Transformer-Based Dual-Branch Multiscale Fusion Network for Pan-Sharpening Remote Sensing Images

Zixu Li, Jinjiang Li, Lu Ren, and Zheng Chen

***Abstract*—Due to the limitations of satellite sensors, we can only obtain MS images and PAN images separately. The focus of our attention is to utilize the pan-sharpening method to generate the high-resolution multispectral (HRMS) images. In this article, we proposed the dual-branch multiscale fusion network, which based on the spatial-spectral transformer to comprehensively capture the information contained in MS images and PAN images at different scales. The architecture of our network consists of three parts: during the feature extraction and image fusion stage, we first independently apply upscaling and downscaling operations to the MS and PAN images. Subsequently, we concatenate the images from the two distinct branches and input them into the shallow feature extraction module individually. And then we input them into our adaptive feature extraction block to further extract the crucial details of the images using the attention mechanism. The images a various scales in different branches are then passed through three spectral transformer and three spatial transformer modules to perform a comprehensive extraction of both spatial and spectral characteristics. Finally, the residual local feature module is utilized during the image reconstruction part to deeply extract intricate information from the images and obtain the final HRMS fused image. We have conducted both simulated and real experiments on the benchmark datasets QB and WV2. The final qualitative and quantitative comparative results demonstrate that our innovative method outperforms the current SOTA methods.**

***Index Terms*—Attention mechanism, convolutional neural network, pansharpening, transformer.**

I. Introduction

REMOTE sensing image technology has received widespread attention since its inception, the application scope includes target detection, image segmentation, image fusion, etc, affecting all aspects of human production activities. [1], [2] And due to technical defects, We are unable to acquire the multispectral images at high resolution on the existing remote sensing sensors simultaneously, how to make good use of the advantages of the respective sensors to obtain the results we want to become the focus of our attention. [3] Through the pan-sharpening method, We have the capability to combine the MS images captured by the sensors with PAN images to generate the targeted HRMS images [4]. As a foundational and continually evolving area of research, the pan-sharpening method has been developed for nearly 40 years, which extracts and fuses the rich multispectral distributions contained in MS images and the rich spatial structural information contained in PAN images to obtain the HRMS images we need [5], [6]. It has also been proved through theory and experiment that the method is widely used in computer vision applications, such as coastal zone monitoring [7], land change detection [8], anomaly detection [9], etc. The pan-sharpening method consists of three principal components described as follows.

Founded on the CS transformation, which is often referred to as the spectral modification method, the initial MS image is spectrally transformed and the panchromatic image is decomposed into multiple bands. The fusion process involves substituting the abundant spatial information within the PAN image with the spatial details from the MS image, ultimately yielding high-resolution multispectral images. Owing to the uncomplicated nature and efficiency of this approach, many well-known and efficient algorithms have been derived, such as Gram-Schmidt (GS) [10], principal component analysis (PCA) [11], Wavelet Transform-Based [12], intensity-hue-saturation (IHS) [13] methods. In addition, the nonlinear PCA [14] and the nonlinear IHS [15] methods are employed to discern finer details within the images, the various types of feature components of the image are also more easily identified. Despite the ease of implementation and the substantial improvement in spatial information acquisition achieved by methods based on the CS transform, they ultimately lead to the spectral information loss and distortion.

The spatial layout of high-resolution PAN images is deconstructed in MRA-based pan-sharpening methods, employing wavelet transformation or the generation of a Laplace pyramid, and Embeds the separated spatial information into the LRMS images, resulting in images endowed with abundant spatial data and an evenly distributed spectral profile. The MRA-based pan-sharpening methods are also simple to implement. The traditional MRA algorithms includes three steps: multiresolution image decomposition, image fusion, and reconstruction. It is now more widely understood that a unified framework should be proposed to avoid the complex three-step processing. The improved adaptive intensity-hue-saturation (IAIHS) [16],

Manuscript received 18 September 2023; revised 6 November 2023; accepted 9 November 2023. Date of publication 13 November 2023; date of current version 29 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61772319, Grant 62002200, Grant 62202268, and Grant 61972235, in part by the Shandong Natural Science Foundation of China under Grant ZR2023MF026 and Grant ZR2022MA076, and in part by the Yantai science and technology innovation development plan under Grant 2022JCYJ031. *(Corresponding author: Lu Ren.)*

Zixu Li is with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China (e-mail: a1525992325@163.com).

Jinjiang Li, Lu Ren, and Zheng Chen are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: lijinjiang@gmail.com; renlu@sdtbu.edu.cn; chenzheng@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3332459

effectively enhances the spatial-spectral information present in the fused images. This enhancement is made possible through the creation of a correlation weight matrix that establishes a relationship between the spatial structure in the PAN images and the spectral attributes of the MS images. Other methods include generalized Laplacian pyramid (GLP)-based methods [17], [18], [19], additive wavelet luminance proportional (AWLP) [17], [18] methods, curvelet transform [20] method, high-pass filtering (HPF) [21], and other methods.

The VO-based pan-sharpening methods is formed on the basis of variational theory [22], the main content is to find the best fusion transformation function that satisfies specific constraints. The P+XS [23] method aroused a strong reaction once it was put forward, which made the VO-based pan-sharpening methods gain more attention and development. Sparse representations (SR) [24] exploit the connection between LRMS images and PAN images to construct dictionary elements and finally inject high-resolution spatial information into MS images. By combining SR and wavelet transform, The fine-grained and coarse-grained information of the image is improved in stages, which is more efficient than the traditional SR [25].

Over the last few years, due to the continuous growth and advancement in science and technology, based on deep learning (DL) pan-sharpening methods have gradually achieved great success in the direction of remote sensing image processing. In contrast to the aforementioned traditional techniques, the basic principle of the methods is to learn and construct the network parameters between the observation samples and the fused images. In addition, the network's structure can be further enhanced by persistently optimizing the loss function between samples and the fused images, which is particularly efficient when dealing with large-scale datasets. However, there are exists some problems when training the network, how to construct a better and more efficient network structure with more reasonable network parameters, these two factors determine the performance of the network. In this article, we propose a two-branch multiscale fusion network structure based on the transformer module for pan-sharpening, we demonstrate the performance of this method, and carry out relevant qualitative and quantitative analyses and comparative experiments on two benchmark datasets, this article chiefly contributes to the field are as follows:

1) We propose the spectral transfomer to enhance the extraction of spectral information from images. This is accomplished by performing self-attention computations in the spectral domain while implementing the multihead mechanism in the spatial domain.
2) To address the limitations of the transformer module in capturing fine-grained details, we have utilized multiscale band/patch embeddings for extracting multiscale spectral/spatial information from images. Leading to a significant improvement in the overall network's performance.
3) We introduce the AFEB module into the feature extraction process so that the network model focuses on the important information within the images while disregarding the irrelevant information for the current task.

The rest of the article is organized as follows: Section II mainly describes the background of the DL-based pan-sharpening methods, the attention mechanism, and the transformer modules. Section III provides a more detailed and specific analysis of the network structure. Section IV provides an extensive overview of the experimental component of this article, showcasing the effectiveness of our proposed method. This is substantiated through qualitative and quantitative comparisons of fusion results against nine current SOTA techniques on the QB and the WV2 datasets. Finally, Section V concludes this article.

II. Related Work

A. DL-Based Pansharpening

With the explosion of DL in recent years, the potential for its application in the field of remote sensing imagery is also steadily advancing. Zhu et al. [26] established a network structure with multiple layers and formed the foundation of the deep network structure. Masi et al. [27] proposed a pan-sharpening neural network (PNN) had three network layers and efficiently extracted image features but struggled with deeper feature information. Scarpa et al. [28] proposed a target-adaptive pan-sharpening method using residual networks to deal with the discrepancy between the dataset and various sensors during training. The DL pan-sharpening method assumes a complex nonlinear relationship between observation samples and fused images. [29] The DNN method learns the parameters between these samples and fused images. In conducting the experiments, in the absence of the actual HRMS images, we acquired the training samples following the Wald [30] protocol. Wei et al. [31] integrated a residual network into the network structure, enabling the network to extract deeper-level feature information from images through nonlinear transformations. Xu et al. [32] constructed a pan-sharpening method based on edge information-guided pan-sharpening using sparse coding matrices, which further expanded the pan-sharpening content. Yang et al. [33] introduced the PanNet network, which combines residual networks with an up-sampled MS image to bolster the network's proficiency in feature extraction from images. This strategy facilitates the extraction of spatial details, while retains the spectral distribution. Based on the CoF fusion algorithm, Tan et al. [34] proposed the CoF-MSMG-PCNN method, which the PAN image is decomposed into three scales, and then fused with the HIS-converted MS image and reconstructed to attain the eventual fused image. The multiscale image feature extraction module (MSDCNN) proposed by Yuan et al. [35] employed multiscale convolutional operations to capture diverse image features, enhancing texture details within the image by integrating a residual network. Zhange et al. [36] introduced the TDNet, which includes three structures: bidilevel, bidibranch, and bidirectional. This network effectively captures the full richness of spatial architecture and subsequently integrates this information into the MS image during fusion stage. Ma et al. [37] extended the pan-sharpening research by proposing an unsupervised adversarial network-based fusion method, which does not require observation samples in the entire training

process and retains spatial-spectral information through the constructed spatial-spectral generator. On this basis, Zhou et al. [38] proposed a PGMAN framework for pan-sharpening, which further extends its application in GAN networks. Uezato et al. [39] proposed a guided deep decoder network as a general prior, which performed well in various image fusion tasks.

B. Attention Mechanism

Attention mechanisms have made rapid progress in DL applications due to their powerful feature extraction capabilities, and they have achieved significant success in various upstream and downstream tasks. Such as semantic segmentation [40], natural language processing [41], and image processing [42]. Attention mechanisms enable models to prioritize the most relevant information for the current task and disregard less important data. This enhances computational efficiency and reduces parameter complexity. Currently, spatial attention, channel attention, and hybrid attention mechanisms are commonly used. Hu et al. [43] presented the innovative "squeeze-and-extraction network (SENet)" network. This structure incorporates a global compression part for feature maps within individual channels. The learned weights are then employed to extract and activate features contained in feature maps. Finally, these learned weights are applied to the original feature maps, either amplifying or suppressing specific channel features. Li et al. [44] proposed Selective Kernel Networks (SKNet), in which the attention mechanisms are primarily utilized to adaptively select convolutional layers with various kernel sizes. This process aids in determining feature weights at different scales, allowing the network to selectively fuse multiscale information. Wang et al. [45] proposed. the efficient channel attention (ECA), which prioritizes the significance of individual channels. Furthermore, the ECA module avoids interchannel interactions, thereby enhancing the network's performance under limited computational resources. [46] The spatial information network (SPANet) additionally introduces an additional branch for handling spatial information and integrates the extracted spatial information into the primary feature extraction branch, consequently enhancing the model's effectiveness in visual task [47]. By combining channel attention and spatial attention mechanisms, the CBAM framework intelligently prioritizes different channels and spatial domains, thereby enhancing its capability in feature extraction. [48] The Bam module dynamically adjusts feature channel importance using learned weight ratios to enhance overall network performance.

C. Transformer Module

In the field of image processing, while CNNs primarily focus on local feature extraction, the transformer module enables the model to capture feature dependencies over long distances and simultaneously process inputs from all locations. Transformer network has already proved their advantages in processing sequential data. [49] The Spectralformer network generates grouped spectral embeddings by extracting local sequence information from neighboring bands in a multispectral image. It also incorporates a cross-layer connection to adaptively learn cross-layer fusion information. Many researchers have also tried to integrate the benefits of CNN and the transformer model to design a more efficient network structure, He et al. [50] presented the spatial-spectral transformer, which employs a tailored CNN framework to extract spatial features and incorporates an adjusted transformer module to capture the spectral characteristics present in the image. [51] The Swin transformer leveraged the strengths of both CNN and transformer by performing self-attention computations within fixed-size windows. It also utilized sliding window operations, including nonoverlapping and overlapping cross-windows, progressively increases the model's perception of the image at each layer. Zhang et al. [52] proposed a multiscale spatial-spectral interactive attention mechanism that combines globally and locally extracted features from LRMS and PAN images. This effectively enhances information complementarity while reducing redundant features. Li et al. [53] treats the degradation process of HRMS images as a unified variational problem and introduces a local-global transformer for the prior image denoising module, the final fused image is achieved after a series of iterative approximations.

The algorithmic inference process of our overall framework, as shown in Algorithm 1.

Algorithm 1: The Implementation Process of our DMFN Model.

Input: $\alpha, \beta(\ MS\ \ PAN\ )$
Output: $F^{Out}(\ the\ final\ pansharpening\ image\ )$
// $step1 : The\ connection\ of\ images$
$*Shallow\ feature\ extraction$
$\mathbf{F_1} = Concat(\boldsymbol{\alpha}, \beta_{down}), \mathbf{F_2} = Concat(\alpha_{up}, \boldsymbol{\beta})$
end
// $step2 : The\ feature\ extraction\ of\ the\ network$
$*Shallow\ Feature\ fusion$
$\gamma = SFE(F_1), \eta = SFE(F_2)$
$\gamma' = AFEB(\gamma), \eta' = AFEB(\eta)$
$*Deep\ Feature\ extraction$
For i $in\ 3$ **do**
$\sigma = SPET(\gamma')$
$\upsilon = SPAT(\eta')$
end
end
// $step3 :\ The\ feature\ fusion\ and\ the\ image\ reconstruction\ of\ the\ network$
$\sigma^{up} = Upsample(\sigma)$
$F^{\delta} = Concat(\sigma^{up}, \upsilon)$
For i $in\ 3$ **do**
$F^{R} = RDLFB(F^{\delta})$
end
$F^{Out} = ESA(F^{R})$
end
// $step4 :$
$Calculate\ the\ loss\ to\ optimize\ training\ process$
$l_{loss} = MSE(x, y) + \mu L_{ap}(x, y)$

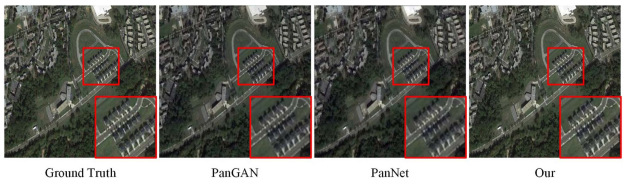


Fig. 1. Fusion result image of pan-sharpening methods.

III. Model Construction

The following section will meticulously explain the implementation algorithms and the fundamental architecture of the Pan-sharpening method as shown in Fig. 2. In the first subsection, we analyze the overall design idea and structure of the DMFN network. In the second subsection, we explain several important modules introduced in the paper. And in the third subsection, we summarize the relevant loss functions set in this article.

A. DMFN Framework

Fig. 2 illustrates the primary structure of our DMFN method, which is divided into three critical components: feature extraction, image fusion, and image construction.

Feature extraction and image fusion part: we first use bilinear interpolation for the LRMS image $\alpha \in R^{h\times\omega\times B}$(h,ω,B symbolize the height, width, and spectral number of the LRMS image, respectively) and the PAN image $\beta \in R^{H\times W}$, which execute up-sampling and down-sampling operations to obtain $\alpha_{\text{up}} \in R^{H\times W\times B}$ and $\alpha_{\text{up}} \in R^{H\times W\times B}$. It can be articulated using the subsequent equation.

$$\alpha_{\text{up}} = Up(\alpha), \beta_{\text{down}} = \text{Down}(\beta) \tag{1}$$

where the $Up(\cdot)$ and $Down(\cdot)$ operations denote the up-sampling and down-sampling operations of bilinear interpolation. Then, we connect α and β_{down} to get $\alpha_{\text{cat}} \in R^{h\times\omega\times(B+1)}$, and connect β and α_{up} to get $\beta_{\text{cat}} \in R^{H\times W\times(B+1)}$, which can be expressed as follows:

$$\alpha_{\text{cat}} = \text{Concat}\left(\alpha, \beta_{\text{down}}\right), \beta_{\text{cat}} = \text{Concat}\left(\beta, \alpha_{\text{up}}\right). \tag{2}$$

The $Concat(\cdot)$ operation represents the connection between the channel dimensions, the connection between the transmembrane states can realize the cross-dimensional information interaction between the two branches.

Since the convolution operation can easily and efficiently map the image from low to high dimensions, in the shallow feature extraction module, we adopt multiscale convolution to capture the image features at different levels to enhance the model's ability to generalize effectively while gradually expanding the sensory field. The shallow feature extraction module consists of 2-D convolution operations with convolution kernels of 3, 5, and 7, respectively, the output channels are 16, 16, and 32 and step sizes of 1, 2, and 3. The outputs F_s and F_s' can be formulated using the equation provided as follows:

$$F_s = \text{SFE}\left(\alpha_{\text{cat}}\right), F_s' = \text{SFE}\left(\beta_{\text{cat}}\right) \tag{3}$$

where $SFE(\cdot)$ denotes the shallow feature extraction module, and then we input the obtained feature images into the adaptive feature extraction block (AFEB) on each branch, respectively, for further processing to obtain $F_\alpha \in R^{h\times\omega\times C}$ and $F_\alpha' \in R^{H\times W\times C}$, which can be symbolized by the equation that follows:

$$F_\alpha = \text{AFEB}\left(F_S\right), F_\alpha' = \text{AFEB}\left(F_S'\right). \tag{4}$$

The output feature images are then fed into our introduced L=1,2,3 spectral transformer (SPET) and spatial transformer (SPAT) for deep feature extraction. Before F_α is input to SPET, the spectral embedding is set to $B_L^0 \in R^{C\times D_{spe}}$ for the SPET feature extraction module in the Lth layer, where $C = 16 \times 2^{L-1}$ is the number of channels, and the D_{spe} is set to 32. Similarly before F_α' is input to SPAT, we set the patch embedding for its SPAT module in the Lth layer to $P_L^0 \in R^{N\times D_{\text{spa}}}$, and D_{spa} is set to 256. Finally, we input to each SPET and SPAT module to obtain the intermediate values $B_L^i (i = 1, \ldots, 5)$ and $P_L^j (j = 1, \ldots, 5)$, which are expressed as the following equations:

$$B_L^i = \text{SPET}_i\left(B_L^{i-1}\right), i = 1 \ldots 5 \tag{5}$$

$$P_L^j = \text{SPAT}_j\left(P_L^{j-1}\right), j = 1 \ldots 5. \tag{6}$$

$SPET_i(\cdot)$ and $SPAT_j(\cdot)$ denote the SPET and SPAT modules in layer i and layer j, respectively. The learned weights are used to aggregate the extracted multiscale features to obtain $F_{\text{sum}}^{\text{spe}} \in R^{h\times\omega\times C}$ and $F_{\text{sum}}^{\text{spa}} \in R^{H\times W\times C}$.

Shallow feature extraction can extract the shallow information from the image through a straightforward and efficient process, while SPET and SPAT modules are more focused on extracting the intricate feature characteristics within the image, so in this article, we adopt the method of combining the shallow feature extraction module and the deep feature extraction module (includes SPAT and SPET) with the skip connection operation to ensure that the network model obtains the shallow features while maintaining the deep feature information, which can be illustrated through the equation that follows:

$$F^{\text{spe}} = F_{\text{sum}}^{\text{spe}} + F_\alpha, F^{\text{spa}} = F_{\text{sum}}^{\text{spa}} + F_\alpha'. \tag{7}$$

In order to ensure that the image sizes on the two branches are consistent before image reconstruction, we use a subpixel convolutional layer to execute an up-sampling procedure on the final feature image F^{spe} to obtain $F_{\text{up}}^{\text{spe}}$, and a spatial-spectral feature map $F \in R^{H\times W\times 2C}$ is obtained by connecting $F_{\text{up}}^{\text{spe}}$ and F^{spa}. Subsequently, utilize it as input for image reconstruction to generate the HRMS image.

The image reconstruction part: The images after the linking operation are inputed into our proposed residual local feature module (RLFM) to further extract and preserve the image's local structural characteristics. The residual dense local feature block (RDLFB) discards multiple feature distillation connections and simply adopts six stacked convolution operations and ReLU activation function for local feature extraction, which greatly reduces the computation time while maintaining the capacity of the model. After three RDLFB in sequence, it is then passed

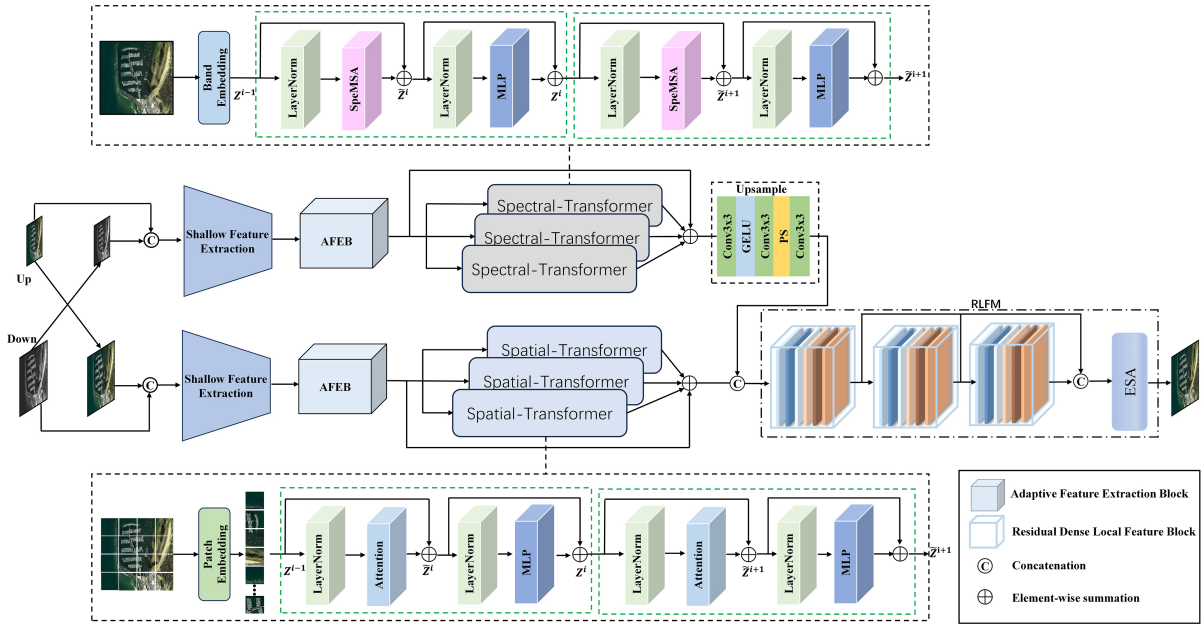


Fig. 2. Overall framework of DMFN network for remote sensing image fusion.

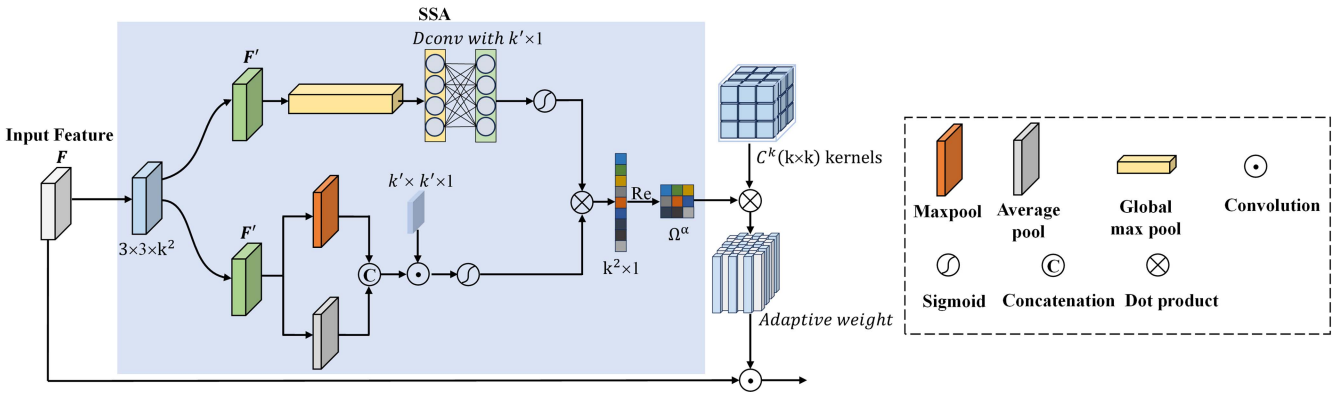


Fig. 3. Architecture of AFEB.

into the enhanced spatial attention (ESA) module for further processing, ultimately resulting in the HRMS image.

B. Adaptive Feature Extraction Block

As depicted in Fig. 3, aim to narrow the perceptual gap between the estimated spatial-spectral details and the authentic images, we propose the AFEB module in this article. First, the feature information is extracted from different local regions of the image using a convolutional kernel $K \times K(C^k)$, and the parameters are adaptively adjusted according to the learning weights Ω^α. In order to improve the learning efficiency, we set K to 3 here. The learning weight Ω^α is obtained by constructing the spectral-spatial attention (SSA) mechanism to combine the spatial-channel attention, the specific structure of SSA is shown as follows.

The purpose of SSA is to acquire the weight matrix Ω^α that matches the size of convolution kernel K $\times$ K. First, the input feature map F is proceeded through a $3 \times 3 \times K^2$ convolution layer, which fuses the channel information while reducing the dimension of the feature map. Then, the spectral and spatial weights are obtained by using spectral and spatial attention, respectively. Spectral weights correspond to the spectral attributes of each spectrum, while spatial weights signify the spatial characteristics within each spectral channel. Each spectrum of the image has unique spatial-spectral information, With the aim of enhancing the harmonization between spatial and spectral information, we sequentially carry out the inner product of the two weights, we obtain the interleaved attention weight matrix Ω^α after perfume the reshape operation. The whole process can be mathematically defined by the following equation:

$$
\begin{aligned}
F' &= \mathrm{Conv}(F) \\
O_\Psi^\tau &= \mathrm{Sigmoid}\left(\mathrm{Conv}_{1D}\left(GMP\left(F'\right)\right)\right) \\
O_\Psi^\pi &= Sigmoid\left(Conv\left(Cat\left(MP\left(F'\right), AP\left(F'\right)\right)\right)\right) \\
\Omega^\alpha &= Re\left(O_\Psi^\tau \otimes O_\Psi^\pi\right)
\end{aligned} \tag{8}
$$

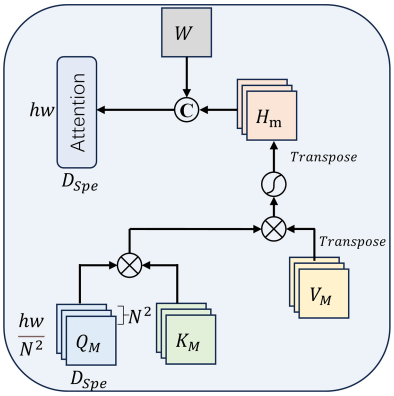


Fig. 4. Architecture diagram of the proposed spectral transformer.

where F denotes the local feature map input to the AFEB, O_Ψ^τ, O_Ψ^π denotes the spectral and spatial attention respectively, $\text{Conv}_{1D}(\cdot)$ denotes the 1-D convolution, $\text{GMP}(\cdot)$ represents the application of global max-pooling, $MP(\cdot)$ and $AP(\cdot)$ denote the max-pooling and average-pooling operations, respectively, $\otimes$ denotes the inner product operation, $Re(\cdot)$ denotes the reshape operation, and $\odot$ denotes the convolution operation. The adaptive local convolution kernel are obtained by weighting Ω^α and C^k with the corresponding weight values of each local. The convolution computation on the feature map is executed using the adaptive local convolution kernel we have obtained to get the features under different local regions, and the final output O_a is presented as follows:

$$O_a = \Omega^\alpha \otimes C^k \odot F. \tag{9}$$

C. Spectral Multihead Self-Attention

In SPET, the spectral self-attention mechanism is cleverly designed to obtain the correlation between spectra by calculating the self-attention on the spectral dimensions, as shown in Fig. 4, we first obtain the query matrix $Q \in \Bbbk^{hw \times D_{Spe}}$, the key matrix $K \in \Bbbk^{hw \times D_{Spe}}$ and the value matrix $V \in \Bbbk^{hw \times D_{Spe}}$ by the the trainable linear transformation computation shown as follows:

$$Q = B_l^i W^Q, K = B_l^i W^K, V = B_l^i W^V \tag{10}$$

where W^Q, W^K, and W^V are learnable mapping matrices, and then a scalar dot product function is applied to the query, key, and value, defined as follows:

$$\text{Attention}(Q, K, V) = V \cdot \left(\text{softmax} \left(\frac{K^T Q}{\sqrt{D_{\text{spe}}}} \right) \right). \tag{11}$$

The $Attention(\cdot)$represents the scalar dot product computation function, where the multihead attention mechanism is used as in VIT to maximize the ability to extract features. However, unlike the VIT network, we assign Q, K, and V to M^2 heads in the spatial domain and set M to 2 for all datasets. The spectral multihead self-attention (SpeMSA) is defined as follows:

$$\begin{aligned} \text{Head}_n &= \text{Attention}\left(Q_n, K_n, V_n\right), N = 1 \cdots M^2 \\ \text{SpeMSA}\left(Q_n, K_n, V_n\right) &= \text{Concate}_{n=1}^{M^2}\left(\text{Head}_n\right) W. \end{aligned} \tag{12}$$

Q_n, K_n, and V_n denote the query matrix, key matrix, and value matrix obtained by training linear transformation at the nth. $SpeMSA(\cdot)$ denotes the spectral multihead self-attention computation function, $Head_n$ denotes the nth head, and $W \in \Bbbk^{D_{spe} \times D_{spe}}$ denotes the transformation matrix obtained by learning the parameters.

D. Multiscale Embeddings

1) Multiscale Band Embedding: For more accurate spectral information extraction from the image spectrum, we propose the SPET module. We introduce the multiscale spectral embedding for multiscale spectral feature extraction. First, the convolution process is carried out using a kernel of dimension 3×3 to the different scales images with the output channels of C. Then, we input these feature maps into the respective SPET to capture multiscale spectral details. We aggregate the extracted multiscale spectral features and use the skip connection operation to acquire the eventual output $F_{\text{spe}}^{\text{Sum}}$, which can be expressed using the following mathematical expression:

$$F_{\text{spe}}^{\text{Sum}} = \text{Add}(\Sigma_{l=1}^{L=3}\text{SPET}, O_a). \tag{13}$$

The $Add(\cdot)$ represents the element-wise summation operation.

2) Multiscale Patch Embedding: We use the SPAT module to capture the spatial characteristics embedded in the image, The SPAT structure is similar to the VIT structure, the image is cut into fixed-size patches and embedded linearly into the sequence for processing. The embedding of patches at different scales obtain feature information at different fine level, so we use multiscale patches to further enrich the spatial details of the image. As depicted in the bottom of Fig. 2, we first split the image into patches with different sizes, then input these different patches into different SPAT as the embedding sequences to further extract the image's spatial characteristics across multiple scales, combining the skip connection and aggregation operations to acquire the ultimate output result $F_{\text{spa}}^{\text{Sum}}$, which can be formally expressed by the equation

$$F_{\text{spa}}^{\text{Sum}} = Add(\Sigma_{l=1}^{L=3} SPAT, O_a) \tag{14}$$

E. Residual Local Feature Module

As depicted in Fig. 5, the RLFM consists of residual dense local feature block (RDLFB) and enhanced spatial attention (ESA), which significantly reduces the computation time while maintaining the model capacity. Among them, the RDLFB mainly consists of several successive stacked convolutional operations and activation functions as deep local feature extraction, and the skip connection are employed to propagate over multiple layers while mitigating the fading problem, the complete procedure can be mathematically summarized using the following equation:

$$\begin{aligned} F_i &= \text{Concate}\left(\text{ReLU}\left(\text{Conv}\left(x_i\right)\right), x_i\right) \\ F_{\text{out}} &= \text{Add}\left(\Sigma_{l=1}^{L=6}\text{Concate}\left(\text{ReLU}\left(\text{Conv}\left(x_L\right)\right), x_L\right), x_1\right) \end{aligned} \tag{15}$$

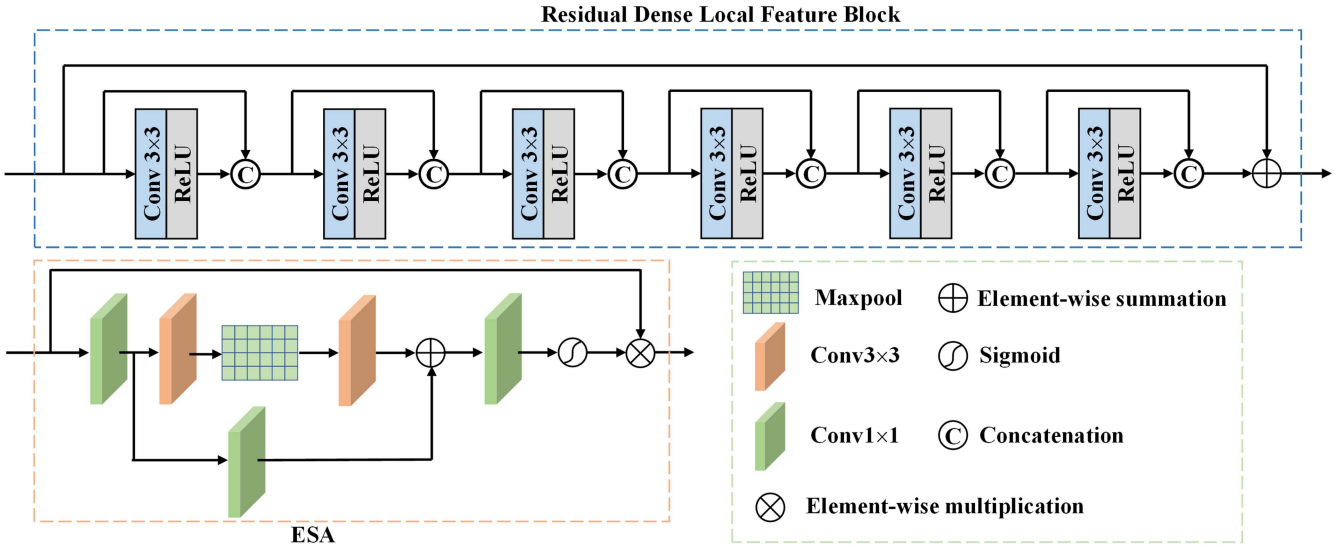


Fig. 5. Architecture of residual dense local feature block (top) and the enhanced spatial attention (down).

where x_i denotes the input at different depths, F_i denotes the output at each stage, $Conv$ denotes the convolution operation with kernel size of 3×3, and the final output F_{out} is obtained through the successive level aggregation and skip connection operation.

The obtained F_{out} is used as the input to the ESA module for further processing to obtain the final HRMS image, as shown in the bottom of Fig. 3. Through convolution operation with a kernel size of 1×1, we downscaled the input feature F_{out} and yielded F_{Conv1}, and in order to reduce the spatial size of the feature map to obtain the sensory field of a larger spatial range, the feature map is sequentially downscaled by the convolution and pooling operations. After the convolution and pooling operations, the output is generated by performing element-wise multiplication between the feature map and the input, and the whole process is expressed as the following equation:

$$\begin{aligned} F_{\text{Conv1}} &= \text{Conv}_{1\times1}\left(F_{\text{out}}\right) \\ F_{\Delta} &= \text{Conv}_{3\times3}\left(\text{MP}\left(\text{Conv}_{3\times3}\left(F_{\text{out}}\right)\right)\right) \\ F_{\Lambda} &= Sigmoid\left(Conv_{1\times1}\left(Add\left(Conv_{1\times1}\left(F_{Conv1}\right), F_{\Delta}\right)\right)\right) \\ F_{\text{final}} &= F_{\text{out}} \otimes F_{\Lambda} \end{aligned} \tag{16}$$

where $Conv_{1\times1}(\cdot)$ and $Conv_{3\times3}(\cdot)$ denotes convolution operations with kernel sizes of 1×1 and 3×3, respectively, $MP(\cdot)$ denotes Maxpool pooling operation, $Add(\cdot)$ denotes an element summation operation, $Sigmoid(\cdot)$ denotes Sigmoid activation function, and $\otimes$ denotes an inner product operation.

F. Loss Function

By calculating the loss between the fused image and the real image in the DMFN network for propagation to continuously optimize the network performance. The MSE loss function is robust for small error values and its squared curve is more smoother in the vicinity, which will not have a significant impact like the L1 [54] loss function. Resulting in the direction of the training model gradually deviates from the target, so we adopts the MSE loss function. In addition, thanks to the Laplace loss function [55], It provides enhanced robustness against abnormal value compared to the MSE loss function, so that the fine-grained features and spectral characteristics of the predicted images closely resemble those of the target image. we combine the two different loss functions in this article to train our DMFN network, the formula is specified as follows:

$$l_{\text{final}} = \text{MSE}(x, y) + \gamma L_{ap}(x, y). \tag{17}$$

x denotes the fused image and y denotes the truth ground. The MSE function is defined as follows:

$$L_{\text{MSE}} = \frac{1}{n}\Sigma_{i=1}^{i=n}\left(y_i - \widetilde{y_i}\right)^2. \tag{18}$$

n represents the count of observations, $(\cdot)^2$ denotes the square of the error between the predicted image and the real image, y_i represents the predicted image, $\widetilde{y_i}$ denotes the true image, and the squared difference of all the samples is summed and averaged to obtain the final loss, and the Laplace loss function is defined as follows:

$$L_{ap} = \Sigma_{j=1}^{n} 2^{j-1} \mid L^j(\beta) - L^j\left(\widetilde{\beta}\right) \mid_1 \tag{19}$$

where $L^j(\widetilde{\beta})$ denotes the result of the jth layer in the slap loss function, and $L^j(\beta)$ denotes the result of the real image in the jth layer, the computed multilayer results are summed.

IV. Experimental Results

The DMFN network is trained and validated on two benchmark datasets. First, we will explain in detail about the benchmark datasets and evaluation metrics used in our experiments, demonstrate the enhancement of the overall network performance brought about by the modules introduced in our experiments through the ablation experiments. Finally, we conduct

a comprehensive analysis, encompassing both qualitative and quantitative aspects, as we compare our method with nine SOTA pan-sharpening approaches with and without reference images, those experiments were carried out using NVIDIA Titan RTX and used pytorch framework.

A. Datasets and Metrics

Experiments are executed on both the QB and WV2 datasets, where the QB dataset includes four bands, and the spatial resolutions provided by the MS image and the PAN image are 2.44 and 0.61 m, respectively. The WV2 dataset includes eight bands, but our experiments only consider four of them, the red, blue, green, and NIR spectral bands. When it comes to spatial resolution, the MS image provides a coarser 4 m, while the PAN image provides a finer 1 m resolution. The DMFN network is trained according to the Wald protocol, the training rounds are 100. The Adam optimizer is used and the learning rate is adjusted according to training rounds to maintain the stability of the training process with the initial value of 0.0001. The final results are evaluated by the qualitative and quantitative analyses. Qualitative analysis is to zoom in the regional features of the fused images through visual operations and calculate the residual images with the real images, which observe the different performance of the methods more intuitively. Quantitative analysis is used to judge the differences between the pan-sharpening methods by comparing several common evaluation metrics, including the reference metrics Universal Image Quality Index (UIQI) [56], Spatial Correlation Coefficient (SCC) [56], Spectral Angle Mapper (SAM) [56], Erreur Relative Globale Adimen-sionnelle de Synthese (ERGAS) [57], No- Reference Evaluation Metrics including Composite Evaluation Index (QNR) [58], Spatial Distortion Index(D_s) [58], Spectral Distortion Index(D_λ) [58].

1) *ERGAS:* ERGAS is used to assess the fusion quality of remote sensing images, taking full account of the importance of each frequency band and the influence of spatial resolution, the formula for calculating the ERGAS metric are as follows:

$$\mathrm{ERGAS}(F,G) = 100 \cdot \frac{Re_{\mathrm{pan}}}{Re_{\mathrm{ms}}} \sqrt{\frac{1}{L}\Sigma_{i=1}^{L} \frac{\mathrm{RMSE}^2\left(F_b, G_b\right)}{M\left(F_b^2\right)}} \tag{20}$$

where F and G represents the fused image and real image, respectively, Re_{pan} and Re_{ms} refers the spatial resolution for both the PAN and MS images, F_b denotes the constituent bands of the fused image, G_b represents the constituent bands of the real image, $M(F_b^2)$ indicates the mean bands of the fused image, $RMSE^2(F_b, G_b)$ denotes the mean square of each band in the fused image and real image. Smaller ERGAS values correspond to higher quality in the fused image, and the best value is 0.

2) *SAM:* The SAM are used to compare the angular difference between two spectral vectors to measure the spectral similarity between pixels or regions. A reduction in the angle corresponds to an escalation in spectral similarity, and the best value is 0. The calculation formula described as follows:

$$\mathrm{SAM} = \arccos\left(\frac{\alpha \cdot \beta}{\| \alpha \| \| \beta \|}\right) \tag{21}$$

where α and β denote the corresponding spectral vectors between the real image and the fused image, respectively, $(\cdot)$ denote the inner product operation of the vectors, $\| \cdot \|$ denote the modulus of the vectors. As the SAM value approaches 0, it indicates a higher degree of spectral fidelity in the fused image.

3) *Q:* The Q is used to compare the similarity of structural and spectral information between the fused image and the real image.A higher Q value signifies an enhanced resemblance between the fused image and the real image, and the best value is 1. Its calculation formula is expressed by the following equation:

$$Q = \frac{4 \cdot \mu_{xy}}{\mu_x \mu_y} \cdot \frac{\eta_x \eta_y}{\eta_x^2 + \eta_y^2} \cdot \frac{\mu_x \mu_y}{\mu_x^2 + \mu_y^2} \tag{22}$$

where μ_*, η_*, and μ_{xy} denotes the standard deviation, mean deviation, and covariance between the fused image and the real image.

4) *SCC:* The primary purpose of the SCC is to assess the spatial similarity between the fused image and the original image. When the SCC value approaches 1, it indicates a higher degree of spatial structural richness in the fused image.

5) *QNR, D_S and D_λ:* These three metrics are mainly used for no-reference measurements, in which QNR mainly evaluates the clarity and noise level of the image by quantifying the signal-to-noise ratio in the image, D_λ represents the spectral distortion of the image, and D_s represents the spatial information distortion rate of the image. The QNR is calculated based on D_λ and D_s, defined as follows:

$$\mathrm{QNR} = \left(1 - D_\lambda\right)^\gamma \left(1 - D_s\right)^\delta \tag{23}$$

The fused image's spatial-spectral information becomes more abundant as QNR gets closer to 1, and the ideal value of D_s and D_λ is 0.

B. Ablation Experiment

In this section, to demonstrate that our introduced modules bring performance enhancement to our proposed DMFN network, we conduct ablation experiments of AFEB and enhanced spatial attention (ESA) modules on four different network on benchmark datasets QB and WV2. The ablation experiments are conducted on four different network. We find the best network through qualitative and quantitative comparative analyses.

1) *DMFN(orin):* This network serves as the initial network structure for the experiments, the ordinary convolutional operations are used to replace the AFEB and ESA modules, which demonstrate the performance enhancement brought by the introduced modules to the network.

2) *DMFN(AFEB):* This network only introduces the AFEB module to process the feature maps based on the initial network, which highlights the performance enhancement brought by the module to the network.

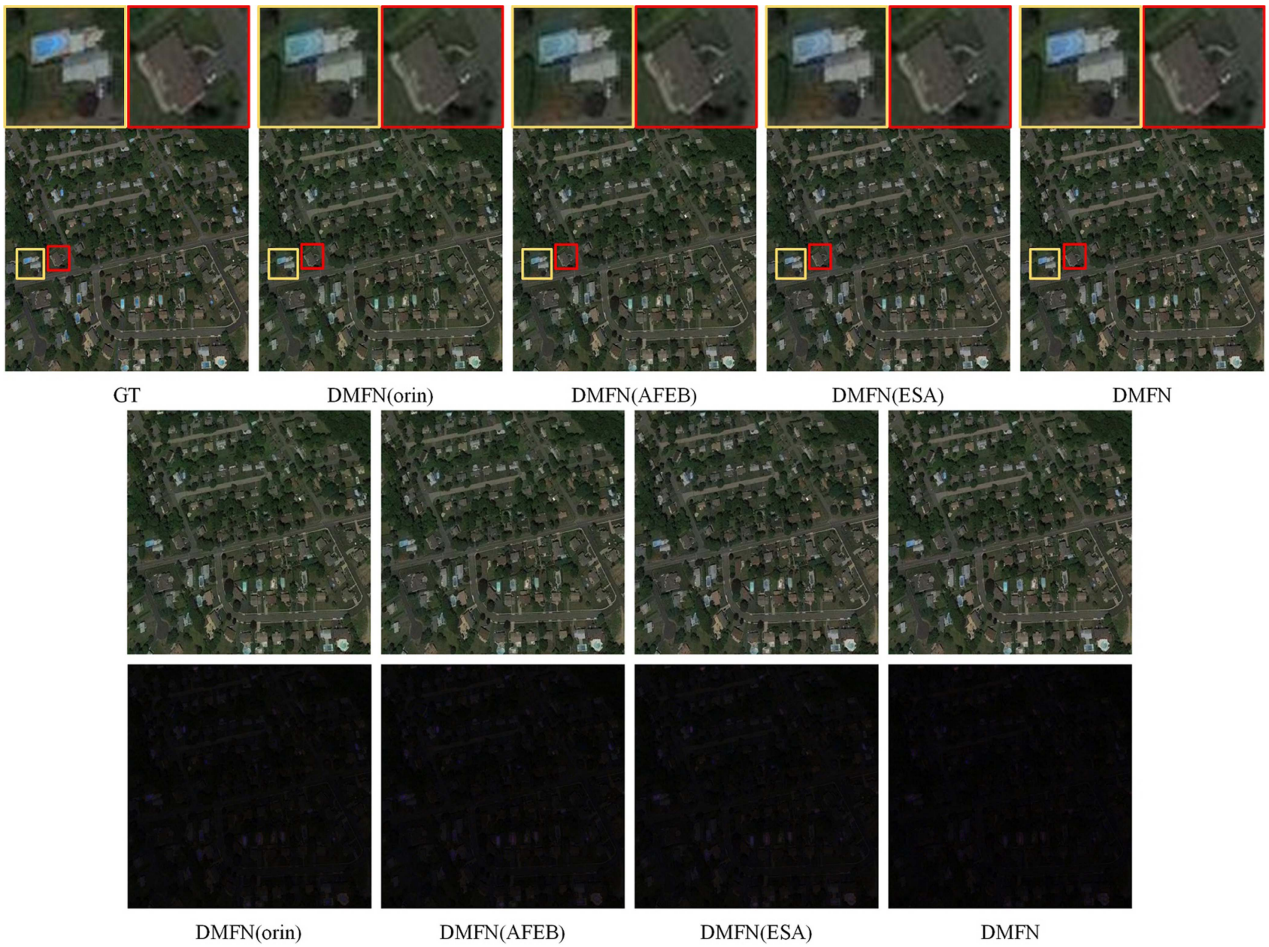


Fig. 6. Ablation experiments were conducted on QB dataset using the four network. The first line shows the enlarge detail of the red rectangle and yellow rectangle, and the second line shows the pansharpening fusion images of the four models,the third line shows the residual images of the four pansharpening results with ground truth image.

3) *DMFN(ESA):* In contrast to the above network model, we only introduce the ESA module based on the initial network model to further prove the capability of the module.
4) *DMFN:* As the network introduced in this article, the DMFN network structure introduced the AFEB module in the feature extraction and image fusion stages to use the attention mechanism on the input images of different scales, which emphasize the crucial aspects of the image and mitigate irrelevant details. The ESA module is also introduced in the image reconstruction process, which enhances the extraction of image features with little increase in the capacity of the model, thus generating the final fused image.

We performed ablation experiments on the QB dataset with the four network and obtained the fusion images and the residual images with the real images as shown in Fig. 6, where the red rectangles and the yellow rectangles are the results of zooming in on some areas in the image.

The yellow rectangles include roofs, pools, and forest areas, while the red areas contain roofs, cars, and forest parts. The spectral distortion of the image generated by DMFN(orin) is more serious as can be seen in the image. Zooming in the yellow and red rectangles also reveals that the pools and roofs are dark and the spatial structure is incomplete. Zooming in the residual images, we can discern that the residuals are most obvious in DMFN(orin). In contrast, the quality of the fused images generated by DMFN(AFEB) and DMFN(ESA) is evidently improved. By zooming in the yellow rectangle in DMFN(AFEB), we can see that the color of the pool and the roof are significantly improved compared with DMFN(orin). Zooming in the red rectangle in the figure, we can see that the spatial structure of the roof is richer, and the details of the contours have been supplemented. The fused image of DMFN(ESA) is also obviously clearer than DMFN(orin), zooming in the yellow image we can see that the spectral distribution of the pool is more uniform than DMFN(AFEB), and the spatial distribution of the roof is obviously richer than that of DMFN(orin) from the zoomed-in red rectangle. Zooming in the residual images, we can see that the spatial structure information of the fused image is complete and the spectral distribution is uniform, which is closer to the real image. Zooming in the yellow rectangle and the red rectangle in the figure, we can also see that the color distribution of the pool and the roof has been greatly improved. And it is clearer and more complete than the fused image of DMFN(AFEB) and DMFN(ESA) network, better than the other network models mentioned above.

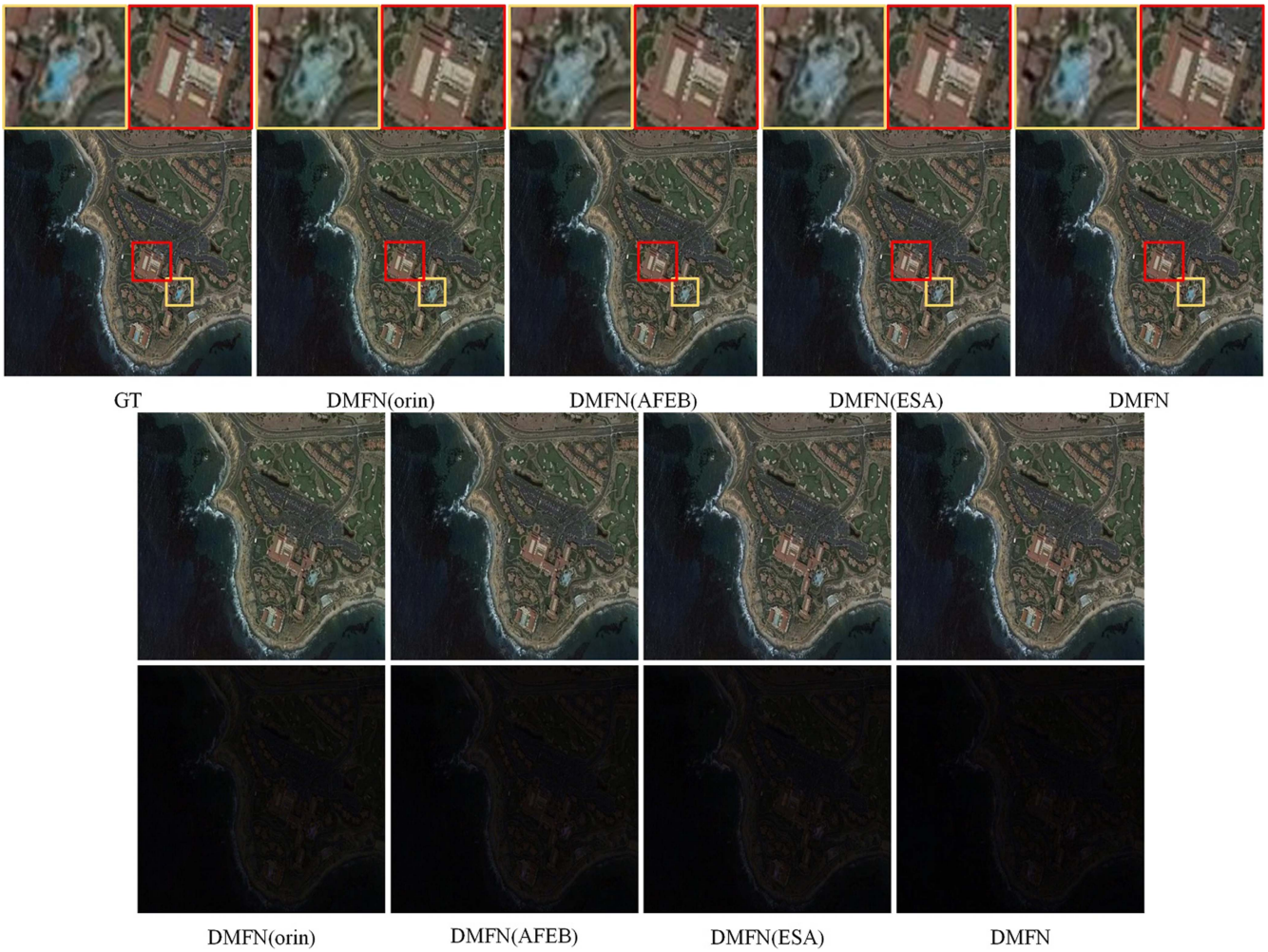


Fig. 7. Ablation experiments were conducted on WV2 dataset using the four network. The first line shows the enlarge detail of the red rectangle and yellow rectangle, and the second line shows the pansharpening fusion images of the four models,the third line shows the residual images of the four pansharpening results with ground truth image.

We conducted ablation experiments on the above four network on the WV2 dataset and obtained the fusion images and the residual images with the real images as shown in Fig. 7, the red rectangles and the yellow rectangles are the results of zooming in some areas in the images.

The yellow rectangle includes rivers and buildings, and the red rectangle mainly includes house roofs. The zoomed-in residual image show that the DMFN(orin) fusion image has the most serious problems of spatial information loss and spectral distortion. The yellow and red rectangles also show that the river and the roofs of the buildings have serious color distortion, the whole image is dark and the spatial information is incomplete, it is almost impossible to see the roofs clearly after the zoomed-in operation. In contrast, DMFN(AFEB) and DMFN(ESA) have improved the ability of the network to extract the spatial-spectral information of the image in a certain extent, and it is evident that the spatial structure and spectral information of the fused images of these two models are richer than the DMFN(orin) from the enlarged residuals. The spectral distributions of the river area are more uniform from the enlarged yellow picture, the enlarged red rectangle also reveals that the DMFN(ESA) fusion image is richer in spatial structure information than DMFN(AFEB), and the details of the contours of the red roofs are better complemented with a uniform spectral distribution. Compared with the above model, the fused image produced by our innovative DMFN algorithm closely approximates the real image. A deeper analysis of the residual image under magnification reveals that our method not only maintains comprehensive spatial details, but also achieves a more even distribution of spectral information. From the enlarged yellow rectangle, we can see that the color distribution of the river area is greatly improved, and the spatial details of the roofs are also significantly improved. From the enlarged red rectangle, we can also find that both the spatial and spectral distributions of the roofs are enriched, the details of the roofs' contours are clearly visible, which is better than the other network structures mentioned above.

We carry out the quantitative analysis of the ablation experiments on the benchmark datasets QB and WV2 and the comparative results are presented in Tables I and II, in which the best values of the SCC and Q are 1, and the best values of the SAM and ERGAS are 0. It can be seen that all the metrics of DMFN(AFEB) and DMFN(ESA) are superior than

TABLE I
QUANTITATIVE COMPARISON OF FOUR METHODS ON THE QB DATASET

Method	QB			
	SAM(0)	ERGAS(0)	SCC (1)	Q (1)
DMFN(orin)	0.037	1.1341	0.9945	0.9626
DMFN(AFEB)	0.0334	1.0485	0.995	0.965
DMFN(ESA)	0.0335	1.0423	0.9952	0.9648
DMFN	**0.0304**	**0.9733**	**0.9958**	**0.9672**

We bolded the best result.

TABLE II
QUANTITATIVE COMPARISON OF FOUR METHODS ON THE WV2 DATASET

Method	WV2			
	SAM(0)	ERGAS(0)	SCC (1)	Q (1)
DMFN(orin)	0.0358	1.1315	0.995	0.9679
DMFN(BFE)	0.0314	1.0251	0.9952	0.9697
DMFN(BFE+BRB)	0.0316	1.0205	0.9951	0.9694
DMFN	**0.0285**	**0.9509**	**0.9957**	**0.972**

We bolded the best result.

TABLE III
COMPARASION OF PARAMETERS AND FLOATING-POINT OPERATIONS OF DL MODEL

Models	PanNet	MSDCNN	GPPNN	LGT	MSIT	Ours
Params	0.069	0.229	0.120	0.174	295.250	13.519
FLOPS	8.287	13.980	10.266	18.5471	133.641	109.56

DMFN(orin), which indicates AFEB and ESA modules can provide improvement to the performance of the model. Boosting the model's aptitude for extracting spatial structure and spectral distribution from the image. Our DMFN method achieves the best results in all the metrics, which indicates that our method has the strongest ability to extract spatial details and recover spectral distribution, the fused image is closer to the real image.

C. Experiment and Evaluations

We will perform a qualitative and quantitative assessment of the method presented in this article, contrasting it with nine current SOTA techniques using both simulated and real data, which are three classical methods: IHS, GSA, MTF_GLP_HPM, and five DL-based methods: PanNet, MSDCNN, GDD [39], MSIT [52], LGT [53], GPPNN. The datasets and codes of the abovementioned methods are all open source projects, the parameter values are established following the recommendations detailed in this article.

To comprehensively analyze the performance of each network from different perspectives, we also conducted a comparative analysis of the complexity (Params) and computational costs (FLOPS) of each network structure. The Table III displays the numerical values associated with the various DL networks used in this article. It is evident that our proposed network model does not exhibit an advantage in terms of parameter count and computational cost. However, as demonstrated in the subsequent comparative analysis experiments, this network has achieved a balance between performance and efficiency.

1) Simulated Data Experiment: The above nine pan-sharpening methods are experimented on the benchmark datasets QB and WV2 following the Wald protocol, and the HRMS images are used as our real reference images for comparison and analysis.

Fig. 8 illustrates the fusion images of all the above pan-sharpening methods on the simulated dataset QB and the zoomed-in images of some regions in the figure, the main content of the figure includes the roof, container, road, and other regions, the yellow rectangle is the zoomed-in area of the container in the figure, and the red rectangle is the zoomed-in area of the roof and the road in the figure. The GDD and MSIT methods exhibit the most severe spectral distortion, resulting in an overall image over-brightness. Upon closer examination of the magnified yellow rectangle in the image, we can observe a significant loss of color in the containers and a pronounced blurring of their overall outlines, making it challenging to discern finer details. Similarly, when we focus on the magnified red rectangle, a substantial disruption in the spatial structure of the rooftops becomes evident, with nearly all edge details being heavily blurred, leading to a significant loss of spectral information. In contrast, the ability of IHS, PanNet, MTF_GLP_HPM, and GSA methods to extract the information embedded in images is greatly improved, but from the magnified regional images, there are still exists spectral distortion and spatial distortion phenomenon, in which the yellow rectangle of the container area of the color distribution and spatial information of the IHS, PanNet, and GSA are better than MTF_GLP_HPM. Observing the zoomed-in red rectangle, it is also found that the roof color remains severely skewed, the specific spatial contour details are blurred, and the spatial characteristics have been seriously impaired. The ability of MSDCNN, LGT, and GPPNN methods to extract spectral information is significantly better than any of the above pan-sharpening methods, from the zoomed-in yellow rectangle, the spectral information of the container is obviously recovered, and the spatial structure information is greatly enriched, so that the details of the container's contour can be observed. From the enlarged red rectangle, we can also find that a noticeable enhancement in the roof's color is apparent when contrasted with the method mentioned above, and the spatial profile information of the highway is further expanded. The comparative analysis indicates that the GPPNN pan-sharpening method outperformes in capturing both spatial and spectral information when compared to the MSDCNN and LGT methods. In contrast, our proposed DMFN method outperforms all the previously mentioned pansharpening methods in extracting both spatial and spectral information. The magnified images vividly depict the spatial distribution details of containers and roof while retaining the rich spectral information contained in the MS image, closer to GT images. It achieves optimal performance when compared to the aforementioned methods.

Fig. 9 illustrates the residual images with the real image, it is evident that the residual values of MTF_GLP_HPM, GSA, and IHS methods are the most obvious, which indicates that these pan-sharpening methods have more serious spectral distortion problems and lose more spatial information. In contrast,

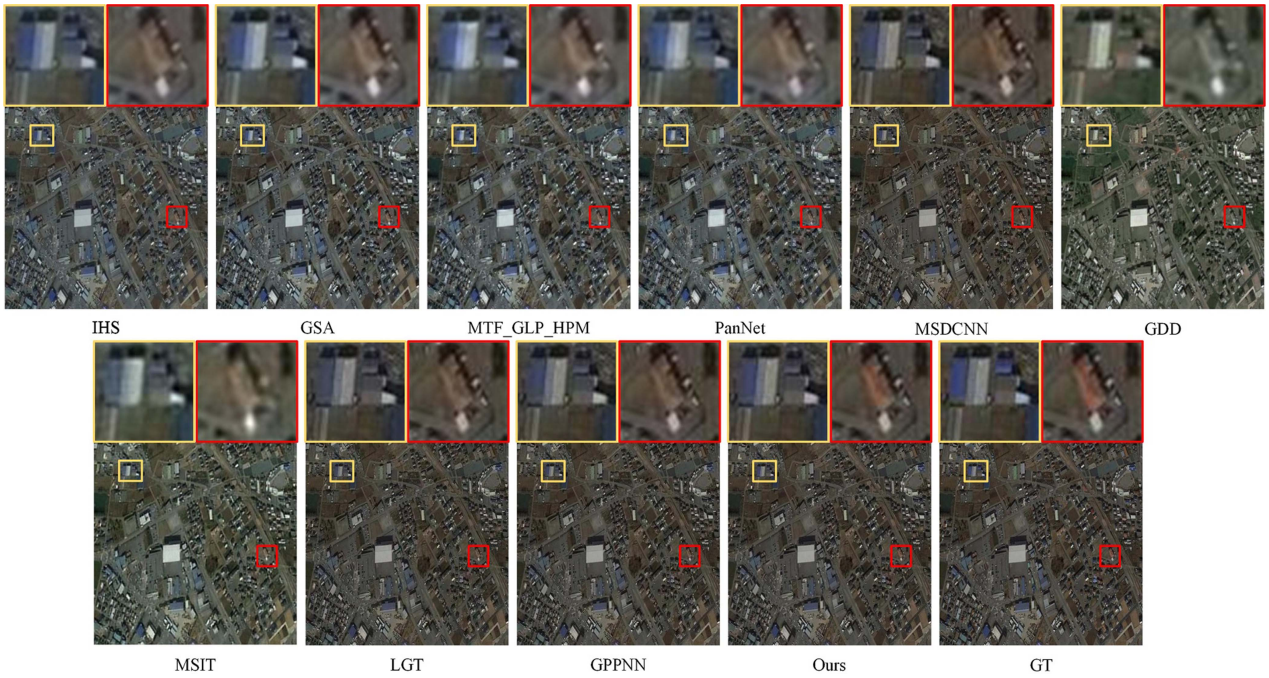


Fig. 8. Fusion images of nine pansharpening methods on the QB simulation dataset. The yellow rectangle and red rectangle are the detailed enlarged images,you can zoom in to visualize the image details.

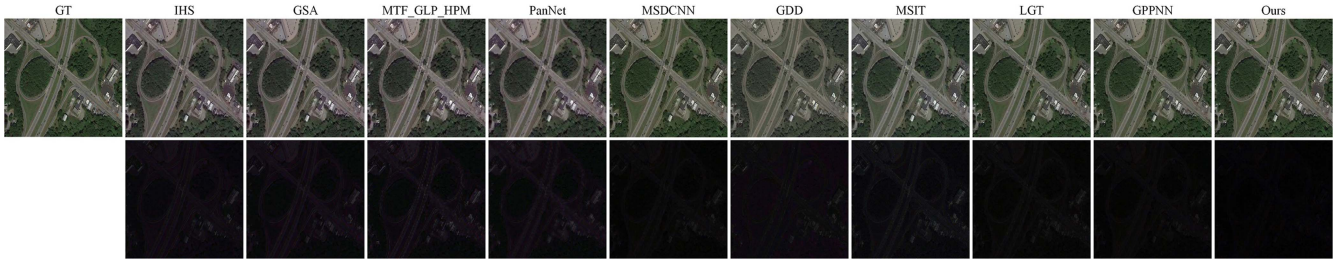


Fig. 9. Fusion results of nine pan-sharpening methods on the QB simulation dataset and the residual images with ground truth images. The images can be zoomed in to view the details.

the residual values of PanNet, GDD, and MSIT methods are significantly reduced, the PanNet method exhibits the most obvious residual values, indicating more severe spectral distortion and significant disruption of the spatial structure compared to the other two methods. While the GDD method keeps more spatial information, which explains the less apparent residual values. The MSDCNN, LGT, and GPPNN methods capture the spatial intricacies and spectral information of the image more deeply, the zoomed-in residual image shows a few obvious white contours, the rest of the image all black background. The residual value of GPPNN is the smallest, which indicates that the GPPNN method extracts the relevant spatial structure details of the image as completely as possible and maintains the spectral information, performing the best among the three methods. Zooming in the residual image generated by our proposed DMFN method, we will find that the whole background is almost black without any white contour, which indicates that our method proves proficient in the preservation of spectral distribution and the extraction of spatial details to the maximum extent.

Fig. 10 exhibits the fused images on the simulated dataset WV2 compared with the other pan-sharpening techniques as well as the zoomed-in images of some areas in the images. The main contents of the figure are wheat fields, roofs, and roads. The yellow rectangle are the zoomed-in areas of roofs and the red rectangle are the zoomed-in areas of wheat fields. The fusion images of GDD and MSIT fusion image has serious spectral distortion problem, and the spatial structure suffer damaged. In addition, observing the enlarged yellow rectangle, it is apparent that the roof's color experiences significant distortion, and there is partial blurring in its spatial structure, which hinders the clear visibility of specific structural details on the roof. In the enlarged red rectangle, the color of the wheat field is heavily distorted, with an overall brightening and blurring of the specific contours. In contrast, the IHS, GSA, MTF_GLP_HPM, and PanNet methods are more capable of extracting the spatial structure and spectral distribution of the image. Observing the enlarged yellow rectangle reveals that the spatial distribution of roof details is rich, and the spectral information distribution is more uniform than the other three methods, shows a better

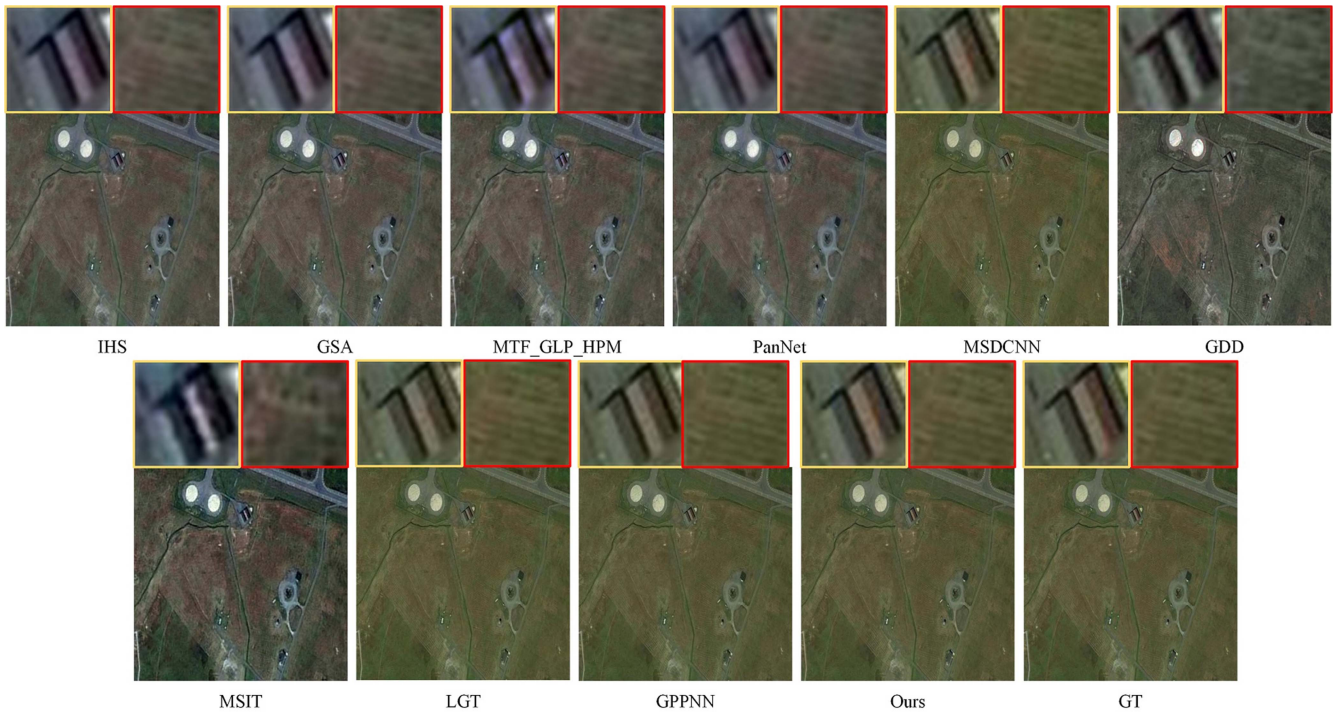


Fig. 10. Fusion images of nine pansharpening methods on the WV2 simulation dataset. The yellow rectangle and red rectangle are the detailed enlarged images,you can zoom in to view image details.

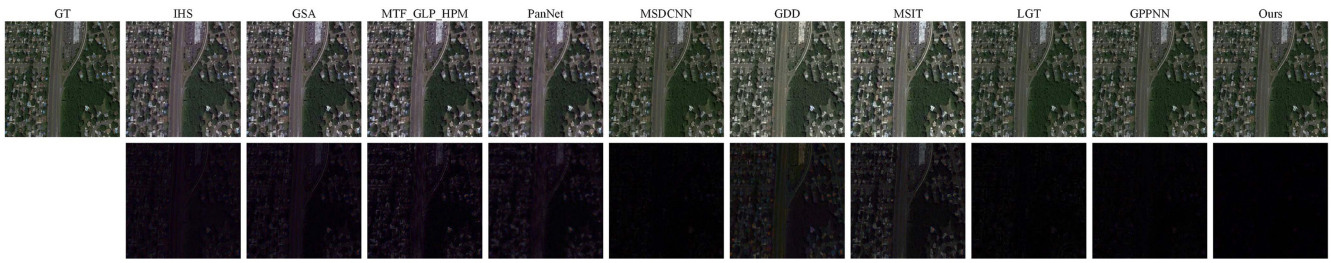


Fig. 11. Fusion results of nine pan-sharpening methods on the WV2 simulation dataset and the residual images with ground truth images. The images can be zoomed in to see the details.

performance. The extraction of the spectral distribution information by the MSDCNN, LGT, and GPPNN methods is obviously stronger than the abovementioned methods, the overall image color tends to be more similar to the real image. What can be seen in the zoomed-in yellow rectangle is that the spatial details of the roof are more complete, the contours are clearly visible and the spectral distribution is uniform. From the enlarged red rectangle, we can see that the color distribution of the wheat field is uniform, and the spatial structure is clear and complete. From the enlarged image, we can also found that the color distribution of the roof and the wheat field of MSDCNN is more uniform and richer than the other two methods, which indicates that the performance of MSDCNN is better than the LGT and GPPNN methods. Our proposed DMFN method has the best performance in extracting spatial detail information and spectral distribution information. The zoomed-in yellow rectangle shows that the spatial detail structure of the roof is basically complete and the spectral distribution is more uniform. The zoomed-in red rectangle also shows that the wheat field is full of hues, the spectral distribution is uniform and the specific details of the distribution can be seen clearly. The ability to extract the spatial structure and spectral distribution information is better than any of the above pan-sharpening methods.

Fig. 11 shows the final fused image and the residual image with the real image. Compared with other pan-sharpening methods, the residual value of GDD and MSIT pan-sharpening methods is the most obvious. The color of the fused image is bright, which reveals that the loss of spatial structure is notably obvious compared with the other methods. The whole road and house contours are clearly visible in the image, means that there is still a large loss of value compared with the real image. Compared with MTF_GLP_HPM and PanNet methods, although the IHS and GSA fused images are darker with fewer regions of residual values. However, when zooming in on certain areas within the residual images, it becomes evident that there is a significant loss of spatial structural information in the images. The fusion images from MSDCNN, LGT, and GPPNN methods closely

TABLE IV
QUANTITATIVE COMPARISON ON THE QB DATASET

Method	QB			
	Q ↑	SCC ↑	ERGAS ↓	SAM ↓
IHS	0.7865	0.9389	4.5934	0.1826
GSA	0.8632	0.9666	5.2084	0.0724
MTF_GLP_HPM	0.8204	0.9567	5.548	0.0718
PanNet	0.815	0.9531	5.4895	0.0745
GDD	0.8185	0.9679	7.8380	0.07793
MSDCNN	0.9555	0.9921	1.3268	0.0332
MSIT	0.7369	0.9576	6.7982	0.0355
LGT	0.9481	0.9937	1.1671	0.0355
GPPNN	0.9639	0.9950	1.0422	0.0336
Ours	**0.9671**	**0.9959**	**0.9733**	**0.0304**

We bolded the best result and Underlined the second value.

TABLE V
QUANTITATIVE COMPARISON ON THE WV2 DATASET

Method	WV2			
	Q ↑	SCC ↑	ERGAS	SAM ↓
IHS	0.8928	0.9325	5.3827	0.1546
GSA	0.8377	0.9632	6.1055	0.0885
MTF_GLP_HPM	0.7918	0.9531	6.4917	0.0888
PanNet	0.7824	0.9479	6.372	0.0902
GDD	0.7923	0.9644	9.4575	0.0955
MSDCNN	0.9578	0.9920	1.3004	0.0317
MSIT	0.6864	0.9569	8.7491	0.0748
LGT	0.9520	0.9932	1.1646	0.0339
GPPNN	0.9681	0.9951	1.0035	0.0311
Ours	**0.9721**	**0.9957**	**0.7909**	**0.0285**

We bolded the best result and Underlined the second value.

resemble the real image. From the residual images, it is evident that except for a few white regions, the rest of the background appears black. This indicates a uniform spectral distribution and rich spatial details in the images, surpassing the other mentioned pansharpening methods. The residual values in the GPPNN method are notably smaller than those in the MSDCNN and LGT methods, implying that the GPPNN method excels in capturing both spatial and spectral information from the images compared to the other two methods. In contrast, when observing the fused image and the residual map of our method, it is found that the residual map's background all black. The zoomed-in residual image contains almost no residual values, and the fused image closely approximates the real image, which indicates that the loss of this method in extracting the spatial structural and the spectral distribution information is almost negligible. Surpasses all the aforementioned pan-sharpening approaches.

Furthermore, we conduct a quantitative comparison and analysis of the fusion outcomes for all the compared methods on two simulated datasets QB and WV2, where we select four metrics as the reference metrics, namely ERGAS, SCC, SAM, and Q. the ideal value for SAM and ERGAS is 0, and the best value for Q and SCC is 1. Table IV shows the results of the experimental comparison with nine pan-sharpening methods on the QB dataset, we bolded the best value and underlined the second value of each metric. It becomes apparent that our method surpasses all others, consistently ranking first in all four metrics. SAM is 0.0028 higher than the second MSDCNN, ERGAS is 0.0689 higher than GPPNN, SCC is 0.0009 higher than GPPNN, and Q is 0.0032 higher than it. In general, the images fused by our our method show the best performance.

The WV2 dataset was used to conduct experiments comparing our method with nine other pan-sharpening techniques, and the results are displayed in Table V. It is apparent that except for the SAM, which is 0.0026 lower than the GPPNN method, our method performer the best in the other metrics. The ERGAS is 0.2126 higher than the GPPNN method, the SCC is 0.0006 higher than the GPPNN method, the Q is 0.004 higher than the GPPNN method. To sum it up, our approach boasts the best performance, indicating that the DMFN network achieves superior results in extracting information in contrast to other methods.

2) Real Data Experiment:: We also conduct qualitative and quantitative comparative experimental analyses of these nine different pan-sharpening methods on the benchmark datasets of QB and WV2. The real QB dataset was used to generate the fusion results depicted in Fig. 12 for various pan-sharpening methods, which includes the areas of rooftops, highways, automobiles forests. The yellow and red rectangles shows the magnified results of the pool and the forest areas. We can see that the GDD and MSIT methods has serious problems of spectral distortion and the loss of spatial information. From the zoomed-in red rectangle, we can also find that the structure of the forest area is seriously twisted and the forest area is seriously blurred, we can hardly see any details of the specific contours. From the zoomed-in yellow rectangle, we find that the pool is seriously distorted and the spatial structure is seriously damaged. In contrast, the GSA and PanNet methods make up for the spatial and spectral loss of the image to a certain extent. From the enlarged yellow rectangle, we can see that the color distribution of the pool has been improved, and the peripheral distribution contours can be seen vaguely. From the enlarged red rectangle, we can find that the spatial details of the forest region have been enriched. The IHS and MTF_GLP_HPM methods obtain the spatial structural and spectral distribution of the image more efficient, as can be seen from the enlarged yellow rectangle, the spectral distribution of these two methods is closer to the MS image with better spatial detail distribution, and the contour detail distribution of the pools is obviously clearer. Observing the enlarged red rectangle also reveals that the IHS method blurs the forest region more than the PAN image, and the spatial detail information of the forest region in the MTF_GLP_HPM method is closer to the PAN image, but the color distribution is obviously darker than the MS image. MSDCNN, LGT, and GPPNN methods further improve the ability of the nerwork to extract the information of the image. From the zoomed-in yellow rectangle, we can find that the spectral distribution of the pool region is greatly improved, its spatial contour details are constantly enriched compared with the PAN image,and the area around the pool can be observed more clearly. From

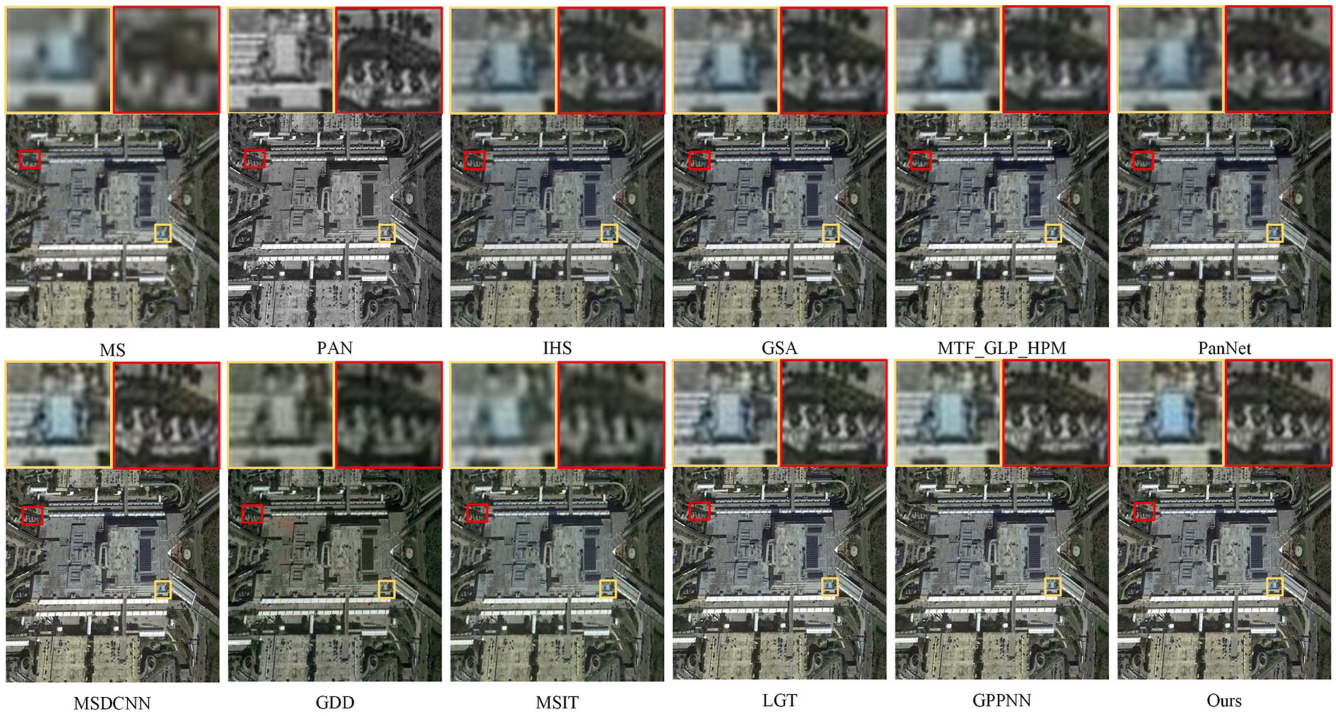


Fig. 12. Fusion images of nine pansharpening methods on the QB real dataset. The yellow rectangle and red rectangle are the detailed enlarged images,you can zoom in to view image details.

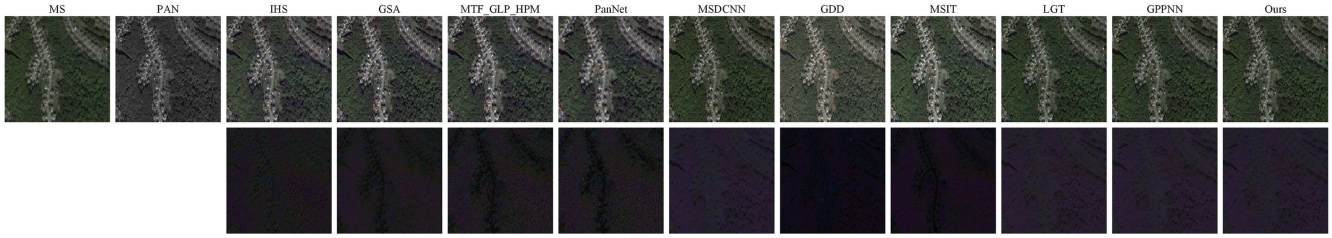


Fig. 13. Fusion results of nine pan-sharpening methods on the QB real dataset and the residual images with PAN images. The images can be zoomed in to view the details.

the enlarged red rectangle, we can also see that the spectral distribution and spatial structure information of the forest region are further expanded. The spatial details of the forest region in the MSDCNN and the GPPNN methodS are richer than the LGT, which is closer to the PAN image. Observing the fused images using our method, we can find that our method effectively obtains the spatial structure contained in PAN image, while recovers the spectral distribution of the MS image to the largest extent. Observing the enlarged yellow and red rectangles, we can find that the spectral distribution of the pool and the forest region is uniform, the spatial details are clear and the contour distribution has been further demonstrated. Our approach outperforms the pan-sharpening methods mentioned above.

Fig. 13 shows the fused images of these nine pan-sharpening methods and the residual images with the real images. It is apparent that the residual value of the GDD, MSIT methods are the most, which indicates that the spatial structure details of the image has been seriously damaged. The loss of spectral information is more serious, and the contours of the highway and the forest area can be clearly seen from the residuals. In contrast, the residual values of GSA, PanNet, MTF_GLP_HPM, and IHS methods are reduced, and the spatial structure of the images is closer to the PAN images. there still exists serious spectral distortion, compared to the MS images, the images manifest a darker color tone. The traces of the road's residual value has been much decreased and the richness of spatial information exceeds that of the methods mentioned above. By enlarging the residual images, we find that the residual values of MSDCNN, LGT, and GPPNN methods are obviously reduced, we can hardly see any obvious residuals in the highway or the forest area, which indicates that the ability of these methods to obtain spatial characteristic information is obviously better than the abovementioned methods. The color distribution shown a significant enhancement in comparison to MS images and the residual images are much more smoother. Correspondingly, the fused image of our method is closer to the real image, the whole residual image is in a smooth state, and the residual values can hardly be seen in the zoomed-in residual image, which indicates

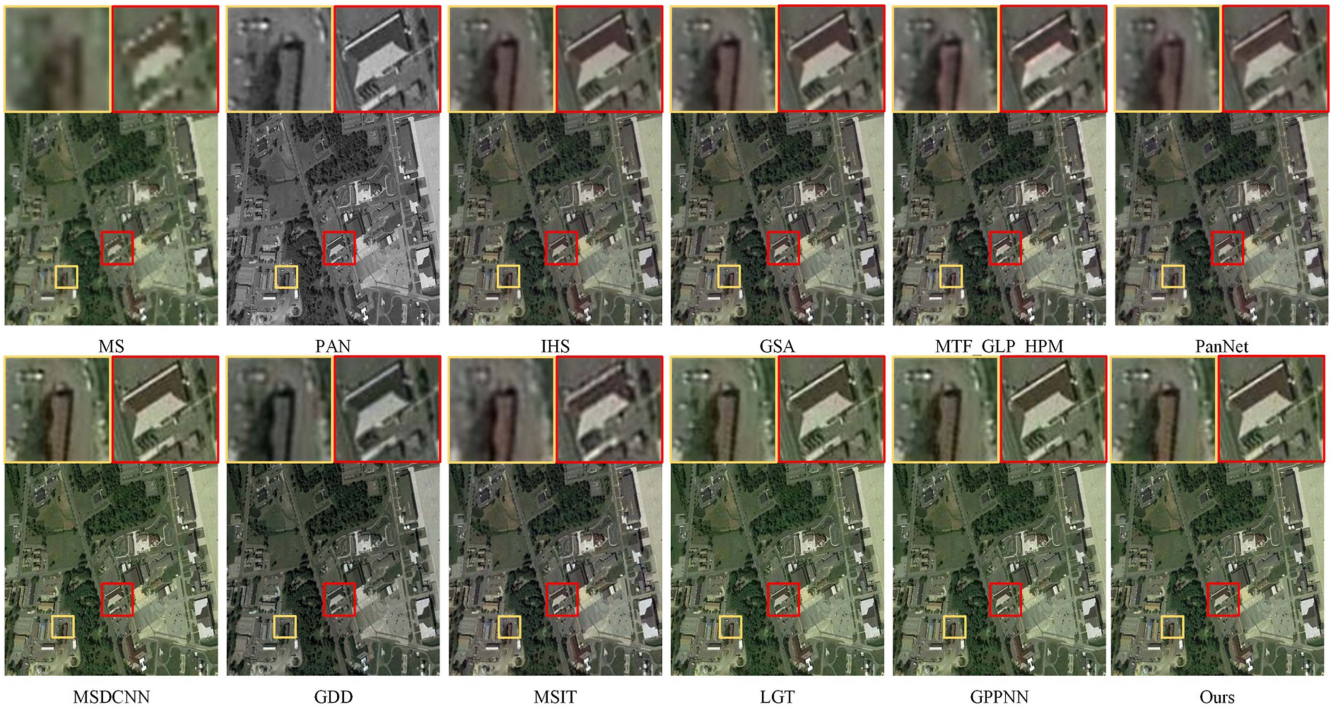


Fig. 14. Fusion images of nine pansharpening methods on the WV2 real dataset. The yellow rectangle and red rectangle are the detailed enlarged images,you can zoom in to see image details.

that our method comprehensively captures the spatial intricacies of the PAN image while also adeptly preserving the rich spectral distribution details of the MS image.

In Fig. 14, the fusion outcomes of nine pan-sharpening techniques applied to the WV2 real dataset are depicted, which contains the roof, forest, and car areas. the yellow and red rectangles are the zoomed-in results in the figure. we can see from the fused image that the GDD and MSIT methods suffer from serious spectral distortion and the loss of spatial information. Zoomed-in the yellow rectangle, we can find that the car and forest areas are seriously blurred, the colors of the cars and forests are pale compared with the MS image, and the spatial structure details inherent in the PAN image is seriously lost. Zooming in the red rectangle reveals that the roof's overall outline is notably well-preserved, but the roof is not as complete as the PAN image. Although the overall outline of the roof is relatively complete, the spatial details are damaged and the blurring is serious. In contrast, the IHS, PanNet, MTF_GLP_HPM, and GSA methods can better extract the spatial structure and spectral information of the image. Observing the enlarged yellow rectangle, we can find that the spectral distribution of the forest and the car is more balanced, the distribution of the specific contours is also clearer. The fused image generated by the MTF_GLP_HPM method is obviously clearer than the other three methods, retains plentiful spatial details in the PAN image better. From the enlarged red rectangle, we can see that the spatial distribution of the roof is effectively preserved, but the spectral information distribution is still insufficient compared with the MS image. We can see that the MSDCNN, LGT, and GPPNN methods have further improved the spatial details extraction and the spectral information recovery. By observing the enlarged yellow rectangle, it is evident that the spectral distribution of the forest and the car area is almost the same as the MS image. At the same time, the rich spatial structural information contained in the PAN image has been retained. From the zoomed-in red rectangle, we can also see that the spatial details of the roof are effectively preserved, and the distribution of the specific contours is almost the same as the PAN image. In contrast, the fused image of our method effectively captures the abundant spatial details embedded in the PAN image and efficiently recovers the rich spectral distribution information contained in the MS image. Observing the zoomed-in yellow and red rectangles, we can find that the spatial structure of the roof and the forest area is effectively preserved with clear contours and uniform color distribution, which is better than any of the above methods.

Fig. 15 shows the fused images of these ten pan-sharpening methods and their residual images with the real images. Notably, the GDD and MSIT methods exhibit the most striking residual artifacts, suggesting a substantial deterioration in spatial structural information and the subsequent degradation of spectral content in the fused images. The overall colors of the fused images are darker than the MS image. In contrast, the residual maps of GSA, IHS, MTF_GLP_HPM, and PanNet are more smoother, and the zoomed-in yellow rectangle show that the spatial information of the highway area is more complete than the above methods, and contains richer spatial information than the MS image. But there are still existed serious spectral distortions, the overall image is darker and the residual values in the residual maps are more obvious, which indicates that these methods lose a lot of details when extracting the spatial structure information. On the other hand, the MSDCNN, LGT, and GPPNN methods have enhanced their ability to extract image spectral

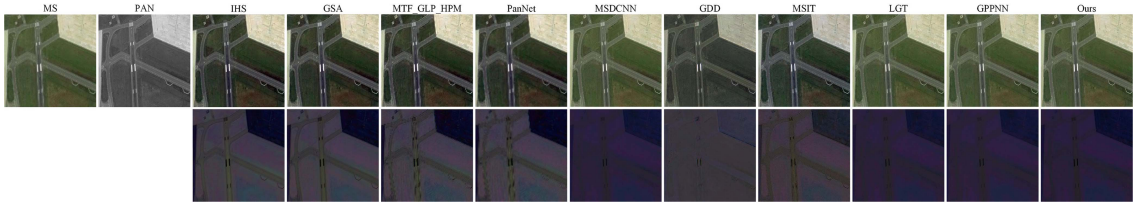


Fig. 15. Fusion results of nine pan-sharpening methods on the WV2 real dataset and the residual images with PAN images. The images can be zoomed in to see the details.

TABLE VI
QUANTITATIVE COMPARISON OF NINE METHODS ON TWO REAL DATASETS QB AND WV2

Method	QB			WV2		
	$D_S \downarrow$	$D_\lambda \downarrow$	QNR ↑	$D_S \downarrow$	$D_\lambda \downarrow$	QNR ↑
IHS	0.0634	0.0623	0.9285	0.1056	0.0367	0.8714
GSA	0.0346	0.0167	0.951	0.0478	0.0326	0.9312
MTF_GLP_HPM	0.0191	0.0251	0.9463	0.0248	0.0219	0.9426
PanNet	0.0214	0.0219	0.9585	0.0301	0.0238	0.9404
GDD	0.0276	0.0390	0.9348	0.0426	0.0539	0.9068
MSDCNN	0.0239	0.0311	0.9458	0.0306	0.0477	0.9236
MSIT	0.0215	0.0454	0.9344	0.0293	0.0643	0.9087
LGT	0.0226	0.0343	0.9440	0.0307	0.0511	0.9201
GPPNN	0.0187	0.0161	0.9591	0.0284	**0.0172**	0.9438
Ours	**0.0156**	**0.0148**	**0.9684**	**0.0195**	0.0212	**0.9512**

We bolded the best result and Underlined the second value.

information. It can also be noticed from the residual images that the residual values for these methods are not very significant, the fused image contains rich spatial details compared with the PAN image, and the specific spectral distribution information is closer to the MS image. The image's spatial and spectral distribution characteristics are fully exploited by our method. From the residual image, it becomes apparent that our method yields results devoid of any residuals. The DMFN method not only preserves the abundant spectral data within the MS image, but also captures the extensive spatial details from the PAN image, which is better than the abovementioned pan-sharpening methods.

We also carry out quantitative and comparative experimental analysis of these nine pan-sharpening techniques on the benchmark datasets QB and WV2, as depicted in Table VI, we use the no-reference metrics D_S, D_λ, and QNR as the comparisons, the optimal values of D_λ and D_S are 0 and the best value of QNR is 1. We bolded the optimal value and underlined the second in the table. We can see that our method achieves best in all metrics in the QB dataset. In the WV2 dataset, except for the D_λ is 0.004 lower than that of GPPNN, the rest metrics performes best. Overall, our method is more proficient than the other nine pan-sharpening methods and effectively obtains both the spatial and spectral details within the image.

In both qualitative and quantitative assessments on simulated and real datasets, our method consistently exhibits superior performance in recovering the spectral distribution and spatial information inherent in MS and PAN images, surpassing other pan-sharpening approaches.

V. CONCLUSION

We present a novel pan-sharpening methodology called DMFN in this articlle, whose full name is transformer-based dual-branch multiscale fusion network for pan-sharpening remote sensing. The network model is separated into three components: feature extraction, image fusion, and image reconstruction. We commence by interconnecting the MS image and PAN image through up-sampling and down-sampling operations within distinct branches, and then after our shallow feature extraction, it is inputted into the AFEB to combine the spatial-channel attention, only focus on the important information of the image. Then, the feature maps of multiple scales in the two branches are input into each (spatial transformer) SPAT and (spectral transformer)SPET for spectral and spatial information extraction, after the Upsample operation on the final output of SPET, it is connected with the final output of SPAT and input into the residual dense local feature block To capture the image's deeper characteristics again, and get our final fusion image after the process of ESA module.To establish the superiority of our method, we begin with ablation experiments on the QB and WV2 benchmark datasets. We then proceed to comprehensive qualitative and quantitative assessments, comparing our method with nine recent pan-sharpening techniques on both simulated and real datasets. The above experiments prove that our DMFN method adeptly captures the abundant spatial details inherent in PAN images, while preserving the consistent spectral characteristics found in MS images.

REFERENCES

[1] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, 2013.
[2] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[3] H. Song, B. Huang, Q. Liu, and K. Zhang, "Improving the spatial resolution of Landsat TM/ETM through fusion with spot5 images via learning-based super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1195–1204, Mar. 2014.

[4] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[5] R. A. Schowengerdt, "Reconstruction of multispatial, multispectral image data using spatial frequency content," *Photogrammetric Eng. Remote Sens.*, vol. 46, no. 10, pp. 1325–1334, 1980.

[6] R. Haydn, "Application of the IHS color transform to the processing of multisensor data and image enhancement," in *Proc. Int. Symp. Remote Sens. Arid Semi-Arid Lands*, Cairo, Egypt, 1982, Art. no. 1982.

[7] D. Sylla, A. Minghelli-Roman, P. Blanc, A. Mangin, and O. H. F. d'Andon, "Fusion of multispectral images by extension of the pan-sharpening arsis method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1781–1791, May 2014.

[8] P. Du, S. Liu, J. Xia, and Y. Zhao, "Information fusion techniques for change detection from multi-temporal remote sensing images," *Inf. Fusion*, vol. 14, no. 1, pp. 19–27, 2013.

[9] Y. Qu, H. Qi, B. Ayhan, C. Kwan, and R. Kidd, "Does multispectral/hyperspectral pansharpening improve the performance of anomaly detection?," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 6130–6133.

[10] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms + pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[11] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 1, pp. 339–348, 1989.

[12] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat TM and spot panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[13] W. Carper et al., "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogrammetric Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, 1990.

[14] G. A. Licciardi, M. M. Khan, J. Chanussot, A. Montanvert, L. Condat, and C. Jutten, "Fusion of hyperspectral and panchromatic images using multiresolution analysis and nonlinear PCA band reduction," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–17, 2012.

[15] M. Ghahremani and H. Ghassemian, "Nonlinear IHS: A promising method for pan-sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1606–1610, Nov. 2016.

[16] Y. Leung, J. Liu, and J. Zhang, "An improved adaptive intensity–hue–saturation method for the fusion of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 985–989, May 2014.

[17] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[18] X. Otazu, M. González-Audícana, O. Fors, and J. Núñez, "Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

[19] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.

[20] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[21] P. Chavez et al., "Comparison of three different methods to merge multiresolution and multispectral data- landsat TM and spot panchromatic," *Photogrammetric Eng. Remote Sens.*, vol. 57, no. 3, pp. 295–303, 1991.

[22] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

[23] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for p xs image fusion," *Int. J. Comput. Vis.*, vol. 69, pp. 43–58, 2006.

[24] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[25] Y. Liu and Z. Wang, "A practical pan-sharpening method with wavelet transform and sparse representation," in *Proc. IEEE Int. Conf. Imag. Syst. Technol.*, 2013, pp. 288–293.

[26] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[27] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.

[28] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[29] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[30] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[31] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[32] S. Xu et al., "Deep convolutional sparse coding network for pansharpening with guidance of side information," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[33] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5449–5457.

[34] W. Tan, P. Xiang, J. Zhang, H. Zhou, and H. Qin, "Remote sensing image fusion via boundary measured dual-channel PCNN in multi-scale morphological gradient domain," *IEEE Access*, vol. 8, pp. 42540–42549, 2020.

[35] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[36] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, and G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9088–9101, Nov. 2022.

[37] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "PAN-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inform. Fusion*, vol. 62, pp. 110–120, 2020.

[38] H. Zhou, Q. Liu, and Y. Wang, "Pgman: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, 2021.

[39] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 87–102.

[40] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[41] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[42] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[44] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

[46] M. Jaderberg et al., "Spatial transformer networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[48] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[49] D. Hong et al., "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5518615.

[50] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 498.

[51] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf Comput. Vis.*, 2021, pp. 10012–10022.

[52] F. Zhang, K. Zhang, and J. Sun, "Multiscale spatial–spectral interaction transformer for pan-sharpening," *Remote Sens.*, vol. 14, no. 7, 2022, Art. no. 1736.

[53] M. Li, Y. Liu, T. Xiao, Y. Huang, and G. Yang, "Local-global transformer enhanced unfolding network for pan-sharpening," 2023, *arXiv:2304.14612*.

[54] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

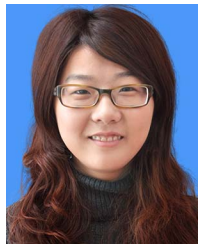
[55] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1701–1710.
[56] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[57] J. Pushparaj and A. V. Hegde, "Evaluation of pan-sharpening methods for spatial and spectral quality," *Appl. Geomatics*, vol. 9, pp. 1–12, 2017.
[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

Zixu Li received the bachelor's degree in software engineering from the School of Business, Jiangxi University of Science and Technology, Nanchang, China in 2022. He is currently working towatrd the master's degree with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, Shandong, China.

His research interests include computer graphics, computer vision, and image progressing.

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 with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, China. He is currently a Professor with 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.



Lu Ren received the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2021.

She is currently a Lecturer with Shandong Technology and Business University. Her current research interests include sentiment analysis and text mining.

Zheng Chen received the B.S. and M.S. degrees from Shandong Agricultural University and Shandong Normal University, Jinan, China, in 2012 and 2015, respectively. He received the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2022.

He is currently a Lecturer with Shandong Technology and Business University. His research interests include computer vision, hand pose estimation and hand shape recovery.