Crossed Dual-Branch U-Net for Hyperspectral Image Super-Resolution

Jingyi Zhang, Jianjun Liu, *Member, IEEE*, Jinlong Yang, and Zebin Wu, *Senior Member, IEEE*

***Abstract*—Hyperspectral images have gained great achievements in many fields, but their low spatial resolution limits the effectiveness in applications. Hyperspectral image super-resolution has emerged as a popular research trend, where high-resolution hyperspectral images are obtained via combining low-resolution hyperspectral images with high-resolution multispectral images. In this process of multimodality data fusion, it is crucial to ensure effective cross-modality information interaction. To generate higher quality fusion results, a crossed dual-branch U-Net is proposed in this article. In specific, we adopt U-Net architecture and introduce a spectral–spatial feature interaction module to capture cross-modality interaction information between two input images. To narrow the gap between downsampling and upsampling processes, a spectral–spatial parallel Transformer is designed as skip connection. This novel design simultaneously learns the long-range dependencies both on spatial and spectral information and provides detailed information for final fusion. In the fusion stage, we adopt a progressive upsampling strategy to refine the generated images. Extensive experiments on several public datasets are conducted to prove the performance of the proposed network.**

***Index Terms*—Hyperspectral image (HSI), multispectral image, super-resolution transformer, U-Net.**

I. Introduction

HYPERSPECTRAL images (HSIs) have extensive spectral bands that carry a wealth of spectral information, which enables them to identify object materials. Accordingly, HSIs have been employed in the field of computer vision, including image classification [1], [2] and environmental monitoring [3] to anomaly detection [4], [5], etc. However, limited by sensor devices, obtaining HSIs with high spatial and spectral resolution concurrently is difficult. To alleviate this problem, some methods for HSI super-resolution have been suggested [6], [7], [8]. There are usually two solutions, namely, single image and fusion-based HSI super-resolution. The fusion-based strategy obtains high-resolution hyperspectral (HRHS) images by merging low-resolution hyperspectral (LRHS) images with high-resolution multispectral (HRMS) images, which are preferred solutions. The existing fusion-based methods for HSI super-resolution generally fall into four classes: component substitution (CS), multiresolution analysis (MRA), model-based, and deep learning-based methods.

The CS methods attempt at generating HRHS images simply by replacing the spatial details in LRHS images with the corresponding HRMS images [9], [10]. The MRA methods extract spatial detail information of HRMS images by multiresolution decomposition, and yield HRHS images by incorporating the obtained spatial details into LRHS images [11], [12]. Both CS and MRA methods have advantages, such as low computational cost and fast implementation, but they often suffer from spectral or spatial distortion.

Model-based methods typically establish an optimization function to model the fusion problem, and the function is solved with iterative algorithms [13], [14], [15], [16], [17], [18]. In HSI super-resolution tasks, the optimization function generally includes two parts: data fidelity terms and regularization terms. The data fidelity terms mainly serve to stabilize the model and reduce the differences between input and output images in spatial and spectral information. The regularization terms constrain the fusion result based on some prior knowledge. These prior knowledge are often based on the latent statistics of HSI, such as sparsity prior [19], [20], low-rank prior [21], and total variation prior [22]. Model-based methods have the advantage of interpretability, but they usually rely too much on handcrafted priors, resulting in many parameters need to be tuned.

Deep learning-based methods have attracted extensive interest from researchers owing to their powerful feature extraction capabilities. These methods typically build end-to-end deep neural networks to effectively learn the underlying relationships of inputs and outputs. In recent years, many CNNs have been employed in HSI super-resolution tasks, such as ResNet [23], U-Net [24], [25], DenseNet [26], and GAN [27]. Because of the limited size of receptive field in convolution operation, CNNs fail to effectively utilizing global information. To overcome this disadvantage, Transformer has been developed and become a promising solution [28]. Transformer relies on the self-attention mechanism to handle the long-range dependencies in images, which has been applied successfully in HSI super-resolution tasks.

Manuscript received 1 August 2023; revised 11 September 2023 and 9 November 2023; accepted 14 December 2023. Date of publication 21 December 2023; date of current version 4 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62071204 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20201338. *(Corresponding author: Jianjun Liu.)*

Jingyi Zhang, Jianjun Liu, and Jinlong Yang are with the Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214126, China (e-mail: 18438859733@163.com; liuofficial@163.com; yjlgedeng@163.com).

Zebin Wu is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zebin.wu@gmail.com).

Data is available online at https://github.com/liuofficial/CD-UNet

Digital Object Identifier 10.1109/JSTARS.2023.3345411

There are two shortcomings in current deep learning-based methods. One is that these methods cannot fully utilize local and global features. The other is that these methods often ignore the correlations of spectral information, leading to suboptimal fusion. Given the aforementioned issues, we propose a crossed dual-branch U-Net for HSI super-resolution based on CNN and Transformer. Specifically, we design two CNN-based branches to fully extract local and shallow features of images. To make full use of these features, a feature interaction module that consists of convolution operations and matrix multiplications is designed to merge these spectral and spatial features and generate interaction information. In particular, we propose a spectral–spatial parallel Transformer (SSPT) that includes a spatial self-attention and a spectral self-attention, which both considers the spatial correlations and spectral correlations. This study's major contributions are described in the following.

1) A novel HSI super-resolution method named crossed dual-branch U-Net is proposed, which combines CNN and Transformer to effectively utilize local details and global information.
2) To facilitate the interaction of information between branches, we introduce a spectral–spatial feature interaction module (SFIM), hence improving quality of fusion.
3) We introduce an SSPT as skip connection to supplement global relevance features, which models global spatial information and takes into account the dependencies between adjacent spectral bands.

The rest of this article is organized as follows. In Section II, we review existing works in HSI super-resolution. Section III mainly describes the proposed network and its components. The presentation and analysis of the experimental results are discussed in Section IV. Finally, Section V concludes this article.

II. Related Works

We give brief review of the model-based and deep learning-based methods of HSI super-resolution.

A. Model-Based Methods

In general, model-based methods are summarized within two categories, nonfactorization- and factorization-based methods. Nonfactorization-based methods aim to obtaining target images via prior knowledge. For example, Wei et al. [29] utilized the probability information within the scene and proposed a Bayesian fusion method. A fast fusion method integrating Sylvester equation was presented, which dramatically reduced computational complexity [30]. Factorization-based methods mainly decompose the target image and then build an optimization model to solve. The factorization-based methods include matrix factorization-based methods [31], [32], [33], [34] and tensor decomposition-based methods [16], [35], [36], [37], [38], [39], [40], [41], [42]. Matrix factorization-based methods primarily transform the fusion task into an estimation of the spectral basis and corresponding coefficients. Dian et al. [32] formulated an optimization model in conjunction with sparse prior and estimated the spectral basis and coefficients simultaneously. Considering the subspace low-rank relationships between HRMS/LRHS images, Xue et al. [21] proposed a subspace clustering-based approach that formulated a variational optimization model. Since the original HSIs are considered as 3-D cubes, the tensor decomposition-based methods could better handle multidimensional information. Examples of popular tensor decomposition methods include Tucker decomposition, CP decomposition, and tensor-ring decomposition. For example, Jin et al. [38] presented a tensor network by fusing the high-order tensors that correspond to LRHS and HRMS images, designing a new regularization term named weighted graph regularization. In response to the noise and nonsmooth problems, Guo et al. [39] inserted two different operators to design a tensor decomposition network. Based on tensor-ring decomposition, He et al. [40] designed a model that iteratively obtain corresponding core tensors from LRHS and HRMS images. A regularization method was proposed by Xu et al. [42], which integrated two priors simultaneously to estimate tensor subspace and tensor coefficients and obtained excellent super-resolution results.

B. Deep Learning-Based Methods

Deep CNNs have powerful feature extraction capabilities and are extensively used in variety of deep learning tasks. In the last few years, many efficient HSI super-resolution methods that use CNNs have been proposed [43], [44], [45], [46], [47], [48], [49], [50]. Yang et al. [43] introduced a network with two branches, where one branch was dedicated to extracting spatial features of HRMS image while the other branch was involved in extracting spectral features of LRHS image. To fully utilize multiscale features, Zhan et al. [44] raised a network incorporating octave convolution with attention mechanism and designed a multisupervised loss function. For a further improvement in the interpretability of pure deep networks, model-driven methods have been suggested. Specifically, these methods solve the iterative algorithm by building a deep network [45]. Combining effective mathematical theoretical guidance, Dong et al. [46] suggested a dual spatial–spectral optimization strategy and introduced two optimization branches based on spatial and spectral priors, respectively. Based on U-Net architecture, Wang et al. [49] proposed a novel approach incorporating spectral and spatial attention that employed dense multiscale link as skip connection to obtain finer feature information. Ran et al. [51] presented a fusion network enabling to solve different resolution augmentation tasks, and incorporated multiscale high-resolution guidance to yield promising fusion results.

Transformer was initially applied in natural language processing. Due to the outstanding performance, it is gradually introduced to other fields as well [52], [53]. Likewise, many HSI super-resolution methods-based Transformer has also been raised [54], [55], [56], [57]. In the beginning, Hu et al. [54] directly fed the upsampled LRHS image concatenated with HRMS image to vision Transformer and achieved excellent results. Wang et al. [55] presented a Transformer-based network that utilized cross-attention for information fusion and enabled

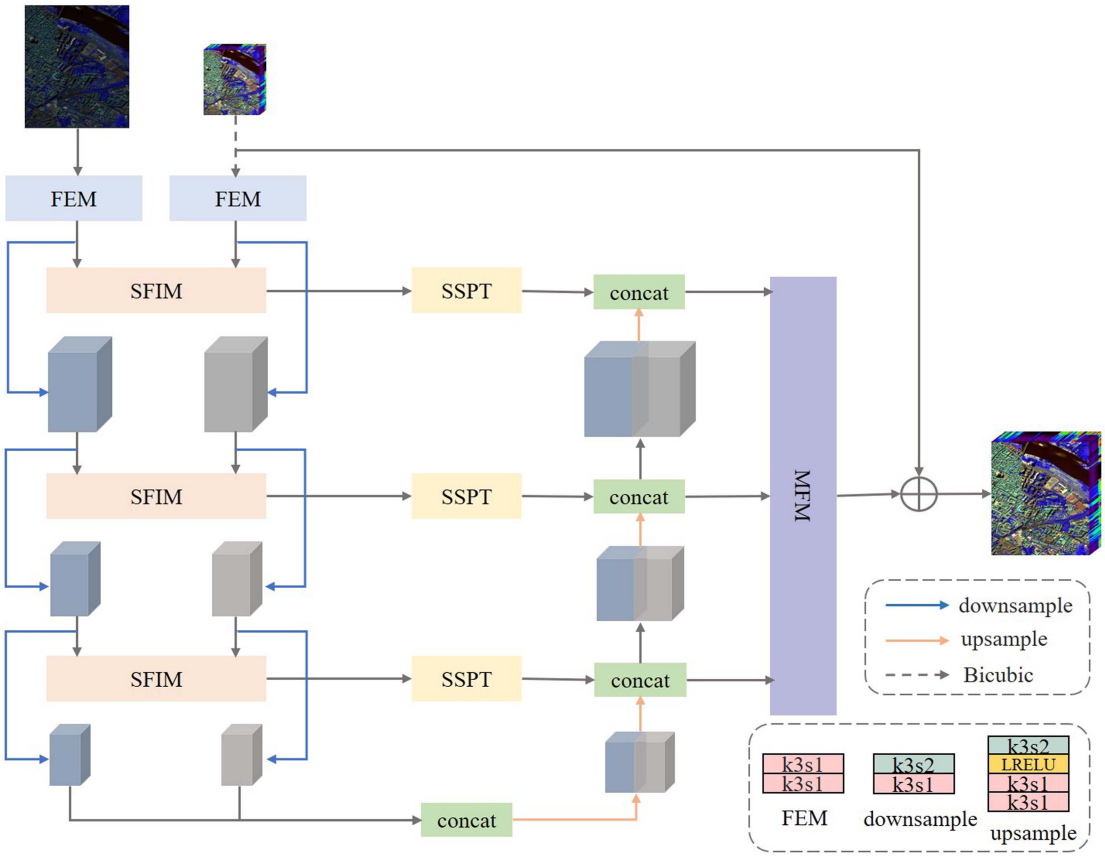


Fig. 1. Illustration of the proposed cross dual-branch U-Net. The structures of FEM, upsampling, and downsampling are shown at the bottom right, where k represents the kernel size, and s presents the stride size. LReLU indicates LeakyReLU.

multilevel feature extraction and aggregation. A novel pyramid network was proposed based on window self-attention by Deng et al. [56], they considered information interaction between patches and solved computational complexity problem by fixing a smaller window size.

III. Methodology

In this section, we provide a thorough overview of the proposed network and loss function.

A. Overall Network Architecture

For brevity, the LRHS image is represented by $\mathbf{Y} \in \mathbb{R}^{h\times w\times C}$, where $h \times w$ and C correspond to its spatial resolution and band number, respectively. $\mathbf{Z} \in \mathbb{R}^{H\times W\times c}$ indicates the HRMS image, and H, W, and c stand for its height, width, and number of bands, respectively. The HRHS image to be generated is denoted as $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$. The primary goal of our method is to generate HRHS images that share as much spectral information as possible with the input LRHS images and spatial information with the input HRMS images.

The proposed network is illustrated in Fig. 1, which mainly contains four primary modules: feature extraction module (FEM), SFIM, SSPT, and multiscale fusion module (MFM). To match the size of two inputs, we first upsample the LRHS images by commonly used bicubic interpolation. Then, two FEMs are employed to extract features from the upsampled LRHS image and HRMS image, where FEM is composed of two same 3×3 convolutional layers with a stride of 1, and these spectral and spatial features are fused by designed SFIM to realize cross-modality information interaction. Motivated by U-Net, we introduce an SSPT as skip connection that allows the network to capture long-range dependencies and compensate information loss. Finally, MFM gradually incorporates multiscale fusion information by continuous stacking and upsampling, the numbers of channels for each feature map are 64, 96, and 128, respectively. The proposed network achieves a compromise of spectral and spatial information, generating accurate and high-quality HRHS images.

B. Spectral–Spatial Feature Interaction Module

HSIs are considered as integrated data cubes of imagery and spectrum, both spectral and spatial information are important. LRHS images have richer spectral information, while HRMS images contain more spatial information. To integrate these spectral information and spatial information effectively, we introduce an SFIM at each scale. The feature maps of LRHS and HRMS images are denoted as Y_i and Z_i, respectively. The details of SFIM are shown in Fig. 2. In SFIM, we first extract

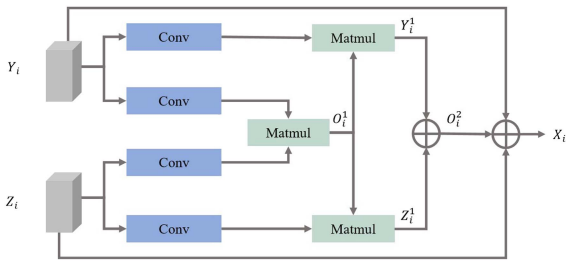


Fig. 2. Illustration of the SFIM, where $\oplus$ represents addition.

spectral features from Y_i and spatial features from Z_i by using a 3×3 convolution operation. After that, the extracted features are fused by performing a matrix multiplication to obtain interaction feature $\mathbf{O}_i^1$. The formula of $\mathbf{O}_i^1$ can be summarized as follows:

$$\mathbf{O}_i^1 = \text{Matmul}(\text{Conv}(\mathbf{Y}_i), \text{Conv}(\mathbf{Z}_i)) \tag{1}$$

where Matmul represents matrix multiplication.

Notably, considering $\mathbf{O}_i^1$ as the first-order interaction feature, we can further obtain the second-order interaction feature by performing similar operations. Specifically, same convolutional operations are performed on $\mathbf{Y}_i$ and $\mathbf{Z}_i$ again, and then, a matrix multiplication is performed between the obtained features and $\mathbf{O}_i^1$, resulting $\mathbf{Y}_i^1$. In the same way, we can also get $\mathbf{Z}_i^1$. At last, $\mathbf{Y}_i^1$ and $\mathbf{Z}_i^1$ are added to generate second-order interaction feature $\mathbf{O}_i^2$. The process of obtaining $\mathbf{O}_i^2$ is formulated as

$$\mathbf{Y}_i^1 = \text{Matmul}(\text{Conv}(\mathbf{Y}_i), \mathbf{O}_i^1) \tag{2}$$

$$\mathbf{Z}_i^1 = \text{Matmul}(\text{Conv}(\mathbf{Z}_i), \mathbf{O}_i^1) \tag{3}$$

$$\mathbf{O}_i^2 = \mathbf{Y}_i^1 + \mathbf{Z}_i^1. \tag{4}$$

Besides, $\mathbf{Y}_i$ and $\mathbf{Z}_i$ are added together to retain detailed information, which enhances the network's capability to preserve spatial and spectral information.

C. Spectral–Spatial Parallel Transformer

Transformer is well known for its ability of capturing long-distance dependencies in spatial locations. Given that spectral and spatial information are both important for HSIs, we design a spectral–spatial parallel Transformer, which takes both spatial global correlations and spectral correlations into consideration.

As we can see in Fig. 3, SSPT includes a spectral self-attention, a spatial self-attention, and a feedforward network. Taking the spatial self-attention as an example, the input represented by $\mathbf{X}_i \in \mathbb{R}^{H \times W \times C}$ is first projected and reshaped into $\mathbf{P} \in \mathbb{R}^{HW \times C}$, and then $\mathbf{P}$ is projected into $\mathbf{K} \in \mathbb{R}^{HW \times C}$, $\mathbf{Q} \in \mathbb{R}^{HW \times C}$, and $\mathbf{V} \in \mathbb{R}^{HW \times C}$ by the linear layers. The spatial self-attention can be formulated as

$$\mathbf{P} = \text{Reshape\&Linear}(\mathbf{X_i}) \tag{5}$$

$$\mathbf{Q} = \mathbf{P}\mathrm{W}^{\mathrm{Q}}, \mathbf{K} = \mathbf{P}\mathrm{W}^{\mathrm{K}}, \mathbf{V} = \mathbf{P}\mathrm{W}^{\mathrm{V}} \tag{6}$$

$$\text{Attention} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_k}}\right)\mathbf{V} \tag{7}$$

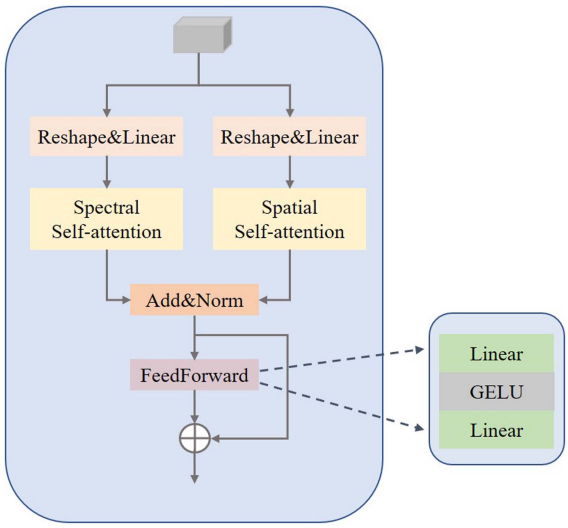


Fig. 3. Illustration of the SSPT.

where $\mathbf{Q}$ indicates the query matrix, $\mathbf{K}$ denotes the key matrix, and $\mathbf{V}$ represents the value matrix, respectively. Their corresponding learnable projection matrices are represented by W^Q, W^K, and $\mathrm{W}^V \in \mathbb{R}^{C \times C}$.$d_k$ corresponds to the dimension of $\mathbf{K}$.$\mathbf{Q}\mathbf{K}^{\mathbf{T}}$ calculates the attention score by dot product.

Multihead attention divides the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into multiple heads, each of which calculates self-attention and captures different aspects of information in data. Specifically, the self-attention is computed h times in parallel with h being the number of heads, and then, these heads are combined to obtain multihead attention. The multihead self-attention is formulated as follows:

$$\text{MultiHead} = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{i}\right)\mathrm{W}^{\mathrm{O}} \tag{8}$$

$$\text{head}_\text{i} = \text{Attention}\left(\mathbf{Q}\mathrm{W}_\text{i}^{\mathrm{Q}}, \mathbf{K}\mathrm{W}_\text{i}^{\mathrm{Q}}, \mathbf{V}\mathrm{W}_\text{i}^{\mathrm{Q}}\right), \text{i} = 1, 2, \ldots, \text{h} \tag{9}$$

where W^{O} is a learnable projection matrix, and h is set to be 4 in experiments.

The spectral self-attention calculates the spectral correlations among pixels, and the spatial self-attention calculates spatial correlations among spectral bands. Their calculation processes are similar, and their corresponding illustrations are shown in Fig. 4. Different from the spatial self-attention, the three learnable projection matrices of spectral self-attention are reshaped into $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{C \times HW}$.

D. Multiscale Fusion Module

After the above process, we obtained spectral–spatial feature maps at different scales. The sizes of these feature maps are 64, 32, and 16, respectively. In order to make full use of these feature maps and generate HRHS image, we adopted a progressive fusion strategy and designed the MFM whose specific structure is depicted in Fig. 5.

In the MFM, feature maps from different scales are gradually upsampled by a block consisting of a 2×2 transposed convolutional layer and a 3×3 convolutional layer, where their strides

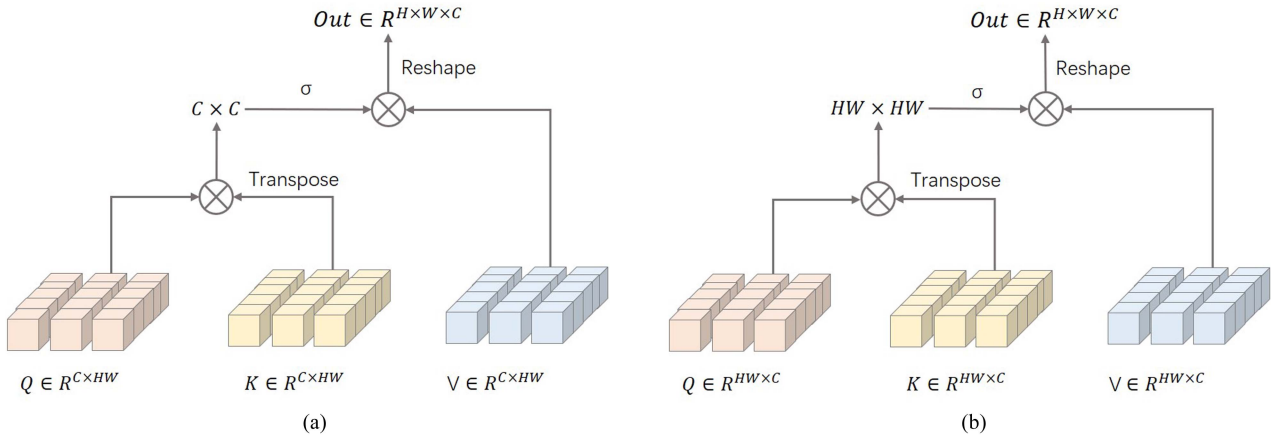


Fig. 4. Illustration of spectral and spatial self-attention. (a) Spectral multihead self-attention. (b) Spatial multihead self-attention, where $\otimes$ represents matrix multiplication.

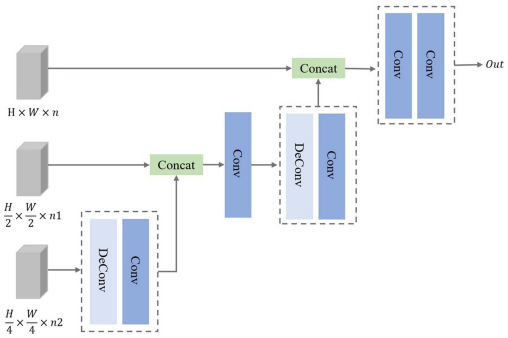


Fig. 5. Illustration of the MFM.

are set to 2 and 1. The upsampled feature maps are concatenated to generate HRHS image. This strategy is inspired by Zhang et al. [58], and we employ transposed convolutional operation for upsampling and achieve a better super-resolution performance. In Section IV, we have done ablation experiments with direct upsampling strategy and demonstrate the effectiveness of this progressive strategy.

E. Loss Function

We adopt L1 loss as loss function, which is commonly used to compute the difference of the desired image and fused images at pixel level. The formula of L1 loss function is as follows:

$$\text{Loss} = \|\mathbf{X} - \mathbf{X}'\|_1 \tag{10}$$

where $\mathbf{X}$ and $\mathbf{X}'$ denote the reference and fused images, respectively.

IV. Experiments Results

In this section, we present the datasets employed in our work as well as their data processing procedures, and present general evaluation metrics. In addition, we performed several ablation experiments and comparative experiments to evaluate the superiority of our approach.

A. Datasets Introduction

1) CAVE: The CAVE dataset[1] includes 32 HSIs, each image contains 31 spectral bands with resolution of 512 × 512. In this dataset, the wavelength ranges from 400 to 700 nm, and the spectral resolution is 10 nm. In our experiments, the former 20 images were assigned to the training set while the rest 12 images were devoted to the test set.

2) Harvard: The Harvard dataset[2] includes 50 different scenes with a resolution of 1392 × 1040. For every image, there are 31 spectral bands. Its wavelength ranging from 420 to 720 nm with a spectral resolution of 10 nm. In our experiments, we cropped the upper left corner of each image, resulting images with size of 1360 × 1024. We selected former 30 images as training set while the rest images were selected as test set.

3) Pavia Center (PC): The PC dataset[3] consists of HSIs with size of 1096 × 640 and band number of 115, which was captured by ROSIS sensors. After removing 13 noisy bands, there are 102 bands left. In our experiments, we cropped 40 nonoverlapping subimages of size 128 × 128 from the original image. The first 28 subimages were organized as training set, and the rest were organized as test set.

Three simulated datasets were processed following Ranchin and Wald's protocol [59]. Specifically, we performed a Gaussian filter with a scale factor of 4 on the original HSIs to generate LRHS images, and the HRMS images were generated via spectral response function (SRF). For the first two datasets, the corresponding SRF was derived from a Nikon D700 camera, while the corresponding SRF was from the IKONOS satellite for the PC dataset. In our experiments, we cropped image patches of size 64 × 64 and 16 × 16 from the observed HRMS and LRHS images as input.

[1][Online]. Available: https://www1.cs.columbia.edu/CAVE/databases/multispectral/

[2][Online]. Available: http://vision.seas.harvard.edu/hyperspec/download.html

[3][Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_University_scene

B. Quantitative Assessment Metrics

1) Spectral Angle Mapper (SAM): SAM is a commonly used metric that quantifies the image quality in terms of spectral dimension. The lower SAM value indicates the lower spectral distortion

$$\mathrm{SAM}(\mathbf{X},\mathbf{X}')=\frac{1}{WH}\sum_{\mathrm{i}=1}^{W}\sum_{\mathrm{j}=1}^{H}\arccos\left(\frac{\mathbf{X}(\mathrm{i},\mathrm{j})\cdot\mathbf{X}'(\mathrm{i},\mathrm{j})}{\|\mathbf{X}(\mathrm{i},\mathrm{j})\|_2\|\mathbf{X}'(\mathrm{i},\mathrm{j})\|_2}\right). \tag{11}$$

2) Peak Signal-to-Noise Ratio (PSNR): PSNR is a general metric to calculate pixel similarities between a pair of images. Higher values of PSNR indicate better results

$$\mathrm{PSNR}(\mathbf{X},\mathbf{X}')=10\lg\left(\frac{\max(\mathbf{X})^2}{\frac{1}{\mathrm{HWC}}\|\mathbf{X}-\mathbf{X}'\|_F^2}\right) \tag{12}$$

where $\max(\mathbf{X})$ presents the largest pixel value in $\mathbf{X}$.

3) Root-Mean-Squared Error (RMSE): RMSE can access the average difference between $\mathbf{X}$ and $\mathbf{X}'$ in pixel wise. Its value range is from 0 to 1, and the smaller RMSE value indicates the better result

$$\mathrm{RMSE}(\mathbf{X},\mathbf{X}')=\sqrt{\frac{\sum_{\mathrm{k}=1}^{\mathrm{C}}\sum_{\mathrm{i}=1}^{\mathrm{H}}\sum_{\mathrm{j}=1}^{\mathrm{W}}\left(\mathbf{X}_{\mathrm{k}}(\mathrm{i},\mathrm{j})-\mathbf{X}'_{\mathrm{k}}(\mathrm{i},\mathrm{j})\right)^2}{HWC}} \tag{13}$$

where $\mathbf{X}_{\mathrm{k}}(\mathrm{i},\mathrm{j})$ denotes pixel value at position (i,j) of the kth band of $\mathbf{X}$.

4) Erreur Relative Globale Adimensionnelle De Synthse (ERGAS): ERGAS is an evaluation index that is used to assess the overall quality of image. The higher ERGAS value indicates the superior fusion quality

$$\mathrm{ERGAS}(\mathbf{X},\mathbf{X}')=\frac{100}{r}\sqrt{\frac{1}{\tilde{\mathrm{C}}}\sum_{\mathrm{k}=1}^{\mathrm{C}}\frac{\|\mathbf{X}_{\mathrm{k}}-\mathbf{X}'_{\mathrm{k}}\|_F^2}{\mu^2\left(\mathbf{X}_{\mathrm{k}}\right)}} \tag{14}$$

where r denotes downsampling factor while μ is a function that calculates mean value.

5) Structure Similarity Index Measure (SSIM): SSIM evaluates the structural similarity between two images. The higher SSIM value suggests the better quality of fused image

$$\mathrm{SSIM}(\mathbf{X},\mathbf{X}')=\frac{1}{C}\sum_{\mathrm{k}}^{C}\frac{\left(2\mu_{\mathbf{X}_{\mathrm{k}}}\mu_{\mathbf{X}'_{\mathrm{k}}}+a_1\right)\left(2\sigma_{\mathbf{X}_{\mathrm{k}}\mathbf{X}'_{\mathrm{k}}}+a_2\right)}{\left(\mu^2_{\mathbf{X}_{\mathrm{k}}}+\mu^2_{\mathbf{X}'_{\mathrm{k}}}+a_1\right)\left(\sigma^2_{\mathbf{X}_{\mathrm{k}}}+\sigma^2_{\mathbf{X}'_{\mathrm{k}}}+a_2\right)} \tag{15}$$

where a_1 and a_2 are constants, $\mu_{\mathbf{X}_{\mathrm{k}}}$ and $\mu_{\mathbf{X}'_{\mathrm{k}}}$ denote the mean values of $\mathbf{X}_{\mathrm{k}}$ and $\mathbf{X}'_{\mathrm{k}}$, respectively, and $\sigma_{\mathbf{X}_{\mathrm{k}}}$ and $\sigma_{\mathbf{X}'_{\mathrm{k}}}$ present the standard value of $\mathbf{X}_{\mathrm{k}}$ and $\mathbf{X}'_{\mathrm{k}}$, respectively. $\sigma_{\mathbf{X}_{\mathrm{k}}\mathbf{X}'_{\mathrm{k}}}$ is the covariance between $\mathbf{X}_{\mathrm{k}}$ and $\mathbf{X}'_{\mathrm{k}}$.

C. Comparison Methods

To thoroughly exhibit the effectiveness of our approach, we conducted comparisons against eight methods, including four model-based methods, namely, FUSE[4] [30], HySure[5] [60], CNMF[6] [13], and GSA[4] [61], and four deep learning-based methods, namely, SSR-Net[7] [62], HSRNet[8] [63], GuidedNet[9] [51], and MCT-Net[10] [55]. For fairness in comparison, all experiments were implemented using the same training set and testing set. The four deep learning-based methods were all executed in a Pytorch framework with a GeForce GTX 3090Ti 24 GB GPU. During the training process, we chose Adam optimizer and trained for 200 epochs, and learning rate was set to 0.0001. The four model-based methods were implemented in MATLAB 2019a. Parameter settings of all comparison methods were consistent with their respective original papers.

TABLE I
ABLATION STUDY OF THE SFIM AND SSPT ON THE CAVE DATASET

Component		Quantitative Metrics				
SFIM	SSPT	SAM	PSNR	RMSE	EGRAS	SSIM
		3.718	39.832	0.0120	3.4875	0.9867
✓		2.281	50.942	0.0033	0.9290	0.9989
	✓	2.127	51.357	0.0031	0.9284	0.9989
✓	✓	**1.983**	**51.676**	**0.0030**	**0.9105**	**0.9990**

The best results are marked in bold.

D. Ablation Study

In this section, we conducted multiple ablation experiments on SSPT and its components, SFIM as well as MFM on the CAVE dataset. All ablation experiments were conducted under the same environmental settings.

1) Influence of Components: We investigated the influences of some important modules in the model by removing them individually. From the results presented in Table I, we observed that the quantitative metrics significantly declined when either SFIM or SSPT was removed. When SSPT was removed, all metrics went worse, with particularly substantial changes in PSNR and SAM, which indicates that SSPT served an essential role in capturing both spatial and spectral information from a global perspective. Similarly, the absence of SFIM leads to suboptimal fusion results, which proved that effective cross-modality information interaction can enhance performance. In conclusion, SFIM and SSPT are both effective for the proposed network, and the network performs best when SFIM and SSPT are employed simultaneously.

2) Influence of Self-Attention: Different attention mechanisms are employed in SSPT. The results were displayed in Table II. When SSPT only contained a spatial self-attention, the values of PSNR and SAM both decreased, indicating that the ability of spectral self-attention for extracting global spectral features. Similarly, when SSPT only consisted of a spectral self-attention, the values of PSNR, SAM, and EGRAS show significant fluctuations, which demonstrated the effectiveness of spatial self-attention in capturing global spatial feature. The

[4][Online]. Available: https://openremotesensing.net/wp-content/uploads/2017/11/HSMSFusionToolbox.zip

[5][Online]. Available: https://github.com/alfaiate/HySure

[6][Online]. Available: https://naotoyokoya.com/assets/zip/CNMFMATLAB.zip

[7][Online]. Available: https://github.com/hw2hwei/SSRNET

[8][Online]. Available: https://liangjiandeng.github.io/ProjectsRes/HSRnet2021tnnls.html

[9][Online]. Available: https://github.com/Evangelion09/GuidedNet

[10][Online]. Available: https://github.com/wxy11-27/MCT-Net

TABLE II
ABLATION STUDY OF THE SELF-ATTENTION IN SFIM ON THE CAVE DATASET

Transformer		Quantitative Metrics				
Spectral	Spatial	SAM	PSNR	RMSE	EGRAS	SSIM
✓		2.160	51.496	0.0031	0.9304	0.9989
	✓	2.183	51.514	0.0033	0.9846	0.9987
✓	✓	**1.983**	**51.676**	**0.0030**	**0.9105**	**0.9990**

The best results are marked in bold.

TABLE III
ABLATION STUDY OF THE RECONSTRUCTION METHOD ON THE CAVE DATASET

Reconstruction	Quantitative Metrics				
	SAM	PSNR	RMSE	EGRAS	SSIM
Direct upsampling	2.177	51.409	0.0031	0.9209	**0.9990**
Progressive upsampling	**1.983**	**51.676**	**0.0030**	**0.9105**	**0.9990**

The best results are marked in bold.

TABLE IV
QUANTITATIVE RESULTS OF THE COMPARISON EXPERIMENTS ON THE CAVE DATASET

Method	SAM	PSNR	RMSE	EGRAS	SSIM
Best Values	0	$+\infty$	0	0	1
FUSE	4.020	37.886	0.0137	3.3525	0.9757
Hysure	5.324	40.769	0.0094	2.4734	0.9816
CNMF	4.731	40.024	0.0107	2.6347	0.9821
GSA	4.590	41.304	0.0090	2.3225	0.9835
SSR-Net	3.476	46.210	0.0058	1.7427	0.9971
HSRNet	2.486	49.325	0.0039	1.1774	0.9982
Guided-Net	2.888	48.201	0.0049	1.5021	0.9977
MCT-Net	2.584	49.371	0.0039	1.1919	0.9983
Ours	**1.983**	**51.676**	**0.0030**	**0.9105**	**0.9990**

The bold fonts represent the best values.

optimal fusion results are obtained when SSPT included both them.

3) Influence of MFM: Two different image reconstruction approaches are compared, one is directly upsampling images to the same size and subsequently fuse, the other is progressively upsampling and fuse. The ablation experiments on the reconstruction approach were conducted, and the results are presented in Table III. It is obvious that the direct upsampling achieves worse fusion results compared with progressive upsampling, which indicates that more information is lost during cross-scale fusion. Therefore, progressive upsampling and fusion were found to be more effective at preserving information and achieving better fusion results.

E. Results of Comparison Experiments on Simulated Datasets

1) Results on CAVE: Table IV gives the results on the CAVE dataset, where optimal results are bolded. What we can conclude is that our method outperforms in all quantitative evaluation metrics. This suggests that our method could improve spatial resolution while retaining spectral information. For a more intuitive representation of the reconstruction results of each method, we display the fused images and their corresponding error images on the sponges image in Fig. 6. We have marked the meaningful areas with red boxes. The error images can visualize the difference that exists between the reference and fused images. It is evident that spectral distortion and detail loss are common problems in HySure, CNMF, SSR-Net, and Guided-Net, and our method has the optimal fusion quality

TABLE V
QUANTITATIVE RESULTS OF THE COMPARISON EXPERIMENTS ON THE HARVARD DATASET

Method	SAM	PSNR	RMSE	EGRAS	SSIM
Best Values	0	$+\infty$	0	0	1
FUSE	3.422	37.479	0.0146	3.1269	0.9617
Hysure	3.258	40.305	0.0111	2.6508	0.9708
CNMF	2.851	42.537	0.0089	2.0189	0.9793
GSA	2.883	42.660	0.0085	1.9541	0.9787
SSR-Net	2.538	44.018	0.0075	1.8815	0.9920
HSRNet	2.327	45.348	0.0066	1.6134	0.9929
Guided-Net	2.432	44.551	0.0073	1.8139	0.9919
MCT-Net	2.339	44.974	0.0068	1.7000	0.9925
Ours	**2.219**	**45.770**	**0.0062**	**1.5650**	**0.9934**

The bold fonts represent the best values.

TABLE VI
QUANTITATIVE RESULTS OF THE COMPARISON EXPERIMENTS ON THE PC DATASET

Method	SAM	PSNR	RMSE	EGRAS	SSIM
Best Values	0	$+\infty$	0	0	1
FUSE	6.615	31.412	0.0270	4.4395	0.9235
Hysure	6.869	32.805	0.0230	3.8661	0.9425
CNMF	4.034	34.251	0.0197	3.3296	0.9524
GSA	5.007	35.845	0.0162	2.6861	0.9653
SSR-Net	3.286	38.218	0.0124	1.9230	0.9876
HSRNet	2.658	43.917	0.0064	1.5033	0.9939
Guided-Net	2.751	43.730	0.0065	1.5552	0.9937
MCT-Net	2.609	44.326	0.0063	1.4916	0.9942
Ours	**2.429**	**45.170**	**0.0060**	**1.4279**	**0.9946**

The bold fonts represent the best values.

among all comparison methods. The PSNR values of all bands are plotted in Fig. 7(a), where we can notice that our proposed method has the highest PSNR values on all bands, demonstrating the superiority of our method.

2) Results on Harvard: Table V illustrates the results for all comparison methods on the Harvard dataset. On all quantitative evaluation indicators, our network all obtains the best results, followed by Guided-Net. There are significant differences between model-based methods and deep learning-based methods on the Harvard dataset. We pick the imgf1 from the Harvard dataset for visualization in Fig. 8. What we can learn from the images is that there exists obvious distortions of FUSE, Hysure, CNMF, GSA, and MCT-Net, while our method has the best visualization results with the least amount of differences. Fig. 7(b) shows the PSNR values of all spectral bands. Although there is an overall decreasing trend in the PSNR values on the Harvard dataset, the optimal performance is achieved by our method. This suggests that our method is able to recover in parallel with spatial and spectral information.

3) Results on PC: The results of all the methods on the PC dataset are presented in Table VI. From the table, we find that GSA performs best in terms of PSNR, while CNMF performs best on SAM metric among the model-based methods. Our method obtains better values than all the other comparison methods on five metrics, followed by Guided-Net. Fig. 9 gives the fused images and their corresponding error images on band 61 of the nine methods. From the visualized results, we can learn that the model-based methods universally suffer from serious spectral and spatial distortion, followed by SSR-Net, HSRNet, and MCT-Net. Guided-Net and our method achieve better fusion

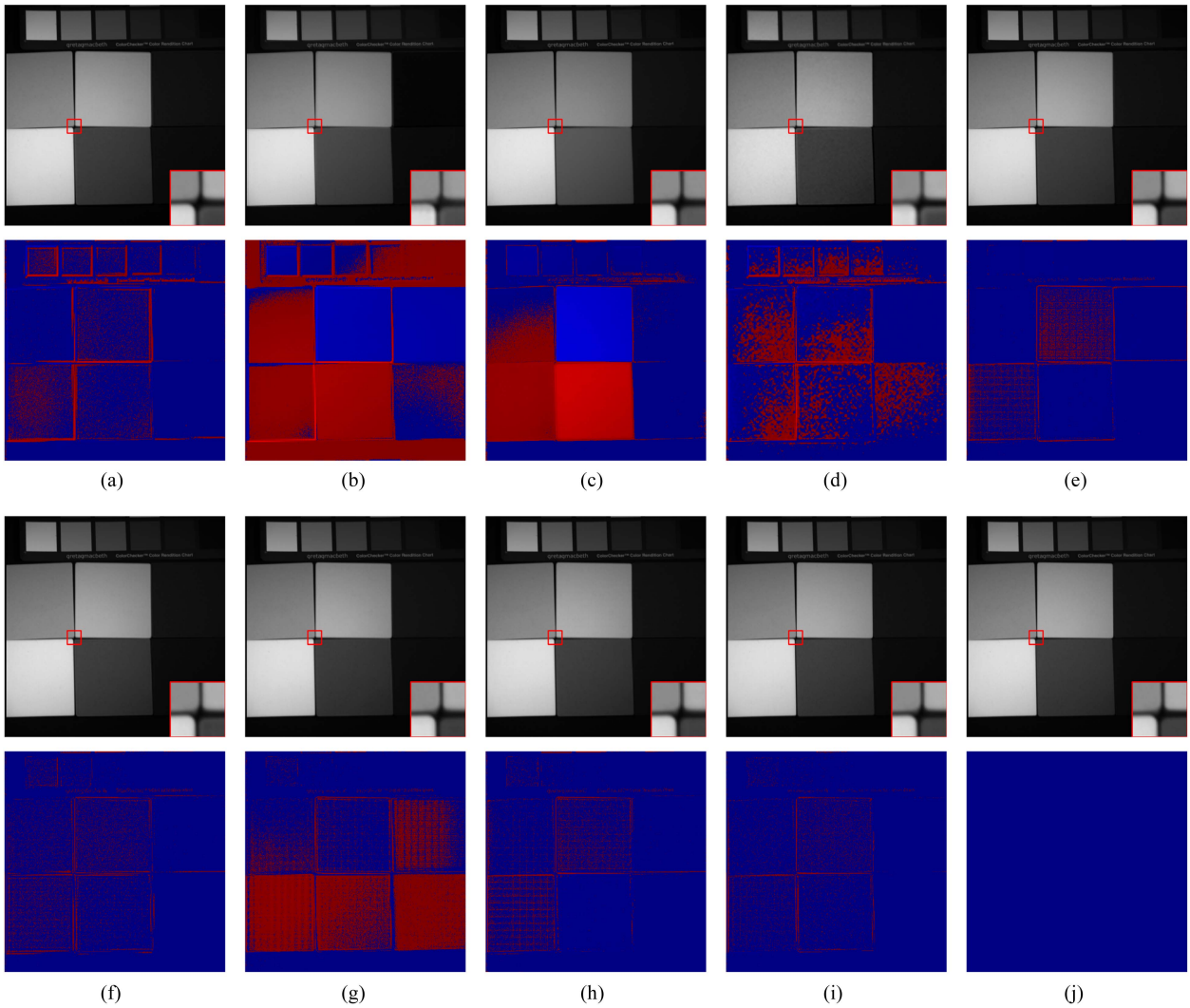


Fig. 6. Visual results of the CAVE dataset at band 19. (a) FUSE. (b) HySure. (c) CNMF. (d) GSA. (e) SSR-Net. (f) HSRNet. (g) Guided-Net. (h) MCT-Net. (i) Ours. (j) Ground truth.

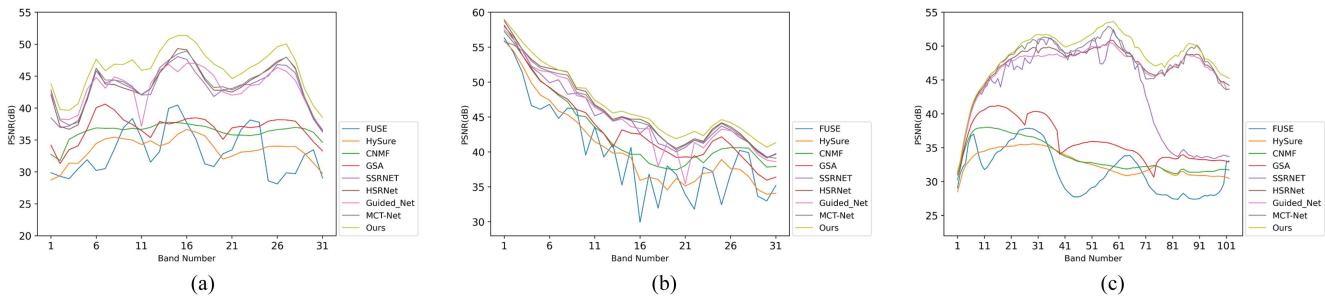


Fig. 7. PSNR as a function of spectral band. (a) CAVE dataset. (b) Harvard dataset. (c) PC dataset.

results. Fig. 7(c) provides a comparison of the PSNR on each spectral band of all methods. The difference between the fusion quality of model-based and deep learning-based approaches is obvious. Among deep learning-based approaches, our proposed method yields better quantitative and qualitative results on the PC dataset than other methods.

F. Experimental Results on Real Dataset

We performed further experiments on WV2 dataset to demonstrate the effectiveness of our proposed method in real-world scenarios. The WV2 dataset consists of an LRHS image and an RGB image with sizes of 419 × 658 × 8

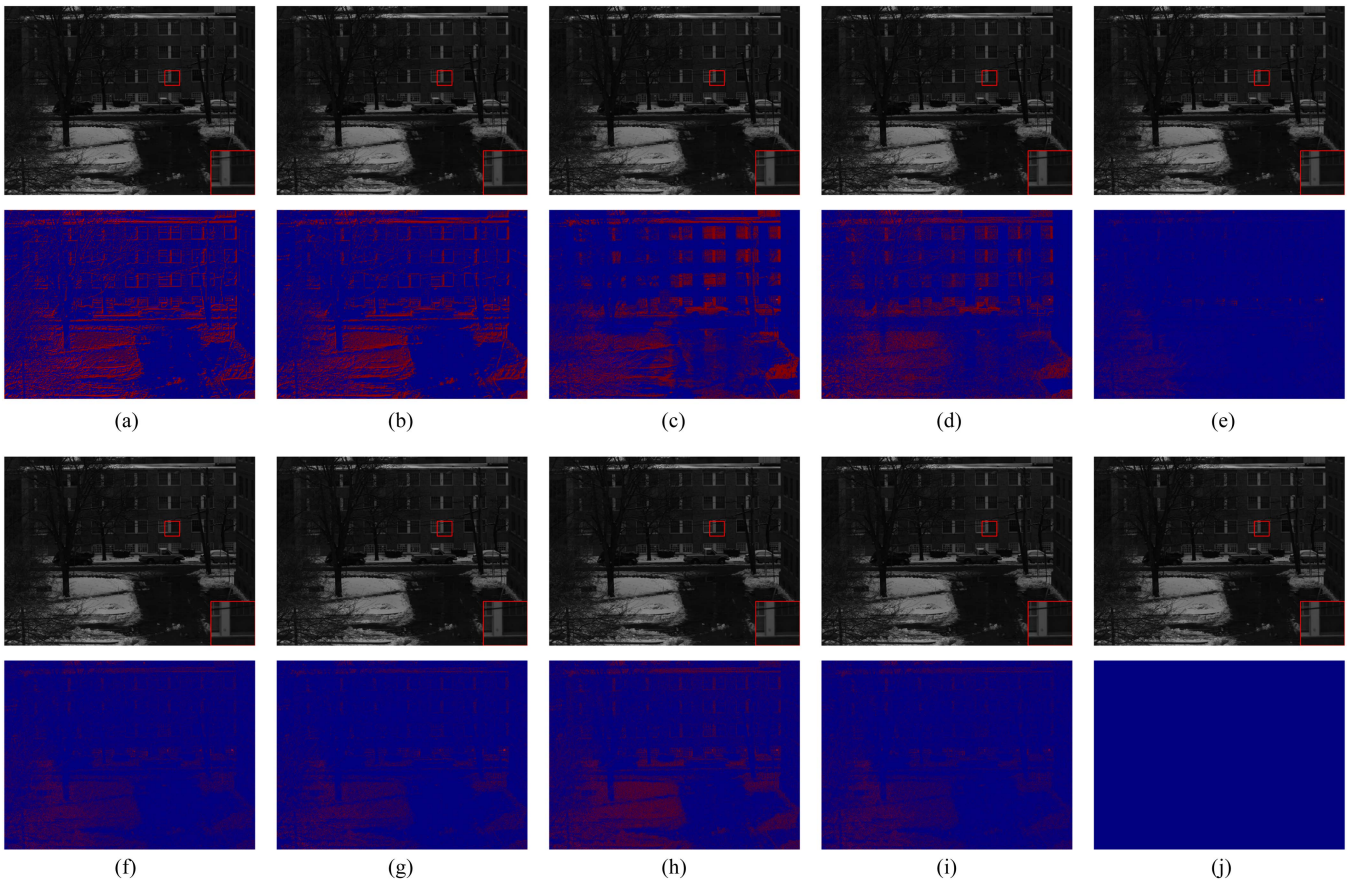


Fig. 8. Visual results of the Harvard dataset at band 23. (a) FUSE. (b) HySure. (c) CNMF. (d) GSA. (e) SSR-Net. (f) HSRNet. (g) Guided-Net. (h) MCT-Net. (i) Ours. (j) Ground truth.

TABLE VII
NUMBER OF PARAMETERS, FLOPS, AND TEST TIME OF THE DEEP LEARNING-BASED METHODS

Method	PARAMS	FLOPs	TIME
SSR-Net	26.1 K	425.1 M	2.4 s
HSRNet	264.1 K	16.0 G	4.1 s
Guided-Net	1.4 M	25.9 G	6.4 s
MCT-Net	941.7 K	1.2 G	10.6 s
Ours	8.1 M	17.3 G	6.3 s

and $1676 \times 2632 \times 3$, respectively. In our experiments, we cropped four sets of nonoverlapping images, where the size of HRMS images was 512×512 and the size of LRHS images was 128×128. The first three subimages were treated as training set and the rest one as test set. Since there are no available reference images, we regenerated experimental data according to Ranchin and Wald's protocol [59]. Specifically, we regarded the original images as reference and generated the HRMS and LRHS images using filters estimated by HySure [60]. In the training phase, we cropped HRMS and LRHS images with patch sizes of 32 and 8. In testing phase, we directly fed the original images into the network.

Fig. 10 illustrates the visualization results on the WV2 dataset. The meaningful regions are zoomed in red boxes. From the visualization results especially the error images, it is apparent that our method yields best visual effects in details and is closest to the original LRHS image. The outperformance in real scenarios further confirms the contributions of our method.

G. Computational Efficiency

To provide a comprehensive comparison, it is necessary to analyze the efficiency and computational cost of deep learning-based methods. Table VII displays the specific values of the number of parameters, FLOPs, and the testing time for deep learning-based methods. From the results in Table VII, we can learn that the proposed method has a higher number of parameters than other deep learning-based methods. The FLOPs of our model are lower than Guided-Net and slightly higher than HSRNet. Because the proposed model is composed of multiple SSPTs, which inevitably leads to suboptimal computational costs. The test time for a single image of our method is shorter than that of Guided-Net and MCT-Net, but longer than that of SSR-Net and HSRNet.

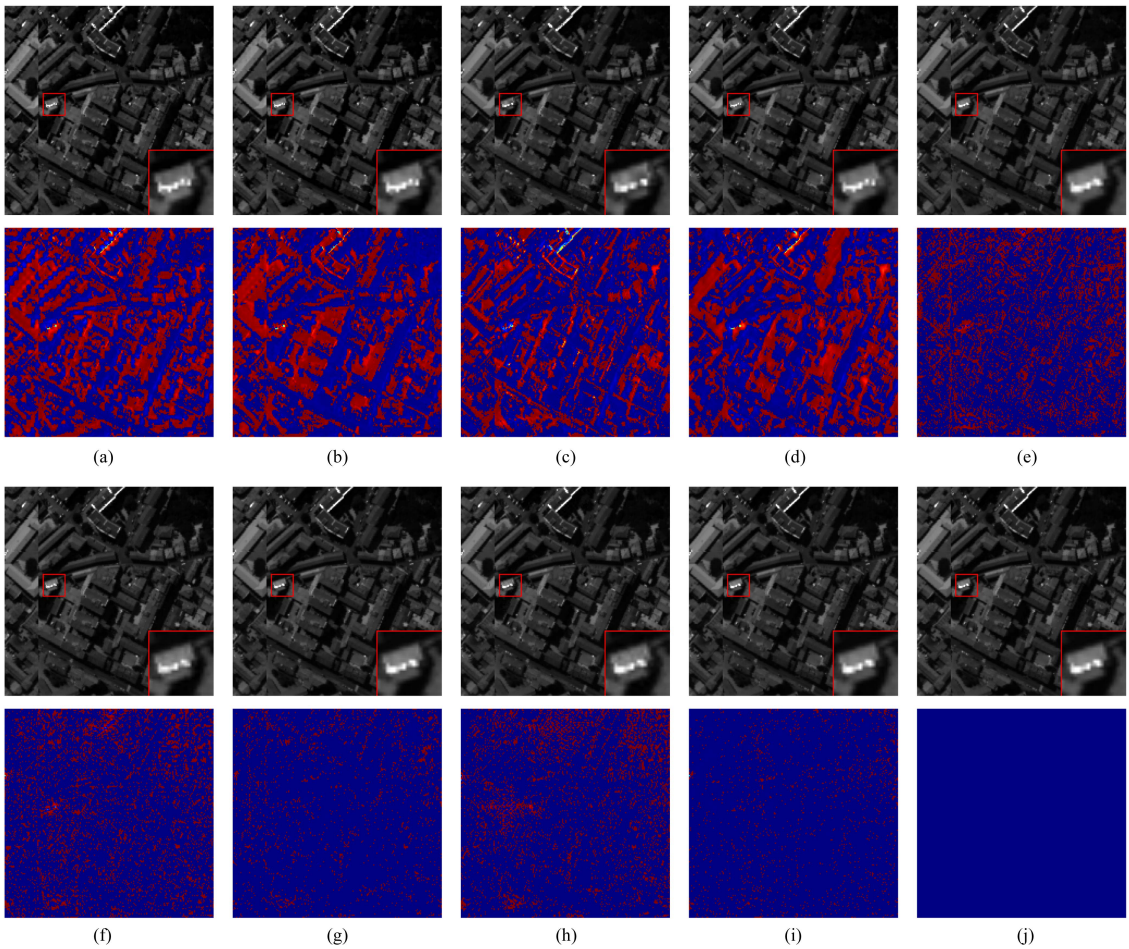


Fig. 9. Visual results of the PC dataset at band 61. (a) FUSE. (b) HySure. (c) CNMF. (d) GSA. (e) SSR-Net. (f) HSRNet. (g) Guided-Net. (h) MCT-Net. (i) Ours. (j) Ground truth.

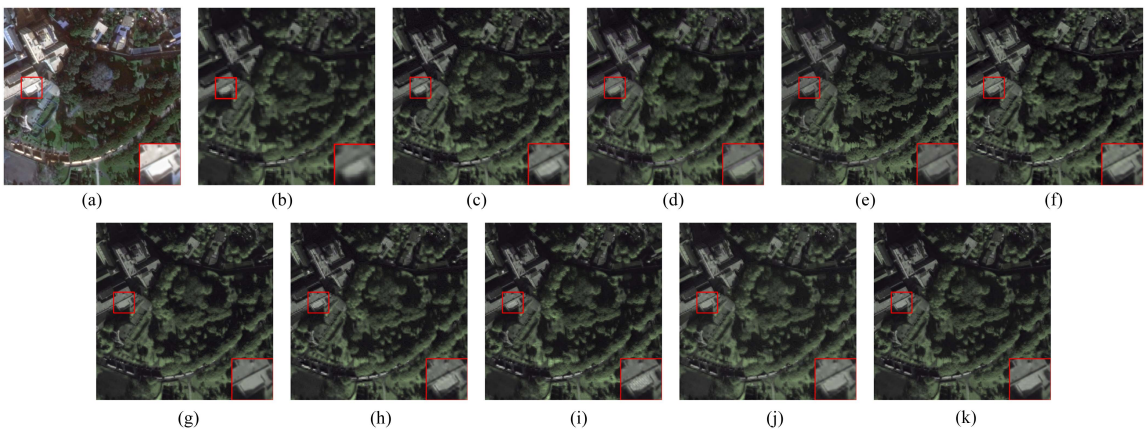


Fig. 10. Visual results of the WV2 dataset. (a) Real HR-MS image. (b) Real LR-HS image. (c) FUSE. (d) Hysure. (e) CNMF. (f) GSA. (g) SSR-Net. (h) HSRNet. (i) Guided-Net. (j) MCT-Net. (k) Ours.

V. Conclusion

This article proposes a crossed dual-branch U-Net for HSI super-resolution. The network adopts a dual-branch structure based on U-Net, focusing on extracting spatial features in HRMS images and spectral features in LRHS images, respectively. An SFIM is designed between the two branches to achieve cross-modality information interaction. Specially, we introduce an SSPT as skip connection, which can efficiently supplement correlative features and contributes to restore detailed information in the upsampling process. Finally, we employ a fusion strategy of progressive upsampling to further enhance the final fusion quality. Extensive comparison and ablation experiments are conducted on different datasets, where all outcomes

confirm that our approach is outperforming many advanced techniques.

Although our method achieved excellent fusion results, the network contains multiple Transformer modules, resulting in an excessive amount of parameters and high computational complexity. In future work, we will strive to achieve the balance between the performance and computational costs.

REFERENCES

[1] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[2] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[3] L. Yi, J. M. Chen, G. Zhang, X. Xu, X. Ming, and W. Guo, "Seamless mosaicking of UAV-based push-broom hyperspectral images for environment monitoring," *Remote Sens.*, vol. 13, no. 22, 2021, Art. no. 4720.

[4] H. Su, Z. Wu, H. Zhang, and Q. Du, "Hyperspectral anomaly detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 64–90, Mar. 2022.

[5] X. Hu et al., "Hyperspectral anomaly detection using deep learning: A review," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 1973.

[6] D. Sara, A. K. Mandava, A. Kumar, S. Duela, and A. Jude, "Hyperspectral and multispectral image fusion techniques for high resolution applications: A review," *Earth Sci. Inform.*, vol. 14, no. 4, pp. 1685–1705, 2021.

[7] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, 2023.

[8] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, 2023.

[9] 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.

[10] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875, Jan. 4, 2000.

[11] 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.

[12] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2014.

[13] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[14] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[15] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[16] R. A. Borsoi, C. Prévost, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, "Coupled tensor decomposition for hyperspectral and multispectral image fusion with inter-image variability," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 702–717, Apr. 2021.

[17] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4381–4398, Jul. 2020.

[18] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, Jul. 2020.

[19] W. Dong et al., "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[20] X. Li, Y. Zhang, Z. Ge, G. Cao, H. Shi, and P. Fu, "Adaptive non negative sparse representation for hyperspectral image super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4267–4283, Apr. 2021.

[21] J. Xue, Y.-Q. Zhao, Y. Bu, W. Liao, J. C.-W. Chan, and W. Philips, "Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, Feb. 2021.

[22] Y. Wang, X. Chen, Z. Han, and S. He, "Hyperspectral image super-resolution via nonlocal low-rank tensor approximation and total variation regularization," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1286.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*2015, pp. 234–241.

[25] Z. Liu, Y. Zheng, and X.-H. Han, "Deep unsupervised fusion learning for hyperspectral image super resolution," *Sensors*, vol. 21, no. 7, 2021, Art. no. 2348.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[27] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[29] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multi-band images," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1117–1127, Sep. 2015.

[30] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.

[31] M. A. Veganzones, M. Simoes, G. Licciardi, N. Yokoya, J. M. Bioucas-Dias, and J. Chanussot, "Hyperspectral super-resolution of locally low rank images from complementary multisource data," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 274–288, Jan. 2016.

[32] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial-spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262–270, 2019.

[33] A. Camacho, E. Vargas, and H. Arguello, "Hyperspectral and multispectral image fusion addressing spectral variability by an augmented linear mixing model," *Int. J. Remote Sens.*, vol. 43, no. 5, pp. 1577–1608, 2022.

[34] T. Gelvez-Barrera, H. Arguello, and A. Foi, "Joint nonlocal, spectral, and similarity low-rank priors for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5537112.

[35] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4747–4760, Nov. 2020.

[36] M. Zare, M. S. Helfroush, K. Kazemi, and P. Scheunders, "Hyperspectral and multispectral image fusion using coupled non-negative tucker tensor decomposition," *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 2930.

[37] H. Xu, M. Qin, S. Chen, Y. Zheng, and J. Zheng, "Hyperspectral-multispectral image fusion via tensor ring and subspace decompositions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8823–8837, 2021.

[38] D. Jin, J. Liu, J. Yang, and Z. Wu, "High-order coupled fully connected tensor network decomposition for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 6014105.

[39] H. Guo, W. Bao, K. Qu, X. Ma, and M. Cao, "Multispectral and hyperspectral image fusion based on regularized coupled non-negative block-term tensor decomposition," *Remote Sens.*, vol. 14, no. 21, 2022, Art. no. 5306.

[40] W. He, Y. Chen, N. Yokoya, C. Li, and Q. Zhao, "Hyperspectral super-resolution via coupled tensor ring factorization," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108280.

[41] Y. Bu et al., "Hyperspectral and multispectral image fusion via graph Laplacian-guided coupled tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 648–662, Jan. 2021.

[42] T. Xu, T.-Z. Huang, L.-J. Deng, and N. Yokoya, "An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5529316.

[43] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 800.

[44] T. Zhan et al., "A novel cross-scale octave network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5545816.
[45] J. Liu, D. Shen, Z. Wu, L. Xiao, J. Sun, and H. Yan, "Patch-aware deep hyperspectral and multispectral image fusion by unfolding subspace-based optimization model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1024–1038, Jan. 2022.
[46] W. Dong, T. Zhang, J. Qu, Y. Li, and H. Xia, "A spatial–spectral dual-optimization model-driven deep network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5542016.
[47] D. Shen, J. Liu, Z. Wu, J. Yang, and L. Xiao, "ADMM-HFNet: A matrix decomposition-based deep approach for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5513417.
[48] S. Liu, S. Liu, S. Zhang, B. Li, W. Hu, and Y.-D. Zhang, "SSAU-Net: A spectral–spatial attention-based U-Net for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5542116.
[49] X. Wang, X. Wang, K. Zhao, X. Zhao, and C. Song, "FSL-UNet: Full-scale linked UNet with spatial–spectral joint perceptual attention for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5539114.
[50] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522412.
[51] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023.
[52] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5528715.
[53] M. Pu, Y. Huang, Y. Liu, Q. Guan, and H. Ling, "EDTER: Edge detection with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1402–1412.
[54] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6012305.
[55] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl.-Based Syst.*, vol. 264, 2023, Art. no. 110362.
[56] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503715.
[57] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, 2023.
[58] 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. [Online]. Available: https://www.mdpi.com/2072-4292/14/7/1736
[59] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogrammetric Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, 2000.
[60] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
[61] 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.
[62] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-Net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.
[63] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

Jingyi Zhang received the B.S. degree in computer science and technology from the Henan Polytechnic University, Jiaozuo, China, in 2021. She is currently working toward the master's degree in electronic information with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China.

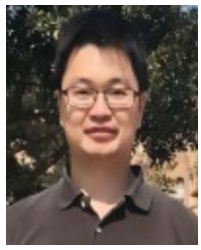
Her research interests include deep learning and hyperspectral image fusion.

Jianjun Liu (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2014, respectively.

From 2018 to 2020, he was a Postdoctoral Researcher with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. He is currently an Associate Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include the areas of hyperspectral image classification, superresolution, spectral unmixing, sparse representation, computer vision, and pattern recognition.

Jinlong Yang received the Ph.D. degree in pattern recognition and intelligence system from Xidian University, Xi'an, China, in 2012.

He is currently an Associate Professor with Jiangnan University, Wuxi, China. His research interests include target tracking, information fusion, and signal processing.



Zebin Wu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2008, respectively.

He is currently a Professor with the School of Computer Science, Nanjing University of Science and Technology. His research interests include hyperspectral image processing, high-performance computing, and computer simulation.