PATN: Polarized Attention based Transformer Network for Multi-focus image fusion

Pan Wu[1], Zhen Hua[1,2*], and Jinjiang Li[1,2]
[1] School of Computer Science and Technology, Shandong Technology and Business University
Yantai, 264005 China
[e-mail:2020410012@sdtbu.edu.cn]
[2] School of Information and electronic engineering, Shandong Technology and Business University
Yantai, 264005 China
[e-mail: huazhen@sdtbu.edu.cn]
*Corresponding author: Zhen Hua

Received May 30, 2022; revised December 19, 2022; accepted March 22, 2023; published April 30, 2023

Abstract

In this paper, we propose a framework for multi-focus image fusion called PATN. In our approach, by aggregating deep features extracted based on the U-type Transformer mechanism and shallow features extracted using the PSA module, we make PATN feed both long-range image texture information and focus on local detail information of the image. Meanwhile, the edge-preserving information value of the fused image is enhanced using a dense residual block containing the Sobel gradient operator, and three loss functions are introduced to retain more source image texture information. PATN is compared with 17 more advanced MFIF methods on three datasets to verify the effectiveness and robustness of PATN.

Keywords: Image fusion, Multi-focus, Polarized Attention, Transformer

http://doi.org/10.3837/tiis.2023.04.011 **ISSN : 1976-7277**

1. Introduction

Since the development of image processing, multi-focus image fusion has existed as a branch of image fusion. Full-focus images are difficult to obtain due to the depth of field of optical shooting lenses or natural environmental conditions. For this problem, multi-focus image fusion technology plays a great role, which extracts the shallow features of these images from a pair or a group of captured images, adopts the corresponding algorithm to fuse the extracted focusing features, smooths out the boundaries of the focusing and out-of-focus regions, retains the basic information in the source image, and obtains the final fully-focused clear image to ensure that all target objects are in focus.

The fused images obtained from image fusion are used by researchers in fields such as image segmentation and target recognition, and image fusion techniques and methods in other fields also promote each other. Multi-focus image fusion methods based on consistency verification play a role in image segmentation, where consistency verification yields the initial decision map of the corresponding source image, and image segmentation techniques are used to process the target image in the image that needs to be segmented.

From these two points, multi-focus image fusion as a branch in the field of image fusion, the methods in this field are in urgent need of improvement. Infrared and visible image fusion, multi-exposure image fusion, multi-spectral image fusion, multi-modal image fusion, and multi-focus image fusion contribute to the development of the broad category of image fusion, and the commonality of all these branches is that the source images are synthesized with higher quality and better results by image fusion algorithms. Multi-focus image fusion is characterized by the fact that the input pair of source images is of the same scene, and there are focused and out-of-focus regions in the scene, and the focused and out-of-focus regions are realized to be fully complementary. Due to the limitation of equipment and knowledge, many previous methods are performed based on traditional algorithms, and the final results obtained are poor.

Fig. 1. Test results of PATN on the Lytro dataset. From left to right are: Near-focused, Far-focused, Decision-Map, and Fused images.

With the development of deep learning in the field of digital image processing, convolutional neural networks have played an important role in the field of image fusion. For large-scale training, CNNs are more suitable for image vision processing tasks because of the unique superiority of convolutional neural networks compared with traditional methods, so that the images processed by CNN-based networks contain more information in the source images. At this stage, most of the existing MFIF methods are based on CNNs, but CNNs

cannot establish long-range dependencies and cannot focus on global feature information, while the Transformer mechanism can make up for this deficiency of CNNs, so the Transformer mechanism is introduced into the image fusion field. This ensures that the features of the source image are perfectly mapped to the fused image, and the fused image finally obtained contains more original information. At present, the main problems of multi-focus image fusion are: 1) the CNN-based MFIF method cannot focus on the global features, resulting in the fused image retaining less overall original information; 2) the MFIF method that introduces the attention mechanism focuses too much on local features, resulting in the final image having more prominent local features. However, the dividing line of the focus/off-focus region is obvious and the visual effect is poor.

To address the current limitations of CNN in the field of multi-focus image fusion, we introduce Transformer to multi-focus image fusion and propose a network based on the U-Net framework combining CNN and Transformer to handle the multi-focus image fusion task. The model proposed in this paper has the following four advantages.

1) An efficient architecture is designed for multi-focus image fusion. The architecture uses a U-shaped Transformer as the core component for multi-focus image fusion, shallow feature encoding and high-level feature recovery decoding space. That is, the proposed network is used for the fusion task of multi-focused images based on the U-Net framework using the Transformer capable of focusing on the operating principle of long-range features, combined with the characteristics of CNN networks, the proposed network is used for the fusion task of multi-focus images to obtain the desired fused images.

2) To ensure that global features and local features are attended to simultaneously, we introduce a polarized self-attention mechanism to attend to the detailed features in the original image. Multi-focus image fusion can be regarded as both a classification task and a regression task, introducing an attention mechanism that focuses on both channel features and spatial features, using sliding windows with Transformer to stitch the upper window features, making attention calculations on pixel points, expecting pixel targets to be clearly extracted, and thus achieving a high-quality pixel regression task.

3) In order to enhance the extraction of decision map and image edge information by the network, we use the Sobel operator with direction as the gradient operator of the network to calculate the approximate gradient of the image gray scale. Because the whole fusion image is the final result of the fusion of a set of source images, including both the clear target in the focus region and the junction line between the out-of-focus region and the focus region, in order to ensure a good visual effect, the fusion image should be as smooth and unobtrusive as possible, so the Sobel operator is used as the gradient operator to sense the change of image grayscale while describing the detail information on the fine-grained space.

4) In order to comprehensively examine the performance of the proposed network, we simultaneously evaluate the network in terms of objective quantitative metrics and subjective visual perception, and compare it with a variety of existing MFIF methods on three datasets, and the comparison results all reveal the natural advantages of our framework for the task of generating fused images.

2. Related Works

The background related to our proposed method and what contributed to the progress of our work will be presented in this subsection, mainly including traditional methods in the field of multi-focus image fusion, deep learning methods and the U-based Transformer method used in this paper.

2.1 Multi-focus image fusion based on traditional methods

MFIF methods based on traditional methods can be basically divided into three categories: transform domain methods[1], [2], [3] that follow transform rules, spatial domain methods[4], [5], [6] that process images based on the same space, and methods that combine transform and spatial domains.Methods based on the transform domain are generally easy to implement and are more robust to processing noise [7]. Discrete Wavelet Transform (DWT) [8], higher-order singular value decomposition and Non-Subsampled Contour wavelet Tansform (NSCT) [9] are all methods for multi-focus image fusion processing based on the transform domain. Such methods often suffer from the problem that the weight coefficients in the training process are difficult to optimize, and the obtained fused images are also prone to The obtained fused images are also prone to blurring and low contrast problems. The spatial domain-based methods are used to use image blocks for multi-focus image fusion processing, where the input image is cut into smaller image blocks, and the fused image is obtained by using the features in the image plus the fusion strategy and adopting the designed fusion rules. Methods based on image denoising [10], PCA transform methods, and guided filtering (GF) [11] are all methods for image fusion based on the spatial domain. Block effects and boundary blurring problems are more likely to occur in such methods.

Among the transform domain methods, the most common are multi-scale transforms such as wavelet transform and Laplace transform, which are prone to produce more pronounced flicker or jitter in the fused images. The final fusion effect of the spatial domain-based methods depends largely on the accuracy of the mask image obtained during processing, which generally includes the HIS transform and PCA transform, etc. The HIS transform is generally applicable to color images with relatively high resolution, while the PCA transform has a better fusion effect under the premise that the color of the source image is closer, which is the limitation of these methods.

2.2 Multi-focus image fusion based on deep learning

Due to the powerful performance of deep learning in visual processing tasks, researchers have introduced deep learning to the field of multi-focus image fusion. Deep learning algorithms learn a large amount of data that needs to be processed by machines[12], gradually fitting an optimized model whose dynamic convolution process can easily extract the features that need to be learned. According to the dynamic learning process of deep learning[13], supervised and unsupervised models have been developed one after another, and the field of multi-focused image fusion presents a promising form.

The parallel processing of the input data using the direct mapping relationship between the source image and the focus is the hallmark of CNN applications in the field of multi-focus image fusion. GAN networks, two networks confront each other and learn from each other, which train data based on unsupervised methods to generate fused images directly, but the generalizability and practicality of the networks need to be considered. A pixel-level CNN network (p-CNN) based on domain information is proposed to solve the problem of artifacts in fused images, in which the pixel focus level in the source image is precisely measured by a focus/off-focus mask, which alleviates the artifacts in multi-focus fused images to some extent. The DSIFT method starts from SIFT descriptors, and the pixel activity level is reflected by local feature descriptors. The refinement of the decision map using local focus and matching features facilitates the reflection of local features of the fused image. Based on the rethinking of image fusion, the PMGI method is used as a general image fusion framework in the field of multi-focus image fusion, which unifies the relationship between the intensity ratio and texture

information of the source image by exchanging the information of the extracted gradient direction and that of the image intensity direction.

Convolutional neural network-based methods can focus on the corresponding features in multi-focus image fusion tasks with the help of attention mechanism, but the limitation of CNN is that it cannot provide long-term global attention to the features, which can easily lead to the emergence of problems such as large-scale blurring or small-area artifacts in the fused image, while some algorithmic methods tend to pay excessive attention to some details in the image and ignore the overall visual effect of the fused image.

2.3 Transformer Model

Transformer has achieved great success in the field of natural language processing and has caught up with CNN structures in image processing tasks. transformer finds different angular relationships between different input sequences by using a multi-headed attention mechanism, which in turn builds long-term dependence on the model. Transformer's flexible use of the sliding window mechanism to segment the input image into smaller image blocks is effective for downstream tasks of image tasks. The use of Transformer in the areas of target detection and image segmentation, where the focus is on downstream tasks, is even more self-evident.

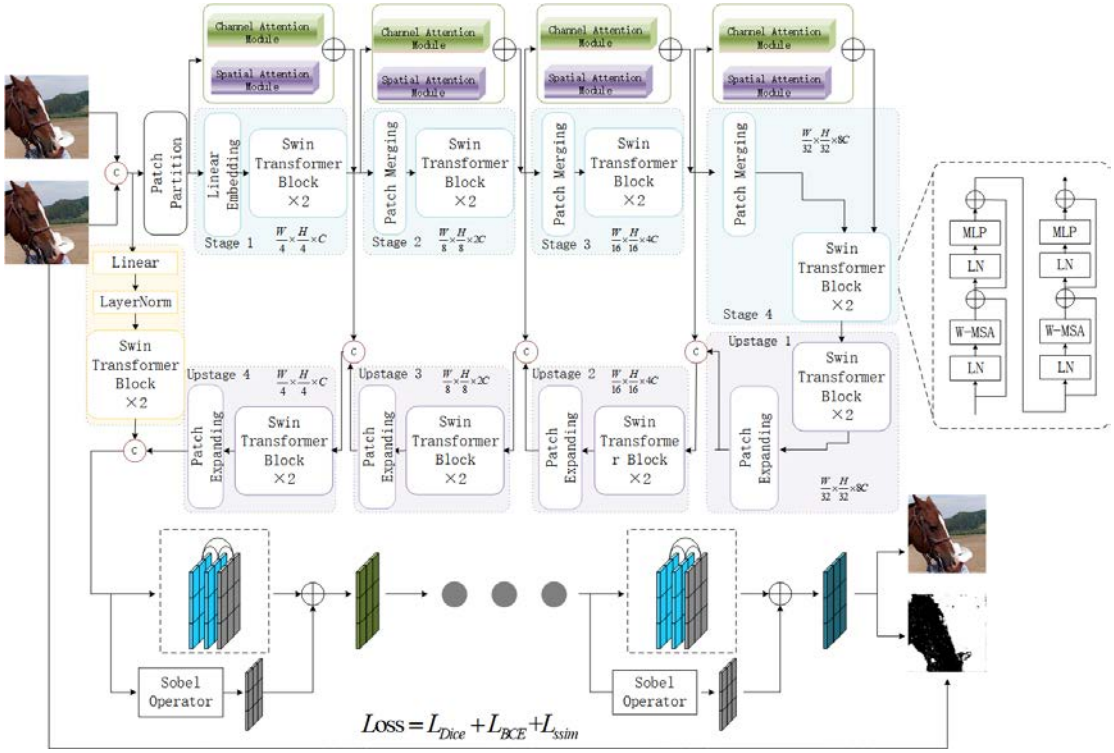


Fig. 2. Overall architecture of PATN. PATN mainly contains Transformer for extracting deep features, channel attention and spatial attention for extracting local features, and continuous dense residual blocks for enhancing edge information.

Some of the researches tried to solve the inherent defects of CNN networks by deepening the depth of the network model, but the results were not satisfactory. Transformer adopts a different calculation method and operation method from CNN, which has awakened the vitality in the field of computer vision and opened new ideas in the direction of CV. From there,

Transformer has created a boom in the field of computer vision, such as medical image segmentation, vehicle lane line detection, target detection, etc. In these fields, Transformer has proven that the results obtained using the Transformer mechanism are better, no less than the processing results of CNN, or even better than the results using CNN methods.

Based on the powerful modeling capability of the Transformer mechanism and taking advantage of the U-Net architecture for image tasks, inspired by both, we introduce the Transformer [14] based on the U-Net framework for image fusion to handle multi-focus image fusion tasks, ensuring that the obtained fused images have the advantage of both global and local features.

3. Method

The proposed approach will be described in detail in this section, including the modules used in the network architecture and the overall network architecture design, and the proposed overall network architecture is shown in **Fig. 2**.

3.1 Framework overview

As shown in **Fig. 2**, our proposed network architecture is based on the U-Net combined with Transformer architecture. We use a large dataset of segmented images to train the network. In the feature extraction phase, we use four identical stages for feature extraction of the input source images, all four stages include Linear Embedding and two consecutive Swin Transformer[15] blocks. Before entering these four stages, the designed network performs a cutting operation on the image after Concatenate to cut the image into smaller image blocks, while processing the height and width of the input image, which is reflected in the dimensionality of the Stage 1 in our network becomes $\frac{H}{4}\times\frac{W}{4}\times C$, and the corresponding dimensionality of each subsequent stage is reduced to achieve the purpose of reducing the resolution of the processed feature map. To enhance the extraction of features in the source image by the network, we add PSA polarized attention at each stage of feature extraction, a mechanism shown in **Fig. 2**, which contains both channel attention and spatial attention to capture the pixel information contained in the image from multiple perspectives. PSA [16] has two implementations, parallel and series, and its parallel approach is borrowed in our network to enhance the extraction of pixel information.

In the fusion phase of the network, we use four ascending dimensional descending channel operations to fuse the extracted features. Similar to the feature extraction phase, all four phases consist of a Patch Expanding and two consecutive Swin Transformer blocks, and these architectures are designed to ensure that information between different windows in the two-layer module can be interacted with, while different queries can share the same set of keys when doing self-attentive computation, which enhances the usefulness of the network. The information from the feature extraction stage and the fusion stage is fused and fed to the feature enhancement stage, which consists of four consecutive and dense residual modules using Sobel operator as the gradient operator [17]. The use of the dense residual module facilitates the complementary image information, and the Sobel operator can enhance the edge information of the image from two gradient directions to enrich the deep detail feature information in the feature map at fine granularity for the purpose of image reconstruction. The process of feature enhancement is described in detail in Section 3.3, and here we just describe the general process of network fusion of images.

3.2 Transformer-Based U-Net Framework

3.2.1 U-Net framework for multi-focus image fusion

The U-Net-based framework for multi-focus image fusion is shown in **Fig. 3**. We input a pair of source images to the feature extraction stage of the network after Concatenate, and this stage performs the upscaling operation on the feature channel dimensions to deepen the channel dimension extraction of features, refine the features to be extracted, and enrich the multi-channel feature map information. Secondly, the obtained feature maps are input to the feature fusion stage of the network, where the extracted features are upsampled to recover the features contained in the final desired fused image, and the channel dimension is downsampled at the same time. Finally, the integrated feature maps are transported to the feature enhancement stage, and the edge features of the images are enhanced using the corresponding feature enhancement operations, so that the fused images contain more original information and are more consistent with human visual perception.

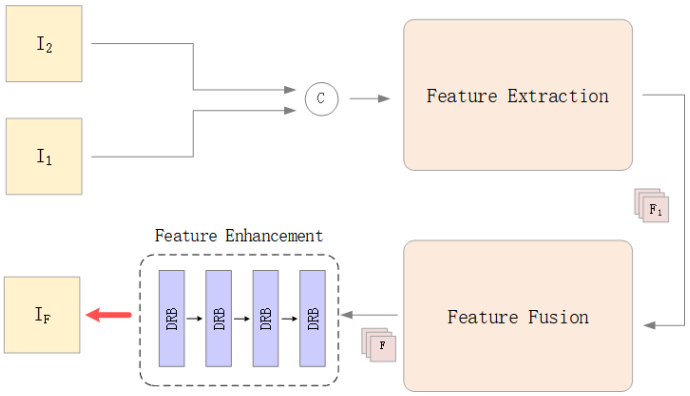


Fig. 3. General procedure of PATN architecture. PATN consists of feature extraction phase, feature fusion phase and feature enhancement phase.

3.2.2 A-Trans Block: a module for feature extraction

Inspired by Swin Transformer and PSA polarized attention mechanism, we use a parallel PSA mechanism to perform residual attention attention on the overall module containing Patch Embedding and two consecutive Swin Transformer Blocks, i.e., Attention-Transformer Block. This module combines the advantages of Transformer and CNN mechanisms to simultaneously feature attention to the phase of feature extraction from channel attention and spatial attention. Specifically, the CNN module consists of parallel PSAs. This module combines the advantages of Transformer and CNN mechanisms to simultaneously feature attention to the phase of feature extraction from channel attention and spatial attention. Specifically, the CNN module consists of parallel PSAs. The dimensionality of the input feature map $X \in \Re^{C_{in} \times H \times W}$ before entering the PSA is the same as that entering the Transformer, and the features enter the PSA internally to maintain high resolution, PSA is introduced to enhance the nonlinearity of the Transformer module and fit the output distribution with higher delicacy. And at the output, the feature tensor is $Z \in \Re^{C_{out} \times H \times W}$ $Z \in \Re^{C_{out} \times H \times W}$, that is, in the first stage of feature extraction, $C_{in} = C_{out} = 96$, after the PSA mechanism $Z = PSA(X) \times X$, the detailed PSA processing process channel number change

is shown in **Fig. 4**. Among them, for the channel attention, the input feature map is set as $PSA^{ch}(X) \in \Re^{c\times 1\times 1}$,and the feature map is calculated as:

$$PSA^{ch}(X) = F_{SG}[W_{z|\theta_1}((\sigma_1(W_v(X)) \times F_{soft\max}(\sigma_2(W_q(X)))))] \tag{1}$$

where , W_q, W_v represents the convolution kernel 1×1 for the convolution layer, σ_1 , σ_2 represents the reshape operation, $F_{soft\max}$ represents the softmax operator, and $\times$ represents the dot product operation in the matrix, while $F_{soft\max}(X) = \sum_{i=1}^{k} \frac{e^{x_i}}{\sum_{m=1}^{k} e^{x_m}} x_i$. In Stage 1, all other feature channels are $C/2 = 48$, and only the corresponding channel branching channels are $Z^{ch} = PSA^{ch}(X) \mathrm{e}^{ch} X \in \Re^{C\times H\times W}$, and e^{ch} denotes the operator in the channel multiplication operation.

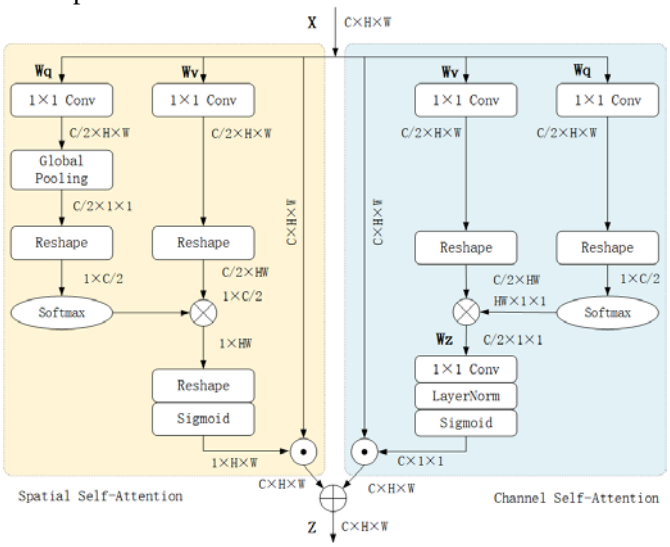


Fig. 4. PSA module, which consists of channel attention and spatial attention mechanisms, with the detailed process channel number variation shown in **Fig. 4**.

Also, for the other spatial channel of PSA note that the calculation is:

$$F_{GlobalP}(X) = \frac{1}{H \times W} \sum_{H}^{i=1} \sum_{W}^{j=1} X(:, i, j) \tag{2}$$

$$F = F_{soft\max}(\sigma_1(F_{GlobalP}(W_q(X)))) \tag{3}$$

$$PSA^{sp}(X) = F_{SG}[\sigma_3 F \times \sigma_2(W_v(X))] \tag{4}$$

where , W_q, W_v represents the convolution layer with a convolutional kernel of 1×1, σ_i represents the i reshape operations, F_{GP} represents a global average pooling operation, and for the spatial branch output of the feature map is calculated as:

$$Z^{sp} = PSA^{sp}(X) \mathrm{e}^{sp} X \in \Re^{C\times H\times W} \tag{5}$$

Similar to the channel branch attention, e^{sp} denotes the operator of the spatial multiplication operation. The final attention branch parallel composition we use is:

$$Z^{ch} = PSA^{ch}(X) \mathrm{e}^{ch} X \tag{6}$$

$$Z^{sp} = PSA^{sp}(X) \mathrm{e}^{sp} X \tag{7}$$

$$PSA(X) = Z^{ch} + Z^{sp} \tag{8}$$

After the PSA module, the input features are added with the features that go through the Transformer module to complete a stage of extracting features, using four stages to extract features for the whole feature extraction stage. We set the Transformer Block layer parameter for each stage as $layers = (2,2,2,2)$, and the structure of two consecutive Swin Transformer Blocks is shown in **Fig. 2**. Among them, the module is calculated as follows:

$$\hat{z}^{l} = W - MSA(LN(z^{l-1})) + z^{l-1} \tag{9}$$

$$z^{l} = MLP(LN(\hat{z}^{l})) + \hat{z}^{l} \tag{10}$$

$$\hat{z}^{l+1} = SW - MSA(LN(z^{l})) + z^{l} \tag{11}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{12}$$

Where $\hat{z}^{l}$ and z^{l} represent the output of the sliding window W-MSA and MLP module layer l, respectively, while the self-attentive mechanism can be calculated by the following equation:

$$Attention(Q,K,V) = SoftMax(\frac{QK^{T}}{\sqrt{d}} + B)V \tag{13}$$

Where $Q,K,V \in R^{M^2\times d}$ represent the query, key and value matrices respectively, M^2 denote the number of patches in the window, d denotes the dimension of the query or key, and value is from the bias matrix $\hat{B} \in R^{(2M-1)\times(2M+1)}$.

3.2.3 DRB: Dense Residual Module

In order to perform the enhancement operation on the feature map output from the feature fusion stage, we introduce the dense residual module containing the Sobel operator as the gradient operator. The residual module consists of a convolutional layer with a convolutional kernel size of 3×3 and the Leaky ReLU activation function for successive dense connection operations, the obtained densely connected features are convolved by 1×1, and the image gradient is computed using the conventional Sobel operator. The feature map obtained from the gradient calculation is convolved by 1×1, and the residuals are summed with the densely

connected module to input the edge-enhanced feature map F^{m+1} , which is calculated as:

$$F^{m+1} = DRB(F^m) = conv^2(F^m) \oplus conv(\nabla F^m) \tag{14}$$

Where, F^m is the input feature map, $conv^2$ represents two consecutive cascaded convolution operations in the map, and ∇ is the Sobel gradient operation used.

The final feature enhancement phase of the network, we use four consecutive dense residual modules, combined with the loss function we use, to output the final fused image and its decision map, and the DRB module is shown in **Fig. 5** in its exact composition.

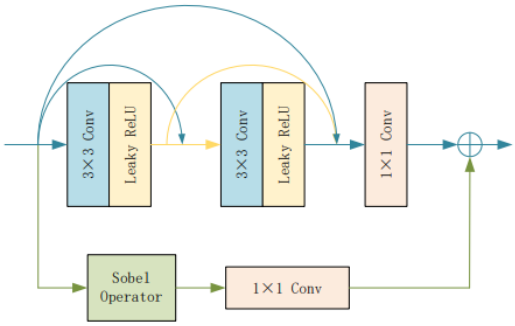


Fig. 5. DRB dense residual module. The module consists of a Sobel gradient operator, a 3×3 convolution, a Leaky ReLU activation function, and a 1×1 convolution dense connection.

3.3 Loss function

The purpose of processing the multi-focus image fusion task is to obtain high quality fused images to achieve our function optimization objective, but the optimization objective is difficult to measure directly, so the loss function is a metric that reflects the optimal objective. In the multi-focus image fusion task, the performance of the fused image includes the ambiguity of the image pixels and the similarity of the pixel information, we introduce three loss function optimization proposed network models based on the consideration of the image information itself, and the overall consideration of the network, in short, the loss function can be expressed as:

$$L_{sum} = L_{Dice} + L_{BCE} + L_{ssim} \tag{15}$$

In the above equation, we introduce the Dice loss for medical image segmentation. The Dice loss does not require a new trade-off for the judgment of imbalance, while the similarity between two images can be calculated to predict the pixel activity level, and then the focus/out-of-focus status of the target can be judged. For images in multi-focus, including focused and out-of-focus regions, the multi-focus image can be regarded as a dichotomous image, and then the BCE dichotomous loss optimization model is introduced. The BCE loss values are obtained by entropy, summation, and averaging operations for the predicted corresponding position points, as calculated in (16).

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c s_i^c}{\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^{c2} + \sum_{i=1}^{N}\sum_{c=1}^{C} s_i^{c2}} \tag{16}$$

Where $\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^{c2}$ and $\sum_{i=1}^{N}\sum_{c=1}^{C} s_i^{c2}$ denote the true frame of the image and the decision diagram, respectively, using the function minimum optimization model, so the lower L_{Dice} indicates the better effect.

The introduction of BCE loss can help predict the focus/out-of-focus state of the target and assist the model in determining the accuracy of the decision map, which is calculated as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{17}$$

Where N is the total number of samples, y_i is the category to which the ith sample belongs, and p_i is the decision diagram for the ith sample.

In addition, in order to enhance the pixel features of image recovery, measure the structural information of images, and judge the similarity between images, we introduce SSIM loss to help model optimization. The structural similarity SSIM is calculated as shown below.

$$SSIM(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^{\ 2} + \mu_y^{\ 2} + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^{\ 2} + \sigma_y^{\ 2} + C_2} \cdot \frac{\sigma_{x,y} + C_3}{\sigma_x^{\ 2}\sigma_y^{\ 2} + C_3} \tag{18}$$

Where, x and y denote the reference image and the fused image, respectively, μ_x, μ_y represents the mean of x, y, $\sigma_x^{\ 2}$ and $\sigma_y^{\ 2}$ represents the variance about x and y, σ_{xy} is the covariance about x and y ,where C_1, C_2 and C_3 are constants used to stabilize the denominator when it is close to 0. When the two images are infinitely close, the value of SSIM tends to 1 more, and on the contrary, SSIM tends to 0. The specific SSIM loss function is defined as:

$$L_{ssim} = 1 - SSIM(x,y) \tag{19}$$

We introduce these three losses to optimize the training model and enhance the reconstruction information of the pixels after image fusion. When any of the defined loss functions disappears in proportion, the performance of the network will be affected accordingly, weakening the final fused image quality. We set the hyperparameter value of all loss functions to 1 in this paper, so that the importance of each loss function is the same, and also find that each loss we introduce is indispensable for the optimization of the network model proposed in this paper.

4. Experiments

In this section, we evaluate our proposed method PATN based on three publicly available datasets and compare it with seventeen more advanced MFIF methods in the field of multi-focus image fusion, including BFMF (2017) [18], CNN (2017) [19], CSR (2016) [20], DCT_Corr (2018), DSITF (2015) [21], DRPL (2020) [22], ECNN (2019) [23], FusionDN (2020) [24], GCF (2020) [25], GD (2016) [26], GFDF [27], MFF-GAN (2021) [28], GFF [29], IFCNN (2020) [30], MADCNN (2019) [31], MWGF (2015) [32], MGFF (2019) [33], PCANet (2019) [34], PMGI (2020) [35], SESF (2020) [36], and U2Fusion (2020) [37]. Among them,

methods such as GFF, CSR and MWGF belong to transform domain methods, methods such as DSIFT and GFDF belong to spatial domain methods, and methods such as CNN, ECNN and U2 belong to deep learning based methods. We introduce the datasets used in the experiments, the details of the network architecture training, the conduct of the comparison experiments, the selection of evaluation metrics and the ablation experiments respectively in subsections. After our comparative experiments, the results prove the effectiveness and generalization of our proposed method and show that the fusion performance of the experimental method in this paper is better than other MFIF methods.

4.1 Dataset

Since multi-focus image fusion is a means of integrating information by our comprehensive use, multi-focus images require that the input images are multiple sets of images from the same scene focused on different targets, and datasets that meet the requirements are difficult to obtain, so we use the COCO2014 dataset[38] commonly used for semantic segmentation to train our network. The dataset used for image segmentation generally includes the binary segmentation map of the source image, and for the multi-focus image fusion domain, we generate the fused images by:

$$I_A = I_{clear} \text{ e } M_A + I_{blur} \text{ e } M_B \tag{20}$$

$$I_B = I_{clear} \text{ e } M_B + I_{blur} \text{ e } M_A \tag{21}$$

Where I_A and I_B are a pair of fused images, and I_{clear} and I_{blur} are a pair of source images, and M_A and M_B are complementary binary segmentation maps and conform to the relation $M_A + M_B = 1$.

We resize and grayscale the COCO dataset, which is then fed to the network for training. To consider the full performance of the network, we test the network using three MFIF datasets, including the lytro dataset[39], the MFFW dataset[40], and the tsai dataset. Among them, the lytro dataset contains 20 pairs of multi-focus images with a uniform size of 520 × 520. Since the out-of-focus diffusion effect (DSE) is not obvious in the lytro dataset, Xu et al. constructed a new benchmark MFIF dataset. The MFFW dataset contains 13 pairs of source images, and each pair of source images varies in size. Meanwhile, the tsai dataset contains 12 pairs of multi-focused images with different sizes of source images. The experiments in this paper test the average metric values for comparison on all three datasets mentioned, and the comparative performance on the datasets is shown in Section 4.4.

4.2 Experimental Details

The training process takes iterative training according to the characteristics of the network. The training process is run based on NVIDIA 2080 SUPER GPU using the Pytorch framework. Adam optimizer is used and hyeprtparameters are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$ .The initial value of the learning rate is 2×10^{-4} and the batch size is set to 8. The learning rate decreases with the training characteristics of the network throughout the epoch process in order to find the optimal model parameters. The learning rate is reduced to 8×10^{-5} , 5×10^{-5} and 2×10^{-5} in turn. The Transformer sliding window size is also set to 7 and the downscaling_factors parameter is (2,2,2,2), and the training process continues for 80 Epochs

to reach near-optimal values, which in turn stops the redundant training process. During the training process, the network loss decreases quickly, the model converges quickly, and the network is more stable.

4.3 Ablation Experiments

In order to ensure the effectiveness of each module of the proposed network, we conduct ablation experiments to test the help of the introduced modules on the network performance.

4.3.1 PSA module

We introduce the PSA mechanism to focus on the channel features and spatial features of the image feature map to strengthen the network's ability to extract the source image information and thus retain more source information. PSA can enhance the information at the pixel level, and we test the performance of the network with this mechanism removed. As shown in **Fig. 6**, when the mechanism is absent, there is more noise in the decision map of the fused image, while the division between the focused and out-of-focus regions is more obvious and the boundary transition is not harmonious, thus, the introduction of the mechanism is effective for the network to extract image features.

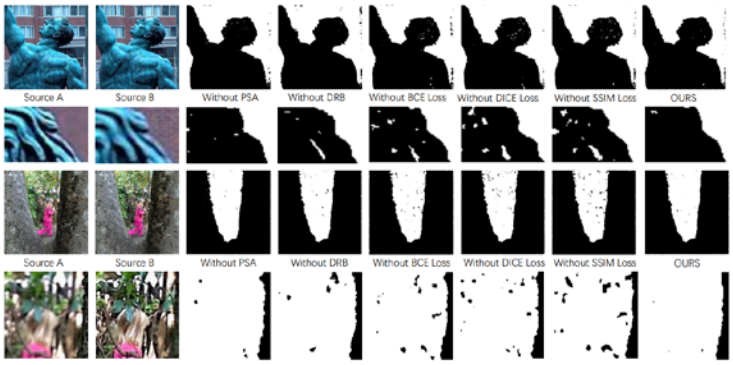


Fig. 6. All ablation experiments performed experimentally. From left to right are: source image A, source image B, resultant decision diagram without PSA module, resultant decision diagram without DRB module, resultant decision diagram without BCE loss, resultant decision diagram without DICE loss, resultant decision diagram without SSIM loss and resultant decision diagram of our experiments.

4.3.2 DRB: Dense Residual Module

Since the image detail information obtained by pixel-level fusion of images is richer than feature-level-based fusion and decision-level-based fusion, but the technical equipment requirements are higher, and the fusion process is not easy to be processed in real time, we introduce the dense residual module based on Sobel gradient operator in the later stage of the network to enhance the information of fused images, which not only ensures the process of network processing, but also ensure the richness of the detail information of the fused image. As shown in **Fig. 6**, there is more noise on the decision map without the DRB module, and also, there is a problem that there is more error information at the edges of the decision map without the DRB to enhance the image edge information.

4.3.3 Loss Function

We use the control variable method to remove the introduced loss functions sequentially, and the three decision maps obtained are shown in **Fig. 6**. We find that the absence of any of the loss terms affects the optimization results of the model, easily introduces wrong detail information, confuses the texture information in the fused images, and thus negatively affects the generated images and decision maps. The experimental results show that only the introduction of a complete loss function term to optimize our training model can guarantee superior fusion results.

4.4 Comparison Experiments

Considering the special characteristics of multi-focus image fusion, i.e., the MFIF method is based on a test set without GroundTruth for performance testing, there is no fixed single metric, so based on this characteristic, we consider the performance of the MFIF method from two perspectives: quantitative objective metrics and qualitative subjective visual experience.

4.4.1 Objective evaluation on the Lytro dataset

We tested all 20 pairs of images on the Lytro dataset, measuring the mean value of the method on this dataset and selecting the statistics of six quantitative metrics, as shown in **Fig. 7**. The measures include Cvejie's metric Q_{cv} [41], gradient-based Q_{xy} [42], SSIM metric-based Q_Y [43], phase-coherence-based Q_{p} [44], spatial frequency error ratio-based Q_Z and spatial frequency Q_{SF} [45], which are measured as follows:

1) Cvejie's metric Q_{cv}: This metric is based on the similarity of pixel blocks between images as a metric to measure the degree of representation of the fused image to the information contained in the input image, and the larger the value of the metric, the better the subjective quality of the fused image obtained. 2) Gradient-based edge information quantity metric Q_{xy}: The index value is calculated using the Sobel operator to calculate the edge intensity of the corresponding image, i.e., the value of the convolution with the Sobel operator is first calculated as $C_1 = C_A^x(x,y)$, $C_2 = C_A^y(x,y)$, C_1 and C_2 are the convolution values obtained using the Sobel operator in the vertical and horizontal directions, respectively. The edge intensities are $g_A(i,j) = \sqrt{C_1^2 + C_2^2}$,and the direction is $\alpha_A(i,j) = \tan^{-1}(C_1 / C_2)$. The final assessment metrics are: $Q_{xy} = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J}[Q^A(i,j)w^A(i,j) + Q^B(i,j)w^B(i,j)]}{\sum_{i=1}^{I}\sum_{j=1}^{J}(w^A(i,j) + w^B(i,j))}$,where Q^A, Q^B are the edge retention information values of the input image, and w are the weighting coefficients. The specific formula is understood in more detail in [46]. 3) SSIM-based metric Q_Y: The metric value mainly measures the similarity between the redundant and complementary regions, and the larger the metric value indicates the higher quality of the fused image. 4) Phase-coherence-based Q_{p}: The phase-coherence-based metric proposed by zhao et al. reflects the richness of

the image edge information and the image contour corner point information, and the larger the metric value indicates the richness of the fused image information. 5) Spatial frequency error ratio-based Q_Z : This metric reflects the change in the local intensity information of the image, which potentially improves the image quality. 6) Spatial frequency based metric Q_{SF} : This metric measures the activity level of the image by spatial frequency, and the larger the value of the metric, the better the image quality.

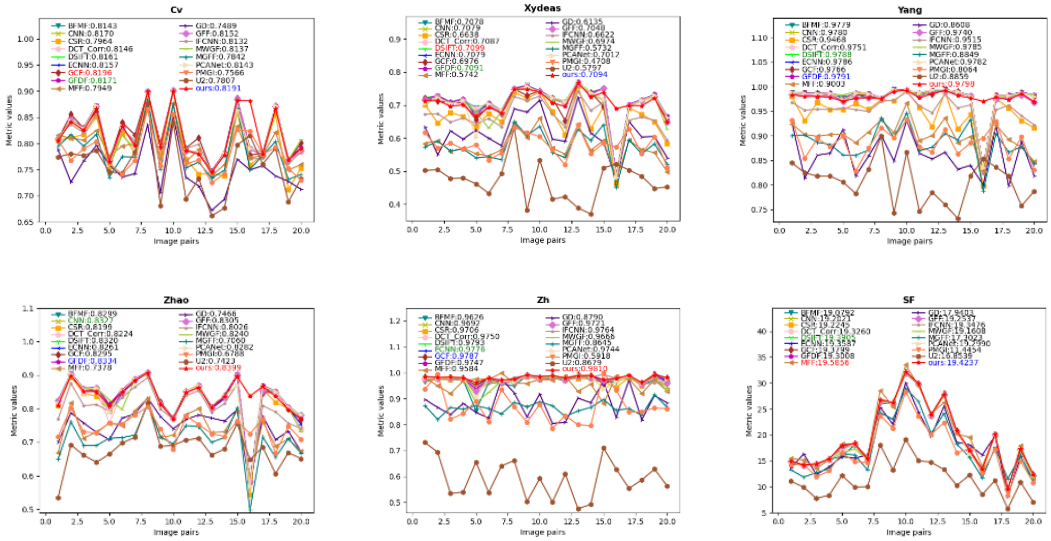


Fig. 7. Comparison of the average metric values of the 17 methods with the method proposed in this paper on the Lytro dataset, with those marked in red representing the best values, those marked in blue representing the second best values, and those marked in green being the third best values.

From **Fig. 7**, we see that PATN is first on average on the Lytro test set on Q_Y , Q_{P} as well as Q_Z , and because PATN introduces the PSA mechanism of local attention and the dense residual blocks of the edge gradient-based Sobel operator to enhance the edge information, the fusion results of the images are rich in information at the target edges and contour corner points, and the intensity information of the image local changes can be captured sensitively. Meanwhile, the second highest mean value is accounted for on Q_{cv} , Q_{xy} and Q_{SF} , and these metrics reflect that PATN maintains better edge information and richer pixel information in the fused images compared with other MFIF methods.

4.4.2 Objective evaluation on the MFFW dataset

We tested the full dataset from the multi-focused image public test set MFFW, and also introduced two new metrics, for the metric based on multi-scale measurement of edge information preservation Q_{PWW} [47] and the metric based on pixel-level metric $Q_{AB/F}$, and the metric values on the MFFW dataset are shown in **Fig. 8**. It can be seen that are optimal on Q_{PWW} , $Q_{AB/F}$, Q_{P} and Q_Z , while ranking second on Q_{cv} and Q_{xy} . The results show that PATN perceives the image edge information and fuses the visual information of the image are beyond the other MFIF methods.

However, from the results of the average index values of all MFIF methods used in this paper on the MFFW dataset, the overall average value on the MFFW dataset is lower than that on the Lytro dataset, and we analyze two multi-focus datasets, both of which are composed of focused and out-of-focus regions. The distinction between foreground and hindground regions on the MFFW dataset is not as clear as on the Lytro dataset. This may be one of the reasons for this difference in results.

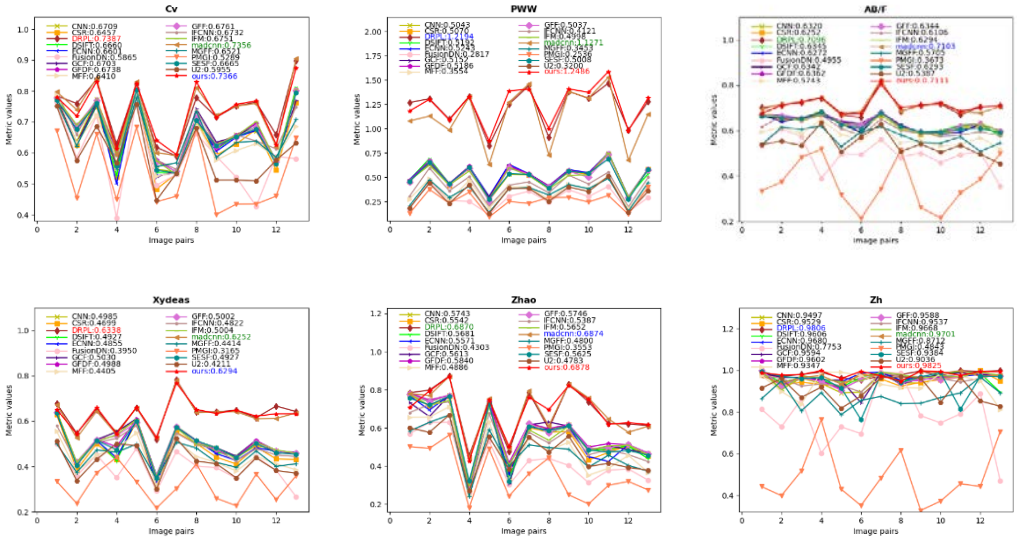
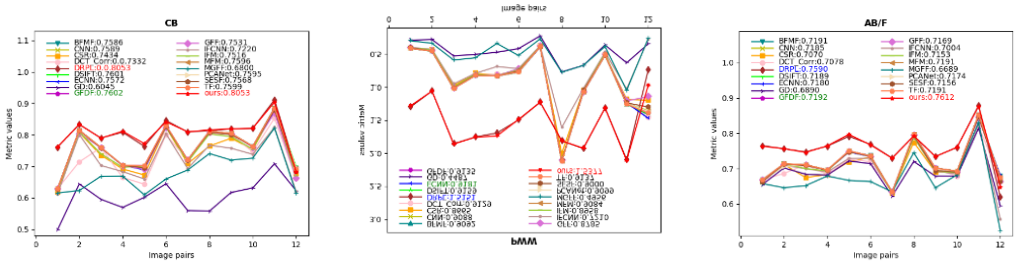


Fig. 8. Comparison of the average index values of the 17 methods with the method proposed in this paper on the MFFW dataset, with those marked in red representing the best values, those marked in blue representing the second best values, and those marked in green being the third best values.

4.4.3 Objective evaluation on the TSAI dataset

In addition to the Lytro dataset and the MFFW dataset, we also test the performance of the method on the TSAI dataset, while we test the method on the metrics of the human vision system model-based measures Q_{CB} [48], and the results of the metrics on the TSAI test set are shown in **Fig. 9**. It can be seen that the mean value of PATN is first for all metrics, and the superior performance of the method is quantitatively and objectively reflected, which is sufficient to prove the robustness of our proposed framework. The TSAI dataset has similarity with the Lytro dataset, and the focused and out-of-focus regions are clearer, so the evaluation results of the MFIF method on these two datasets are better than those on the MFFW dataset.



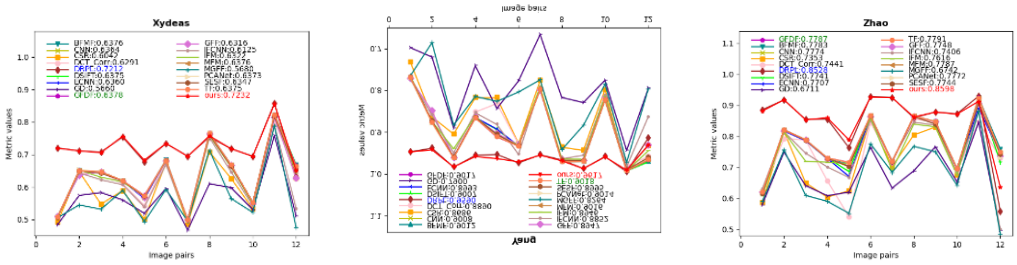


Fig. 9. Comparison of the average index values of the 17 methods with the method proposed in this paper on the TSAI dataset, with those marked in red representing the best value, those marked in blue representing the second best value, and those marked in green being the third best value.

4.4.4 Subjective visual evaluation on MFIF testsets

We selected several sets of images from the test results of the three test sets for visual comparison of the methods, and considered various MFIF methods from the perspective of subjective visual experience. In **Fig. 10**, we find that the gradient-based GFDF method and PMGI are worse in overall visual sensory, appearing with dark image colors that do not match the color information of the real source images. GD, PMGI, and U2Fusion show poor results in the difference maps, all showing color distortion problems, while both foreground and hindground regions appear in the difference maps. CNN, CSR, DSIFT, ECNN, GD, and MGFF methods failed to accurately retain the detail information of the boundary line of the focus/out-of-focus area. The "white dots" on the "grass" next to the "baseball" appear blurred. Meanwhile, the BFMF, DCT_Corr, GCF, GFF and PCANet methods did not fuse the "white dot" next to the "shoulder" well, and some of the fusion results did not even show the small target directly, and the artifacts were extremely serious. PATN's fusion result of the target next to the "baseball" was complete and clear, and the overall vision was good.

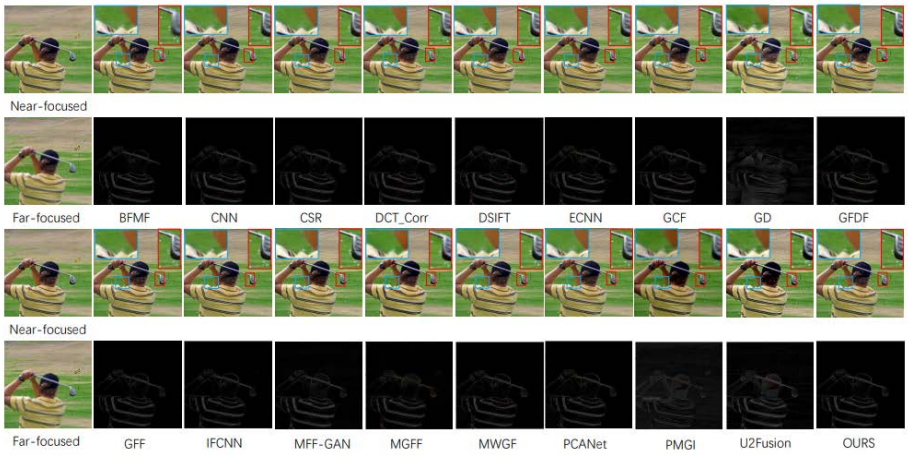


Fig. 10. Visual comparison plots of the 18 methods on Lytro-01. The first and third rows show the result plots for each method, and the second and fourth rows show the result difference plots for each method.

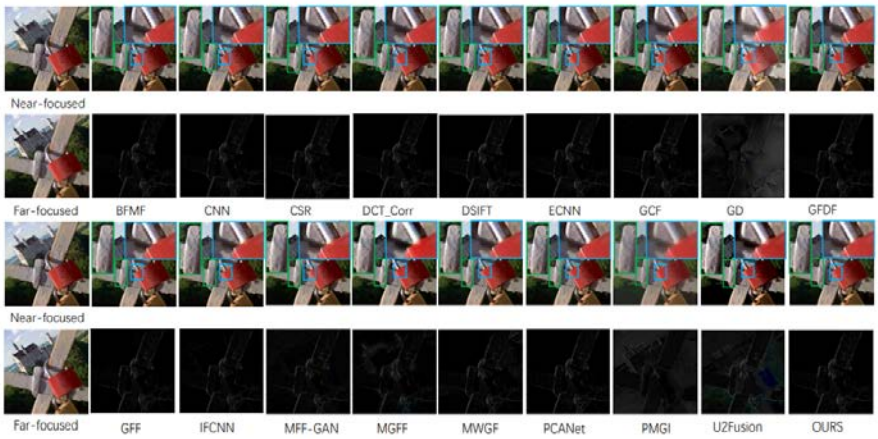


Fig. 11. Visual comparison plots of the 18 methods on Lytro-06. The first and third rows show the result plots for each method, and the second and fourth rows show the result difference plots for each method.

In **Fig. 11**, it is clear from the differential map that GD is exposed. And MGFF, MWGF, IFCNN, PMGI and U2Fusion appear in the differential map in both the front and rear fields.

The left side of the "cross" of BFMF and GFDF is not shown in the differential map, and the PMGI as a whole shows a wide range of blurring. CNN, CSR, DCT_Corr, DSIFT, ECNN, GCF, GFF and MFF-GAN are not clearly blended at the "iron lock" in the blue frame, and the boundary between the front and back view areas is blurred, without good visual perception. PATN, on the other hand, blends clearly at the "cross" without extensive blurring and with soft colors.

In **Fig. 12**, still as in Lytro-06, on Lytro-11, GD, MGFF, PMGI, and U2Fusion contain both focused and out-of-focus regions in the differential map, with poorer fusion. BFMF, CSR, DSIFT, GFDF, MGFF, PCANet and PMGI are blurred in the "camera" in the yellow box, i.e., there is still some room for improvement in the processing of small detail targets in these methods. Other deep learning-based MFIF methods, such as CNN, ECNN, IFCNN and MFF-GAN, are quite effective in fusing out-of-focus regions of the image, and there is no color distortion, and the differential detail information of out-of-focus regions is also more complete in the differential map.

Fig. 12. Visual comparison plots of the 18 methods on Lytro-11. The first and third rows show the result plots for each method, and the second and fourth rows show the result difference plots for each method.

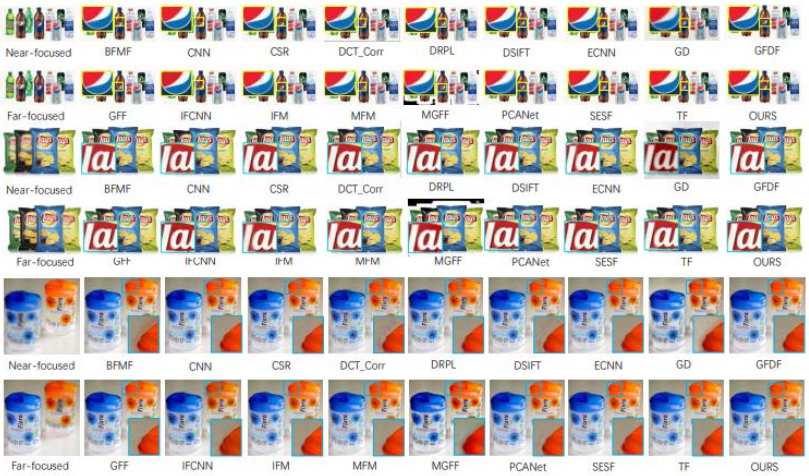


Fig. 13. Visual comparison graph of three images selected on the TSAI dataset based on 18 methods.

We tested on the TSAI dataset, and the TSAI dataset contains more targets in some images compared to the lytro dataset, and some of the focused and out-of-focus regions are blurred, and the source input images are of more general quality. As shown in **Fig. 13**, on the TSAI first image data, BFMF, DCT_Corr, ECNN, GD, IFM, MFM, MGFF, and PCANet methods all showed extensive blurring at the focus/out-of-focus boundary line, and poor visual results for image fusion in out-of-focus regions. The other deep learning-based MFIF methods produced okay visual results, and the unsupervised SESF fusion was also quite good. However, on the second image data of TSAI, deep learning-based MFIF methods such as CNN, DRPL, ECNN, IFCNN, TF, and scale invariance-based DSIFT all showed artifacts visible to the naked eye, resulting in visual indistinctness. Gradient-based methods such as GD, GFDF, and GFF even showed severe colour distortion and blurred patches, resulting in poor subjective visual perception. On the third image of TSAI data, DCT_Corr shows obvious block effect, while DSIFT and PCANet show "halo phenomenon", and the junction of focus/out-of-focus area of the fused image shows obvious and poor fusion effect. From the enlarged figure, we can see that the BFMF, CSR, ECNN, GD, IFCNN, MGFF, and SESF methods do not deal well with the DSE at the boundary of the "bottle cap" in the blue box. PATN showed neither colour distortion nor extensive blurring and small artifacts, and the fusion results were good on the TSAI dataset.

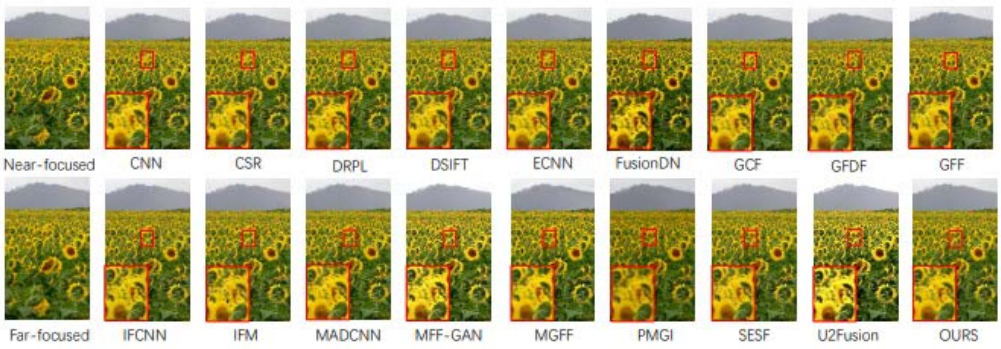


Fig. 14. Visual comparison chart of 18 methods on MFFW-03. We zoom in on the local details as shown in the figure.

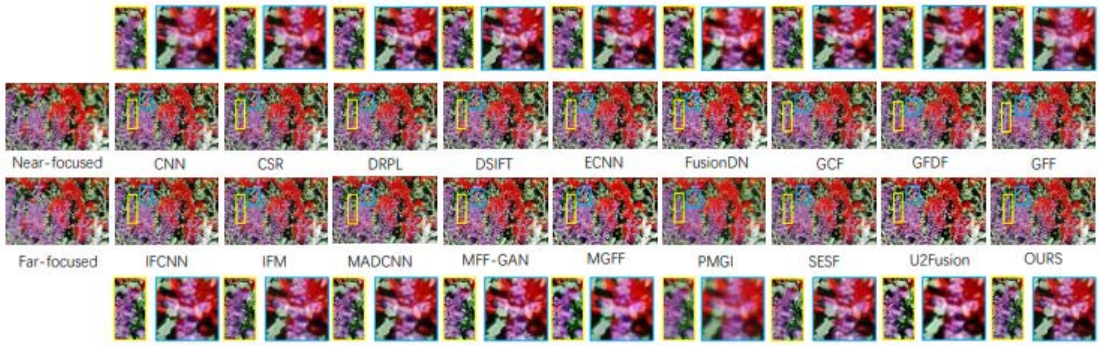


Fig. 15. Visual comparison plots of the 18 methods on MFFW-05. The second and third rows are plots of the results of the various methods, and the first and fourth rows are detailed plots of the various methods being partially enlarged.

In **Fig. 14**, we find that the overall colour of PMGI and FusionDN is darker and less saturated, while the colour of the generic fusion framework U2Fusion local target is oversaturated and the local colour information is over-captured, resulting in a visual effect like overexposure. The image detail information is also blurrier using MFIF methods with fewer loss terms, such as ECNN and MADCNN, probably because the generalization ability of these methods on the MFFW dataset is not better represented. At the same time, the generalization ability shown by the methods that do not use the consistency check operation, such as DRPL, IFCNN, and MADCNN, is also more general. In **Fig. 15**, the deep learning-based CNN, ECNN, FusionDN, and MADCNN methods show a wide range of blurring, and the visual results obtained by the deep learning-based MFIF on this test image are poor. The gradient-based GFF and GCF methods show a wide range of blurring in the enlarged image area we selected, while MGFF and IFM show unclear petals in the "flower" area. The overall visual effect of PMGI is not good. The overall visual effect of the unsupervised SESF-based method is quite good, and there are no small artifacts, which can indicate that the post-processing of the image helps the fusion effect of the deep learning-based method.

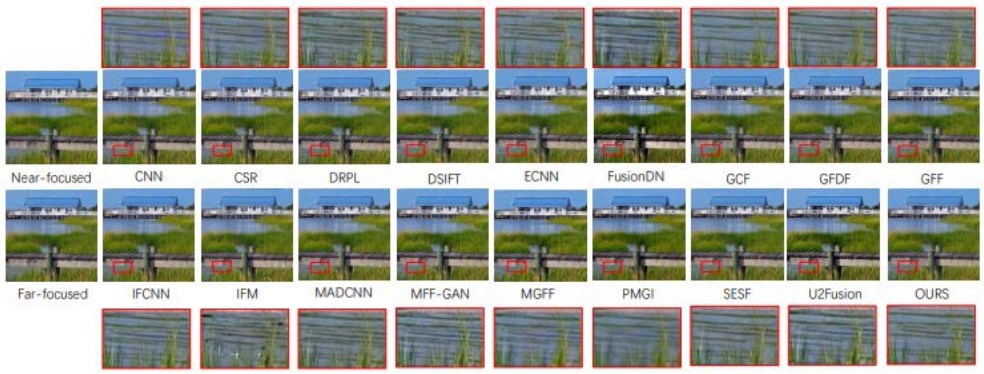


Fig. 16. Visual comparison plots of the 18 methods on MFFW-08. The second and third rows are plots of the results of the various methods, and the first and fourth rows are detailed plots of the various methods being partially enlarged.

In **Fig. 16**, FusionDN is visually dark and the image color is distorted, IFM shows wrong fusion information in the local area, PMGI shows a wide range of blurring at the "waterline". GCF has a lot of "water lines" that are not shown in the enlarged image, and MFIF based on deep learning, such as CNN, DRPL, IFCNN and MADCNN methods have comfortable overall color and clear fusion of detailed information, and the fusion effect is quite good, while ECNN has "fault " phenomenon. In general, the performance of general image fusion frameworks is not as good as the performance of methods specifically designed for the MIFIF task, such as the U2Fusion method, which is worse than the other MFIF methods. Our method uses multiple function loss terms to ensure that the trained model is more robust on different datasets, and we use consistency tests to correct the obtained decision maps. The final performance shows that the robustness and generalization of our method are better.

5. Conclusion

In this paper, a Transformer-based U-shaped architecture PATN is proposed to handle multi-focus image fusion tasks. PATN extracts deep features of images based on the long-distance dependence property of Transformer mechanism, and introduces PSA parallel module to focus on both channel information and spatial information of features, so that deep and shallow features of images can be perfectly aggregated. The deep and shallow features of the image are perfectly aggregated. We then input the obtained feature maps to the feature enhancement stage of PATN, and introduce the DRB dense residual module to enhance the edge information of the image to ensure that the obtained fused image has correct and rich edge information. In addition, we introduce multiple losses in order to retain more texture detail information. The generalization and effectiveness of PATN are verified on three multi-focus image fusion datasets, and the superior performance of PATN is demonstrated by comparing it with 17 existing MFIF methods.

Acknowledgement

The authors acknowledge the National Natural Science Foundation of China (62002200, 62202268 and 61972235), the Shandong Natural Science Foundation of China (ZR2021MF107, ZR2022MA076) Youth Innovation Technology Project of Higher School in Shandong Province under (2021KJ069, 2019KJN042) and Yantai science and technology innovation development plan(2022JCYJ031).

References

[1] B. Yang and S. Li, “Multi-focus image fusion and restoration with sparse representation,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010. Article(CrossRef Link)

[2] Ba Virisetti, P. D, “Multi-sensor image fusion based on fourth order partial differential equations,” in *Proc. of 20th International Conference on Information Fusion (Fusion)*, IEEE, 2017. Article(CrossRef Link)

[3] H. Li, B. S. Manjunath, and S. K. Mitra, “Multi-sensor image fusion using the wavelet transform,” *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, May 1995. Article(CrossRef Link)

[4] S. Li, J. T. Kwok, and Y. Wang, “Combination of images with diverse focuses using the spatial frequency,” *Inf. Fusion*, vol. 2, no. 3, pp. 169–176, Sep. 2001. Article(CrossRef Link)

[5] I. De and B. Chanda, “Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure,” *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, Apr. 2013. Article(CrossRef Link)

[6] X. Bai, Y. Zhang, F. Zhou, and B. Xue, “Quadtree-based multi-focus image fusion using a weighted focus-measure,” *Inf. Fusion*, vol. 22, pp. 105–118, Mar. 2015. Article(CrossRef Link)
[7] W. Pan, Z. Zhao, and W. Huang, “Video Moment Retrieval With Noisy Labels,” *Inf. Fusion, IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022. Article(CrossRef Link)
[8] P. Hill, M. E. Al-Mualla, and D. Bull, “Perceptual image fusion using wavelets,” *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1076–1088, Mar. 2017. Article(CrossRef Link)
[9] Q. Zhang and B.-L. Guo, “Multi-focus image fusion using the non-subsampled contourlet transform,” *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, Jul. 2009. Article(CrossRef Link)
[10] S. Li, X. Kang, J. Hu, and B. Yang, “Image matting for fusion of multi-focus images in dynamic scenes,” *Information Fusion*, vol.14, no.2, pp. 147-162, 2013. Article(CrossRef Link)
[11] S. Li, X. Kang, and J. Hu, “Image Fusion With Guided Filtering,” *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864-2875, 2013. Article(CrossRef Link)
[12] L. Ma, Y. Zheng, and Z. Zhang, “Motion Stimulation for Compositional Action Recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-1, 2022. Article(CrossRef Link)
[13] L. Fu, D. Zhang, and Q. Ye, “Recurrent thrifty attention network for remote sensing scene recognition,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8257-8268, 2021. Article(CrossRef Link)
[14] H. Cao, Y. Wang, J. Chen, D. Jiang, and M. Wang, “Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation,” *Image and Video Processing*, May.2021. Article (CrossRef Link)
[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhang, “Swin transformer: hierarchical vision transformer using shifted windows,” *Computer Vision and Pattern Recognition*, Mar.2021. Article (CrossRef Link)
[16] H. Liu, F. Liu, X. Fan, and D. Huang, “Polarized self-attention: towards high-quality pixel-wise regression,” 2021. Article(CrossRef Link)
[17] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28-42, 2022. Article(CrossRef Link)
[18] Y. Zhang, X. Bai, and T. Wang, “Boundaryfinding based multi-focus image fusion through multi-scale morphological focus-measure,” *Information fusion*, vol. 35, pp. 81–101, 2017. Article(CrossRef Link)
[19] 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. Article(CrossRef Link)
[20] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, “Image fusion with convolutional sparse representation,” *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016. Article(CrossRef Link)
[21] W. Shuping, and Zengfu, “Multi-focus image fusion with dense SIFT,” *Information Fusion*, vol.23, pp.139-155, May. 2015. Article (CrossRef Link)
[22] J Li, X. Guo, G. Lu, B. Zhang, and D. Zhang, “DRPL: Deep regression pair learning for multi-focus image fusion,” *IEEE Transactions on Image Processing*, vol.29, pp.4816-4831, Mar. 2020. Article (CrossRef Link)
[23] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, “Ensemble of cnn for multi-focus image fusion,” *Information Fusion*, vol.51, pp.201-214, Nov. 2019. Article (CrossRef Link)
[24] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, “FusionDN: A Unified Densely Connected Network for Image Fusion,” in *Proc. of the AAAI Conference on Artificial Intelligence*, Vol.34, no.7, pp.12484-12491, 2020. Article (CrossRef Link)
[25] 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. Article(CrossRef Link)
[26] S. Paul, I. S. Sevcenco, and P. Agathoklis, “Multi-exposure and multi-focus image fusion in gradient domain,” *Journal of Circuits, Systems and Computers*, vol. 25, no. 10, pp. 1650123.1-1650123.18, 2016. Article(CrossRef Link)

[27] 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. Article(CrossRef Link)
[28] Z. A. Hao, A. Zl, B. Zs, X. A. Han, and A. Jm, "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, Feb.2021. Article (CrossRef Link)
[29] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013. Article(CrossRef Link)
[30] 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. Article(CrossRef Link)
[31] R. Lai, Y. Li, J. Guan, and A. Xiong, "Multi-scale visual attention deep convolutional neural network for multi-focus image fusion," *IEEE Access*, vol. 7, pp. 114385–114399, 2019. Article(CrossRef Link)
[32] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Information Fusion*, vol. 20, pp. 60–72, 2014. Article(CrossRef Link)
[33] 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, Dec 2019. Article(CrossRef Link)
[34] X. Song and X.-J. Wu, "Multi-focus Image Fusion with PCA Filters of PCANet," in *Proc. of MPRSS 2018: Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, Springer, pp. 1–17, 2019. Article(CrossRef Link)
[35] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12797-12804, 2020. Article(CrossRef Link)
[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 Computing and Applications*, vol. 33, pp. 5793-5804, 2021. Article(CrossRef Link)
[37] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: a unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502-518, 2022. Article (CrossRef Link)
[38] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft coco: common objects in context," in *Proc. of European Conference on Computer Vision, Springer International Publishing*, pp 740–755, 2014.Article(CrossRef Link)
[39] M. Nejati, "Lytro multi focus dataset," Isfahan Univ. Technol., Isfahan, Iran, Tech. Rep., 2015. Article(CrossRef Link)
[40] S. Xu, X. Wei, C. Zhang, J. Liu, J. Zhang, "MFFW: A new dataset for multi-focus image fusion," 2020. Article(CrossRef Link)
[41] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *International journal of signal processing*, vol. 2, no. 3, pp. 178–182, 2008. Article(CrossRef Link)
[42] C. S. Xydeas and P. V. V., "Objective image fusion performance measure," *Military Technical Courier*, vol. 36, no. 4, pp. 308–309, 2000.
[43] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, no. 2, pp. 156–160, 2008. Article(CrossRef Link)
[44] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 6, pp. 1433–1447, 2007. Article(CrossRef Link)
[45] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, vol. 43, no. 12, pp. 2959–2965, 1995. Article(CrossRef Link)

[46] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.34, no.1, pp. 94-109, 2012. Article(CrossRef Link)
[47] P. W. Wang, and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Proc. of International Conference on Signal Processing IEEE*, 2008. Article(CrossRef Link)
[48] C. Yin, and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image & Vision Computing*, vol.27, no.10, pp.1421-1432, 2009. Article(CrossRef Link)

Pan Wu received her bachelor's degree from the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China in 2020. Currently studying for a master's degree in the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, Shandong. Her research interests include computer graphics, computer vision, and image processing.

Zhen Hua received the B.S. and M.S. degrees in electrical automation from Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, the Ph.D. degree in electronic information engineering from China University of Mining and Technology, Beijing, China, in 2008. She is currently a professor at Shandong Technology and Business University. Her research interests include computer aided geometric design, information visualization, virtual reality, and image processing.

Jinjiang Li received the B. S. and M. S. degrees in computer science from Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, the Ph. D. degree in computer science from Shandong University, Jinan, China, in 2010. From 2004 to 2006, he was an assistant research fellow at the institute of computer science and technology of Peking University, Beijing, China. From 2012 to 2014, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, China. He is currently a Professor at the school of computer science and technology, Shandong Technology and Business University. His research interests include image processing, computer graphics, computer vision, and machine learning.