IEICE TRANS. INF. & SYST., VOL.E107–D, NO.11 NOVEMBER 2024

PAPER

Multi-Focus Image Fusion Algorithm Based on Multi-Task Learning and PS-ViT

Qinghua WU[†a)] ***and*** **Weitong LI**[†b)]**,** ***Nonmembers***

SUMMARY Multi-focus image fusion involves combining partially focused images of the same scene to create an all-in-focus image. Aiming at the problems of existing multi-focus image fusion algorithms that the benchmark image is difficult to obtain and the convolutional neural network focuses too much on the local region, a fusion algorithm that combines local and global feature encoding is proposed. Initially, we devise two self-supervised image reconstruction tasks and train an encoder-decoder network through multi-task learning. Subsequently, within the encoder, we merge the dense connection module with the PS-ViT module, enabling the network to utilize local and global information during feature extraction. Finally, to enhance the overall efficiency of the model, distinct loss functions are applied to each task. To preserve the more robust features from the original images, spatial frequency is employed during the fusion stage to obtain the feature map of the fused image. Experimental results demonstrate that, in comparison to twelve other prominent algorithms, our method exhibits good fusion performance in objective evaluation. Ten of the selected twelve evaluation metrics show an improvement of more than 0.28%. Additionally, it presents superior visual effects subjectively.

key words: *multi-focus image fusion, multi-task learning, PS-ViT, spatial frequency*

1. Introduction

Due to the depth-of-field (DOF) limitations of optical lenses, it is challenging for cameras to capture objects at different DOF in a single image [1]. Multi-focus image fusion (MFIF) is a significant image enhancement technique holding substantial application value in various domains. This approach combines distinct focus information in multiple source images of the same scene to create an all-in-focus image.

Over the past few years, deep learning-based algorithms have progressively emerged as the dominant force in image fusion. According to the adopted network architectures, they can be classified into methods based on auto-encoder, convolutional neural network (CNN), and generative adversarial network (GAN). Guo et al. [2] introduced a method named FuseGAN, which utilized conditional GAN (cGAN). This approach established an adversarial relationship by using human-annotated mask maps and generator-produced mask maps as positive and negative samples, which guided the generative network to enhance the detection of focus areas. Nevertheless, adversarial loss based on the L_2-norm might magnify image distinction, resulting in training instability. Due to the robust feature learning capability, CNN-based methods can extract more information when compared to traditional methods. Liu et al. [3] were the first to apply CNN to MFIF, learning the direct mapping between source images and focus map. This approach distinguished whether image patches were in focus, eliminating the need for manual design of activity level measurement and fusion rule. Guo et al. [4] proposed a fully convolutional network-based method that used the entire image for model training to acquire an initial decision map. Further refinement of the decision map was achieved through the fully connected conditional random fields. However, the approaches to generate decision maps for MFIF often struggle to classify regions near the focus/defocus boundary (FDB). Additionally, post-processing is frequently needed for generating decision maps, which introduces complexity to the methods. Simultaneously, owing to the lack of large-scale standard multi-focus image datasets for training, algorithms usually face overfitting issues or require intricate parameter optimization.

In addition, the convolutional and pooling operations of CNN may lead to the loss of positional information, making it challenging to capture global information. It is worth noting that, in multi-focus image fusion, global information can compensate for the lack of local information in textureless regions. In Depth-from-Focus [5], classical methods initially use focus measures (FMs) to extract sharpness and subsequently utilize Markov Random Fields (MRF) for semi-global belief updating. These methods require considerations of kernel size and assumptions about the depth's smoothness. Surh et al. [6] proposed a ring difference filter that combines the advantages of local and non-local FMs through a distinctive ring and disk structure. By incorporating information from a relatively large window of adjacent pixels and introducing a gap space to disregard certain areas of the window, this approach enhances robustness to noise and helps create more natural and smooth transitions in depth maps. Inspired by the above, we model global and local information in our network by introducing the PS-ViT module [7] into MFIF. The PS-ViT module is combined with the dense connection module in the encoder, which employs an iterative progressive sampling strategy. The model is trained on a natural image dataset using multi-task learning. In the fusion stage, the encoder extracts deep features from two source images. Subsequently, image metrics are applied to evaluate the activity level and merge the deep features. Ultimately, the decoder is utilized to reconstruct the fused image. Experimental results indicate that the proposed

Manuscript received February 27, 2024.
Manuscript revised May 24, 2024.
Manuscript publicized July 11, 2024.
[†]School of Information Engineering, Guangdong University of Technology, Guangdong, China.
a) E-mail: 2112203049@mail2.gdut.edu.cn
b) E-mail: liweitong@gdut.edu.cn
DOI: 10.1587/transinf.2024EDP7046

method demonstrates superior fusion performance in objective and subjective assessments. The contributions of this paper are as follows.

1. Considering the characteristics of multi-focus images, we introduce two image transformation techniques: Gaussian-Gamma transformation and PatchShuffle-NonLinear transformation. To improve the network's proficiency in capturing the distinctive features of multi-focus images, we train an encoder-decoder network through multi-task learning and use different loss functions for each task within the network.
2. To harness both local and global information during the feature extraction process, we integrate the dense connection module with the PS-ViT module in the encoder. This combination compensates for the limitation of CNN in capturing global information and guiding the network's attention to the focus areas of images. Moreover, the utilization of residual connections in the encoder output section helps prevent information loss, enabling the network to make better use of both low-level and high-level features.

2. Related Work

2.1 Spatial Domian Methods

Traditional MFIF algorithms can be divided into two categories, including transform domain and spatial domain methods. Transform domain-based methods convert source images into a designated feature space to acquire transformation coefficients. Following this, a fusion strategy is utilized to merge the coefficients, and the fused image is generated through inverse transformation. However, these algorithms may experience information loss during the transformation process, reducing the clarity of the fused images.

Spatial domain-based algorithms directly select pixels or image blocks from the source images that are relatively sharper for fusion. In contrast with transform domain-based algorithms, these methods can better retain the focus information of the source images. General image metrics in this category include energy of gradient (EOG), energy of lap (EOL), sum-modified-Laplacian (SML) and spatial frequency (SF) [8]. Among these measurements, SF indicates the level of grayscale variation in the image and can provide insights into the image's clarity. Li et al. [9] segmented the source images into several blocks of fixed size and used SF to assess each block's activity level. Then, a threshold-based fusion rule was employed to obtain the fused blocks.

2.2 Global Feature

Recently, significant progress has been achieved in image fusion, attributed to the powerful feature extraction and representation capabilities of deep learning. Zhang et al. [10] proposed IFCNN, a universal image fusion framework based on CNN. This method was trained in the end-to-end manner, eliminating the necessity for post-processing operations. For CNN, convolution operations typically pay more attention to local regions, and a global understanding of the entire image requires a series of time-consuming down-sampling and convolution processes. Throughout this process, there is a risk of losing edge information from the source images, and features like color and texture details in local regions may disrupt the global semantic information.

To address this issue, Xiao et al. [11] introduced a U-Net with global feature encoding designed for MFIF. This model incorporates a global feature pyramid extraction (GFPE) module and a global attention connection upsample (GACU) module, enabling the segmentation of focused and defocused regions from a global view. The GFPE module enables the network to capture image features at different scales, while the GACU module optimizes the feature upsampling process through global average pooling and attention weighting. The global information extracted by these two modules focuses more on hierarchical feature fusion from local to global. Qu et al. [12] introduced TransMEF, a novel network for multi-exposure image fusion that integrates CNN and transformer architecture. This approach considers the long-range dependencies present in the source images, thereby boosting the model's ability to extract features. By considering the relationships between all regions in the image, the self-attention mechanism can improve the model's perception of the global context. Therefore, in our method, we focus more on using the self-attention mechanism to enhance the network's understanding of the global structure of the image, capturing the spatial relationships and contextual information within the image.

2.3 Vision Transformer

The Vision Transformer (ViT) module [13] is predominantly employed in image classification tasks. The Transformer Encoder Layer comprises a multi-head self-attention (MHA) and a feed-forward unit. Given the input matrix X, queries matrix Q, keys matrix K and values matrix $V \in R^{L\times D}$, with L being the sequence length and D being the dimension, the output of self-attention mechanism is shown in Eq. (1).

$$Attn(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{1}$$

where Q^T represents the transpose of Q, and $softmax(\cdot)$ is the normalization procedure applied over each row of the input matrix. MHA divides attention computation into M subspaces, which can be expressed as:

$$\begin{aligned} MHA(X) &= Concat(H_1, H_2, \cdots, H_M)W^o \\ H_i &= Attn(XW_i^Q, XW_i^K, XW_i^V) \end{aligned} \tag{2}$$

where $W^o \in R^{D\times D}$ is a learnable linear projection. W_i^Q, W_i^K and $W_i^V \in R^{D\times \frac{D}{M}}$ are the linear projections for the queries, keys and values of the i-th head respectively. The feed-forward network consists of two linear transformation

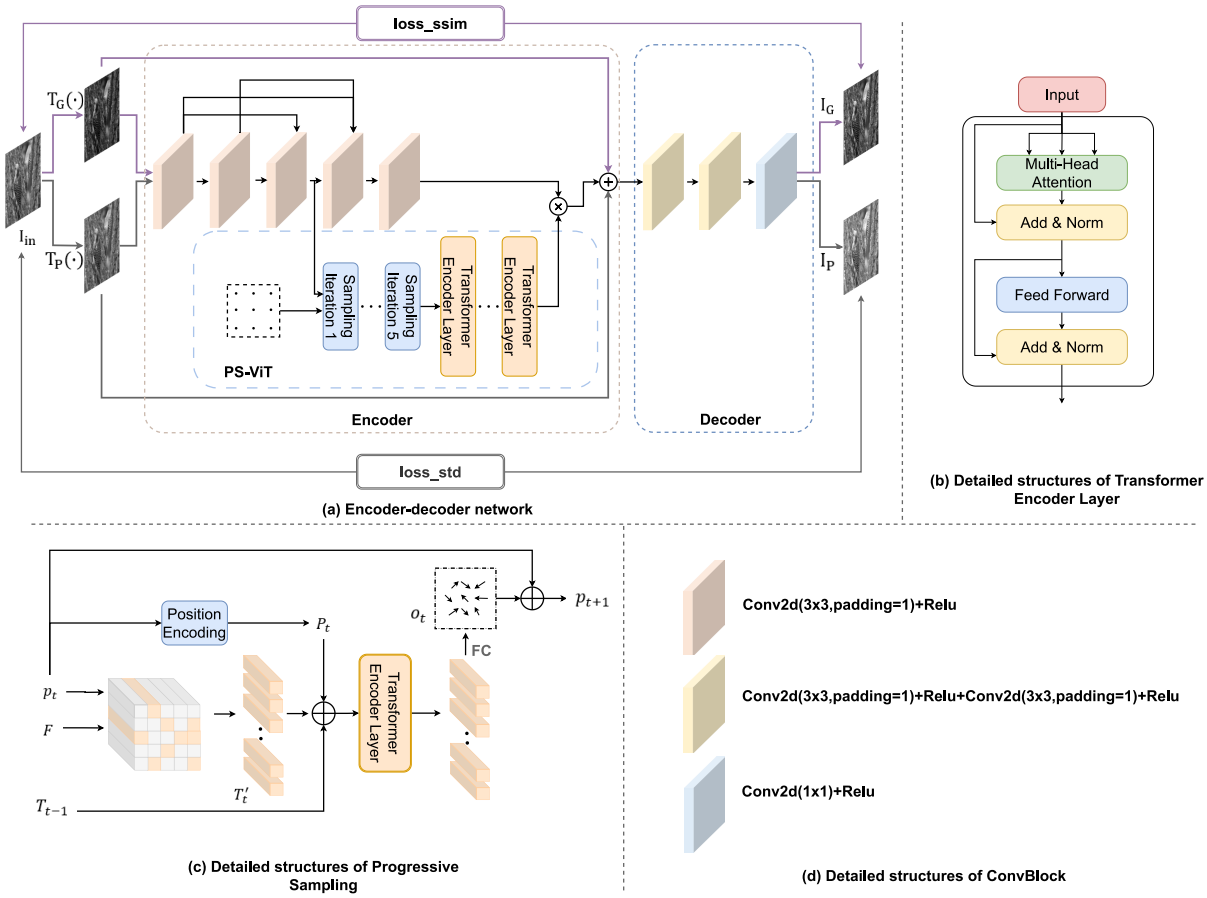


Fig. 1 Framework of proposed method. (a) The proposed encoder-decoder network. The purple lines represent the image reconstruction task based on the Gaussian-Gamma transformation, and the gray lines represent the image reconstruction task based on the PatchShuffle-NonLinear transformation. (b) Detailed structures of Transformer Encoder Layer. (c) Detailed structures of Progressive Sampling module. (d) Detailed structures of ConvBlock.

layers and a non-linear activation function, the latter being a Gaussian Error Linear Unit (GELU).

3. Proposed Method

Figure 1 shows our MFIF framework. In the training phase, we employ Gaussian-Gamma and PatchShuffle-NonLinear transformations on input images to facilitate better learning of features in multi-focus images. The model involves a multi-task learning strategy, training an encoder-decoder network with distinct loss functions for each task. To effectively leverage both local and global information in images and guide the network's attention towards focus regions, we integrate the dense connection module with the PS-ViT module in the encoder. During the fusion stage, SF is used to measure the activity level of features extracted by the encoder. The fusion rule (elementwise-max) is then applied to derive the feature mapping for the fused image. Ultimately, the decoder is employed for feature reconstruction to generate the fused image.

3.1 Architecture of Proposed Network

In the ViT module, it is general to segment images into tokens of fixed length and then utilize a transformer encoder to learn the relationships between the tokens, which may destroy the structure of the image and introduce interference signals. To solve this problem, Yue et al. [7] proposed the PS-ViT module, which employs an iterative progressive sampling strategy to locate discriminative regions, as illustrated in Fig. 1 (c). At each iteration, the current iteration's output tokens are used to predict a set of sampling offsets, which are then utilized to update the sampling positions for the next iteration. We integrate it into MFIF, as shown in Fig. 1 (a). Within the encoder, we merge the dense connection module with the PS-ViT module, concatenating the feature mappings derived from both modules. Subsequently, these interconnected feature mappings are fed into the decoder to capture local and global information about the image.

The dense connection module is a primary component for learning local information in images. It comprises five convolutional layers linked in sequence. Dense connections

are incorporated into the first four convolutional layers, allowing the output of each layer to be transmitted to all subsequent layers. This design maximizes the utilization of information from earlier convolutional layers, enhancing the network's capability to tackle intricate tasks. It facilitates valuable information and gradient propagation, mitigates vanishing gradient during model training and contributes to parameter reduction. Every convolutional layer employs a 3×3 convolutional kernel and incorporates a ReLU activation function. Leveraging the effectiveness of the convolutional operator in modeling spatial local context, the deep features extracted by the initial three convolutional layers in the dense connection module serve as the input feature maps for the first iteration. These feature maps are subsequently fed into the PS-ViT module.

PS-ViT is constructed with two key modules, Progressive Sampling and Transformer, dedicated to grasping global information from images. In the initial iteration of the Progressive Sampling module, sampling positions are determined through uniform interval sampling. The sampling tokens from the input feature map, position embeddings corresponding to the current sampling positions, and output tokens from the previous iteration are combined element-wise. This combined information is then input into a Transformer Encoder Layer to generate the output tokens for the current iteration. Formally,

$$\begin{aligned} P_t &= W_t p_t \\ X_t &= T_t^{'} \oplus P_t \oplus T_{t-1} \\ T_t &= Transformer(X_t),\ t \in \{1, \ldots, N\} \end{aligned} \tag{3}$$

where W_t is the linear transformation matrix that projects the sampling points p_t to the positional embeddings P_t, all iterations share the same W_t. $T_t^{'}$ signifies the sampled tokens at the iteration t. T_{t-1} is the tokens predicted by the Progressive Sampling module at the iteration $t-1$ and $\oplus$ indicates the element-wise addition. As positional information is already incorporated into the output tokens from the last iteration during sampling, there is no requirement to introduce positional embeddings when the tokens are fed into the Transformer module. The iteration number of the Progressive Sampling module is 5, and the PS-ViT module comprises 14 Transformer Encoder Layers.

Moreover, we apply residual connections to the encoder output section, where input information is directly added to the output. This design facilitates the direct passage of lower-level feature information to higher layers, helping mitigate issues like vanishing and exploding gradients. The decoder is composed of three convolutional layers. The initial two layers utilize a 3×3 convolutional kernel with a ReLU activation function, while the last layer employs a 1×1 convolutional kernel.

3.2 Multi-Task Learning

Throughout the training phase, we use two distinct processing techniques on input images to enhance the network's ability to learn features of multi-focus images. The input images are all 8-bit images. The Gaussian-Gamma transformation is utilized for acquiring scene content and brightness information, while the PatchShuffle-NonLinear transformation is employed to grasp structural information and contrast details.

(1) Gaussian-Gamma transformation

The source image I_{in} is subjected to a blurring operation through Gaussian filtering, yielding a blurred image I_b. Formally,

$$\begin{aligned} I_b &= G * I_{in} \\ G(x, y) &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \end{aligned} \tag{4}$$

where $*$ signifies the convolution operation, G stands for the Gaussian kernel, and σ represents the standard deviation of the Gaussian filter. We set σ as a randomly sampled value from a uniform distribution in the range $[0.5, 1.0]$. To preserve abundant information and maintain uniform brightness [14] in the fused image, following Gaussian blur, we utilize a Gamma-based transformation to adjust the brightness of the source images. This approach enables the network to learn scene content and brightness information from images with diverse blur and brightness levels. The Gamma-based transformation is expressed as:

$$\widetilde{u} = 255 \times \left(\frac{u}{255}\right)^{\gamma} \tag{5}$$

while u and $\widetilde{u}$ denote the original and transformed pixel values respectively, while γ is a randomly selected value uniformly sampled from the interval $[1 + 0.5 \times \sigma, 1 + 2 \times \sigma]$.

(2) PatchShuffle-NonLinear transformation

We use a regularization method named PatchShuffle [15] to process the source image I_{in}. After randomly choosing ten image blocks of size $h \times w$ from I_{in}, the elements within this block undergo shuffling. The values h and w are randomly sampled from the set of positive integers in the range $[1, 25]$. The random permutations within the patch ensure that the transformed images retain nearly identical global structures to the original ones while introducing diverse local variations. To make the edge detail information richer, nonlinear contrast enhancement is applied after PatchShuffle, modifying the brightness differences between various regions in the image. This adjustment assists the network in learning structural and contrast information. The nonlinear contrast enhancement is presented in Eq. (6):

$$\widetilde{v} = 255 \times \alpha \times \log_2\left(1 + \frac{v}{255}\right) \tag{6}$$

where $\widetilde{v}$ and v represent the transformed and original pixel values respectively, while α is a randomly chosen value drawn uniformly from the range $[0.9, 1.0]$.

Figure 1 (a) illustrates the Gaussian-Gamma transformation denoted as $T_G(\cdot)$, and the PatchShuffle-NonLinear

Fig. 2 Transformations of original images. The first row shows the original images, the second row shows the images after the Gaussian-Gamma transformation, and the third row shows the images after the PatchShuffle-NonLinear transformation.

transformation denoted as $T_P(\cdot)$. The transformed outcomes are shown in Fig. 2, with the first row displaying the input image, the second row showing the image after the Gaussian-Gamma transformation, and the third row exhibiting the image following the PatchShuffle-NonLinear transformation. The red boxes in the third row highlight several representative subregions after the PatchShuffle transformation, where the pixels within the regions are visibly shuffled. The Gaussian-Gamma transformation can alter the blurriness of the original images, simulate different levels of defocus effect, adjust image brightness, and highlight details in bright areas, thereby aiding the model in learning the differences between focused and defocused regions. The PatchShuffle-NonLinear transformation introduces rich variations locally in the images, where patches at the same original position share the same weights across different iterations, and adjusts image contrast. It can help the model capture common features between images with different focuses, maintains scene coherence and visual consistency, and extract subtle features and edge information.

3.3 Loss Function

Multi-task learning simultaneously learns multiple related tasks to improve the model's performance on each task by transferring information between them, allowing the network to learn more generalized feature representations. Based on the sharing of inputs and outputs among different tasks, multi-task learning can be classified into three different categories: multi-input single-output (MISO), single-input multi-output (SIMO), and multi-input multi-output (MIMO) [16]. In the MISO case, multiple data sources map to a single output. In the SIMO case, all tasks share the same input to predict different types of outputs. In the MIMO case, multiple input sources are used to predict multiple outputs.

Our network adopts the MIMO mode of multi-task learning, targeting two types of inputs to generate two outputs similar to the source images. To enhance the model's ability to learn the unique characteristics of each task, specific loss functions are applied to individual tasks. Gaussian filtering blurs the details in the images, while the structural similarity (SSIM) loss [17] measures structural similarity by comparing the mean, variance, and covariance of pixels within a local window, effectively reflecting changes in image details. Therefore, we use the SSIM loss L_{ssim} in the image reconstruction task based on the Gaussian-Gamma transformation. The rearrangement of image patches causes changes in the local structure of the image, and the standard deviation loss, by considering the range of pixel distribution, helps the model learn and quantify the uncertainty introduced by structural adjustments. Thus, we use the standard deviation loss L_{std} in the image reconstruction task based on the PatchShuffle-NonLinear transformation. We construct the overall loss using a weighted sum to unify the loss scales and optimize both tasks simultaneously. The overall loss function is shown in Eq. (7):

$$Loss = L_{ssim} + L_{std} \tag{7}$$

L_{ssim} quantifies the structural dissimilarity between the reconstructed image I_G and the input image I_{in}. Its function expression is:

$$\begin{aligned} &L_{ssim} = 1 - SSIM(I_G, I_{in}) \\ &SSIM(I_G, I_{in}) = \\ &\sum_{g,x} \frac{2\mu_g\mu_x + C_1}{\mu_g^2 + \mu_x^2 + C_1} \cdot \frac{2\sigma_g\sigma_x + C_2}{\sigma_g^2 + \sigma_x^2 + C_2} \cdot \frac{\sigma_{gx} + C_3}{\sigma_g\sigma_x + C_3} \end{aligned} \tag{8}$$

where $\frac{2\mu_g\mu_x+C_1}{\mu_g^2+\mu_x^2+C_1}$, $\frac{2\sigma_g\sigma_x+C_2}{\sigma_g^2+\sigma_x^2+C_2}$ and $\frac{\sigma_{gx}+C_3}{\sigma_g\sigma_x+C_3}$ evaluates the similarity in brightness, contrast and structural information. g and x correspond to image blocks of I_G and I_{in} within a sliding window. σ_{gx} represents the covariance between g and x, while σ_g and σ_x denote the standard deviations of g and x, respectively. Additionally, μ_g and μ_x signify the means of g and x. The constants C_1, C_2, and C_3 are introduced to prevent division by zero.

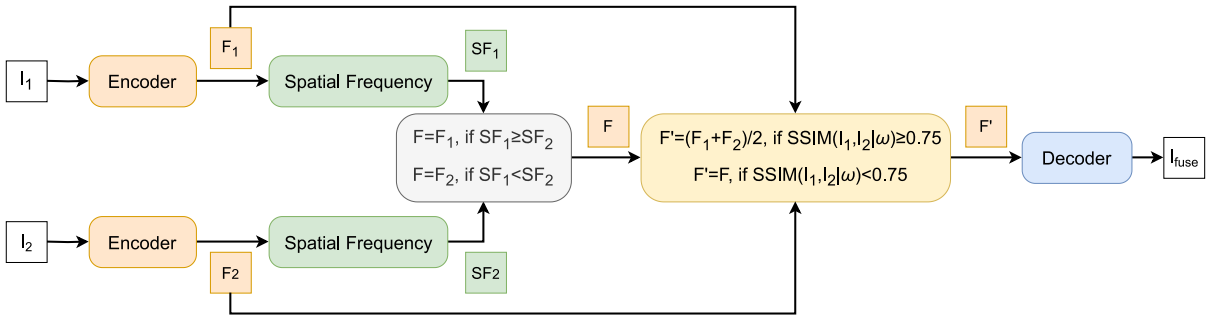


Fig. 3 The image fusion architecture.

L_{std} captures the diversity in data distribution between the reconstructed image I_P of size $m \times n$ and the input image I_{in}. Its function expression is:

$$\begin{aligned} I_{diff}(i,j) &= |I_P(i,j) - I_{in}(i,j)| \\ \mu &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{diff}(i,j) \\ L_{std} &= \sqrt{\frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}[I_{diff}(i,j)-\mu]^2} \end{aligned} \tag{9}$$

Employing standard deviation from the difference image of I_P and I_{in} as a loss function provides insight into the extent of dissimilarity between two images, emphasizing subtle distinctions rather than just average differences. Throughout the optimization process, model parameters are adjusted by minimizing L_{std} to enhance the similarity between I_P and I_{in}.

3.4 Fusion Rule

Figure 3 illustrates the specific architecture of image fusion. During the fusion stage, the SF is calculated on a pixel-by-pixel basis to measure the activity level. Let $F(x,y)$ denote the feature vector extracted by the encoder for each pixel, where (x,y) represents the coordinates of the pixel within the image. The SF is specifically expressed as:

$$\begin{aligned} &RF(x,y) = \\ &\sqrt{\sum_{-r\le a,b\le r}[F(x+a,y+b)-F(x+a,y+b-1)]^2} \\ &CF(x,y) = \\ &\sqrt{\sum_{-r\le a,b\le r}[F(x+a,y+b)-F(x+a-1,y+b)]^2} \\ &SF(x,y) = \sqrt{\frac{(RF(x,y))^2+(CF(x,y))^2}{(2r+1)^2}} \end{aligned} \tag{10}$$

Where RF and CF correspond to the frequencies of the row and column vectors respectively, with r denoting the kernel radius.

The encoder extracts high-dimensional features for every pixel in the image, capturing its intricate details. When two source images I_1 and I_2 are fed into the pre-trained encoder, it produces two deep feature maps F_1 and F_2. The activity levels of F_1 and F_2 are measured using SF with $r = 5$, and the maximum activity level strategy is applied to determine the feature mapping for each pixel, resulting in the initial fused feature map F. This strategy ensures the retention of robust feature information from the source images. Subsequently, different operations are conducted in various local regions based on the similarity between the source images. In redundant regions, where $\mathrm{SSIM}(I_1, I_2 \mid \omega)$ is greater than or equal to 0.75, the average of F_1 and F_2 is taken as the local feature. In complementary regions, where $\mathrm{SSIM}(I_1, I_2 \mid \omega)$ is less than 0.75, the local feature is F. This process generates the final fused feature map F′. The parameter ω is a 11×11 window that moves pixel by pixel from the top left to the bottom right, and in each sliding window, the pixel considered is located at the center of the window. The redundant regions indicate areas with similar or repeated information between the two source images, while complementary regions signify areas with distinct yet complementary content across the two source images [18]. Finally, the decoder reconstructs F′ to generate the fused image.

4. Experiments

4.1 Experimental Settings

The training of the encoder-decoder network is conducted on the PASCAL VOC 2012 dataset [19], with 13701 images as a training set and 3424 images for validation. All images are converted to grayscale and resized to 256×256. During the training phase, the ADAM optimizer is used along with the cosine annealing learning rate adjustment strategy. The initial learning rate is configured as 1×10^{-4}, weight decay is set at 0.0005, and the batch size is defined as 4. For the evaluation in the testing stage, the Lytro [20] and MFI-WHU [21] datasets are utilized. The Lytro dataset contains 20 pairs of multi-focus images, while the MFI-WHU dataset comprises 120 multi-focus images pairs. The network's code implementation is developed using the PyTorch framework, and training is executed on an NVIDIA RTX 3090 GPU.

4.2 Managing RGB Input

For color image fusion, the initial step involves converting source images from RGB to the YCbCr color space. Following this, our method is employed to fuse the Y-channel of the source images. The information in the Cb and Cr channels is then fused using a conventional weighted averaging approach. The formula is as follows:

$$C = \frac{C_1|C_1 - \tau| + C_2|C_2 - \tau|}{|C_1 - \tau| + |C_2 - \tau|} \tag{11}$$

where the notation $|\cdot|$ signifies the absolute value function. C represents either the Cb or Cr channel of the fused image, while C_1 and C_2 correspond to the Cb or Cr channels of the two source images. The parameter τ is set as 128. The ultimate step involves the conversion of the fusion images back to the RGB color space.

4.3 Objective Image Fusion Quality Metrics

To conduct a comprehensive comparison with other fusion methods, we have chosen 12 objective evaluation metrics across five aspects, which are (1) information theory-based metrics including entropy (EN) [22], mutual information (MI) [23], fusion artifacts ($N^{AB/F}$) [24] and Tsallis entropy-based metric (TE) [25], (2) image feature-based metrics involving average gradient (AG) [26], spatial frequency (SF) [8], standard deviation (SD) [27], edge intensity metric (EI) [28] and linear index of fuzziness (LIF) [29], (3) as an image structure similarity-based metric, structural similarity index measure (SSIM) [17], (4) as a correlation-related metric, correlation coefficient (CC) [30], (5) as a human perception-inspired fusion metric, visual information fidelity (VIF) [31].

We compared the proposed method with 12 representative MFIF methods. Among them, NSCT [32] and MWGF [33] are classified as transform domain-based approaches, while GFDF [34] and BRW [35] are spatial domain-based methods. Additionally, CNN [3], IFCNN [10], SESF [36], SDNet [37], U2Fusion [14], GACN [38], MFIFGAN [39] and R-PSNN [40] are deep learning-based methodologies. All comparative methods are configured with default parameters and utilize the training models provided by the original authors, ensuring conformity with the outcomes presented in the original papers.

4.4 Qualitative Comparisons

To qualitatively illustrate the effectiveness of our approach, we choose four representative images. The fusion results are presented in Fig. 4 and Fig. 5. Regions with differences in the fused images are delineated with rectangles and further magnified for detailed examination. Upon observation, it becomes evident that our method excels in preserving details of the source images, which included information in the vicinity of FDB. Furthermore, it effectively retains texture information, contributing to an enhancement in overall image quality.

Analysis of Fig. 4 reveals that in the first set of results, NSCT exhibits unclear edges around the fence. Moreover, the fence in MWGF appears blurred, failing to retain foreground details effectively. Misclassification near FDB has a notable impact on the fusion images, particularly for GFDF,

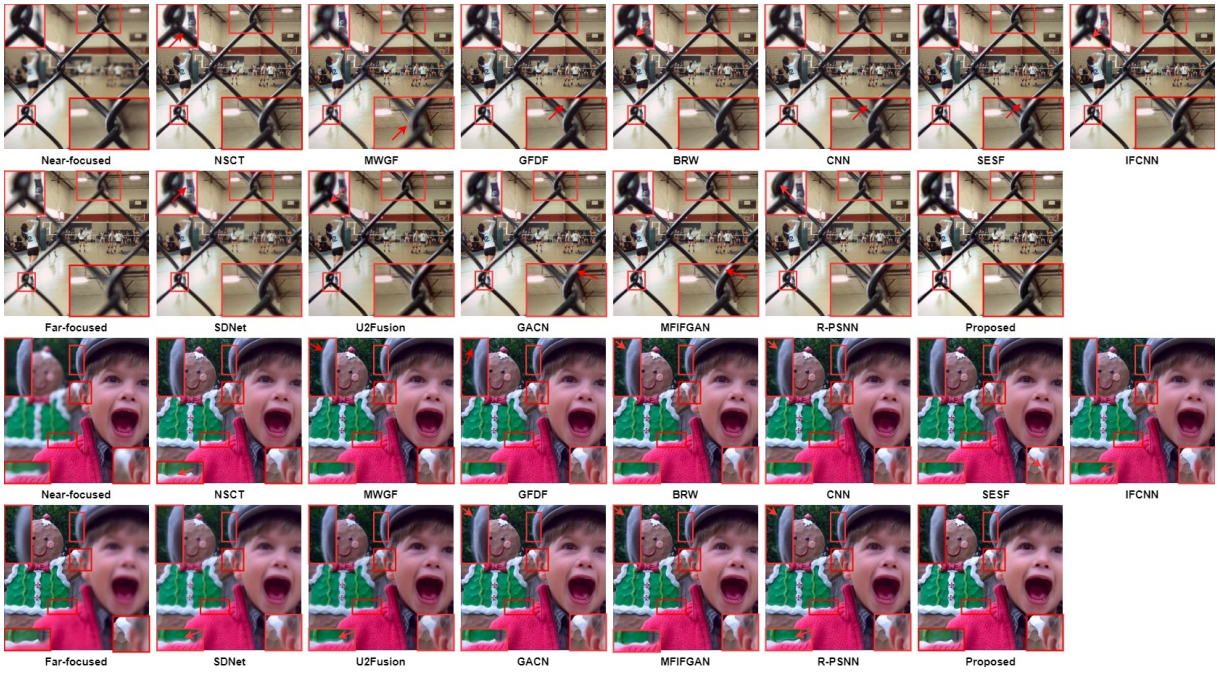


Fig. 4 Qualitative results on the Lytro dataset.

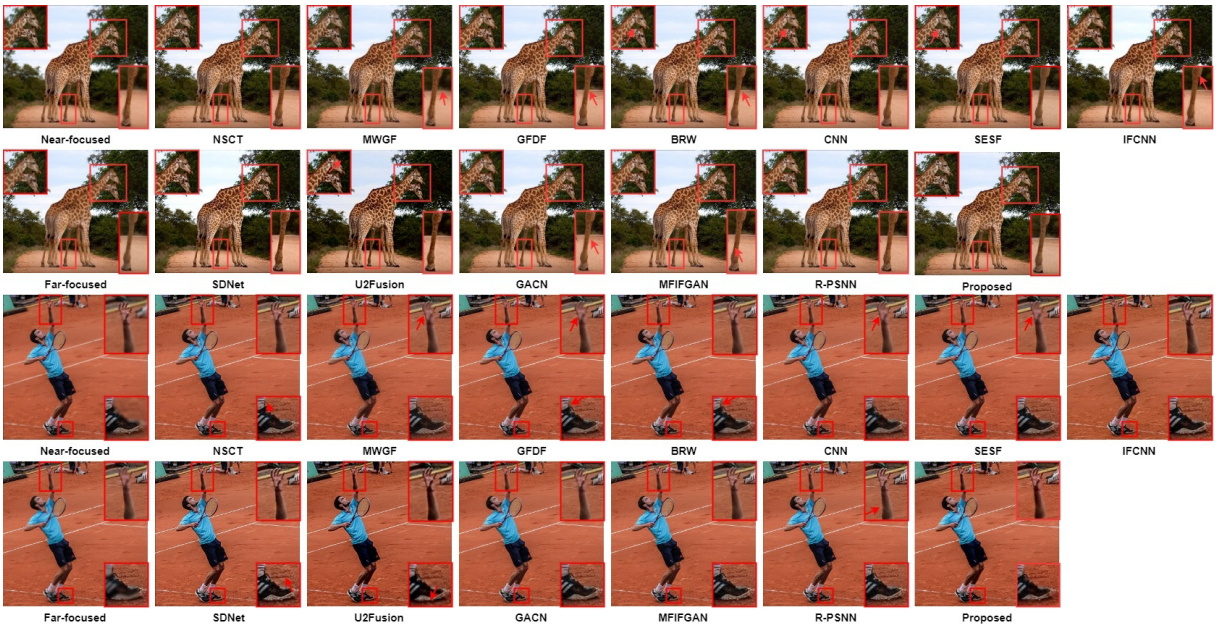


Fig. 5 Qualitative results on the MFI-WHU dataset.

CNN and SESF, leading to the omission of a pipe on the ceiling. The shoes at the base of the fence lack clarity in BRW, IFCNN and U2Fuison, indicating inadequate preservation of minor clear areas from the source images. Additionally, the clarity of socks adjacent to the fence is compromised in the fusion result of SDNet. GACN, MFIFGAN, and R-PSNN exhibit a loss of some details in the fence, accompanied by white artifacts along its edge. In the second set of results, artifacts are evident at the shoulder edge in the fused images of NSCT, IFCNN, SDNet, U2Fusion, and R-PSNN. Additionally, the boundary of the child's hat appears blurred in MWDF, GFDF, BRW, CNN, and GACN. SESF exhibits minor areas of missing details around the child's ear. Notably, MFIFGAN and R-PSNN display prominent white artifacts along the edge of the child's hat in their fusion results. In contrast, our method excels in preserving details near FDB, providing excellent overall visual perception, and minimizing the impact of blurring.

Analyzing the first group of results from Fig. 5, it is evident that the fusion result of MWGF has blurred background information, while the fusion outcomes of GFDF, BRW, IFCNN and GACN lack the intricate patterns on the giraffe's leg. Furthermore, fused images of BRW, CNN, and SESF reveal small blocks' identification errors, leading to an unclear far-focused region under the giraffe's neck. The fusion result of U2Fusion lacks detail in the leaves, and the fused image of MFIFGAN shows blurriness on the giraffe's legs. Moving to the second group of results, misjudgment in the palm area is evident in MWGF, GFDF, CNN, and SESF, resulting in a blurred appearance. Additionally, the shoelace near the sock is fuzzy in the fusion results of GFDF and BRW. Although NSCT provides clearer details, it still exhibits artifacts around the shoelace. Compared to the source images, the grooves on the floor appear darker in the fusion result of SDNet. U2Fusion introduces black shadows beneath the shoe. Lastly, a slight blurring is observed at the edge of the arm in R-PSNN's fusion result. In contrast, the proposed method excels in detecting the focused area and simultaneously exhibits superior retention of the texture details throughout the entire image.

4.5 Quantitative Comparisons

Table 1 presents the average evaluation metrics for all fused images in the Lytro dataset. Obviously, compared to 12 representative algorithms, our approach exhibits significant advantages in information entropy, image features and human perception. Relative to the SESF algorithm, our method demonstrates a decrease of 2.63% and 0.71% in the $N^{AB/F}$ and LIF metrics, respectively. Conversely, the SD and VIF metrics experience increases of 7.96% and 13.65%. Contrasting with the R-PSNN algorithm, our approach yields improvements of 7.81%, 5.75%, and 5.86% in the SF, AG, and EI metrics, respectively. Compared to the NSCT algorithm, our method exhibits an improvement of 0.39% in both the EN and MI metrics. In comparison to the IFCNN algorithm, our approach achieves a 0.28% increase in the TE metric. Regarding the SSIM metric, our method ranks third.

Table 2 provides the average evaluation metrics for all fused images in the MFI-WHU dataset. Like the Lytro dataset, our method exhibits superiority over other comparative algorithms in metrics related to information entropy and image features. Compared with the SDNet algorithm, our approach yields a 0.45% reduction in the LIF metric, accompanied by improvements of 2.27%, 1.68%, 1.54%, and 0.80% in the AG, SF, EI, and SD metrics, respectively. Relative to the U2Fusion algorithm, our method results in enhancements of 0.47% in both the EN and MI metrics, with

Table 1 Average scores of fusion results based on Lytro dataset by all algorithms on 12 metrics. The best, the second best, and the third best results are highlighted in bold, double underlining, and underlining, respectively.

Algorithm	EN	MI	$N^{AB/F}$	TE	AG	SF	EI	SD	LIF	SSIM	CC	VIF
NSCT	7.5348	15.0696	0.0185	7.5266	6.7696	19.2488	70.1789	57.4127	0.4097	0.8446	0.9785	1.1255
MWGF	7.5322	15.0644	0.0127	7.5260	6.6415	19.0193	68.9482	57.2624	0.4088	0.8404	0.9776	1.1140
GFDF	7.5311	15.0622	0.0083	7.5246	6.7938	19.3094	70.5255	57.5319	0.4078	0.8422	0.9778	1.1352
BRW	7.5310	15.0618	0.0091	7.5243	6.8002	19.3249	70.5825	57.5398	0.4076	0.8421	0.9778	1.1360
CNN	7.5305	15.0610	0.0100	7.5241	6.7589	19.1880	70.1749	57.4676	0.4080	0.8432	0.9780	1.1289
IFCNN	7.5319	15.0639	0.0153	7.5291	6.8195	19.4004	70.7663	57.5502	0.4081	0.8450	0.9787	1.1338
SESF	7.5323	15.0646	0.0076	7.5258	6.8331	19.4255	70.9117	57.5934	0.4072	0.8407	0.9777	1.1415
SDNet	7.4754	14.9507	0.0275	7.4720	5.8999	16.9492	60.5919	55.3862	0.4209	**0.8618**	**0.9814**	0.9210
U2Fusion	7.4082	14.8164	0.0244	7.4082	5.8394	15.3326	61.8445	53.5791	0.4142	0.8492	0.9786	1.0340
GACN	7.5313	15.0626	0.0082	7.5249	6.8075	19.3255	70.6468	57.4866	0.4081	0.8416	0.9778	1.1329
MFIFGAN	7.5334	15.0667	0.0092	7.5272	6.8408	19.4098	70.9398	57.5388	0.4077	0.8395	0.9776	1.1375
R-PSNN	7.5328	15.0657	0.0151	7.5258	6.8600	19.4569	71.0385	57.4902	0.4083	0.8399	0.9778	1.1329
Proposed	**7.5640**	**15.1279**	**0.0074**	**7.5550**	**7.2547**	**20.9762**	**75.2041**	**62.1752**	**0.4043**	0.8481	0.9783	**1.2973**

Table 2 Average scores of fusion results based on MFI-WHU dataset by all algorithms on 12 metrics. The best, the second best, and the third best results are highlighted in bold, double underlining, and underlining, respectively.

Algorithm	EN	MI	$N^{AB/F}$	TE	AG	SF	EI	SD	LIF	SSIM	CC	VIF
NSCT	7.3263	14.6527	0.0178	7.3208	8.3230	26.7769	78.4451	52.4596	0.4822	0.8331	0.9786	1.1092
MWGF	7.3091	14.6181	0.0169	7.3051	7.9888	26.1615	75.4133	52.0373	0.4844	0.8295	0.9781	1.0737
GFDF	7.3180	14.6360	0.0130	7.3136	8.1952	26.5082	77.3963	52.3168	0.4831	0.8324	0.9786	1.1017
BRW	7.3189	14.6378	0.0128	7.3146	8.2203	26.5594	77.5987	52.3233	0.4829	0.8323	0.9786	1.1032
CNN	7.3140	14.6279	0.0149	7.3096	8.1045	26.2848	76.6053	52.2734	0.4835	0.8330	0.9787	1.0967
IFCNN	7.3286	14.6572	0.0173	7.3254	8.2823	26.1820	78.9436	52.6258	0.4807	**0.8357**	0.9782	1.1742
SESF	7.3183	14.6367	0.0126	7.3132	8.2015	26.6899	77.4357	52.4506	0.4819	0.8303	0.9784	1.1097
SDNet	7.3144	14.6288	**0.0119**	7.2998	8.5774	27.8297	81.0814	58.5839	0.4475	0.7960	**0.9806**	1.3102
U2Fusion	7.3297	14.6595	0.0193	7.3297	7.9727	21.8984	80.8168	55.7383	0.4612	0.7831	0.9742	**1.4661**
GACN	7.3127	14.6255	0.0125	7.3083	8.1202	26.5089	76.7208	52.3417	0.4834	0.8315	0.9785	1.1041
MFIFGAN	7.3207	14.6414	0.0125	7.3164	8.2917	26.8236	78.1827	52.3779	0.4826	0.8299	0.9782	1.1087
R-PSNN	7.3246	14.6491	0.0200	7.3204	8.3389	26.7987	78.5667	52.4422	0.4819	0.8328	0.9786	1.1178
Proposed	**7.3643**	**14.7286**	0.0173	**7.3540**	**8.7724**	**28.2968**	**82.3313**	**59.0524**	**0.4455**	0.8349	0.9804	1.3164

Table 3 Results of the ablation study for PS-ViT and Res (Residual connections) using 20% of the training data. The best, the second best, and the third best results are highlighted in bold, double underlining, and underlining, respectively.

PS-ViT	Res	EN	MI	$N^{AB/F}$	TE	AG	SF	EI	SD	LIF	SSIM	CC	VIF
		7.5546	15.1092	0.0100	7.5447	7.0069	20.3437	72.9451	**62.6894**	0.4036	0.8332	0.9791	1.2881
✓		7.5574	15.1150	0.0081	7.5475	7.1791	20.6591	74.6425	62.6891	**0.4026**	0.8412	**0.9796**	**1.3074**
	✓	7.5611	15.1221	0.0103	7.5505	7.1225	20.1768	74.0113	62.2386	0.4084	0.8404	0.9787	1.2794
✓	✓	**7.5635**	**15.1270**	**0.0079**	**7.5535**	**7.2074**	**20.6984**	**74.7925**	62.4951	0.4051	**0.8436**	0.9784	1.2985

a 0.33% increase in the TE metric. In terms of the metrics SSIM, CC and VIF, the proposed method is the second best. A decrease in the $N^{AB/F}$ metric suggests a reduction in introduced artifacts during the process of fusion. A small LIF indicates that the enhancement of the fused image is good. The other metrics are positively oriented, with higher values indicative of superior performance.

Conclusions drawn from these results indicate that, although our method performs mediocrely on the correlation-related metric, it includes more information and minimizes artifacts. Additionally, our approach stands out in visual information fidelity, indicating that the fused images effectively preserve intricate texture details and align closely with human visual perception. In summary, the proposed method outperforms other comparison approaches in objective assessments.

5. Ablation Experiments

5.1 Ablation Study for PS-ViT and Residual Connections

To verify the effectiveness of the PS-ViT module and residual connections, we conducted ablation experiments using 20% of the training data. The results are presented in Table 3. It can be observed that the addition of the PS-ViT module significantly improves metrics based on image features, indicating that focusing on global features contributes to enriching texture details in the images. The introduction of residual connections shows a noticeable enhancement in information theory-based metrics, implying that the network's learning of residuals helps prevent information loss. Notably, the joint incorporation of the PS-ViT module and residual connections yields the best overall performance. In

Table 4 Results of the ablation study for self-supervised image reconstruction tasks based on Gaussian-Gamma (GG) and PatchShuffle-NonLinear (PN) using 20% of the training data. The best, the second best, and the third best results are highlighted in bold, double underlining, and underlining, respectively.

GG	PL	EN	MI	$N^{AB/F}$	TE	AG	SF	EI	SD	LIF	SSIM	CC	VIF
		7.5553	15.1107	0.0088	7.5471	7.1366	20.6527	74.0369	62.4132	0.3867	0.8391	0.9784	1.2940
✓		7.5572	15.1145	0.0106	7.5489	**7.2342**	20.6425	**74.9105**	59.8271	0.4520	0.8380	0.9751	1.2512
	✓	7.5032	15.0064	0.0080	7.4951	7.1826	**21.0591**	74.4928	**64.2412**	**0.3692**	0.8248	**0.9792**	**1.3362**
✓	✓	**7.5635**	**15.1270**	**0.0079**	**7.5535**	7.2074	20.6984	74.7925	62.4951	0.4051	**0.8436**	0.9784	1.2985

particular, the metric $N^{AB/F}$ experiences a reduction of 21%, while metrics AG, SF and EI demonstrate improvements of 2.86%, 1.74% and 2.53%, respectively.

5.2 Ablation Study for Two Specific Self-Supervised Image Reconstruction Tasks

In this ablation experiment, we affirmed the effectiveness of each self-supervised image reconstruction task and highlighted the advantages of executing them through multi-task learning. As shown in Table 4, the experimental results indicate that the Gaussian-Gamma transformation yields a substantial improvement in information theory-based and image feature-based metrics, suggesting its contribution to the network's understanding of scene content and brightness information. Notably, metrics EN and AG exhibit increases of 0.03% and 1.37%, respectively. The PatchShuffle-NonLinear transformation demonstrates a significant enhancement in metrics based on correlation and human perception, in which metrics CC and VIF are improved by 0.08% and 3.26% individually, indicating its role in facilitating the network's learning of structural-semantic and contrast information. The concurrent execution of both tasks achieves the best overall performance, with four metrics achieving optimal values and three metrics reaching suboptimal values.

6. Conclusion

This paper introduces an innovative MFIF algorithm that integrates the encoding of local and global features. We adopt the multi-task learning approach to train an encoder-decoder network, where the encoder incorporates a dense connection module and a PS-ViT module. This design allows the network to efficiently capture both local and global information in images concurrently. Additionally, leveraging the characteristics of multi-focus images, we have introduced two self-supervised tasks for image reconstruction. In the training phase, the network performs both tasks simultaneously and uses a different loss function for each task. This strategy is instrumental in facilitating the network to capture the distinctive features of multi-focus images. Experimental results confirm that our approach, when compared to prevalent algorithms, successfully preserves intricate details from the source images and significantly improves the clarity of the fused images. Since the simplicity of the employed loss functions in this paper, a crucial future task is devising a robust loss function to enhance the network's capability for feature extraction and improve the edge details in the image. Additionally, our approach doesn't account for the defocus spread effect in the modeling process. Consequently, how to model it from a distribution perspective to enhance the visual quality of the fused images is also a direction for further exploration.

References

[1] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," IEEE Trans. Pattern Anal. Mach. Intell., vol.44, no.9, pp.4819–4838, 2021. DOI: 10.1109/TPAMI.2021.3078906.

[2] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," IEEE Trans. Multimed., vol.21, no.8, pp.1982–1996, 2019. DOI: 10.1109/TMM.2019.2895292.

[3] 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, 2017. DOI: 10.1016/j.inffus.2016.12.001.

[4] X. Guo, R. Nie, J. Cao, D. Zhou, and W. Qian, "Fully convolutional network-based multifocus image fusion," Neural Comput., vol.30, no.7, pp.1775–1800, 2018. DOI: 10.1162/neco_a_01098.

[5] S.K. Nayar and Y. Nakagawa, "Shape from focus," IEEE Trans. Pattern Anal. Mach. Intell., vol.16, no.8, pp.824–831, 1994. DOI: 10.1109/34.308479.

[6] J. Surh, H.-G. Jeon, Y. Park, S. Im, H. Ha, and I.S. Kweon, "Noise robust depth from focus using a ring difference filter," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.6328–6337, 2017. DOI: 10.1109/CVPR.2017.262.

[7] X. Yue, S. Sun, Z. Kuang, M. Wei, P. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," Proc. IEEE/CVF International Conference on Computer Vision, pp.387–396, 2021. DOI: 10.1109/ICCV48922.2021.00044.

[8] A.M. Eskicioglu and P.S. Fisher, "Image quality measures and their performance," IEEE Trans. Commun., vol.43, no.12, pp.2959–2965, 1995. DOI: 10.1109/26.477498.

[9] S. Li, J.T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," Inf. Fusion, vol.2, no.3, pp.169–176, 2001. DOI: 10.1016/S1566-2535(01)00038-0.

[10] 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, 2020. DOI: 10.1016/j.inffus.2019.07.011.

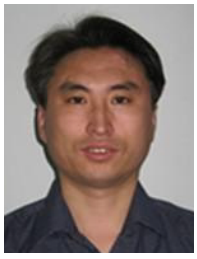
[11] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding U-Net (GEU-NET) for multi-focus image fusion," IEEE Trans. Image Process., vol.30, pp.163–175, 2020. DOI: 10.1109/TIP.2020.3033158.

[12] L. Qu, S. Liu, M. Wang, and Z. Song, "TransMEF: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," Proc. AAAI Conference on Artificial Intelligence, vol.36, no.2, pp.2126–2134, 2022. DOI: 10.1609/aai.v36i2.20109.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020. DOI: 10.48550/arXiv.2010.11929.

[14] 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, 2020. DOI: 10.1109/TPAMI.2020.3012548.
[15] G. Kang, X. Dong, L. Zheng, and Y. Yang, "Patchshuffle regularization," arXiv preprint arXiv:1707.07103, 2017. DOI: 10.48550/arXiv.1707.07103.
[16] K.-H. Thung and C.-Y. Wee, "A brief review on multi-task learning," Multimed. Tools. Appl., vol.77, no.22, pp.29705–29725, 2018. DOI: 10.1007/s11042-018-6463-x.
[17] 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, 2004. DOI: 10.1109/TIP.2003.819861.
[18] S. Li, R. Hong, and X. Wu, "A novel similarity based quality metric for image fusion," 2008 International Conference on Audio, Language and Image Processing, pp.167–172, IEEE, 2008. DOI: 10.1109/icalip.2008.4589989.
[19] M. Everingham, S.M. Ali Eslami, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," Int. J. Comput. Vis., vol.111, pp.98–136, 2015. DOI: 10.1007/s11263-014-0733-5.
[20] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," Inf. Fusion, vol.25, pp.72–84, 2015. DOI: 10.1016/j.inffus.2014.10.004.
[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," Inf. Fusion, vol.66, pp.40–53, 2021. DOI: 10.1016/j.inffus.2020.08.022.
[22] J.W. Roberts, J.A. Van Aardt, and F.B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," J. Appl. Remote Sens., vol.2, no.1, 023522, 2008. DOI: 10.1117/1.2945910.
[23] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," Electron. Lett., vol.38, no.7, pp.313–315, 2002. DOI: 10.1049/el:20020212
[24] V. Petrovic and C. Xydeas, "Objective image fusion performance characterisation," Tenth IEEE International Conference on Computer Vision (ICCV '05) Volume 1, pp.1866–1871, IEEE, 2005. DOI: 10.1109/ICCV.2005.175.
[25] N. Cvejic, C.N. Canagarajah, and D.R. Bull, "Image fusion metric based on mutual information and Tsallis entropy," Electron. Lett., vol.42, no.11, pp.626–627, 2006. DOI: 10.1049/el:20060693.
[26] 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," Opt. Commun., vol.341, pp.199–209, 2015. DOI: 10.1016/j.optcom.2014.12.032.
[27] Y.-J. Rao, "In-fibre Bragg grating sensors," Meas. Sci. Technol., vol.8, no.4, 355, 1997. DOI: 10.1088/0957-0233/8/4/002.
[28] X. Luo, Z. Zhang, C. Zhang, and X. Wu, "Multi-focus image fusion using HOSVD and edge intensity," J. Vis. Commun. Image Represent., vol.45, pp.46–61, 2017. DOI: 10.1016/j.jvcir.2017.02.006.
[29] X. Bai, F. Zhou, and B. Xue, "Noise-suppressed image enhancement using multiscale top-hat selection transform through region extraction," Appl. Optics, vol.51, no.3, pp.338–347, 2012. DOI: 10.1364/AO.51.000338.
[30] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," Inf. Fusion, vol.48, pp.11–26, 2019. DOI: 10.1016/j.inffus.2018.09.004.
[31] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," Inf. Fusion, vol.14, no.2, pp.127–135, 2013. DOI: 10.1016/j.inffus.2011.08.002.
[32] B. Yang, S. Li, and F. Sun, "Image fusion using nonsubsampled contourlet transform," Fourth International Conference on Image and Graphics (ICIG 2007), pp.719–724, IEEE, 2007. DOI: 10.1109/ICIG.2007.124.
[33] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," Inf. Fusion, vol.20, pp.60–72, 2014. DOI: 10.1016/j.inffus.2013.11.005.
[34] 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. DOI: 10.1016/j.image.2018.12.004.
[35] J. Ma, Z. Zhou, B. Wang, L. Miao, and H. Zong, "Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps," Neurocomputing, vol.335, pp.9–20, 2019. DOI: 10.1016/j.neucom.2019.01.048.
[36] 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, pp.5793–5804, 2021. DOI: 10.1007/s00521-020-05358-9.
[37] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," Int. J. Comput. Vis., vol.129, pp.2761–2785, 2021. DOI: 10.1007/s11263-021-01501-8.
[38] 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. DOI: 10.1016/j.neucom.2021.10.115.
[39] Y. Wang, S. Xu, J. Liu, Z. Zhao, C. Zhang, and J. Zhang, "MFIF-GAN: A new generative adversarial network for multi-focus image fusion," Signal Processing: Image Communication, vol.96, 116295, 2021. DOI: 10.1016/j.image.2021.116295.
[40] L. Jiang, H. Fan, J. Li, and C. Tu, "Pseudo-Siamese residual atrous pyramid network for multi-focus image fusion," IET Image Processing, vol.15, no.13, pp.3304–3317, 2021. DOI: 10.1049/ipr2.12326.

Qinghua Wu received the B.E. degree in Guangdong University of Technology, China, in 2022. She is currently pursuing the Master's degree at Guangdong University of Technology, China. Her current research direction is image fusion.



Weitong Li received the M.E. degree in Xidian University, China, in 2000 and Ph.D degree in Harbin Engineering University, China, in 2005. His current research interests include data fusion, image fusion, image quality assessment and others.