FS-Transformer: A new frequency Swin Transformer for multi-focus image fusion

Weiping Jiang[1], Yan Wei[1], and Hao Zhai[1]*
[1]School of Computer and Information Science, Chongqing Normal University
Chongqing, China, 401331
[e-mail: zhaihao@cqnu.edu.cn]
*Corresponding author: Hao Zhai

Received September 19, 2023; revised May 25, 2024; accepted June 24, 2024; published July 31, 2024

Abstract

In recent years, multi-focus image fusion has emerged as a prominent area of research, with transformers gaining recognition in the field of image processing. Current approaches encounter challenges such as boundary artifacts, loss of detailed information, and inaccurate localization of focused regions, leading to suboptimal fusion outcomes necessitating subsequent post-processing interventions. To address these issues, this paper introduces a novel multi-focus image fusion technique leveraging the Swin Transformer architecture. This method integrates a frequency layer utilizing Wavelet Transform, enhancing performance in comparison to conventional Swin Transformer configurations. Additionally, to mitigate the deficiency of local detail information within the attention mechanism, Convolutional Neural Networks (CNN) are incorporated to enhance region recognition accuracy. Comparative evaluations of various fusion methods across three datasets were conducted in the paper. The experimental findings demonstrate that the proposed model outperformed existing techniques, yielding superior quality in the resultant fused images.

Keywords: Multi-focus image fusion, Swin Transformer, Wavelet Transform, Deep learning

http://doi.org/10.3837/tiis.2024.07.010 **ISSN : 1976-7277**

1. Introduction

Multi-focus image fusion is a subfield of image fusion. As the name suggests, the primary function of this tool is to extract sharp pixels from multiple images with varying focal points and then merge them into a single image. Through this technology, the disadvantages caused by the lens depth of field problem can be avoided, enabling the capture of clear images that are more suitable for computer processing and recognition. In recent years, it has been widely used in the fields of military, medical imaging, and computer vision [1-3]. Existing multi-focus image fusion methods can be classified into two categories: traditional methods and non-traditional methods.

Traditional methods are transform domain-based and spatial domain-based methods. The transform domain-based methods mainly achieve image fusion through three steps, namely image transformation, coefficient fusion, and image inverse transformation [4]. The representative methods are Wavelet Transform [5], discrete cosine transform [6] and pyramid transform [7]. Among them, the Wavelet Transform can excellently extract high-frequency detail information in the horizontal, vertical, and diagonal directions. And with the in-depth research on Wavelet Transform, other various Wavelet Transform theory techniques have also emerged. However, the Wavelet Transform is limited to the extraction of detail information, because it cannot extract the anisotropic detail information of the images. In order to obtain better fusion results, methods based on multi-scale geometric analysis have been successively proposed. For example, methods based on bandelet transform [8], shearlet transform [9] and Nonsubsampled Contourlet transform [10] and so on. In general, transform domain-based methods obtained better results, but have high computational complexity. Thus, spatial domain-based methods that directly operate on the image pixels themselves have been proposed one after another. This method first extracts relevant features in the spatial domain to measure the activity level of the source images, and then uses a certain fusion rule to fuse the source image according to the calculated activity. Such methods can be further divided into pixel-based [11], block-based [12] and region-based [13] methods. In short, the core of the spatial domain-based methods is the metric of pixels, blocks, and regions. The method also has some disadvantages, such as, block artifacts, contrast degradation and so on.

Non-traditional methods mainly refer to deep learning-based methods. The simulation of deep learning has made it mainstream in the field of image fusion. The convolution process in CNN facilitates a more comprehensive extraction of localized image characteristics. Thus, the integrity and details of the fusion image structure information can be maintained. Existing CNN-based methods can be further divided into end-to-end based and non-end-to-end based methods. Non-end-to-end methods can in turn be classified as classification-based methods. The first model using CNN in the field was proposed by Liu et al [14]. Since then, ECNN [15], CNN-based methods [16] all classify pixels to obtain an initialization decision map, and then refine it through post-processing technology for image fusion. End-to-end methods, on the other hand, belong to the regression-based methods. This method is different from the classification method, it does not generate a decision map but directly predicts the output fusion images. It learns an end-to-end mapping from the source images to the final fused images. For example, Wang et al. proposed a progressive residual learning network [17]. The model first fused the color information of the source images with the initial fusion block, and then used the enhanced fusion block to fuse the detail features. Huang et al. used GAN to complete multi-focused image fusion, and designed an adaptive weight module, which can guide the generator to adaptively learn the distribution of focused pixels.

While non-traditional approaches have shown improved outcomes in multi-focus image fusion models, a predominant reliance on CNN is observed in their implementation. Nonetheless, a key limitation lies in the convolutional neural network's inadequate understanding of the significance of global information, resulting in a deficiency in modeling such information. Consequently, the loss of global information may lead to a deficiency in color information from the source image within the final fused images, ultimately resulting in chromatic aberration. In order to solve the above problems, inspired by the application of transformer in vision field [18-20], and considering that transformer can make up for the low global accuracy of CNN. Therefore, combining transformer and CNN, a parallel structure of CNN and transformer is proposed. Among them, the advanced Swin Transformer [21] is introduced into the model in the paper. To enhance the quality of the resultant fused images, maximizing the feature information extracted from the source images is imperative. Consequently, the research paper deviates from the initial design of the Swin Transformer by incorporating a frequency domain layer aimed at capturing diverse image components to extract edge and line details. This is followed by the utilization of an attention mechanism to effectively model these features. The primary contribution of the paper can be summarized as outlined below.

(1) An end-to-end multi-focused image fusion method based on Wavelet Transform and Swin Transformer called FS-Transformer is proposed.

(2) A parallel network structure of CNN and Swin Transformer is used. CNN is used to compensate for the missing local information of Swin Transformer and enhance the ability of the model

(3) A combination of frequency domain layer and attention layer is proposed to interrogate the source images from a complex perspective. The frequency domain layer is implemented by Wavelet Transform and inverse Wavelet Transform. The main function of the frequency domain layer is to obtain high-quality features of the image.

(4) Experimental results show that the proposed method improves both visually and in objective metrics compared to existing multi-fusion image fusion methods.

The remainder of the paper is organized as follows, Section II focuses on the application of Wavelet Transform in the field of images and Vision Transformer. Section III then elaborates on the model proposed in this paper. Section IV focuses on presenting the experimental results and demonstrating the effectiveness and superiority of the algorithm using ablation experiments. Finally, section V concludes the paper.

2. Related works

2.1 Wavelet Transform

Wavelet Transform is inherited and developed from the idea of Fourier transform. It is used for time-frequency analysis and decomposes the signal into different frequency domain subbands to obtain the time spectrum. Wavelet Transform has a wide range of tasks in the field of images such as image denoising [22], image compression [23], image segmentation [24] and image fusion [25] fields. The combination of Wavelet Transform and CNN has also been shown to be beneficial for image restoration tasks by Bae et al. [26]. The MWCNN [27] model proposed by Liu et al. achieves feature images size reduction and reconstruction of feature images by introducing the Wavelet Transform and inverse Wavelet Transform, respectively. In [28], the Wavelet Transform technology is also combined with resnet, and this structure has achieved relatively good performance in image recognition. The Wavelet Transform is commonly utilized in image fusion applications, where the fundamental concept involves conducting Wavelet Transform on the original images initially. Subsequently, the transformed coefficients are integrated following specific guidelines, and ultimately, the inverse Wavelet Transform is applied to the combined coefficients. By following these procedures, the original images can be fused to produce the ultimate fused images. Wavelet Transform decomposes the image to extract information components in all image directions, with these components typically containing detailed or comprehensive image information. This property enables a focus on specific image details, aiding in the preservation of intricate image information. Wavelet Transform ensures that the image decomposition process does not result in information loss or the addition of extraneous noise. Upon completion of the decomposition, the transformed coefficients are merged based on predefined rules, and subsequently, the inverse Wavelet Transform is executed on the merged coefficients. Through these outlined steps, the original image can be fused to generate the final fused image.

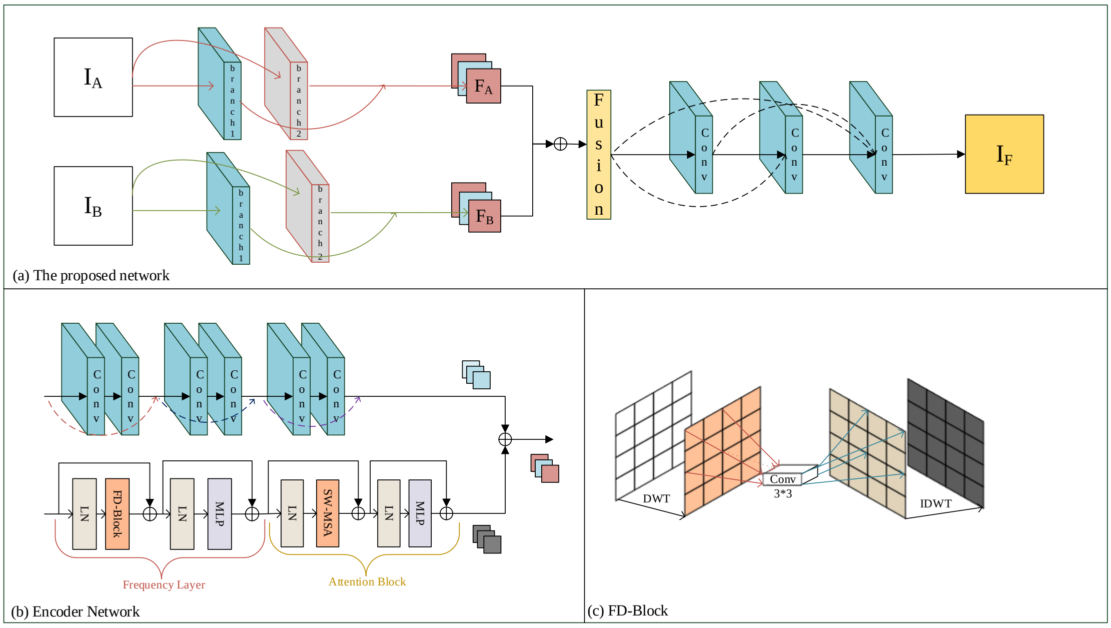
Wavelet Transform can be classified into a class of transform domain-based methods. The transform domain-based methods have long attracted the attention of researchers because it is more consistent with human visual systems and computer processing. Nowadays, there are still many researchers using transform domain techniques to deal with the task of multi-focus image fusion [29-30]. Because of the time-frequency local features of the Wavelet Transform and its ability to extract fine-grained features in the images, more and more image fusion methods based on various Wavelet Transforms have been developed. Thanks to the excellent performance of Wavelet Transform in the field of image fusion, the paper will combine the Wavelet Transform and make full use of the advantages of Wavelet Transform for the modeling ability of model features.

2.2 Transformer

With Transformer, a deep neural network based on self-attention, making a splash in the field of natural language processing. Researchers have followed suit by applying Transformer to the image field. The Vision Transformer [31] proposed by the Google team performed well in image classification and achieved better results. Because of this milestone work, Transformer has also been extended to more fields to solve computer vision tasks. Based on Vision Transformer, Liu et al. proposed a new type of backbone relying on sliding windows, the Swin Transformer [21]. It is precisely because this sliding window-based approach solves the problem that Vision Transformer does not pay enough attention to local areas. Chen et al. proposed a Swin Transformer-based image segmentation method [32]. The Transformer framework proposed by Vibashan et al. for infrared and visible light image fusion performed well [33]. The multi-focus image fusion framework based on Swin Transformer and feedback mechanism [34] proposed by Wang et al. makes the final fused images obtain high fidelity and clarity. Swin Transformer is in image processing can grasp the global information of the image, it deals with the global information. This feature better can enhance the model's ability to extract global relevance and global information. Compared to CNN, this ability makes the model no longer limited to a small region of the image. In practice, CNN and Swin transformer are often combined to complement each other so that the model can get better results. Such, Qu et al. proposed a general framework for various fusion tasks [35]. The framework combines CNN and transformer and this structure achieved better results in fusion tasks.

3. The proposed method

CNN functions by utilizing convolution, which is specifically tailored to the receptive field, allowing CNN to focus more on local information within images. Conversely, the attention mechanism in the transformer model enhances model interpretability and emphasizes global information, albeit with less efficiency in capturing local details compared to CNN. To address the limitations of both CNN and transformer models while leveraging their respective strengths, a parallel transformer model alongside CNN has been proposed. This parallel architecture enables the model to simultaneously prioritize global and local information. The schematic overview of this framework is depicted in **Fig. 1 (a)**. The resultant fused images are generated by feeding source images A and B into the model. The encoder of the model, illustrated in **Fig. 1 (b)**, comprises a CNN branch and an FS-Transformer block, operating in parallel to extract image features. The CNN branch is structured with three residual blocks, each containing two CONV layers, with dimensionality reduction incorporated to ensure consistency between input and output channels. The FS-Transformer block is segmented into the frequency layer and the attention block, with the FD-Block in the frequency layer depicted in **Fig. 1 (c)**.



(a) The proposed network

(b) Encoder Network

(c) FD-Block

Fig. 1. Framework of the proposed method

3.1 Frequency Layer

The utilization of the frequency domain layer aims to analyze the diverse frequency components present in images for the purpose of comprehending local frequency characteristics. Drawing inspiration from the application of Wavelet Transform within the realm of image processing, the frequency domain layer is constructed with Wavelet Transform and its inverse counterpart. Specifically, the Wavelet Transform technique employed in this paper is the well-established Haar Wavelet Transform. Initially, images undergo conversion into the frequency domain space through Wavelet Transform, followed by the application of inverse Wavelet Transform to revert the frequency domain space back to the original physical space. In alignment with the Swin Transformer block's configuration, subsequent to the frequency domain layer, normalization and MLP layers are incorporated. A distinctive feature of this paper is the integration of the frequency domain layer in lieu of the initial self-attention layer within the Swin Transformer block. A comparison between the proposed Swin Transformer block and the conventional version is depicted in **Fig. 1 (b)** and **Fig. 2**, respectively. The conventional Swin Transformer block employs two self-attention mechanisms, which, while effective, may fall short in accurately capturing local features. Consequently, the frequency domain layer, comprising Wavelet Transform and the attention layer, is introduced to concurrently capture local frequency nuances and global features. Central to the frequency layer is the FD-Block, which segregates two subbands, L and H, based on low and high channel filters. Subsequently, LL, LH, HL, and HH subbands are derived from the columns of subbands L and H using low-pass and high-pass filters, with LL representing low-frequency coarse-grained information, LH denoting high-frequency coarse-grained details, HL signifying low-frequency fine-grained data, and HH encapsulating high-frequency fine-grained elements. This meticulous approach ensures comprehensive coverage of image details without information loss, akin to an image decomposition process, with the inverse Wavelet Transform serving as the reconstruction phase. Notably, the frequency domain layer incorporates a 3*3 convolution between the Wavelet Transform and its inverse, enhancing the receptive field and acquiring more robust local information. The associated increase in computational cost and memory usage due to the 3*3 convolution is deemed negligible in light of the benefits it offers.

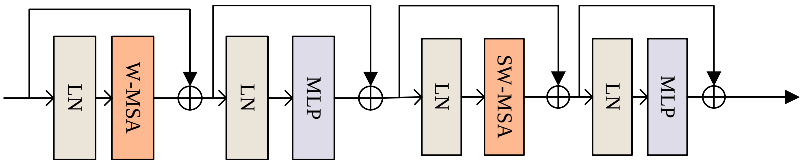


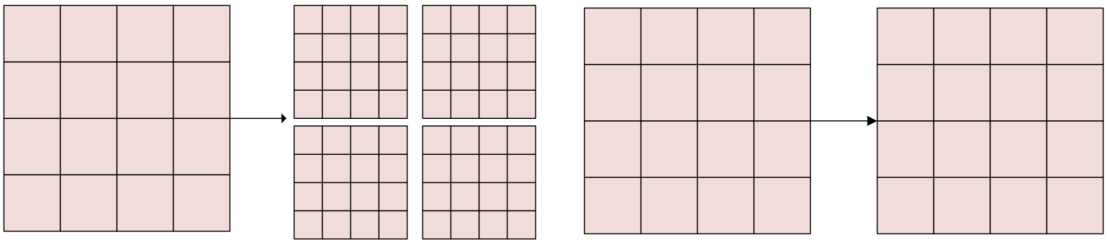
Fig. 2. Swin Transformer block

3.2 Attention Block

The attention layer of the standard Vision Transformer is based on the global calculation of attention, so its computational complexity is high. Instead, Swin Transformer introduces a layered structure approach commonly used in CNN to perform non-overlapping computation, thus reducing the computational complexity. As shown in **Fig. 3 (a)** and **(b)**, it can be seen from the figure that when Swin Transformer constructs feature maps, it adopts the form of window to separate the feature maps, and thus the feature maps constructed by it are hierarchical with the increasing of feature extraction layers. The vision transformer does not manipulate the feature maps. Compared with vision transformer, which calculates the whole map directly, it can greatly reduce the computation amount by calculating for each window. Two multi-head self-attentions are used in the Swin Transformer, the difference is that the former used window-based method (W-MSA), and the latter used shifted window-based method (W-MSA) to compute the attention. The attention layer is followed by the normalization layer and the MLP layer respectively. The mathematical expression of attention is shown in Eq. (1). In the paper, only the shift window based Swin Transformer block is used to implement the attention layer.

$$Attention(Q,K,V) = Soft\max(QK^{T} / \sqrt{d} + B)V \tag{1}$$

Where $Q,K,V \in \mathbb{R}^{M^2 \times d}$ are the query vector, key vector, and value vector, respectively. M^2 is the number of patches in the window. d denotes the feature dimension. The value of the bias matrix is $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.



(a) Swin Transformer

(b) Vison Transformer

Fig. 3. The process of constructing feature maps with different transformers

3.3 Fusion strategy

The features corresponding to each of the two source images, i.e., F_A and F_B, can be obtained through the feature extraction block (Encoder block). These two features are then fused to obtain the final feature map F .Specifically, the mean values of the different dimensions of the feature map are firstly obtained as shown in Eq. (2) and (3).

$$F_1 = mean(F_A) \tag{2}$$

$$F_2 = mean(F_B) \tag{3}$$

Then in the weight map is obtained by Softmax, which is shown in Eq. (4). Where Softmax is calculated as shown in Eq. (5).

$$A_1, A_2 = Soft\max(F_1, F_2) \tag{4}$$

$$A_i = \frac{e^{F_i}}{\sum_{j=1}^{C} e^{F_i}} \tag{5}$$

After obtaining the weight map, it is multiplied with the first feature maps F_A and F_B to obtain FA and FB, which are shown in Eq. (6) and (7). Finally, FA and FB are added together to get the final feature fusion map F as shown in Eq. (8).

$$FA = F_A * A_1 \tag{6}$$

$$FB = F_B * A_2 \tag{7}$$

$$F = FA + FB \tag{8}$$

3.4 Loss function

In the paper, SSIM based on image similarity is used as a loss function training model, SSIM is calculated as shown in Eq. (9), and the loss function is shown in Eq. (10).

$$\mathrm{SSIM}(O,G)=\frac{(2\mu_O\mu_G+C_1)(\sigma_{OG}+C_2)}{(\mu_O^2+\mu_G^2+C_1)(\sigma_O^2+\sigma_G^2+C_2)} \tag{9}$$

$$L_{\text{loss}}=1-\mathrm{SSIM}(O,G) \tag{10}$$

Where μ,σ is the mean and standard deviation respectively. O, G is the predicted output and true result of the model respectively. C_1, C_2 are constant positive numbers. This loss function will respond to the similarity between two images from brightness and contrast as structural information in the image.

4. Experimental Results and Discussions

4.1 Experimental Settings

4.1.1 Training dataset

Since there are not enough original datasets in the field of multi-focus image fusion to train the model, the datasets are made by ourselves. The raw HD images are collected from the DUTS dataset, which is used for salient object detection. Therefore, these datasets not only provided clear images but also masks, as shown in **Fig. 4 (c)** and **(d)**. Previous researchers used Gaussian filtering to blur images by making datasets. Blurring is applied to different areas so that a pair of datasets that can be used for training can be obtained. Considering that the transformer needs enough datasets for training to get better results, 9520 image pairs with different focuses are generated. As shown in **Fig. 4 (a)**, **(b)**, and **(d)**, the source images are input to the model, and then the output fusion results and the real images are lost to train the model. Specifically, the masks are binarized and then different thresholds are selected to blur the clear images to different degrees. Finally, the masks and blurred images are subjected to dot product operation, so that the datasets required for training can be obtained.

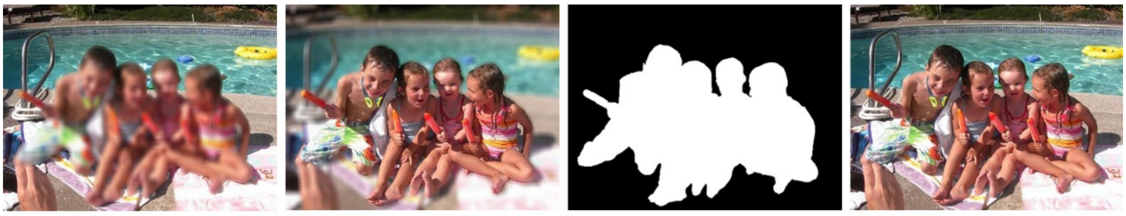

(a) Source A (b) Source B (c) Mask (d) Ground-truth

Fig. 4. Multi-focus image training dataset

4.1.2 Test dataset

Three pairs of datasets are tested to demonstrate the superiority of this experiment's performance. The test datasets are Lytro [37], MFI-WHU [38] and MFFW [39] datasets, and some images of the datasets are shown in **Fig. 5**. The Lytro datasets have a total of 20 pairs of colored images, which are commonly used test datasets in the field. They have a clear foreground and background because they are captured with far and near-focus binocular lens camera. Also, it contains 4 sequences of multi-focus images with 3 focal lengths. The MFI-WHU datasets were produced in 2021 by Zhang et al. It has a total of 120 pairs of images. The MFI-WHU datasets were based on the MS-COCO datasets and the MEF datasets and then synthesized into pairs of images by Gaussian blurring and manually annotated decision maps. Here the proposed method selects 9 pairs of MFI-WHU images for one test. The MFFW datasets were proposed to test whether the model can handle the defocus diffusion effect effectively. Since the Lytro and MFI-WHU datasets are not prominent in testing the impact of the diffuse focus effect on the task, this paper used 13 pairs of images from the MFFW datasets to illustrate that the model proposed in this paper can effectively cope with the diffuse focus problem.

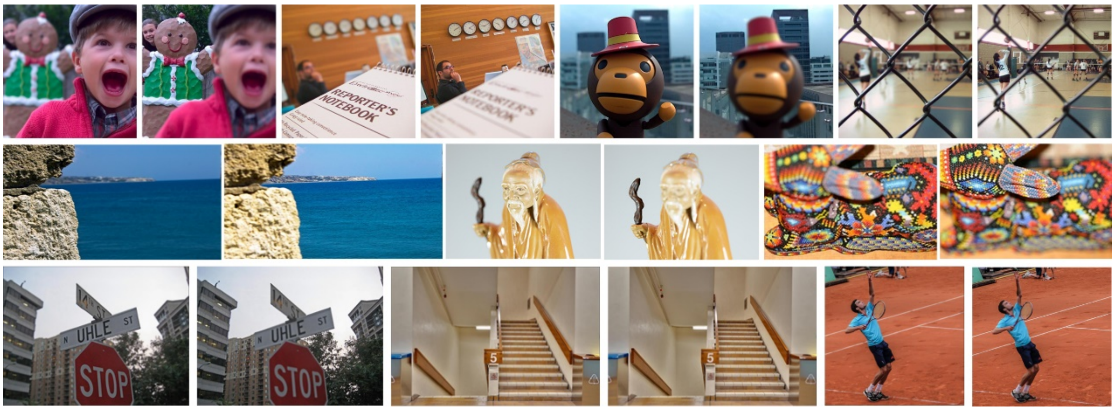


Fig. 5. Multi-focus image test dataset

4.1.3 Experimental details

In this paper, the pytroch framework is used to implement the code, the initial learning rate is 1e-4, and the learning rate is updated every 50 epoch. The epoch is set to 100 and batch_size is set to 8. The whole training and testing process is run on a RTX 3090 GPU (24G) and a 3.07GHz CPU. Each Conv layer in the model is a 3*3 convolution, and in order to ensure the consistency of the input and output channels of each residual block, a 1*1 convolution is used to reduce the dimensionality of the residuals after they are connected. For Attention block, dim=64, attention heads=4, Window size:4*4.

4.1.4 Evaluation metrics

Since none of these datasets mentioned above gives the final fused image, thus for the accuracy and validity of the experiments, the fusion results are evaluated in this paper using a combination of subjective visual effect comparisons and objective metric comparisons. Among them, Q_G [40] is used for the gradient measure, which responds to the richness of image detail information. Q_{NCIE} [41], on the other hand, is the nonlinear correlation information entropy, which measures the degree of dependence between the real and predicted images. Q_P [42] is a fusion metric based on phase coherence that evaluates the extent to which salient features are preserved in the source image. Q_{abf} [43] is a novel criterion for evaluating the quality of fused images, which utilizes a local metric to calculate the extent to which the salient information of an image is represented in the fused image, FMI_{pixel} 、 FMI_{dct} and FMI_{w} [44] are information theory based evaluation metrics that represent pixel, discrete cosine feature and wavelet feature mutual information, respectively. These seven metrics judge the superiority of the proposed model from different evaluation dimensions. Larger values of these seven metrics indicate better results.

4.1.5 Comparison methods

The comparison methods used in the paper include the current kinds of representative multi-focus image fusion methods. They are sparse representation-based methods (ASR [45]), spatial domain-based methods (PCNN [46]), gradient-based methods (GDF [47]), deep learning unsupervised methods (FusionDN [48], MFF-GAN [49]), deep learning supervised methods (MADCNN [50], UFA [51]), deep learning self-supervised methods (SMFuse [52]), Swin Transformer based methods (SwinFusion [53]) and universal fusion frameworks (IFCNN [54] and U2F [55]) It should be noted that the experimental parameters are set according to the values given in the original paper in order to get the best fusion results comparing the methods.

4.2 Comparison of experimental results

4.2.1 Visual effects comparison

As stated in 4.1.2, this paper uses three datasets to test the model and then selects representative images for each dataset separately to demonstrate the subjective visual effects. Taking the Lytro datasets as an example, in order to visualize the advantages and disadvantages of each method, this paper uses a difference plot for comparison, as shown in **Fig. 6**. We subtract the source image from the final fusion result of each method to get the difference map of each method, i.e., **Fig. 6 (a)-(l)**. The fusion result after subtracting the source image B should only have the focused part of the source image A remaining, i.e., only the pixel information of the fence is left. The less background indicates the better result. As can be seen from **Fig. 6**, GDF suffers from image distortion, while MFF-GAN suffers from the problem of not being able to recover the pixels of the source image, and it is obvious from **Fig. 6 (k)** that the difference map of U2F has a chromatic aberration. FusionDN also suffers from the same problem as that of MFF-GAN and U2F, only it is a little bit better than these two methods. SMFuse and SwinFusion, on the other hand, suffer from the problem of misjudgments of focused and de-focus pixels, which cannot accurately distinguish the focused and de-focus regions of the two source images. The misjudgments region of ASR has fewer misjudgments range than that of SMFuse and SwinFusion. Although PCNN, IFCNN and UFA are slightly better, they still misjudge pixels in some regions. Whereas the method proposed in this paper and MADCNN are the most robust and show better experimental results. MADCNN is a decision map-based method whereas the method proposed in this paper is an end-to-end method. Since MADCNN is fused with the decision graph obtained through post-processing, it is expected that its difference graph will get less background information.

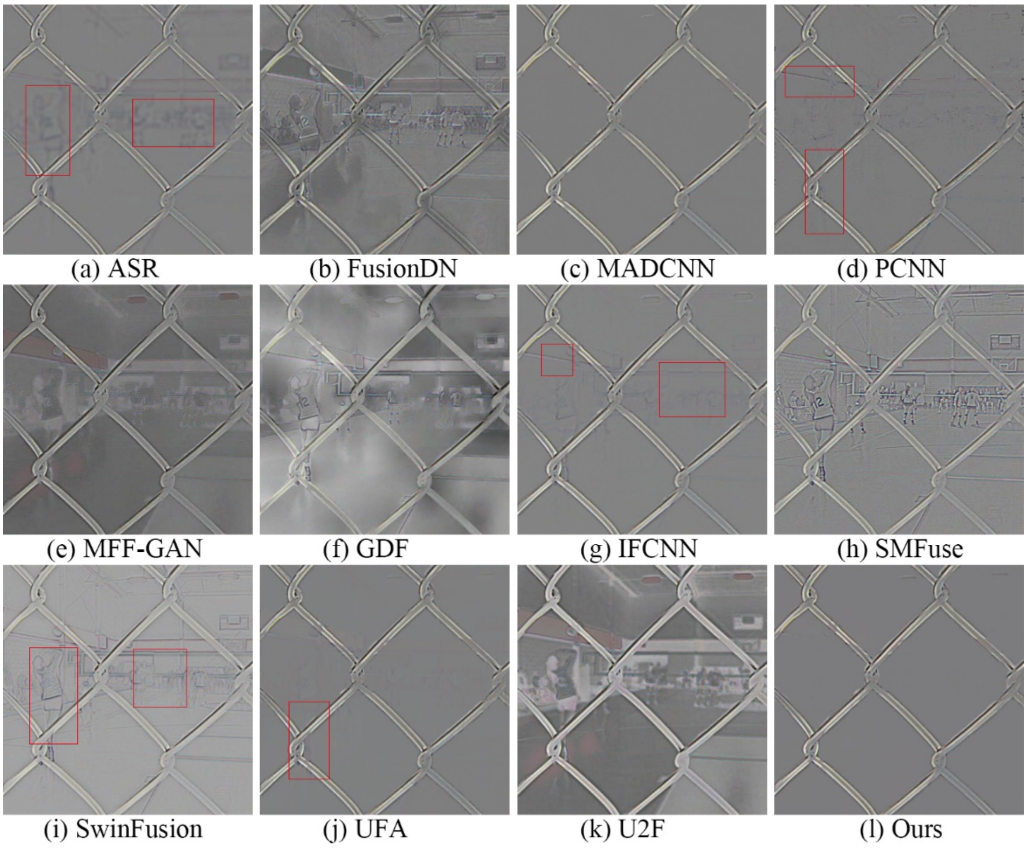
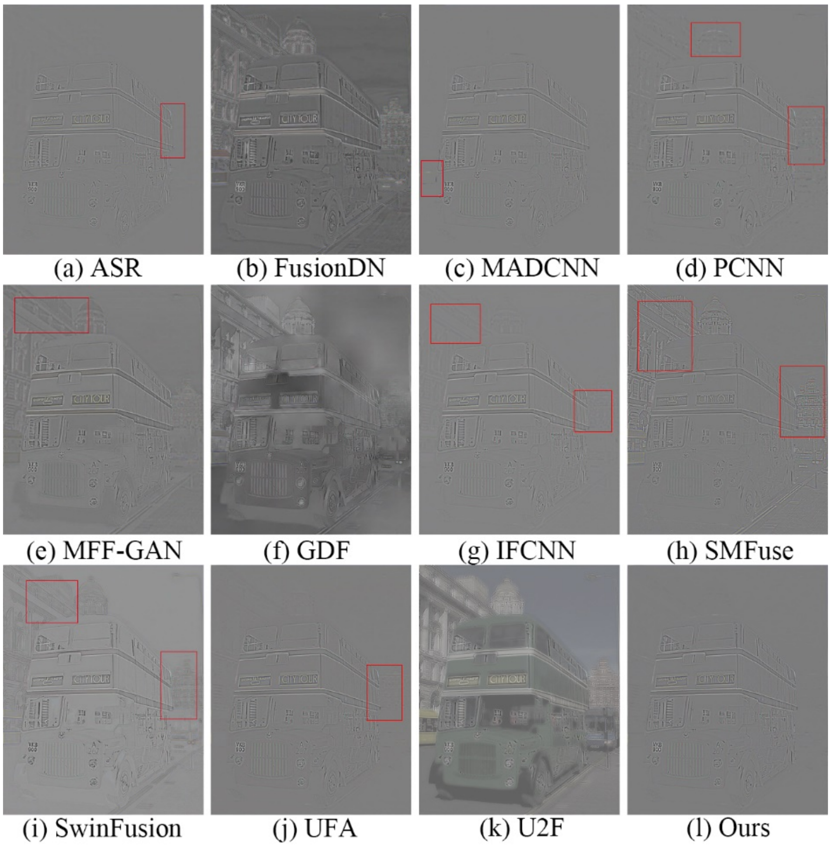


Fig. 6. Difference images with different fusion methods for Lytro

Further, taking the MFI-WHU datasets as an example, the subjective visual comparison effect of different fusion methods is also demonstrated in the form of difference plots. The demonstrated results are shown in **Fig. 7**. The more and more heterogeneous information contained in the difference map indicates that the method is less capable of retaining the information of the source image, i.e., the worse the fusion effect is. From **Fig. 7**, it can be seen that in the MFI-WHU datasets, GDF and U2F still have distortion problems, while U2F has more serious chromatic aberration problems. FusionDN does not have distortion problems but has colors in some parts, which indicates that the origin of pixels in the fused image is not from the source image. PCNN and UFA have artifacts in both in-focus and de-focus and at segmentation boundaries. MFF-GAN, SMFuse and SwinFusion on the other hand wrongly differentiate between a large amount of pixel information. IFCNN and PCNN besides introducing artifacts at segmentation boundaries contain information other than clear pixels in the source image A there is also other cluttered information. ASR and MADCNN have fewer problems than the other methods. However, it can be seen that ASR's lines in the red box are not smooth and even appear jagged. This situation will be reflected as block artifacts in the final fused image. MADCNN shows some messy lines in the box, which indicates that its effect still has some problems and the feature retention ability of the source image still needs to be improved. It can be seen that the method proposed in the paper is not only able to accurately distinguish the focus and de-focus regions, but also can effectively reduce the introduction of artifacts and the loss of detailed information.


(a) ASR (b) FusionDN (c) MADCNN (d) PCNN

(e) MFF-GAN (f) GDF (g) IFCNN (h) SMFuse

(i) SwinFusion (j) UFA (k) U2F (l) Ours

Fig. 7. Difference images with different fusion methods for MFI-WHU

Finally, the MFFW datasets are continued to show the comparison of different visual effects. The demonstrated results are shown in **Fig. 8**. From **Fig. 8**, it can be seen that IFCNN, SMFuse, ASR, FusionDN, MFF-GAN, SwinFusion and U2F have corrosion effects in the upper left corner. GDF and U2F have white lines, these white lines represent artefacts and noise the more white lines indicate the worse the fusion results. In addition, ASR, IFCNN, SMFuse and SwinFusion have misjudged the tree roots in the right half of the region. PCNN and UFA still have the problem of misjudgment in some regions. MADCNN is better compared to the results, but there are still two obvious noise spots. In summary, the final fusion image obtained by the method proposed in this paper is clearer, without noise and artifacts, and the overall effect is more natural and better than other methods.

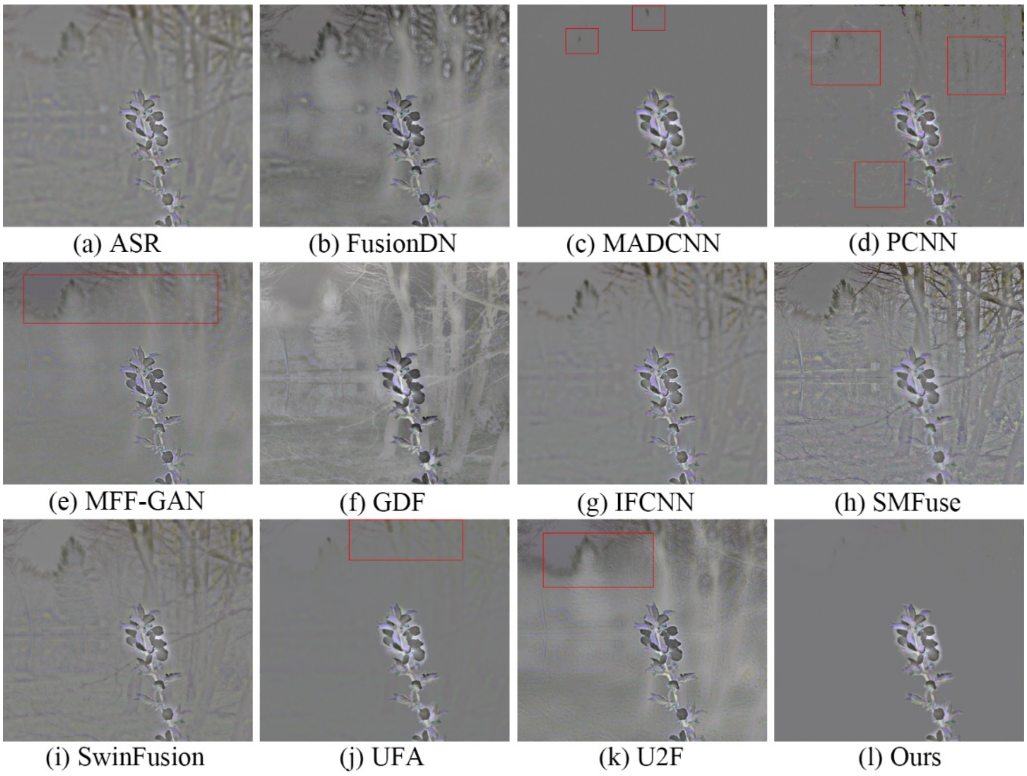


Fig. 8. Difference images with different fusion methods for MFFW

4.2.2 Quantitative comparison

In addition to illustrating the superiority of the proposed model in this paper in terms of subjective vision, it will also be further illustrated in terms of objective metrics. For a total of three testsets i.e. a total of 42 color multi-focus image pairs, the average values of the objective metric results of the different methods are shown in **Tables 1-3**. **Tables 1**, **2** and **3** represent the objective metric values of the different methods on the Lytro, MFI-WHU and MFFW datasets, respectively, with higher values of the evaluation metrics representing better results. The first, second and third values are indicated in the table by orange, blue and green colours respectively. From **Table 1**, it can be seen that the method proposed in this paper has gained the first place in four metrics and the second place in the remaining metrics. The worst performer in the Lytro datasets is SMFuse. while for the MFI-WHU datasets the method proposed in this paper is slightly inferior, in also located in the top three. The best performers on the datasets are ASR, MADCNN and the method proposed in this paper. **Table 3** also visualizes that the method of this paper is in the top three in all the indicators, while ASR has a mediocre performance, and although MADCNN also achieves better results, the gap between the method proposed in this paper and MADCNN is not very big. Moreover, MADCNN performs generally well on the ASR, so the method proposed in this paper is more advantageous from a comprehensive point of view. From the objective metrics comparison table, the method proposed in this paper shows certain superiority. In order to demonstrate more intuitively the superiority of the proposed method in terms of objective metrics, the average values of the different methods on the three test datasets are shown in **Fig. 9**. The horizontal coordinates represent the seven different evaluation metrics and the vertical coordinates represent the values, where higher values of these evaluation metrics represent better results. The comparison methods used and the method proposed in this paper are represented by lines of different colors. It is very obvious from **Fig. 9** that the red line is at the top of the list for each indicator and the red line represents the method proposed in this paper. In conclusion, the method proposed in this paper has more advantages in detail texture discrimination and visual fidelity while obtaining better fusion results.

Table 1. Objective metrics for different methods on Lytro
(Orange, blue, and green respectively represent the first, second, and third.)

	Q_G	Q_{NCIE}	Q_P	Q_{abf}	FMI_{pixel}	FMI_{dct}	FMI_w
ASR	0.7378	0.8298	0.8075	0.7348	0.8987	0.3992	0.5019
FusionDN	0.6018	0.8221	0.6216	0.5949	0.8833	0.2994	0.3779
MADCNN	0.7491	0.8372	0.8277	0.7473	0.8992	0.3959	0.4984
PCNN	0.7081	0.8355	0.7387	0.7042	0.8930	0.3455	0.3939
MFF-GAN	0.6652	0.8238	0.7148	0.6601	0.8915	0.3808	0.4252
GDF	0.7034	0.8139	0.7466	0.6989	0.8887	0.3662	0.4331
IFCNN	0.7337	0.8298	0.8178	0.7296	0.8965	0.3881	0.4567
SMFuse	0.5448	0.8243	0.5970	0.5388	0.8871	0.2901	0.3556
SwinFusion	0.7192	0.8266	0.7715	0.7140	0.8952	0.3774	0.4316
UFA	0.7418	0.8325	0.8241	0.7384	0.8991	0.4102	0.4852
U2F	0.6143	0.8221	0.6657	0.6091	0.8844	0.3069	0.3857
Ours	0.7479	0.8366	0.8353	0.7453	0.8995	0.4142	0.5124

Table 2. Objective metrics for different methods on MFI-WHU
(Orange, blue, and green respectively represent the first, second, and third.)

	Q_G	Q_{NCIE}	Q_P	Q_{abf}	FMI_{pixel}	FMI_{dct}	FMI_w
ASR	0.7337	0.8358	0.7317	0.7287	0.8827	0.4470	0.6038
FusionDN	0.4852	0.8184	0.5327	0.4771	0.8569	0.3182	0.4159
MADCNN	0.7322	0.8356	0.7323	0.7252	0.8815	0.4445	0.5823
PCNN	0.6753	0.8341	0.6820	0.6663	0.8788	0.4014	0.4983
MFF-GAN	0.6489	0.8203	0.6500	0.6397	0.8751	0.4295	0.4867
GDF	0.6899	0.8128	0.6517	0.6792	0.8716	0.4078	0.4527
IFCNN	0.6997	0.8256	0.7163	0.6910	0.8795	0.4253	0.4938
SMFuse	0.5435	0.8217	0.5623	0.5321	0.8670	0.3361	0.4341
SwinFusion	0.6873	0.8233	0.6463	0.6967	0.8778	0.4285	0.5053
UFA	0.7236	0.8311	0.7274	0.7164	0.8804	0.4478	0.5403
U2F	0.5848	0.8185	0.6339	0.5744	0.8678	0.3685	0.4243
Ours	0.7302	0.8346	0.7264	0.7239	0.8810	0.4485	0.5495

Table 3. Objective metrics for different methods on MFFW
(Orange, blue, and green respectively represent the first, second, and third.)

	Q_G	Q_{NCIE}	Q_P	Q_{abf}	FMI_{pixel}	FMI_{dct}	FMI_w
ASR	0.6433	0.8196	0.5530	0.6278	0.8799	0.2922	0.3009
FusionDN	0.5114	0.8182	0.4304	0.4956	0.8651	0.2438	0.2750
MADCNN	0.7198	0.8258	0.6874	0.7103	0.8838	0.3825	0.4167
PCNN	0.6411	0.8262	0.5151	0.6299	0.8705	0.2862	0.3332
MFF-GAN	0.5905	0.8179	0.4887	0.5744	0.8742	0.2855	0.2885
GDF	0.6915	0.8127	0.6386	0.6796	0.8789	0.3664	0.3852
IFCNN	0.6867	0.8218	0.6680	0.6765	0.8806	0.3796	0.4101
SMFuse	0.5132	0.8201	0.4521	0.4968	0.8743	0.2501	0.2739
SwinFusion	0.6912	0.8211	0.6464	0.6788	0.8816	0.3850	0.3933
UFA	0.6815	0.8228	0.6382	0.6729	0.8811	0.4025	0.4174
U2F	0.5537	0.8171	0.4784	0.5387	0.8690	0.2466	0.2733
Ours	0.6973	0.8255	0.6665	0.6894	0.8823	0.4071	0.4446

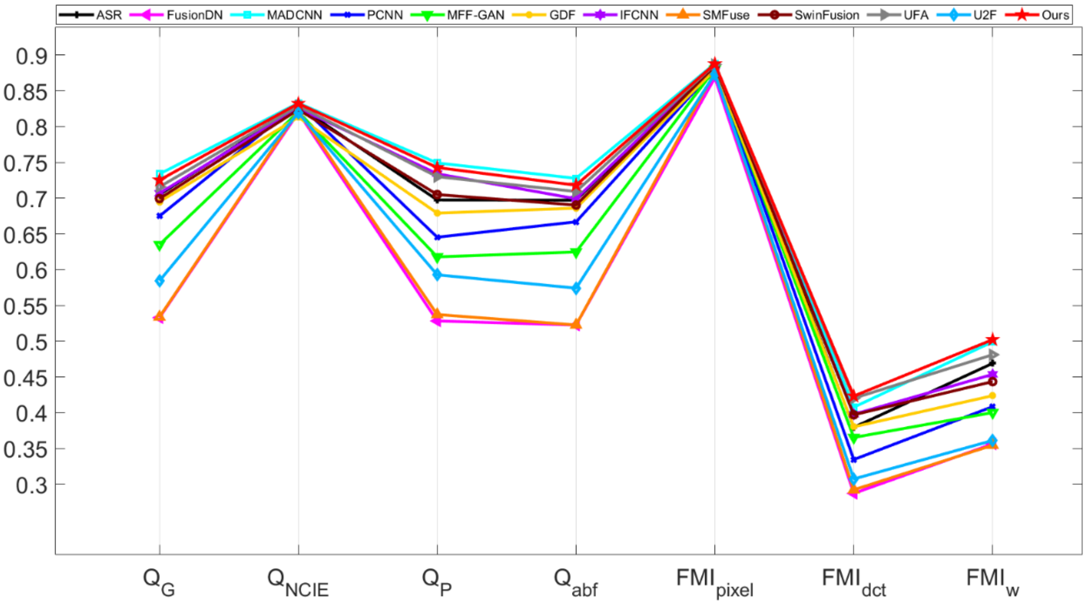


Fig. 9. Average of different methods on three testsets

4.3 More analysis

4.3.1 More fusion results

In order to further illustrate the effectiveness of the proposed approach, image pairs from three datasets were chosen based on the criteria outlined in Section 4.2.1. The resulting fused images, denoted as (a)-(l), were generated by combining source images A and B using various techniques. Subsequently, difference maps were created by subtracting source image B from the fused image. To enhance clarity, the difference maps within the red-boxed regions were magnified to facilitate a more intuitive comparison of the different methods. A proficient method should yield a difference map containing solely the information from source image A. The outcomes of this analysis are presented in **Fig. 10**. The fusion results reveal that the proposed method effectively concentrates all distinct pixels from source images A and B. This observation suggests that the fusion model introduced in this paper not only performs well on the previously examined image pairs but also produces exceptional visual outcomes on additional image pairs.

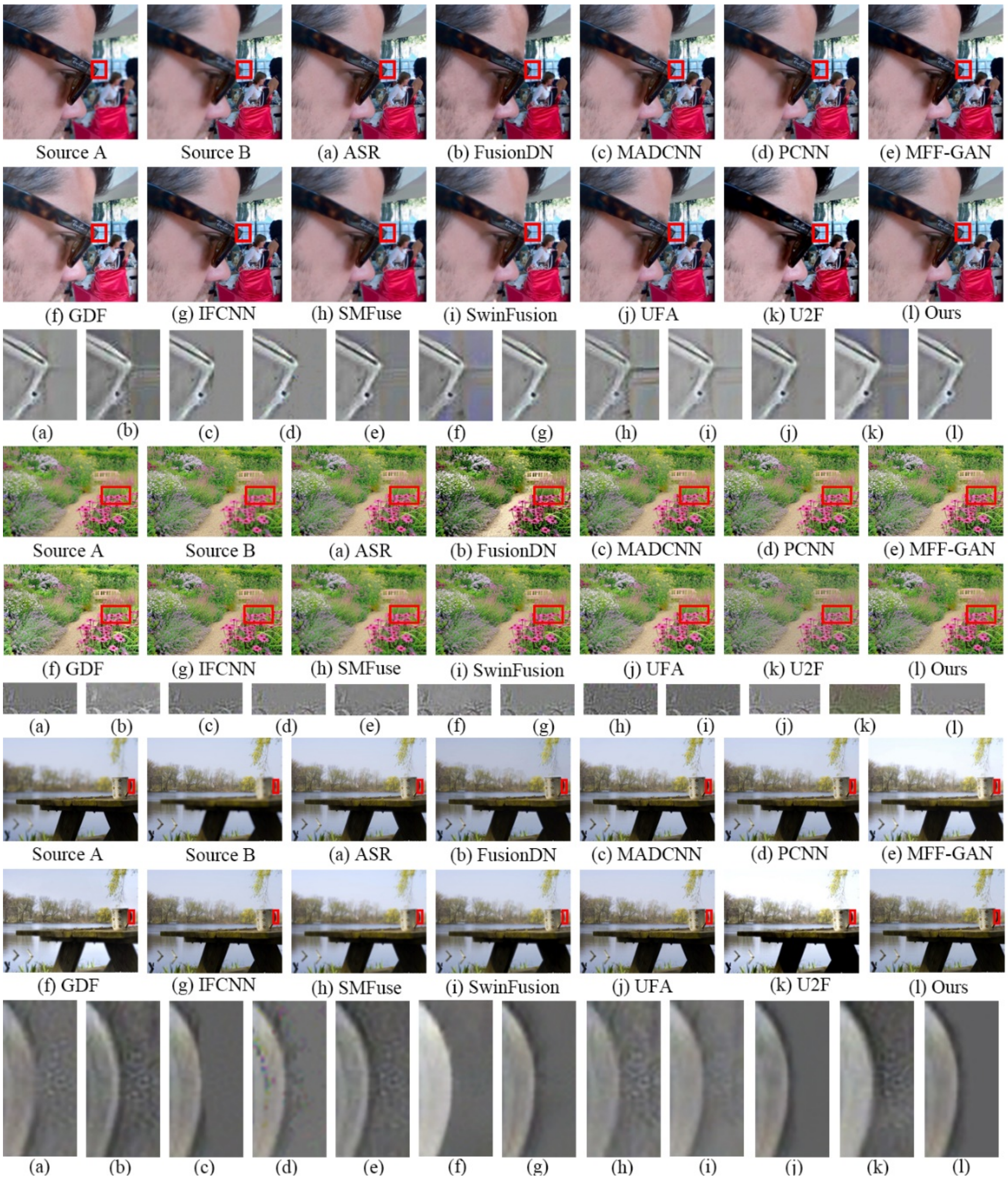
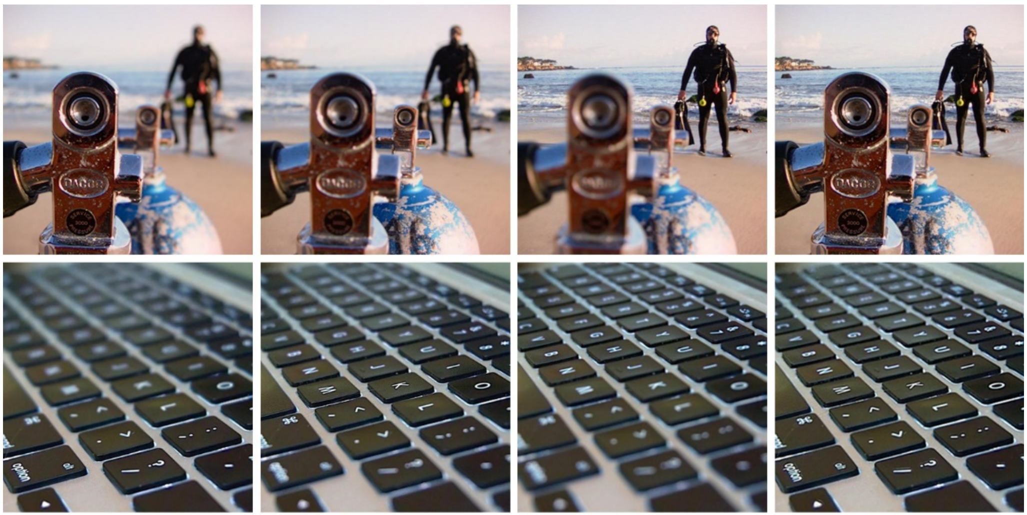


Fig. 10. More fusion results

4.3.2 Fusion results of three source images

The Lytro datasets not only provide two source images for testing, but also several image pairs with different regional focuses for further validation by the researcher. Thus, in the paper, the image pairs are selected for further validation. This experiment shows that the proposed method is not only limited to the fusion of two source images, but also applicable to the fusion of more than two images. The experimental results are shown in **Fig. 11**. In the **Fig. 11 (a)**, **(b)**, **(c)** and **(d)** are the source image A, source image B, source image C and fusion results. Obviously, using the fusion method proposed in this paper, the focused regions of the source images can all be integrated into the final images.



(a) Source A (b) Source B (c) Source C (d) Fusion image

Fig. 11. Fusion results of three multi-focus source images

4.3.3 Ablation experiments

In order to verify the validity of the combination of Wavelet Transform and Swin Transformer proposed in the paper. Experiments are also conducted on three datasets, Lytro, MFI-WHU and MFFW. Different combinations are experimented and then a comparison of the performance is made. These combinations include, (1) the original Swin Transformer architecture, (2) the attention layer in the front and the frequency domain layer is in the back, that is, the reverse architecture, and (3) the frequency domain layer is in the front, and the attention is in the back, which is the architecture proposed in the paper. The results of the objective metrics for these three network models are shown in **Table 4**, with the best values of the metrics indicated in red. It can be seen intuitively from the table that the architecture proposed in this paper is superior to the other two in terms of metrics. This is because the proposal of the frequency domain layer makes up for the shortcoming of local information loss caused by the attention layer to a certain extent. As a result, the method proposed in this paper is better than the transformer block that only uses the attention layer. Furthermore, the local information is first extracted and then the correlation of features is analyzed from the global level, thus helping the model to prepare to capture local features and global attributes. If the attention layer is used first, although the effect of this form is slightly improved compared with the original block, it still has its limitations. This is because the frequency domain layer cannot accurately process the global attribute information extracted by the attention layer. Therefore, the network proposed in the paper is reasonable and effective, effectively considering both local and global feature information.

Table 4. Objective metrics for different network on multi-focus image fusion pairs (red represent the first.)

	Q_G	Q_{NCIE}	Q_P	Q_{abf}	FMI_{pixel}	FMI_{dct}	FMI_w
Swin Transformer	0.7219	0.8313	0.7390	0.7161	0.8874	0.4218	0.4898
reverse network	0.7230	0.8320	0.7419	0.7170	0.8873	0.4220	0.5007
Ours	0.7251	0.8322	0.7427	0.7196	0.8876	0.4232	0.5021

5. Conclusion

The paper introduces a novel parallel model that integrates Wavelet Transform and Swin Transformer block with CNN for image fusion. This model adopts an end-to-end approach for fusion, enabling direct fusion of images through network training, thereby eliminating the need for subsequent operations. Comparative and ablation experiments demonstrate the superiority of the proposed WS-Transformer over the conventional Swin Transformer and existing multi-focus image fusion techniques. While the proposed method has shown promising results in multi-focus image fusion, further comprehensive research is required to explore its applicability in other fusion domains. Future work will focus on enhancing the model's generalization capabilities to facilitate its utilization across diverse fields.

Acknowledgement

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62003065); the Natural Science Foundation of Chongqing (General Program) (Grant No. CSTB2022NSCQ-MSX1231); the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202200564); the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202200504); the Fund Project of Chongqing Normal University (Grant No. 21XLB032).

References

[1] Li Q, Yang X, Wu W, Liu K, Jeon G, “Multi-Focus Image Fusion Method for Vision Sensor Systems via Dictionary Learning with Guided Filter,” *Sensors*, vol.18, no.7, 2018. Article(CrossRefLink)

[2] Ghandour, C. et al., “Applying medical image fusion based on a simple deep learning principal component analysis network,” *Multimedia Tools and Applications*, vol.83, pp.5971-6003, 2024. Article(CrossRefLink)

[3] Wang, Zeyu et al., “When Multi-Focus Image Fusion Networks Meet Traditional Edge-Preservation Technology,” *International Journal of Computer Vision*, vol.131, pp.2529-2552, 2023. Article(CrossRefLink)

[4] Wu, Pan et al., “Multi-focus image fusion: Transformer and shallow feature attention matters,” *Displays*, vol.76, 2023. Article(CrossRefLink)

[5] Bhat, Shiveta, and Deepika Koundal, “Multi-focus Image Fusion using Neutrosophic based Wavelet Transform,” *Applied Soft Computing*, vol.106, 2021. Article(CrossRefLink)

[6] Nie, Xixi et al., “A focus measure in discrete cosine transform domain for multi-focus image fast fusion,” *Neurocomputing*, vol.465, pp. 93-102, 2021. Article(CrossRefLink)

[7] Zhao, D. D., and Y. Q. Ji, “Multi-focus image fusion combining regional variance and EAV,” *Chinese Journal of Liquid Crystals and Displays*, vol.34, no.3, pp.278-282, 2019. Article(CrossRefLink)

[8] X. Qu, J. Yan, G. Xie et al., “A novel image fusion algorithm based on bandelet transform,” *Chinese Optics Letters*, vol.5, no.10, pp.569-572, 2007. Article(CrossRefLink)

[9] C. Duan, S. Wang, Q. Huang, “A Novel Multi-Focus Image Fusion Method Based on Dual-Tree Shearlet Transform,” *International Journal of Advanced Robotics Systems*, vol.12, no.6, 2015. Article(CrossRefLink)

[10] K. Wang, W. Li, X. Li, “Color image fusion algorithm based on NSCT and fuzzy logic,” *Electronic Science and Technology*, vol.29, no.4, pp.107-110, 2016. Article(CrossRefLink)

[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(CrossRefLink)

[12] Aslantas V., Kurban R., “Fusion of multi-focus images using differential evolution algorithm,” *Expert Systems with Applications*, vol.37, no.12, pp.8861-8870, 2010. Article(CrossRefLink)

[13] X. Bai, Y. Zhang, F. Zhou, and B. Xue, “Quadtree-based multi-focus image fusion using a weighted focus-measure,” *Information Fusion*, vol.22, pp.105-118, 2015. Article(CrossRefLink)

[14] 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(CrossRefLink)

[15] M. Amin-Naji, A. Aghagolzadeh, M. Ezoji, “Ensemble of CNN for multi-focus image fusion,” *Information Fusion*, vol.51, pp.201-214, 2019. Article(CrossRefLink)

[16] W.W. Kong, Y. Lei, “Multi-focus image fusion through DCNN and ELM,” *Electronics Letters*, vol.54, no.22, pp.1282-1284, 2018. Article(CrossRefLink)

[17] Wang, Haoran, Zhen Hua, and Jinjiang Li, “Two-stage progressive residual learning network for multi-focus image fusion,” *IET Image Processing*, vol.16, no.3, pp.772-786, 2022. Article(CrossRefLink)

[18] Ding, Cong, Ru Xue, and Shiming Niu, “Multibiometric Images Encryption Method Based on Fast Fourier Transform and Hyperchaos,” *International Journal of Bifurcation and Chaos*, vol.33, no.7, 2023. Article(CrossRefLink)

[19] Fumihiko Uesugi, “Novel image processing method inspired by Wavelet Transform,” *Micron*, vol.168, 2023. Article(CrossRefLink)

[20] Perez, Ronal A. et al., “Nonlinear Encryption for Multiple Images Based on a Joint Transform Correlator and the Gyrator Transform,” *Sensors*, vol.23, no.3, 2023. Article(CrossRefLink)

[21] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *Proc. of 2021 IEEE/CVF International Conference on Computer Vision*, pp.9992-10002, 2021. Article(CrossRefLink)

[22] Tian, Chunwei et al., “Multi-stage image denoising with the Wavelet Transform,” *Pattern Recognition*, vol.134, 2023. Article(CrossRefLink)

[23] Xue, Dongmei et al., “aiWave: Volumetric Image Compression With 3-D Trained Affine Wavelet-Like Transform,” *IEEE Transactions on Medical Imaging*, vol.42, no.3, pp.606-618, 2023. Article(CrossRefLink)

[24] Kumar, D. Maruthi, D. Satyanarayana, and M. N. Giri Prasad, “An improved Gabor Wavelet Transform and rough K-means clustering algorithm for MRI brain tumor image segmentation,” *Multimedia Tools and Applications*, vol.80, pp.6939-6957, 2021. Article(CrossRefLink)

[25] Zhang, Zhao et al., “An image fusion algorithm based on iterative Wavelet Transform for space non-cooperative targets,” *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol.237, no.10, pp.2228-2239, 2023. Article(CrossRefLink)

[26] Bae, W., Yoo, J., Ye, J. C., “Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification,” in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.1141-1149, 2017. Article(CrossRefLink)

[27] Liu, Pengju et al., “Multi-Level Wavelet Convolutional Neural Networks,” *IEEE Access*, vol.7, pp.74973-74985, 2019. Article(CrossRefLink)

[28] Oyallon, E., Belilovsky, E., Zagoruyko, S, “Scaling the Scattering Transform: Deep Hybrid Networks,” in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.5619-5628, 2017. Article(CrossRefLink)

[29] Peng, Hong et al., “Multi-focus image fusion approach based on CNP systems in NSCT domain,” *Computer Vision and Image Understanding*, vol.210, 2021. Article(CrossRefLink)

[30] Li, Xiaosong et al., “Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal,” *Signal Processing*, vol.184, 2021. Article(CrossRefLink)

[31] Dosovitskiy, Alexey et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proc. of International Conference on Learning Representations (ICLR)*, 2021. Article(CrossRefLink)

[32] Wei, Chen et al., “High-Resolution Swin Transformer for Automatic Medical Image Segmentation,” *Sensors*, vol.23, no.7, 2023. Article(CrossRefLink)

[33] Vs, Vibashan et al., “Image fusion transformer,” in *Proc. of 2022 IEEE International Conference on Image Processing (ICIP)*, pp.3566-3570, 2022. Article(CrossRefLink)

[34] Wang, Xuejiao, Zhen Hua, and Jinjiang Li, “Multi-focus image fusion framework based on transformer and feedback mechanism,” *Ain Shams Engineering Journal*, vol.14, no.5, 2023. Article(CrossRefLink)

[35] Qu, L., Liu, S., Wang, M., Li, S., Yin, S., Qiao, Q., Song, Z., “Transfuse: A unified transformer-based image fusion framework using self-supervised learning,” *arXiv:2201.07451*, 2022. Article(CrossRefLink)

[36] Liu, L., Liu, J., Yuan, S., Slabaugh, G., Leonardis, A., Zhou, W., Tian, Q., “Wavelet-Based Dual-Branch Network for Image Demoireing,” in *Proc. of Computer Vision – ECCV 2020*, pp.86-102, 2020. Article(CrossRefLink)

[37] https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

[38] https://github.com/HaoZhang1018/MFI-WHU

[39] Xu, Shuang et al., “MFFW: A new dataset for multi-focus image fusion,” *JOURNAL OF LATEX CLASS FILES*, vol.14, no.8, 2020. Article(CrossRefLink)

[40] C. S. Xydeas, V. Petrovic, “Objective image fusion performance measure,” *Electronics Letters*, vol.36, no.4, pp.308-309, 2000. Article(CrossRefLink)

[41] Q. Wang, Y. Shen, J. Jin, “Performance evaluation of image fusion techniques,” *Image fusion: algorithms and applications*, pp.469-492, 2008. Article(CrossRefLink)

[42] Zhao, Jiying, Robert Laganiere, and Zheng 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(CrossRefLink)

[43] M. Emmerich, A. Deutz, and N. Beume, “Gradient-Based/Evolutionary Relay Hybrid for Computing Pareto Front Approximations Maximizing the S-Metric,” in *Proc. of International Workshop on Hybrid Metaheuristics*, pp.140-156, 2007. Article(CrossRefLink)

[44] M. Haghighat, M.A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. of 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, pp.1-3, 2014. Article(CrossRefLink)

[45] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Processing*, vol.9, no.5, pp.347-357, 2015. Article(CrossRefLink)

[46] Mingrui C, Junyi Y, Guanghui C, "Multi-focus image fusion algorithm using LP transformation and PCNN," in *Proc. of 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp.237-241, 2015. Article(CrossRefLink)

[47] 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, 2016. Article(CrossRefLink)

[48] Xu, Han et al., "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(CrossRefLink)

[49] Zhang, Hao et al., "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol.66, pp.40-53, 2021. Article(CrossRefLink)

[50] Lai, Rui et al., "Multi-Scale Visual Attention Deep Convolutional Neural Network for Multi-Focus Image Fusion," *IEEE Access*, vol.7, pp.114385-114399, 2019. Article(CrossRefLink)

[51] 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(CrossRefLink)

[52] Ma J, Le Z, Tian X, Jiang J, "SMFuse: Multi-Focus Image Fusion Via Self-Supervised Mask-Optimization," *IEEE Transactions on Computational Imaging*, vol.7, pp.309-320, 2021. Article(CrossRefLink)

[53] Ma, Jiayi et al., "SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer," *IEEE/CAA Journal of Automatica Sinica*, vol.9, no.7, pp.1200-1217, 2022. Article(CrossRefLink)

[54] Y. Zhang, Y. Liu, P. Sun et al., "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol.54, pp.99-118, 2020. Article(CrossRefLink)

[55] Zang Y, Zhou D, Wang C et al., "UFA-FUSE: A Novel Deep Supervised and Hybrid Model for Multifocus Image Fusion," *IEEE Transactions on Instrumentation and Measurement*, vol.70, pp.1-17, 2021. Article(CrossRefLink)

Weiping Jiang received the B.E.(2021) degree from Anhui Normal University, China. She is a M.E. student, Department of Computer and Information Science, Chongqing Normal University.

Yan Wei received the M.E.(2001) and Ph.D.(2010) degrees from Chongqing University, China. He is a Professor, Department of Computer and Information Science, Chongqing Normal

Hao Zhai received the M.E.(2015) degree from Chongqing Normal University, China. And Ph.D(2020) degree from Nanjing University of Aeronautics and Astronautics. He is a teacher, Department of Computer and Information Science, Chongqing Normal University.