}}^{\text{LR}}$.

By weighted combining these three losses together, the final loss for FAPN is

$$L_{\text{FAPN}} = \lambda_1 L_m^{\text{LR}} + \lambda_2 L_{\text{gd}}^{\text{LR}} + \lambda_3 L_{\text{ssim}}^{\text{LR}} \tag{18}$$

where λ_1, λ_2, and λ_3 are nonnegative parameters. According to [15], we set them to 0.8, 0.1, and 0.1, respectively.

2) Loss Functions for SRN: Let the ground truth of $\mathbf{I}_1^{\text{HR}}$ and $\mathbf{I}_2^{\text{HR}}$ be $\mathbf{I}_{\text{GT1}}^{\text{HR}}$ and $\mathbf{I}_{\text{GT2}}^{\text{HR}}$, respectively. Then, the generated high-resolution images $\mathbf{I}_1^{\text{HR}}$ and $\mathbf{I}_2^{\text{HR}}$ can be measured through

$$L^{\text{HR}} = \left\|\mathbf{I}_1^{\text{HR}} - \mathbf{I}_{\text{GT1}}^{\text{HR}}\right\|_1 + \left\|\mathbf{I}_2^{\text{HR}} - \mathbf{I}_{\text{GT2}}^{\text{HR}}\right\|_1. \tag{19}$$

In addition, to encourage the focused information to be transported from source images to the fused image, we further introduce $L_{\text{fus}}^{\text{HR}}$ to optimize the decoder D_2

$$\begin{aligned} L_{\text{fus}}^{\text{HR}} = &\left\|\mathbf{I}_{\text{fus}}^{\text{HR}} - \mathbf{I}_{\text{GT}}^{\text{HR}}\right\|_1 + \left\|\text{Resize}\left(\mathbf{M}_1^{\text{pred}}\right) \odot \left(\mathbf{I}_{\text{fus}}^{\text{HR}} - \mathbf{I}_{\text{GT1}}^{\text{HR}}\right)\right\|_1 \\ &+ \left\|\text{Resize}\left(\mathbf{M}_2^{\text{pred}}\right) \odot \left(\mathbf{I}_{\text{fus}}^{\text{HR}} - \mathbf{I}_{\text{GT2}}^{\text{HR}}\right)\right\|_1 \end{aligned} \tag{20}$$

where $\mathbf{I}_{\text{GT}}^{\text{HR}}$ is the all-in-focus source image with high resolution. Here, the last two terms in (20) are used to make the network further focus on the clear pixels.

E. Training Strategy

In our training phase, L_{FAPN} is first exploited for training FAPN. We then fix weights in it and train the SRN by utilizing L^{HR} to enforce SRN to gain the capability of spatial resolution enhancement. Finally, $L_{\text{fus}}^{\text{HR}}$ is introduced to jointly train SRN and D_2, so that D_2 is capable of getting the all-in-focus image with high resolution.

IV. Experiments

To quantitatively substantiate the effectiveness of the proposed method, experiments are conducted in this section. The generation of the training set, implementation details, and evaluation metrics is first described. Then, the comparisons with the state-of-the-arts methods as well as the ablation study are analyzed.

A. Training Set

For the training of FAPN, we generate the training set through a synthetic way. Specifically, being like [15], 200 all-in-focus images are first collected from ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) [54]. For each raw image, nine subimages whose sizes are 128×128 are randomly cropped. Denoting a 128×128 image as $\mathbf{I}$, a pair of multifocus images are then obtained through

$$\begin{aligned} \mathbf{I}_1 &= \mathbf{M} \odot \mathbf{I} + (1 - \mathbf{M}) \odot f_{\text{blur}}(\mathbf{I}) \\ \mathbf{I}_2 &= \mathbf{M} \odot f_{\text{blur}}(\mathbf{I}) + (1 - \mathbf{M}) \odot \mathbf{I} \end{aligned} \tag{21}$$

where $\mathbf{M}$ is the randomly generated binary decision by using the "findContours" function in OpenCV and f_{blur} is the blur function. For the training of SRN, being like the existing super-resolution methods, we regard the DIV2k set [55] as the training set, in which each raw image is separated into nine 128×128 subimages. According to (21), the multifocus images are also generated. For the testing, being the same to the existing multifocus image fusion approaches, the Lytro dataset [44] is used for quantitative and qualitative evaluation. Note that, due to the limited space, in this article, we only visualize the results of five pairs from the Lytro dataset.

B. Implementation Details

In our train stage, the Adam optimizer is adopted for the model optimization. For the training of FAPN, four pairs of multifocus images are used as the inputs in each iteration. Specifically, the maximum epoch number is set to 500, and the learning rate is primarily set to 10^{-4}, which is then reduced to 5×10^{-5} after 200 epochs. For the SRN training, the batch size is also set to four pairs of multifocus images with the size of $128/t \times 128/t$ (where t is the scaling factor). For the training rate, it is reduced from 10^{-3} to 5×10^{-4} after 400 epochs, and the maximum epoch is 500. After achieving the pretrained models of FAPN and SRN, D_2 and SRN are jointly fine-tuned with 500 epochs, where the learning rate is set to 5×10^{-4}.

C. Evaluation Metrics

In this article, six metrics, which have been widely used in the existing methods, are adopted for quantitative evaluation, including normalized mutual information Q_{MI} [56], nonlinear correlation information entropy Q_{NICE} [57], gradient-based metric Q_G [58], phase congruency-based metric Q_P [59], Piella's metric Q_S [60], and Yang's metric Q_Y [61], [62]. Specifically, Q_{MI} is used to measure how much information is obtained from the source images; Q_{NICE} exploits the nonlinear entropy function to measure the correlation between the fused image and source images; Q_G is used to measure how much gradient-related information is obtained from the source

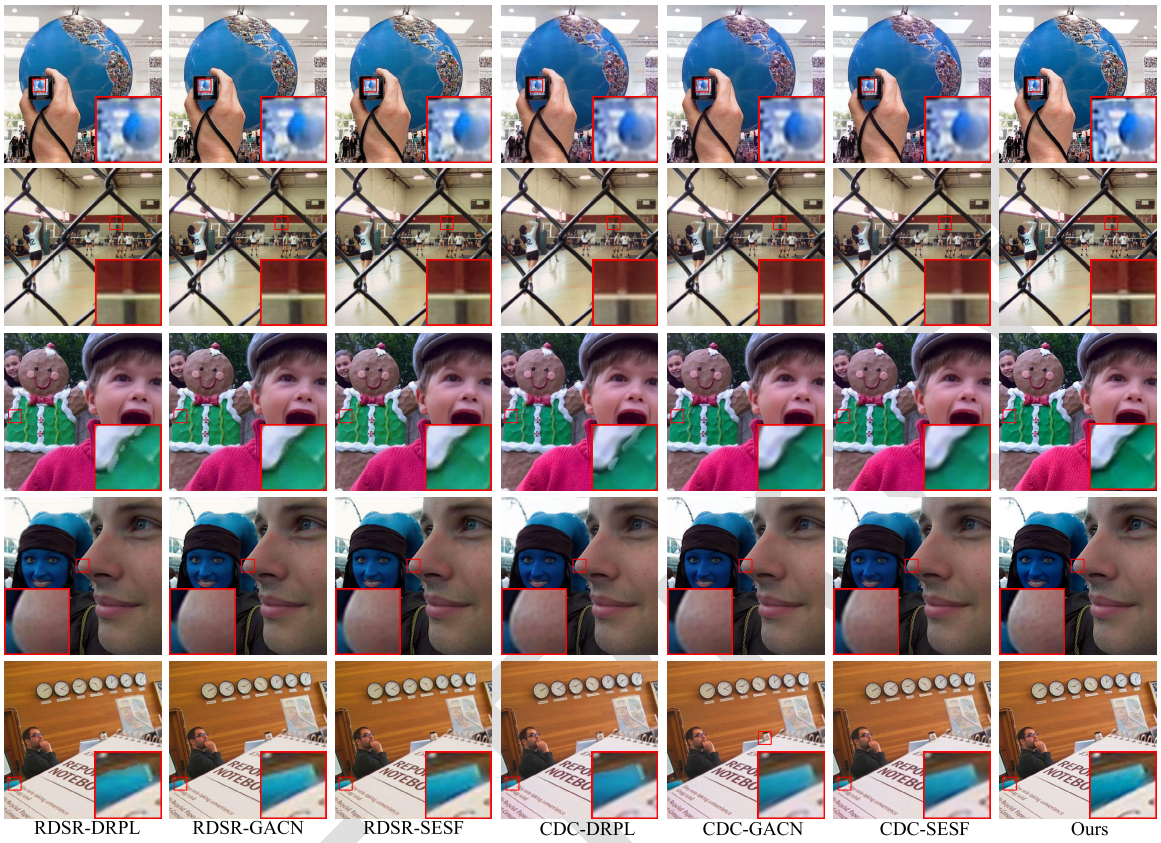


Fig. 5. Images generated by RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, CDC-SESF, and our proposed method when the scaling factor is set to 2×.

images; Q_P is used to compare the local cross correlation of corresponding feature maps of input images and the fused output; Q_S is used to measure how much saliency information is transported from the source images; Q_Y is used to measure the similarity of the local region from inputs and the output according to different strategies. For all these six metrics, the large the value is, the better the performance is. Note that, since these metrics are primarily designed for gray images, here, we, respectively, do the measurement based on the RGB three channels and only display their averaged value.

D. Comparisons With State of the Arts

To substantiate the outstanding performance of our proposed method, extensive comparisons between it and existing state of the arts are conducted in this section. Since the existing multifocus image fusion methods fail to jointly consider the image super-resolution, we additionally use the super-resolution methods for image upsampling. Except of these two-step combinations, we also make comparisons with the existing multisource image fusion to further show that our proposed method is particularly adaptive for multifocus image fusion.

Directly applying image super-resolution methods followed by multifocus image fusion methods is the most straightforward strategy for image super-resolution and fusion. Here, we select residual dense super-resolution (RDSR) network [63] and component divide-and-conquer (CDC) network [64] as the super-resolution methods. Simultaneously, the most state-of-the-art multifocus image fusion methods, such as deep regression pair learning (DRPL) [15], SESF-Fuse (SESF) [65], and gradient-aware cascade net (GACN) [48], are applied to achieve the fusion task. According to different combinations, the comparison approaches can be represented as RDSR-DRPL, RDSR-SESF, RDSR-GACN, CDC-DRPL, CDC-SESF, and CDC-GACN. Note that, the scaling factors for all images are set to 2×, 3×, and 4×, and their corresponding results are displayed in Figs. 5–7, respectively. It is true that aforementioned comparison methods do achieve super-resolution and fusion tasks, while their obtained images are remarkably inferior to that computed by our method. In particular, when the scaling factor is set to 4×, images achieved by RDSR-DRPL, RDSR-GACN, and RDSR-SESF not only suffer from the color's degradation, being far different from that of source images, but also are too smooth, which indicates that the texture details are missed. Referring to CDC-DRPL, CDC-GACN, and CDC-SESF, although they outperform RDSR-DRPL, RDSR-GACN, and RDSR-SESF in visualization, the details of the reconstructed images are still far below compared with that gained by our method, further substantiating that the joint learning for super-resolution and fusion is much more adaptive than the two-step strategies.

In contrast to RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, and CDC-SESF, Table I lists the averaged quantitative values on the Lytro dataset. It is easy to observe that our presented approach gains the most outstanding performance on all metrics. Referring to Q_G, Q_P, Q_Y, and Q_{MI}, there are remarkable enhancements compared with these six comparison methods. For instance, where the scaling factor is 2×, our proposed method gets as high as 0.4757, 0.5605, 0.8451, and 0.7849, while the best results obtained by comparison methods are only 0.4128, 0.4796,

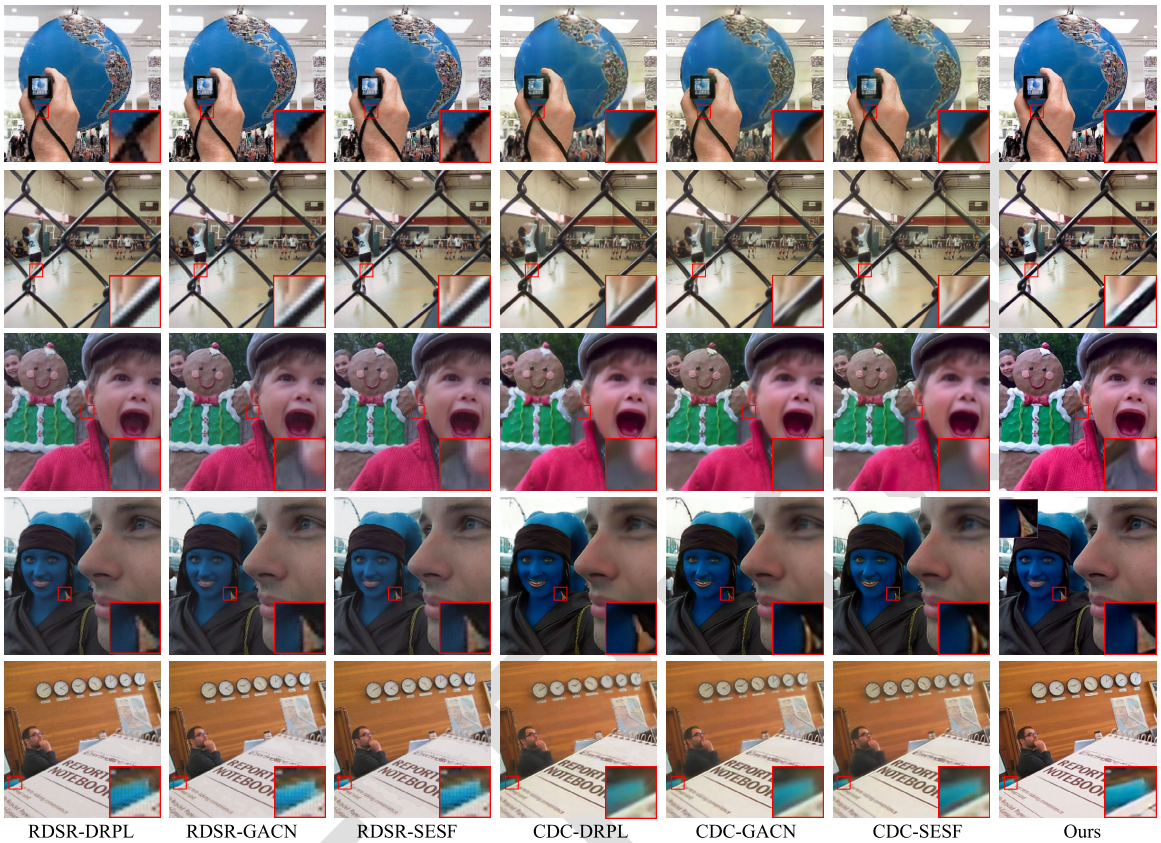


Fig. 6. Images generated by RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, CDC-SESF, and our proposed method when the scaling factor is set to 3×. The magnified area at the top-left corner of the last column is the ground truth obtained from the focus area of the source image with the same resolution. If there is only one magnified area in the fusion result of the proposed method, it means that this area is located in the junction of focus and defocus pixels in the source image, and there is no all-in-focus ground truth.

0.7645, and 0.7442, being much inferior. For Q_{NICE} and Q_S, our presented method also achieves more or less improvement. Besides, when the fusion is first conducted followed by super-resolution, our proposed approach continues achieving our superiority compared with these two-step methods. Due to the limitation of the space, please find the fused images as well as their quantitative evaluations in the Supplementary Material.

To further prove the effectiveness of the proposed method, experiments based on end-to-end fusion model combined with super-resolution model are performed in this section. Specifically, an end-to-end multifocus image fusion model called MFFGAN [28] is combined with RDSR, CDC, and DIP flow-based kernel prior USRNet (DFUNet) [66]. As shown in Figs. 8–10, our proposed method performs better than compared methods, which indicates that our method is also superior to the end-to-end fusion method followed with super-resolution models. The objective evaluation data listed in Table I confirm this conclusion.

E. Ablation Study

In this article, the affinity estimation (FAPN), shallow- and deep-feature combination (SDFC), and GEB are jointly exploited for multifocus image fusion and super-resolution. To demonstrate the effectiveness of each one, they are first removed from our proposed method, and this version is denoted as "baseline." We then add them into "baseline" one by one, which are denoted as "baseline + FAPN," "baseline + FAPN + SDFC," and "baseline + FAPN + SDFC + GEB," respectively. Moreover, we additionally substantiate the significance of joint multifocus image fusion and super-resolution learning, but not two separative steps. Note that here the scaling factor is set to 2× in all experiments.

1) Effectiveness of Affinity Perception: To efficiently detect the focused pixels, FAPN is presented to enhance the discriminability between the focused and defocused parts. As shown in Fig. 11, it is obvious that "baseline + FAPN" gains much better visualization compared with "baseline," indicating the significance of the affinity perception. Referring to their quantitative comparison, as listed in Table II, "baseline + FAPN" also achieves improvement on the most of metrics.

2) Effectiveness of SDFC: To exploit both texture and context information for image super-resolution, the shallow features are combined with deep features, followed by an attention computation. From Fig. 11, we can see that SDFC further contributes to the image performance improvement. Specifically, the enlarged details on the third column enjoy clearer visualization. Furthermore, Table II also indicates the agreement of the comparison result mentioned above.

3) Effectiveness of Gradient Embedding: To prevent from the upsampled image to be smooth, GEB is designed to embed the gradient map into the network for texture detail preservation. Fig. 11 and Table II experimentally substantiate the importance of GEB, in which the fused images attain better visualization and higher quantitative values on the most of metrics.

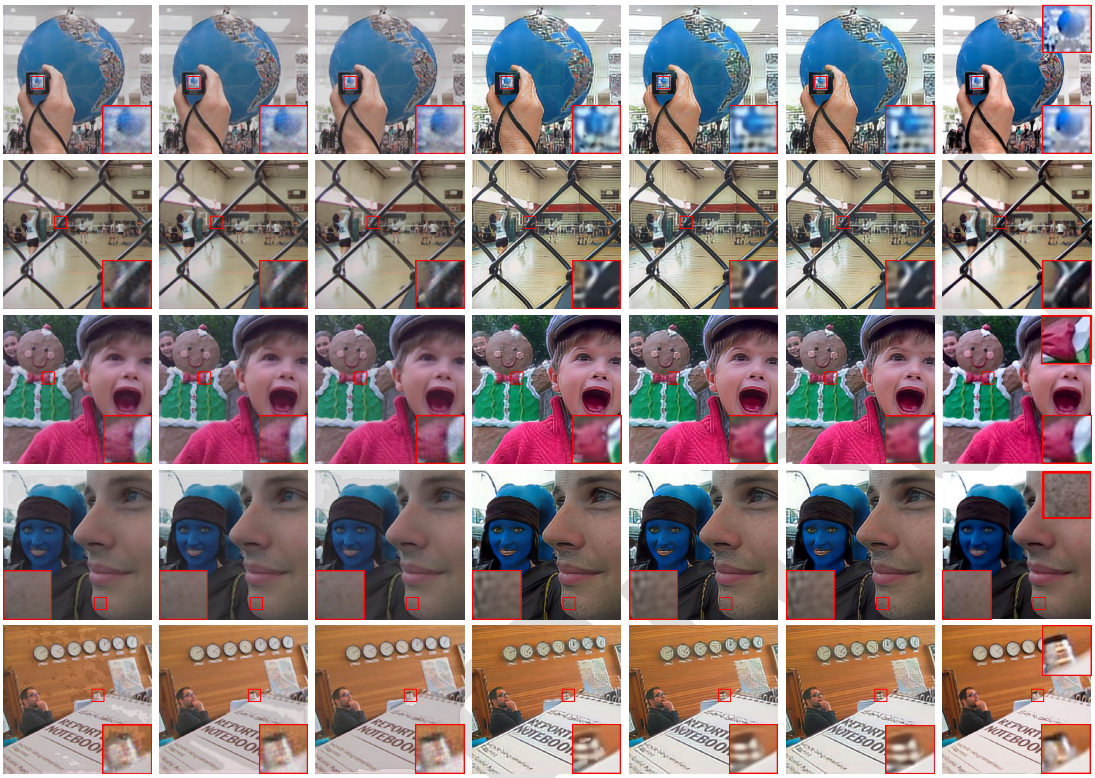


Fig. 7. Images generated by RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, CDC-SESF, and our proposed method when the scaling factor is set to 4×. The magnified area at the top-right corner of the last column is the ground truth obtained from the focus area of the source image with the same resolution. If there is only one magnified area in the fusion result of the proposed method, it means that this area is located in the junction of focus and defocus pixels in the source image, and there is no all-in-focus ground truth.

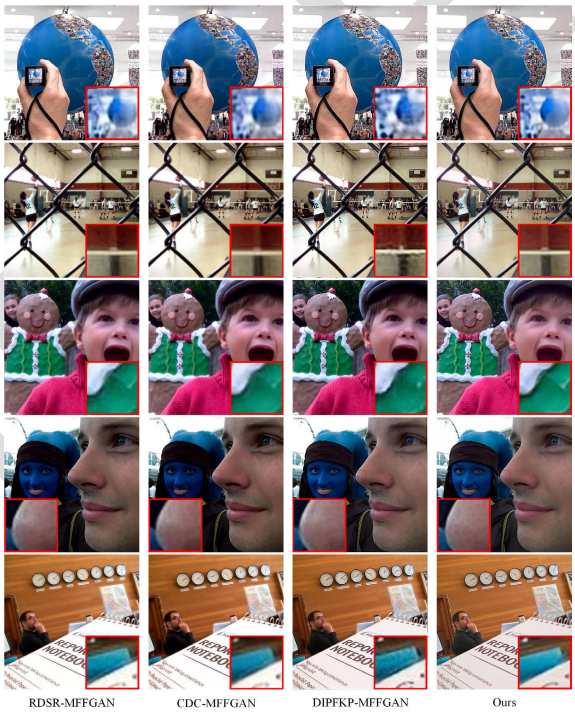


Fig. 8. Fusion results generated by the super-resolution followed by end-to-end fusion method when the scaling factor is set to 2×.

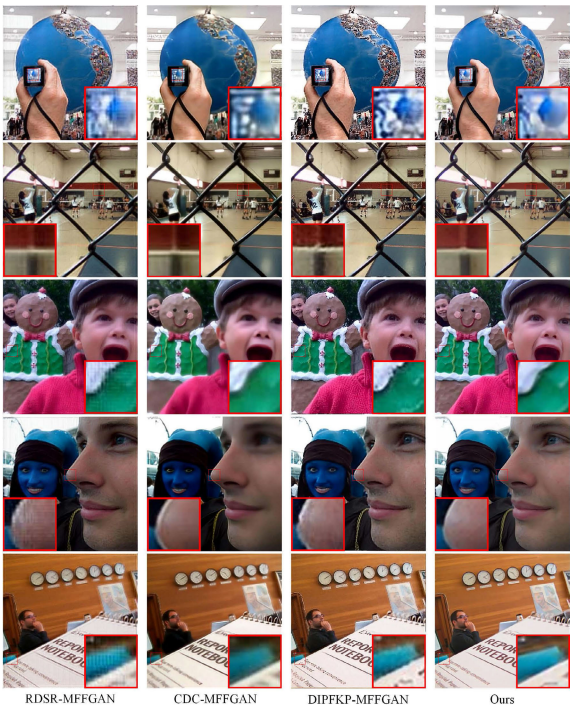


Fig. 9. Fusion results generated by the super-resolution followed by end-to-end fusion method when the scaling factor is set to 3×.

4) Effectiveness of Joint Learning: The more remarkable contribution of our method is primarily to achieve multifocus image fusion and super-resolution in a joint way. To further prove the superiority of this joint learning, we also conduct experiments in which the image fusion and

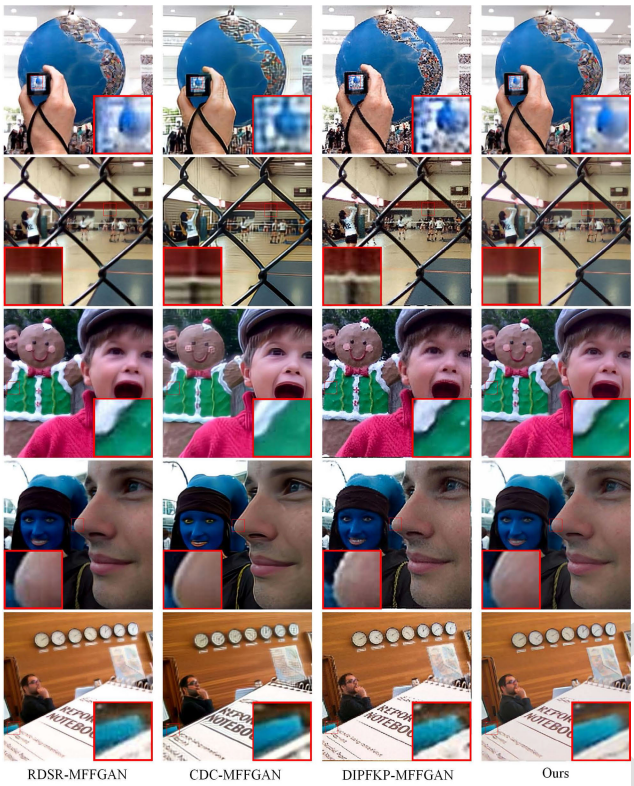


Fig. 10. Fusion results generated by the super-resolution followed by end-to-end fusion method when the scaling factor is set to 4×.

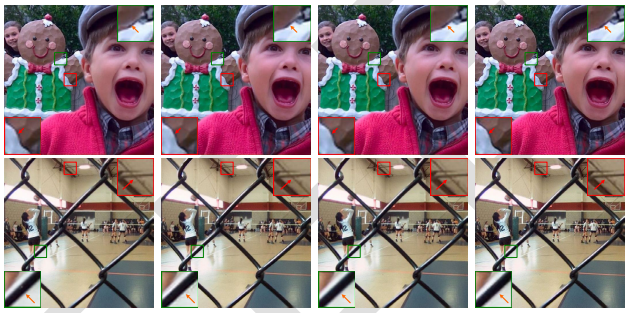


Fig. 11. Images generated by "baseline," "baseline + FAPN," "baseline + FAPN + SDFC," and "baseline + FAPN + SDFC + GEB," respectively (from left to right).

image super-resolution are individually implemented. Here, let "SR + Fusion" denote that multiple images are first upsampled through our SRN, which are then followed by our image FN. Of course, "Fusion + SR" means the inverse implementation. Obviously, results displayed in Fig. 12 demonstrate the effectiveness of our joint learning strategy. As we can see, images fused by our method can preserve much better texture details, while these computed by "SR + Fusion" and "Fusion + SR" suffer from more or less blur. Besides, Table III also lists six types of metrics associated with Fig. 12, further demonstrating the superiority of our presented approach.

5) Effectiveness of SAM Followed by CAM: Attention can highlight the discrimination features in focus detection and improve the quality of the final fusion result. To prove the effectiveness of the SAM and CAM, we sequentially remove them from the proposed method. As shown in Fig. 13, after

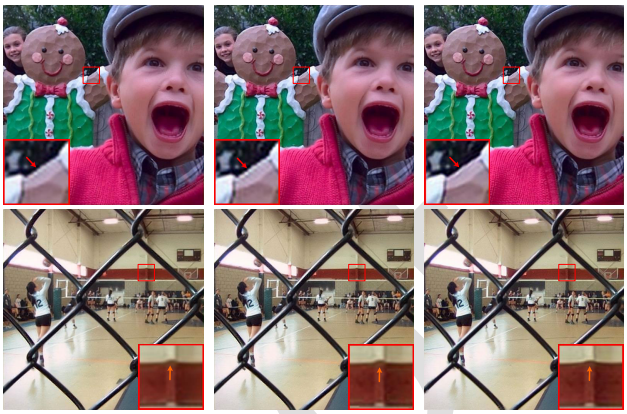


Fig. 12. Images generated by "SR + Fusion," "Fusion + SR," and our proposed method, respectively (from left to right).

Fig. 13. Effectiveness of attention. From left to right, the results after removing SAM and CAM, the results after removing only the SAM, the results after removing only the CAM, the results when the CAM is followed by the SAM, and the results of our method.

removing the SAM and CAM, some information from the out-of-focus area is introduced into the fusion result. When only attention module is removed, the fusion result is relatively better. If both SAM and CAM are introduced and CAM is followed by SAM, the fusion performance only meets slight improvement in the objective evaluation (as listed in Table IV). By contrast, both the visual quality and the objective evaluation of the fusion results generated by our proposed method are the best, demonstrating the effectiveness of SAM followed by CAM.

F. Discussion on the Effectiveness of Comparison Methods

Since this article is the first one to propose a joint framework for multifocus image fusion and super-resolution, we adopt the most common step-by-step approaches to obtain high-resolution fusion results for performance comparison. In this article, we select RDSR [63] and CDC [64] as the super-resolution methods, and DRPL [15], SESF [65], and GACN [48] as the fusion methods. Then, the methods RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, and CDC-SESF are constructed for experimental comparisons. Thus, it is necessary to prove that these compared methods also have excellent performance. To this end, we additionally use the classic bicubic interpolation and the current DFUNet [66] as super-resolution methods. They are then combined with DRPL, CDC, and SESF to generate (Bic-DRPL, Bic-GACN, and Bic-SESF) and (DFUNet-DRPL, DFUNet-GACN, and DFUNet-SESF), respectively.

As shown in Fig. 14 and Table V, the comparison methods based on current super-resolution approaches (e.g., DFUNet,

TABLE I

QUANTITATIVE AVERAGE VALUES OBTAINED RDSR-DRPL, RDSR-GACN, RDSR-SESF, CDC-DRPL, CDC-GACN, CDC-SESF, RDSR-MFFGAN, CDC-MFFGAN, DFUNET-MFFGAN, AND OUR METHOD UNDER DIFFERENT SCALING FACTORS

Task	Method	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
× 2	RDSR-DRPL	0.8209	0.4039	0.4608	0.8776	0.7478	0.7417
	RDSR-GACN	0.8210	0.4128	0.4796	0.8861	0.7645	0.7442
	RDSR-SESF	0.8210	0.4121	0.4794	0.8853	0.7641	0.7436
	CDC-DRPL	0.8203	0.3546	0.4002	0.8414	0.6990	0.7271
	CDC-GACN	0.8203	0.3615	0.4277	0.8510	0.7138	0.7286
	CDC-SESF	0.8203	0.3614	0.4263	0.8508	0.7140	0.7282
	RDSR-MFFGAN	0.8196	0.3998	0.4448	0.8354	0.6828	0.7194
	CDC-MFFGAN	0.8190	0.3529	0.3802	0.8054	0.7899	0.6941
	DFUNet-MFFGAN	0.8154	0.2274	0.1392	0.6839	0.4760	0.5749
	Ours	**0.8226**	**0.4757**	**0.5605**	**0.8998**	**0.8451**	**0.7849**
× 3	RDSR-DRPL	0.8172	0.2462	0.1961	0.7373	0.5124	0.6355
	RDSR-GACN	0.8172	0.2476	0.2029	0.7433	0.5151	0.6348
	RDSR-SESF	0.8171	0.2482	0.2026	0.7438	0.5168	0.6340
	CDC-DRPL	0.8162	0.2205	0.1367	0.6939	0.5131	0.6053
	CDC-GACN	0.8159	0.2185	0.1476	0.6985	0.5140	0.5959
	CDC-SESF	0.8158	0.2209	0.1454	0.6985	0.5179	0.5928
	RDSR-MFFGAN	0.8160	0.2410	0.1862	0.6969	0.5212	0.5993
	CDC-MFFGAN	0.8164	0.2397	0.1485	0.6741	0.6361	0.6124
	DFUNet-MFFGAN	0.4909	0.8133	0.1837	0.0593	0.5428	0.3360
	Ours	**0.8211**	**0.3285**	**0.2952**	**0.8222**	**0.6808**	**0.7454**
× 4	RDSR-DRPL	0.8155	0.2117	0.1174	0.6600	0.4293	0.5769
	RDSR-GACN	0.8154	0.2117	0.1181	0.6622	0.4272	0.5720
	RDSR-SESF	0.8153	0.2116	0.1182	0.6619	0.4272	0.5709
	CDC-DRPL	0.8134	0.2039	0.0826	0.6094	0.4103	0.4972
	CDC-GACN	0.8132	0.2036	0.0858	0.6091	0.4101	0.4897
	CDC-SESF	0.8131	0.2035	0.0843	0.6077	0.4085	0.4871
	RDSR-MFFGAN	0.8185	0.2671	0.1796	0.7124	0.5852	0.4999
	CDC-MFFGAN	0.8112	0.1748	0.0340	0.4100	0.4428	0.3973
	DFUNet-MFFGAN	0.8120	0.1602	0.0320	0.4495	0.2402	0.4315
	Ours	**0.8193**	**0.2718**	**0.1947**	**0.7627**	**0.5952**	**0.6935**

TABLE II

QUANTITATIVE VALUES OBTAINED "BASELINE," "BASELINE + FAPN," "BASELINE + FAPN + SDFC," AND "BASELINE + FAPN + SDFC + GEB," RESPECTIVELY

Task	Method	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
Ly03	Baseline	0.8207	0.4833	0.5481	0.8929	0.8361	0.7514
	Baseline+FAPN	0.8205	0.4915	0.5592	0.8930	0.8536	0.7478
	Baseline+FAPN+SDFC	0.8209	0.4946	0.5715	0.8968	0.8554	0.7576
	Baseline+FAPN+SDFC+GEB	**0.8213**	**0.5154**	**0.6068**	**0.9028**	**0.8828**	**0.7672**
Ly05	Baseline	0.8243	0.4251	0.5700	0.8819	0.7994	0.8113
	Baseline+FAPN	0.8243	0.4185	0.5795	0.8819	0.8029	0.8084
	Baseline+FAPN+SDFC	**0.8249**	0.4254	0.5827	0.8857	0.8085	**0.8206**
	Baseline+FAPN+SDFC+GEB	0.8245	**0.4510**	**0.6025**	**0.8866**	**0.8402**	0.8121

RDSR, and CDC) are superior to Bic-DRPL, Bic-GACN, and Bic-SESF. In addition, we also verify the superiority of DRPL, GACN, and SESF. To this end, we choose U2Fuison [67], which is an unsupervised image fusion method, as the baseline

Fig. 14. Effectiveness of compared methods. For each of "Ly03" and "Ly05," the first row from left to right shows the results generated, respectively, by Bic-DRPL, Bic-GACN, and Bic-SESF; the second row shows the results generated by DFUNet-DRPL, DFUNet-GACN, and DFUNet-SESF; the third row shows the results generated by RDSR-DRPL, RDSR-GACN, and RDSR-SESF; the fourth row shows the results generated by CDC-DRPL, CDC-GACN, and CDC-SESF. The areas in the blue and yellow boxes in the fusion results are the ground truths obtained from the focus area of the source image with the same resolution.

TABLE III

QUANTITATIVE VALUES OBTAINED "SR + FUSION," "FUSION + SR," AND OUR PROPOSED METHOD, RESPECTIVELY

Task	Method	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
Ly03	SR-Fusion	0.8204	0.4885	0.5620	0.8931	0.8461	0.7455
	Fusion-SR	0.8205	0.4780	0.5591	0.8914	0.8393	0.7474
	Ours	**0.8213**	**0.5154**	**0.6068**	**0.9028**	**0.8828**	**0.7672**
Ly05	SR-Fusion	0.8231	0.4197	0.5620	0.8732	0.7934	0.7793
	Fusion-SR	0.8232	0.4065	0.5578	0.8728	0.7825	0.7829
	Ours	**0.8245**	**0.4510**	**0.6025**	**0.8866**	**0.8402**	**0.8121**

TABLE IV

EFFECTIVENESS OF THE ATTENTION MECHANISM IN THIS ARTICLE. "W/O" DENOTES WITHOUT

Methods	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
w/o SAM+CAM	0.8220	0.4652	0.5301	0.8952	0.8310	0.7691
w/o SAM	0.8229	0.4667	0.5420	0.8939	0.8371	0.7894
w/o CAM	0.8224	0.4700	0.5409	0.8977	0.8367	0.7788
CAM+SAM	0.8224	0.4728	0.5424	0.8981	0.8379	0.7797
SAM+CAM	0.8226	0.4757	0.5605	0.8998	0.8451	0.7849

Fig. 15. Effectiveness of compared methods. Images from left to right are fused by U2Fusion, DRPL, GACN, and SESF.

for comparison. From the visual quality of the enlarged local area in Fig. 15 and the objective evaluation results in Table VI, it can be seen that DRPL, GACN, and SESF have better performance than U2Fusion. Thus, our constructed two-step method is reasonable for comparison.

TABLE V

EFFECTIVENESS OF COMPARISON METHODS USED IN SECTION IV-D. THE QUANTITATIVE AVERAGE VALUES OF THE FUSION RESULTS GENERATED BY DIFFERENT FUSION METHODS (SCALING FACTOR IS 2×) ARE LISTED

Methods	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
Bic-DRPL	0.8146	0.3318	0.3323	0.5653	0.4210	0.5422
Bic-GACN	0.8147	0.3227	0.3088	0.5247	0.3918	0.5481
Bic-SESF	0.8145	0.3375	0.3565	0.5706	0.4315	0.5401
DFUNet-DRPL	0.8154	0.2327	0.1445	0.7070	0.5160	0.5726
DFUNet-GACN	0.8155	0.2342	0.1454	0.7084	0.5085	0.5741
DFUNet-SESF	0.8154	0.2343	0.1454	0.7052	0.5066	0.5709
RDSR-DRPL	0.8209	0.4039	0.4608	0.8776	0.7478	0.7417
RDSR-GACN	0.8210	0.4128	0.4796	0.8861	0.7645	0.7442
RDSR-SESF	0.8210	0.4121	0.4794	0.8853	0.7641	0.7436
CDC-DRPL	0.8203	0.3546	0.4002	0.8414	0.6990	0.7271
CDC-GACN	0.8203	0.3615	0.4277	0.8510	0.7138	0.7286
CDC-SESF	0.8203	0.3614	0.4263	0.8508	0.7140	0.7282

TABLE VI

QUANTITATIVE AVERAGE VALUES OF THE FUSION RESULTS GENERATED BY U2FUSION, DRPL, GACN, AND SESF

Methods	Q_{NICE}	Q_G	Q_P	Q_S	Q_Y	Q_{MI}
U2Fusion	0.8217	0.5285	0.6964	0.8775	0.8557	0.7775
DRPL	0.8343	0.6896	0.8170	0.9400	0.9713	0.9678
GACN	0.8367	0.6981	0.8221	0.9404	0.9753	1.0561
SESF	0.8362	0.6967	0.8216	0.9394	0.9755	1.0519

V. CONCLUSION

In this article, we primarily propose an end-to-end network that simultaneously achieves multifocus image fusion and super-resolution for low-resolution images. A focus perception module is studied by calculating the affinity of feature maps under different receptive fields, so that the focus pixels-related features are enhanced and have been more discriminative. Simultaneously, referring to the super-resolution, features from the shallow layers and deep layers are jointly exploited, so that the texture information in the shallow layers and the context information in the deep layers are efficiently extracted. To enable the upsampled image to contain abundant edge details, a gradient map-based subnetwork is also embedded. Compared with methods, which implement fusion and super-resolution independently, our proposed method achieves two tasks in parallel way, avoiding enlarging artifacts caused by the inferior output of image fusion or super-resolution. Besides, in contrast to the existing joint learning approaches, which are generally for multisource image fusion and super-resolution, our presented method is particularly designed for multifocus images, being more adaptive. Experimental results on the real-world dataset, including outdoor and indoor images, substantiate the superiority of our proposed methods both qualitatively and quantitatively. However, we also notice that at the junction of the focus and defocus areas, there is a performance degradation. This is also a challenge to most methods, and we will focus on this problem in our future works.

REFERENCES

[1] Y. Liu, L. Wang, J. Cheng, C. Li, and X. Chen, "Multi-focus image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 64, pp. 71–91, Dec. 2020.

[2] 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.

[3] J. Liu, S. Li, H. Liu, R. Dian, and X. Wei, "A lightweight pixel-level unified image fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 11, 2023, doi: 10.1109/TNNLS.2023.3311820.

[4] Z. Zhang et al., "A componentwise approach to weakly supervised semantic segmentation using dual-feedback network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7541–7554, Sep. 2023.

[5] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[6] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5802–5811.

[7] S. Li et al., "Logical relation inference and multiview information interaction for domain adaptation person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 12, 2023, doi: 10.1109/TNNLS.2023.3281504.

[8] X. Li, H. Li, Z. Yu, and Y. Kong, "Multifocus image fusion scheme based on the multiscale curvature in nonsubsampled contourlet transform domain," *Opt. Eng.*, vol. 54, no. 7, Jul. 2015, Art. no. 073115.

[9] S. Li and B. Yang, "Multifocus image fusion by combining curvelet and wavelet transform," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1295–1301, Jul. 2008.

[10] X. Li, F. Zhou, H. Tan, Y. Chen, and W. Zuo, "Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal," *Signal Process.*, vol. 184, Jul. 2021, Art. no. 108062.

[11] Q. Jiang et al., "Two-scale decomposition-based multifocus image fusion framework combined with image morphology and fuzzy set theory," *Inf. Sci.*, vol. 541, pp. 442–474, Dec. 2020.

[12] H. Li, H. Qiu, Z. Yu, and B. Li, "Multifocus image fusion via fixed window technique of multiscale images and non-local means filtering," *Signal Process.*, vol. 138, pp. 71–85, Sep. 2017.

[13] H. Li, X. Liu, Z. Yu, and Y. Zhang, "Performance improvement scheme of multifocus image fusion derived by difference images," *Signal Process.*, vol. 128, pp. 474–493, Nov. 2016.

[14] H. Li, X. Li, Z. Yu, and C. Mao, "Multifocus image fusion by combining with mixed-order structure tensors and multiscale neighborhood," *Inf. Sci.*, vols. 349–350, pp. 25–49, Jul. 2016.

[15] J. Li et al., "DRPL: Deep regression pair learning for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4816–4831, 2020.

[16] Q. Zhang and M. D. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2045–2058, May 2016.

[17] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao, "Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion," *IEEE Trans. Instrum. Meas.*, vol. 69, pp. 1082–1102, 2020.

[18] Y. Zhang, M. Yang, N. Li, and Z. Yu, "Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion," *Signal Process.*, vol. 167, Feb. 2020, Art. no. 107327.

[19] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[20] X. Li, F. Zhou, and H. Tan, "Joint image fusion and denoising via three-layer decomposition and sparse representation," *Knowl.-Based Syst.*, vol. 224, Jul. 2021, Art. no. 107087.

[21] H. Li, X. He, Z. Yu, and J. Luo, "Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior," *Inf. Sci.*, vol. 523, pp. 14–37, Jun. 2020.

[22] J. Li et al., "Relaxed asymmetric deep hashing learning: Point-to-angle matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4791–4805, Nov. 2020.

[23] M. Li, K. Zhang, J. Li, W. Zuo, R. Timofte, and D. Zhang, "Learning context-based nonlocal entropy modeling for image compression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1132–1145, Mar. 2023.

[24] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[25] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108792.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

[28] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, Feb. 2021.

[29] X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Inf. Fusion*, vol. 92, pp. 127–138, Apr. 2023.

[30] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2388–2400, Jun. 2021.

[31] H. Wu, J. Gui, J. Zhang, J. T. Kwok, and Z. Wei, "Feedback pyramid attention networks for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4881–4892, Mar. 2023.

[32] W. Shi, F. Tao, and Y. Wen, "Structure-aware deep networks and pixel-level generative adversarial training for single image super-resolution," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[33] L. Yu et al., "Learning to super-resolve blurry images with events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10027–10043, Jan. 2023.

[34] G. Wu, J. Jiang, and X. Liu, "A practical contrastive learning framework for single-image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 10, 2023, doi: 10.1109/TNNLS.2023.3290038.

[35] W.-Y. Hsu and P.-W. Jian, "Wavelet pyramid recurrent structure-preserving attention network for single image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 13, 2023, doi: 10.1109/TNNLS.2023.3289958.

[36] H. Li, Z. Yu, and C. Mao, "Fractional differential and variational method for image fusion and super-resolution," *Neurocomputing*, vol. 171, pp. 138–148, Jan. 2016.

[37] H. Yin, S. Li, and L. Fang, "Simultaneous image fusion and super-resolution using sparse representation," *Inf. Fusion*, vol. 14, no. 3, pp. 229–240, Jul. 2013.

[38] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 4070–4083, 2021.

[39] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 671–679.

[40] A. Ellmauthaler, C. L. Pagliari, and E. A. B. da Silva, "Multiscale image fusion using the undecimated wavelet transform with spectral factorization and nonorthogonal filter banks," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1005–1017, Mar. 2013.

[41] Y. Yang, S. Tong, S. Huang, and P. Lin, "Multifocus image fusion based on NSCT and focused area detection," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2824–2838, May 2015.

[42] M. N. Do and M. Vetterli, "The finite ridgelet transform for image representation," *IEEE Trans. Image Process.*, vol. 12, no. 1, pp. 16–28, Jan. 2003.

[43] H. Hariharan, A. Koschan, and M. Abidi, "The direct use of curvelets in multifocus fusion," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 2185–2188.

[44] 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.

[45] Q. Zhang, G. Li, Y. Cao, and J. Han, "Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107325.

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

[47] X. Yan, S. Zulqarnain Gilani, H. Qin, and A. Mian, "Unsupervised deep multi-focus image fusion," 2018, *arXiv:1806.07272*.

[48] B. Ma, X. Yin, D. Wu, X. Ban, and H. Huang, "Gradient aware cascade network for multi-focus image fusion," Tech. Rep., 2020.

[49] H. Ma, Q. Liao, J. Zhang, S. Liu, and J.-H. Xue, "An a-Matte boundary defocus model-based cascaded network for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8668–8679, 2020.

[50] X. Nie, B. Hu, and X. Gao, "MLNet: A multi-domain lightweight network for multi-focus image fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 5565–5579, 2023.

[51] Z. Wang, X. Li, H. Duan, and X. Zhang, "A self-supervised residual feature learning model for multifocus image fusion," *IEEE Trans. Image Process.*, vol. 31, pp. 4527–4542, 2022.

[52] H. Li, M. Yang, and Z. Yu, "Joint image fusion and super-resolution for enhanced visualization via semi-coupled discriminative dictionary learning and advantage embedding," *Neurocomputing*, vol. 422, pp. 62–84, Jan. 2021.

[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[54] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[55] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 114–125.

[56] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, Mar. 2002.

[57] Q. Wang, Y. Shen, and J. Jin, "Performance evaluation of image fusion techniques," *Image Fusion, Algorithms Appl.*, vol. 19, pp. 469–492, Jun. 2008.

[58] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.

[59] J. Zhao, R. Laganière, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *Int. J. Innov. Comput., Inf. Control*, vol. 3, no. 6, pp. 1433–1447, Dec. 2007.

[60] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. Int. Conf. Image Process.*, Jul. 2003, p. 173.

[61] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Inf. Fusion*, vol. 9, no. 2, pp. 156–160, Apr. 2008.

[62] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, 2011.

[63] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[64] P. Wei et al., "Component divide-and-conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 101–117.

[65] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, and M. Mukeshimana, "SESF-fuse: An unsupervised deep model for multi-focus image fusion," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5793–5804, Jun. 2021.

[66] J. Liang, K. Zhang, S. Gu, L. V. Gool, and R. Timofte, "Flow-based kernel prior with application to blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10596–10605.

[67] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

Huafeng Li received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from Chongqing University, Chongqing, China, in 2009 and 2012, respectively.

He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His research interests include image processing, computer vision, and information fusion.

Ming Yuan received the B.E. degree in software engineering from the Qingdao University of Science and Technology, Qingdao, Shandong, China, in 2019, and the M.S. degree in software engineering from Kunming University of Science and Technology, Kunming, Yunnan, China, in 2022.

His research interests include computer vision and machine learning.

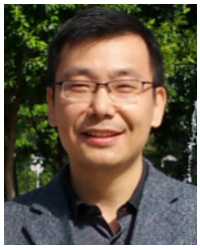
Jinxing Li (Member, IEEE) received the B.Sc. degree from the Department of Automation, Hangzhou Dianzi University, Hangzhou, China, in 2012, the M.Sc. degree from the Department of Automation, Chongqing University, Chongqing, China, in 2015, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China, in 2018.

He worked with The Chinese University of Hong Kong at Shenzhen, Shenzhen, China, from 2019 to 2021. He is currently an Associate Professor with Harbin Institute of Technology at Shenzhen, Shenzhen. His research interests are pattern recognition, deep learning, medical biometrics, and machine learning.

Yu Liu (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the Department of Biomedical Engineering, Hefei University of Technology, Hefei. His research interests include image processing, computer vision, information fusion, and machine learning. In particular, he is interested in image fusion, image restoration, visual recognition, and deep learning.

Dr. Liu was a recipient of the IET Image Processing Premium (Best Paper) Award in 2017 and the IEEE Instrumentation and Measurement Society Andi Chi Best Paper Award in 2020. He was identified as a Clarivate Highly Cited Researcher in 2023. He is serving as an Editorial Board Member for *Information Fusion* and an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.



Guangming Lu (Senior Member, IEEE) received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in computer science and engineering from Harbin Institute of Technology (HIT), Harbin, China, in 1998, 2000, and 2005, respectively.

He is currently a Professor with Harbin Institute of Technology at Shenzhen, Shenzhen, China. His current research interests include pattern recognition, image processing, and automated biometric technologies and applications.

Yong Xu (Senior Member, IEEE) received the B.S. and M.S. degrees from the Air Force Institute of Meteorology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, in 2005. AQ:9

He is currently a Professor with Harbin Institute of Technology at Shenzhen, Shenzhen, China. His current interests include pattern recognition, biometrics, machine learning, and video analysis.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005.

He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His main research interests include natural language process, image processing, and machine learning.

David Zhang (Life Fellow, IEEE) received the degree in computer science from Peking University, Beijing, China, the M.Sc. degree in computer science and the Ph.D. degree from Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994. AQ:10

He is currently a Chair Professor with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong at Shenzhen, Shenzhen, China. His research interests are medical biometrics and pattern recognition.