A Diffusion Model-Assisted Multiscale Spectral Attention Network for Hyperspectral Image Super-Resolution

Kaiqi He, Yiheng Cai, *Member, IEEE*, Shengjun Peng, and Meiling Tan

***Abstract*—Hyperspectral images (HSIs) can reflect the spectral characteristics of objects in multiple bands, which can be used in various tasks, including classification, material detection and identification, and geological exploration. However, due to hardware limitations, spatial data have commonly been partially discarded to obtain more spectral information. Therefore, the enhancement of spatial resolution is often contemplated through the application of super-resolution algorithms. In view of this, this study proposes a diffusion model-assisted multiscale spectral attention network (DMSANet) to increase the HSI resolution in the spatial dimension while preserving spectral information as much as possible. For the first time, a diffusion model is combined with deep networks to solve the HSI super-resolution problem, which enhances the spatial texture details of the output image using a layer-by-layer super-resolution mechanism of Markov chains. In addition, a multiscale attention block that can integrate multiple receptive fields to extract spectral features of HSIs is designed, which enhances spectral information details. Extensive evaluations and comparisons on three benchmark datasets demonstrate that the proposed DMSANet can achieve superior performance compared with the existing methods.**

***Index Terms*—Convolutional super-resolution network, diffusion model, image fusion, super-resolution hyperspectral image (HSI).**

I. Introduction

UNLIKE red-green-blue (RGB) images, hyperspectral images (HSIs) possess dozens to even hundreds of bands, offering imperceptible features to the human eye. In recent years, hyperspectral imaging technology has been widely used in various fields, such as urban planning [1], land cover classification [2], and image classification [3]. However, in the HSI collection process, it is challenging to achieve a balance between spectral and spatial information due to hardware limitations. In contrast to data in the spatial dimension, spectral dimension data are characterized by greater richness. During the acquisition process of HSIs, there is typically a heightened emphasis on capturing data in the spectral dimension, resulting in a resolution reduction in the spatial dimension. To compensate for this shortcoming, various hyperspectral super-resolution techniques have been employed to augment pixel quantity, providing additional data for subsequent hyperspectral tasks [4].

The HSI super-resolution reconstruction techniques based on deep learning can enhance the spatial resolution of low-resolution hyperspectral images (LrHSIs) by interpolating them with prior knowledge, thus ensuring that the resultant HSIs after super-resolution possess higher spectral curve accuracy and more spatial texture details. The HSI super-resolution methods can be roughly categorized into two types: fusion-based super-resolution methods and single-image-based super-resolution methods [5]. Fusion-based super-resolution methods use high-resolution multispectral images (HrMSIs) as auxiliary data [6], [7], [8] and combine spectral features of HSIs with spatial features of MSIs as a basis for high-quality interpolation operations on LrHSIs to generate high-resolution hyperspectral images (HrHSIs) [9], [10], [11], [12]. In general, the HSI fusion super-resolution methods have shown promising results, but they typically require using auxiliary images. However, in practical scenarios, it is nearly impossible to capture the HSI and MSI of the same object from the same perspectives simultaneously. Therefore, the application of these methods faces certain challenges in real-world situations.

In single-image-based super-resolution methods, only an LrHSI is used as input data. Since there are many similarities between natural images and HSIs, many methods developed for natural image processing can be applied to single-HSI super-resolution, yielding excellent results [13], [14], [15]. However, these methods cannot effectively meet the requirement of the HSI super-resolution task for increasing the number of spatial pixels while ensuring quality in the spectral dimension. Meanwhile, since most of the existing methods [16], [17] use deep convolution network models to extract features, the shallow features might not be well preserved, making it challenging to recover complex texture structures using these features. Aiming to address the aforementioned problems, this article proposes a diffusion model-assisted multiscale spectral attention network (DMSANet) that uses the spatial texture generation capability of the diffusion model to enhance spatial details and adopts a multiscale spectral attention block (MSAB) to adapt to the sparse and low-rank spectral characteristics of

Manuscript received 15 January 2024; revised 1 March 2024 and 31 March 2024; accepted 3 April 2024. Date of publication 9 April 2024; date of current version 25 April 2024. This work was supported by the National Key Research and Development Program of China under Grant 2017YFC1703302. *(Corresponding author: Yiheng Cai.)*

The authors are with the Department of Information Science, Beijing University of Technology, Beijing 100124, China (e-mail: hekaiqi@emails.bjut.edu.cn; caiyiheng@bjut.edu.cn; pengshengjun@emails.bjut.edu.cn; tanmeiling_1@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3386702

HSIs, thus enhancing the credibility of reconstructed spectral information.

In summary, the main contributions of this work are as follows.

1) To the best of the authors' knowledge, the diffusion model has been used for the first time to assist deep neural networks in single-image-based HSI super-resolution. By employing a diffusion model that can gradually denoise white noise, the diffusion-based approach is employed for layer-by-layer super-resolution, generating spatial information with richer and more realistic textures.
2) A network structure consisting of three branches, namely, the bicubic interpolation upsampling branch, diffusion model reconstruction branch (DRB), and spectral feature extraction branch (SFEB), is designed to obtain upsampling basic images and adjust spatial and spectral residual details.
3) The MSAB, which can use 3-D convolution kernels of different sizes to consider spectral information of different receptive fields, is proposed. In this way, more representative spectral features can be extracted, and redundant information on the feature matrix can be removed by combining the attention mechanism.

The rest of this article is organized as follows. Section II reviews the related work on single-image-based super-resolution methods and a diffusion model. Section III describes in detail the proposed DMSANet and MSAB. Section IV presents the results of related experiments and ablation experiments. Finally, Section V concludes this article.

II. Related Work

A. Single-Image-Based Super-Resolution Methods

Single-image-based super-resolution methods usually use the generative adversarial network (GAN)-based approaches, spectral–spatial feature separation, attention mechanisms, and nonnegative matrix feature decomposition to extract abstract features from HSIs, which are then used as a basis for realizing the HrHSI generation. Shi et al. [15] used the GAN as a basis of the autoencoder structure to generate coupled components to regularize the generated samples. Li et al. [18] proposed a spectral band attention mechanism based on GAN, which guides the training of the generative network by imposing a series of spatial and spectral constraints. Liu et al. [19] proposed the EUNet, which uses a depth separable convolution, to separate the spatial and spectral features, and this model focuses on interspectral correlations using a lightweight spectral attention mechanism. Zheng et al. [20] proposed a feature extraction module that extracts features from each band of an image independently and then fuses them while using the residual images to establish multipath connections to acquire features at different levels. Hu et al. [21] proposed an aggregation network based on multiscale feature fusion, where the aggregation of multiscale features expanded the network's receptive field. Li et al. [22] proposed the grouped deep recursive residual network (GDRRN), embedding a grouped recursive module in the global residual structure and designing a new loss function by combining the MSE and SAM. Mei et al. [23] proposed the 3DFCNN model, where the 3-D convolution block exploits both the spatial context of neighboring pixels and the spectral correlation of neighboring bands. Yao et al. [24] proposed a channel MLP-based method, using local and global spectral integrating blocks for grouping and shifting operations to capture local and global spectral correlations. Hu et al. [25] combined a collaborative nonnegative matrix decomposition strategy with a deep feature extraction network's output to reduce spectral distortion and texture blurring. Xu et al. [26] employed an alternating upsampling and downsampling U-Net to allocate the learning tasks of complex mapping relationships to each stage of the U-Net and proposed a spatial–spectral interaction transformer structure to learn complementary information between spatial and spectral dimensions. Zhao et al. [27] designed the network model that focuses on dual feature-guided spatial feature extraction, which includes feature aggregation guidance and gradient texture attention guidance. Zhang et al. [28] proposed spectral correlation coefficients for spectra, replacing the original attention matrix and incorporating inductive biases into the model to facilitate training, and applied them to the self-attention kernel of the network. Zhang et al. [29] proposed an innovative network based on explicit neural representations, enabling continuous functions to map spatial coordinates to pixel values in a content-aware manner, and in addition, periodic spatial encoding was used. Zhang et al. [30] constructed a smoothed function based on smoothed particle hydrodynamics and designed a smooth convolution block in the network to use spectral correlations for a more comprehensive acquisition of spectral information. However, most existing methods that employ attention mechanisms commonly use single-scale 3-D convolutional kernels to generate feature matrices. This results in a joint weighting of spatial and spectral data, causing spatial features to influence spectral features during the convolution process. Moreover, this type of feature extraction approach lacks a decoder that can independently recover spatial details, resulting in the loss of spatial details in the output image.

B. Diffusion Model Development and Application

The idea of the diffusion model was first proposed in 2015 by Sohl-Dickstein et al. [31], which progressively removes the noise added to an image using the Markov chain, thus restoring images with high-quality spatial details. Nichol and Dhariwal [32] found that the diffusion model can be trained using fewer forward sampling processes assisted by negligible differences in training results, thus improving the training speed of the model. Salimans and Ho [33] proposed an advanced parameterization method for a diffusion model to address the drawback of a long training sampling time, thus reducing training time while enhancing stability. In recent work, the diffusion model has been trained with fewer steps, higher accuracy, and better scalability than in [34], and has been widely used in RGB image generation, natural language processing, and other related fields.

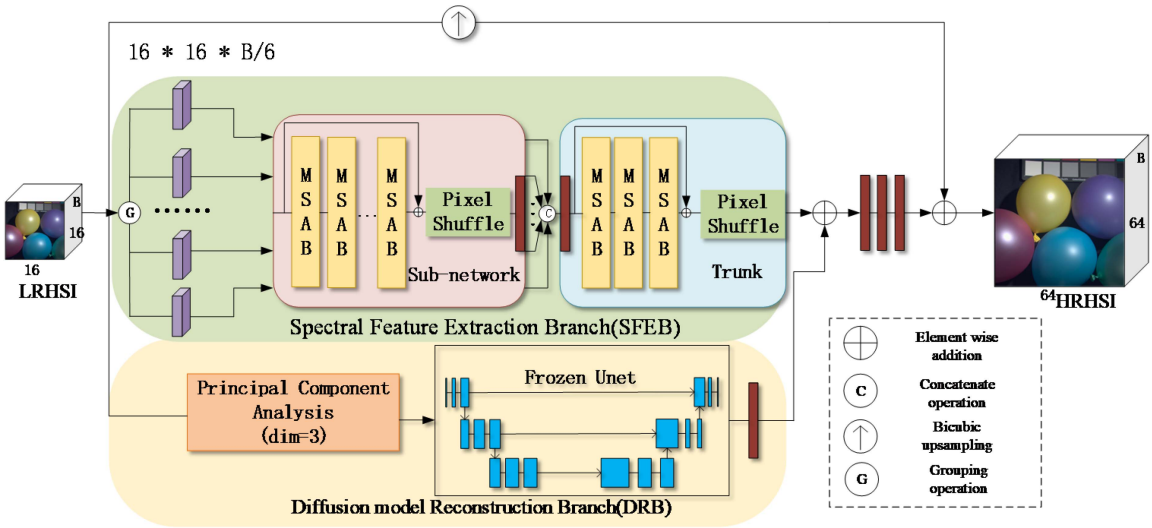

Fig. 1. General overview of the DMSANet.

In the field of image processing [35], [36], [37], [38], the diffusion model can generate images with better texture details than the GAN. Moreover, as the diffusion model does not engage in adversarial processes, it avoids various problems, such as mode collapse [39] and convergence failures during training. Due to these advantages, the diffusion model has been widely applied to the RGB image super-resolution tasks. Rombach et al. [40] introduced a cross-attention layer in the model architecture to construct a potential diffusion model, which improved the diffusion model's performance and allowed the high-resolution synthesis of RGB images. Saharia et al. [38] proposed the SR3 model that can enhance RGB images up to eight times using the upsampling results of low-resolution images as a guide, and this complements numerous spatial resolutions and incorporates many reliable details. Meanwhile, the diffusion model has also been applied to hyperspectral fusion. Shi et al. [41] developed a method that used HrMSIs to guide the diffusion model to generate high-quality spatial texture details for fusion-based HSI super-resolution. Consequently, the diffusion model is suitable for HSI super-resolution tasks, and the further transfer of the model to single-image-based HSI super-resolution could also be effective in generating spatial dimensional detail.

III. Proposed Method

A. Overall Network Architecture

Most existing methods use deep networks to integrate an LrHSI into a feature matrix, which is then used for the reconstruction of super-resolution images [42], [18], [21], [43]. However, deep networks with multiple layers (usually convolutional-based network architecture) might face certain problems, such as excessive compression of features, resulting in insufficient spatial details in reconstructed images. In addition, during the convolutional decoding process from overly compressed abstract features, grid artifacts might appear in regions with sharp changes, such as object edges. To reduce the impact of these issues, in the proposed DMSANet, the diffusion model uses a stepwise super-resolution mechanism to recover spatial information and reconstruct texture details. Moreover, inspired by research in [44], this study employs a strategy of separating spectral and spatial features to generate a higher quality spectral curve. The overview of the DMSANet is shown in Fig. 1.

The DMSANet includes three branches: the bicubic interpolation upsampling branch, the SFEB, and the DRB. The bicubic interpolation upsampling branch generates super-resolution base images, and the SFEB and the DRB are used to provide spectral and spatial details, respectively. By performing elementwise addition, which is commonly used in data fusion to integrate information on each pixel without expanding the data volume, between the spectral residual feature maps output by the SFEB and the spatial residual feature maps output by the DRB, the overall residual features are obtained to adjust the base images. The reason bicubic interpolation is selected in this study to generate base images is because it has a wide range of applications in image super-resolution tasks [45] and does not cause problems of severe texture blurring and jagged images. The SFEB adopts the MSAB to focus on multiscale feature extraction in the spectral dimension, and the DRB uses the stepwise super-resolution mechanism of a diffusion model to achieve spatial dimension detail generation. The details generated by the two branches are smoothed by a convolutional network and added to the base image to obtain the final super-resolution HSI. The parameters of network modules comprising the network are given in Table I. Two pixel-shuffle operations are used in the SFEB, meaning that the feature maps in the SFEB are upsampled twice in the spatial dimension rather than in a single step, which can provide better results. In the DRB, the "principal component analysis (PCA) & bicubic upsampling" block preprocesses the input HSI, and UNet represents the network used for predicting noise in the diffusion model; all diffusion model's backward processes use the same UNet. Assume that $X \in \mathbb{R}^{B\times hw}$ and $\hat{Y} \in \mathbb{R}^{B\times HW}$ represent the input LrHSI and the output reconstructed HrHSI, respectively, B is the number of spectral bands, hw and HW represent the number of spatial pixels in the

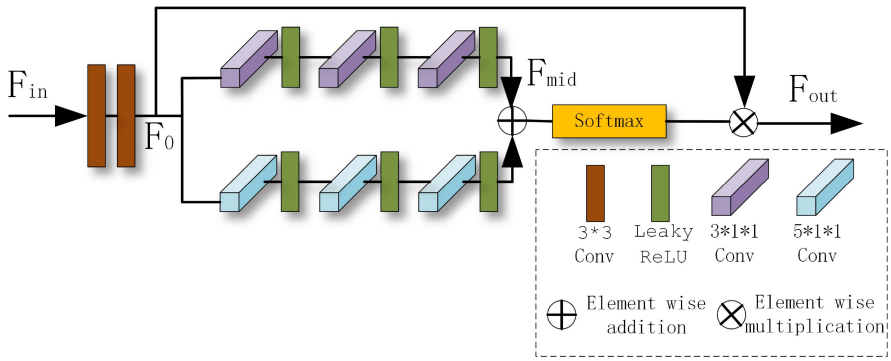

Fig. 2. Structure of MSAB.

TABLE I
STRUCTURE OF DMSANET

Branch	Module	Numbers
SFEB	Grouping	6
	MSAB(Subnetwork)	8
	MSAB(Trunk)	3
	Pixel Shuffle	2
DRB	PCA & Bicubic upsamlping	1
	Unet	1
	Reshape_layer	1

LrHSI and HrHSI, respectively, and $HW > hw$. The process of recovering the ideal $\hat{Y}$ using DMSANet can be described as follows:

$$\hat{Y} = f_{\text{UP}}(X) + f_{\text{INT}}(f_{\text{DRB}}(X) + f_{\text{SFEB}}(X)) \tag{1}$$

where $f_{\text{UP}}()$ is the bicubic interpolation upsampling operation, $f_{\text{DRB}}()$ denotes the DRB, $f_{\text{SFEB}}()$ denotes the SFEB, and $f_{\text{INT}}()$ stands for the integration of the information from the two branches.

B. Spectral Feature Extraction Branch

The SFEB is designed to extract the main spectral features of the input LrHSIs while considering their sparse and low-rank characteristics. The SFEB generates a residual image that corrects the spectral details of the base image. As shown in Fig. 1, the SFEB is the second branch of the DMSANet. Due to the higher correlation between adjacent spectral bands, the input HSI is grouped along the spectral dimension for better extraction of data features from the neighboring bands. Each group of data is input into a different subnetwork, where weight sharing is performed, which facilitates the integration of diverse spectral features learned by different subnetworks across various spectral bands and also increases the robustness of the SFEB while reducing the number of network parameters.

The subnetwork consists of the MSAB and a pixel shuffle module. The grouped images are subjected to a series of MSABs to obtain spectral attention features. Then, the pixel shuffle module is used to perform the upsampling operation on the features to obtain intermediate-resolution features for the subsequent concatenating operation. In addition, to preserve high-frequency information on HSIs, where spectral curves can undergo significant variations, residual structures are also employed to assist the subnetwork. The output features of each subnetwork are concatenated to obtain the intermediate features, which ensures that the features can be integrated in the same dimension but in no particular order. The concatenated features are then fed to the trunk network, which incorporates the MSAB for feature re-extraction and decoding to generate a residual image in the spectral dimension.

The core of the SFEB is the proposed MSAB, whose structure is shown in Fig. 2. To adapt to the above-mentioned characteristics better, the MSAB uses a multiscale 3-D convolution, which combines receptive fields of different sizes and makes the most of information on spectrally adjacent bands. Furthermore, the attention mechanism introduced in the MSAB conducts convolutional operations only in the spectral dimension, thus avoiding excessive attention to the spatial features, which is different compared with the existing spectral attention mechanisms that tend to extract both spatial- and spectral-dimension information simultaneously, providing less well-directed attention toward spectral information. The MSAB not only assists in obtaining a better understanding of the overall structure and content of the spectrum but also reduces the transmission of redundant information through the weighting operation of the weight matrix. The MSAB operation is defined as follows:

$$\begin{aligned} F_0 &= f_{\text{Conv2D}}(F_{\text{in}}) \\ F_{\text{mid}} &= f_{\text{Conv3D}(3\times1\times1)}(F_0) + f_{\text{Conv3D}(5\times1\times1)}(F_0) \\ F_{\text{out}} &= \text{Softmax}(F_{\text{mid}}) * F_0. \end{aligned} \tag{2}$$

First, F_{in} undergoes convolution processing to obtain F_0. Next, spectral feature extraction is applied to F_0 at different scales, specifically $5 \times 1 \times 1$ and $3 \times 1 \times 1$. The extracted feature maps of different scales are summed up, and then the Softmax operation is performed on them to obtain the weight matrix, which is multiplied by the initial processed feature F_0 to obtain the MSAB output F_{out}.

Overall, the application of attention mechanisms and residual connections in SFEB can mitigate the interference of redundant spectral features and alleviate the excessive compression of information in deep networks, thus limiting the loss of spectral information.

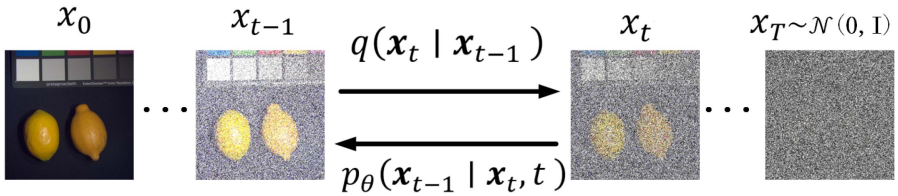

Fig. 3. Forward and backward processes in the diffusion model.

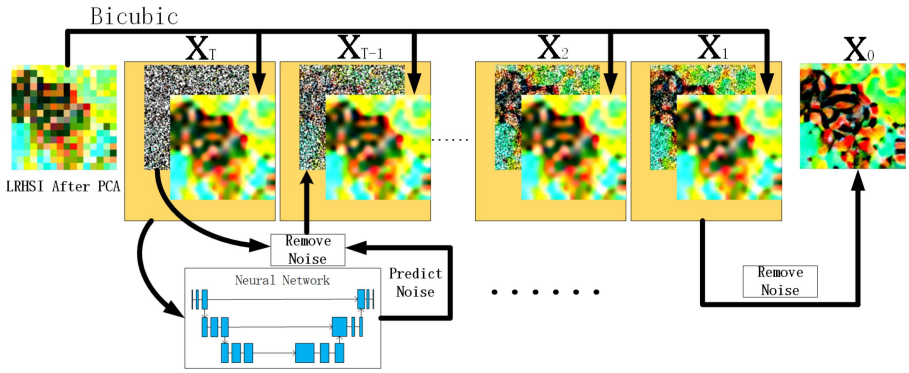

Fig. 4. Super-resolution generation process for diffusion modeling.

C. Diffusion Reconstruction Branch

The role of the DRB is to provide spatial texture details of the base image, which is a necessary supplement to the function of the SFEB. To generate HSIs with a high spatial resolution, the diffusion model with strong generative power is introduced, leading to the DRB design. By adopting a progressive denoising mechanism, the diffusion model can gradually incorporate details into an image, possessing powerful image generation capabilities, which have been applied to tasks involving RGB image generation. Due to its excellent ability to generate detailed textures and its successful application to RGB image generation, the diffusion model can be applied to the HSI super-resolution tasks, providing spatial details of the base image.

The diffusion model involves the gradual fitting and restoration of information, which includes two processes, namely, the forward process q, and the backward process p, as illustrated in Fig. 3. In the forward process, noise is gradually added to the original image. In the backward process, the same UNet is used to iteratively predict and remove noise from the input Gaussian white noise, thus recovering the original image. The diffusion model uses the nonlocal means algorithm for denoising, which can better preserve high-frequency data, such as object edges and textures, compared with the existing methods that tend to oversmooth the edges and textures of images, which can retain more spatial information. At the same time, the diffusion model adopts an iterative approach for denoising and detail recovery, allowing for the model to learn and understand the local and global information on an image better, thus enhancing the completeness of spatial information by updating and optimizing the image in each iteration. The diffusion model employs a nontraditional neural network decoding approach and reconstructs images from a global perspective, allowing for the incorporation of additional information, such as scene structure and layout, which is vital for recovering spatial details of HSIs. Based on the advantages of the diffusion model, it is introduced into the DRB, and the workflow of the DRB is shown in Fig. 4. To enhance the quality of spatial details, the input LrHSI is initially subjected to the PCA with the component number of three. Then, the image size is increased from $16 \times 16 \times 3$ to $64 \times 64 \times 3$ by upsampling. The processed image serves as a conditional guide for the UNet, preserving a significant portion of spectral information [46] and guiding spatial reconstruction. The spatial dimensions of the initial Gaussian white noise x_T are set to match the target super-resolution image's size. Subsequently, the reverse process of the diffusion model is performed using the UNet model to predict progressively the noise to be removed at each step. In each step, the spatial information is enhanced from x_t, thus obtaining x_{t-1}, $t \in (1,\ T)$. The final iteration of the diffusion model yields x_0. Finally, by expanding the channel number through convolutional modules, the spatial residual images are obtained.

The reconstruction process at any stage of the DRB is defined by (3)

$$x_{t-1} = \mathrm{Remove}(f_{\mathrm{UNet}}(\mathrm{Concat}[x_t, f_{\mathrm{up}}(f_{\mathrm{PCA}}(\mathrm{LRHSI}))])) \mathrm{From}\, x_t. \tag{3}$$

Conditional guidance is achieved by concatenating x_t with the guiding image in the channel dimension. The concatenated image is then input into the UNet for noise prediction at each stage. In (3), $f_{\mathrm{UNet}}()$ is the UNet, concat() is the concatenating operation in the channel dimension, $f_{\mathrm{up}}()$ is the bicubic interpolation upsampling operation, and $f_{\mathrm{PCA}}()$ is the PCA; Remove operation indicates the removal of noise computed by the network from the image of the previous stage, which enhances the details of the super-resolution image. The UNet in the DRB is pretrained to

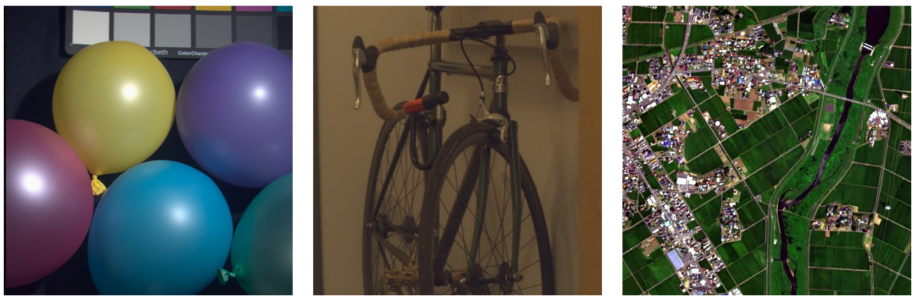

Fig. 5. Some RGB images corresponding to HSIs on three datasets. (a) CAVE dataset. (b) Harvard dataset. (c) Chikusei dataset.

acquire spatial super-resolution capabilities; namely, the input of the diffusion model is a low-resolution feature, and the output is a high-resolution feature. During the DMSANet's training process, other convolutional kernels in the DRB are trained to output spatial-domain residual images, while the weights of the UNet remain fixed.

IV. Experiments

To validate the proposed method's effectiveness, it was compared with the existing methods on three datasets: the CAVE dataset, the Harvard dataset, and the Chikusei dataset. In addition, ablation experiments were performed on the CAVE dataset to confirm the effectiveness of the core component of the proposed DMSANet. Furthermore, to demonstrate the adaptability of the proposed approach to different scale factors, spatial super-resolution experiments at 2x, 3x, and 4x were conducted on each dataset. Detailed information on experimental settings and data analysis is provided in the following sections.

A. Datasets

The three datasets were selected because they include varying spectral bands, object categories, and acquisition methods, providing a comprehensive basis for demonstrating the robustness of the DMSANet. The RGB images corresponding to hyperspectral data from the datasets are illustrated in Fig. 5.

1) CAVE Dataset: This dataset[1] has HSI in bands ranging from 400 to 700 nm. (31 bands in total) [47]. The spectral dimension of this dataset was collected in steps of 10 nm by a charge-coupled device camera, where the size of each HSI is $512 \times 512 \times 31$.

2) Harvard Dataset: This dataset[2] has HSI in bands from 400 to 700 nm (31 bands in total). It consists of 77 HSI of real indoor or outdoor scenes. The size of each of these HSI is $1040 \times 1392 \times 31$.

3) Chikusei Dataset: This dataset[3] was produced by Yokoya and Iwasaki [48], it has bands from 343 to 1018 nm (128 bands in total), has a spatial resolution of 2.5 m, and contains both urban and rural areas. The size is $2517 \times 2335 \times 128$.

B. Experimental Setup

Training and testing were conducted on multiple datasets. For the CAVE and Harvard datasets, five images were randomly selected as test samples, namely, five HSIs with the size of $512 \times 512 \times 31$, and the rest were used for training. In addition, in the Chikusei dataset, four $512 \times 512 \times 128$ regions were randomly selected as test samples, and the rest were used for training.

First, the size of each training and testing image was limited to $512 \times 512 \times B$, where B denoted the number of bands. Then, the training and testing images were cut in a top-to-bottom and left-to-right manner, and the size of the cut was $W \times H \times B$, where H and W were selected according to different super-resolution multiples. The cut images were downsampled by bicubic downsampling to obtain LrHSIs, which were used as the input for network training and testing.

The training process of the diffusion model requires using many samples, but due to the limited volume of available HSI data, the diffusion model was pretrained for the super-resolution task on the Flickr Faces HQ dataset. Through transfer learning, the network could accumulate knowledge on the representation of object features. Furthermore, by fixing the weight parameters of the pretrained diffusion model, the powerful spatial dimensionality super-resolution capability was retained. Then, it is combined with the DRB to provide spatial texture details for the super-resolution base image. The initial learning rate of all networks was set to 1e-4, the batch size was three, and the L1 loss function was used; each method selected the best results from 100 epochs. The training process was performed on a PC equipped with the NVIDIA GeForce RTX 2080ti GPUs using PyTorch as a framework.

C. Evaluation Metrics

To test the effectiveness of the proposed DMSANet quantitatively, this study selected five objective metrics for testing, namely, SAM, PSNR, ERGAS, SSIM, and RMSE, which were

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

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

[3][Online]. Available: https://naotoyokoya.com/ Download.html

TABLE II
ABLATION STUDY OF DMSANET STRUCTURES

	DMSANet	DRB(*W*)	Spectral attention(*W*)	Multi-scale attention structure(*W*)	5×1×1(W)	3×1×1(W)
SAM	2.94	3.19	3.27	3.12	3.12	3.17
PSNR	38.2	37.46	35.99	37.59	37.46	37.48
ERGAS	2.950	3.221	3.900	3.171	3.231	3.225
SSIM	0.996	0.995	0.993	0.995	0.995	0.995
RMSE	0.0140	0.0154	0.0168	0.0159	0.0157	0.0158

defined as follows:

$$\mathrm{SAM} = \arccos\left(\frac{< \hat{Y}, Y >}{\| \hat{Y} \|_2 \| Y \|_2}\right) \tag{4}$$

$$\mathrm{PSNR} = \frac{1}{B}\sum_{l=1}^{B} 10\log_{10}\left(\frac{\mathrm{MAX}_l^2}{\mathrm{MSE}_l}\right) \tag{5}$$

$$\mathrm{ERGAS}\left(Y, \hat{Y}\right) = \frac{100}{k}\sqrt{\frac{1}{C}\sum_{i=1}^{B}\frac{\| \hat{Y}_i - Y_i \|_2^2}{\widetilde{\mu}_{Y_i}^2}} \tag{6}$$

$$\mathrm{SSIM}\left(Y_i, \hat{Y}_i\right) = \frac{\left(2\widetilde{\mu}_{Y_i}\widetilde{\mu}_{\hat{Y}_i} + c_1\right)\left(2\sigma_{Y_i,\hat{Y}_i} + c_2\right)}{\left(\widetilde{\mu}_{Y_i}^2 + \widetilde{\mu}_{\hat{Y}_i}^2 + c_1\right)\left(+\sigma_{\hat{Y}_i}^2 + c_2\right)} \tag{7}$$

$$\mathrm{RMSE}\left(Y, \hat{Y}\right) = \sqrt{\frac{\| Y - \hat{Y} \|_F^2}{B \times WH}} \tag{8}$$

where $\hat{Y}$ is the super-resolution image generated by the network and Y is the ground truth (GT) of the HrHSI. $<,>$ denotes the dot product operation, and $\|\|_2$ is the l_2 paradigm. B is the number of spectral bands, and MSE_l is the mean squared error of $\hat{Y}$ versus Y in the lth band, and MAX_l is the maximum value of the lth band. k is the magnification, $\widetilde{\mu}_{Y_i}$ and $\widetilde{\mu}_{\hat{Y}_i}^2$ denote the variance of Y and $\hat{Y}$in the ith band, σ_{Y_i} and $\sigma_{\hat{Y}_i}$ denote the variance of the ith band of Y and $\hat{Y}$, respectively, $\sigma_{Y_i,\hat{Y}_i}$denotes the standard deviation of the ith band of Y and $\hat{Y}$, and $\|\|_F$ denotes F paradigms. The SAM metric reflects the degree of spectral matching between Y and $\hat{Y}$, and the smaller its value, the better the two are spectrally matched, and the closer their spectra are to each other. The PSNR metric mainly measures the distortion degree of an image and how much information can be conveyed, and the larger its value, the more information is conveyed. The ERGAS metric considers both spatial and spectral resolution variations, and the smaller its value, the better performance is achieved. The SSIM metric is used to measure the degree of similarity between the two images in terms of luminance, contrast, and structure, and the higher its value, the smaller the image distortion. Finally, the RMSE metric measures the mean square error between Y and $\hat{Y}$, and the smaller the RMSE value, the higher the prediction accuracy of the model.

D. Ablation Experiments

In this study, ablation experiments were conducted to validate the effectiveness of the DRB and MSAB. The results of ablation experiments for each structure are presented in Table II, where "DRB (*W*)" represents the exclusion of the DRB, "spectral attention (*W*)" indicates the removal of the attention structure from the MSAB, and "multiscale attention structure (*W*)" signifies the application of a $3\times 3\times 3$ convolution to compose the spectral attention module instead of the multiscale structure in the MSAB; "$5\times 1\times 1$(*W*)" denotes the elimination of the multiscale $5\times 1\times 1$ branch from the MSAB, and "$3\times 1\times 1$ (*W*)" indicates the removal of the $3\times 1\times 1$ branch from the MSAB. As given in Table II, after moving the DRB, all objective metrics deteriorated. The ERGAS value deteriorated significantly, indicating that the diffusion model could provide a large number of spatial details in the HSI super-resolution tasks. However, upon the removal of the DRB, there was a certain degree of deterioration in the SAM value, which meant that the DRB not only enhanced the spatial details of the HSI but also could transmit partial spectral information. The results related to SFEB demonstrated that the MSAB could achieve better abstraction of features compared with the traditional attention modules, providing higher quality feature maps for the decoder network. Moreover, based on the results, removing any of the spectral dimension convolution branches would cause a significant deterioration in the SAM value. This demonstrated that the collaboration of the two scales of convolution could effectively extract the primary spectral information, thus confirming the effectiveness of the MSAB.

E. HSI Super-Resolution Experiments and Results

In these experiments, the proposed DMSANet was compared with the EUNet [19], GDRRN [22], SSMLP [24], SSPSR [49], EDSR [50], and bicubic interpolation methods on the CAVE, Harvard, and Chikusei datasets. The experiments used three scaling factors: 2x, 3x, and 4x. The bold and underlined values in all tables represent the best and suboptimal results, respectively.

1) CAVE: Five images were randomly selected from the CAVE dataset as a testing set. The experimental results are given in Table III, where it can be seen that the DMSANet achieved better results than the other models on the CAVE dataset regarding different metrics at different scaling factors. Compared with the SSPSR, which extracted spatial features first and then the spectral features, the DMSANet exhibited obvious improvements in the SAM and PSNR metrics. This indicated that the DMSANet for separating spectral and spatial features during the extraction

TABLE III
COMPARATIVE RESULTS ON THE CAVE DATASET

Scale	Metrics	Bicubic	DMSANet	SSPSR	EUNet	EDSR	GDRRN	SSMLP
2x	SAM	2.03	**1.45**	1.78	1.51	1.97	1.60	1.62
	PSNR	40.82	**49.57**	46.74	47.64	45.52	49.43	46.87
	ERGAS	2.620	**0.771**	1.208	0.987	1.420	0.817	1.107
	SSIM	0.998	**0.999**	0.999	0.999	0.999	0.999	0.999
	RMSE	0.0110	**0.0033**	0.0049	0.0042	0.0059	0.0035	0.0047
3x	SAM	2.74	**1.66**	1.96	1.76	2.08	1.72	1.85
	PSNR	37.27	**41.03**	39.83	40.61	40.13	38.92	39.04
	ERGAS	3.470	**1.360**	1.540	1.420	1.490	1.850	1.710
	SSIM	0.995	**0.998**	0.998	0.998	0.998	0.997	0.997
	RMSE	0.0140	**0.0091**	0.0100	0.0095	0.0100	0.0120	0.0110
4x	SAM	3.66	**2.94**	3.31	3.25	3.62	3.16	3.40
	PSNR	36.13	**38.20**	37.96	37.25	37.76	36.38	35.91
	ERGAS	3.860	**2.950**	3.022	3.250	3.070	3.740	3.810
	SSIM	0.994	**0.996**	0.995	0.995	0.995	0.994	0.993
	RMSE	0.0160	**0.0140**	0.0150	0.0150	0.0150	0.0160	0.0180

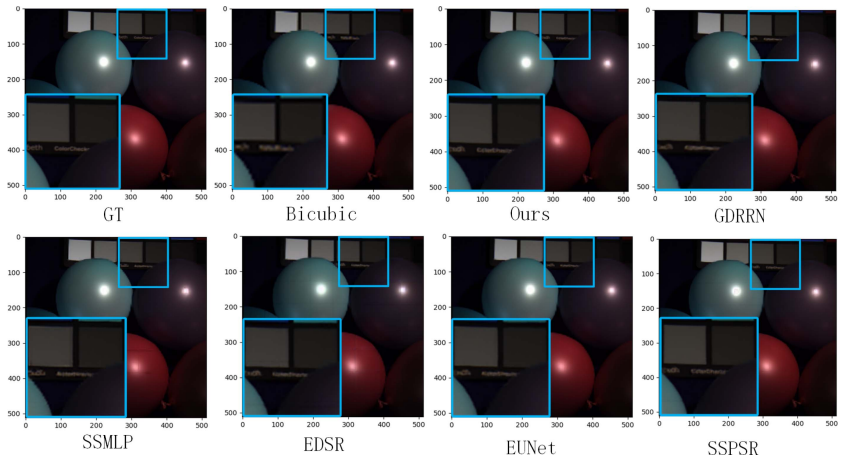

Fig. 6. Generated image of the CAVE dataset (R:10 G:20 B:30).

process was very effective, and the three-branch structure was better suited for specialized information processing. In addition, compared with the state-of-the-art method, the EUNet, which combined spectral attention mechanism and spectral correlation, the DMSANet achieved significant improvements in the SAM and ERGAS indicators, indicating that the DMSANet structure could better extract HSI features compared with the EUNet structure. Furthermore, the DMSANet outperformed the classical EDSR composed of 3-D convolutions in all objective indicators, indicating that the DMSANet could better solve the HSI super-resolution problems, such as reconstructing spatial and spectral details, than the EDSR. Furthermore, compared with the GDRRN that used recursive residual convolution modules to synthesize spectral features, the DMSANet achieved a significant improvement in the PSNR metric, indicating that compared with GDRRN's feature extraction method, the DMSANet could transmit much more information. Finally, compared with the state-of-the-art method SSMLP, which used fully connected layers to extract global spectral features, the DMSANet achieved significant improvements in the SAM and ERGAS metrics. This indicated that the DMSANet structure was more suitable for the HSI feature processing, including grouping operations of the SFEB and multiscale convolution of spectral dimensions for local spectral feature extraction than the SSMLP structure.

The reconstructed images are shown in Fig. 6, where it can be seen that compared with the other methods, the DMSANet generated sharper object contours, such as letter edges, demonstrating the ability of the diffusion model to reconstruct texture data. Particularly, due to the application of the diffusion model for gradual super-resolution of images, details were continuously added to the image, generating higher quality spatial details in the decoding form different from convolution. In addition, as shown in Fig. 7, the hot maps were used to visualize the errors of the reconstructed images in the spectral dimension for all the methods. These maps were generated by mapping the MSE errors of all spectra of a spatial pixel into the color bars. The deeper the red color in the plot, the higher the spectral error, and the closer the color in the plot to the black color, the lower the error. In Fig. 7, it can be observed that the spectral error of the DMSANet was very small in most of the regions, and even at the object edges with less favorable spectral reconstruction results, the DMSANet still performed better than the comparison methods. This was because the three-branch structure had specialized processing of spectral information, which improved the reliability of spectral dimension reconstruction.

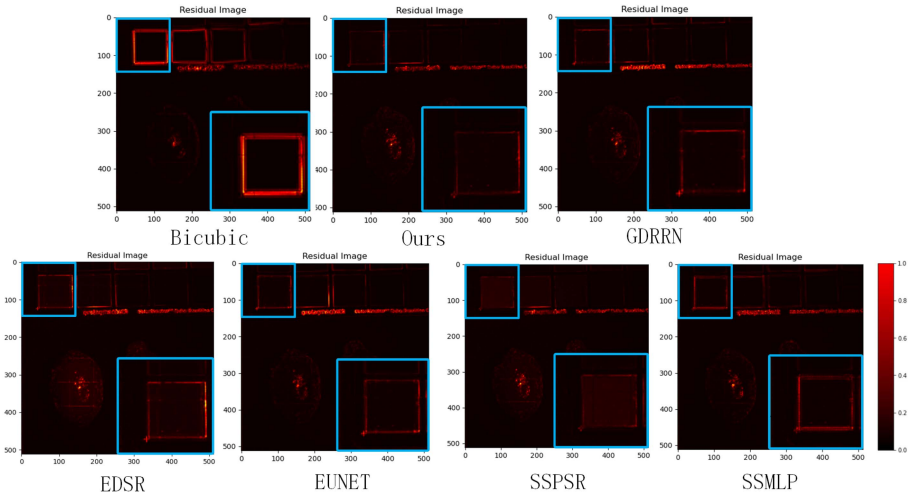

Fig. 7. Corresponding hot map for each method in the CAVE.

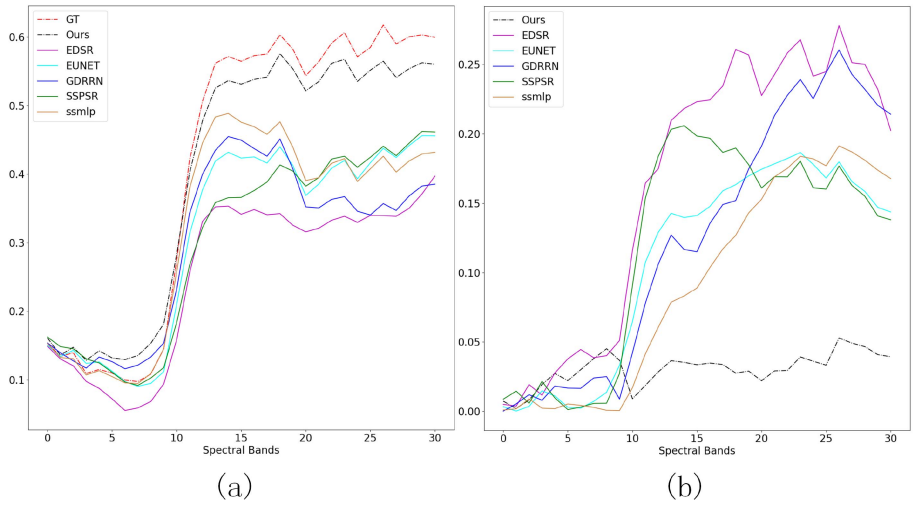

Fig. 8. Spectral curves generated by different methods of the CAVE dataset. (a) Spectral curve of random point. (b) Absolute MSE curve of spectral dimension compared with GT.

Furthermore, a point was randomly selected from the test dataset to validate the reconstruction effect of spectral curves, and the results of different methods are shown in Fig. 8. As presented in Fig. 8, compared with the comparison methods' results, the output result of the DMSANet was closer to the GT curve with the smallest error. This could be attributed to the introduction of the MSAB, where the extraction of multiscale spectral information could ensure the quality of constructed spectral curves.

2) Harvard: The quantitative experiments were also conducted on the Harvard dataset, and the results are given in Table IV. From the table, the DMSANet obtained the best results under all three scale factors among all the models, which demonstrated that the proposed DMSANet could provide much spatial information while maintaining spectral quality.

The DMSANet achieved considerable results in preserving texture details of super-resolution reconstruction. The visualization results of the reconstructed HSI are shown in Fig. 9, where it can be seen that the DMSANet's super-resolution results had very clear texture details, demonstrating good reflection effects on the support part of the chair and sharper contours on the cushion part. It should be noted that due to the split processing of the Harvard dataset, splicing gaps could often appear in the reconstructed images (the edges of each cut block can be seen from the zoomed image), but the DMSANet remained largely unaffected by this, which indicated that the diffusion model exhibited a favorable handling effect on the edges of the reconstructed image. The hot map in Fig. 10 shows that the spectra constructed by the comparison methods exhibited significant errors at the cutting boundaries of the images, while the super-resolution result generated by the DMSANet remained smooth at the boundaries. The spectral curves reconstructed by all the methods for the randomly selected point are shown in Fig. 11, where it can be seen that the DMSANet could generate the spectral curve that was the closest to the GT curve, indicating its stronger ability in spectral curve reconstruction.

TABLE IV
COMPARATIVE RESULTS ON THE HARVARD DATASET

Scale	Metrics	Bicubic	DMSANet	SSPSR	EUNet	EDSR	GDRRN	SSMLP
2x	SAM	2.69	**2.08**	2.20	2.14	2.35	2.21	2.16
	PSNR	39.96	**44.56**	44.02	44.31	43.45	43.87	43.76
	ERGAS	5.182	**3.270**	3.716	3.368	3.851	3.481	3.492
	SSIM	0.986	**0.995**	0.994	0.995	0.994	0.994	0.994
	RMSE	0.0162	**0.0084**	0.0087	0.0086	0.0090	0.0093	0.0090
3x	SAM	3.18	**2.17**	2.40	2.20	2.45	2.68	2.69
	PSNR	38.01	**44.03**	43.68	43.89	43.52	43.15	43.17
	ERGAS	6.090	**3.234**	3.386	3.466	3.409	3.620	3.618
	SSIM	0.970	**0.993**	0.993	0.993	0.992	0.991	0.991
	RMSE	0.0180	**0.0085**	0.0088	0.0086	0.0088	0.0092	0.0091
4x	SAM	3.58	**2.54**	3.01	2.79	3.72	2.77	2.61
	PSNR	34.55	**41.60**	40.62	41.22	39.70	41.05	41.35
	ERGAS	8.074	**3.615**	4.248	3.820	5.410	3.850	3.821
	SSIM	0.932	**0.988**	0.985	0.987	0.977	0.987	0.987
	RMSE	0.0336	**0.0130**	0.0145	0.0135	0.0175	0.0139	0.0137

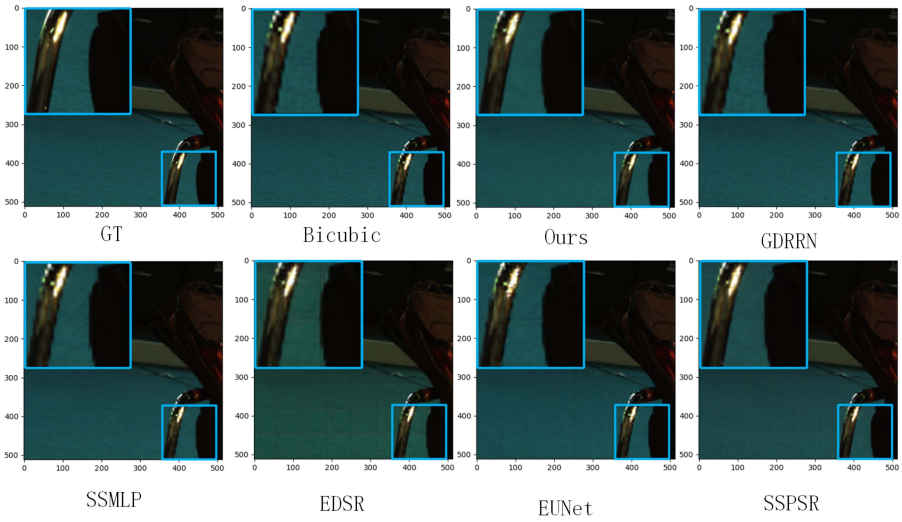

Fig. 9. Generated image of the Harvard dataset (R:10 G:20 B:30).

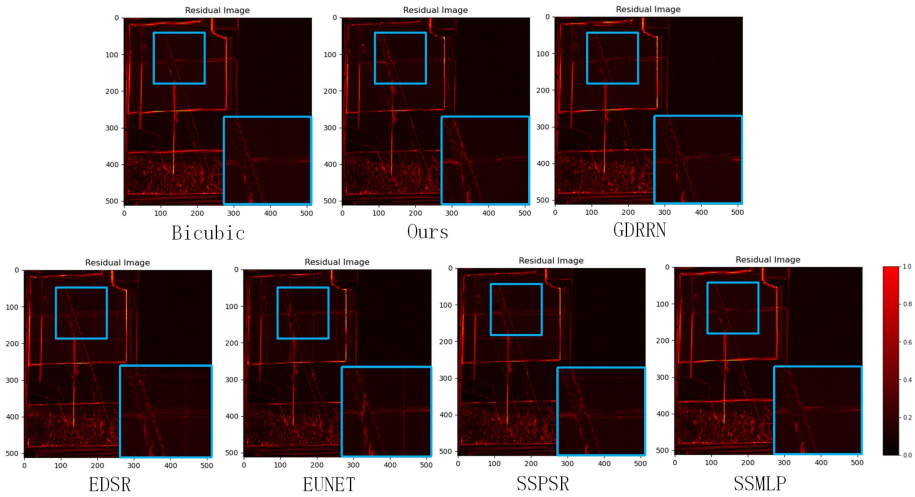

Fig. 10. Corresponding hot map for each method in the Harvard.

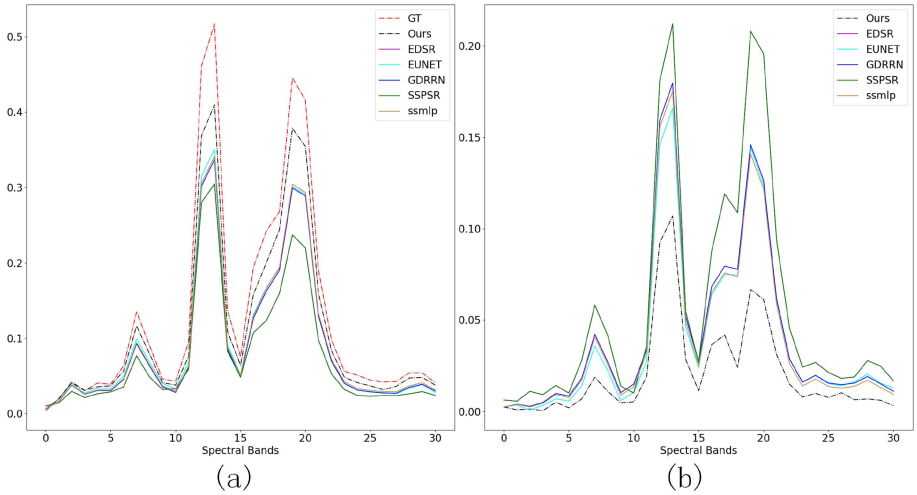

Fig. 11. Spectral curves generated by different methods of Harvard dataset. (a) Spectral curve of random point. (b) Absolute MSE curve of spectral dimension compared with GT.

TABLE V
COMPARATIVE RESULTS ON THE CHIKUSEI DATASET

Scale	Metrics	Bicubic	DMSANet	SSPSR	EUNet	EDSR	GDRRN	SSMLP
2x	SAM	1.67	**1.42**	1.53	1.52	1.61	1.62	1.49
	PSNR	45.09	**46.04**	45.7	45.71	45.81	45.1	45.61
	ERGAS	6.025	**5.460**	5.798	5.719	5.936	6.000	5.582
	SSIM	0.990	**0.993**	0.992	0.991	0.992	0.990	0.991
	RMSE	0.0070	**0.0063**	0.0063	0.0064	0.0062	0.0070	0.0067
3x	SAM	2.74	**2.14**	2.22	2.41	2.48	2.68	2.19
	PSNR	40.11	**44.96**	43.62	43.43	43.52	43.07	44.09
	ERGAS	6.580	**5.975**	6.290	6.546	6.432	6.656	6.229
	SSIM	0.967	**0.994**	0.990	0.992	0.992	0.991	0.993
	RMSE	0.0122	**0.0079**	0.0082	0.0088	0.0089	0.0092	0.0084
4x	SAM	3.68	**3.04**	3.13	3.17	3.2	3.27	3.28
	PSNR	37.02	**37.81**	37.53	37.75	37.66	37.52	37.32
	ERGAS	7.510	**6.868**	7.122	6.980	7.055	7.139	7.225
	SSIM	0.934	**0.943**	0.940	0.942	0.941	0.939	0.938
	RMSE	0.0170	**0.0155**	0.0160	0.0156	0.0157	0.0160	0.0165

3) Chikusei: To demonstrate the effectiveness of the proposed method on datasets with a larger number of bands, additional experiments were conducted on the Chikusei dataset, which had a total of 128 bands. From this dataset, four regions were selected as a testing set; the experimental results are given in Table V. Due to the presence of numerous spectral bands in the Chikusei dataset, the redundancy of HSI was high, making feature extraction of spectral information more difficult. However, the DMSANet still performed the best among all the models, which indicated its effectiveness.

The super-resolution results in Fig. 12 show that among all the models, the DMSANet achieved the best subjective results, including the best continuity of the playground runway and the least distortion of the lake texture (see enlarged area in Fig. 12); this indicated that the DMSANet had a strong texture detail recovery ability. The hot map results are shown in Fig. 13; by zooming into specific regions in Fig. 13, it can be observed that the spectral errors reconstructed by the DMSANet remained minimal among all the comparison methods, which demonstrated that the DMSANet could achieve excellent spectral reconstruction even in the presence of a large number of spectral bands. The spectral reconstruction results obtained using a randomly selected point from the Chikusei dataset are shown in Fig. 14 where it can be seen that the spectral curve of the DMSANet's output result was the closest to the GT curve among all the models, with the smallest absolute value error.

TABLE VI
MODEL COMPLEXITY AND CONVERGENCE

Methods	Number of total params(M)	Convergent epoch
DMSANet(ours)	6.03	19
EDSR	27.71	23
EUNet	0.28	20
GDRRN	0.03	19
SSPSR	26.08	25
SSMLP	8.13	24

4) Model Analysis: The DMSANet achieved the best HSI super-resolution results on datasets among all models, which demonstrated its effectiveness in learning common features and patterns in HSI data. To evaluate the DMSANet comprehensively, this study analyzed the DMSANet's complexity and training convergence speed, as given in Table VI. The results

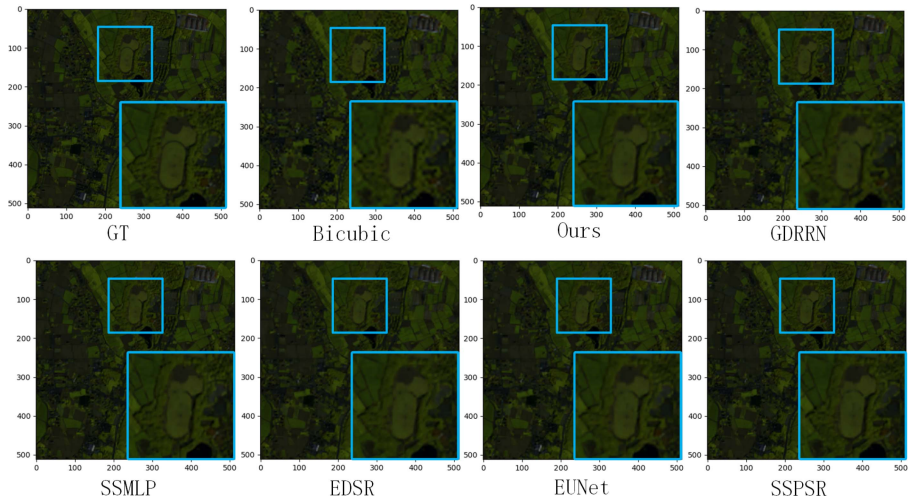

Fig. 12. Generated image of the Chikusei dataset (R:10 G:20 B:30).

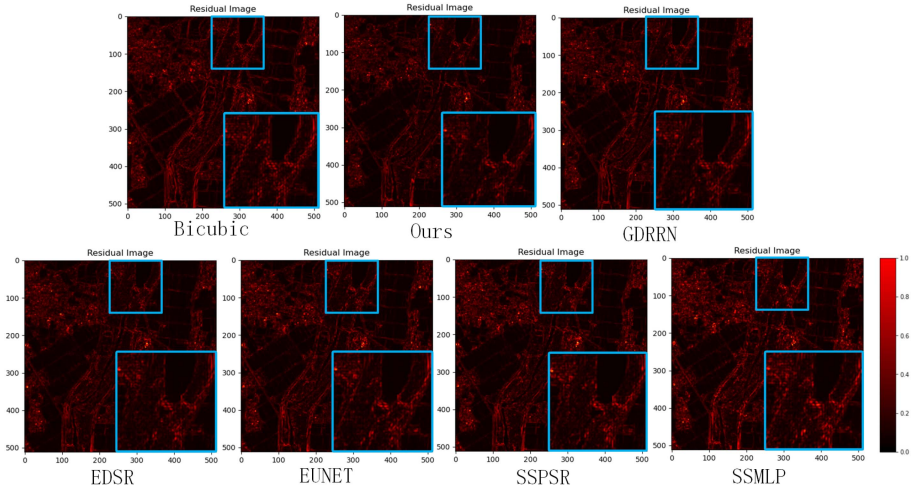

Fig. 13. Corresponding hot map for each method in the Chikusei.

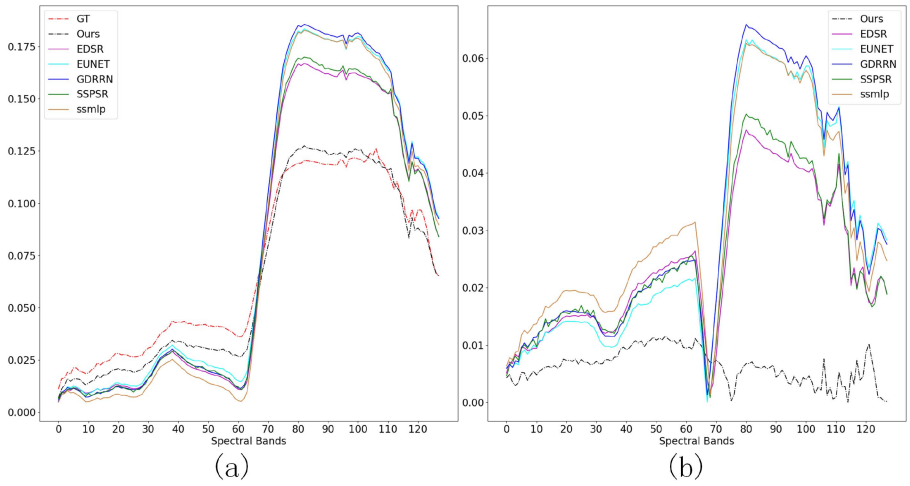

Fig. 14. Spectral curves generated by different methods of Chikusei dataset. (a) Spectral curve of random point. (b) Absolute MSE curve of spectral dimension compared with GT.

TABLE VII
NOISE ROBUSTNESS

	DMSANet	SSPSR	EUNet	EDSR	GDRRN	SSMLP
SAM	**6.84**	6.95	9.2	7.46	7.72	8.23
PSNR	**36.70**	36.68	34.26	31.44	35.77	34.9
ERGAS	3.439	**3.403**	4.803	6.26	3.987	4.216
SSIM	**0.995**	0.994	0.990	0.983	0.993	0.992
RMSE	0.0164	**0.0162**	0.0211	0.0285	0.0169	0.0203

in Table VI demonstrate that the DMSANet had significantly fewer parameters compared with the EDSR and SSPSR, and its number of parameters was only slightly higher than those of the EUNet and GDRRN. In addition, among all the comparative methods, the DMSANet exhibited the fastest convergence speed during training and showed no signs of overfitting, which indicated its excellent trainability. Overall, the DMSANet achieved the best performance on all datasets with the fewest training epochs among all models and maintained a reasonable model complexity.

Furthermore, Gaussian noise was added to the input HSI to validate the noise robustness of the DMSANet, and the objective metrics of the output are given in Table VII, where bold values indicate the best objective metric and underlined values denote the second-best objective metric. As given in Table VII, the DMSANet achieved satisfactory super-resolution results even when Gaussian noise was added to the input data, indicating its strong noise robustness and ability to adapt to different data acquisition environments.

V. CONCLUSION

This article introduces the DMSANet for spatial super-resolution reconstruction of HSIs. By introducing the diffusion model, the DMSANet enhances the reconstruction ability of texture information. The MSAB design can extract the main part of spectral information, ensuring the accuracy of spectral reconstruction. Through the three-branch network structure, spectral and spatial residual features are applied to the base image to generate a high-quality super-resolution HSI.

The experimental results on the CAVE, Harvard, and Chikusei datasets indicate that the proposed DMSANet can achieve better results than most recent HSI super-resolution methods. This proves that the diffusion model-assisted deep network has a great application value in HSI super-resolution reconstruction, particularly under suitable network structure design conditions. In the future, diffusion model-based super-resolution deep learning algorithms for HSI could have great prospects.

REFERENCES

[1] U. Heiden, K. Segl, S. Roessner, and H. Kaufmann, "Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data," *Remote Sens. Environ.*, vol. 111, no. 4, pp. 537–552, 2007.

[2] L.-J. Ferrato and K. W. Forsythe, "Comparing hyperspectral and multispectral imagery for land classification of the Lower Don River, Toronto," *J. Geogr. Geol.*, vol. 5, no. 1, pp. 92–107, Feb. 2013. [Online]. Available: http://dx.doi.org/10.5539/jgg.v5n1p92

[3] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.

[4] L. Loncan et al., "Hyperspectral Pansharpening: A Review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.

[5] 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.

[6] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522412.

[7] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1585–1594.

[8] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 6012305.

[9] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[10] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, "Deep recursive network for hyperspectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1233–1244, Aug. 2020.

[11] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.

[12] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.

[13] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, May 2017.

[14] Y. Li, J. Hu, X. Zhao, W. Xie, and J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[15] Y. Shi, L. Han, L. Han, S. Chang, T. Hu, and D. Dancey, "A latent encoder coupled generative adversarial network (LE-GAN) for efficient hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5534819.

[16] J. Jia, L. Ji, Y. Zhao, and X. Geng, "Hyperspectral image super-resolution with spectral–spatial network," *Int. J. Remote Sens.*, vol. 39, no. 22, pp. 7806–7829, 2018.

[17] J. Hu, T. Li, M. Zhao, F. Wang, and J. Ning, "A gated content-oriented residual dense network for hyperspectral image super-resolution," *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3378.

[18] 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.

[19] D. Liu et al., "An efficient unfolding network with disentangled spatial-spectral representation for hyperspectral image super-resolution," *Inf. Fusion*, vol. 94, pp. 92–111, 2023.

[20] K. Zheng et al., "Separable-spectral convolution and inception network for hyperspectral image super-resolution," *Int. J. Mach. Learn. Cybern.*, vol. 10, pp. 2593–2607, 2019.

[21] J. Hu, Y. Tang, and S. Fan, "Hyperspectral image super resolution based on multiscale feature fusion and aggregation network with 3-D convolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5180–5193, Sep. 2020.

[22] Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data*, 2018, pp. 1–4.

[23] 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, 2017, Art. no. 1139.

[24] Y. Yao, J. Hu, Y. Liu, and Y. Zhao, "Spectral-spatial MLP network for hyperspectral image super-resolution," *Remote Sens.*, vol. 15, no. 12, 2023, Art. no. 3066.

[25] J. Hu, X. Jia, Y. Li, G. He, and M. Zhao, "Hyperspectral image super-resolution via intrafusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7459–7471, Oct. 2020.

[26] Q. Xu, S. Liu, J. Wang, B. Jiang, and J. Tang, "AS ^{3} ITransUNet: Spatial–spectral interactive transformer U-net with alternating sampling for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5523913.

[27] M. Zhao, J. Ning, J. Hu, and T. Li, "Attention-driven dual feature guidance for hyperspectral super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5525116.

[28] 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.

[29] K. Zhang, D. Zhu, X. Min, and G. Zhai, "Implicit neural representation learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Dec. 2022, Art. no. 5500212.

[30] M. Zhang, J. Xu, J. Zhang, H. Zhao, W. Shang, and X. Gao, "SPH-Net: Hyperspectral image super-resolution via smoothed particle hydrodynamics modeling," *IEEE Trans. Cybern.*, early access, Oct. 31, 2023, doi: 10.1109/TCYB.2023.3323374.

[31] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[32] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[33] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," 2022, *arXiv:2202.00512*.

[34] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023.

[35] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18208–18218.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[37] D. Bau et al., "Seeing what a GAN cannot generate," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4502–4511.

[38] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[39] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 8780–8794, 2021.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[41] S. Shi, L. Zhang, and J. Chen, "Hyperspectral and multispectral image fusion using the conditional denoising diffusion probabilistic model," 2023, *arXiv:2307.03423*.

[42] 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.

[43] S. Mei, R. Jiang, X. Li, and Q. Du, "Spatial and spectral joint super-resolution using convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4590–4603, Jul. 2020.

[44] X. Zhang et al., "Spectral–spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2021, Art. no. 5512115.

[45] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2020.

[46] S. Meng, L.-T. Huang, and W.-Q. Wang, "Tensor decomposition and PCA jointed algorithm for hyperspectral image denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 897–901, Jul. 2016.

[47] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[48] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, vol. 5, 2016.

[49] 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, May 2020.

[50] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.