Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 4155948, 16 pages
https://doi.org/10.1155/2023/4155948

*Research Article*

# Two-Stage Focus Measurement Network with Joint Boundary Refinement for Multifocus Image Fusion

**Hao Zhai**,[1] **Xin Pan**,[1] **You Yang**,[2] **Jinyuan Jiang**,[1] **and Qing Li**[1]

[1]*College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China*
[2]*National Center for Applied Mathematics in Chongqing, Chongqing 401331, China*

Correspondence should be addressed to Xin Pan; 15984189028@163.com

Received 25 April 2023; Revised 27 June 2023; Accepted 17 August 2023; Published 31 August 2023

Academic Editor: Vittorio Memmolo

Focus measurement, one of the key tasks in multifocus image fusion (MFIF) frameworks, identifies the clearer parts of multifocus images pairs. Most of the existing methods aim to achieve disposable pixel-level focus measurement. However, the lack of sufficient accuracy often gives rise to misjudgments in the results. To this end, a novel two-stage focus measurement with joint boundary refinement network is proposed for MFIF. In this work, we adopt a coarse-to-fine strategy to gradually achieve block-level and pixel-level focus measurement for producing more fine-grained focus probability maps, instead of directly predicting at the pixel level. In addition, the joint boundary refinement optimizes the performance on the focused/defocused boundary component (FDB) during the focus measurement. To improve feature extraction capability, both CNN and transformer are employed to, respectively, encode local patterns and capture long-range dependencies. Then, the features from two input branches are legitimately aggregated by modeling the spatial complementary relationship in each pair of multifocus images. Extensive experiments demonstrate that the proposed model achieves state-of-the-art performance in both subjective perception and objective assessment.

## 1. Introduction

Images have become the most common and important information tool for people to observe the world, benefits from the development of imaging technology, photographic equipment, and the popularity of smartphones. However, due to the limited depth of field of optical lenses, only the targets within the focal distance can be clearly captured in a single shot, and others will eventually appear with varying degrees of blur. It becomes a challenging task to capture an all-in-focus image which can describe the whole scene in a more clear, comprehensive and realistic way. To address this issue, MFIF technique combines distinct focus parts of multiple images to obtain an all-in-focus image, effectively extending the depth of field. Nowadays, it has been widely applied to microscopy, digital photography, and video surveillance.

Conventional MFIF methods can generally be divided into two categories: spatial domain-based methods and transform domain-based methods. Spatial domain-based methods operate on the source images in a straightforwardly manner by selecting the clearer subimages, which are partitioned by fixed size windows or image segments. Research on spatial domain can be traced back to Li et al. [1], who proposed to fuse multifocus image based on a block partitioning scheme. Since then, block-based methods have dominated spatial domain-based methods. To achieve more effective activity-level measurement, some methods [2] based on the Laplacian energy algorithm has emerged. Several works also attempted to take multiple focus measures instead of single ones and designed a corresponding fusion rule to integrate the multiple results. The above methods divide the source image with a fixed size, which is highly affected by select of size. Thus, some approaches [3] adopted the bee algorithm to choose the optimal size, while others focused more on the relationship between adjacent blocks rather than treating them independently. To further improve the flexibility of image decomposition, the linear spectral clustering-based method [4] segment the source image into irregular regions but not fixed size blocks. Recently, pixel-based methods, such as local binary pattern-based [5], dense scale-invariant feature transform-based [6], and Hessian matrix-based [7], have become the research trend due to their superiority in obtaining precise pixel-level weighted maps.

Transform domain-based methods involve three steps: image decomposition, coefficient fusion, and inverse transform. First, the pixels of the source image are represented as coefficients of transform domain. Second, a predefined fusion rule is used to fuse the coefficients of different images. In the end, the fused coefficients are reconstructed by means of the inverse transform to produce the fused image. Research on multiscale decomposition has become the mainstream of transform domain-based methods with the emergence of multiscale analysis theories such as image pyramids and wavelets transform. Meanwhile, there are some methods based on other multiscale image decomposition. Among them, Li et al. [8] introduced the discrete wavelet transform (DWT) for image fusion for the first time and proposed an influential fusion framework consisting of three procedures: focus measurement, fusion rule, and consistency verification. Constrained by the ability of the wavelet transform to extract direction information in two-dimensional space, some methods [9, 10] utilized the pulse-coupled neural network (PCNN) and its improved versions to design the fusion rule. Within the framework of basic detail decomposition, Bavirsetti and Dhuli [11] introduced saliency detecting algorithms as the fusion rule. Some methods [12, 13] have also introduced guided filter algorithms to improve the performance of the basic detail decomposition. Sparse representation-based methods [14] are adopted to resolve the natural sparsity of features. These methods are in accordance with the physiological characteristics of the human visual system. Moreover, some methods [15, 16] attempt to combine several transform domain algorithms to synthesize their respective advantages during the fusion.

Conventional methods rely heavily on the focus measurement method and fusion rule. When faced with complex real-world scenarios, the hand crafted ones could not meet the requirements of producing high-quality all-in-focus images. This issue is further exacerbated by the fact that the connection between them is not taken into account in conventional methods. With the rise of deep learning, researchers have been investigating deep learning methods to the problem of MFIF. Some methods [17–22] treat focus measurement as a binary classification task and attempt to train a classification model to recognize the focused and defocused regions in the source images. To produce final fused results, the postprocessing steps in pixel-level spatial domain-based methods are essential, followed by these methods. With the aim of achieving an end-to-end fusion scheme, some methods [23, 24] use regression models to directly learn the mapping from the source images to the fused images. Since 2017, over 70 deep learning-based MFIF methods have been proposed [25]. They have demonstrated significant improvements in fusion quality compared to conventional methods. However, the promotion of their performance is limited, which occurs because most of them heavily rely on the results of one-shot pixel-level focus measurement results. Due to the inadequate performance, the results of one-shot are unreliable and lead to the significant degradation of the fused images. In addition, pixel-level focus measurement is prone to noise, especially from similar pixels in the focused and defocused regions. To address the above challenges, we propose a two-stage focus measurement network based on the encoder-decoder architecture for MFIF. The encoder with a Siamese network structure employs CNNs to encode local patterns in the early stages and transformers to capture global context relationships in the last stages. There is also a FAM designed for more legitimate mixing of features from different encoder branches. However, due to the multilevel downsampling as the network deepens, deep features have lost fundamental spatial details, making it hard to recognize the FDB. To this end, we decompose the focus measurement procedure into two stages with a coarse-to-fine strategy. First, deep features are fed into the coarse focus measurement where the HiLo block is utilized to capture global context information across different frequency domains. The analysis of the high- and low-frequency information can help us to achieve the focus measurement at block level. As for the fine focus measurement, an auxiliary boundary detection branch is added to extract more boundary related details from shallow features for refinement of the boundary. On this basis, we further integrate the deep features refined in the previous stage to achieve pixel-level prediction in order to produce more fine-grained focus probability maps for MFIF.

The main contributions of this work include as follows:

(1) We propose a two-stage focus measurement network for MFIF, consisting of a CFM and a FFM. In this way, the MFIF is formulated as two-stage process that achieves block-level and pixel-level predictions step-by-step with the coarse-to-fine strategy. Joint boundary refinement implemented by boundary detection branch has also improved the quality of boundary in the fusion images.

(2) LITv2 network is introduced as the backbone of the encoder. Compared to the models that only use CNN or transformer, LITv2 has the ability to both encode local patterns and model long-range dependency with lower computational cost and parameters. Furthermore, we have made some appropriate modifications to it, making it more suitable for the MFIF task.

(3) To make effective use of the mutual correlation and difference in each pair of multifocus images, FAM is designed to model the spatial complementary relationship through spatial attention mechanisms.

(4) For better supervise the proposed model, we collect two public datasets for salient object detection and construct a high-resolution large-scale multifocus image dataset.

Rest part of this paper is organized as follows. In Section 2, an overview of relevant research and vision transformer are briefly present. Section 3 introduces the structure of the proposed model in detail. Section 4 gives some subjective visual effect and objective evaluation results on public test sets, as well as conducting the ablation experiment to prove the effectiveness of specific modules. Lastly, we have concluded this paper in Section 5.

## 2. Related Works

*2.1. Deep Learning-Based Methods.* In 2017, Liu et al. [17] first introduced convolutional neural networks (CNNs) into the field of multifocus image fusion. Trained CNN demonstrates superior activity level measurement and fusion rule by learning the mapping from source images to focus maps. Guo et al. [18] proposed a fully convolutional network for MFIF. Notably, they eliminate the fully connected layer in the network and generate a focus map of the same size as input for MFIF in the way of segmentation. Attention mechanisms as a powerful tool in deep learning can effectively capture focus information for subsequent fusion, thus Zang et al. [19] proposed a novel unified fusion attention module to obtain informative fusion images not via simple element fusion operations in pervious works. Guan et al. [20] proposed to adopt nested connection structures and dilated convolutions to extract multiscale features. Expect CNN, generative adversarial network (GAN) has been applied to MFIF. In MFF-GAN [21], generator was utilized to produce fusion results of the same distribution as all-in-focus images and constructed an adversarial game to enhance texture details. To achieving precise edges while preserving the original texture, Li et al. [22] formulated fusion as adversarial learning between each pair of multifocus image features. Moreover, Xiao et al. [23] believed that defocused images are degraded from latent all-in-focus images and concocted a mathematical degradation model with the deep learning technique. Transformer, as a competitor to CNN, has also been introduced into MFIF with its global receptive field to model long-range correlations. Ma et al. [24] designed a long-distance cross-domain attention module based on Swin transformer, which generalized the image fusion of multiple scenes into a unified framework with structure maintenance, detail preservation, and appropriate intensity control.

Above methods have not paid enough attention to the defocus spread effects in multifocus image and obtained fused images with rough boundary. In this paper, two-stage focus measurement adopts the strategy from coarse-to-fine to extract informative spatial details in shallow feature extraction. In addition, auxiliary task branches for boundary detection are designed to augment boundary quality.

*2.2. Vision Transformer.* Vision transformer (ViT) is the first model that introduced self-attention mechanisms to computer vision tasks. Dosovitskiy et al. [26] proposed a ViT model that achieved significant improvement for image recognition. Since then, there are extensive research studies that attempted to improve ViT on diverse aspects. For instance, some work [27] sought to innovate locality into ViT to enhance its ability for encoding details. Inspired by the pyramid hierarchical structure in CNNs for dense prediction tasks, there is also a prevailing trend to introduce it into ViT [28]. However, the quadratic computational complexity brought by self-attention mechanisms brings high computational costs and memory consumption so that DeiT [29] introduced knowledge distillation methods to improve ViT's train efficiency. Swin transformer [30] proposed a locally windowed self-attention mechanism to reduce computational complexity. ViT has shown outstanding dominance in image processing, but it is quite difficult to train a well-performed ViT model, especially for MFIF, which lacks large-scale datasets with ground-truth. Furthermore, MFIF is still a high-resolution and intensive prediction task. It is unimaginable that normal ViT brings computing costs and operational burden. Therefore, we employ LiTv2 [31], which takes into account the operation efficiency and transformer's long-distance modeling capability for MFIF. LiTv2 replaces the early-stage multihead self-attention (MSA) with convolution, which not only improves the representation capacity of local features, but also avoids the computational costs and memory consumption of MSA in processing large-size feature maps in the early stages.

## 3. Proposed Method

*3.1. Overview.* In this paper, we formulate the MFIF as a two-stage focus measurement process. We build our model base on the encoder-decoder architecture, consisting of encoder, feature aggregation module (FAM), coarse focus measurement module (CFM), and fine focus measurement module (FFM), as shown in Figure 1. First, source images A and B are input into the different branch of encoder to extract multilevel features as the operation in the solid light blue box. The encoder adopts Siamese network structure, which the upper and lower branches share weights. Then, FAM models the spatial complementary relationships from different input branches to fuse features in each stage. In the decoder, CFM and FFM reach the precise prediction that guides the fusion of each pair images from block level to pixel level step-by-step as the operation in the solid blue box and the solid pink box. The following is a detailed discussion for each module.

*3.2. Encoder.* Encoder consists of two networks with the same structure and shares weight parameters during training. To conform to the characteristics of MFIF, we made some modifications to the LITv2 network, as shown in Figure 2. First, convolution with stride 4 is employed to split inputs into nonoverlapping patches and project the initial feature dimension from 3 to $C_1$, serving as the initial input for subsequent pipeline. Previous studies [32] prove that the both CNNs and transformers still focus on local patterns in the shallow layers, showing that the use of self-attention at early stages may be unnecessary. Therefore, MSAs in the early two

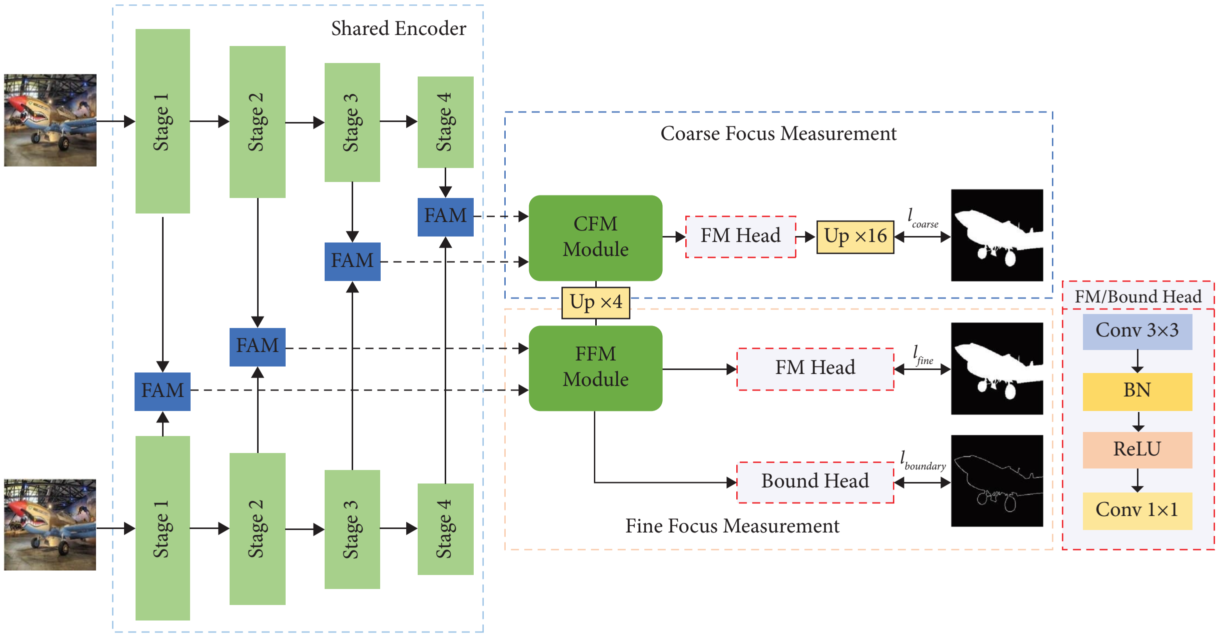

Figure 1: Overview of the proposed model.

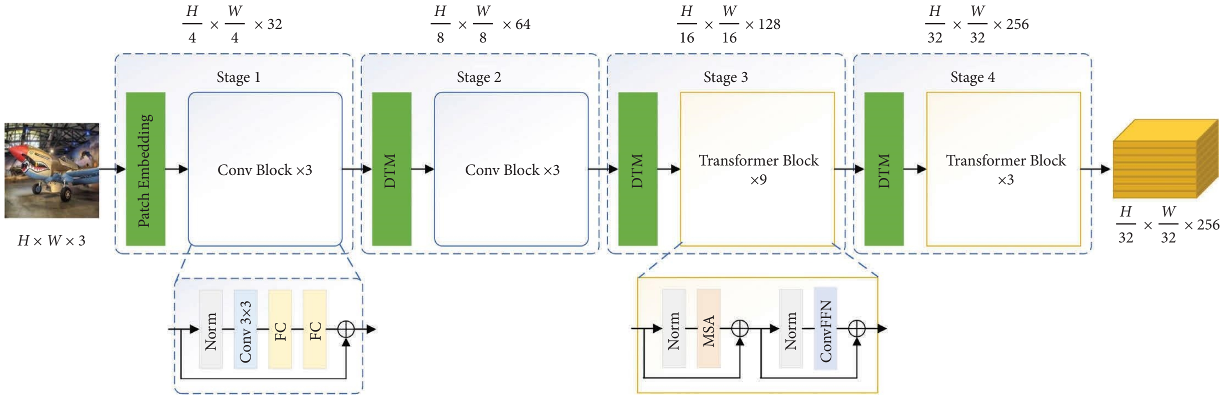
$\frac{H}{4} \times \frac{W}{4} \times 32$
$\frac{H}{8} \times \frac{W}{8} \times 64$
$\frac{H}{16} \times \frac{W}{16} \times 128$
$\frac{H}{32} \times \frac{W}{32} \times 256$
Stage 1
Stage 2
Stage 3
Stage 4
$H \times W \times 3$
Patch Embedding
Conv Block ×3
DTM
Conv Block ×3
DTM
Transformer Block ×9
DTM
Transformer Block ×3
$\frac{H}{32} \times \frac{W}{32} \times 256$
Norm
Conv 3x3
FC
FC
Norm
MSA
Norm
ConvFFN

Figure 2: Architecture of modified LITv2.

stages are replaced by the conv module, which consists of a 3 × 3 DWConv layer, two FC layers, followed by the LN normalization layer and GELU nonlinearity. The last two stages employ standard transformer blocks, including an MSA and a ConvFFN. MSA is responsible for capturing long-distance dependencies. Followed by ConvFFN adopts a zero-padding DWConv layer to incorporate the implicitly learned position information. We compare the computational consumption of the LITv2 network with other SOTA methods and mark the minimal values in bold. As shown in Table 1, with this structure, the LITv2 network can effectively increase the receptive field while reducing computational burden.

The entire model adopts the hierarchical pyramid structure which is divided into 4 stages, and the number of blocks in each stage is 3 : 3 : 9 : 3. A separate deformable token merging (DTM) module is used at each stage to scale feature size and expand dimension. Compared with ordinary convolution, the learning of offset grid in deformable convolution can adaptively select the sampling position of convolution, and search more informative local window when merging patches. Each stage feature representation of the two input branches is $\{F_i^A, F_i^B\}$, where $i \in \{1, 2, 3, 4\}$.

Table 1: Comparison results of computational consumption.

| Methods | Param (*M*) | FLOPs (*G*) | TrainMem (*GB*) |
|---|---|---|---|
| ResNet-50 [33] | 26 | 4.1 | 7.9 |
| ConvNext-Ti [34] | 28 | 4.5 | 8.3 |
| PVT-S [35] | 25 | 3.8 | 6.8 |
| Swin-Ti [30] | 28 | 4.5 | 6.1 |
| CVT-13 [36] | **20** | 4.5 | 6.1 |
| LITv2-S [31] | 28 | **3.7** | **5.1** |

The bold values in Table 1 denote the minimal parameter, computational consumption, and memory consumption of all the comparison methods.

*3.3. FAM.* In previous methods, there are two common approaches to fuse features from different input branches together: (1) features are fused straightforward in element-wise style such as element-wise maximum, element-wise sum, and element-wise average operation and (2) concatenation along the channel dimension. As the channel concatenation increases memory consumption, while element-wise fusion rules need to be selected based on the characteristics of the image dataset. Above all, they both ignore mutual correlation and difference in the process of fusion. Hence, the FAM mines the complementary information in each pair of multifocus images with the help of spatial attention mechanisms and fuse twofold features together, as shown in Figure 3.

Specifically, in the FAM, feature maps from branch A ($F_i^A$) and branch B ($F_i^B$) are normalized to [0, 1] by the max-min normalization. The complementary version ($\widehat{F}_i^B$) is obtained by 1 minus the normalized $F_i^B$. The FAM combines the $F_i^A$ and $\widehat{F}_i^B$ by multiplying the corresponding position elements as

$$x = \text{Norm}\left(F_i^A\right) \times \left(1 - \text{Norm}\left(F_i^B\right)\right). \tag{1}$$

The multiplication operation can enhance the focused information in the feature while reducing attention to defocused information. Then, the spatial attention module is applied to generate spatial attention map ($W_i^A$) for input branch A, which can enhance spatial focus information and suppress irrelevant regions. Specifically, max pooling (MAP) and avg pooling (AAP) are applied to the inputs among the channel dimension, and convolution on the concatenated feature maps compress the number of channels to 1. Followed by the sigmoid layer projects it to $W_i^A$, which is formulated as

$$W_i^A = \delta\left(\text{Conv}_{7\times 7}\left[\text{MaxPool}(x), \text{AvgPool}(x)\right]\right), \tag{2}$$

where $\delta, \text{Conv}_{7\times 7},$ and $[\bullet]$ denote the sigmoid function, a $7 \times 7$ convolution layer, and concatenation operation, respectively. And i denotes theindex of stage. Then, the obtained attention maps mix with preceding features to enhance the focus component. And the attention maps in the other branch can be generated by switching the order of inputs while calculating.

*3.4. Coarse-to-Fine Focus Measurement.* Due to the rich frequency-domain information present in natural images, where high-frequency components capture local details of objects (e.g., line and shape) and low-frequency components encode global structures (e.g., texture and color), frequency domain analysis has been a mainstream method for MFIF. Most existing deep learning methods for MFIF have not considered the characteristics of different frequencies in feature maps. To comprehend high/low frequency in feature maps, we introduce the HiLo [31] block.

HiLo block allocates a MSA calculation process to two paths: high-frequency attention (Hi-Fi) and low-frequency attention (Lo-Fi). Hi-Fi captures high-frequency interactions by local self-attention with initial-resolution feature maps; Lo-Fi captures low-frequency interactions by global attention with downsampled feature maps. The specific structure is shown in Figure 4, where the upper branch denotes the high-frequency attention and lower branch denotes the low-frequency attention.

The same number of heads in an MSA separates into two groups based on a split ratio $\alpha$ in the HiLo block, where $(1-\alpha)N_h$ heads are allocated to Hi-Fi and the remaining heads for Lo-Fi. In Hi-Fi, a simple nonoverlapping window partitioning method is first applied to the feature maps. Then, to capture high-frequency information in Input $X$, local self-attention is applied to each local window (e.g., $2 \times 2$ windows). For Lo-Fi, since mean filtering is a low-pass filter, average pooling is employed to obtain the low-frequency signal of the input feature map for each local window. Next, the feature maps after average pooling are projected to keys $K_2 \in \mathbb{R}^{H/2\times W/2\times C}$ and values $V_2 \in \mathbb{R}^{H/2\times W/2\times C}$, queries $Q \in \mathbb{R}^{H\times W\times C}$ still come from $X$. Low-frequency information in can be effectively obtained with standard self-attention. Finally, the outputs of HiLo are the concatenation of the outputs from the Hi-Fi and Lo-Fi paths, which is formulated as

$$\begin{aligned} \text{HiLo}(X) &= \left[\text{Hi-Fi}\left(Q_1, K_1, V_1\right)\right); \text{Lo-F}i\left(Q, K_2, V_2\right)] \\ &= \left[\text{LMSA}\left(Q_1, K_1, V_1\right); \text{MSA}\left(Q, K_2, V_2\right)\right], \end{aligned} \tag{3}$$

where LMSA andM SA denote local self-attention and standard self-attention, respectively.

For the coarse focus measurement stage, deep features are integrated by the CFM. The details of CFM are illustrated in Figure 5. First, feature maps of last two stages in the encoder $F_4$ and $F_3$ are input into CFM, and the FAM is applied to fuse multibranch features. Also, the HiLo block employs two disparate paths to disentangle high/low frequencies in the feature maps, which can effectively simulate the relationships between their different frequencies. First of all, the sizes of $F_4$ are too small to capture frequency-related information owing to multiple spatial reductions so that we select $F_3$ to extract rich high/low-frequency attention results. For a multifocus image, the frequency domain information in the focus area and the defocus area has different distribution. The clearer part contains more details which have greater changes, resulting in abundant low-frequency information, while the more blurred part displays more high-frequency information due to the concentration of pixel values. To make full use of frequency attention information, we apply two HiLo block to distinguish features from branch A and branch B. And the outputs are multiplying with $F_4$. Then, the integrated features after concatenating are fed to the FM head to generate focus probability map. By analyzing the different frequency domains of features, we can achieve block-level focus measurement of the image. The coarse focus measurement can be formulated as follows:

$$\begin{cases} F^B = \text{HiLo}\left(F_3^B\right) * \text{Up}\left(F_4^B\right), \\ F^A = \text{HiLo}\left(F_3^A\right) * \text{Up}\left(F_4^A\right), \\ D^{\text{coarse}} = \text{FM}\left(\left[F^A, F^B\right]\right), \end{cases} \tag{4}$$

where FM denotes focus measurement head, comprising a $3 \times 3$ convolution layer, BN, and ReLU, followed by a $1 \times 1$ convolution layer to reduce the number of channels to 1 for predicting.

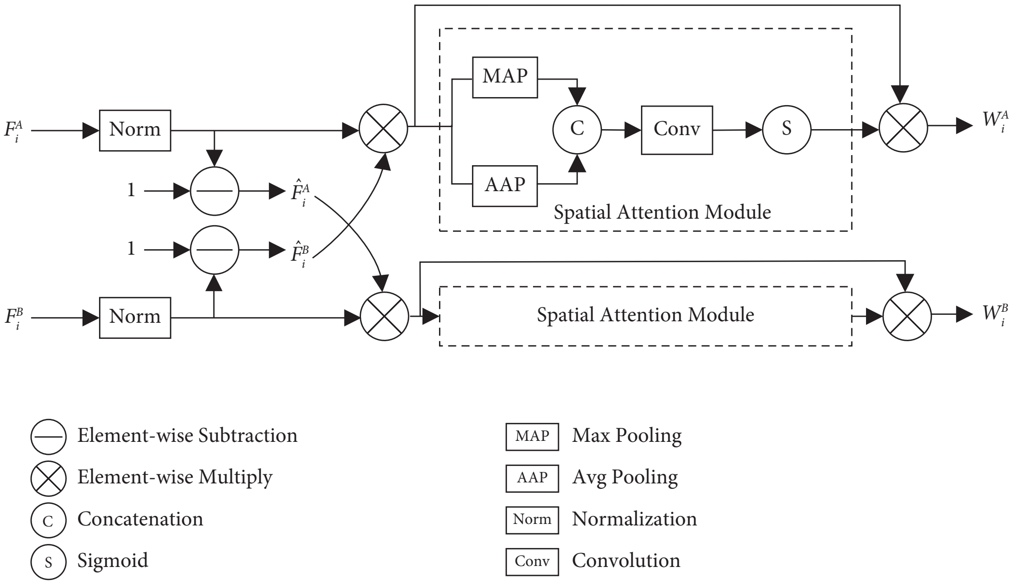

Figure 3: The multiple input branches feature fusion approaches in the FAM.

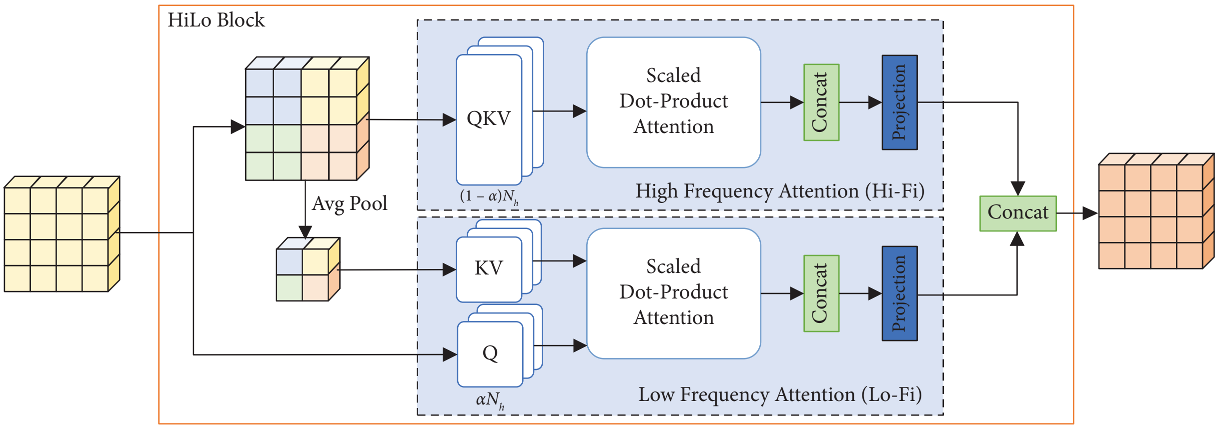
HiLo Block
QKV
Scaled Dot-Product Attention
Concat
Projection
$(1-\alpha)N_h$
High Frequency Attention (Hi-Fi)
Avg Pool
KV
Q
$\alpha N_h$
Scaled Dot-Product Attention
Concat
Projection
Low Frequency Attention (Lo-Fi)
Concat

Figure 4: Framework of HiLo block.

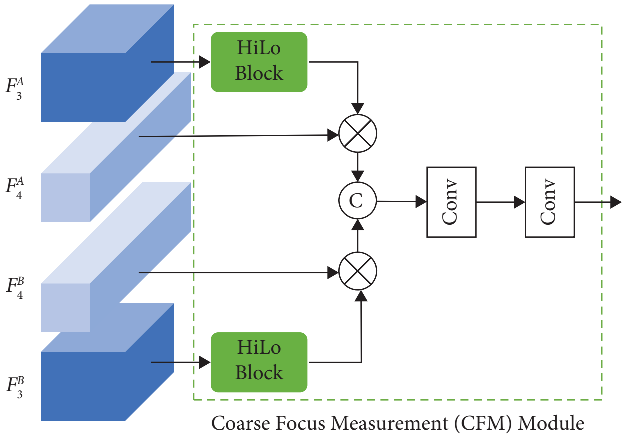

Figure 5: An illustration of the proposed CFM.

Compared with pixel-level focus measurement, block-level focus measurement is more robust [37]. However, as the coarse focus probability map is 1/16 the size of the input image, amplifying by a factor of 16 is prerequisite to guide the fusion of source image. In the procedure of expanding the sizeof focus probability map, the rate of misjudgment is also rapidly increasing, and it is apparent that the handling of local details becomes more tough. To improve the quality of coarse focus map, we proposed the fine focus measurement to estimate the focus map in a finer manner.

For the fine focus measurement stage, we can comprehend different stage feature information via fusing multiscale feature representation [38, 39]. To this end, shallow features from the early two stages of encoder are first integrated in the same process of CFM. The recent research shows that spatial positional information hidden in features can be learned by zero-padding [40]. Thus, in the early stage of encoder, depth-wise separable convolutional layer which initialize with zero-padding is employed to extract spatial details. Low-level features contain affluent spatial detailed information, which is crucial to understand the boundaries between focused and defocused regions. At this juncture, if we output the focus measurement results directly with the swallow features which includes extensive spatial details, a substantial amount of shape and structural of objects in the source image will persist. To suppress the impact of irrelevant texture, HiLo block is applied to $F_2$ which is less affected by redundant textures. By analyzing the frequency of $F_2$, we can exploit more important spatial details. Next, $F_4$ after frequency analysis guides the network to extract more boundary related features from $F_1$ as $F^s$ patial. Then, $F^c$ oarse and $F^s$ patial are amplified to the same size as inputs. We combine the features by the way of concatenating to predict it in pixel-level. Deep features can distinguish focused and defocused blocks, and on the other hand, shallow features guide the learning of boundary. The pixel-level focus prediction process can be formulated as

$$D^{\text{fine}} = \text{FM}\left(\left[\text{UP}\left(F^{\text{spatial}}\right), \text{UP}\left(F^{\text{coarse}}\right)\right]\right), \tag{5}$$

where UP denotes the upsample operation which is implemented via bilinear interpolation.

*3.5. Joint Boundary Refinement Branch.* As is well known, multifocus images can be traced back to: due to the limited depth range of the optical lens, some parts are clear and others are blur in the images which obtained in a single shot of the camera. In real images, influenced by the defocus diffusion effect, there is no clear boundary between the clear and blurred regions because the clear pixels overlap with the blurred pixels in this area even some are expanding outward. The focused/defocused boundary (FDB) is the most severely affected part of the defocused diffusion effect. These parts seriously confuse the judgment of the model, leading to the lack of clear boundaries in the decision maps generated by most methods. The previous approach ignored this phenomenon and treated the FDB on an equal footing with other regions. Misjudgment is more likely to occur when processing FDBs, generating final results with fuzzy boundaries.

Inspired by the above discussion, we construct an additional boundary detection branch for the boundary refinement. At the same time of focus measurement, the auxiliary branch aims to estimate the FDB maps for comparison with the boundary ground-truth. By jointly training a boundary detection branch, we can improve the boundary quality and optimized the handling of FDB. Boundary prediction maps can be obtained by

$$D^{\text{bound}} = BD\left(F^{\text{spatial}}\right), \tag{6}$$

where BD denotes boundary detection head, which has the same structure of FM head. Furthermore, Figure 6 shows some predictions results from different stages of our model on the three test datasets: from left to right, (A) near-focused image, (B) far-focused image, (C) coarse focus map, (D) fine focus map, and (E) bound map.

*3.6. Loss Function.* In the procedure of focus measurement from coarse-to-fine, the generation of decision map is regarded as a dense binary prediction task. L1 loss function is capable of calculating the error between decision map and ground-truth at each position, which can effectively supervise the training for proposed model. Thus, we devise L1 to optimize the generation of the decision map. For coarse focus measurement tasks, the loss function can be calculated as

$$L_{\text{coarse}}\left(D^{\text{coarse}}, \text{GT}\right) = \sum_{i=1}^{H\times W} \left|\text{UP}\left(D_i^{\text{coarse}}\right), \text{GT}_i\right|, \tag{7}$$

where $|,|$ denotes the mean absolute error. Also, fine focus measurement loss function can be calculated as

$$L_{\text{fine}}\left(D^{\text{fine}}, \text{GT}\right) = \sum_{i=1}^{H\times W} \left|D_i^{\text{fine}}, \text{GT}_i\right|. \tag{8}$$

Boundary detection task is a class-imbalance problem, as the number of boundaries in the ground-truth is much less than the nonboundary. Dice loss function [41] can evaluate the similarity between the prediction and label. The definition of Dice loss is

$$L_{\text{dice}}\left(D^{\text{boundary}}, \text{GT}^{\text{boundary}}\right) = 1 - \frac{2\sum_i^{H\times W} D_i^{\text{boundary}} \text{GT}_i^{\text{boundary}} + \varepsilon}{\sum_i^{H\times W}\left(D_i^{\text{boundary}}\right)^2 + \sum_i^{H\times W}\left(\text{GT}_i^{\text{boundary}}\right)^2 + \varepsilon}, \tag{9}$$

where $\varepsilon$ are smooth factor and set to 1. Particularly, it is not sensitive to the number of foreground and background samples which mean it can alleviate the class-imbalance problem. It focuses more on mining the foreground area during training, but gradient during training is unstable so that we compose BCE loss and Dice loss as the boundary loss $L_{\text{boundary}}$, which is formulated as follows:

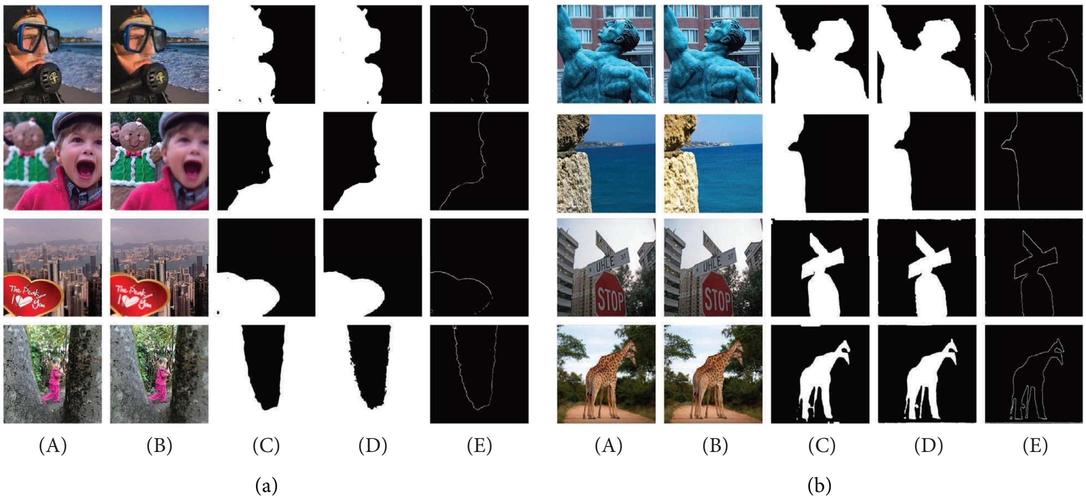

Figure 6: The prediction maps from different stages of our model.

$$L_{\text{boundary}}\left(D^{\text{boundary}}, \text{GT}\right) = L_{\text{dice}}\left(D^{\text{boundary}}, \text{GT}\right) + L_{\text{bce}}\left(D^{\text{boundary}}, \text{GT}\right). \tag{10}$$

Finally, the total objective function for the proposed model is a weighted sum of all subloss terms in equations (8) and (9):

$$L_{\text{total}} = \alpha_1 L_{\text{coarse}} + \alpha_2 L_{\text{fine}} + \alpha_3 L_{\text{boundary}}, \tag{11}$$

where $\alpha_1, \alpha_2,$ and $\alpha_3$ are set to 1, 1, and 0.5.

*3.7. Fusion Scheme.* The outputs of our proposed model are input into the postprocessing steps to refine the focus map for guiding the fusion of source images. As shown in Figure 7, the dashed yellow box is fused image. Notably, the outputs of fine focus measurement stage represent the focus probability of each pixel in $I^A$ which ranges from 0 to 1. In order to guide the fusion of multiple input images, we need to convert the focus probability map into a decision map via binary segmentation. For segmenting the focused and defocused regions in source images, the probability values above 0.5 are set to 1 and the values below 0.5 are set to 0.

It is inevitable that there may still be some misclassified pixels in the binary segmentation maps. Therefore, we need some postprocessing methods to improve the quality of the decision map. In this paper, we used a small region filtering algorithm to refine the decision map to remove small noise. Finally, the weights after postprocessing and the pixel values in the source image are combined by the weighted averaging strategy to produce the final all-in-focus image. Fused images $(I^{\text{fuse}})$ can be obtained as follows:

$$I^{\text{fuse}} = I_i^A * D_i^{\text{final}} + I_i^B * \left(1 - D_i^{\text{final}}\right). \tag{12}$$

## 4. Experiments

In this section, we first describe the experimental setups in detail. Next, we compare the proposed model with several state-of-the-art methods on both subjective visual effect and objective evaluation. And last, the ablation experiments on the proposed model have shown the effectiveness of each module.

### 4.1. Experimental Setups

*4.1.1. Training Dataset.* In the field of multifocus image fusion, there is a lack of large-scale real image datasets with labeled data. To better supervise the training process, we collect six commonly applied public datasets from salient object detection tasks: DUT-RGB [42], HKU-IS [43] to construct the training dataset. First, we filtrate some low-grade samples to balance the overall distribution of datasets. Next, the initial images, as shown in Figure 8(a), and corresponding mask images, as shown in Figure 8(d), are cropped to 256 × 256. Then, a Gaussian filtering kernel with a window size of 7 and a standard deviation of 2 is adopted to simulate the process from focus to defocus, generating two different levels of blurred versions in total. In addition, the $\alpha$-Matte model [44] was used to simulate the defocus diffusion effect to obtain pairs of multifocus images including the clear foreground and blurred background of $I^A$, as shown in Figure 8(b), and its complementary image $I^B$, as shown in Figure 8(c).

After above operations, 14704 pairs of multifocus images are produced as the whole training dataset. To avoid overfitting, we also do some data augmentation in the training process, i.e., randomly flip the inputs vertically or horizontally. As for supervising the auxiliary boundary detection task, we apply the Canny algorithm to masks for generating boundary labels, as shown in Figure 8(e).

*4.1.2. Testing Datasets.* To validate the effectiveness of the proposed model for MFIF, we choose three public image datasets as the test sets, i.e., Lytro datasets which contains 20 pairs of 520 × 520 natural multifocus images, MFI-WHU which contains 30 pairs of high-resolution multifocus images, and MFFW which contains 13 pairs of irregular

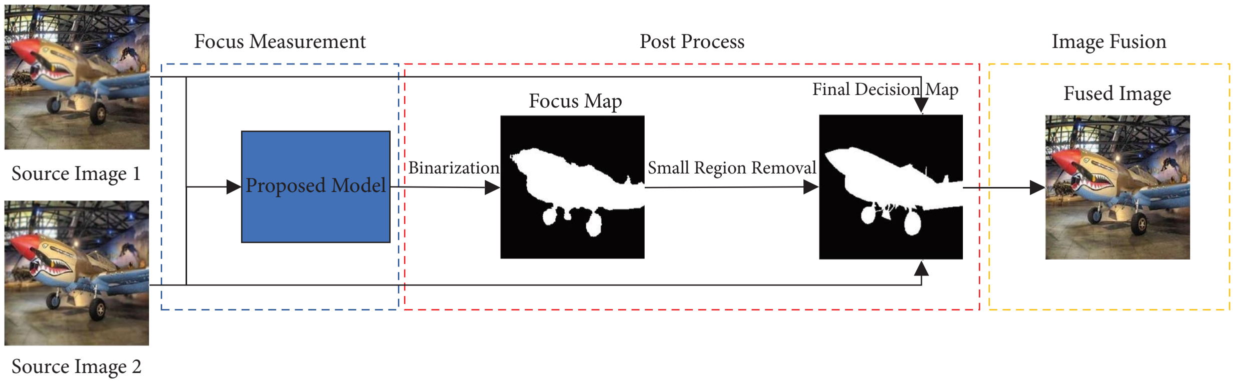

Figure 7: Schematic of the multifocus image fusion.

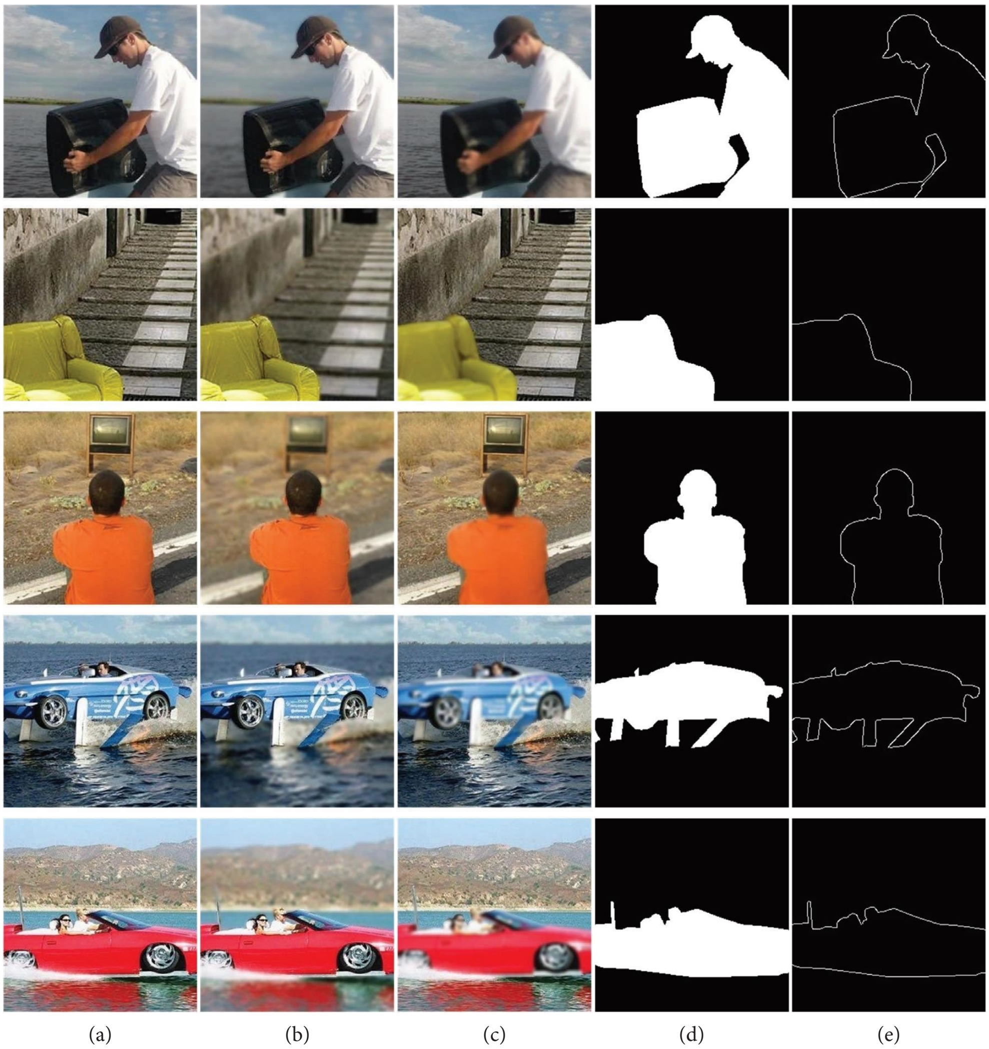

Figure 8: Examples from our generated training dataset.

multifocus images. The three testing datasets, as shown in Figure 9 are all widely used in MFIF task.

*4.1.3. Implement Details.* The proposed model is implemented in the PyTorch framework and trained and tested on a platform with Intel(R) Core (TM) i9-10900X CPU @ 3.70 GHz and NVIDA GeForce RTX 3090. At training time, the AdamW optimizer with a poly decay learning rate scheduler and an initial learning rate of 0.0001 was employed to optimize the proposed model. The total batch size is set as 24. In the training process, the proposed model is trained on the training multifocus image datasets for 50 epochs. We

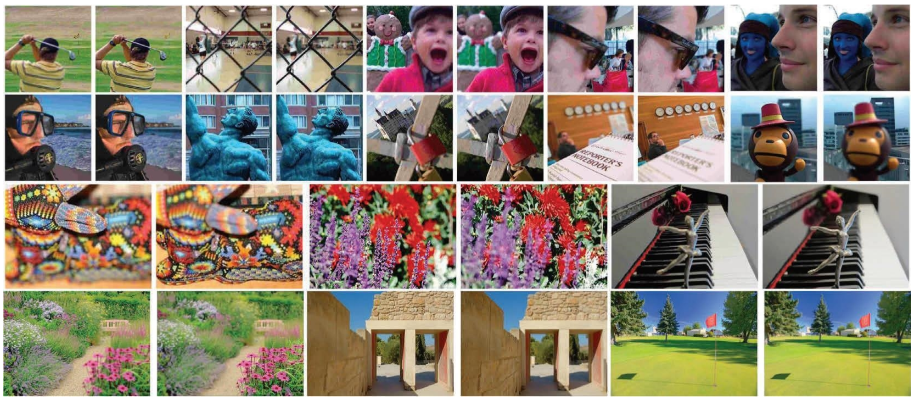

Figure 9: Examples of the Lytro, MFFW, and MFI-WHU datasets.

mainly optimize $L_{\text{coarse}}$ and $L_{\text{fine}}$ to get better focus maps. Here, $L_{\text{boundary}}$ optimizing boundary between focused and defocused region. With this coarse-to-fine training strategy, we can not only predict from the image block, but also accurately classify each pixel.

#### 4.1.4. Comparison Settings

*(1) Evaluation Metrics*. Since there are no available ground-truth images for MFIF algorithms to reference, the evaluation of MFIF methods becomes a tough task. In general, we will evaluate the performance of MFIF methods by objective evaluation metrics. To have comprehensive and objective performance comparison, we have selected six metrics in total to evaluate the quality of the fused images: information theory-based metric: feature mutual information (FMI_$p$ [45]) and phase congruency ($Q_P$ [46]); correlation-based metric: image feature-based metrics: average gradient ($Q_{AG}$ [47]) and gradient-based similarity ($Q_{AB/F}$ [48]); image structural similarity-based metric: structural similarity index measure ($Q_{\text{SSIM}}$ [49]); and human perception inspired fusion metric: visual information fidelity (VIFF [50]). For each of the above metrics, larger scores indicate better fusion performance.

*4.1.5. Comparative Methods.* The proposed method is compared with 9 SOTA methods, which include two conventional methods (i.e., MGFF [51] and GFDF [52]), seven DL-based methods (i.e., GCF [53], CNN [17], IFCNN [54], PMGI [55], MSFIN [56], GACN [57], and SwinFusion [24]). Some of the fusion results are available in [25], and others are publicly available online.

### 4.2. Experimental Results and Discussion

*4.2.1. Visual Quality Comparison.* Frist, we have chosen three testing examples (i.e., "Girl" in the Lytro dataset, "Old man" in the MFFW dataset, and "Bear" in the MFI-WHU dataset) to compare the subjective visual quality of different methods. To distinguish the differences between source images and fused images, we mark a specific region with the red circle and enlarge it in the lower-left corner of the fused image. Furthermore, we create a pseudocolor image based on the absolute difference value of the specific region, which is displayed in the lower-right corner. The color value in the pseudocolor image indicates the degree of difference. In general, the absolute difference value of focused region in pseudocolor image should be pure blue.

Figure 10 shows the fused images of different methods on the "Girl" example from the Lytro dataset. The lower-left corner of the images is an enlarged view of the red framed area, and the lower-right corner is a pseudocolor image to distinguish the difference between source image and fused image. There are some unexpected blurry artifacts around the fence regions in the fusion results of MGFF, PMGI, GACN, and SwinFusion. The redder regions in the pseudocolor images indicate that those methods are very likely to make misjudgments around the area. In contrast, the focused region in the pseudocolor image of our model is basically pure blue which means there are no so many misjudgments. Moreover, our model can detect accurately boundaries so that the pixels of around boundary are obviously divided into two parts.

The fused images of different methods on the "Old man" example from the MFFW dataset are shown in Figure 11, due to the presence of a large background in this image, most of them are very close in the visual effect. However, we can see the difference clearly in the pseudocolor images. There are a lot of noise and no clear boundary in the results of MGFF, GFDF, IFCNN, PMGI, GACN, and SwinFusion. Their performance will descend because those methods are unable to deal with very similar clear and blurry pixels. Unlike the previous methods, our model via boundary refine process improves this issue greatly.

Figure 12 shows the fused images of different methods on the "Bear" example from the MFI-WHU dataset. We can know that the "Bear" in the source image is in the wild, which leads it to be more relevant to reality. One thing that is similar to the previous "Old man" is that they have a large area of same color in the foreground and background, making it more difficult to separate the focused and defocused regions. Within all the methods, MGFF, IFCNN, PMGI, and SwinFusion are not good at handling with the "Bear." Our model and other methods achieve significantly performance.

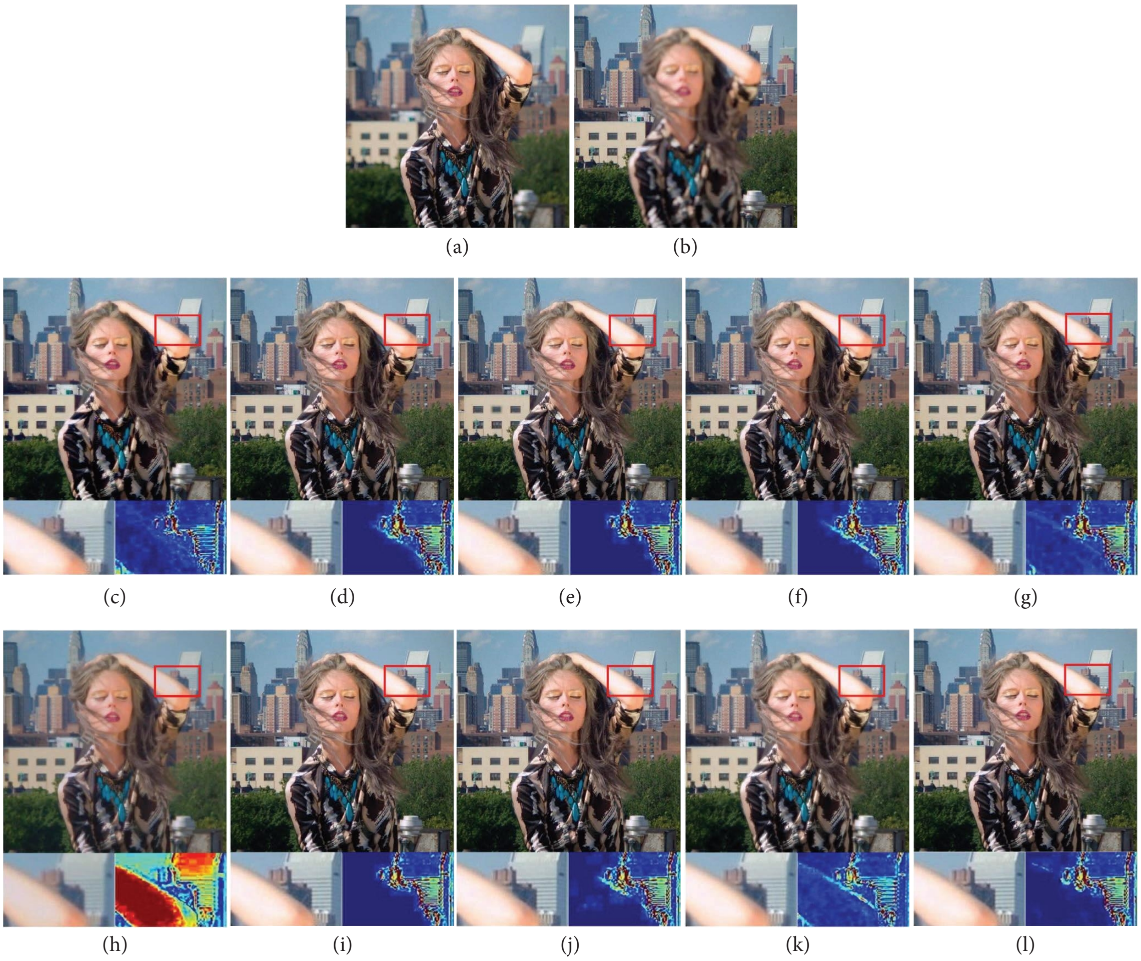

Figure 10: Comparison of the fused images between our model and the SOTA methods on "Girl" of Lytro dataset. (a) Source image 1, (b) source image 2, (c) MGFF, (d) GFDF, (e) GCF, (f) CNN, (g) IFCNN, (h) PMGI, (i) MSFIN, (j) GACN, (k) SwinFusion, and (l) ours.

*4.2.2. Quantitative Evaluation Comparison.* To further validate the effectiveness of our proposed model, we have also conducted a quantitative comparison experiment on objective evaluation metrics. All the quantitative evaluation results of 10 methods on the three testing datasets are listed in Tables 2–4, respectively. For each metric, the average score over all test images in each dataset is exhibited, and the top three of each metric among the 10 methods are shown in the following brackets. The highest score among all the methods is marked in bold which means the best fusion effect amongthe comparison methods in the three testing datasets.

From Table 2, we can see the superior performance of our model on FMI_$p$, $Q_{AG}$, $Q_{AB/F}$, and $Q_{\text{SSIM}}$ on the Lytro dataset. As for the $Q_P$ and VIFF, the performance of our model is not in the top three. But the average score over the 20 pairs of testing images is very close to the top three. Considering that the comparison methods represent the state of the art in MFIF, the proposed model also achieves appreciable performance on the two metrics. For the MFFW dataset, the results on Table 3 show that we have won the first place on FMI_$p$, $Q_P$, $Q_{AB/F}$, $Q_{\text{SSIM}}$, and $Q_{AG}$. Only on VIFF, our model wins the second places. Among three testing datasets, our model makes the best performance on the MFFW dataset. For the MFI-WHU dataset, the results on Table 4 show that we have won the first place on $Q_{AB/F}$ and $Q_{\text{SSIM}}$. For the $Q_{AG}$ and VIFF, the ranks of our model are in the top three. But for the FMI_$p$ and $Q_P$, the average scores are not in top three. Overall, even though our model is not the one who performs best in each metric, we are still able to achieve a very competitive performance with the SOTA methods.

*4.2.3. Extended Experiments.* To validate the generalization of our model, we conduct the extend experiments on the dataset with three source images. Specifically, we first fuse two of the source images as before and then fuse this intermediate result with the last source image to produce the final fused image. Figure 13 shows the results of three multifocus dataset. It can be seen that the fused images of our model contain all the focused regions in the source images, which is an all-in-focus image with high visual quality. Even the number of multifocus images more than two, the proposed model still owns the capability of handling this situation via this extended experiment.

*4.3. Ablation Experiments.* In this section, to verify the effectiveness of the proposed modules in our method, we conduct a series of ablation experiments on the Lytro

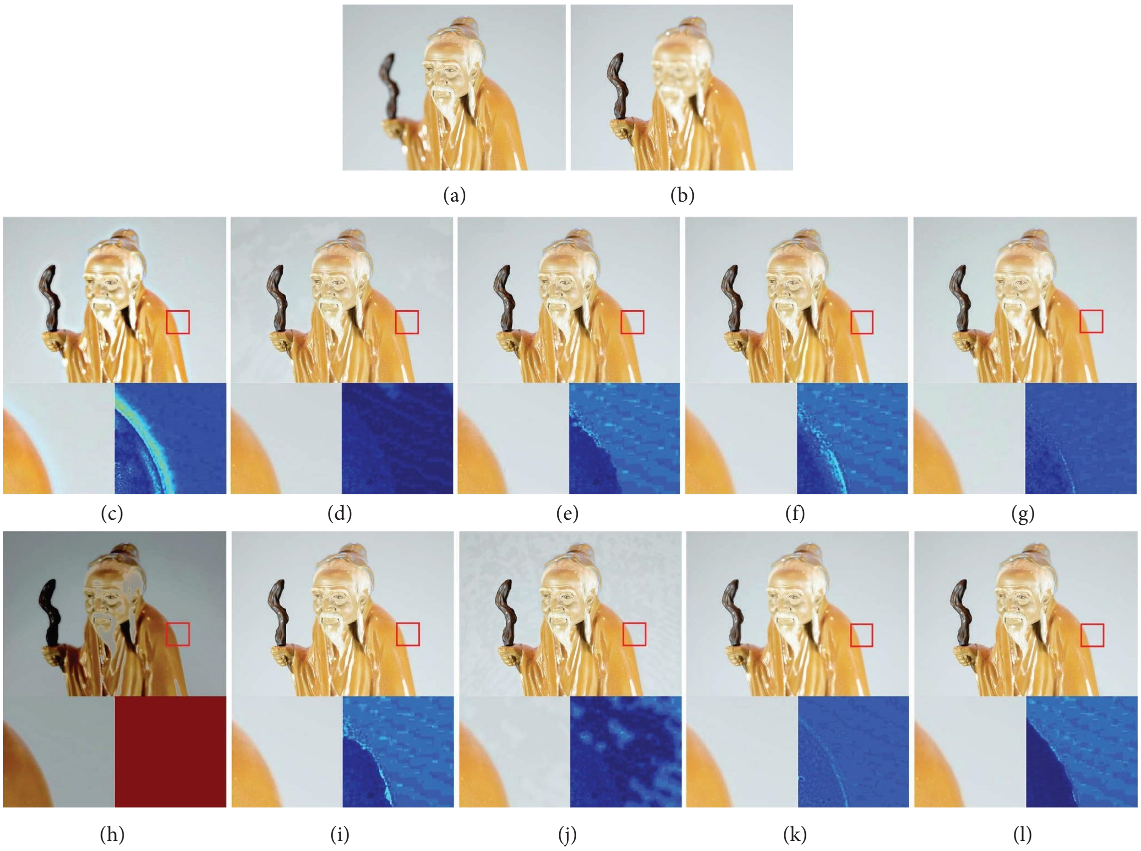

Figure 11: Comparison of the fused images between our model and the SOTA methods on “Old man” of MFFW dataset. (a) Source image 1, (b) source image 2, (c) MGFF, (d) GFDF, (e) GCF, (f) CNN, (g) IFCNN, (h) PMGI, (i) MSFIN, (j) GACN, (k) SwinFusion, and (l) ours.

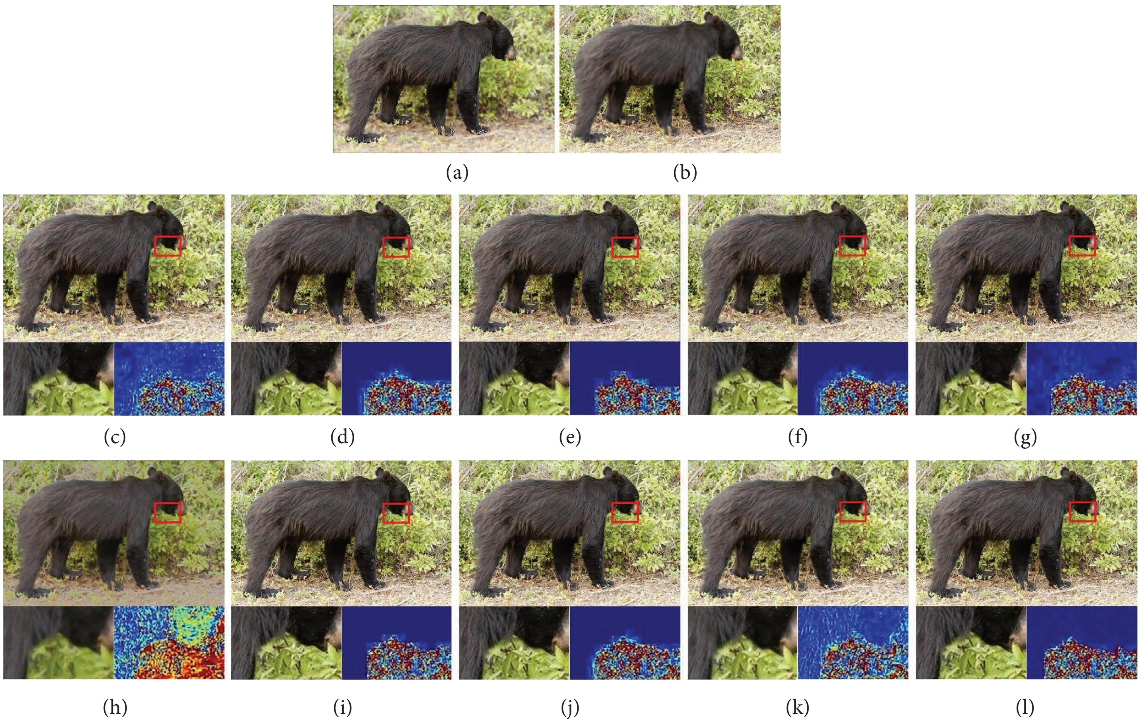

Figure 12: Comparison of the fused images between our models and the SOTA methods on “Bear” of MFI-WHU dataset. (a) Source image 1, (b) source image 2, (c) MGFF, (d) GFDF, (e) GCF, (f) CNN, (g) IFCNN, (h) PMGI, (i) MSFIN, (j) GACN, (k) SwinFusion, and (l) ours.

Table 2: Average scores of objective assessment of different methods on the Lytro dataset.

| Methods | FMI_*p* | $Q_P$ | $Q_{AG}$ | $Q_{AB/F}$ | $Q_{\text{SSIM}}$ | VIFF |
|---|---|---|---|---|---|---|
| MGFF | 0.8883 | 0.7067 | 6.0644 | 0.6563 | 0.8868 | **0.9853** (1) |
| GFGF | 0.9012 (3) | **0.8459** (1) | 6.7980 | 0.7591 | 0.9873 (3) | 0.9455 (3) |
| GCF | **0.9013** (1) | 0.8421 | 6.8114 | 0.7573 | 0.9852 | 0.9450 |
| CNN | 0.9010 | 0.8449 (2) | 6.7703 | 0.7581 | 0.9861 | 0.9437 |
| IFCNN | 0.8967 | 0.8047 | 6.8299 (3) | 0.7260 | 0.9531 | 0.9415 |
| PMGI | 0.8814 | 0.4612 | 3.5654 | 0.3861 | 0.6756 | 0.6034 |
| MSFIN | 0.9011 | 0.8443 (3) | 6.8314 (2) | 0.7592 | 0.9874 (2) | 0.9466 (2) |
| GACN | 0.9010 | 0.8429 | 6.8063 | 0.7597 (3) | 0.9865 | 0.9426 |
| SwinFusion | 0.8952 | 0.7715 | 5.9323 | 0.7140 (2) | 0.9061 | 0.9050 |
| Ours | **0.9014** (1) | 0.8435 | **6.8414** (1) | **0.7620** (1) | **0.9887** (1) | 0.9444 |

The bold values in the Table 2 denote that the corresponding method obtains the highest values in the metric which means this method achieves the best fusion results in this metric. Also, the numbers in the bracket denote the ranks of all the comparison methods in the metric.

Table 3: Average scores of objective assessment of different methods on the MFFW dataset.

| Methods | FMI_*p* | $Q_P$ | $Q_{AG}$ | $Q_{AB/F}$ | $Q_{\text{SSIM}}$ | VIFF |
|---|---|---|---|---|---|---|
| MGFF | 0.8716 | 0.4801 | 6.9713 | 0.5705 | 0.7269 | **0.9803** (1) |
| GFGF | 0.8812 | 0.5841 (3) | 7.5270 | 0.6363 | 0.7984 | 0.8299 |
| GCF | 0.8785 | 0.5613 | 7.7072 (2) | 0.6342 | 0.8023 | 0.8182 |
| CNN | 0.8813 (3) | 0.5744 | 7.4252 | 0.6321 | 0.7989 | 0.8259 |
| IFCNN | 0.8763 | 0.5388 | 7.6141 | 0.6107 | 0.7781 | 0.8367(3) |
| PMGI | 0.8675 | 0.3554 | 4.1518 | 0.3673 | 0.5651 | 0.6048 |
| MSFIN | 0.8799 | 0.5673 | 7.6195 | 0.6286 | 0.8040 (3) | 0.8300 |
| GACN | 0.8808 | 0.5695 | 7.6243 (3) | 0.6347 (3) | 0.7775 | 0.8257 |
| SwinFusion | 0.8816 (2) | 0.6464 (2) | 6.5421 | 0.6788 (2) | 0.8442 (2) | 0.8169 |
| Ours | **0.8859** (1) | **0.7180** (1) | **7.7211** (1) | **0.7207** (1) | **0.9516** (1) | 0.8422 (2) |

The bold values in the Table 3 denote that the corresponding method obtains the highest values in the metric which means this method achieves the best fusion results in this metric. Also, the numbers in the bracket denote the ranks of all the comparison methods in the metric.

Table 4: Average scores of objective assessment of different methods on the MFi-WHU dataset.

| Methods | FMI_*p* | $Q_P$ | $Q_{AG}$ | $Q_{AB/F}$ | $Q_{\text{SSIM}}$ | VIFF |
|---|---|---|---|---|---|---|
| MGFF | 0.8667 | 0.6964 | 7.1443 | 0.6386 | 0.9219 | 0.9507 |
| GFGF | 0.8786 (3) | **0.7883** (1) | 8.1339 | 0.7331 (2) | 0.9885 (3) | 0.9847 |
| GCF | 0.8781 | 0.7866 | 8.1463 (3) | 0.7309 (3) | 0.9875 | 0.9831 |
| CNN | **0.8788** (1) | 0.7874 (2) | 8.0703 | 0.7296 | 0.9880 | 0.9833 |
| IFCNN | 0.8737 | 0.7699 | **8.2545** (1) | 0.6940 | 0.9597 | **1.0028** (1) |
| PMGI | 0.8558 | 0.4750 | 4.6884 | 0.4178 | 0.7029 | 0.7809 |
| MSFIN | 0.8782 | 0.7857 | 8.1227 | 0.7294 | 0.9886 (2) | 0.9870 (2) |
| GACN | **0.8788** (1) | 0.7873 (3) | 8.0491 | 0.7271 | 0.9884 | 0.9847 |
| SwinFusion | 0.8717 | 0.7478 | 6.8843 | 0.6782 | 0.9123 | 0.9438 |
| Ours | 0.8781 | 0.7861 | 8.2033 (2) | **0.7370** (1) | **0.9891** (1) | 0.9867 (3) |

In Table 4, the horizontal axis is the metric value, and the vertical axis is the comparison method. We have listed six metrics for each method and compared with ten methods on MFI-WHU dataset. To show the fusion effect, we have highlighted the highest metrics values in bold of ten methods. So, the values in bold mean that the best fusion effect in Table 4.

dataset. Table 5 shows all the results of ablation experiments in detail, where the highest scores are highlighted in bold to display the influence of the proposed module on the final fusion. At the beginning of ablation experiments, the baseline model consists only of encoder, the CFM and FAM. Then, the FFM is added into the decoder part as the second row. It can be seen from it that there has been an improvement in performance with the FFM. In the next step, the joint boundary refine (JBR) is installed in the baseline model as the third row. The results of the third row prove that JBR is beneficial to our proposed model. For the verification of the FAM, we remove it as the fourth row of the table. In the results, we can see that the performance without the FAM has decreased.

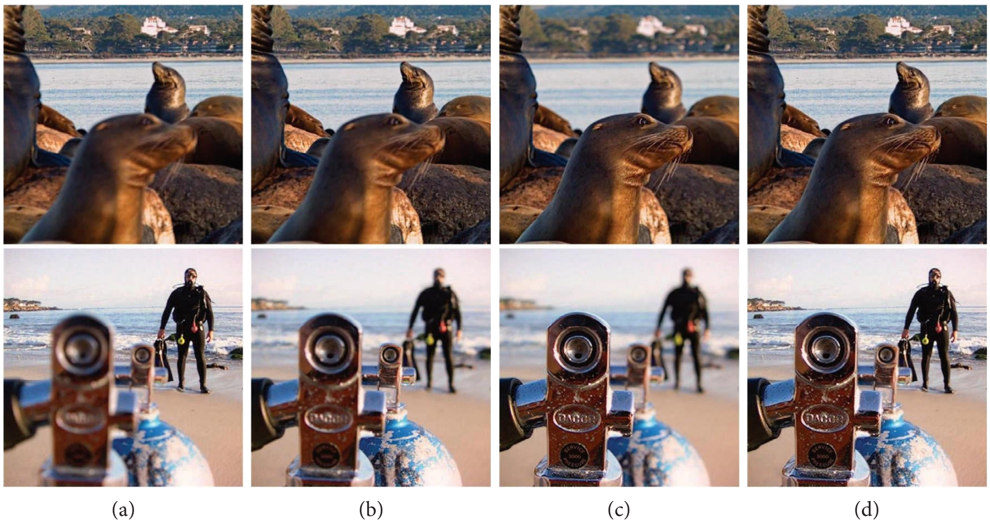

Figure 13: Fused results on three source images: (a) source image 1, (b) source image 2, (c) source image 3, and (d) fused image.

Table 5: Average scores of objective evaluation results in the ablation experiments on the Lytro dataset.

| Baseline | FFM | JBR | FAM | FMI_$p$ | $Q_{AG}$ | $Q_{AB/F}$ | VIFF |
|---|---|---|---|---|---|---|---|
| √ | | | √ | 0.9010 | 6.7929 | 0.7584 | 0.9387 |
| √ | √ | | √ | 0.9011 | 6.7959 | 0.7584 | 0.9393 |
| √ | √ | √ | | 0.9002 | 6.8358 | 0.7539 | 0.9443 |
| √ | √ | √ | √ | **0.9014** | **6.8414** | **0.7620** | **0.9444** |

The bold values denote the highest metric values in the ablation experiments to verify the effectiveness of our proposed module on the Lytro dataset.

## 5. Conclusion

In this paper, we propose a deep learning method based on the encoder-decoder architecture for MFIF and introduce a two-stage focus measurement network to generate satisfactory fusion results. The straightforward one-shot pixel-level focus measurement was implemented as a progressively increasing process from block-level to pixel-level in the coarse-to-fine strategy. The inspiration for our work is to achieve focus measurement from different granularities, which not only fully utilizes features at each level but also avoids the instability of single-focus measurement. In addition, to improve the performance about FDB that suffers severely from defocus spread effect, an additional auxiliary branch is designed to estimate the boundaries in the input image. The experimental results show that the proposed model demonstrates superior performance in both subjective visual effects and objective evaluation metrics. In the future, we will continue to optimize the model's design to improve prediction accuracy to meet the requirements of the human visual system and computer-related processing tasks.

## Data Availability

The datasets generated during the current study are available from the corresponding author upon request. Code is available at https:/github.com/panxin904/MFLIT.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant no. 62003065), the Science and Technology Research Project of Chongqing Municipal Education Commission (Grant nos. KJQN202200564 and KJZD202200504), the Fund project of Chongqing Normal University (Grant no. 21XLB032), and the Chongqing Education Science 14th Five Year Plan Project (Grant no. 2022-576).

## References

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

[2] X. Jin, D. Zhou, S. Yao et al., "Multi-focus image fusion method using S-PCNN optimized by particle swarm optimization," *Soft Computing*, vol. 22, no. 19, pp. 6395–6407, 2018.

[3] A. Banharnsakun, "Multi-focus image fusion using best-so-far abc strategies," *Neural Computing & Applications*, vol. 31, no. 7, pp. 2025–2040, 2019.

[4] J. Duan, L. Chen, and C. P. Chen, "Multifocus image fusion with enhanced linear spectral clustering and fast depth map estimation," *Neurocomputing*, vol. 318, pp. 43–54, 2018.

[5] W. Yin, W. Zhao, D. You, and D. Wang, "Local binary pattern metric-based multi-focus image fusion," *Optics & Laser Technology*, vol. 110, pp. 62–68, 2019.

[6] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Information Fusion*, vol. 23, pp. 139–155, 2015.

[7] B. Xiao, G. Ou, H. Tang, X. Bi, and W. Li, "Multi-focus image fusion by hessian matrix based decomposition," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 285–297, 2020.

[8] 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.

[9] J. Qian, L. Yadong, D. Jindun, F. Xiaofei, and J. Xiuchen, "Image fusion method based on structure-based saliency map and FDST-PCNN framework," *IEEE Access*, vol. 7, pp. 83484–83494, 2019.

[10] Z. Dong, C. S. Lai, D. Qi, Z. Xu, C. Li, and S. Duan, "A general memristor-based pulse coupled neural network with variable linking coefficient for multi-focus image fusion," *Neurocomputing*, vol. 308, pp. 172–183, 2018.

[11] D. P. Bavirisetti and R. Dhuli, "Multi-focus image fusion using multi-scale image decomposition and saliency detection," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 1103–1117, 2018.

[12] Y. Zhang, W. Wei, and Y. Yuan, "Multi-focus image fusion with alternating guided filtering," *Signal, Image and Video Processing*, vol. 13, no. 4, pp. 727–735, 2019.

[13] F. Zhou, X. Li, J. Li, R. Wang, and H. Tan, "Multifocus image fusion based on fast guided filter and focus pixels detection," *IEEE Access*, vol. 7, pp. 50780–50796, 2019.

[14] X. Ma, S. Hu, S. Liu, J. Fang, and S. Xu, "Multi-focus image fusion based on joint sparse representation and optimum theory," *Signal Processing: Image Communication*, vol. 78, pp. 125–134, 2019.

[15] A. Vishwakarma and M. K. Bhuyan, "Image fusion using adjustable non-subsampled shearlet transform," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 9, pp. 3367–3378, 2019.

[16] W. Liu and Z. Wang, "A novel multi-focus image fusion method using multiscale shearing non-local guided averaging filter," *Signal Processing*, vol. 166, Article ID 107252, 2020.

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

[18] X. Guo, R. Nie, J. Cao, D. Zhou, and W. Qian, "Fully convolutional network-based multifocus image fusion," *Neural Computation*, vol. 30, no. 7, pp. 1775–1800, 2018.

[19] Y. Zang, D. Zhou, C. Wang, R. Nie, and Y. Guo, "UFA-FUSE: a novel deep supervised and hybrid model for multifocus image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–17, 2021.

[20] Z. Guan, X. Wang, R. Nie, S. Yu, and C. Wang, "NCDCN: multi-focus image fusion via nest connection and dilated convolution network," *Applied Intelligence*, vol. 52, no. 14, pp. 15883–15898, 2022.

[21] 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," *Information Fusion*, vol. 66, pp. 40–53, 2021.

[22] H. Li, W. Qian, R. Nie, J. Cao, and D. Xu, "Siamese conditional generative adversarial network for multi-focus image fusion," *Applied Intelligence*, vol. 53, no. 14, pp. 17492–17507, 2023.

[23] Y. Xiao, Z. Guo, P. Veelaert, and W. Philips, "DMDN: degradation model-based deep network for multi-focus image fusion," *Signal Processing: Image Communication*, vol. 101, Article ID 116554, 2022.

[24] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica.*, vol. 9, no. 7, pp. 1200–1217, 2022.

[25] X. Zhang, "Deep learning-based multi-focus image fusion: a survey and a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4819–4838, 2022.

[26] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[27] L. Yuan, Y. Chen, T. Wang, and Y. Weihao, "Tokens-to-token vit: training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, Montreal, BC, Canada, October 2021.

[28] W. Wang, E. Xie, X. Li, and D. P. Fan, "Pyramid vision transformer: a versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, Montreal, BC, Canada, October 2021.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning*, pp. 10347–10357, July 2021.

[30] Z. Liu, Y. Lin, and Y. Cao, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, Montreal, BC, Canada, October 2021.

[31] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with hilo attention," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[32] J. B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," 2019, https://arxiv.org/abs/1911.03584.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, July 2016.

[34] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, June 2022.

[35] W. Wang, E. Xie, X. Li, D. P. Fan, and L. Shao, "Pyramid vision transformer: a versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE International Conference on Computer Vision*, Montreal, BC, Canada, October 2021.

[36] H. Wu, B. Xiao, and N. Codella, "CvT: introducing convolutions to vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision*, Montreal, BC, Canada, October 2021.

[37] S. Xu, L. Ji, Z. Wang et al., "Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1561–1570, 2020.

[38] J. Cheng, X. Peng, X. Tang, W. Tu, and W. Xu, "MIFNet: a lightweight multiscale information fusion network," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5617–5642, 2022.

[39] Y. Ren, H. Ren, C. Shi et al., "Multistage semantic-aware image inpainting with stacked generator networks," *International Journal of Intelligent Systems*, vol. 37, no. 2, pp. 1599–1617, 2022.

[40] M. A. Islam, S. Jia, and N. D. Bruce, "How much position information do convolutional neural networks encode?" 2020, https://arxiv.org/abs/2001.08248.

[41] F. Milletari, N. Navab, and S. A. V. N. Ahmadi, "Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, Stanford, CL, USA, October 2016.

[42] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency

detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7254–7263, Seoul, Korea (South), October 2019.

[43] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5455–5463, Boston, MA, USA, June 2015.

[44] H. Ma, Q. Liao, J. Zhang, S. Liu, and J. H. Xue, "An $\alpha$-Matte Boundary Defocus Model-Based Cascaded Network for Multi-Focus Image Fusion$\alpha$-matte boundary defocus model-based cascaded network for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8668–8679, 2020.

[45] M. Haghighat and M. A. Razian, "Fast-FMI: non-reference image fusion metric," in *Proceedings of the 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–3, Astana, Kazakhstan, October 2014.

[46] J. Zhao, R. Laganiere, 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, 2007.

[47] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Optics Communications*, vol. 341, pp. 199–209, 2015.

[48] C. S. Xydeas and V. S. Petrovic, "Objective pixel-level image fusion performance measure," *Sensor Fusion: Architectures, Algorithms, and Applications IV*, vol. 4051, pp. 89–98, 2000.

[49] 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.

[50] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

[51] D. P. Bavirisetti, G. Xiao, J. Zhao, R. Dhuli, and G. Liu, "Multi-scale guided image and video fusion: a fast and efficient approach," *Circuits, Systems, and Signal Processing*, vol. 38, no. 12, pp. 5576–5605, 2019.

[52] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Processing: Image Communication*, vol. 72, pp. 35–46, 2019.

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

[54] 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.

[55] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12797–12804, 2020.

[56] Y. Liu, L. Wang, J. Cheng, and X. Chen, "Multiscale feature interactive network for multifocus image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–16, 2021.

[57] B. Ma, X. Yin, D. Wu, H. Shen, X. Ban, and Y. Wang, "End-to-end learning for simultaneously generating decision map and multi-focus image fusion result," *Neurocomputing*, vol. 470, pp. 204–216, 2022.