A Hybrid Spectral Attention-Enabled Multiscale Spatial–Spectral Learning Network for Hyperspectral Image Superresolution

Wenjing Wang, Tingkui Mu, Qiuxia Li, Haoyang Li, and Qiujie Yang

***Abstract*—Hyperspectral image superresolution (HSI-SR) has become an essential step of data preprocessing for tasks such as classification and change detection in remote sensing. For HSI-SR tasks, the state-of-the-art methods lie in how to learn effective spatial and spectral characteristics. More deeply mining the shallow and deep spatial–spectral features is vital for the performance improvement of HSI-SR. Hereby, we provide an end-to-end multiscale spatial–spectral network (M3SN) driven by a hybrid spectral attention mechanism (HSAM) for HSI-SR, aiming to fully dig shallow and deep spatial–spectral characteristics. Precisely, considering the importance of shallow spatial–spectral features at early stages, a multiscale information block consisting of a 3-D convolution, three parallel 2-D convolutions with different scale sizes, and SE-Net is first designed to extract informative multiscale shallow spatial–spectral features and recalibrate the channel weights of features. Then, a dual-path multiscale spatial–spectral feature block is set up to explore the deep spatial and spatial–spectral features. While one path using 2-D convolutions extracts spatial features, the other path employing 3-D-Res2Net as well as an updated HSAM module mines multiscale deep spatial–spectral features. Finally, we design a multiscale fusion block based on the channel reduction-scaling operation to fuse the extracted hierarchical feature maps for the final reconstruction. It is demonstrated that the M3SN outperforms existing methods by extensive experiments on four publicly available hyperspectral datasets.**

***Index Terms*—Convolutional neural network (CNN), dual-path framework (DPF), hybrid spectral attention mechanism (HSAM), hyperspectral image superresolution (HSI-SR), multiscale spatial–spectral features.**

I. Introduction

HYPERSPECTRAL image (HSI) obtained by hyperspectral sensor consists of hundreds of continuous bands. It is a 3-D data cube with a set of images recording reflectivity or radiance at different wavelengths [1]. The high spectral resolution of HSI can reflect the tiny spectral characteristics of the measured object [2]. Therefore, HSI has been widely used for medical diagnosis [3], plant disease detection [4], and mineral exploration [5], etc. Hyperspectral sensors have many continuous narrow bands while the photon energy is constant, so the average number of photons reaching each narrow band is limited. As a result of this inherent limitation, the spatial resolution of HSI is generally low, and details are rarely captured, which limits the performance of high-level tasks, such as classification [6], change detection [7], [8], [9], small object detection [10] etc. Generally, the spatial resolution of HSI is enhanced by improving hardware facilities or utilizing algorithms. The former is expensive and has high requirements on existing engineering techniques. The latter is hyperspectral image superresolution (HSI-SR), which aims to recover high-resolution (HR) HSI from low-resolution (LR) HSI and keep the spectrum undistorted simultaneously.

The mainstream HSI-SR methods can be divided into multimodal-fusion-based [11], [12] and single image superresolution (SISR). For the former one, high spatial resolution images such as panchromatic (PAN) images [13], RGB [14], or multispectral Images (MSI) [15] are fused with HSI. Conventional methods can be divided into three categories, the Bayesian-based methods such as HySure [16], the matrix-factorization-based methods such as CNMF [17], and the representative-based methods such as LTTR [18]. However, obtaining high spatial resolution images and HSIs corresponding to the same scene is difficult in practical engineering. As for SISR, traditional methods can be categorized into interpolation-based methods and regularization-based methods, which mainly solve the ill-posed inverse problem to get HR HSIs. The former primarily includes bilinear, bicubic, and Nearest interpolation, while the latter includes Markov random fields [19], total variation (TV) regularization [20], sparse regularization [21], and the self-similarity prior [22]. However, traditional methods rely on heuristic and manually designed prior information, which has weak representation capabilities. In addition, the necessity of optimal parameter adjustments for different images and devices limits the performance and practicality of these methods.

In recent years, the powerful representation capabilities [23] of deep learning, especially convolutional neural network (CNN), make it have great potential in the field of image

Manuscript received 16 October 2023; revised 5 February 2024 and 27 April 2024; accepted 21 May 2024. Date of publication 31 May 2024; date of current version 14 June 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62175196, in part by the Shaanxi Province Key Research and Development Program under Grant 2021GXLH-Z-058, and in part by the Shaanxi Fundamental Science Research Project for Mathematics and Physics under Grant 22JSY020. *(Corresponding authors: Tingkui Mu; Qiujie Yang.)*

Wenjing Wang, Tingkui Mu, Qiuxia Li, and Haoyang Li are with the MOE Key Laboratory for Nonequilibrium Synthesis and Modulation of Condensed Matter, Research Center for Space Optics and Astronomy, School of Physics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wjwang@stu.xjtu.edu.cn; tkmu@mail.xjtu.edu.cn; lqx0324@stu.xjtu.edu.cn; lihaoyang@stu.xjtu.edu.cn).

Qiujie Yang is with the Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China (e-mail: yqj488112gxx@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3407953

processing. Many CNN models have been proposed in the field of natural image superresolution (SR) [24], [25], [26]. For example, SRCNN [24] is the first model for SR. Since then, many typical networks have been proposed one after another, such as EDSR [25], and VDSR [26], which have achieved satisfactory results in the field of natural images. Therefore, many literatures have begun to use CNN-based methods for HSI-SR. Unlike natural images, HSI is a 3-D data cube with many bands, making extracting feature information difficult. Due to the neglect of spectral information, the model designed for natural images is less effective for HSI. Therefore, developing a CNN-based model for HSI-SR has become a research hotspot in recent years [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. In addition, due the capability in establishing long-range dependencies and capturing spectral features, Transformer has emerged as a research hotspot in the field of HSI-SR [47], [48], [49].

However, there are three challenges in the CNN-based HSI-SR. The first one is how to model the mapping relationship between LR and HR HSIs without auxiliary HR images. Therefore, the critical task is to design a rational network capable of fully exploring spatial and spectral characteristics to enhance the ability to model complex relationships effectively [27]. To tackle this challenge, various approaches have been proposed. One notable strategy involves leveraging typical basic blocks such as Res2Net [50], renowned for their ability to extract multiscale features at a finer granularity level, thereby achieving the excellent performance. In addition, combining 3-D and 2-D convolutions is preferred to fully explore the spatial and spectral information, MCNet [28] and ERCSR [29], proposed by Li et al. [28], [29], are two typical networks. While MCNet used multiple 2-D and 3-D units to share spatial parameters, ERCSR alternately used 3-D and 2-D units to fully consider the relationship between vertical, horizontal, and spectral dimensions. However, both of them ignored the second-order covariance statistics that have shown the importance on HSI-SR [31], [32]. Furthermore, the dual-path framework (DPF) is an alternative method for improving performance. Li et al. [39] proposed a DPF to parallelly learn spatial–spectral information and spatial information using a spatial–spectral learning subnetwork and a spatial reconstruction subnetwork, respectively. Although the performance is improved, the two features are not interactive in the middle layer resulting in the underutilization of complementary information.

The second challenge lies in how to mitigate spectral distortion. For instance, SRCNN was extended to msiSRCNN for processing multispectral images in a band-by-band manner [30], which resulted in severe spectral distortion since the spectral information was ignored. To tackle this challenge, recent researches have explored various methodologies, including imposing constraints on the spectral domain through 3-D convolution, spatial–spectral TV loss functions, and gradient maps. However, despite these efforts, maintaining spectral fidelity remains an ongoing challenge. Considering the strong correlation between adjacent bands and the capability of 3-D convolution can extract both spatial and spectral features, Mei et al. [37] developed a 3-D fully convolution neural network (3D-FCNN) that achieved good performance. However, naive 3-D convolution cannot explore the spectral correlation sufficiently and simple MSE loss function cannot gauge spectral distortion accurately [34]. Jiang et al. [38] designed SSPSR to explore the spatial–spectral features and focus on spectral fidelity by spatial–spectral TV loss function. Zhao et al. [40] used the spectral gradient information of HSI as prior information to constrain the spectrum that achieved good result. In addition, some literatures [31], [32] have demonstrated the vital contribution of the second-order statistics to mine more spectral characteristics and then facilitate spectral consistency.

The final challenge concerns the treatment of spectral bands within general CNN architectures. Since general CNN treats each spectral band equally, the difference between spectral bands cannot be well distinguished, which would limit the representation ability of network [31], [32], [33], [34], [35], [36], [37]. Therefore, a preferable approach is to treat each spectral band individually. Inspired by that human vision selectively pays attention to regions with significant features [51], attention mechanisms have become prevalent in this field. Currently, the attention mechanisms used in SR tackle two main considerations. One is that the network focuses on high-frequency information, which is easily lost in the propagation process and difficult to recover. The other is that the contribution of each channel to reconstruction is inconsistent and the weight of each channel should be recalibrated [31], [32]. Many attention mechanisms have been applied to the field of HSI-SR [31], [32], [33], [34], [35]. LN-atten-CNN [35] obtained global attention information and improved the representation ability by max pooling in the spectral dimension and average pooling in the spatial dimension. Dong et al. [52] proposed a CFDcagaNet to provide more realistic spectral guidance via a content-aware attention mechanism, which has achieved better results. Furthermore, the representation abilities of attention mechanisms can be enhanced by exploring the second-order statistical features [25], [32], [36], [53], [54]. Bryan et al. [53] modeled long-term relationships using the attention mechanism with second-order statistics, achieving excellent performance. Dai et al. [54] proposed the second-order channel attention to represent channel characteristics, where the first-order pooling operation is replaced by the second-order covariance pooling. For HSI-SR, just using the first-order statistical feature or the second-order statistical feature alone is not enough to express the internal feature of images. Therefore, it is better to combine the first and second statistical features. Hu et al. [31] combined the second-order covariance information and the first-order statistics for HSI-SR and achieved good results. But there is still a limitation in the excitation step because using two 3-D convolutions for capturing channel-wise relationships is not accurate, which would lead to the weights of channels not recalibrating accurately.

In this article, we design an end-to-end trainable network called a multiscale spatial–spectral network (M3SN) driven by a hybrid spectral attention mechanism (HSAM).

First, for shallow feature extraction (SFE), a multiscale information block (MSIB) is designed to extract multiscale shallow spatial–spectral features. MSIB is realized by a 3-D convolution, three parallel 2-D convolutions with different scale kernels, and

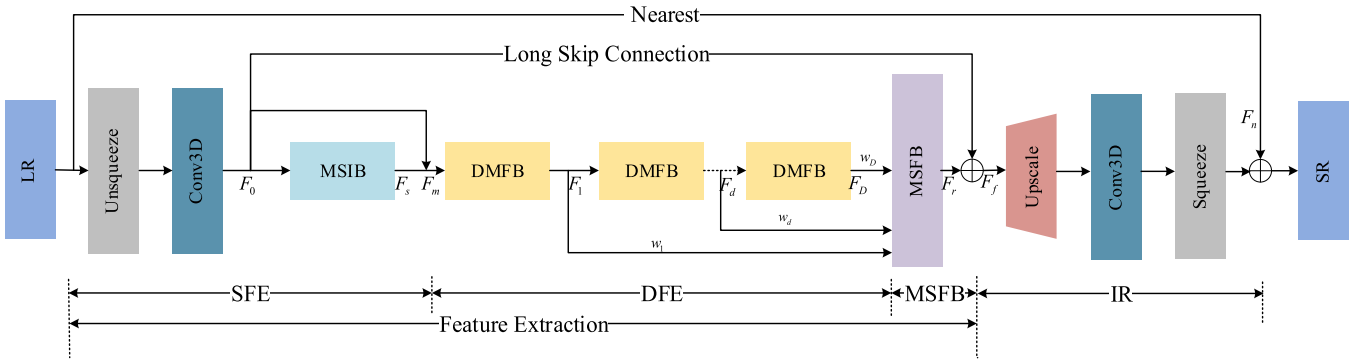

Fig. 1. Overview of M3SN. The same-colored boxes indicate the same operation. The main components of our M3SN are: 1) MSIB extracts multiscale shallow spatial-spectral features by a 3-D convolution, three parallel 2-D convolution operations of different size kernels, and SE-Net; 2) DMFB extracts multiscale deep spectral–spatial features for accurate reconstruction by 2-D convolutions and 3D-Res2Net equipped with HSAM in a dual-path manner; and 3) MSFB fuses the different hierarchical features of each DMFB.

SE-Net. Second, for deep feature extraction (DFE), multiple dual-path multiscale spatial–spectral feature blocks (DMFBs) are stacked in a cascaded manner to explore spatial–spectral and spatial features. The DMFB includes two paths: spatial features extraction (SpaFE) and multiscale spatial–spectral feature extraction (MSFE). SpaFE consists of two 2-D convolutions. MSFE consists of 3D-Res2Net and HSAM, 3D-Res2Net replaces 2-D convolution in the original Res2Net with 3-D convolution. HSAM combines first-order statistics and second-order statistics to enhance the useful spectral features and suppress the useless ones. Then, a multiscale fusion block (MSFB) is used to aggregate different hierarchical features of each DMFB, which can improve the reuse rate of features. Finally, the image reconstruction (IR) is completed through deconvolution.

In conclusion, the main contributions of this article can be summarized as follows.

1) A new CNN-based HSI-SR method, M3SN, is proposed. To fully enhance the learning ability of spatial features using spatial–spectral features, the deep multiscale spatial–spectral features and spatial features are explored in a dual-path learning manner.
2) Considering the importance of shallow features, MSIB consisting of a 3-D convolution, three parallel 2-D convolutions with different size of kernels, and SE-Net are utilized to obtain rich multiscale shallow spatial–spectral features and recalibrate the weights of spectral bands.
3) To fully use spectral information and avoid the spectral distortion, the HSAM is designed to pay attention simultaneously to the first-order statistics and the second-order covariance statistics. Then, an enhanced excitation module makes it predict the weight of channel more accurately. It increases the diversity of spectral features and the learning ability of discriminative spectral features. Furthermore, 3D-Res2Net equipped with HSAM is powerful to explore the deep multiscale spatial–spectral features and spectral correlation.

The rest of this article is organized as follows. In Section II, we describe the proposed method, M3SN, including the network structure and description. Section III presents extensive experiments and ablation analyses on four public datasets. Finally, Section IV concludes this article.

II. Proposed Method

In this section, we specify the proposed model M3SN in detail. It is an end-to-end model dedicated to mining and utilizing the multiscale spatial–spectral features of HSI.

A. Network Structure

As shown in Fig. 1, M3SN mainly consists of four parts: SFE, DFE, MSFB, and IR. For HSI, define $I_{\mathrm{LR}} \in R^{B\times W\times H}$ and $I_{\mathrm{SR}} \in R^{B\times rW\times rH}$ as the input LR and SR HSI, respectively, where W and H are width and height, B is the total number of bands, and r is the scale factor.

SFE is mainly to obtain informative shallow spatial–spectral features F_m. The input LR is reshaped into four dimensions ($1\times B\times W\times H$) to exact the initial shallow feature F_0 by a 3-D convolution layer. Then, MSIB is followed to extract the multiscale shallow spatial–spectral features F_m as shown in Fig. 2. The aforementioned process can be formulated as

$$F_0 = f_{\mathrm{Conv3D},3}\left(\mathrm{unsqueeze}\left(I_{\mathrm{LR}}\right)\right) \tag{1}$$

$$F_m = H_{\mathrm{MSIB}}\left(F_0\right) + F_0 \tag{2}$$

where unsqueeze$(\cdot)$ is used to expand I_{LR}, $f_{\mathrm{Conv3D},3}$ denotes a 3-D convolution layer, and $H_{\mathrm{MSIB}}(\cdot)$ represents a composite operation of 3-D convolution, 2-D convolution, and SE-Net.

DFE is designed to extract deep spatial–spectral features. It contains several DMFBs. Specifically, the output F_d of the d_{th} DMFB can be denoted as

$$F_d = H_{\mathrm{DMFB},d}\left(H_{\mathrm{DMFB},d-1}\left(\cdots H_{\mathrm{DMFB},1}\left(F_m\right)\cdots\right)\right) \tag{3}$$

where $H_{\mathrm{DMFB},d}$ represents the operations of the d_{th} DMFB, a composite operation of 3-D convolution, 2-D convolution, and HSAM. The diagram is presented in Fig. 3. We will discuss and analyze the impact of the number of DMFBs in the network in Section III.

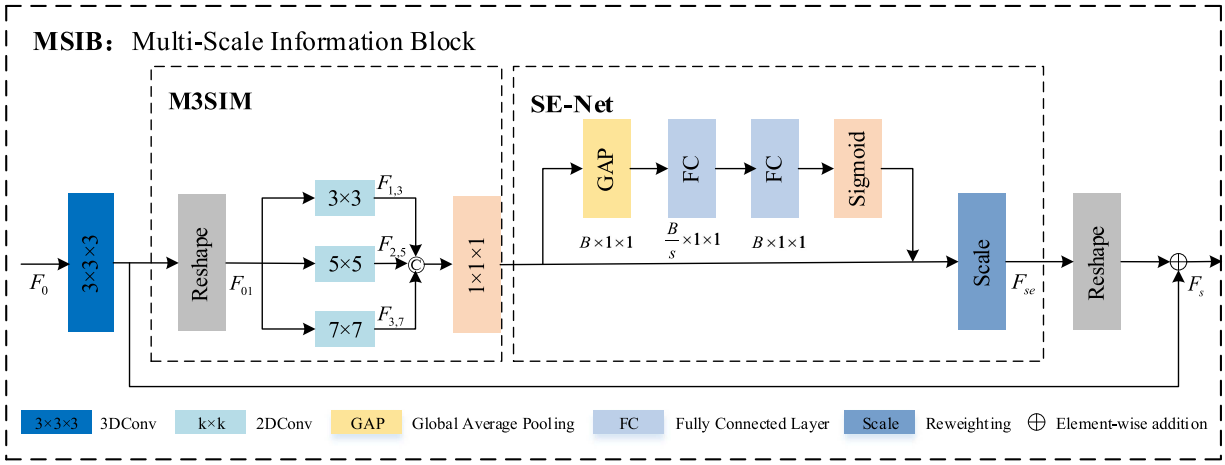

Fig. 2. Structure of the MSIB, which is composed of M3SIM and SE-Net. M3SIM consists of a 3-D convolution layer and three 2-D convolutions with different scale kernels.

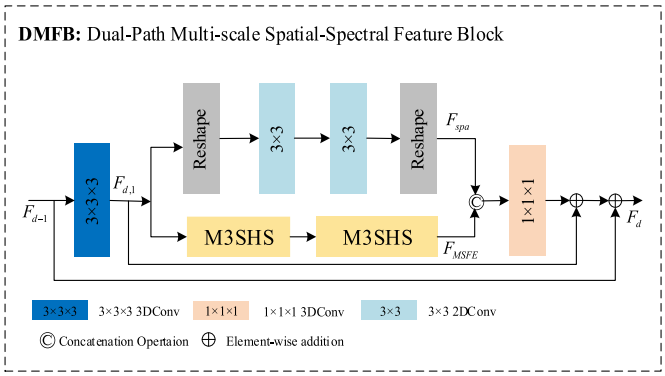

Fig. 3. Structure of the DMFB. It consists of MSFE and SpaFE. The upper branch is called SpaFE and the lower branch is called MSFE.

The hierarchical features from D DMFBs can be expressed as $[F_1, \ldots, F_D]$ and fused by the MSFB, which will be introduced in Section III. Then, the shallow features F_0 and deep fused features are added via long skip connection, the output F_f is described as

$$F_f = H_{\text{MSFB}}\left([F_1, \ldots, F_D]\right) + F_0 \tag{4}$$

where H_{MSFB} represents the operations of MSFB, which is a composite operation of 3-D convolution and SE-Net.

With respect to IR, F_f is up-sampled by a transposed 3-D convolution according to scale factor r, which is followed by a 3-D convolution layer. Since the input and output images are so differently that F_n is introduced to up-sample the input LR to narrow the gap by Nearest operation [29], which greatly alleviated the burden on the network. After a squeeze operation, the feature map is reshaped into three dimensions ($B \times W \times H$), the output I_{SR} is obtained by

$$I_{SR} = \text{squeeze}\left(f_{\text{Conv3D}}\left(\text{up}\left(F_f\right)\right)\right) + F_n \tag{5}$$

where $\text{up}(\cdot)$ is a 3-D transposed convolution layer, $f_{\text{Conv3D}}(\cdot)$ is a 3D convolution layer, and $\text{squeeze}(\cdot)$ is squeeze function.

B. Multiscale Information Block (MSIB)

In the SGARDN [32], the idea that effective shallow features can improve the reconstruction ability of the network is confirmed. That is, shallow features at multiscale need to be acquired at the early stages in the network. Furthermore, inspired by the [6] that multiscale spatial features can be extracted by 2D convolutions with different size kernels, we design MSIB for multiscale shallow spatial–spectral features extraction at the early stages of the network.

As shown in Fig. 2, MSIB is divided into two parts, multiscale shallow spatial–spectral information module (M3SIM) and SE-Net [55].

1) Multiscale Shallow Spatial–Spectral Information Module (M3SIM): The initial spatial–spectral feature F_{01} is first extracted by a 3-D convolution layer, which can be formulated as

$$F_{01} = \text{reshape}\left(f_{\text{Conv3D},3}\left(F_0\right)\right). \tag{6}$$

Then, the multiscale shallow spatial features are extracted by using N parallel 2-D convolutions with kernels $k \times k$. In M3SIM, we set $N = 3$ and $k = 3, 5, 7$, respectively. The output feature maps $F_{n,k}$ of nth can be expressed as

$$F_{n,k} = f_{\text{Conv2D},k}\ \left(F_{01}\right) \tag{7}$$

where $f_{\text{Conv2D},k}(\cdot)$ represents 2-D convolution with kernel $k \times k$. $F_{1,3}$,$F_{2,5}$, and $F_{3,7}$ are concatenated in spectral dimension, followed by a $1 \times 1 \times 1$ convolution layer, which can reduce the number of feature maps. The output F_s can be denoted as

$$F_s = f_{\text{Conv2D},1}\ \left(\text{concat}\left(F_{1,3}, F_{2,5}, F_{3,7}\right)\right). \tag{8}$$

2) SE-Net: Since CNN treats channels equally in the process of extracting features, however, each channel contains different spatial information and contributes inconsistently to the network. Therefore, SE-Net is introduced to adjust the weight of spectral bands to improve the representation of MSIB. The SE-Net consists of two steps, squeeze and excitation. First, the squeeze operation is utilized to obtain global distribution of channel-wise responses through shrinking each 2-D spatial

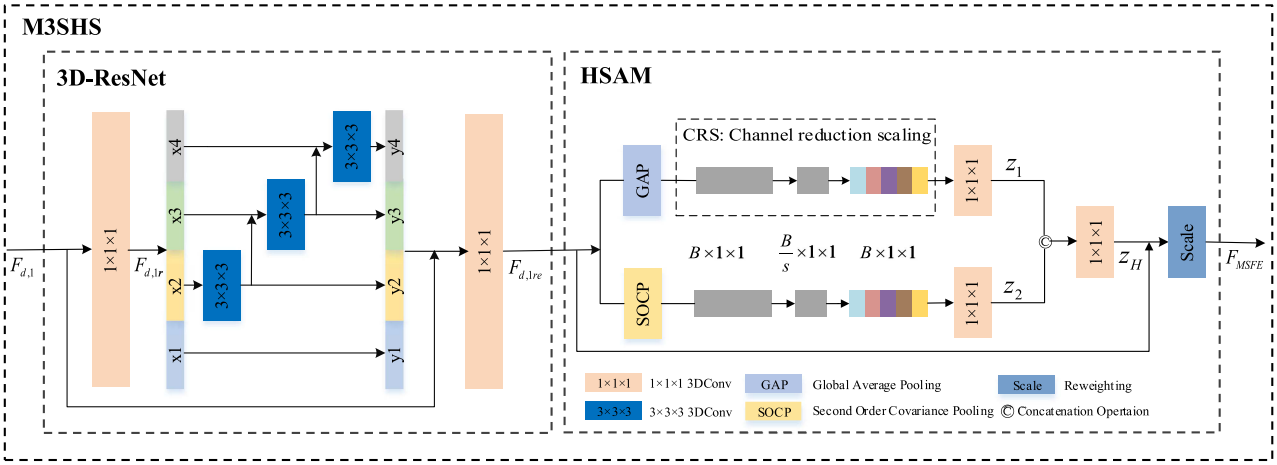

Fig. 4. Structure of M3SHS formed by 3D-Res2Net with HSAM.

dimension into a real number by global average pooling (GAP)

$$F_G = f_{sq}\ (F_s) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_s\,(i,j)\,. \tag{9}$$

The feature map F_G is reshaped from $F_G \in R^{B\times W\times H}$ to $F_G \in R^{B\times 1\times 1}$.

Then, the excitation operation is used to predict the weight of each channel via two fully connected layer and nonlinear layers. This process can be described as

$$w = f_{\text{ex}}\ (F_G, \omega) = \sigma\ (\mu\,(F_G, \omega)) = \sigma\,(\omega_2 \mu\,(\omega_1 F_G)) \tag{10}$$

where μ and σ are the ReLU activation function and Sigmoid activate function, respectively. ω_1 and ω_2 are the weights of two consecutive fully connected layers.

Finally, the output feature F_{se} is obtained by multiplying the learned weight w and original feature map F_s as follows:

$$F_{\text{se}} = f_{\text{scale}}\ (F_s, w) = w \cdot F_s. \tag{11}$$

The output feature map F_s of MSIB is obtained by a reshape operation and local residual learning, which can improve the representation ability of the network.

C. Dual-Path Multiscale Spatial–Spectral Feature Block (DMFB)

Some literatures [28], [29], [39] show that just spatial–spectral features extraction is not conducive to the reconstruction. We need to enhance the ability to extract features in the spatial domain. Therefore, a DPF is used to extract spatial–spectral features and spatial features simultaneously so that spatial–spectral and spatial features can complement each other. As shown in Fig. 3, $F_{d,1}$ is extracted by a $3 \times 3 \times 3$ convolution layer

$$F_{d,1} = f_{\text{Conv3D},1}\ (F_{d-1})\,. \tag{12}$$

Then, DFP is followed to obtain the deep feature F_d. It includes SpaFE branch and MSFE branch.

1) Spatial Features Extraction (SpaFE): SpaFE consists of two 2-D convolution layers, aiming to acquire spatial features. The output $\boldsymbol{F_{\text{spa}}}$ is obtained by

$$F_{\text{spa}} = \text{reshape}\,(f_{\text{Conv2D},3}\,(f_{\text{Conv2D},3}\,(\text{reshape}\,(F_{d,1}))))\,. \tag{13}$$

2) Multiscale Spatial–Spectral Feature Extraction (MSFE): MSFE consists of two M3SHSs focusing more on exploring multiscale deep spatial–spectral features. As shown in Fig. 4, M3SHS consists of 3D-Res2Net and HSAM, which is designed to obtain more representative multiscale spatial–spectral features.

3D-Res2Net [31] replaces 2-D convolution in original Res2Net [28] with 3-D convolution, which is capable of acquiring multiscale spatial–spectral features of different receptive fields. First, $F_{d,1r}$ is obtained by a $1\times1\times1$ convolution layer

$$F_{d,1r} = f_{\text{Conv3D},1}\ (F_{d,1})\,. \tag{14}$$

Then, $F_{d,1r}$ is divided into q subsets defined as $x_i \in [x_1, x_2, \ldots, x_q]$. The subsets have the same size, and the number of channels is $1/q$ of the input features. The multiscale feature map y_i is obtained according to

$$y_i = \begin{cases} x_i, & i = 1 \\ \mu\,(f_{\text{Conv3D},3}\,(x_i + y_{i-1})), & 1 < i \le q \end{cases}. \tag{15}$$

Therefore, 3D-Res2Net can obtain feature combinations with different numbers and different receptive field sizes. Then, we concatenate $y_i$in spectral dimension, and a $1\times1\times1$ 3-D convolution layer for dimensionality reduction is followed. The local residual connections in 3D-Res2Net can capture both detailed and global information. The output feature map $F_{d,1re}$ can be denoted as

$$F_{d,1re} = f_{\text{Conv3D},1}\ (\text{concat}\,[y_i]) + F_{d,1})\,. \tag{16}$$

We mainly concern two points. First, the convolution kernel equally treats all spectral bands that have different contributions to the reconstruction, which limits the ability of reconstruction. Second, just using the first-order statistics would limit the modeling capability, so the second-order covariance statistics need to be concerned to capture spectral correlation. Besides, the insufficient exploration of spectral correlation that is usually described via covariance would lead to spectral distortion [31],

[32]. Therefore, for sake of considering the dependencies among spectral channel, an enhanced HSAM based the mixed attention in the 3-D-MSMAB [31] is proposed. In the squeeze module of mixed attention, the first-order statistics and second-order covariance statistics are explored. In the excitation module, two 3-D convolution layers with size of $1\times1\times1$ are used to generate the weight vector. However, two 3-D convolutions for capturing channel-wise relationships are not accurate. Therefore, we replace the first 3-D convolution with channel-reduction-scaling (CRS) operation. The weights among different spectral channels are adaptively recalibrated, which can improve the learning ability of discriminative spectral features and further enhances the learning ability of the network.

As shown in Fig. 4, the first-order statistics and second-order covariance statistics are explored in a dual-path learning manner. This structure ensures the flexibility to explore new spectral features. Specifically, for the first-order attention, in the squeeze module, the global spatial information z_s is obtained by GAP. In the excitation module, the CRS operation and a 3-D convolution layer with size of $1\times1\times1$ are used to get a B-dimensional attention vector z_1

$$z_1 = f_{\text{Conv3D},1}\ (\mu\,(w_{\text{cs}}\,(\sigma w_{\text{cd}}\,(z_s)))) \tag{17}$$

where w_{cd} denotes the channel reduction layer with reduction scale $s = 16$ and w_{cs} denotes the channel scaling layer. For the second-order covariance statistics, the spectral correlation is explored by second-order covariance pooling. Specifically, we compute the covariance matrix $b_n \in R^{B\times B}$ for the different channels to obtain spectral correlation. Then, the mean of row vector B is calculated to obtain the spectral descriptor z_n. Finally, the CRS operation and a 3-D convolution layer with size of $1\times1\times1$ in the excitation module are performed to obtain the attention vector z_2 as

$$z_2 = f_{\text{Conv3D},1}\ (\mu\,(W_{\text{cs}}\,(\sigma W_{\text{cd}}\,(z_n)))). \tag{18}$$

Then, the weight z_H is obtained by integrating B-dimensional attention vector z_1 and z_2, and a $1\times1\times1$ 3D convolution layer is followed to reduce the number of feature maps

$$z_H = \sigma\,(f_{\text{Conv3D},1}\,(\text{concat}\,[z_1, z_2])). \tag{19}$$

The weight z_H is used to scale the input feature map $F_{d,1re}$ to obtain the F_{MSFE}. Finally, F_{spa} and F_{MSFE} are concatenated in the spectral domain. A 3-D convolution layer for feature maps reduction is followed. Through two residual learning operations, the ability of network representation is improved. The output feature map F_d of DMFB can be described as

$$F_d = (f_{\text{Conv3D},1}\,(\text{concat}\,([F_{\text{spa}}, F_{\text{MSFE}}])) + F_{d,1}) + F_{d-1}. \tag{20}$$

D. MSFB

In the earlier network such as MCNet [28], the output features at each inherited feature level are directly connected by concatenate operation, ignoring the inconsistency of features at each level. Inspired by MSFMNet [42], we introduce channel reduction-scaling operation to treat features at each level differently. The network structure diagram is shown in Fig. 5.

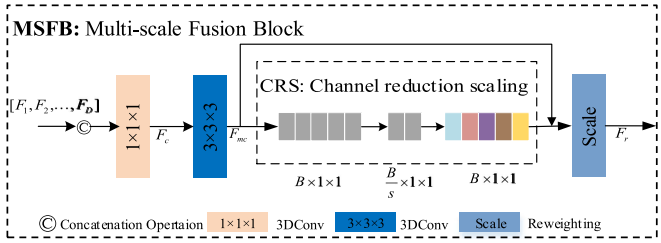

Fig. 5. Structure of the MSFB.

Suppose there are D DMFBs, the ith feature map from DMFB is noted as $F_i \in R^{B\times W\times H}$ where $i \in [1, 2, \ldots, D]$. First, D feature maps are stacked in spectral dimension. A 3-D convolution layer with a size of $1\times1\times1$ is followed for dimensionality reduction processing to obtain $F_c \in R^{B\times W\times H}$

$$F_c = f_{\text{Conv3D},1}\ (\text{concat}\,[F_1, F_2, \ldots, F_D]). \tag{21}$$

Next, a 3-D convolution layer with a size of $3\times3\times3$ is followed to extract spatial–spectral features, which improves the reusability of features to obtain F_{mc}

$$F_{\text{mc}} = f_{\text{Conv3D},1}\ (F_c). \tag{22}$$

The weights of each channel are obtained by CRS operation that differentiates the features of different channels. Finally, the weights are used to scale F_{mc} by the channel-wise multiplication to obtain the F_r.

E. Loss Function

General image restoration tasks mainly take L_1 as loss function [28], [29], which can effectively preserve spatial information. For HSI, the spectral constraint is also significant. A joint L_1 and spatial–spectral TV L_{SSTV} is proposed in [38], which can guarantee both spatial and spectral features. Therefore, we use the same loss function as in the literature [32], which can be described as

$$L_{\text{total}} = L_1 + \alpha L_{\text{SSTV}} \tag{23}$$

where α is a factor used to balance the contributions of the two losses, and we set it to be a constant, $\alpha = 1\times10^{-3}$. The L_1 loss function can be defined as

$$L_1 = \frac{1}{N}\sum_{n=1}^{N} \|I_{\text{HR}}^n - H_{\text{Net}}\,(I_{\text{LR}}^n)\,\|_1 \tag{24}$$

where I_{HR}^n and $H_{\text{Net}}(I_{\text{LR}}^n)$ are the nth ground-truth and reconstructed HSI, respectively. N is the number of training batches. L_{SSTV} can be defined as

$$L_{\text{SSTV}} = \frac{1}{N}\sum_{n=1}^{N} (\nabla_h \|H_{\text{Net}}\,(I_{\text{LR}}^n)\,\|_1 + (\nabla_w \|H_{\text{Net}}\,(I_{\text{LR}}^n)\,\|_1 + \nabla_c \|H_{\text{Net}}\,(I_{\text{LR}}^n)\,\|_1) \tag{25}$$

where ∇_h, ∇_w, and ∇_c are the horizontal, vertical, and spectral gradients of $H_{\text{Net}}(I_{\text{LR}}^n)$.

III. Experiments

In this section, we verify the effectiveness of M3SN qualitatively and quantitatively. Specifically, first, four publicly available datasets are presented. Second, the implementation details and evaluation metrics of the experiments are described. Then, M3SN is compared with other state-of-the-art methods. Finally, the validity analysis of the M3SN is performed.

A. Experimental Settings

1) Datasets: We conduct experiments on four publicly available datasets, CAVE [56], Chikusei [57], Pavia Center, and Botswana. They are collected by different camera sensors and do not have the same properties. The CAVE contains rich synthetic scenes but has few bands, while the Chikusei, Pavia Center, and Botswana are real hyperspectral remote sensing scenes that have more bands but contain a single scene.

The CAVE dataset is collected by the Cooled CCD camera from 400 to 700 nm at 10 nm. It has 31 bands with the spatial resolution of 512 × 512. The dataset consists of 31 scenes covering a wide variety of real-world materials and objects.

The Chikusei dataset is gathered by Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan, on July 29, 2014. It has 128 bands in the spectral range from 363 to 1018 nm, which consists of 2517 × 2335 pixels and the ground sampling distance is 2.5 m.

The Pavia Center dataset is acquired by the reflective optics system imaging spectrometer sensor over Pavia, northern Italy. It is a hyperspectral remote sensing dataset with 1096 × 715 spatial resolution for a geometric resolution of 1.3 m and has 102 spectral bands after removing water absorption bands. The Botswana dataset is collected by Hyperion sensor on NASA EO-1 satellite over the Okavango Delta, Botswana in 2001–2004, which has 30-m pixel resolution in 242 bands over 400–2500 nm. The spatial resolution is 1476 × 256. Finally, 145 bands are remained after removing uncalibrated and noisy bands.

2) Implementation Details: For 32HSIs in CAVE with the size of 31 × 512 × 512, we randomly select 70% of the dataset as the training set, 20% as the validation set, and 10% as the test set. For Chikusei dataset, due to the lack of edge information in the original image, we crop the edge part of the image, and the images with 128 × 2034 × 2048 are obtained. Rows 1–128 are used as test set, which is cut into 16 subimages with size of 128 × 128 × 128. Rows 129–256 are used as the validation set, and the rest is used as the training set. For Pavia Center dataset, rows 1–143 are used as the test set, which is cut into five subimages with size of 102 × 143 × 143, rows 144 – 364 are for the validation set, and the rest are used as the training set. For Botswana dataset, rows 1 – 1034 are used as training set, rows 1035–1347 are used as the validation set, and the rest are test dataset with the size of 145 × 128 × 256.

Data augmentation is performed on training sets of four datasets. For the CAVE, 24 patches are randomly selected from each picture from the training set. For the Chikusei, Pavia Center, and Botswana, 32 patches are randomly selected from the training set. After obtaining patches, each patch is scaled by 1, 0.75, and 0.5. The scaled images are rotated by 270°, 180°, and 90°. Horizontal and vertical mirroring operations are performed. We set the number of filters to 32. The Adam optimizer is utilized in the network, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A total of 200 epochs are trained. The number of batch sizes is set to 12 for CAVE, four for Chikusei, Pavia Center, and Botswana. We initialize the learning rate of all layers to 1×10-4 and halve by every 35 epochs. All experiments are performed on Pytorch framework with NVIDIA GeForce RTX 3090 GPU.

3) Evaluation Metrics: Six widely quantitative image quality assessment metrics are used to evaluate the performance of the M3SN, including peak signal noise ratio (PSNR), structure similarity index measurement (SSIM), spectral angle mapper (SAM), root mean square error (RMSE), erreur relative globale adimensionnelle de synthse (ERGAS), and universal image quality index (UIQI). PSNR, SSIM, and RMSE are commonly used to evaluate the spatial quality of reconstructed images. SAM, ERGAS, and UIQI are usually used to assess image fusion quality. PSNR, SSIM, and UIQI are spatial measurements, the higher values indicate spatial reconstruction is good. SAM is used to measure the spectral quality, the lower the value, the less distortion of spectral reconstruction. ERGAS and RMSE are global measurements, the ideal value is 0. We let the referenced HSI and reconstructed HSI be denoted as $I_{\mathrm{HR}} \in R^{B\times rW\times rH}$ and $I_{\mathrm{SR}} \in R^{B\times rW\times rH}$, respectively, the aforementioned metrics are expressed as

$$\mathrm{PSNR} = \frac{1}{B}\sum_{b=1}^{B} 10\log_{10}\left(\frac{\mathrm{MAX}_b^2}{\mathrm{MSE}_b}\right) \tag{26}$$

$$\mathrm{MSE}_b = \frac{1}{\mathrm{WH}}\sum_{w=1}^{W}\sum_{h=1}^{H}\|I_{\mathrm{HR}} - I_{\mathrm{SR}}\|^2 \tag{27}$$

$$\mathrm{SSIM} = \frac{1}{B}\sum_{b=1}^{B}\frac{2\mu^b_{I_{\mathrm{HR}}}\mu^b_{I_{\mathrm{SR}}} + c_1}{\left(\mu^b_{I_{\mathrm{HR}}}\right)^2 + \left(\mu^b_{I_{\mathrm{SR}}}\right)^2 + c_1}\frac{2\mu^b_{I_{\mathrm{HR}}I_{\mathrm{SR}}} + c_2}{\left(\mu^b_{I_{\mathrm{HR}}}\right)^2 + \left(\mu^b_{I_{\mathrm{SR}}}\right)^2 + c_2} \tag{28}$$

$$\mathrm{SAM} = \cos^{-1}\left(\frac{\langle I_{\mathrm{SR}}, I_{\mathrm{HR}}\rangle}{\|I_{\mathrm{SR}}\|_2\|I_{\mathrm{HR}}\|_2}\right) \tag{29}$$

$$\mathrm{RMSE} = \frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{\sum_{w=1}^{W}\sum_{h=1}^{H}\left(I_{\mathrm{HR}} - I_{\mathrm{SR}}\right)^2}{w\times h}} \tag{30}$$

$$\mathrm{ERGAS} = 100r\sqrt{\frac{1}{B}\sum\nolimits_{b=1}^{B}\left(\frac{\mathrm{RMSE}_b}{\mu^b_{I_{\mathrm{HR}}}}\right)^2} \tag{31}$$

$$\mathrm{UIQI} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sigma^b_{I_{\mathrm{HR}}I_{\mathrm{SR}}}}{\sigma^b_{I_{\mathrm{HR}}}\sigma^b_{I_{\mathrm{SR}}}}\frac{2\mu^b_{I_{\mathrm{HR}}}\mu^b_{I_{\mathrm{SR}}}}{\left(\mu^b_{I_{\mathrm{HR}}}\right)^2 + \left(\mu^b_{I_{\mathrm{SR}}}\right)^2}\frac{2\sigma^b_{I_{\mathrm{HR}}}\sigma^b_{I_{\mathrm{SR}}}}{\left(\sigma^b_{I_{\mathrm{HR}}}\right)^2 + \left(\sigma^b_{I_{\mathrm{SR}}}\right)^2} \tag{32}$$

where $\langle , \rangle$ is the dot product operation, r denotes scale factor, and $\mu^b_{I_{\mathrm{HR}}}$ and $\mu^b_{I_{\mathrm{SR}}}$ represent the mean of I_{HR} and I_{SR} for the bth band, respectively. $\sigma^b_{I_{\mathrm{HR}}}$ and $\sigma^b_{I_{\mathrm{SR}}}$ are the variance of I_{HR} and I_{SR} for the bth band, and $\sigma^b_{I_{\mathrm{HR}}I_{\mathrm{SR}}}$ is the covariance of I_{HR} and I_{SR} for the bth band. c_1 and c_2 are two constants.

TABLE I
QUANTITATIVE EVALUATION ON THE CAVE DATASET OF STATE-OF-THE-ART ALGORITHMS BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI FOR SCALE FACTORS 2, 3, AND 4

Scale factor		Bicubic	CNMF	HySure	LTTR	SRCNN	MCNet	SSPSR	ERCSR	M3SN
×2	PSNR	41.252	41.027	41.249	40.861	42.548	42.575	43.921	**45.079**	**45.429**
	SSIM	0.9837	0.9805	0.9822	0.9809	0.9853	0.9859	0.9884	**0.9890**	**0.9891**
	SAM	3.134	4.459	3.813	3.547	3.570	3.118	2.835	**2.729**	**2.737**
	RMSE	0.036	0.038	0.036	0.038	0.032	0.032	0.026	**0.024**	**0.024**
	ERGAS	6.098	4.733	4.619	4.795	9.132	10.646	**4.139**	**4.251**	5.260
	UIQI	0.981	0.969	0.973	0.979	0.980	0.982	0.986	**0.987**	**0.987**
×3	PSNR	37.024	36.885	37.022	37.226	39.133	37.242	/	**41.671**	**41.700**
	SSIM	0.9678	0.9630	0.9658	0.9648	0.9729	0.9689		**0.9812**	**0.9812**
	SAM	3.793	5.306	4.520	3.969	4.247	3.908		**3.190**	**3.191**
	RMSE	0.055	0.058	0.055	0.054	0.045	0.053		**0.035**	**0.035**
	ERGAS	81.359	8.441	17.716	58.321	6.704	18.746		**5.006**	**4.971**
	UIQI	0.971	0.954	0.954	0.969	0.970	0.970		**0.981**	**0.981**
×4	PSNR	34.991	34.998	34.977	36.058	36.829	35.313	39.161	**39.298**	**39.486**
	SSIM	0.9516	0.9493	0.9479	0.9558	0.9581	0.9534	0.9726	**0.9737**	**0.9739**
	SAM	4.190	5.157	5.239	4.288	5.045	4.315	3.625	**3.513**	**3.521**
	RMSE	0.071	0.072	0.072	0.063	0.059	0.069	0.045	**0.044**	**0.044**
	ERGAS	10.233	8.751	8.242	10.102	6.790	15.574	7.468	**5.392**	**5.240**
	UIQI	0.962	0.938	0.928	0.963	0.959	0.962	0.977	**0.977**	**0.977**

4) Competing Methods: Eight competitive methods that consist of four traditional methods and four CNN-based methods are compared with our M3SN, including Bicubic, CNMF [17], HySure [16], LTTR [18], SRCNN [24], MCNet [28], SSPSR [38], and ERCSR [29]. Among these, Bicubic is a classical interpolation method. CNMF based on matrix decomposition, HySure based on Bayesian learning, and LTTR based on representation learning are the traditional fusion-based methods. SRCNN is the first CNN-based SR method for natural images. MCNet, SSPSR, and ERCSR are the state-of-the-art CNN-based HSI-SR methods with competing performance.

It should be noted that CNMF, HySure, and LTTR belong to the fusion-based methods that require auxiliary HR images such as PAN or MSI. Therefore, it would be unfair to just feed them auxiliary HR images in inference phase. To let all methods accept the same input information, we just use the bicubic interpolation to up-sample the I_{LR} to the desired auxiliary HR images for CNMF, HySure, and LTTR methods [40].

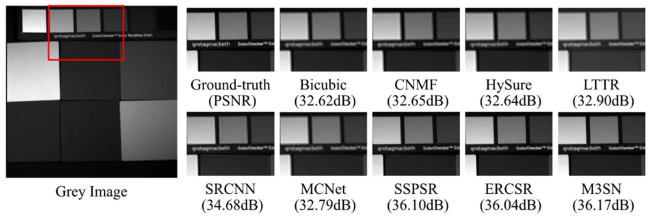

Fig. 6. Visual results of spatial SR of sponges_ms for the 20th band on the CAVE dataset at scale factor 4. (The parentheses indicate the PSNR value.)

B. Comparisons to State of the Arts

We compare M3SN with eight existing methods, including Bicubic, CNMF, HySure, LTTR, SRCNN, MCNet, SSPSR, and ERCSR. Four benchmark datasets CAVE, Harvard, Pavia Center, and Botswana are employed to verify the effectiveness of the M3SN for different scale factors.

1) Experiments on the Synthesized CAVE Dataset: Table I presents the quantitative evaluation results for scale factors 2, 3, and 4. Red represents the best result and blue represents the second-best result. At scale factor 2, the M3SN has achieved the best results on PSNR, SSIM, RMSE, and UIQI, while the second-best result on SAM. As the scale factor increases, M3SN has achieved the best results on PSNR, SSIM, RMSE, ERGAS, and UIQI, while achieved the second-best result on SAM.As for the fusion-based methods, the CNMF, HySure, and LLTR have achieved the poor performance, mainly because there are no auxiliary HR images. As for CNN- based methods, the SRCNN ignores the spectral dimension, the spatial–spectral information is not underutilization. Although MCNet considers spatial and spectral dimensions, it does not fully utilize 2-D units just with a simple stack. SSPSR adopts RBs that is based on 2-D network, leading to explore spatial and spectral information not effectively. ERCSR achieves second-best results by cross-utilizing spatial–spectral information, the second-order spectral correlation is not explored. In contrast, our M3SN has achieved the best result in the spatial indexes such as PSNR and SSIM at all scale factors, mainly because it can fully consider the multiscale shallow and deep spatial–spectral features, explore more effective spatial features, which is important for improving the sharpness of reconstruction results.

Figs. 6–9 also support the quantitative results. The reconstructed HSI of various algorithms and detailed comparison

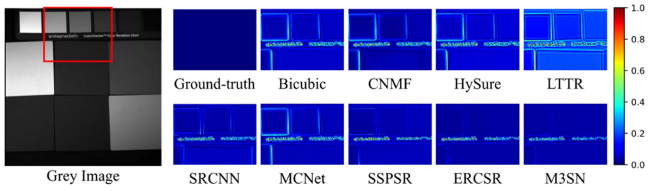

Fig. 7. Absolute error map of sponges_ms at scale factor 4.

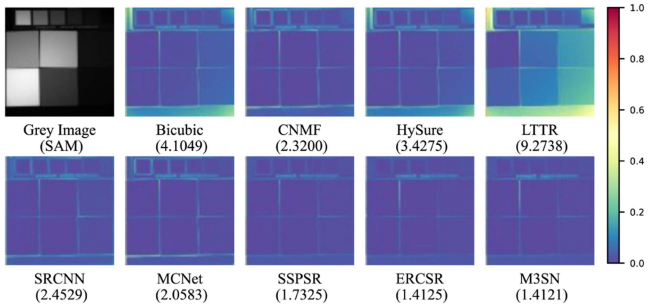

Fig. 8. Spectral angle visualization map of sponges_ms at scale factor 4. (The parentheses indicate the SAM value.)

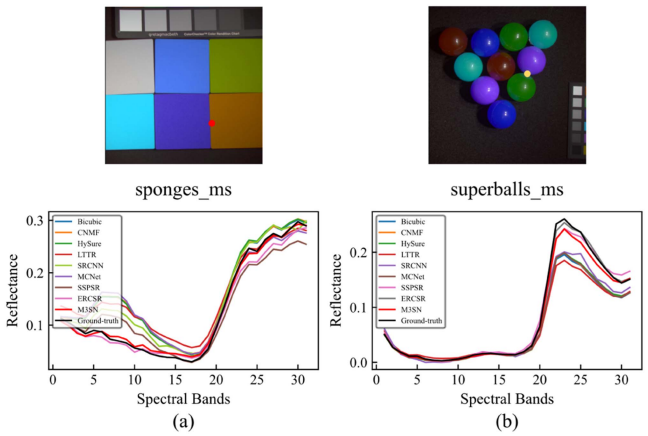

Fig. 9. Sample spectral curves of the two pictures were selected on the CAVE at scale factor 4. (a) Spectrum of the red point. (b) Spectrum of the yellow point.

images for the 20th band of sponges_ms at scale factor 4 is shown in Fig. 6. The area marked with a red rectangle shows where we select to enlarge. As can be intuitively seen from Fig. 6, in the image restored by the M3SN model, the edges of rectangles with different gray levels are distinct, without any blurring, and the overall visual effect is the sharpest with well-defined edges. To better illustrate the spatial quality of the reconstructed images, we also calculate the absolute error map between the reconstructed and referenced HSI by various algorithms. The bluer the images, the smaller the absolute error, and the better the reconstruction quality. As shown in Fig. 7, the M3SN recovers the detailed information better and obtains a lower absolute error. Compared with other algorithms, M3SN has the largest blue area in the reconstructed image, especially at the edge and character positions of the color chart rectangle. This indicates that M3SN can well reconstruct the high-frequency

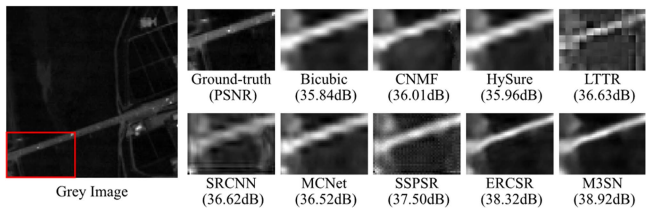

Fig. 10. Visual results of spatial SR of Test area 7 for the fifth band on the Chikusei dataset at scale factor 4. (The parentheses indicate the PSNR value.)

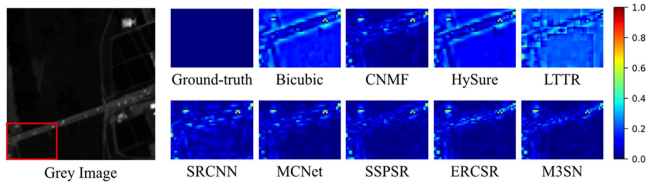

Fig. 11. Absolute error map of Test area 7 band on the Chikusei dataset at scale factor 4.

information in the image. As shown in Fig. 7, the M3SN recovers the detailed information better and obtains a lower absolute error. Specifically, the square-frame artifacts in LTTR demonstrated that the fusion-based approaches have a strong dependence on the auxiliary HR images. The reconstructed images of Bicubic, SRCNN, MCNet, and SSPSR have blurred edges.

For HSI-SR, keeping the spectrum undistorted is also a measure of model performance. To this end, the SAM value of each pixel is visualized more intuitively at scale factor 4 as shown in Fig. 8. However, just using the SAM value for illustrating the spectral distortion is limited [58]. For a better illustration, we take the spectral curve as a supplementary description. The spectral curve of selected pixel in the sponges_ms and superballs_ms are shown in Fig. 10. As seen, the spectral curve is closer to the corresponding ground truth, which shows that the M3SN achieves good spectral fidelity.

2) Experiments on the Real-Scenario Chikusei Dataset: To comprehensively demonstrate the effectiveness of the M3SN, the Chikusei dataset that belongs to hyperspectral remote sensing data is used to verify the performance of real HSI. Table II shows the quantitative evaluation results of various algorithms at scale factors 2, 3, and 4. Compared with existing methods, our M3SN achieves superior performance in all evaluation criteria for the different scale factors, especially at scale factor 2, six evaluation metrics ($+0.226$dB, $+0.0008$, -0.0510, 0, -0.043, $+0.001$) are significantly better than the second performance algorithm ERCSR. The visual results of the spatial SR of Test area 7 are displayed in Fig. 10. The lower left area is selected to display. It can be seen that our M3SN achieves a clearer road edge, whereas other competing methods fail to recover the details. The corresponding absolute error maps are shown in Fig. 11. As seen, the M3SN has the slightest absolute error, which indicates a better reconstruction ability. The SAM value shown in Fig. 12 can reflect the spectral distortion of different algorithms. The 58th, 35th, and 22nd spectral bands is used to synthesize the pseudocolor images for Test area 1 and Test area

TABLE II
QUANTITATIVE EVALUATION ON THE CHIKUSEI DATASET OF STATE-OF-THE-ART ALGORITHMS BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI FOR SCALE FACTORS 2, 3, AND 4

Scale factor		Bicubic	CNMF	HySure	LTTR	SRCNN	MCNet	SSPSR	ERCSR	M3SN
×2	PSNR	42.360	42.059	41.808	40.389	42.692	43.729	43.210	**45.178**	**45.404**
	SSIM	0.9716	0.9712	0.9722	0.9532	0.9745	0.9796	0.9780	**0.9849**	**0.9860**
	SAM	1.859	2.212	2.396	2.920	2.749	1.730	2.325	**1.495**	**1.444**
	RMSE	0.010	0.010	0.010	0.012	0.009	0.008	0.009	**0.007**	**0.007**
	ERGAS	3.220	2.492	2.551	3.018	2.317	2.063	2.199	**1.768**	**1.725**
	UIQI	0.993	0.990	0.986	0.983	0.989	0.994	0.993	**0.996**	**0.997**
×3	PSNR	38.655	38.609	38.427	39.317	39.273	39.244	/	**40.638**	**40.671**
	SSIM	0.9337	0.9351	0.9377	0.9413	0.9434	0.9428		**0.9565**	**0.9569**
	SAM	2.815	3.026	3.596	2.816	3.365	2.724		**2.378**	**2.375**
	RMSE	0.014	0.014	0.015	0.013	0.013	0.013		**0.011**	**0.011**
	ERGAS	4.926	3.711	3.779	3.435	3.434	3.457		**3.000**	**2.991**
	UIQI	0.982	0.980	0.967	0.981	0.980	0.984		**0.989**	**0.989**
×4	PSNR	36.633	36.410	36.790	36.922	36.986	37.085	37.511	**38.248**	**38.358**
	SSIM	0.8978	0.8974	0.9050	0.9013	0.9059	0.9073	0.9166	**0.9260**	**0.9279**
	SAM	3.617	3.803	3.818	4.047	4.083	3.511	3.623	**3.107**	**3.020**
	RMSE	0.018	0.019	0.018	0.018	0.018	0.018	0.017	**0.015**	**0.015**
	ERGAS	6.202	4.794	4.567	4.508	4.480	4.422	4.240	**3.940**	**3.899**
	UIQI	0.972	0.968	0.961	0.964	9.967	0.973	0.976	**0.980**	**0.982**

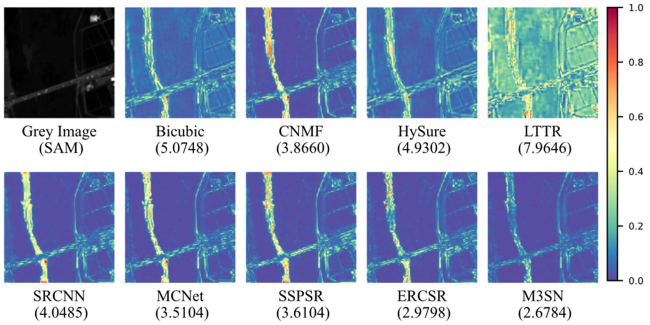

Fig. 12. Spectral angle visualization map of the Test area 7 at scale factor 4. (The parentheses indicate the SAM value.)

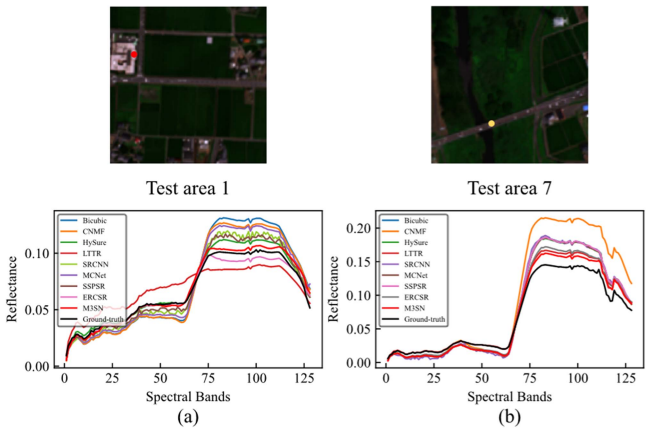

Fig. 13. Sample spectral curves of the two pictures were selected on the Chikusei dataset at scale factor 4. The first line is pseudo color image which is synthesized by the 58th, 35th, and 22nd bands. (a) Spectral curve of the red point. (b) Spectral curve of the yellow point.

7 as shown in Fig. 13. We can find that our M3SN achieves the best spectral fidelity, while other competing methods result in some obvious spectral distortion. Besides, the spectral curve of the selected pixel is depicted. It can be seen that M3SN has a very low SAM, and the spectral curve is closer to the ground-truth value, indicating that M3SN has a good advantage in maintaining spectral fidelity.

3) Experiments on the Real-Scenario Pavia Center Dataset: We further conduct experiments on the Pavia Center dataset to verify our M3SN performance on different sensors. Similar to the Chikusei dataset, Table III shows the quantitative evaluation results of various algorithms for different scale factors. Our M3SN exhibits excellent performance at all scale factors, indicating that M3SN significantly outperforms other competing methods in terms of spatial resolution improvement and spectral fidelity. Figs. 14–17 show the qualitative results. Fig. 14 is a visual representation of the reconstructed HSI of Test area 4 on the Pavia Center. Our M3SN achieves a clearer edge than other algorithms. Fig. 15 shows the absolute error map where M3SN achieves the smallest error. In Figs. 16 and 17, the SAM is visualized and the spectral curves of selected pixels are displayed. It can be seen that M3SN obtains a lower SAM value, and the spectral curves are closer to ground truth. That is, M3SN has superior spectral fidelity.

4) Experiments on the Real-Scenario Remote Sensing Botswana Dataset: Further experiments are performed on the Botswana dataset to verify the performance of the M3SN on real-scenario remote sensing HSIs. The two traditional methods (i.e., Bicubic and HySure) and two CNN-based methods (i.e.,

TABLE III
QUANTITATIVE EVALUATION ON THE PAVIA CENTER DATASET OF STATE-OF-THE-ART ALGORITHMS BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI FOR SCALE FACTORS 2, 3, AND 4

Scale factor		Bicubic	CNMF	HySure	LTTR	SRCNN	MCNet	SSPSR	ERCSR	M3SN
×2	PSNR	32.982	32.883	33.425	32.641	34.001	34.279	34.424	**35.835**	**35.847**
	SSIM	0.9183	0.9145	0.9249	0.9066	0.9315	0.9391	0.9390	**0.9539**	**0.9541**
	SAM	4.854	5.726	5.272	5.753	5.986	4.908	5.209	**4.317**	**4.311**
	RMSE	0.149	0.152	0.142	0.156	0.134	0.129	0.127	**0.108**	**0.108**
	ERGAS	5.575	4.329	4.030	4.367	3.807	3.637	3.619	**3.074**	**3.072**
	UIQI	0.980	0.969	0.972	0.976	0.976	0.983	0.982	**0.988**	**0.988**
×3	PSNR	29.949	29.953	30.280	30.656	30.864	30.632	/	**31.730**	**31.821**
	SSIM	0.8358	0.8364	0.8461	0.8540	0.8653	0.8607		**0.8879**	**0.8895**
	SAM	5.917	6.549	6.266	6.073	6.655	6.176		**5.555**	**5.489**
	RMSE	0.212	0.213	0.205	0.196	0.192	0.196		**0.173**	**0.171**
	ERGAS	7.825	5.964	5.707	5.422	5.340	5.456		**4.807**	**4.747**
	UIQI	0.963	0.953	0.950	0.964	0.963	0.966		**0.975**	**0.976**
×4	PSNR	28.214	28.253	28.304	29.125	29.030	28.739	29.288	**29.581**	**29.599**
	SSIM	0.7607	0.7647	0.7621	0.7984	0.8010	0.7878	0.8113	**0.8212**	**0.8217**
	SAM	6.638	7.139	7.181	6.776	7.231	6.879	6.548	**6.342**	**6.335**
	RMSE	0.259	0.259	0.257	0.233	0.236	0.244	0.228	**0.221**	**0.220**
	ERGAS	9.521	7.242	7.194	6.440	6.544	6.745	6.382	**6.115**	**6.095**
	UIQI	0.947	0.934	0.918	0.953	0.948	0.950	0.958	**0.961**	**0.962**

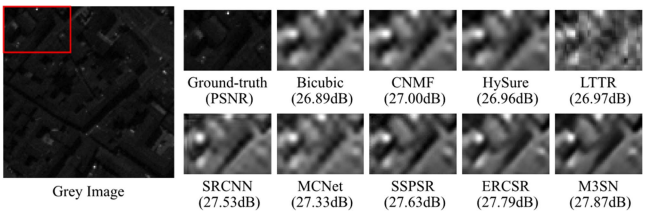

Fig. 14. Visual results of spatial SR of Test area 4 for the 20th band on the Pavia Center dataset at scale factor 4. (The parentheses indicate the PSNR value.)

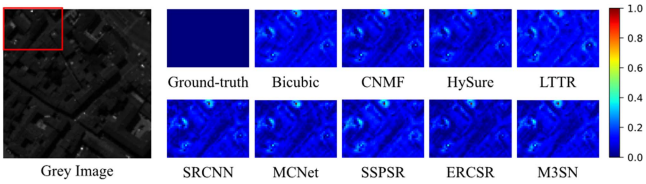

Fig. 15. Absolute error map of Test area 4 on the Pavia Center dataset at scale factor 4.

SRCNN andERCSR) are selected for comparison. The quantitative results in Table IV show that the performance of our M3SN is best in all scales. Figs. 18 and 19 show the visual results of spatial SR and the absolute error map, respectively, which also demonstrates that the M3SN achieves the best result. Furthermore, in Figs. 20 and 21, our M3SN obtains lowest spectral distortion and the spectral curves of selected pixels are closer to the ground truth. The Bicubic, HySure, and SRCNN have obvious spectral distortion. In all, M3SN achieved best

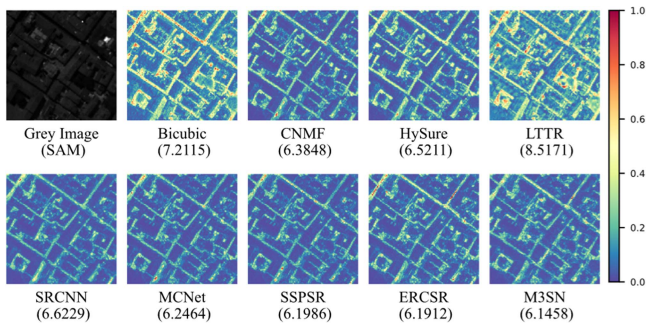

Fig. 16. Spectral angle visualization map of Test area 4 on the Pavia Center dataset at scale factor 4. (The parentheses indicate the SAM value.)

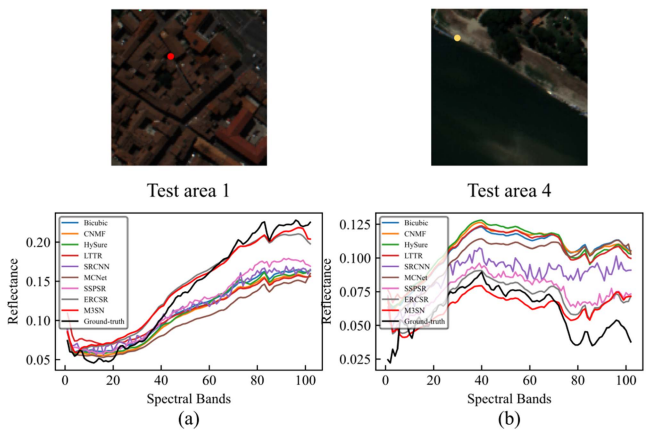

Fig. 17. Sample spectral curves on the Pavia Center dataset at scale factor 4. (a) Spectral curve of the red point. (b) Spectral curve of the yellow point.

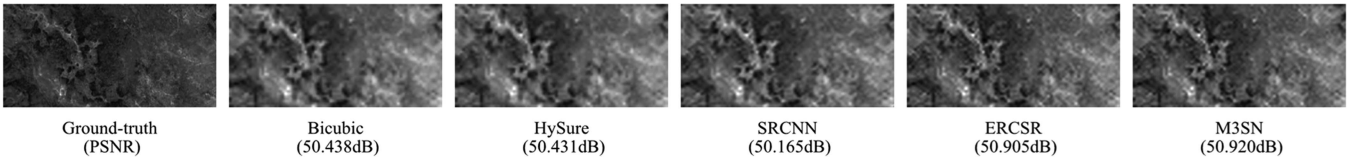

Fig. 18. Visual results of spatial SR of Test area 1 for the 20th band on the Botswana dataset at scale factor 3. (The parentheses indicate the PSNR value.)

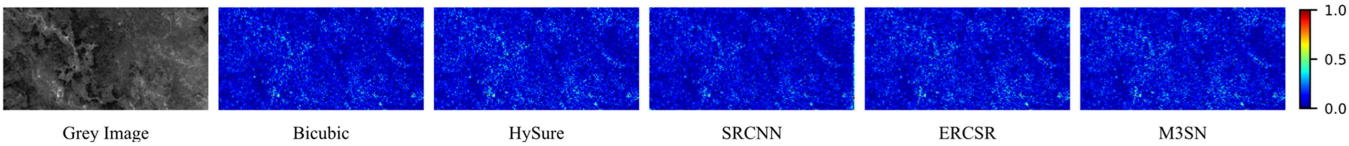

Fig. 19. Absolute error map of Test area 1 on the Botswana dataset at scale factor 3.

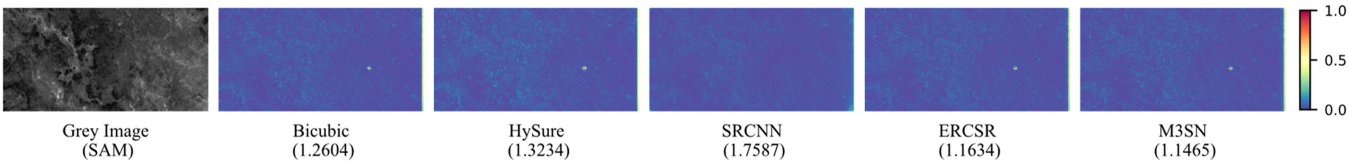

Fig. 20. Spectral angle visualization map of Test area 1 on the Botswana dataset at scale factor 3. (The parentheses indicate the SAM value.)

TABLE IV
QUANTITATIVE EVALUATION ON THE BOTSWANA DATASET OF STATE-OF-THE-ART ALGORITHMS BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI FOR SCALE FACTORS 2, 3, AND 4

Scale factor		Bicubic	HySure	SRCNN	ERCSR	M3SN
×2	PSNR	51.561	51.551	50.350	**52.352**	**52.399**
	SSIM	0.9942	0.9943	0.9936	**0.9951**	**0.9952**
	SAM	1.207	1.296	1.759	**1.153**	**1.147**
	RMSE	0.006	0.006	0.007	**0.006**	**0.006**
	ERGAS	1.560	1.563	1.807	**1.421**	**1.411**
	UIQI	0.995	0.995	0.993	**0.996**	**0.996**
×3	PSNR	50.438	50.431	50.165	**50.905**	**50.920**
	SSIM	0.9910	0.9911	0.9909	**0.9919**	**0.9920**
	SAM	1.307	1.362	1.701	**1.276**	**1.270**
	RMSE	0.008	0.008	0.008	**0.007**	**0.007**
	ERGAS	1.745	1.747	1.796	**1.653**	**1.650**
	UIQI	0.997	0.997	0.996	**0.998**	**0.998**
×4	PSNR	48.652	48.649	48.290	**49.033**	**49.048**
	SSIM	0.9889	0.9890	0.9888	**0.9897**	**0.9897**
	SAM	1.564	1.609	1.843	**1.537**	**1.534**
	RMSE	0.009	0.009	0.009	**0.009**	**0.009**
	ERGAS	2.514	2.610	2.242	**2.068**	**2.066**
	UIQI	0.993	0.993	0.992	**0.994**	**0.994**

results in both spatial and spectral dimensions on the remote sensing Botswana HSIs.

5) Sensitivity Analysis of Different Sizes Images: Our preliminary analysis indicates that although we mainly focus on the SR

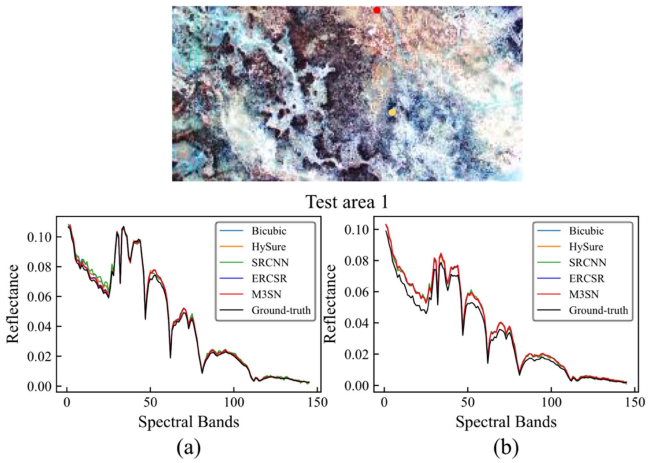

Fig. 21. Sample spectral curves on the Botswana dataset at scale factor 3. (a) Spectral curve of the red dot. (b) Spectral curve of the yellow dot.

reconstruction of small image areas, the network structure can theoretically handle input images of any size. This is because our network structure adopts the design principle of layer-by-layer convolution, aiming to learn the mapping relationship between local image features. The relationship should remain consistent across different image sizes. To verify this opinion, the experimental conclusions on the Houston dataset at a scale factor of 4 were validated. We fed the large areas and small areas into the network for inference, the results and efficiency are consistent with existing experimental conclusions. Specifically, in the inference stage, we used two modes: the overall mode and the patch mode. In the overall mode, a large image with

TABLE V
QUANTITATIVE EVALUATION ON THE HOUSTON DATASET OF STATE-OF-THE-ART ALGORITHMS FOR TWO MODES BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI AT SCALE FACTORS 4

	Mode	Bicubic	CNMF	HySure	LTTR	SRCNN	MCNet	SSPSR	ERCSR	M3SN
PSNR	overall	36.987	36.99	36.946	36.553	37.804	37.525	38.154	**38.558**	**38.562**
	patch	37.084	37.14	37.071	36.756	37.805	37.577	38.159	**38.616**	**38.639**
	▲	0.003	0.004	0.003	0.006	0.000	0.001	0.000	0.002	0.002
SSIM	overall	0.9269	0.9285	0.9278	0.9087	0.9383	0.9337	0.9408	**0.9440**	**0.9446**
	patch	0.9271	0.9293	0.9284	0.9130	0.9382	0.9337	0.9409	**0.9441**	**0.9450**
	▲	0.0002	0.0009	0.0006	0.0047	-0.0001	0.0000	0.0001	0.0001	0.0004
SAM	overall	2.422	2.557	2.527	2.979	2.515	2.310	2.312	**2.104**	**2.092**
	patch	2.426	2.500	2.559	2.864	2.559	2.330	2.342	**2.114**	**2.100**
	▲	0.002	0.023	0.013	0.040	0.017	0.008	0.013	0.005	0.004
RMSE	overall	0.070	0.070	0.070	0.074	0.064	0.066	0.061	**0.058**	**0.058**
	patch	0.066	0.065	0.066	0.069	0.061	0.062	0.058	**0.055**	**0.055**
	▲	0.057	0.068	0.057	0.068	0.047	0.061	0.049	0.056	0.056
ERGAS	overall	2.597	2.594	2.604	2.734	2.357	2.452	2.278	**2.172**	**2.168**
	patch	2.623	2.606	2.632	2.720	2.406	2.487	2.324	**2.207**	**2.201**
	▲	0.010	0.005	0.011	0.005	0.020	0.014	0.020	0.016	0.015
UIQI	overall	0.991	0.991	0.991	0.990	0.993	0.992	0.994	**0.994**	**0.994**
	patch	0.991	0.991	0.991	0.991	0.993	0.993	0.994	**0.994**	**0.994**
	▲	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000

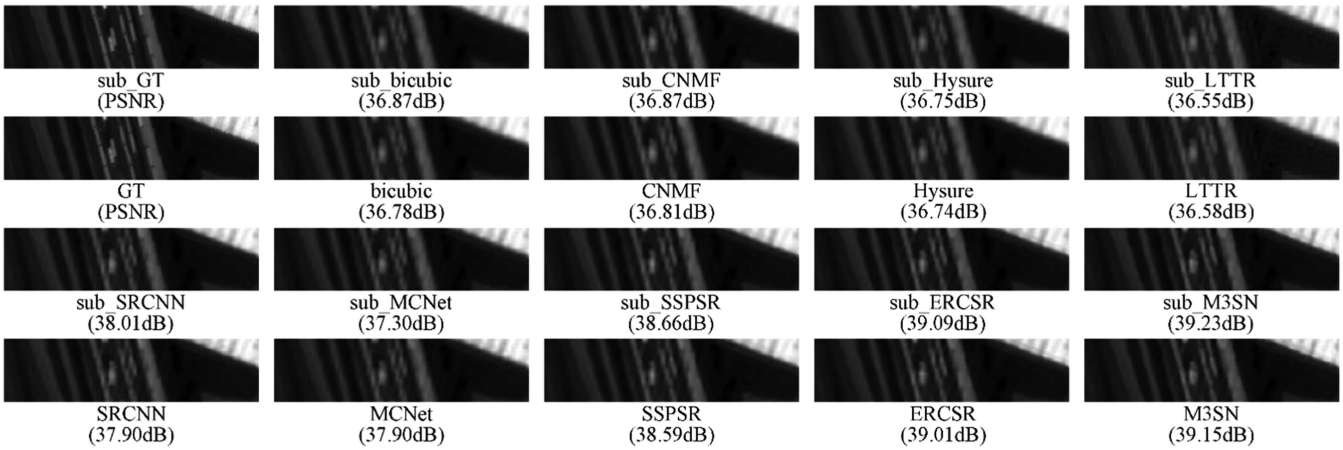

Fig. 22. Visual results of spatial SR of subimage and the corresponding areas in the large images for the 5th band on the Houston dataset at scale factor 4. (The parentheses indicate the PSNR value.)

the size of 512 × 512 × 50 was directly input into the network for inference. In the patch mode, the large image was divided into four small subimages with the size of 128 × 512 × 50 for inference, and the evaluation values were then averaged. We compared the accuracy and error of the two modes. As shown in Table V, ▲ represents the relative error, the relative error of the PSNR is between 0 and 0.006, the relative error of SSIM is between 0.0001 and 0.0009, the relative error of SAM is between 0 and 0.023, the relative error of RMSE is between 0.047 and 0.068, the relative error of ERGAS is between 0and 0.020, and the relative error of UIQI is between 0 and 0.002. All the relative errors are within acceptable ranges. Therefore, the impact of inputting the whole large image and dividing it into small patches on image SR quality can be ignored. Besides, our proposed method is superior or equal to the second-best method ERCSR in six evaluation indicators, which further demonstrates the superiority of the M3SN in improving hyperspectral image SR tasks. In addition, we presented the visual results of spatial SR of subimages (sub_M3SN.etc) with the size of 64 × 128 × 50 and the corresponding areas in the large images (M3SN.etc) in Fig. 22. Furthermore, the absolute error maps are shown in Fig. 23, while the SAM visualization maps are shown in Fig. 24. It can be seen from Fig. 22 that the M3SN has significant advantages in preserving subtle structures in

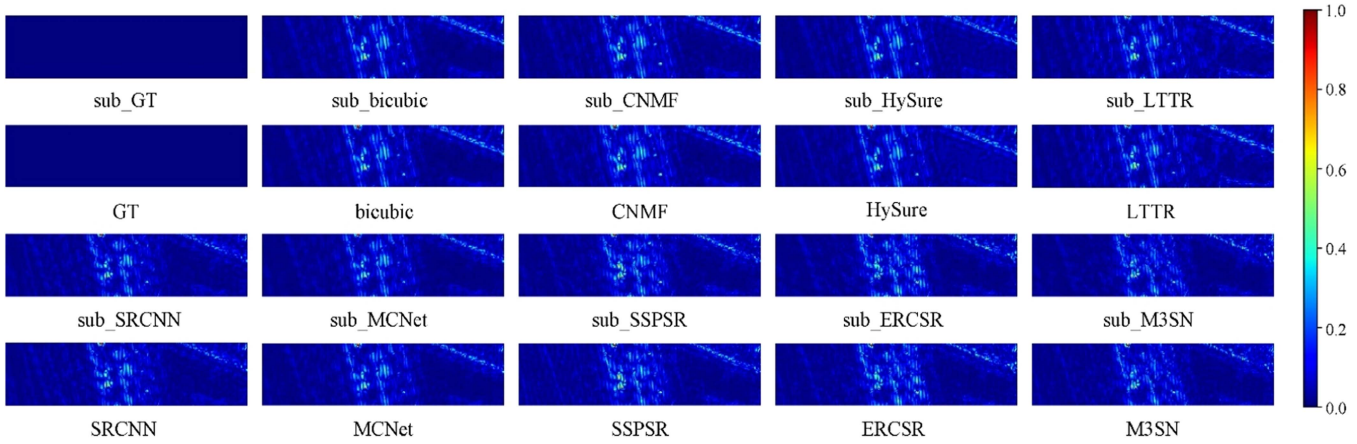

Fig. 23. Absolute error map of subimages and the corresponding areas in the large images on the Houston dataset at scale factor 4.

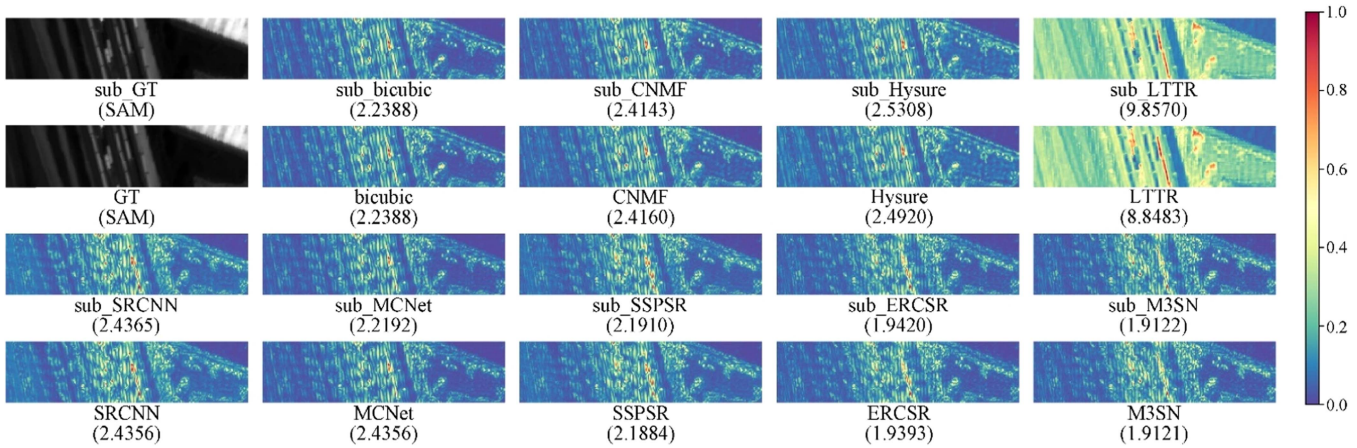

Fig. 24. Spectral angle visualization map of subimage and the corresponding areas in the large images on the Houston dataset at scale factor 4. (The parentheses indicate the SAM value).

images, such as roads, and furthermore, M3SN has minimum spectral distortion. Moreover, a proximate PSNR and SAM across different input image sizes suggests that the quality of the reconstructed image remains relatively stable regardless of the input size. Therefore, the impact of input image size on the inference results is minimal. In short, the proposed method M3SN is superior to the competitors in both spatial information recovery and spectral information preservation.

6) Time Analysis: We compare the prediction time for different algorithms (deep-learning-based methods) on the Chikusei test dataset at scale factor 2. The test dataset consists of 16 HSIs with a size of $128 \times 64 \times 64$. We calculate the mean PSNR of the whole dataset. Fig. 25 shows the relationship between time and PSNR.

We calculate the mean PSNR of the whole dataset. Fig. 25 shows the relationship between time and PSNR. As seen, M3SN has obtained the highest PSNR value with acceptable time. Because the second-order attention mechanism is introduced into M3SN, which not only improves the learning ability of discriminative spectral features but also increases the amount of calculation.

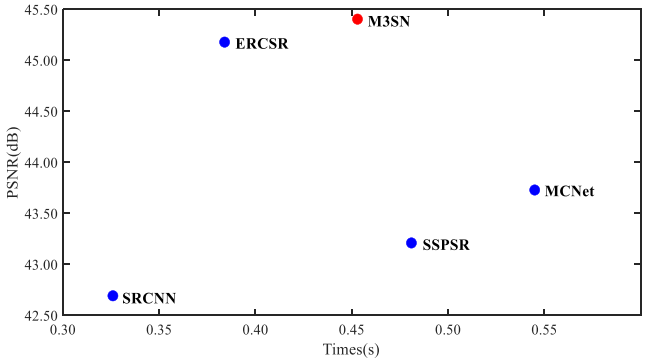

Fig. 25. Time-PSNR on the Chikusei dataset at scale factor 2.

7) Study of DMFB: The DMFB is the main component of the M3SN. Therefore, we verify the effectiveness by analyzing its components: MSFE and SpaFE. The experiments are conducted on the Pavia Center at scale factor 4. First, we verify the case of just MSFE, just SpaFE, and both (MSFE+SpaFE). Table VI details the experimental performance under the same parameter

TABLE VI
QUANTITATIVE EVALUATION ON THE PAVIA CENTER DATASET OF ABLATION INVESTIGATIONS BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI AT SCALE FACTOR 4

Type	PSNR	SSIM	SAM	RMSE	ERGAS	UIQI
MSFE	29.584	0.8215	6.363	0.221	6.109	0.962
SpaFE	29.555	0.8206	6.330	0.221	6.141	0.961
MSFE+SpaFE	**29.599**	**0.8217**	**6.335**	**0.220**	**6.095**	**0.962**

TABLE VII
QUANTITATIVE EVALUATION ON THE PAVIA CENTER DATASET OF DIFFERENT ATTENTION MECHANISM BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI AT SCALE FACTOR 4

Type	PSNR	SSIM	SAM	RMSE	ERGAS	UIQI
R3D_1	29.592	0.8215	6.361	0.221	**6.094**	0.962
R3D_2	29.588	0.8214	6.342	0.221	6.108	0.962
R3D_H	**29.599**	**0.8217**	**6.335**	**0.220**	6.095	**0.962**
R3D_M	29.576	0.8214	6.367	0.221	6.122	0.962

settings. In the case of MSFE, the effect worsens because the network does not mine enough spatial information. In the case of SpaFE, the performance becomes terrible due to SpaFE only exploring spatial information and completely ignoring the spectral information. The performance of MSFE+SpaFE is greatly improved. That is, dual-path learning enables spatial–spectral features and spatial features to complement each other. It also improves the learning ability of the spatial domain at the same time. Second, the main component M3SHS of MSFE is verified. We verify the cases of only 3D-Res2Net (R3D) and 3D-Res2Net with different attention mechanisms, including the first-order attention mechanism (R3D_1), second-order attention mechanism branch (R3D_2), and HSAM (R3D_H). Besides, since our HSAM is improved from the mixed attention in the 3-D-MSMA, we also verify the case of 3D-Res2Net and mixed attention (R3D_M). Keeping other components of the network and parameter settings unchanged, we conduct the experiments on the Pavia Center.

The experimental results are shown in Table VII. R3D_H achieves the best result. Regarding SAM, R3D_2 is 0.019 lower than R3D_1, indicating that the second-order attention mechanism has a stronger ability to explore spectral correlations. R3D_H is 0.026 lower than R3D_1 and 0.007 lower than R3D_2, indicating that HSAM has a stronger ability to improve the learning ability of discriminative spectral features. Compared to R3D_M, we can see that our HSAM has achieved excellent results, which demonstrated that HSAM overcomes the limitation in the 3-D-MSMA and its ability of modeling the relationship is enhanced.

8) Parameter Analysis: In this section, we analyze the number of parallel layers in the M3SIM on four datasets. The basic parameters include the number of parallel layers (denote as N for short) used in the M3SIM and the number of DMFBs (denote as D for short) used in the DFE. First, we let N change from 2 to 5 by copying and removing the medium branch layer. We display the curves of PSNR, SSIM, and SAM of different layers for all

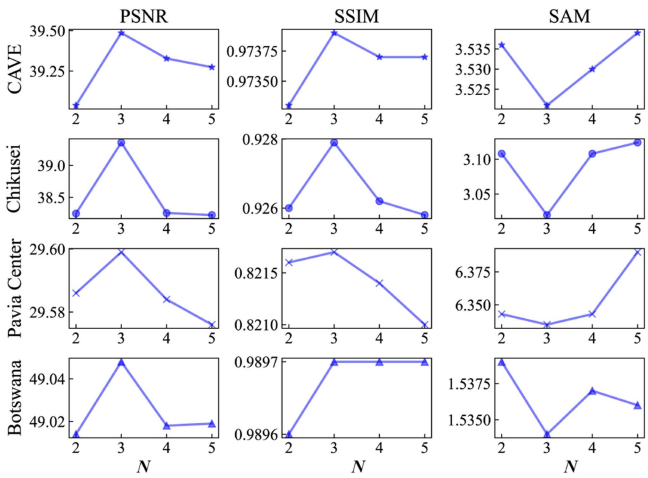

Fig. 26. Curves of PSNR, SSIM, and SAM for different number of parallel layers in the MSIM on the four datasets.

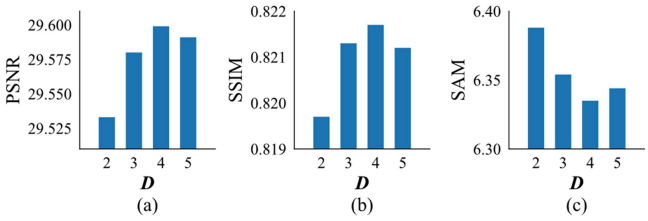

Fig. 27. Histogram of PSNR, SSIM, and SAM for different number D of DMFBs in the DFE on the Pavia Center dataset.

TABLE VIII
QUANTITATIVE EVALUATION ON THE PAVIA CENTER DATASET OF DIFFERENT ATTENTION MECHANISM BY AVERAGE PSNR /SSIM/SAM/RMSE/ERGAS/UIQI AT SCALE FACTOR 4

Architecture	1	2	3	4	5
DMFB	√	√	√	√	√
LSC	×	√	√	√	√
Nearest	×	×	√	√	√
MSIB	×	×	×	√	√
MSFB	×	×	×	×	√
PSNR	29.479	29.502	29.572	29.594	**29.599**

datasets as depicted in Fig. 26. As seen, the inflection point $N = 3$ achieved the best performance.

In addition, we let D change from 2 to 5 for comparison. The experimental results are shown in Fig. 27. As the number D of DMFB increases from 2 to 4, the performance of the network goes better. However, the performance is decreasing from 4 to 5. Increasing the number of DMFB does not improve the performance of the network. That is, the performance of the network tends to be saturated for $D = 4$ DMFBs. As a result, we set $N = 3$ and $D = 4$ in the experiments.

9) Ablation Investigation: In this section, different component combinations are set to analyze the performance of our M3SN. Table VIII presents the results of the ablation investigation on the effect of DMFB, LSC, Nearest, MSIB, and MSFB.

Through these operations, the contribution of each component to the network performance can be drawn. The network containing only DMFB produces the worse performance. Then, the LSC and Nearest operation are added to the network, the value of the PSNR is increased greatly, which means that the LSC can improve the representative ability and the Nearest operation can alleviate the burden of the network effectively. When we add MSIB to the network, the performance is further improved, which confirms the importance of obtaining multiscale shallow spatial–spectral features at the early stage of the network. Finally, we add MSFB to the network, the PSNR gets better, that is, MSFB effectively improves the utilization of different levels of features.

IV. Conclusion

In this article, a novel HSI-SR learning network M3SN is proposed. It integrates hybrid 3-D and 2-D convolutions, 3D-Res2Net, and HSAM. The multiscale spectral–spatial features can be effectively extracted by hybrid 3-D convolution and 2-D convolutions of different sizes kernel. Guided by SE-Net, multiscale shallow spatial–spectral features obtained at earlier stages are integrated to obtain more representative spatial–spectral feature learning. The multiscale deep spatial–spectral features and spatial features are learned separately in a dual-path learning way, strengthening the learning ability of spatial domain. When 3D-Res2Net is equipped with HSAM simultaneously driven by the first-order spectral statistics and the second-order covariance statistics, the ability of discriminant spectral features is learned to improve the reconstruction ability. Sufficient ablation experiments verify the contribution of each component to the network performance. Intensive experimental results show that our M3SN is superior to the competitive methods on synthetic scenes and real HSI datasets in terms of kinds of evaluation criteria, visual effect, and spectral fidelity.

References

[1] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image super resolution by transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, May 2017, doi: 10.1109/JSTARS.2017.2655112.

[2] Q. Li, Q. Wang, and X. Li, "An efficient clustering method for hyperspectral optimal band selection via shared nearest neighbor," *Remote Sens.*, vol. 11, no. 3, Feb. 2019, Art. no. rs11030350, doi: 10.3390/rs11030350.

[3] J. Lin et al., "Dual-modality endoscopic probe for tissue surface shape reconstruction and hyperspectral imaging enabled by deep neural networks," *Med. Image Anal.*, vol. 48, pp. 162–176, Aug. 2018, doi: 10.1016/j.media.2018.06.004.

[4] N. Zhang, G. Yang, Y. Pan, X. Yang, L. Chen, and C. Zhao, "A review of advanced technologies and development for hyperspectral-based plant disease detection in the past three decades," *Remote Sens.*, vol. 12, no. 19, Sep. 2020, Art. no. rs12193188, doi: 10.3390/rs12193188.

[5] I. C. Contreras Acosta, M. Khodadadzadeh, and R. Gloaguen, "Resolution enhancement for drill-core hyperspectral mineral mapping," *Remote Sens.*, vol. 13, no. 12, Jun. 2021, Art. no. rs13122296, doi: 10.3390/rs13122296.

[6] H. Gong et al., "Multiscale information fusion for hyperspectral image classification based on hybrid 2D-3D CNN," *Remote Sens.*, vol. 13, no. 12, Jun. 2021, Art. no. 2268, doi: 10.3390/rs13122268.

[7] Q. Li et al., "Unsupervised hyperspectral image change detection via deep learning self-generated credible labels," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9012–9024, Sep. 2021, doi: 10.1109/JSTARS.2021.3108777.

[8] Q. Li et al., "A superpixel-by-superpixel clustering framework for hyperspectral change detection," *Remote Sens.*, vol. 14, no. 12, Jun. 2022, Art. no. rs14122838, doi: 10.3390/rs14122838.

[9] Q. Li, T. Mu, A. Tuniyazi, Q. Yang, and H. Dai, "Progressive pseudo-label framework for unsupervised hyperspectral change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 127, Mar. 2024, Art. no. 103663, doi: 10.1016/j.jag.2024.103663.

[10] L. Yan, M. Yamaguchi, N. Noro, Y. Takara, and F. Ando, "A novel two-stage deep learning-based small-object detection using hyperspectral images," *Opt. Rev.*, vol. 26, no. 6, pp. 597–606, Dec. 2019, doi: 10.1007/s10043-019-00528-0.

[11] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926, doi: 10.1016/j.jag.2022.102926.

[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, Feb. 2022, Art. no. 6007305, doi: 10.1109/LGRS.2022.3151779.

[13] Q. Xu, W. Qiu, B. Li, and F. Gao, "Hyperspectral and panchromatic image fusion through an improved ratio enhancement," *J. Appl. Remote Sens.*, vol. 11, no. 1, Mar. 2017, Art. no. 015017, doi: 10.1117/1.JRS.11.015017.

[14] K. Li, D. Dai, and L. V. Gool, "Hyperspectral image super-resolution with RGB image super-resolution as an auxiliary task," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 4039–4048, doi: 10.1109/WACV51458.2022.00409.

[15] J. Xiao, J. Li, Q. Yuan, M. Jiang, and L. Zhang, "Physics-based GAN with iterative refinement unit for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6827–6841, May 2021, doi: 10.1109/JSTARS.2021.3075727.

[16] M. Simões, 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, doi: 10.1109/TGRS.2014.2375320.

[17] 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, doi: 10.1109/TGRS.2011.2161320.

[18] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019, doi: 10.1109/TNNLS.2018.2885616.

[19] J. Chen, J. Nunez-Yanez, and A. Achim, "Video super-resolution using generalized gaussian Markov random field," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 63–66, Feb. 2012, doi: 10.1109/LSP.2011.217859.

[20] A. Marquina and S. J. Osher, "Image super-resolution by TV-regularization and Bregman iteration," *J. Sci. Comput.*, vol. 37, no. 3, pp. 367–382, Dec. 2008, doi: 10.1007/s10915-008-9214-8.

[21] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010, doi: 10.1109/TIP.2010.2050625.

[22] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011, doi: 10.1109/TIP.2011.2108306.

[23] S. Anwar, S. Khan, and N. J. A. C. S. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, Apr. 2020, doi: 10.1145/3390462.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/tpami.2015.2439281.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1132–1140, doi: 10.1109/CVPRW.2017.151.

[26] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654, doi: 10.1109/CVPR.2016.182.

[27] J. Yang, Y.-Q. Zhao, J. C.-W. Chan, and L. Xiao, "A multi-scale wavelet 3D-CNN for hyperspectral image super-resolution," *Remote Sens.*, vol. 11, no. 13, Jun. 2019, Art. no. 1557, doi: 10.3390/rs11131557.

[28] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1660, doi: 10.3390/rs12101660.

[29] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8693–8703, Oct. 2021, doi: 10.1109/TGRS.2020.3047363.
[30] L. Liebel and M. Körner, "Single-image super resolution for multispectral remote sensing data using convolutional neural networks," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLI-B3, pp. 883–890, Jun. 2016, doi: 10.5194/isprs-archives-XLI-B3-883-2016.
[31] J. Hu, Y. Tang, Y. Liu, and S. Fan, "Hyperspectral image super-resolution based on multiscale mixed attention network fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 5511405, doi: 10.1109/LGRS.2021.3124974.
[32] D. Liu, J. Li, and Q. Yuan, "A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7711–7725, Sep. 2021, doi: 10.1109/TGRS.2021.3049875.
[33] X. Dou, C. Li, Q. Shi, and M. Liu, "Super-resolution for hyperspectral remote sensing images based on the 3D attention-SRGAN network," *Remote Sens.*, vol. 12, no. 7, Apr. 2020, Art. no. 1204, doi: 10.3390/rs12071204.
[34] J. Li et al., "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4304–4318, Jun. 2020, doi: 10.1109/TGRS.2019.2962713.
[35] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Hybrid local and nonlocal 3-D attentive CNN for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1274–1278, Jul. 2021, doi: 10.1109/LGRS.2020.2997092.
[36] J. Li et al., "Deep hybrid 2-D-3-D CNN based on dual second-order attention with camera spectral sensitivity prior for spectral super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 623–634, Feb. 2023, doi: 10.1109/TNNLS.2021.3098767.
[37] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens.*, vol. 9, no. 11, Nov. 2017, Art. no. 1139, doi: 10.3390/rs9111139.
[38] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1082–1096, 2020, doi: 10.1109/TCI.2020.2996075.
[39] Y. Li, S. Chen, W. Luo, L. Zhou, and W. Xie, "Hyperspectral image super-resolution based on spatial-spectral feature extraction network," *Chin. J. Electron.*, vol. 32, no. 3, pp. 415–428, May 2023, doi: 10.1049/cje.2021.00.081.
[40] M. Zhao, J. Ning, J. Hu, and T. Li, "Hyperspectral image super-resolution under the guidance of deep gradient information," *Remote Sens.*, vol. 13, no. 12, Jun. 2021, Art. no. 2382, doi: 10.3390/rs13122382.
[41] J. Hu, T. Li, M. Zhao, and J. Ning, "Hyperspectral image superresolution via deep structure and texture interfusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8665–8678, Aug. 2021, doi: 10.1109/JSTARS.2021.3107311.
[42] J. Zhang, M. Shao, Z. Wan, and Y. Li, "Multi-scale feature mapping network for hyperspectral image super-resolution," *Remote Sens.*, vol. 13, no. 20, Oct. 2021, Art. no. 4180, doi: 10.3390/rs13204180.
[43] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
[44] Q. Wang, Q. Li, and X. Li, "Spatial-spectral residual network for hyperspectral image super-resolution," Jan. 2020, *arXiv:2001.04609*.
[45] X. Wang, Q. Hu, J. Jiang, and J. Ma, "A group-based embedding learning and integration network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541416, doi: 10.1109/TGRS.2022.3217406.
[46] M. A. Haq, S. B. Hadj Hassine, S. J. Malebary, H. A. Othman, and E. M. Tag-Eldin, "3D-CNNHSR: A 3-dimensional convolutional neural network for hyperspectral super-resolution," *Comput. Syst. Sci. Eng.*, vol. 47, no.no. 2, pp. 2689–2705, 2023, doi: 10.32604/csse.2023.039904.
[47] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715, doi: 10.1109/TGRS.2022.3183468.
[48] Y. Long, X. Wang, M. Xu, S. Zhang, S. Jiang, and S. Jia, "Dual self-attention swin transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5512012, doi: 10.1109/TGRS.2023.3275146.
[49] M. Zhang, C. Zhang, Q. Zhang, J. Guo, X. Gao, and J. Zhang, "ESSAformer: Efficient transformer for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23016–23027.
[50] 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, doi: 10.1109/CVPR.2016.90.
[51] M. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.
[52] W. Dong, J. Qu, T. Zhang, Y. Li, and Q. Du, "Context-aware guided attention based cross-feedback dense network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530814, doi: 10.1109/TGRS.2022.3180484.
[53] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3759–3768, doi: 10.1109/ICCV.2019.00386.
[54] T. Dai, J. Cai, Y. Zhang, S. T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11057–11066, doi: 10.1109/CVPR.2019.01132.
[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
[56] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010, doi: 10.1109/TIP.2010.2046811.
[57] N. Yokoya and A. Iwasaki, Airborne hyperspectral data over Chikusei, Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, May 2016.
[58] J. Jia, L. Ji, Y. Zhao, and X. Geng, "Hyperspectral image super-resolution with spectral–spatial network," *Int. J. Remote Sens.*, vol. 39, pp. 1–24, Jun. 2018, doi: 10.1080/01431161.2018.1471546.

Wenjing Wang received the B.E. degree in physics from Guizhou University, Guizhou, China, in 2020. She is currently working toward the M.S. degree in physics with the School of Physics, Xi'an Jiaotong University, Xi'an, China.

Her research interests include hyperspectral remote sensing image super resolution based on deep learning methods.

Tingkui Mu received the Ph.D. degree in physics from Xi'an Jiaotong University, Xi'an, China, in 2012.

He was a Postdoctoral Fellow with the College of Optical Sciences, University of Arizona, Tucson, AZ, USA, between 2014 and 2016. He is currently a Professor with the School of Physics, Xi'an Jiaotong University. He has authored and coauthored more than 90 research papers and 16 authorized invention patents. His research interests include integrated acquisition and processing of multidimensional optical information, computer vision, spectral and polarimetric imaging technique, and sensors.

Dr. Mu is the Director of Shaanxi Provincial Optical Society, the Committee Member of the Optoelectronics Technology Committee of the Chinese Aerospace Society, the Member of the Chinese Optical Society, Chinese Optical Engineering Society, the Shaanxi Provincial Physical Society, the Optical Society of America (OSA), and the Society of Photo-Optical Instrumentation Engineers (SPIE).

Qiuxia Li received the B.E. degree in physics from Yunnan University, Yunnan, China, in 2019. She is currently working toward the Ph.D. degree in physics with the School of Physics, Xi'an Jiaotong University, Xi'an, China.

Her research interests include hyperspectral remote sensing image change detection based on deep learning methods.

Haoyang Li received the B.E. degree in applied physics from the Qingdao University of Science and Technology, Qingdao, China, in 2016. He was currently working toward the postgraduate degree in physics with Xi'an Jiaotong University, Xi'an, China.

His research interests include polarimetric imaging, hyperspectral imaging, and spectropolarimetric imaging.

Qiujie Yang received the B.S. degree in optical information science and technology from the Changchun University of Science and Technology, Changchun, China, in 2012, and the Ph.D. degree in physical electronics from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Associate Researcher with the Shanghai Institute of Technical Physics. His research interests include image processing and Fourier spectroscopy.