Spatial–Spectral Interaction Super-Resolution CNN–Mamba Network for Fusion of Satellite Hyperspectral and Multispectral Image

Guangwei Zhao, Haitao Wu, Dexiang Luo, Xu Ou, and Yu Zhang

***Abstract*—The tradeoff between spatial and spectral resolution in sensor design is inevitable, and spatial–spectral fusion aims to use low spatial resolution hyperspectral image (HSI) and high spatial resolution (HR) multispectral image (MSI) obtained at the same time and in the same area to reconstruct HR HSI. Recently, a large number of deep-learning methods have been applied in this field and achieved success. However, these methods do not fully utilize the characteristics of data for network design, and cannot guarantee effective computational efficiency in extracting local and global features. To solve the above problems, we designed a spatial–spectral interaction super-resolution convolutional neural network (CNN)–Mamba fusion network for satellite HSI and MSI, which uses mutual guidance to improve the spatial and spectral resolution of different data, and obtains the final fused image through feature fusion. In addition, we combined Mamba with CNN to effectively explore global and local features of images. Extensive experiments have proven that our method can reconstruct fused images of high quality and is superior to current state-of-the-art fusion methods.**

***Index Terms*—Convolutional neural network (CNN), fusion, hyperspectral, Mamba, multispectral.**

I. Introduction

SPECTRAL resolution of hyperspectral image (HSI), spanning from 400 to 2500 nm and boasting more than a hundred bands, has made it a powerful tool for environmental monitoring and resource exploration [1], [2], [3]. However, due to sensor design tradeoffs, such data often miss crucial spatial details (the spectral resolution of GF-5, ZY-1 02D from China, or PRISMA satellites from Italy reaches 10–20 nm, with a spatial resolution of only 30 m) [4], [5], [6]. Fortunately, other remote sensing data like multispectral image (MSI) can be acquired from different platforms, offering valuable assistance in enhancing the spatial resolution of HSI [7], [8], [9]. This process, known as spatial spectral fusion, extracts and reconstructs spectral and spatial information from both sources to generate high spatial resolution (HR) HSI [10], [11], [12].

The existing spatial–spectral fusion techniques are categorized into component substitution (CS), multi-resolution analysis (MRA), matrix decomposition (MD), and deep-learning (DL) approaches depending on their fundamental principles [13], [14], [15]. These CS methods assume that the MSI is a subspace of HSI. The forward transformation is performed to obtain the subspace of HSI, and the MSI is used to replace the subspace of HSI for the inverse transformation to improve the spatial resolution of the HSI [16], [17]. Typical methods include the intensity hue saturation (IHS) proposed by Carper et al. [18], which converts images from red–green–blue space to IHS space and uses MSI to replace the intensity component of HSI for inverse transformation to achieve image fusion. Pal et al. [19] introduced principle component analysis into image spatial–spectral fusion and obtained the fused image by replacing the first principle component of HSI with an MSI through inverse transformation. Aiazzi et al. [20] proposed adaptive Gram–Schmidt (GSA), which uses a multiple linear regression algorithm to extract the intensity component of HSI. The CS methods directly replace the component of HSI with MSI, and the fused data have obvious spatial detail information. However, component replacement alters the spectral characteristics of HSI, resulting in severe spectral distortion in fused images [21], [22]. These MRA methods assume that the information differences between images exist at multiple scales, and calculate the difference information between images at different scales through multiscale decomposition. By injecting difference information into HSI, spatial–spectral fusion is achieved [23], [24], [25]. Typical methods include modulation transfer function—generalized Laplacian pyramid (GLP) proposed by Javan et al. [26]. Wady et al. [27] introduced wavelet decomposition theory to achieve image fusion in the frequency domain, and proposed "A Ì Trous" wavelet transform. Thanks to the unchanged spectral characteristic of HSI, the MRA methods have better spectral fidelity. However, some spatial information from MSI has not been injected into the fused image, resulting in limited spatial resolution enhancement [28]. Various matrix factorization techniques, utilizing theories like mixed pixel factoring, low rank factorization, sparse representation, tensor factorization, etc., are employed for image fusion [29],

Received 11 June 2024; revised 23 July 2024 and 1 September 2024; accepted 23 September 2024. Date of publication 26 September 2024; date of current version 21 October 2024. This work was supported in part by the Henan Province Science and Technology Research Projects under Grant 242102210123 and Grant 242102211029, in part by the 2022 Henan Province Graduate Joint Training Base Project under Grant YJS2022JD45, in part by the Computer Basic Education Teaching Research Project under Grant 2024-AFCEC-393, and in part by the National Natural Science Foundation of China under Grant 12071277. *(Corresponding author: Dexiang Luo.)*

Guangwei Zhao, Haitao Wu, and Yu Zhang are with the College of Computing and Artificial Intelligence, Huanghuai University, Zhumadian 463000, China (e-mail: zgw@huanghuai.edu.cn; wht@huanghuai.edu.cn; zhangyujlu@163.com).

Dexiang Luo and Xu Ou are with the Information Centre of Guangxi Medical University, Nanning 530021, China (e-mail: ldx@gxmu.edu.cn; ox@gxmu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3469184

[30], [31], [32], [33]. They split HSI and MSI into spectral and spatial matrices before reintegrating them for fusion. For instance, Yokoya et al. [34] applied image unmixing theory to spatial spectral fusion, merging the endmember matrix of HSI with the abundance matrix of MSI. Nezhad et al. [35] proposed a nonlocal sparse representation method using pixel group theory. Dian et al. [33] developed nonlocal sparse tensor factorization (NLSTF) based on tensor decomposition theory. These methods often employ alternating direction method of multipliers or augmented Lagrangian method (ALM) for optimization, yielding satisfactory fusion results [36], [37], [38], [39]. However, numerous parameters are required, these methods exhibit limited robustness, and be computationally inefficient [40], [41], [42]. On the other hand, DL methods achieve the fusion of test datasets by training the nonlinear regression relationship between Low spatial Resolution (LR) HSI+HR MSI and HR HSI, and mapping network parameters [43], [44], [45]. Network architectures like convolutional neural network (CNN) [46], [47], [48], generative adversarial network (GAN) [49], [50], [51], and transformers [52], [53], [54] are commonly used. For example, Yang et al. [55] designed a deep CNN with two branches for HSI–MSI fusion. Wang et al. [44] proposed a novel variational probability autoencoder framework using CNN for fuse LR HSI and HR MSI. To explore long-range dependencies in the feature space, Ran et al. [56] proposed kernel space nonlocal convolution, which explores nonlocal dependencies in the generated kernel space to utilize this global information to guide the network in extracting image features more flexibly. Peng et al. [57] proposed a spatial–spectral integrated dual U-shaped network U2Net for image fusion. U2Net utilizes spatial U-Net and spectral U-Net to extract spatial details and spectral features, enabling differentiation and hierarchical learning of features from different images. It also introduces a novel spatial–spectral ensemble structure called S2Block to fuse features. Numerous GAN fusion networks have also emerged, including one proposed by Zhu et al. [58] that employs a lightweight adversarial network with quadtree implicit sampling (QIS). Zhang et al. [59] proposed a GAN-based fusion method for HSI and PAN images. Additionally, transformers have become increasingly popular in image fusion recently; Jia et al. [60] proposed a multiscale spatial–spectral transform network (MSST-Net) with two branches—one for spectral feature extraction from HSI and another for spatial feature extraction from MSI. Wang et al. [15] presented a new multilevel cross-transformer (MCT-Net) for HSI and MSI fusion. This MCT-Net comprises a multilevel cross-modal interaction module and a feature aggregation reconstruction module. In summary, DL methods have shown promising results in spatial spectral fusion of HSI and MSI [61], [62], [63]. In addition, in order to improve the computational efficiency of long-range dependency modeling, Mamba networks have recently been applied to remote sensing image processing and have achieved great success [64], [65], [66], [67].

Overall, the following conditions hold.

1) Although component replacement can significantly improve the spatial resolution of HSI, severe spectral distortion is achieved. MRA methods exhibit minimal spectral distortion, but limited spatial enhancement. MD methods have a large number of parameters, poor computational efficiency, and insufficient robustness.
2) DL has recently surpassed traditional methods in image fusion and become a focus in this field. However, current DL spatial spectral fusion methods face challenges, such as when using MSI data to assist HSI for spatial super-resolution (SpaR), significant data differences can lead to damage to the spatial spectral information in the fused image. In addition, the receptive field of convolutional fusion networks is limited and cannot consider the global information correlation in remote sensing images. Although transformers can extract global feature information, their computational efficiency is still a concern.

To tackle these issues, we propose spatial–spectral interaction super-resolution CNN–Mamba fusion network for satellite HSI and MSI (SSRFN). To mitigate the tradeoff between spatial and spectral information due to data differences, we initially downsampled HR MS, and extracted the shallow and deep feature information of LR HSI, LR MSI, and HR MSI. To correct the channel relationship of HR MSI, we propose a change feature extraction module to compute the error information between LR HSI and LR MSI, then use this error information to refine the features of HR MSI in the spectral correction module (SCM). Subsequently, we employ features of LR HSI to guide feature extraction of HR MSI for spectral super-resolution (SpeR), and features of HR MSI to guide feature extraction of LR HSI for SpaR. By focusing on different aspects of super-resolution (i.e., SpeR prioritizes spatial information, while SpaR ensures spectral information), we construct a feature fusion module (FFM) to merge and reconstruct super-resolution results, yielding the final HR HSI with superior precision. Furthermore, to effectively extract local and global features from remote sensing images for precise reconstruction of HR HSI, we integrate Mamba network with CNN, offering satisfactory inference speed and efficient feature extraction. Our main contributions include the following.

1) SSRFN is the first to combine Mamba with CNN for HSI and MSI fusion, effectively achieving the extraction of local and global features. In addition, SpeR and SpaR networks are designed to interactively reconstruct fused images, avoiding spatial spectral distortion caused by a single network.
2) SSRFN proposes an effective difference information extraction module and devises a suitable SCM. Besides, SpaR–SpeR networks are proposed. Moreover, the design of the FFM module not only eliminates redundant feature information but also significantly enhances the spectral and spatial fidelity of fusion results.
3) SSRFN demonstrates comparable fusion performance across simulated and real datasets, and outperforms eight state-of-the-art methods in spatial and spectral fidelity.

The rest of this article is organized as follows. Section II presents our proposed SSRFN in detail. Section III analysis the experimental results on simulated and real datasets. Finally, Section IV concludes this article.

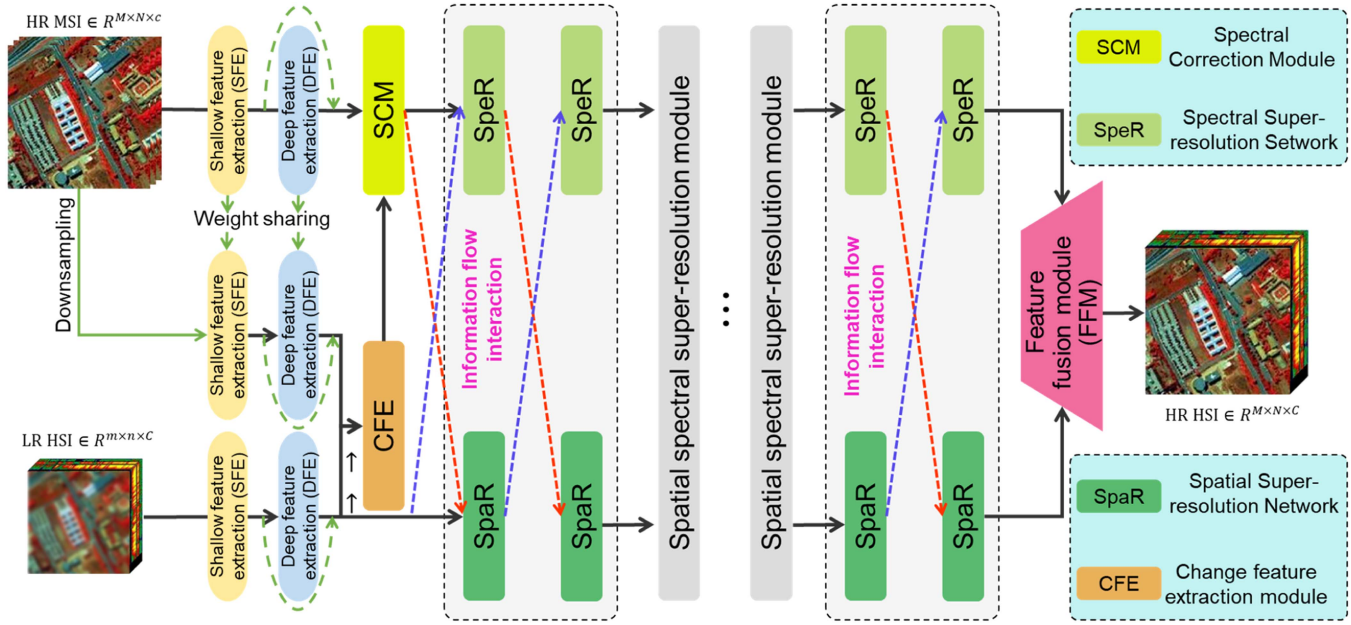

Fig. 1. SSRFN (the size of HSI is $m \times n \times C$, and the size of MSI is $M \times N \times c$, where $M > m$, $N > n$, and $C > c$).

II. Methodology

A. Mamba

State space models (SSMs) have emerged as a competitive backbone network in DL, demonstrating the ability to model long-range dependencies with a linear expansion in sequence length and showing significant potential in image processing. Mamba relies on classical continuous systems, mapping a one-dimensional (1-D) input function or sequence [denoted as $x(t)$] to an output $y(t)$ via an intermediate hidden state $h(t)$. SSMs can be represented by the following linear ordinary differential equations:

$$\begin{aligned} h'(t) &= \boldsymbol{\alpha} h(t) + \boldsymbol{\beta} x(t) \\ y(t) &= \boldsymbol{\phi} h(t) + \boldsymbol{\psi} x(t) \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha}$ represents the state matrix, while $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ denote the projection parameters. Subsequently, a discretization process is typically applied in practical DL algorithms. Specifically, let Δ represent the time-scale parameter for converting continuous parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to their discrete counterparts $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$. A commonly used discretization method is the zero-order hold rule, which is defined as follows:

$$\begin{aligned} \bar{\boldsymbol{\alpha}} &= \exp(\Delta \boldsymbol{\alpha}) \\ \bar{\boldsymbol{\beta}} &= (\Delta \boldsymbol{\alpha})^{-1} (\exp(\Delta \boldsymbol{\alpha}) - \mathbf{I}) \cdot \Delta \boldsymbol{\beta}. \end{aligned} \tag{2}$$

After discretization, the discrete version of (1) with step size Δ can be rewritten in the following recurrent neural network form

$$\begin{aligned} h_k &= \bar{\boldsymbol{\alpha}}\, h_{k-1} + \bar{\boldsymbol{\beta}} x_k \\ y_k &= \boldsymbol{\phi} h_k + \boldsymbol{\psi} x_k. \end{aligned} \tag{3}$$

Additionally, (3) can also be mathematically equivalently transformed into the following CNN form:

$$\begin{aligned} \bar{\mathbf{K}} &\triangleq \left(\boldsymbol{\phi}\bar{\boldsymbol{\beta}}, \boldsymbol{\phi}\overline{\boldsymbol{\alpha}\boldsymbol{\beta}}, \cdots, \boldsymbol{\phi}\bar{\boldsymbol{\alpha}}^{L-1}\bar{\boldsymbol{\beta}}\right) \\ \mathbf{y} &= \mathbf{x} \otimes \bar{\mathbf{K}} \end{aligned} \tag{4}$$

where $\otimes$ denotes the convolution operation, $\mathbf{K}$ is a structured convolutional kernel, and L represents the length of the input sequence $\mathbf{x}$.

B. Overview of SSRFN

Employing MSI as auxiliary information for SpaR of HSI results in a compromise between spatial and spectral information, potentially leading to a reduction of some spatial information in merged images. Furthermore, SpeR does not guarantee accurate transfer of band relationships, resulting in severe spectral distortion. Given these considerations, we propose a spatial–spectral interaction super-resolution CNN–Mamba network for the fusion of HSI and MSI. Fig. 1 illustrates the network architecture of SSRFN. To acquire sufficient prior feature information for precise reconstruction of HR HSI, SSRFN initially executes shallow feature extraction (SFE) and deep feature extraction (DFE) on HR MSI and LR HSI to standardize the number of channels. Specifically, HR MSI is downsampled and the feature information of LR MSI is extracted by using the feature extraction weights of HR MSI. To ensure that the features of HR MSI and LR HSI maintain similar channel relationships, i.e., band relationships, we designed a CFE to extract residual information between LR MSI and LR HSI features, and constructed an SCM to utilize residual information to guide HR MSI features for channel correction. Subsequently, we designed a SpaR network and enhanced the spatial resolution of LR HSI features using modified HR MSI features. Additionally, an SpeR

was constructed, and we used LR HSI features to guide HR MSI features to establish accurate channel relationships. Finally, we constructed a feature fusion network to fuse SpeR and SpaR features to yield the final HR HSI.

C. Network Architecture

1) Shallow Feature Extraction Module (SFE) and Deep Feature Extraction Module (DFE): Remote sensing image possess substantial feature information. To extract comprehensive features and precisely reconstruct the fused results, we established the SFE and DFE modules. SFE can retrieve more minute feature details, ensuring that the network can apprehend more HR detail information. Moreover, as the number of convolutions increases, DFE can ensure that the network extracts extensive LR semantic information.

Specifically, SFE is a residual network composed of four convolutional layers with 3×3 kernel. In the experiment, HR MSI and LR HSI are initially fed into SFE to obtain shallow features, and the feature extraction parameters of HR MSI are shared with LR MSI to procure the feature information of LR MSI. The formula can be expressed as follows:

$$\mathbf{F}_h = \mathrm{Conv}_h^i\left(\mathrm{Conv}_h^{i-1}\left(\mathbf{X}\right)\right) + \mathbf{X},\ i = 1, \cdots 4 \tag{5}$$

$$\mathbf{F}_m = \mathrm{Conv}_m^i\left(\mathrm{Conv}_m^{i-1}\left(\mathbf{Y}\right)\right) + \mathbf{Y},\ i = 1, \cdots 4 \tag{6}$$

$$\mathbf{F}_{lm} = \mathrm{Conv}_m^i\left(\mathrm{Conv}_m^{i-1}\left(\tilde{\mathbf{Y}}\right)\right) + \tilde{\mathbf{Y}},\ i = 1, \cdots 4 \tag{7}$$

where $\mathbf{X}$ is LR HSI, $\mathbf{Y}$ represents HR MSI, $\tilde{\mathbf{Y}}$ is the downsampled $\mathbf{Y}$, $\mathbf{F}_h$, $\mathbf{F}_m$, and $\mathbf{F}_{lm}$ are the features of LR HSI, HR MSI, and LR MSI, respectively, and i represents the ith convolutional layer.

Recently, Mamba networks have gained substantial adoption in image processing and have shown superior performance, attributable to their advantages compared to transformers [68]. The Mamba network can execute rapid inference and its performance escalates linearly with the increase of sequence length. The universal Mamba block can be expressed as $\rho(\mathrm{ssm}(\sigma(\mathrm{Conv}(\rho(\mathbf{I}))))) \otimes (\sigma(\rho(\mathbf{I}))))$, where $\mathbf{I}$ represents the input data, ρ is a multilayer perceptron, σ denotes the sigmoid activation function, ssm is a structured SSM. Nevertheless, Mamba has not been extensively examined in remote sensing image super-resolution. Given that the information in remote sensing images possesses global relevance, the Mamba network can proficiently extract the global features of images. In DFE, we employed the Mamba network to extract global information of images, and considering the local similarity of remote sensing images, we integrated a 3×3 convolutional layer to extract the local information of images. It is noteworthy that convolutional layers and Mamba networks are not two distinct extraction branches, but rather interactive extraction of local and global information. As illustrated in Fig. 2, DFE comprises 15 convolutional layers and 5 Mamba blocks, with a Mamba network interspersed every 3 convolutional layers extracting global features based on local image features and transmitting them downstream. Specific formulas can be represented as

$$F_{du}^j = \mathrm{Conv}_3\left(\mathrm{Conv}_2\left(\mathrm{Conv}_1\left(u\right)\right) + \left(\mathrm{Mamba}\left(u\right)\right)\right) \tag{8}$$

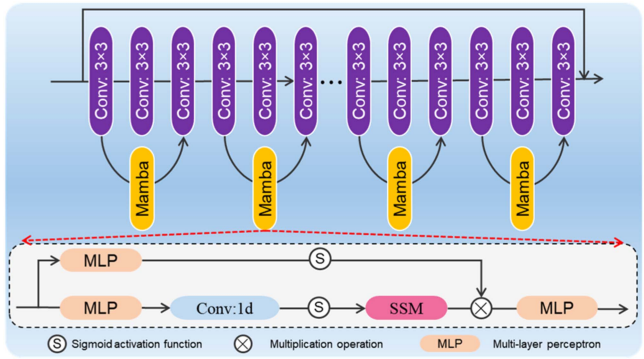

Fig. 2. DFE.

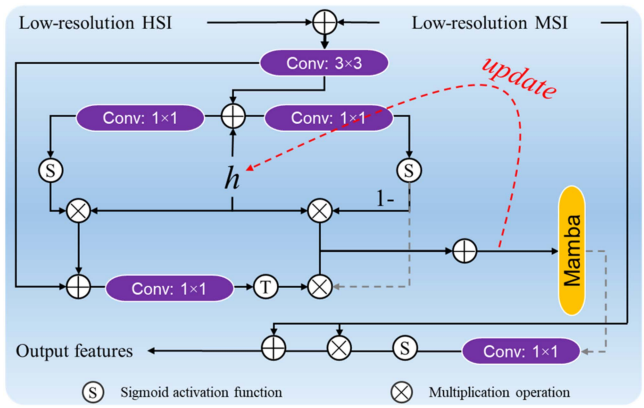

Fig. 3. CFE.

where $u = [\mathbf{F}_h,\ \mathbf{F}_m,\ \mathbf{F}_{lm}]$ is the input data, $\mathbf{F}_{du}$ represents the deep features of the output, and j represents the jth convolutional Mamba block. It is worth noting that the SSM state expansion factor in the Mamba block is set to 64, the local convolution width is 4, and the block expansion factor is set to 2.

2) Change Feature Extraction Module (CFE): Upon extracting both shallow and profound features, although $\mathbf{F}_{dm}$ shows an identical channel count as $\mathbf{F}_{dh}$, inconsistent channel relationships will induce significant SpeR errors. To rectify the channel relationship of $\mathbf{F}_{dm}$, we engineered CFE to extract the error information $\mathbf{E}$ between features using $\mathbf{F}_{dh}$ and $\mathbf{F}_{dlm}$ as input data. As illustrated in Fig. 3, CFE employs the gated recurrent unit (GRU) framework, which adaptively erases similarity information between features via gating operations and preserves difference information between features [69]. Unlike conventional GRU, in this experiment, we devised a cyclic GRU, which continually adjusts network parameters to optimize the ultimate difference information by reinputting each output as a new state h. In addition, we incorporated channel stacking to $\mathbf{F}_{dh}$ and $\mathbf{F}_{dlm}$ as input data and set the number of cyclic iterations to 6. It is worth noting that in CFE, we embedded a Mamba module during the operation process to explore the global correlation of features based on local features. The specific mathematical expression is as follows:

$$P = \sigma\left(\mathcal{T} \odot w_{xp}^i + H^i \otimes w_{hp}^i + b_p^i\right) \tag{9}$$

$$O^i = \sigma\left(\mathcal{T} \otimes w_{xq}^i + H^i \otimes w_{hq}^i + b_q^i\right) \tag{10}$$

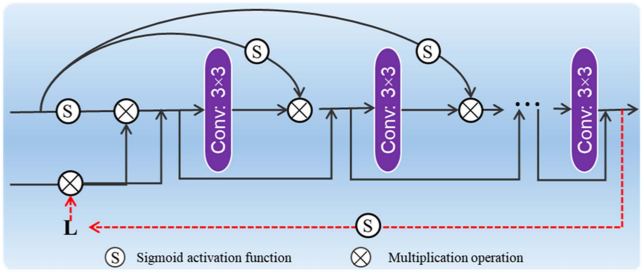

Fig. 4. SCM.

$$\tilde{H}^i = \tan h\left(\mathcal{T} \otimes w_{xh}^i + \left(P^i \odot H^i\right) \otimes w_{hh}^i + b_h^i\right) \tag{11}$$

$$H^i = O^i \odot \tilde{H}^i + \left(1 - O^i\right) \odot \tilde{H}^i \tag{12}$$

where $\mathcal{T}$ represents the input data, P denotes the output of the reset gate, O signifies the output of the update gate, H represents the hidden layer, tan$h$ represents tan$h$ activation function, w denotes the weight, b denotes the bias, $\otimes$ represents the convolution operation, and $\odot$ represents multiplication.

3) Spectral Correction Module (SCM): It is paramount to employ error information to correct channel correlation of $\mathbf{F}_{dm}$. To accurately transmit channel correlations, we designed SCM in the experiment. As depicted in Fig. 4, SCM initially triggers $\mathbf{E}$ via a sigmoid function and multiplies it with $\mathbf{F}_{dm}$ to augment error information in $\mathbf{F}_{dm}$. Subsequently, employing skip connections, the enhanced F_{dm} is added into the original $\mathbf{F}_{dm}$, and a 3×3 convolution is executed to eliminate error information in $\mathbf{F}_{dm}$. The aforementioned process underwent 16 identical correction operations to attain the ultimate optimized $\mathbf{F}_{dm}$. It is noteworthy that our SCM is a recurrent network that utilizes output features to refine the input $\mathbf{F}_{dm}$ for subsequent network parameter update after each output feature is acquired. Specifically, in the experiment, SCM implemented an initialized identity matrix $\mathbf{L}$, and $\mathbf{L}$ was updated for each iteration using the output features attained from the preceding iteration. Specific formulas can be expressed as

$$\mathbf{F}_{dm}^{op} = \text{Conv}\left(\sigma_i^j\left(\mathbf{E}\right) \odot \left(v^j \odot \mathbf{F}_{dm_i}^j\right) + \mathbf{F}_{dm_i}^j\right) \tag{13}$$

where $\mathbf{F}_{dm}^{op}$ represents the final channel corrected $\mathbf{F}_{dm}$, i represents the ith correction operation, and j represents the jth iteration.

4) Spatial Super-Resolution Network (SpaR) and Spectral Super-Resolution Setwork (SpeR): After obtaining the features $\mathbf{F}_{dh}$ and $\mathbf{F}_{dm}^{op}$ of LR HSI and HR MSI, we designed SpaR and SpeR for SpaR of $\mathbf{F}_{dh}$ and SpeR of $\mathbf{F}_{dm}^{op}$, respectively. Fig. 5 shows SpaR. SpaR first performed channel stacking on $\mathbf{F}_{dh}$ and $\mathbf{F}_{dm}^{op}$, and used 3-D convolution with $3 \times 3 \times 1$ convolution kernels and Mamba networks to construct local and global feature extraction modules, respectively. This can be represented as

$$\mathbf{F}_{\text{local}} = \text{3D Conv}\left(\text{cat}\left(\mathbf{F}_{dh}, \mathbf{F}_{dm}^{op}\right)\right) \tag{14}$$

$$\mathbf{F}_{\text{global}} = \text{Mamba}\left(\text{cat}\left(\mathbf{F}_{dh}, \mathbf{F}_{dm}^{op}\right)\right) \tag{15}$$

where $\mathbf{F}_{\text{local}}$ represents a local feature and $\mathbf{F}_{\text{global}}$ is a global feature. To reduce informational redundancy between $\mathbf{F}_{\text{local}}$ and $\mathbf{F}_{\text{global}}$, we implemented a sigmoid activation function for $\mathbf{F}_{\text{local}}$ and $\mathbf{F}_{\text{global}}$ separately, then multiplied the sigmoid parameters of $\mathbf{F}_{\text{local}}$ with $\mathbf{F}_{\text{global}}$ to extract the unique information of $\mathbf{F}_{\text{global}}$, which was subsequently added to $\mathbf{F}_{\text{local}}$. Similarly, the sigmoid parameter of $\mathbf{F}_{\text{global}}$ was multiplied by $\mathbf{F}_{\text{local}}$ to extract the unique information of $\mathbf{F}_{\text{local}}$, which was subsequently added to $\mathbf{F}_{\text{global}}$. This process can be expressed as

$$\mathbf{F}_{f\text{local}} = \sigma\left(\mathbf{F}_{\text{local}}\right) \odot \mathbf{F}_{\text{global}} + \mathbf{F}_{\text{local}} \tag{16}$$

$$\mathbf{F}_{f\text{global}} = \sigma\left(\mathbf{F}_{\text{global}}\right) \odot \mathbf{F}_{\text{local}} + \mathbf{F}_{\text{global}} \tag{17}$$

where $\mathbf{F}_{f\text{local}}$ is the local feature after SpaR and $\mathbf{F}_{f\text{global}}$ represents the global features after SpaR. Finally, after four stages of global and local feature fusion, a $1 \times 1 \times 1$ convolutional layer was used to obtain the final SpaR feature.

It is worth noting that, as shown in Fig. 6, SpeR and SpaR have the same network structure. However, in order to perform SpeR, SpeR used a $1 \times 1 \times 3$ convolution kernel to obtain the final SpeR features.

5) Feature Fusion Module (FFM): As shown in Fig. 7, after obtaining the SpaR features of LR HSI and the SpeR features of HR MSI, we constructed an FFM to fuse $\mathbf{F}_{\text{spa}}$ and $\mathbf{F}_{\text{spe}}$ to obtain the final HR HSI. Specifically, to explore the global correlation between features and reduce information redundancy, FFM first constructed a Mamba cross-attention network to fuse $\mathbf{F}_{\text{spa}}$ and $\mathbf{F}_{\text{spe}}$, and then constructed a Mamba self-attention network to reduce the dimensionality of the fused features and obtain the final HR HSI.

For cross-attention networks, $\mathbf{F}_{\text{spa}}$ was first used to extract features $\mathbf{Q}$ and $\mathbf{K}$ from two Mamba branches, respectively, $\mathbf{F}_{\text{spe}}$ was used to extract feature $\mathbf{V}$ from a Mamba branch, which can be represented as follows:

$$\begin{aligned} \mathbf{Q} &= \text{Mamba}\left(\mathbf{F}_{\text{spa}}\right) \\ \mathbf{K} &= \text{Mamba}\left(\mathbf{F}_{\text{spa}}\right) \\ \mathbf{V} &= \text{Mamba}\left(\mathbf{F}_{\text{spe}}\right). \end{aligned} \tag{18}$$

Then, $\mathbf{K}$ and $\mathbf{V}$ are dot products and subtracted from 1 to obtain attention matrices, and unique feature information of $\mathbf{F}_{\text{spa}}$ is obtained by multiplying the attention matrix with $\mathbf{Q}$. The final fusion feature was obtained by adding the unique feature information of $\mathbf{F}_{\text{spa}}$ into $\mathbf{V}$

$$\mathbf{F}_{\text{hrhsi}} = \left(1 - \text{Softmax}\left(\mathbf{K} \odot \mathbf{V}\right)\right) \odot \mathbf{Q} + \mathbf{V} \tag{19}$$

where $\mathbf{F}_{\text{hrhsi}}$ represents the fused features. However, compared to HR HIS, $\mathbf{F}_{\text{hrhsi}}$ has more channels. We constructed a self-attention mechanism, first extracting three branch Mamba features from $\mathbf{F}_{\text{hrhsi}}$, adjusting the feature channels to calculate the conversion relationship between channels, and finally obtaining the fused image

$$\begin{aligned} \mathbf{A} &= \text{Mamba}\left(\mathbf{F}_{\text{spa}}\right) \\ \mathbf{B} &= \text{Mamba}\left(\mathbf{F}_{\text{spa}}\right) \\ \mathbf{D} &= \text{Mamba}\left(\mathbf{F}_{\text{spe}}\right) \end{aligned} \tag{20}$$

$$\mathbf{Z} = \left(1 - \text{Softmax}\left(\mathbf{B} \odot \mathbf{D}\right)\right) \odot \mathbf{A} + \mathbf{D} \tag{21}$$

where $\mathbf{Z}$ is the final fused image.

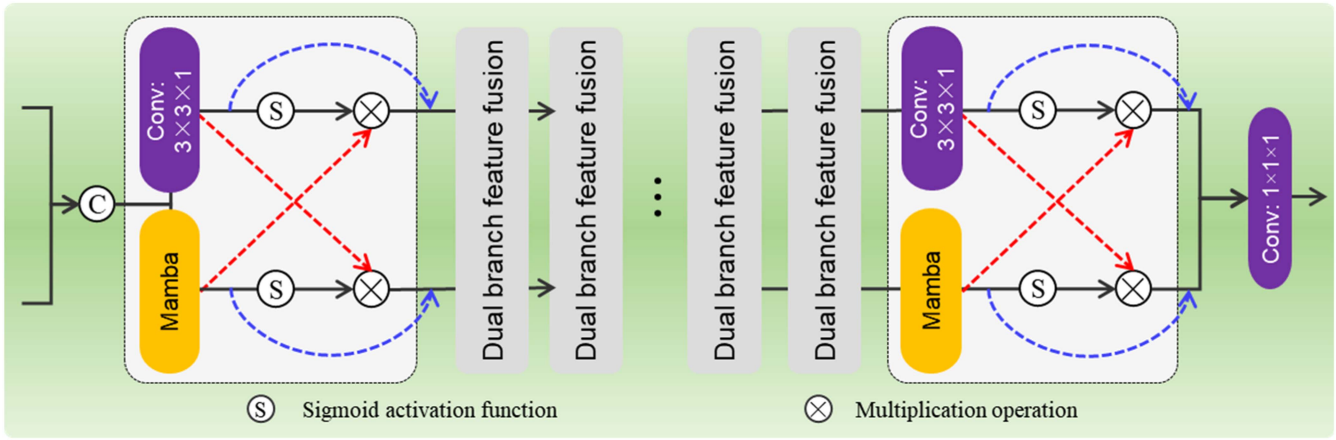

Fig. 5. SpaR.

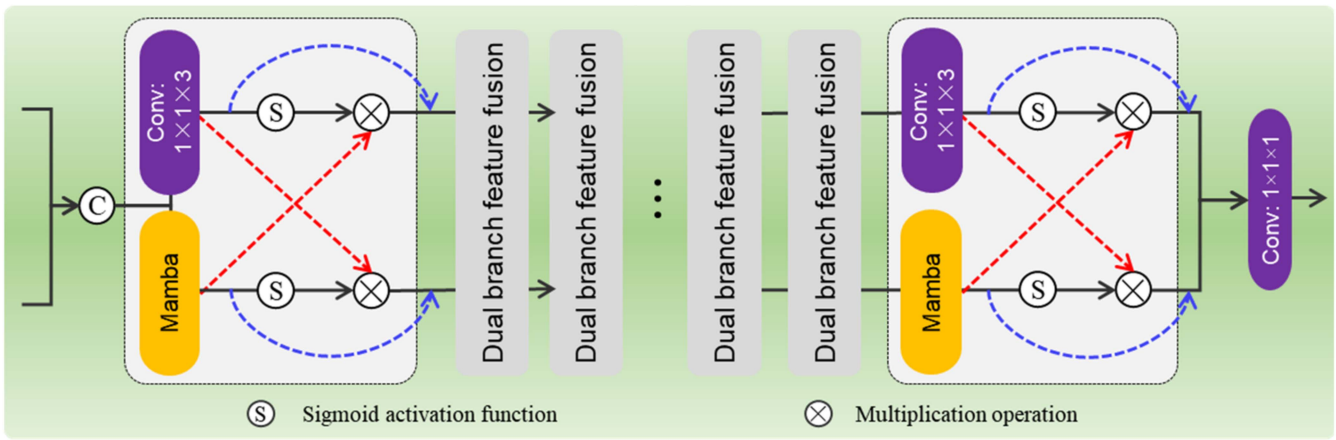

Fig. 6. SpeR.

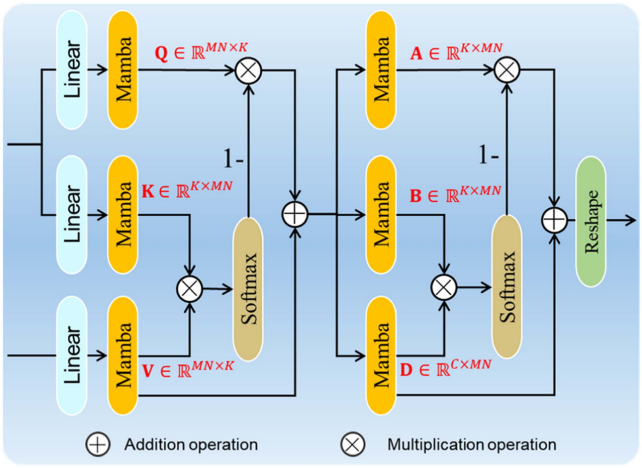

Fig. 7. FFM.

III. Experiments

A. Experimental Setting

To train SSRFN, Adam optimizers with $\alpha = 0.9$ and $\beta = 0.99$ and a batch size of 8 were used, and the learning rate was set to 1e−4. In addition, SSRFN is trained on a Linux computer with 128-GB CPU memory and 1 × GPU GeForce GTX 4090D. SSRFN uses the MSE loss equation for network optimization, and the loss function can be expressed as loss $= \sum_{j=1}^{n} (\mathbf{Z} - \mathbf{GT})^2$, where $\mathbf{GT}$ is the ground truth and n is the number of pixels.

To verify the performance of SSRFN, two simulated and real HSI datasets were used. The HSI of two of the simulated datasets comes from the Chikusei dataset (https://naotoyokoya.com/Download.html) and the Pavia Center dataset (http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_University_scene). The Chikusei dataset was captured by the Headwall Hyperspec-VNIR-C sensor, with a spectral range of 343–1018 nm and a pixel count of 2517 × 2335 × 128. The final 2000 × 2000 × 100 pixels were used for the experiment. The Pavia Center dataset was obtained by ROSIS sensors, with a spectral range of 400–1000 nm and an image size of 512 × 1400 × 115 pixels. The final 512 × 1400 × 100 pixels met the experimental requirements. In order to train SSRFN and quantitatively evaluate the fusion results, the above datasets were used as a reference image in the experiment, and the spectral response function of SPOT-4 multispectral satellite was used to simulate MSI. In addition, Gaussian blur operation and a downsampling factor of 4 are used

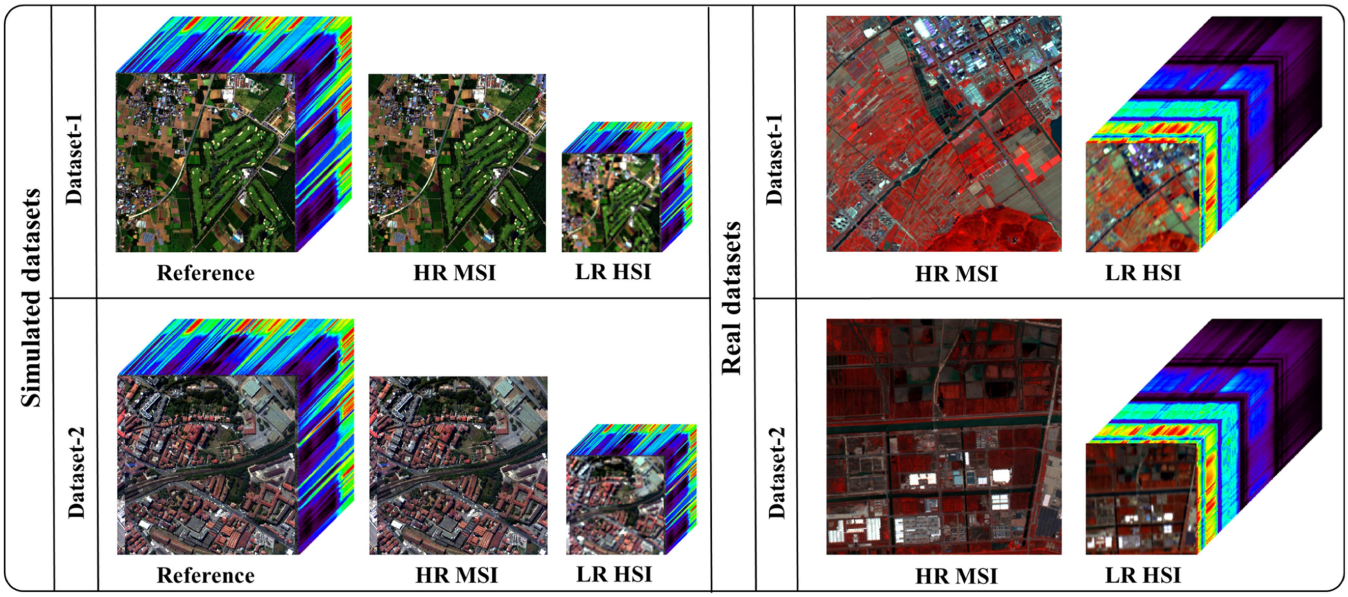

Fig. 8. Experimental datasets.

to simulate HSI. For the above datasets, we cut 512 × 512 × 100 pixels from the Chikusei dataset and 512 × 512 × 100 pixels from the Pavia Center dataset as test samples to verify the fusion performance of SSRFN. The size of the reference training sample, reference validation sample, and reference test sample in the experiment is 512 × 512 × 100 pixels. In order to expand the sample size, a spatial step of 80 was used in the image cropping process, which ensured that the sample size in the experiment was greater than 1000. In the end, 1300 samples from simulated dataset-1 were used for training and 236 samples were used for validation. A total of 1100 samples from simulated dataset-2 were used for training, and 104 samples were used for validation. In addition, two real datasets were used to verify the real application performance of SSRFN. The HSI of these two datasets were obtained from PRISMA hyperspectral satellite sensors, with a spatial resolution of 30 m and a spectral range of 400–2500 nm. After removing the bad bands, a total of 100 bands were used for experiments. The corresponding MSI at that time were obtained from Sentinel-2 A satellite, including four bands with a spatial resolution of 10 m. In the experiment, we trained SSRFN on real datasets using a reduced resolution dataset, and tested it on a full resolution dataset (see Fig. 8).

To evaluate the fusion performance of SSRFN, eight different traditional and DL methods were selected, including CS method: GSA [20], MRA method: GLP [26], MD methods: coupled nonnegative matrix factorization (CNMF) [34], NLSTF [33], and DL methods: spatial–spectral reconstruction network (SSRNET) [70], MSST-Net [60], unsupervised hybrid network of transformer and CNN (UHNTC) [71], multi-input multioutput spatial–spectral transformer (MIMO–SST) [72], Fusionmamba (FMamba) [73], and QIS–GAN [58]. It is worth noting that in the DL method, SSRNET is designed based on CNN network, MSST Net, UHNTC, and MIMO–SST are designed based on transformer network, FMamba is designed based on Mamba network, and QIS–GAN is designed based on GAN network. In order to fairly compare these methods, the same training strategy was used, and the epoch of all methods was set to 200. Other parameters were sourced from relevant references.

In addition, peak signal-to-noise ratio (PSNR) [2], spectral angle map (SAM) [10], Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [23], root mean square error (RMSE) [40], and cross-correlation (CC) [71] were used as quantitative evaluation indicators for the experiment.

B. Experimental Results of Simulated Datasets

Fig. 9 shows the experimental results of the simulated dataset-1. The first row shows the fusion results of different methods, the second row shows a locally enlarged image of the fusion results, and the third row shows SAM maps of the fusion results. GSA and FMamba have good spatial enhancement effect; however, significant spectral distortion occurs in bare areas. GLP produces the worst visualization result, with all spatial information lost. NLSTF and MSST-Net enhance spatial information, but introduce significant noise signals. Other methods have better visualization results, and our method achieves the closest result to the reference image. For SAM maps, GSA, GLP, NLSTF, and MSST-Net exhibit significant spectral errors, the spectral distortion of CNMF mainly exists in the building area, SSRNET, UHNTC, MIMO–ST, FMamba, QIS–GAN, and SSRFN exhibit spectral distortion in water extraction, but our method achieves minimal spectral reconstruction error. Fig. 10 shows the fusion results of the simulated dataset-2. Compared with the reference image, GSA has significant visual errors and severe spectral distortion. GLP and NLSTF exhibit blurring effects in the edge regions of the features. Other methods have visualization results that are consistent with the reference image. However, SAM maps show GSA, NLSTF, and SSRNET exhibit significant

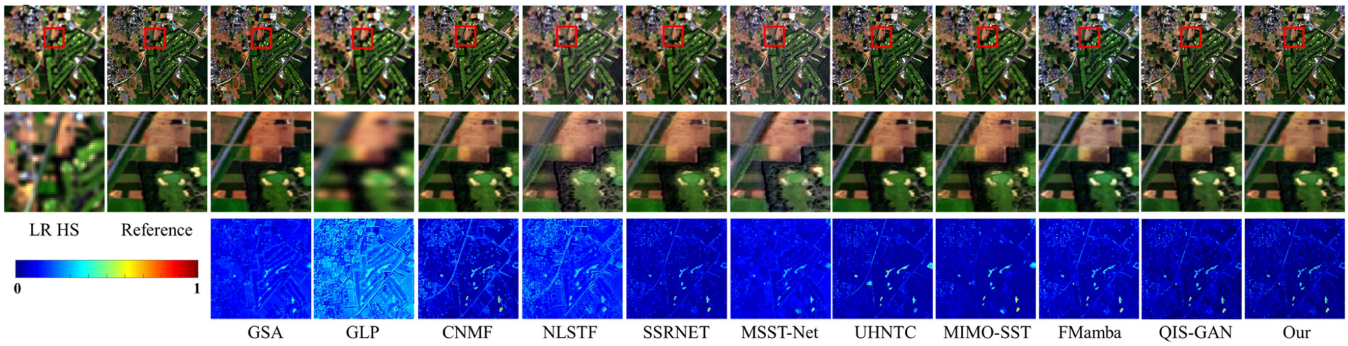

Fig. 9. Experimental results of simulated dataset-1.

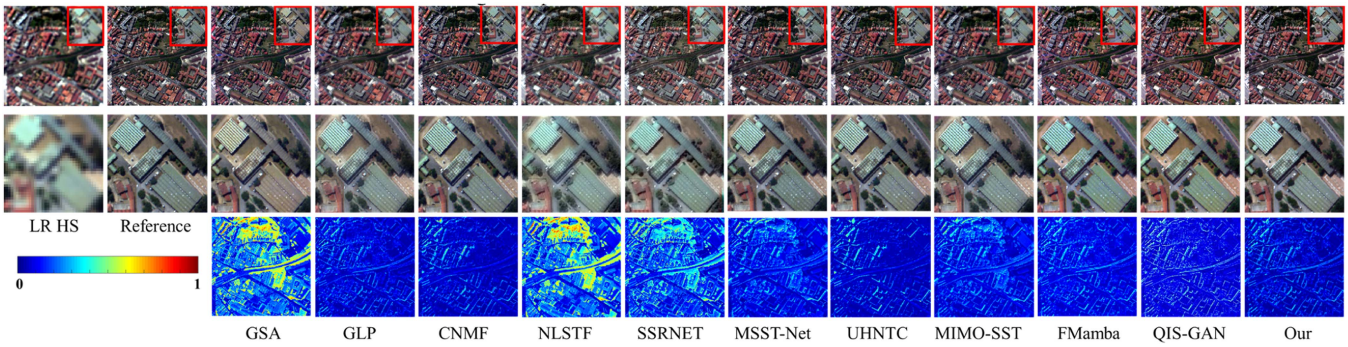

Fig. 10. Experimental results of simulated dataset-2.

TABLE I
QUANTITATIVE RESULTS OF SIMULATED DATASETS

Methods	GSA	GLP	CNMF	NLSTF	SSRNET	MSST-Net	UHNTC	MIMO-SST	FMamba	QIS-GAN	Our
Simulated dataset-1											
PSNR	18.55	18.53	21.37	19.95	25.37	22.66	25.76	28.78	27.61	28.13	29.86
SAM	8.77	13.27	5.11	7.15	2.88	4.48	2.56	2.89	3.11	2.93	2.53
ERGAS	11.23	10.48	8.02	9.56	7.45	7.06	4.22	3.96	4.72	4.38	3.88
CC	0.904	0.871	0.977	0.988	0.983	0.968	0.985	0.968	0.953	0.977	0.988
RMSE	0.103	0.089	0.081	0.088	0.046	0.062	0.062	0.039	0.041	0.037	0.032
Simulated dataset-2											
PSNR	12.93	25.69	26.78	13.70	15.28	25.01	27.80	26.97	28.67	28.23	30.71
SAM	15.61	4.40	4.82	15.38	10.36	7.09	4.90	7.12	4.75	5.32	3.07
ERGAS	14.72	33.21	11.25	13.88	12.30	6.81	7.08	5.99	6.60	6.97	4.59
CC	0.934	0.976	0.983	0.921	0.968	0.983	0.989	0.985	0.089	0.983	0.992
RMSE	0.188	0.097	0.043	0.191	0.138	0.050	0.036	0.039	0.033	0.037	0.028

spectral distortion in vegetation regions, MSST-Net, MIMO–SST, FMamba, and QIS–GAN exhibit spectral reconstruction errors in edge regions of objects, while other methods achieve better spectral fidelity.

In order to fairly compare different methods, Table I presents the quantitative evaluation results of the fusion results of different methods. In simulation dataset-1, PSNR of GSA, GLP, and NLSTF is less than 20, ERGAS is greater than 9, SAM of GSA, GLP, CNMF, NLSTF, and MSST-Net is greater than 4, and GLP achieves the worst CC. The RMSE of GSA, GLP, CNMF, and NLSTF is greater than 0.08. Our method achieves the best results among all indicators. In dataset -2, GSA, NLSTF, and SSRNET achieve the worst PSNR and SAM, with PSNR less than 20, SAM is greater than 10. The ERGAS of GSA, GLP, CNMF, NLSTF, and SSRNET is greater than 10. GSA and NLSTF achieve the worst CC and RMSE. SSRFN has consistently maintained the best results among all indicators. Fig. 11 shows the PSNR and CC of each band in the fusion results, and our method obtains the best results in almost all bands in both datasets.

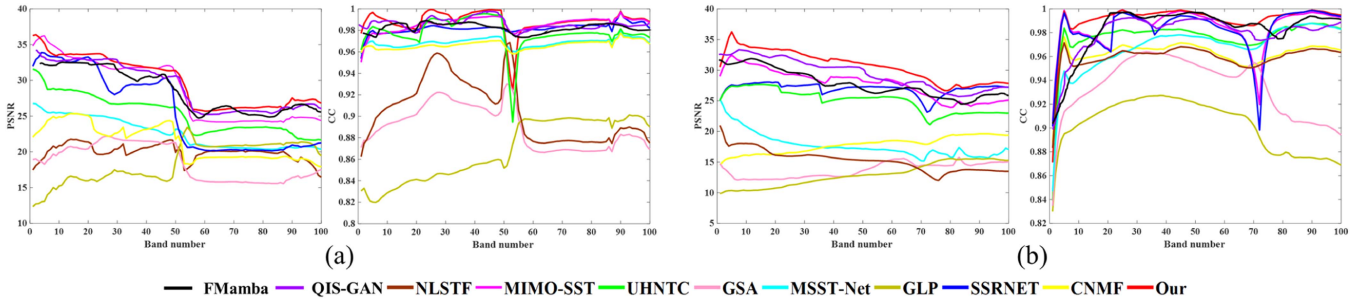

Fig. 11. PSNR and CC of each band in the fusion results.

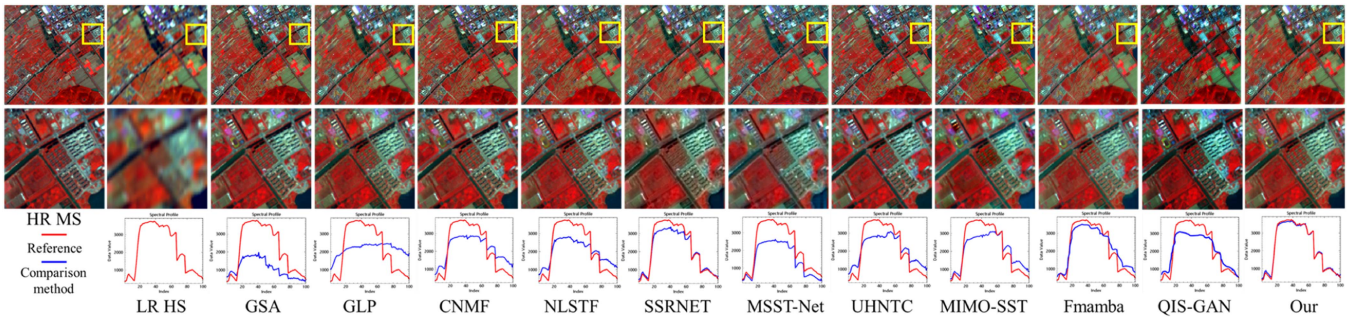

Fig. 12. Experimental results of real dataset-1.

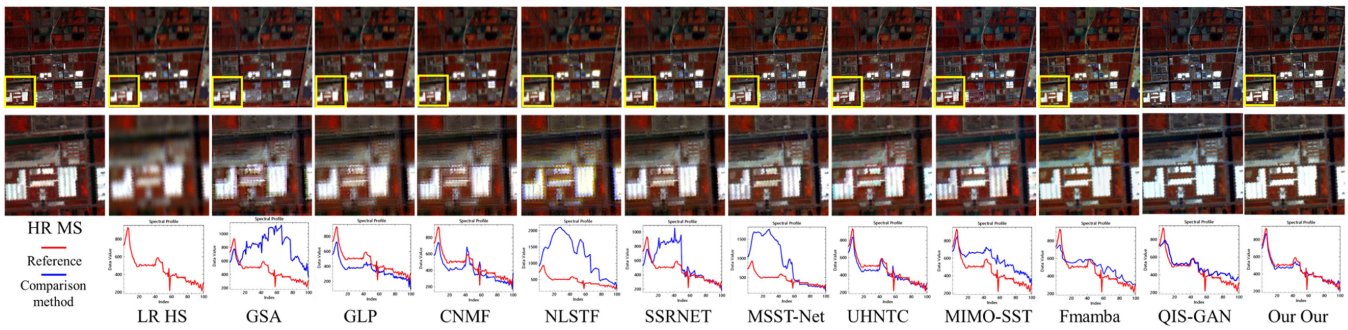

Fig. 13. Experimental results of real dataset-2.

C. Experimental Results of Real Datasets

Fig. 12 shows the experimental results of different methods in the real dataset-1. The first row shows the fusion results of different methods, the second row shows a locally enlarged image of the fusion results, and the third row shows the spectral curve of typical features in the fusion results. All methods achieve better spectral reconstruction results, but there are significant differences in spatial information enhancement. SSRNET and MSST-Net exhibit significant spectral distortion in vegetation regions, with a large amount of detailed information smoothed out. NLSTF loss some detailed information at the edges of features. Due to the lack of reference images, we selected pure pixels that are common to LR HSI and fused images, and compare the spectral curves of these pixels. GLP has severe spectral reconstruction errors, which are completely inconsistent with the trend of the reference curve. The peak and valley values of other methods have significant reconstruction errors, but have a trend consistent with the reference curve. Our method achieves the most consistent results with the reference curve. Fig. 13 shows the fusion results of different methods in real dataset-2. GSA, GLP, CNMF, and NLSTF loss some spatial information at the edges of features, especially for buildings. Compared to other DL methods, SSRNET has the worst spatial enhancement effect. MSST-Net, UHNTC, MIMO–ST, UHNTC, FMamba, QIS–GAN, and SSRFN have good visualization results. In addition, the comparison results of spectral curves show GSA, NLSTF, SSRNET, and MSST-Net exhibit significant spectral errors, with significant errors in both trend and peak values compared to the reference curve. Other benchmark methods obtain better spectral reconstruction results; however, some peak errors exist. Our method still achieves the most consistent results with the reference curve.

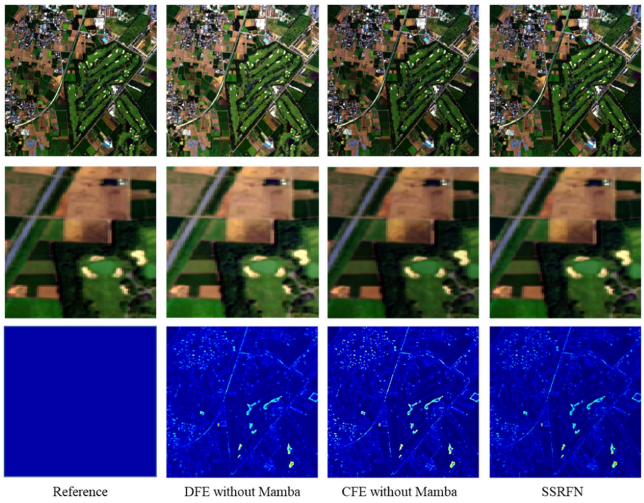

Fig. 14. Mamba ablation experiments in DFE and CFE.

TABLE II
QUANTITATIVE RESULTS OF ABLATION EXPERIMENT OF MAMBA IN DFE AND CEF

Methods	DFE without Mamba	CFE without Mamba	SSRFN
PSNR	27.11	28.91	29.86
SAM	2.90	2.59	2.53

TABLE III
QUANTITATIVE RESULTS OF ABLATION EXPERIMENT OF CNN AND MAMBA

Methods	CNN	Mamba	CNN+Mamba
PSNR	22.11	21.65	29.86
SAM	11.73	14.21	2.53

D. Discussion

To verify the important role of Mamba in DFE and CFE, we performed ablation on Mamba blocks in DFE and CFE, and the visualization results are shown in Fig. 14. When the Mamba module of DFE is ablated, spatial information can still be significantly improved, but some color distortion exists. Moreover, when the Mamba module of CFE is ablated, the fusion result is very similar to SSRFN. Table II shows the quantitative evaluation results, where CFE without Mamba module caused a slight decrease in network performance, and DFE without Mamba module had a significant impact on network performance.

During this experimental process, we integrated CNN with the Mamba network to ascertain the local and global features of the image. To validate the feasibility of this integration of CNN and Mamba, we executed ablation procedures on CNN and Mamba separately. Fig. 15 and Table III exhibit the fusion results of simulated dataset-1. The fused output of CNN outperforms that of the Mamba network markedly, and the amalgamation of CNN and Mamba network escalates PSNR by approximately 7, SAM has been diminished by approximately 9, signifying that the design is plausible.

In the experiment, we designed spectral and SpaR networks to improve the spectral resolution of HR MSI and the spatial resolution of LR HSI, respectively. To verify the rationality of the spatial spectral interaction super-resolution network, we

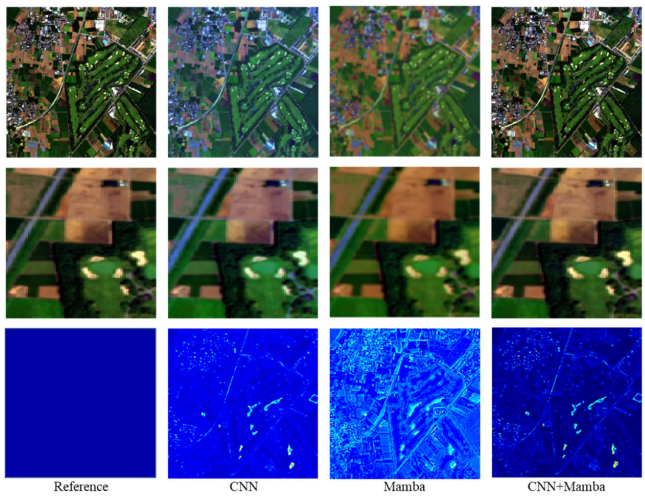

Fig. 15. Results of ablation experiment of CNN and mamba.

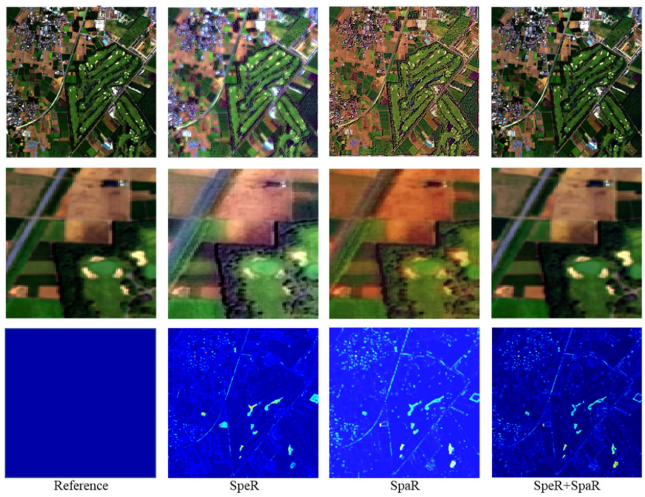

Fig. 16. Results of ablation experiment of SpeR and SpaR.

TABLE IV
QUANTITATIVE RESULTS OF ABLATION EXPERIMENT OF SPER AND SPAR

Methods	SpeR	SpaR	SpeR+SpaR
PSNR	20.19	19.68	29.86
SAM	7.63	12.55	2.53

conducted ablation experiments on SpeR and SpaR networks, respectively. Fig. 16 and Table IV show the experimental results of the simulated dataset-1, and indicate that using only SpeR or SpaR has significant spectral distortion, the combination of SpeR and SpaR effectively improves the fusion performance of the network, PSNR has increased by about 9, and SAM has improved by about 5.

In the experiment, we used a combination of CFE and SCM to correct the channel relationship of HR MSI features. To verify whether this correction can effectively improve the fusion performance of the network, we conducted ablation experiments. Fig. 17 and Table V show the fusion results. When CFE and SCM are ablated, the fusion result has serious spectral reconstruction errors, PSNR decreases by about 12, SAM increased

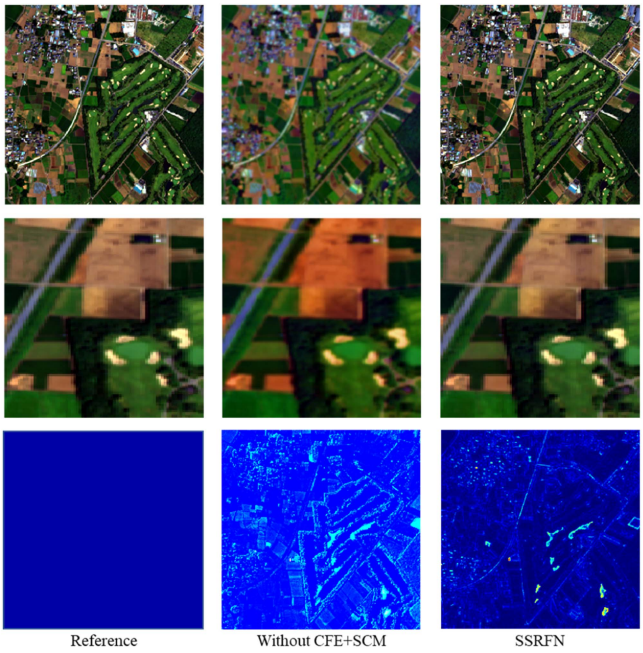

Fig. 17. Results of ablation experiment of CFE and SCM.

TABLE V
QUANTITATIVE RESULTS OF ABLATION EXPERIMENT OF CFE AND SCM

Methods	Without CFE+SCM	SSRFN
PSNR	17.26	29.86
SAM	13.22	2.53

TABLE VI
RUN TIMES FOR THE DIFFERENT METHODS ON THE SIMULATED DATASETS-1

Methods	Train (min) Case 5	Test (min)	FLOPs (G)	Params (M)
GSA	—	2.07	—	—
GLP	—	2.68	—	—
CNMF	—	4.02	—	—
NLSTF	—	7.92	—	—
SSRNET	137	0.37	24.36	5.62
MSST-Net	1075	0.81	426.15	86.54
UHNTC	909	0.79	396.57	67.68
MIMO–SST	1376	0.95	483.05	106.64
FMamba	197	0.61	168.27	15.35
QIS–GAN	373	0.67	308.64	46.75
SSRFN	297	0.59	193.67	38.63

by about 11. The experimental results indicate that CFE and SCM are crucial for the network.

Table VI shows the training and testing time of different methods in the simulated datasets-1, and the computational efficiency of GSA and GLP is superior to other traditional methods. In DL methods, SSRNET, FMamba, QIS–GAN, and SSRFN have achieved high computational efficiency, while MSST-Net, UHNTC, and MIMO–SST have achieved lower computational efficiency in the transformer framework. The experimental results show that the introduction of Mamba effectively extracts global features of images to improve fusion performance while ensuring low computational cost.

IV. CONCLUSION

This article proposes a spatial–spectral interaction super-resolution CNN–Mamba network for fusion of HSI and MSI. SSRFN uses bidirectional guidance to perform SpeR and SpaR on MSI and HSI, respectively. To some extent, this avoids the problem of spatial spectral information balance caused by information differences, effectively improving the fusion quality. Specifically, we designed novel difference information extraction modules and SCMs to correct the feature channels of MSI, and designed SpeR and SpaR networks, using MSI and HSI as auxiliary information for super-resolution of each other. The design of FFM module effectively improves the reconstruction accuracy of fusion results while avoiding information redundancy. In addition, the Mamba network was introduced to combine with CNN to extract local and global features of images, exploring the distribution characteristics of image information while ensuring the computational efficiency of the network. A large number of experiments were conducted on both simulated and real datasets. Experimental results show that our method achieves perfect performance compared to other advanced methods.

REFERENCES

[1] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, 2023.
[2] K. Ren, W. Sun, X. Meng, G. Yang, and Q. Du, "Fusing China GF-5 hyperspectral data with GF-1, GF-2 and Sentinel-2A multispectral data: Which methods should be used?," *Remote Sens.*, vol. 12, no. 5, 2020, Art. no. 882.
[3] N. Aburaed, M. Q. Alkhatib, S. Marshall, J. Zabalza, and H. Al Ahmad, "A review of spatial enhancement of hyperspectral remote sensing imaging techniques," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2275–2300, 2023.
[4] L. Chen et al., "Mapping alteration minerals using ZY-1 02D hyperspectral remote sensing data in coalbed methane enrichment areas," *Remote Sens.*, vol. 15, no. 14, 2023, Art. no. 3590.
[5] X. Hu et al., "High SNR eSWIR image sensor applied in advanced hyperspectral imager (AHSI) aboard China's GaoFen-5 satellite and ZY-1 satellite," *IEEE Sensors J.*, vol. 24, no. 16, pp. 25550–25557, Aug. 2024.
[6] R. U. Shaik, S. Periasamy, and W. Zeng, "Potential assessment of PRISMA hyperspectral imagery for remote sensing applications," *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1378.
[7] Z. Tong, Y. Li, J. Zhang, L. He, and Y. Gong, "MSFANet: Multiscale fusion attention network for road segmentation of multispectral remote sensing data," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 1978.
[8] C. Wang, X. Zhang, W. Yang, X. Li, B. Lu, and J. Wang, "MSAGAN: A new super-resolution algorithm for multispectral remote sensing image based on a multiscale attention GAN network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5001205.
[9] Z. Zhang, L. Wei, S. Xiang, G. Xie, C. Liu, and M. Xu, "Task-driven on-board real-time panchromatic multispectral fusion processing approach for high-resolution optical remote sensing satellite," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7636–7661, 2023.
[10] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519015.
[11] K. Li, W. Zhang, D. Yu, and X. Tian, "HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 30–44, 2022.
[12] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6007305.
[13] 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, 2022, Art. no. 5537112.

[14] W. Sun, K. Ren, X. Meng, G. Yang, J. Peng, and J. Li, "Unsupervised 3D tensor subspace decomposition network for spatial-temporal-spectral fusion of hyperspectral and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[15] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl. Syst.*, vol. 264, 2023, Art. no. 110362.

[16] H. Gao, S. Li, and R. Dian, "Hyperspectral and multispectral image fusion via self-supervised loss and separable loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537712.

[17] R. Ducay and D. W. Messinger, "Image fusion of hyperspectral and multispectral imagery using nearest-neighbor diffusion," *J. Appl. Remote Sens.*, vol. 17, no. 2, 2023, Art. no. 024504.

[18] W. Carper, T. Lillesand, and R. Kiefer, "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.

[19] S. Pal, T. Majumdar, and A. K. Bhattacharya, "ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 61, no. 5, pp. 281–297, 2007.

[20] 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.

[21] W. Sun, K. Ren, X. Meng, C. Xiao, G. Yang, and J. Peng, "A band divide-and-conquer multispectral and hyperspectral image fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502113.

[22] A. Arienzo, B. Aiazzi, L. Alparone, and A. Garzelli, "Reproducibility of pansharpening methods and quality indexes versus data formats," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4399.

[23] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, 2021.

[24] 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.

[25] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517218.

[26] F. D. Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, and A. Stein, "A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 101–117, 2021.

[27] S. Wady, Y. Bentoutou, A. Bengermikh, A. Bounoua, and N. Taleb, "A new IHS and wavelet based pansharpening algorithm for high spatial resolution satellite imagery," *Adv. Space Res.*, vol. 66, no. 7, pp. 1507–1521, 2020.

[28] W. Sun et al., "MLR-DBPFN: A multi-scale low rank deep back projection fusion network for anti-noise hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522914.

[29] 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.

[30] 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.

[31] L. Sui, L. Li, J. Li, N. Chen, and Y. Jiao, "Fusion of hyperspectral and multispectral images based on a Bayesian nonparametric approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1205–1218, Apr. 2019.

[32] Y. Zhou, L. Feng, C. Hou, and S.-Y. Kung, "Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5997–6009, Oct. 2017.

[33] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4469–4480, Oct. 2020.

[34] N. Yokoya, X. X. Zhu, and A. Plaza, "Multisensor coupled spectral unmixing for time-series analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2842–2857, May 2017.

[35] Z. H. Nezhad, A. Karami, R. Heylen, and P. Scheunders, "Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2377–2389, Jun. 2016.

[36] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi-and hyperspectral images using PCA and wavelets," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2652–2663, May 2015.

[37] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Trans. Image Process.*, vol. 29, pp. 116–127, 2020.

[38] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.

[39] C.-H. Lin, F. Ma, C.-Y. Chi, and C.-H. Hsieh, "A convex optimization-based coupled nonnegative matrix factorization algorithm for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1652–1667, Mar. 2018.

[40] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.

[41] R. Guerra, S. López, and R. Sarmiento, "A computationally efficient algorithm for fusing multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5712–5728, Oct. 2016.

[42] X. Li, Y. Yuan, and Q. Wang, "Hyperspectral and multispectral image fusion based on band simulation," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 479–483, Mar. 2020.

[43] W. Sun et al., "Domain transform model driven by deep learning for anti-noise hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5500117.

[44] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[45] R. Lu, B. Chen, Z. Cheng, and P. Wang, "RAFnet: Recurrent attention fusion network of hyperspectral and multispectral images," *Signal Process.*, vol. 177, 2020, Art. no. 107737.

[46] K. Ren et al., "CDFSL: Image registration for spaceborne hyperspectral and multispectral data having large spatial-resolution difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[47] Y. Sun, B. Xue, M. Zhang, G. G. Yen, and J. Lv, "Automatically designing CNN architectures using the genetic algorithm for image classification," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3840–3854, Sep. 2020.

[48] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided CNN for image denoising," *Neural Netw.*, vol. 124, pp. 117–129, 2020.

[49] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.

[50] S. Niu, B. Li, X. Wang, and H. Lin, "Defect image sample generation with GAN for improving defect recognition," *IEEE Trans. Automat. Sci. Eng.*, vol. 17, no. 3, pp. 1611–1622, Jul. 2020.

[51] 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," *Inf. Fusion*, vol. 62, pp. 110–120, 2020.

[52] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[53] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3585.

[54] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[55] 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.

[56] R. Ran, L.-J. Deng, T.-J. Zhang, J. Chang, X. Wu, and Q. Tian, "KNL-Conv: Kernel-space non-local convolution for hyperspectral image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 1–16, 2024.

[57] S. Peng, C. Guo, X. Wu, and L.-J. Deng, "U2net: A general framework with spatial-spectral-integrated double U-Net for image fusion," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 3219–3227.

[58] C. Zhu, S. Deng, Y. Zhou, L.-J. Deng, and Q. Wu, "QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531115.

[59] J. Zhang et al., "Transformer based conditional GAN for multimodal image fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 8988–9001, 2023.

[60] 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.
[61] S. Li, R. Dian, and H. Liu, "Learning the external and internal priors for multispectral and hyperspectral image fusion," *Sci. China Inf. Sci.*, vol. 66, no. 4, 2023, Art. no. 140303.
[62] J. Yang, T. Lin, X. Chen, L. Xiao, and R. Sensing, "Multiple deep proximal learning for hyperspectral-multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[63] 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, 2023, Art. no. 5503715.
[64] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "Mambahsi: Spatial-spectral mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.
[65] C. Wang, J. Huang, M. Lv, H. Du, Y. Wu, and R. Qin, "A local enhanced mamba network for hyperspectral image classification," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 133, pp. 1–11, 2024.
[66] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "Rsmamba: Remote sensing image classification with state space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[67] H. Zhang et al., "A survey on visual mamba," *Appl. Sci.*, vol. 14, no. 13, pp. 1–14, 2024.
[68] G. Sun, Y. Hua, G. Hu, and N. Robertson, "Mamba: Multi-level aggregation via memory bank for video object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2620–2627.
[69] M. S. Ansari, V. Bartoš, and B. Lee, "GRU-based deep learning approach for network intrusion alert prediction," *Future Gener. Comput. Syst.*, vol. 128, pp. 235–247, 2022.
[70] 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.
[71] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
[72] J. Fang, J. Yang, A. Khader, L. Xiao, and R. Sensing, "MIMO-SST: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
[73] X. Ma, X. Zhang, and M.-O. Pun, "RS 3 Mamba: Visual state space model for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

Guangwei Zhao received the Ph.D. degree in operational research and cybernetics from the College of Science, Shanghai University, Shanghai, China, in 2022.

He is currently a Lecturer with the College of Computer and Artificial Intelligence, Huanghuai University, Zhumadian, China. His research interests include deep learning and computer vision.

Dr. Zhao is a Member of the Key Laboratory of Smart Lighting in Henan Province.

Haitao Wu received the Ph.D. degree in computer software and theory from the Wuhan University, Wuhan, China, in 2015.

He is currently a Professor with the College of Computer and Artificial Intelligence, Huanghuai University, Zhumadian, China. His research interests include data mining and intelligent information processing.

Dexiang Luo received the M.S. degree in information and computing science from the Guangxi Minzu University, Nanning, China, in 2009.

He is currently a Senior Engineer with the Information Center, Guangxi Medical University, Nanning, China. His research interests include deep learning and data mining.

Xu Ou received the M.S. degree in computer science and technology from the Guangxi University, Nanning, China, in 2009.

He is currently a Senior Engineer with the Information Center, Guangxi Medical University, Nanning, China. His research interests include deep learning and information security.

Yu Zhang received the Ph.D. degree in computer software and theory from the Jilin University, Changchun, China, in 2018.

He is currently an Associate Professor with the College of Computer and Artificial Intelligence, Huanghuai University, Zhumadian, China. He also serves with the Henan Key Laboratory of Smart Lighting and Henan Joint International Research Laboratory of Behavior Optimization Control for Smart Robots. His research interests include semantic web and natural language processing.