A Group-Based Embedding Learning and Integration Network for Hyperspectral Image Super-Resolution

Xinya Wang⬤, Qian Hu, Junjun Jiang⬤, *Senior Member, IEEE*, and Jiayi Ma⬤, *Senior Member, IEEE*

Abstract— Although natural image super-resolution methods have achieved impressive performance, single hyperspectral image super-resolution still remains a challenge due to the high dimensionality. In recent years, many single hyperspectral image super-resolution methods adopted the group-convolution strategy to design the network for reducing the computational burden. However, these methods still process all spectral bands at once during the deep feature extraction and reconstruction, which increases the difficulty of fully exploring the inherent data characteristic of hyperspectral images. Moreover, the advanced group-based methods make insufficient exploitation of complementary information contained in different bands, resulting in limited reconstruction performance. In this article, we propose a novel group-based single hyperspectral image super-resolution method termed group-based embedding learning and integration network (GELIN) to reconstruct high-resolution images in a group-by-group manner, which alleviates the difficulty of feature extraction and reconstruction for hyperspectral images. Specifically, a spatial–spectral embedding learning module is designed to extract rewarding spatial details and explore the correlations among spectra simultaneously. Considering the high similarity among different bands, a neighboring group integration module is proposed to fully exploit the complementary information contained in neighboring image groups to recover missing details in the target image group. Experimental results on both natural and remote sensing hyperspectral datasets demonstrate that the proposed method is superior to other state-of-the-art methods both visually and metrically.

Index Terms— Group convolution, hyperspectral image, neighboring groups, super-resolution.

I. INTRODUCTION

HYPERSPECTRAL images contain information of the same scene across different bands over a certain electromagnetic spectrum [1]. Each pixel in the hyperspectral image is dispersed to form tens or even hundreds of continuous spectral bands for rich and detailed spectral information, which

Xinya Wang, Qian Hu, and Jiayi Ma are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: wangxinya@whu.edu.cn; huq1an@whu.edu.cn; jyma2010@gmail.com).

Junjun Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangjunjun@hit.edu.cn).

reflects the subtle spectral properties of different materials. Attributed to the powerful diagnostic capability, hyperspectral images have been extensively used in many fields, such as mineral exploration [2], medical diagnosis [3], and plant detection [4]. However, in order to obtain hyperspectral images with high spectral resolution, the spatial resolution usually falls victim. The narrow and dense spectral bands in hyperspectral imaging systems limit the amount of incident energy reaching the sensor, resulting in relatively low spatial resolution of images. Since upgrading hardware technology is difficult and time-consuming, how to increase the spatial resolution of hyperspectral images still remains a challenge.

Super-resolution, as a postprocessing technique that infers the high-resolution (HR) image from its low-resolution (LR) version [5], could enhance the spatial resolution of hyperspectral images without hardware modification. Depending on whether auxiliary information is needed, hyperspectral image super-resolution is divided into two categories: fusion-based hyperspectral image super-resolution and single-image super-resolution. For the former, the LR hyperspectral image is fused with the HR auxiliary image to enhance the spatial resolution. Relying on Bayesian inference, matrix factorization, sparse representation, or leading deep learning techniques [6], [7], [8], [9], [10], these fusion-based methods have become mainstream. However, the HR auxiliary images, such as the panchromatic image and multispectral image, need to be captured at the same scene as LR hyperspectral images, resulting in the impracticality of fusion-based approaches in real scenarios. Although the recent work [11] realized the fusion between HR auxiliary images and LR hyperspectral images without registration, additional auxiliary information is still needed.

Compared with fusion-based methods, single-image super-resolution needs no auxiliary images or other priors, which is more applicable to real scenarios. Traditional methods usually resort to some spatial or spectral statistical distributions (e.g., total variation, sparse regularization, self-similarity, and low-rank approximation) to reconstruct HR hyperspectral images [12], [13], [14]. However, these handcrafted priors cannot fully characterize the intrinsic properties of hyperspectral images. Consequently, the reconstruction performance is often unsatisfactory. Recently, due to the success of the convolutional neural network (CNN) in natural image super-resolution, many CNN-based single hyperspectral image super-resolution
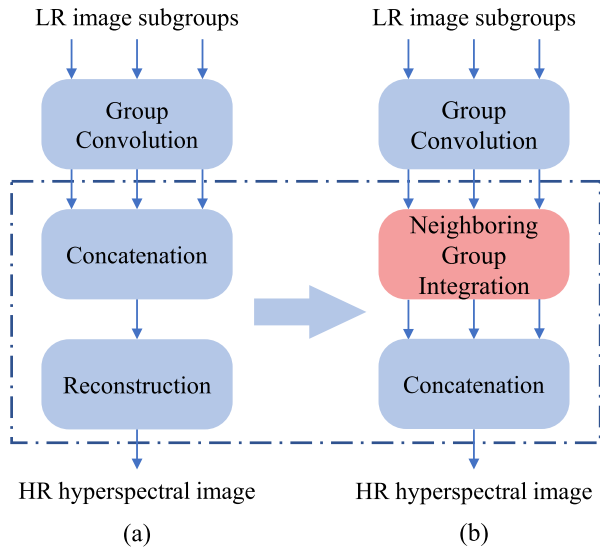
Fig. 1. Comparison of the frameworks for group-based hyperspectral image super-resolution. (a) Framework of the advanced group-based method. (b) Framework of the proposed GELIN, where each image group is super-resolved individually through the proposed NGIM.

methods have been proposed. For CNN-based hyperspectral super-resolution, the most intuitive way is to apply proven natural image super-resolution methods to reconstruct hyperspectral images in a band-by-band manner [15], [16]. However, such a reconstruction manner ignores the data characteristics of hyperspectral images, taking no spectral correlations into consideration. Inevitably, the reconstructed HR hyperspectral images suffer from severe spectral distortion. In order to obtain high spectral fidelity, several methods take the entire LR hyperspectral images as input and impose constraints on the network, e.g., 3-D convolution [17], [18], spectral angle loss [19], and spectral difference [20]. Due to the high spectral dimensionality of hyperspectral images, these methods need larger models (hundreds to thousands of feature maps or deeper networks) to super-resolve all spectral bands simultaneously, but the size of the network is constrained by the limited amount of data. Consequently, the expressiveness of CNN is hindered.

To overcome the drawbacks associated with high spectral dimensionality while exploring spectral correlations, several group-based methods have been proposed [21], [22], [23]. As shown in Fig. 1(a), the framework of group-based methods can be summarized as follows. The LR hyperspectral image is divided into several subgroups for shallow feature extraction, and the intermediate results of each subgroup are concatenated together for deep feature extraction and reconstruction. Although the current advanced group-based methods can greatly decrease the computational complexity, it still processes the whole image (the concatenated intermediate results) at once, which increases the difficulty of deep feature extraction and reconstruction. In a word, the advanced group-based methods are still troubled by the side effects introduced by the high dimensionality of hyperspectral images, resulting in unsatisfactory reconstruction performance.

Except for high spectral dimensionality, the high similarity among bands is another outstanding characteristic of

hyperspectral images. All spectral bands of a hyperspectral image are taken at the same scene, resulting in large spatial information redundancy in hyperspectral data. Since different spectral bands look quite similar, the missing details in the specific band are likely to be complemented by other bands, especially adjacent bands. However, the advanced group-based methods usually simply stack the features of each group together, failing in effectively exploiting the beneficial information contained in neighboring image groups. Consequently, there is still room for further improvement in reconstruction performance.

Taking all the above into account, we propose a group-based embedding learning and integration network (GELIN) for hyperspectral image super-resolution, as shown in Fig. 1(b), which effectively explores and exploits the inherent data characteristic of hyperspectral images. Except for spatial information, spectral information is also crucial for hyperspectral image super-resolution. The high similarity among different bands introduces much complementary information, which contains beneficial details that can be further exploited to boost the super-resolution performance. Moreover, failing to explore the correlation among spectra will result in severe spectral distortion, which is highly detrimental to the subsequent high-level vision tasks, i.e., classification [24], [25] and target detection [26]. Based on the above observations, we design the spatial–spectral embedding learning module (SSELM) to simultaneously explore the spatial and spectral information of hyperspectral images. The advanced self-calibrated convolution (SCConv) [27] is introduced to enable spatial feature extraction with a large receptive field. Similar to many previous group-based methods, SSELM adopts the group-convolution and parameter-sharing strategy to reduce the computational burden. After acquiring the immediate results of each group, the advanced group-based methods usually simply stack them together for further reconstruction, which makes insufficient utilization of the complementary information in hyperspectral data. To practically resolve this problem, we propose a neighboring group integration module (NGIM) to reconstruct each hyperspectral image group individually with the assistance of complementary information in neighboring image groups. Specifically, NGIM extracts the missing details from two information sources, i.e., the target image group and its neighboring image groups. Then, NGIM enhances the difference between the features of two information sources to obtain the discriminative embedding, which represents the complementary information among neighboring image groups. Finally, the discriminative embedding is fused back with the target image group to reconstruct the refined HR hyperspectral image. In this way, not only can we make full use of the complementarity but also alleviate the side effects introduced by the high dimensionality of hyperspectral data. The qualitative and quantitative experiments on both nature hyperspectral datasets and remote sensing hyperspectral datasets demonstrate our superiority over other state-of-the-art methods both visually and metrically.

In summary, our contributions can be concluded as follows.

1) A novel single hyperspectral image super-resolution method GELIN is proposed. Both spatial information

and spectral information are well-considered to boost the final reconstruction performance.

2) Rather than modeling all spectral bands simultaneously, we super-resolve the hyperspectral image in a group-by-group manner, which not only explores the spectral correlations but also alleviates the difficulty of feature extraction and reconstruction.

3) Considering the strong correlations among spectra, we propose an NGIM to exploit the rich complementary information contained in hyperspectral data, leading to better spatial fidelity.

The rest of this article is organized as follows. In Section II, we describe some existing methods related to our work. Section III gives details about the newly proposed GELIN method. Then, the network settings and experimental results involving ablation analysis, qualitative experiments, and quantitative experiments are reported in Section IV. Finally, Section V presents the conclusion.

II. RELATED WORKS

A. Single Natural Image Super-Resolution

In recent years, deep convolutional networks have shown a remarkable capability of recovering the missing details in natural images. As the pioneering work of the natural image super-resolution task, Dong et al. [28] upscaled an input image by bicubic interpolation and trained a CNN with three layers (SRCNN) to recover the missing details, which showed better performance over traditional methods. Later, some methods adopted the residual learning strategy to stable the training process and deepen the networks, including VDSR [29], DRCN [30], EDSR [31], DRRN [32], LapSRN [33], and CRD-Net [34]. Based on the observation that higher peak signal-to-noise ratio (PSNR) does not necessarily equate to a better super-resolution result in perceptual, SRGAN [35] adopted a generative adversarial network for photorealistic results. Moreover, RDN [36], DBPN [37], and RCAN [38] were proposed by introducing dense connections, iterative up-down sampling, and the attention mechanism. To further exploit the potential of the attention mechanism, SAN [39] proposed a second-order attention network for more powerful feature expression, and HAN [40] proposed a holistic attention network to model the interdependencies among different convolutional layers. Most recently, Mei et al. [41] combined nonlocal attention and sparse representation and proposed nonlocal sparse attention to retain long-range modeling capability. GLEAN [42] used pretrained generative adversarial networks as a latent bank to improve the performance of reconstructed images. Although these methods have proven to be efficient in natural image super-resolution, they fail to achieve satisfactory results when applied directly to the hyperspectral image. Reconstructing the hyperspectral image in a bandwise manner, these single natural image super-resolution methods take no spectral information into consideration. As a result, the produced super-resolution results suffer from severe spectral distortion.

B. Single Hyperspectral Image Super-Resolution

In recent years, the single hyperspectral image super-resolution has gained more and more attention due to its feasibility and efficiency. Akgun et al. [43] first reconstructed the HR hyperspectral images by modeling the hyperspectral image acquisition process. Li et al. [44] conducted an analysis on spectral mixture and sparsity of spatial–spectral group and then proposed a novel hyperspectral image super-resolution framework. In [12], the total variation prior was utilized to guide the reconstruction process of hyperspectral images. However, these methods usually resort to handcrafted priors, which are time-consuming and sophisticated to optimize. Moreover, the handcrafted priors take no external information into consideration, limiting the reconstruction performance. Recently, several hyperspectral image super-resolution methods based on deep learning techniques have been proposed. Hu et al. [20] designed a spectral difference network to reduce spectral distortion. Yuan et al. [15] adopted a transfer learning strategy to learn the mapping between LR and HR hyperspectral images and applied nonnegative matrix factorization to enforce collaborations between LR and HR images. Xie et al. [45] blended the feature matrix extracted by a deep neural network with the nonnegative matrix factorization strategy for super-resolving real-scene hyperspectral images. However, these methods exploit spatial and spectral features, leading to spatial–spectral inconsistencies. In recent years, end-to-end methods have become mainstream. Mei et al. [17] designed a 3-D fully CNN to explore the correlations among spectra, but the computational cost is very high. Li et al. [19] developed a grouped deep recursive residual network (GDRRN) and used spectral angle mapper (SAM) loss to reduce spectral distortion. Based on the spectral grouping strategy, Jiang et al. [21] proposed a spatial–spectral prior network (SSPSR) with group convolution and progressive upsampling framework. Li et al. [18] introduced a mixed convolutional network (MCNet) to extract the potential features by 2-D/3-D convolution, but processing all bands simultaneously with 3-D convolution would introduce heavy computational cost, especially for remote sensing hyperspectral images. Wang et al. [16] designed a dual-channel network to reconstruct images in a band-by-band manner and jointly exploited the information from both single band and adjacent bands. However, such a band-by-band scheme requires considerable training time. Most recently, Liu et al. [23] developed a spectral grouping and attention-driven residual dense network (SGARDN), which considered the block characteristic of hyperspectral images. Wang et al. [22] designed a group-based feedback network with a regularization network to explore the coherence among bands and preserve the spatial–spectral structure of the scene. Li et al. [46] proposed the multidomain feature learning strategy to explore spatial and spectral knowledge by sharing spatial information.

Generally, deep learning-based methods achieve better reconstruction results than traditional methods, and group-based methods show better performance over nongroup ones. In most cases, the group-based methods simultaneously model all spectral bands during deep feature extraction and reconstruction, which places great demand on the network size. However, the limited amount of available hyperspectral data cannot support a large-scale network, which becomes a troublesome contradiction. Besides, the high similarity
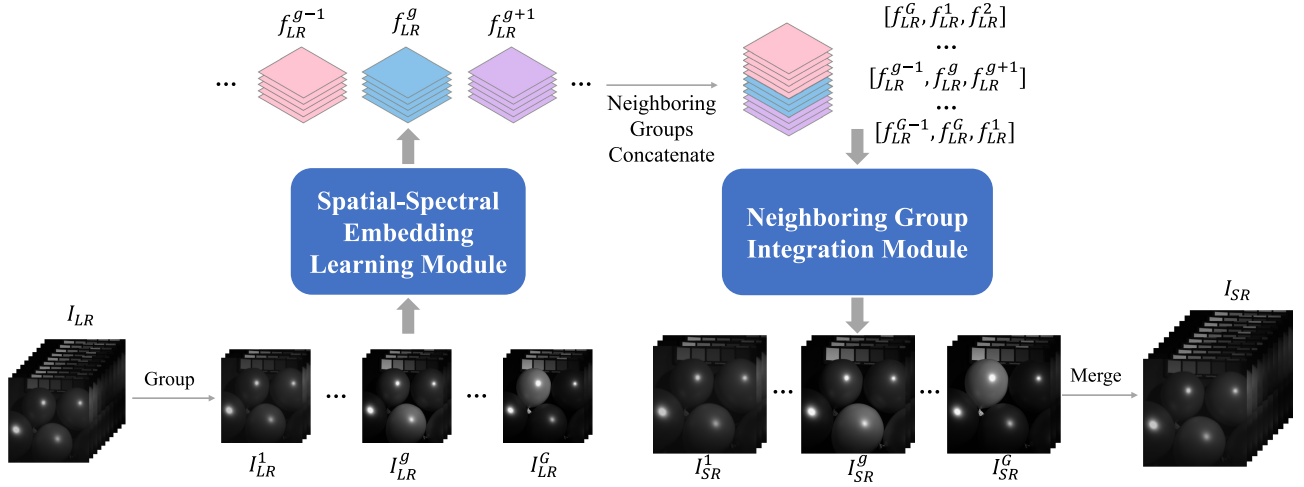
Fig. 2. Architecture of the proposed GELIN consisting of the SSELM and the NGIM.

between different bands is another notable characteristic of the hyperspectral image, which introduces much spatial redundant information that has proven to be rewarding for hyperspectral super-resolution. Nevertheless, the shallow features of each hyperspectral image group are simply concatenated together in the advanced group-based methods. There are insufficient studies on how to effectively exploit spatial redundancy to extract and recover more missing details.

To this end, we propose a novel hyperspectral image super-resolution network to reconstruct HR images in a group-by-group manner. An SSELM is designed to explore the data characteristic of hyperspectral images, which achieves spatial and spectral information extraction simultaneously. Moreover, an NGIM is proposed to reconstruct the target image group with the assistance of its neighboring groups, which exploits the complementarity contained in redundant information to recover more missing details.

III. PROPOSED METHOD

In this section, we will detail the proposed hyperspectral super-resolution method GELIN. As shown in Fig. 2, GELIN mainly consists of two parts: SSELM, exploiting the spatial and spectral information of each hyperspectral image group, and NGIM, fully utilizing the redundant information contained in neighboring image groups to assist the reconstruction process of the target group. In the following, we will introduce the network framework first and then detail SSELM and NGIM.

A. Overall Framework

Denoting the input LR hyperspectral image as $I_{\mathrm{LR}} \in \mathbb{R}^{h \times w \times C}$, its original HR hyperspectral image as $I_{\mathrm{HR}} \in \mathbb{R}^{H \times W \times C}$, and the output super-resolved hyperspectral image as $I_{\mathrm{SR}} \in \mathbb{R}^{H \times W \times C}$, our goal is to predict super-resolution results I_{SR} from LR input I_{LR} through the proposed network, and the output I_{SR} should be consistent with I_{HR} as much as possible in both spatial and spectral. The above procedure can be formulated as

$$I_{\mathrm{SR}} = H_{\mathrm{GELIN}}(I_{\mathrm{LR}}) \qquad (1)$$

where H_{GELIN} denotes the proposed GELIN method.

Among advanced methods, group-based methods have shown superiority over other methods. Following the same insight, we divide the input LR hyperspectral image into G groups, $I_{\mathrm{LR}} = [I_{\mathrm{LR}}^1, \ldots, I_{\mathrm{LR}}^g, \ldots, I_{\mathrm{LR}}^G]$. Specifically, following the settings in SSPSR [20], each image group has c bands, and the neighboring hyperspectral image groups have o overlaps. The settings of c and o are discussed in Section IV. For each group I_{LR}^g, we apply the proposed SSELM to learning its spatial–spectral embedding f_{LR}^g

$$f_{\mathrm{LR}}^g = H_{\mathrm{SSELM}}(I_{\mathrm{LR}}^g) \qquad (2)$$

where H_{SSELM} denotes the proposed SSELM, which consists of one convolution layer for shallow feature extraction and R spatial–spectral embedding learning blocks (SSELBs) for deep feature extraction. Considering the high similarity among different bands, we adopt the same SSELM to process different image groups, which greatly reduces the computational cost.

After acquiring the spatial–spectral embedding of each hyperspectral image group, we reconstruct the final HR hyperspectral image in a group-by-group manner. For fully utilizing the redundant information to recover more missing details, when reconstructing the gth group, the NGIM receives both the embedding of the target group and its neighboring groups and then generates the super-resolved hyperspectral image of the gth group with the assistance of its neighboring groups

$$I_{\mathrm{SR}}^g = H_{\mathrm{NGIM}}\left(f_{\mathrm{LR}}^{g-1}, f_{\mathrm{LR}}^g, f_{\mathrm{LR}}^{g+1}\right) \qquad (3)$$

where H_{NGIM} denotes the function of the proposed NGIM and f_{LR}^{g-1}, f_{LR}^g, and f_{LR}^{g+1} are the features of the $(g-1)$th, gth, and $(g+1)$th image group, respectively.

Finally, the super-resolved hyperspectral images can be obtained by merging each image group and adding the results with global residual (we ignore it in Fig. 2), which can be expressed as

$$I_{\mathrm{SR}} = \left[I_{\mathrm{SR}}^1, \ldots, I_{\mathrm{SR}}^g, \ldots, I_{\mathrm{SR}}^G\right] + I_{\mathrm{LR}} \uparrow \qquad (4)$$

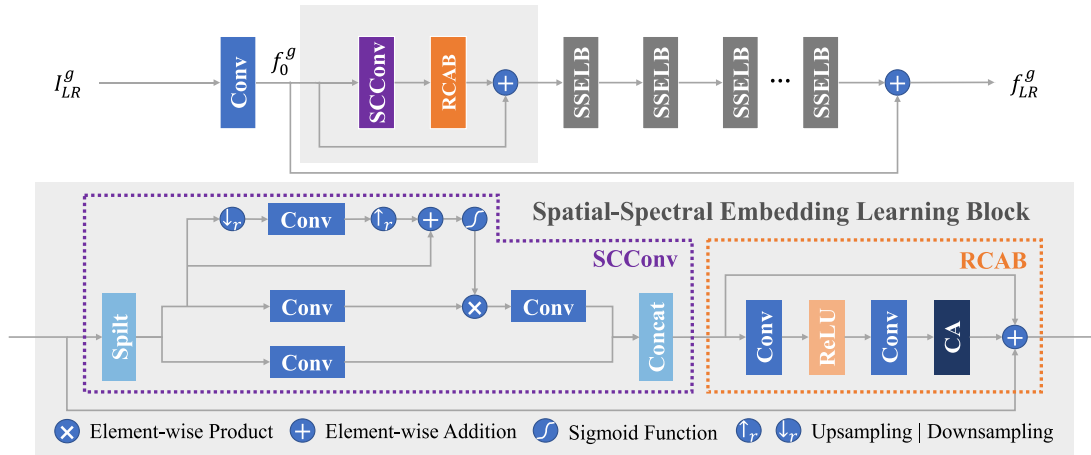where $I_{\mathrm{LR}} \uparrow$ is the bicubic upsampling version of the input LR hyperspectral image.

Fig. 3. Structure of the proposed SSELM, which consists of a normal convolutional layer and several cascaded SSELBs.

B. Spatial–Spectral Embedding Learning Module

In natural image super-resolution, only spatial information needs to be considered. However, in hyperspectral image super-resolution, both spatial information and spectral information are of great importance. Failing to thoroughly exploit spatial and spectral features will result in insufficient detail recovery and undesirable spectral distortion.

For the above reasons, we design an SSELM to extract spatial and spectral features simultaneously, and the architecture of SSELM is shown in Fig. 3. Specifically, SSELM contains one convolutional layer for shallow feature extraction and R SSELBs, which can be expressed as

$$
\begin{aligned}
f_0^g &= H_{\text{sfe}}\left(I_{\text{LR}}^g\right) \\
f_{\text{LR}}^g &= H_{\text{SSELB}_R}\left(H_{\text{SSELB}_{R-1}}\left(\cdots H_{\text{SSELB}_1}\left(f_0^g\right)\cdots\right)\right) + f_0^g
\end{aligned}
\tag{5}
$$

where H_{sfe} refers to a convolutional layer with the kernel size of 3×3 for shallow feature extraction and H_{SSELB_R} refers to the Rth SSELB. Also, we introduce residual learning "$+f_0^g$" to stable and accelerate the training process.

As the building block of SSELM, SSELB is designed to learn spatial–spectral embedding of hyperspectral data. As shown in Fig. 3, we cascade the SCConv and the residual channel attention block (RCAB) to form the SSELB. The RCAB has been used as the building block in hyperspectral image super-resolution for spectral correlations exploration, and however, the spatial information learning ability of the network could be further enhanced since the commonly used 2-D convolutional layer has a relatively limited receptive field. Inspired by SCNet [27], we introduce the SCConv in our building block. The SCConv allows feature extraction with large receptive field, leading to more representative spatial embedding learning. Specifically, for the input features $X \in \mathbb{R}^{h \times w \times C_f}$, the SCConv first uniformly splits it into two parts $X_1, X_2 \in \mathbb{R}^{h \times w \times (1/2)C_f}$, and each part will be processed by different branches. For the first part X_1, the average-pooling operation with ratio r is first conducted to downsample it to the lower spatial resolution space, and then, one convolutional layer with the kernel size of 3×3 is equipped to extract its spatial features in small-scale space. Finally, a bilinear interpolation operator is introduced to map the features back to the original spatial resolution space. The above workflow can be formulated as

$$
X_1' = \text{Up} \uparrow_r \left(H_1(\text{Down} \downarrow_r (X_1))\right)
\tag{6}
$$

where Down $\downarrow_r$ refers to the average-pooling operation with ratio r, H_1 refers to a convolutional layer with the kernel size of 3×3, and Up $\uparrow_r$ refers to the bilinear interpolation operator with ratio r. After collecting the features in the lower spatial resolution space, the formula for calibration operation is defined as

$$
Y_1' = H_2(X_1) \cdot \sigma\left(X_1 + X_1'\right)
\tag{7}
$$

where H_2 refers to a convolutional layer with the kernel size of 3×3 and σ is the sigmoid function. Finally, the output of this branch can be obtained by processing intermediate feature Y_1' with another convolutional 3×3 layer

$$
Y_1 = H_3\left(Y_1'\right).
\tag{8}
$$

As for another part X_2, we simply use a 3×3 convolutional layer to obtain the output of this branch

$$
Y_2 = H_4(X_2)
\tag{9}
$$

and finally, we concatenate Y_1 and Y_2 to obtain the output of the SCConv.

The above self-calibration operation performs spatial feature extraction at two different spaces, which enables each pixel to consider the information both from the original space and the lower spatial resolution space. As for the super-resolution task, a larger receptive field enables the network to effectively search similar image patches for recovering missing details. Moreover, the features in lower spatial resolution space carry much low-frequency information, which allows the network to spend more computational sources on the extraction and recovery of high-frequency details, leading to better reconstruction performance.

Compared with natural images, the correlation among bands is a unique characteristic of hyperspectral images. Failing to mine the spectral information can lead to unsatisfactory

reconstruction results. Taking the spectral pattern into consideration, we choose the widely used RCAB [38] to explore the correlation among different bands. To be specific, each RCAB consists of two convolutional layers with a ReLU layer in the middle, followed by channel attention (CA). The CA first obtains the aggregated information of input features by global pooling and then uses one dimensionality-reduction layer and one dimensionality-increasing layer followed by the sigmoid gating function to capture the cross-channel correlations. Finally, we obtain the weight vector to adaptively rescale the importance of each channel.

Through the well-designed SSELM, we can not only extract spatial information of the hyperspectral image in a large receptive field but also explore the correlations among different spectra, both of which are crucial for hyperspectral image super-resolution.

C. Neighboring Group Integration Module

In order to deal with problems introduced by the high dimensionality of hyperspectral data, many advanced methods [21], [22], [23] adopt the group-convolution strategy to design the network. Compared with nongroup-based methods, these group-based methods have shown better performance in reconstruction quality and computational cost. As mentioned earlier, in advanced group-based methods, the intermediate results generated by group shallow feature extraction are usually stacked together. Then, these methods conduct deep feature extraction and reconstruction on all spectral bands simultaneously, which requires a large network with even thousands of feature maps. However, the limited available data constrain the size of the network, which hinders the characteristics of hyperspectral data from being well-explored. In brief, the advanced group-based methods are still troubled by the high dimensionality of hyperspectral data, which can be further optimized.

Moreover, the strong correlations among different bands have been proven in previous work [23], which should not be ignored in hyperspectral image super-resolution. When reconstructing the target image group, the complementary information among neighboring image groups can contribute to the missing details recovery. To utilize the valuable complementary information, an intuitive way is to directly concatenate the features from two information sources, i.e., neighboring groups and the target group together. However, this approach fails to explicitly represent the complementarity between groups since two information sources are concatenated and treated as one. Consequently, the strong correlations among spectra are not well-exploited, limiting the reconstruction quality of super-resolved hyperspectral images.

To address the above problems, we propose an NGIM to reconstruct the HR hyperspectral image in a group-by-group manner. At the same time, NGIM effectively exploits the complementary information contained in neighboring groups to recover more high-frequency details. The architecture of NGIM is shown in Fig. 4.

Let $f_{\mathrm{LR}}^g \in \mathbb{R}^{h \times w \times C_f}$ denote the embedding of the gth group learned by SSELM. When reconstructing the gth hyperspectral
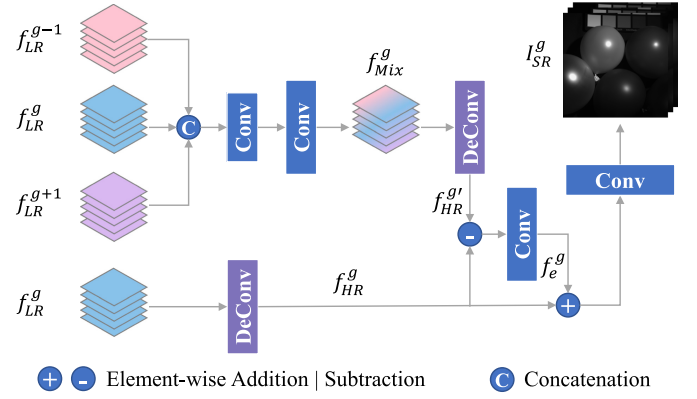

Fig. 4. Architecture of the proposed NGIM. The redundant information contained in neighboring hyperspectral image groups is exploited to complement missing details in the target image group.

image group, the upper branch of NGIM takes both the embedding of the target group and its neighboring groups as input. Then, the upper branch uses two convolutional layers with different numbers of filters to mix these groups. The above workflow can be formulated as

$$f_{\mathrm{mix}}^g = H_U\left(H_D\left(\left[f_{\mathrm{LR}}^{g-1}, f_{\mathrm{LR}}^g, f_{\mathrm{LR}}^{g+1}\right]\right)\right) \tag{10}$$

where f_{mix}^g refers to the mixed embedding and H_D refers to the first convolutional layer that downscales the filter number of the input to $(1/2)C_f$. H_U refers to the followed convolutional layer that upscales the filter number to C_f. Considering the high dimensionality of the input stacked features, we adopt two convolutional layers with different numbers to reduce the parameters.

Then, we upsample f_{mix}^g and f_{LR}^g to obtain their representation in HR space, which can be expressed as

$$f_{\mathrm{HR}}^{g'} = U_1\left(f_{\mathrm{mix}}^g\right) \tag{11}$$

$$f_{\mathrm{HR}}^g = U_2\left(f_{\mathrm{LR}}^g\right) \tag{12}$$

where U_1 and U_2 refer to two upsampling functions that are conducted by the deconvolutional layer.

$f_{\mathrm{HR}}^{g'}$ and f_{HR}^g recover missing details from different sources. The difference between them contains the complementary information, which facilitates the detail recovery of the target image group. In order to adaptively retain the complementary information, we further enhance the difference to obtain the discriminative embedding f_e^g

$$f_e^g = H_e\left(f_{\mathrm{HR}}^{g'} - f_{\mathrm{HR}}^g\right) \tag{13}$$

where H_e refers to the convolutional layer used to enhance the difference.

Finally, we fuse the discriminative embedding f_e^g with f_{HR}^g to borrow the complementary information from neighboring groups to the target group and use one convolutional layer H_{rec} to reconstruct the super-resolution results of the gth hyperspectral image group. The procedure can be formulated as

$$I_{\mathrm{SR}}^g = H_{\mathrm{rec}}\left(f_e^g + f_{\mathrm{HR}}^g\right). \tag{14}$$

$\mathcal{L}_1$	$\mathcal{L}_s$	$\mathcal{L}_g$	PSNR↑	SAM↓	SSIM↑
√			39.9857	3.0145	0.9535
√	√		40.2292	2.8699	0.9537
√		√	40.1505	3.0013	0.9534
√	√	√	40.2676	2.8657	0.9537

D. Loss Function

In advanced super-resolution methods, ℓ_1 and ℓ_2 losses are the most commonly used loss functions. Since the ℓ_2 loss tends to generate oversmooth results, we adopt the ℓ_1 loss to measure the accuracy of reconstructed HR hyperspectral images. The ℓ_1 loss can be formulated as

$$\mathcal{L}_1(\Theta) = \frac{1}{N} \sum_{n=1}^{N} \left\| I_{\text{SR}}^n - I_{\text{HR}}^n \right\|_1 \qquad (15)$$

where I_{SR}^n and I_{HR}^n are the nth reconstructed super-resolved hyperspectral image and original HR hyperspectral image, respectively, N is the number of images in one training batch, and Θ refers to the parameter set of the proposed network.

Although the above ℓ_1 loss can retain the spatial information well, it takes no spectral consistency into consideration. In order to preserve spatial and spectral information simultaneously, we introduce the SAM loss in [19] to preserve the spectral profile of super-resolution results. The SAM loss can be formulated as

$$\mathcal{L}_s(\Theta) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{1}{\pi} \arccos\left(\frac{I_{\text{SR}}^{n,i} \cdot I_{\text{HR}}^{n,i}}{\left\| I_{\text{SR}}^{n,i} \right\|_2 \cdot \left\| I_{\text{HR}}^{n,i} \right\|_2} \right) \qquad (16)$$

where N_p is equal to $H \times W$ and $I_{\text{SR}}^{n,i}$ refers to the ith spectral vector of the nth image.

Furthermore, inspired by [22], we introduce a gradient loss to preserve the sharpness of the reconstructed images in both spatial and spectral domains

$$\mathcal{L}_g(\Theta) = \frac{1}{N} \sum_{n=1}^{N} \left\| \text{M}\left(I_{\text{SR}}^n \right) - \text{M}\left(I_{\text{HR}}^n \right) \right\|_1 \qquad (17)$$

where M computes the gradient value along horizontal, vertical, and spectral dimensions of image.

In summary, the joint loss for the proposed network can be formulated as

$$\mathcal{L}_{\text{joint}}(\Theta) = \mathcal{L}_1 + \lambda_s \mathcal{L}_s + \lambda_g \mathcal{L}_g \qquad (18)$$

where λ_s controls the weight of SAM loss and λ_g controls the weight of gradient loss. In our experiment, we set $\lambda_s = 0.3$ and $\lambda_g = 0.1$ empirically.

In Table I, we report the reconstruction results under different losses. When adopting spectral loss $\mathcal{L}_s$ with pixel loss $\mathcal{L}_1$, both PSNR and SAM metrics show significant improvement. The gradient loss $\mathcal{L}_g$ contributes to the preservation of sharp parts of the image, resulting in notable improvement in the PSNR metric. Finally, we train the network with joint loss to constrain the network in both spatial and spectral domains.

E. Implementation Details

In the proposed network GELIN, the input hyperspectral images are divided into G groups with o overlaps between neighboring groups. Each group has c spectral bands, and we select the last c bands as the last group. We set the kernel size of all convolutional layers to 3×3 except for that in CA, where the size is set to 1×1. Zero padding is applied to keep the spatial size of feature maps. When the scale factor is 4, we use the 8×8 deconvolutional layer with four stridings and two paddings. When the scale factor is 8, we use the 12×12 deconvolutional layer with eight stridings and two paddings. All the deconvolutional layers are followed by parametric rectified linear units (PReLUs). Following the settings in [38], we set the reduction ratio in CA as 16. Considering that the input LR hyperspectral images have a small spatial resolution, we set the downsampling rate in SCCConv as 2. In the proposed NGIM, both the 1st and gth groups have only one neighboring image group, so we use the gth and 1st groups to pad their absented neighboring image group, respectively. For the training phase, we set the mini-batch size of 8 and use the Adam optimizer with an initial learning rate of $1e^{-4}$, which decays by ten times after 8.36k iterations, while the total iteration is 8.8k. Our model is implemented by Pytorch on NVIDIA TITAN V.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Experimental Setup

In this section, we evaluate our method on both nature hyperspectral image datasets, i.e., CAVE dataset [47] and ICVL dataset [48], and remote hyperspectral image datasets, i.e., Chikusei dataset [49] and Reflective Optic System Imaging Spectrometer (ROSIS) dataset. We compare our method with eight state-of-the-art methods, including one single natural image super-resolution method EDSR [31] and six learning-based single hyperspectral image super-resolution methods, i.e., GDRRN [19], SFCSR [16], SSPSR [21], MCNet [18], RFSR [22], and MDFL [46]. Also, we choose the Bicubic interpolation method as the baseline. For EDSR [31], we treat the whole hyperspectral image as input and adjust the channels of the first and last convolutional layers accordingly. For the above methods, we try our best to achieve their best performance.

Six widely used evaluation indices are employed in experiments, including PSNR, SAM, structure similarity (SSIM), cross correlation (CC), erreur relative globale adimensionnelle de synthese (ERGAS), and root-mean-squared error (RMSE). For PSNR and SSIM, we report their mean value over all spectral bands. The best values for these indices are $+\infty$, 0, 1, 1, 0 and 0, respectively.

B. Experimental Results on CAVE Dataset

Captured by the cooled CCD camera with wavelengths ranging from 400 nm to 700 nm, the CAVE dataset contains 32 HR hyperspectral scenes, all of which are of

TABLE II
QUANTITATIVE PERFORMANCE OVER TEN TEST IMAGES ON THE
CAVE DATASET. BOLD REPRESENTS THE BEST RESULT AND
UNDERLINE REPRESENTS THE SECOND BEST

Method	d	PSNR↑	SAM↓	CC↑	RMSE↓	SSIM↑	ERGAS↓
Bicubic	4	35.1478	3.8769	0.9800	0.0204	0.9185	5.5747
EDSR [31]	4	38.9195	3.5364	0.9874	0.0144	0.9477	3.7676
GDRRN [19]	4	37.5605	3.5439	0.9853	0.0162	0.9396	4.3240
SFCSR [16]	4	39.1516	3.1026	0.9872	0.0142	0.9486	3.7333
SSPSR [21]	4	39.0520	3.1893	0.9877	0.0145	0.9493	3.6934
MCNet [18]	4	39.2585	3.0237	0.9878	0.0139	0.9505	3.6804
RFSR [22]	4	38.6029	3.0798	0.9870	0.0146	0.9477	3.9050
MDFL [46]	4	39.0883	3.0668	0.9874	0.0142	0.9493	3.7556
Ours	4	**40.2676**	**2.8657**	**0.9888**	**0.0130**	**0.9537**	**3.3599**
Bicubic	8	30.9401	5.2591	0.9594	0.0316	0.8455	8.6986
EDSR [31]	8	33.8730	4.9678	0.9692	0.0246	0.8829	6.4536
GDRRN [19]	8	32.9514	4.9250	0.9660	0.0265	0.8711	7.0478
SFCSR [16]	8	34.3690	4.3633	0.9709	0.0234	0.8907	6.1633
SSPSR [21]	8	34.2500	4.3306	0.9697	0.0240	0.8865	6.2384
MCNet [18]	8	34.5124	4.3407	0.9709	0.0233	0.8936	6.1249
RFSR [22]	8	34.5331	4.0530	0.9706	0.0231	0.8934	6.0608
MDFL [46]	8	34.4875	4.2917	0.9703	0.0235	0.8924	6.1441
Ours	8	**34.9338**	**3.9810**	**0.9711**	**0.0230**	**0.8953**	**5.9430**

size $512 \times 512 \times 31$. We randomly select 22 scenes from the dataset for training, and the remaining 10 scenes are for testing. When the scale factor is 4, we randomly extract patches with $128 \times 128 \times 31$ pixels as the HR hyperspectral image; when the scale factor is 8, we randomly extract patches with $256 \times 256 \times 31$ pixels. The corresponding LR hyperspectral images are generated by Bicubic interpolation. To investigate the influence of spectral band numbers of each group c and overlaps o, SSPSR [21] has conducted some experiments, which demonstrate that either setting c too large or too small can deteriorate the reconstruction performance, while a moderate c could boost the performance. The results also show that cropping spectrum bands in an overlapping way could benefit the reconstruction quality, but a large o would result in a heavy computational burden. In our experiments, the values of c and o are determined according to the band numbers of the hyperspectral image in each dataset, with the balance of computational cost and reconstruction performance taken into consideration. For the CAVE dataset, each hyperspectral image group has eight spectral bands with two overlaps, i.e., $c = 8$ and $o = 2$. In the SSELM, we set the number of the SSELB as $R = 8$.

In Table II, we compare the proposed method with some state-of-the-art methods. The average values of six evaluation indices for different scale factors are reported. Although EDSR [31] achieves relatively promising results in spatial information recovery, it neglects the spectral information, leading to severe spectral distortion. Considering the high dimensionality of hyperspectral data, GDRRN [19] designed a group-based recursive network with SAM loss to reduce spectral distortion. Following the same insight, SSPSR [21] and RFSR [22] both adopted group strategy and achieved significant improvement. SFCSR [16], MCNet [18], and MDFL [46] combined 2-D and 3-D convolutional layers to explore the correlation among spectral bands. However, 3-D convolution

significantly increases the computational burden, especially when super-resolving remote sensing hyperspectral images at the large scale factor. It can be noticed from Table II that our method significantly outperforms other algorithms in all evaluation indices. The PSNR value of our method is about 1 dB higher for scale factor $4\times$ and 0.4 dB higher for scale factor $8\times$ than that of the second-best method. Due to the proposed SSELM, our method can better extract the spatial and spectral information of hyperspectral data. Moreover, attributed to the proposed NGIM, the target image group can fully exploit the rich spatial redundant information contained in neighboring groups to recover more missing details.

Furthermore, we display the mean absolute differences between reconstructed super-resolved images and original HR images in both spatial and spectral. The mean error maps of four test scenes from the CAVE dataset at scale factors 4 and 8 are shown in Figs. 5 and 6, respectively. The PSNR and SAM values of different competing methods are reported at the bottom of these maps as supplementary materials. From the reconstruction results, we can find that the proposed GELIN method shows superiority over other methods in recovering the missing spatial information of HR hyperspectral images. Compared with the state-of-the-art methods, i.e., MCNet and RFSR, our method provides better recovery of details such as text, edge of feathers, and fruits, which is mainly because the NGIM brings much redundant information in neighboring groups to the target group to assist the reconstruction process. For the spectral fidelity, we demonstrate the mean spectral difference curve of three test scenes from the CAVE dataset in Fig. 7. In order to obtain more reliable conclusions, instead of randomly selecting a few points, we evaluate the spectral reconstruction quality from the perspective of the whole image. As can be seen from Fig. 7, the hyperspectral images reconstructed by our proposed method achieve the best spectral fidelity, while other competing methods suffer from more severe spectral distortion.

C. Experimental Results on Chikusei Dataset

The remote hyperspectral image dataset Chikusei, taken by the Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan, contains 128 spectral bands with a spectral range from 363 to 1018 nm and a spatial size of 2517×2335. Since some edge part of the image contains no information, we extract the central part of the original image with 2048×2048 pixels and then divide it into training, validation, and testing data. To be specific, we crop the bottom region of the image into four nonoverlap hyperspectral images with the size of $512 \times 512 \times 128$ and randomly select three subimages for testing and one for validation. The rest region of the image is cropped to form the training data, which contains 12 nonoverlapping hyperspectral images with $512 \times 512 \times 128$ pixels. When the scale factor is 4, we randomly extract patches with $128 \times 128 \times 128$ pixels as the HR image; when the scale factor is 8, we randomly extract patches with $256 \times 256 \times 128$ pixels. The corresponding LR hyperspectral images are generated by Bicubic interpolation. For the Chikusei dataset, we set the spectral
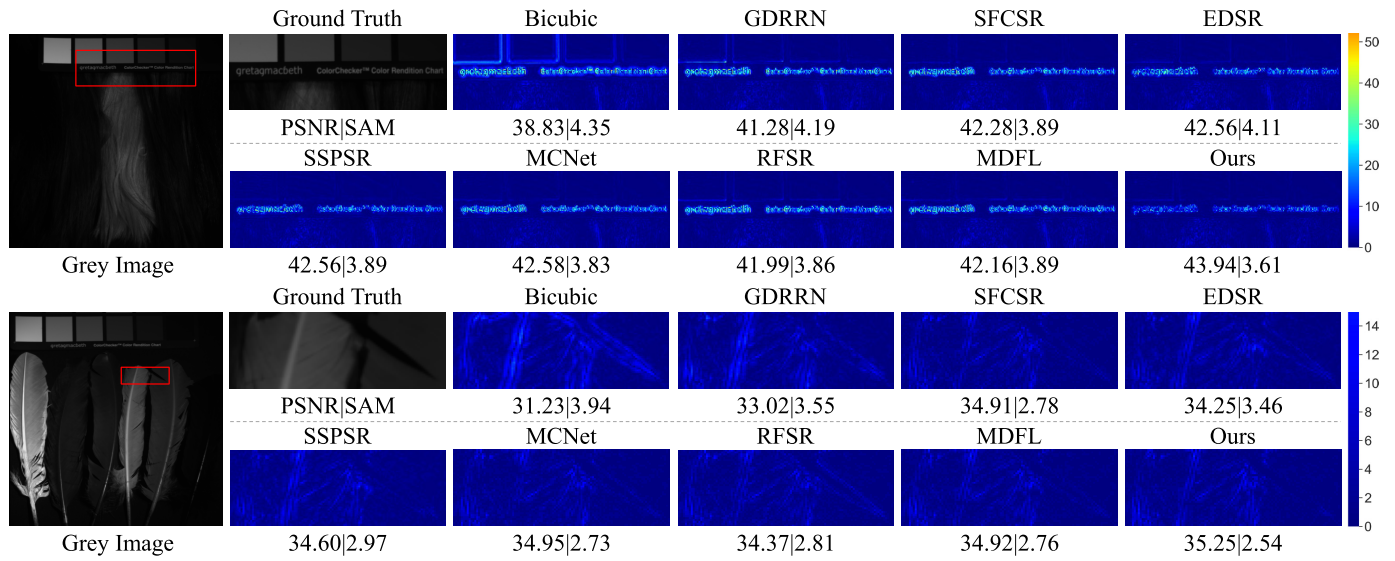
Fig. 5. Mean error maps of two test hyperspectral images in the CAVE dataset at the scale factor 4: hairs and feathers. The corresponding PSNR and SAM values of each comparison method are reported below each map.
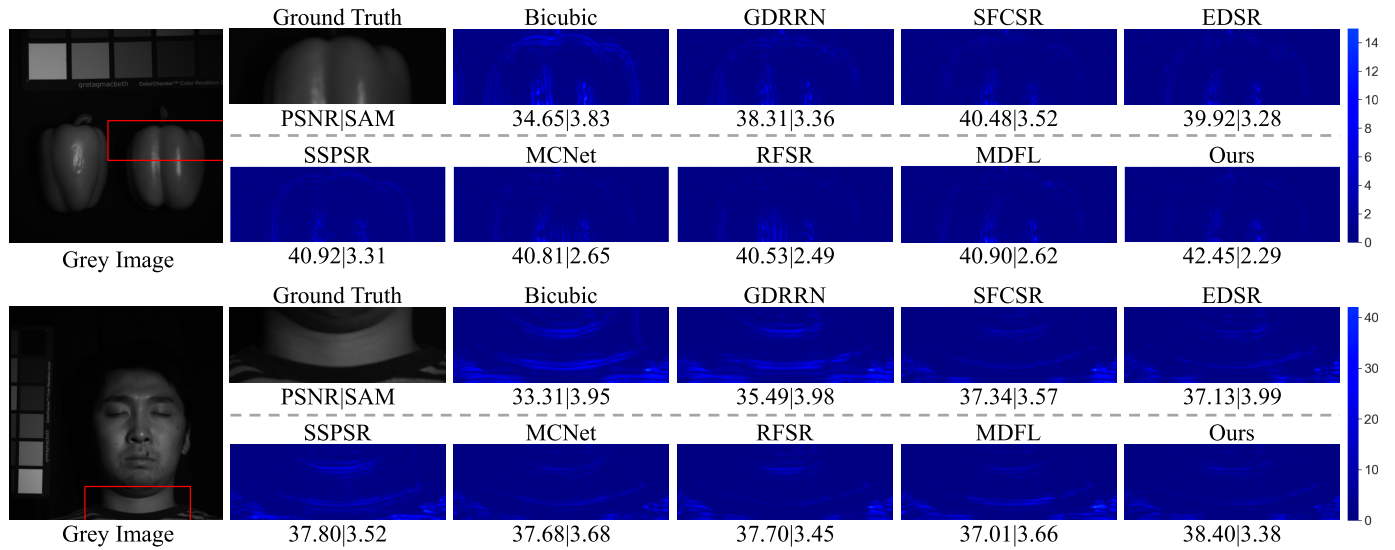

Fig. 6. Mean error maps of two test hyperspectral images in the CAVE dataset at the scale factor 8: real_and_fake_peppers and face. The corresponding PSNR and SAM values of each comparison method are reported below each map.
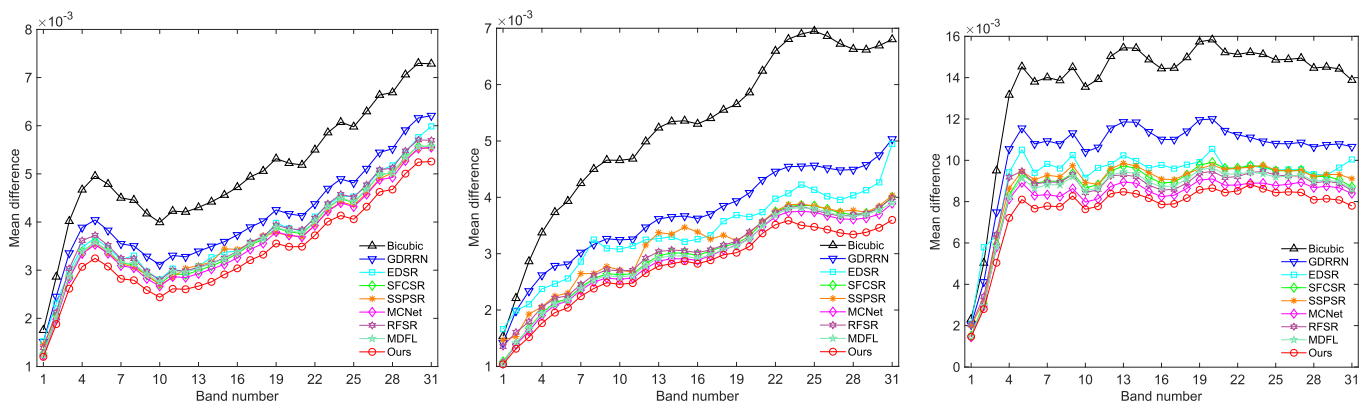

Fig. 7. Mean spectral difference curve of three test hyperspectral images in the CAVE dataset at the scale factor 4: hairs, fake_and_real_food, and chart_and_stuffed_toy.
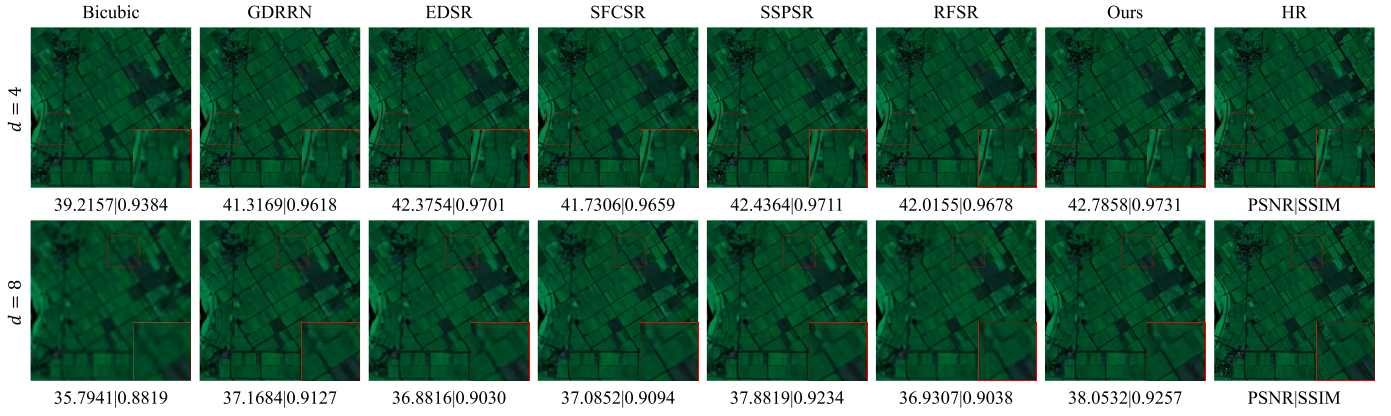
Fig. 8. Reconstructed images of one test hyperspectral image in the Chikusei dataset with spectral bands 70–100–36 as R–G–B at the scale factors 4 and 8. The corresponding PSNR and SSIM values of the comparison method are reported below each image.

TABLE III

QUANTITATIVE PERFORMANCE OVER THREE TEST IMAGES ON THE CHIKUSEI DATASET. BOLD REPRESENTS THE BEST RESULT AND UNDERLINE REPRESENTS THE SECOND BEST

Method	d	PSNR↑	SAM↓	CC↑	RMSE↓	SSIM↑	ERGAS↓
Bicubic	4	40.0063	2.5679	0.9317	0.0127	0.9262	5.5924
EDSR [31]	4	42.7949	1.8298	0.9615	0.0091	0.9617	4.1877
GDRRN [19]	4	41.8253	2.0015	0.9534	0.0103	0.9513	4.5967
SFCSR [16]	4	42.3938	1.9326	0.9589	0.0098	0.9558	4.2473
SSPSR [21]	4	43.1525	1.7398	0.9649	0.0089	0.9632	3.9385
RFSR [22]	4	42.7217	1.8184	0.9611	0.0094	0.9597	4.1342
Ours	4	**43.4615**	**1.6390**	**0.9668**	**0.0087**	**0.9652**	**3.7758**
Bicubic	8	36.6731	3.8477	0.8491	0.0186	0.8596	8.1526
EDSR [31]	8	37.4693	3.6168	0.8732	0.0167	0.8824	7.5653
GDRRN [19]	8	37.7942	3.0708	0.8834	0.0164	0.8906	7.1589
SFCSR [16]	8	37.8655	3.4363	0.8853	0.0167	0.8869	6.9665
SSPSR [21]	8	38.5092	2.8323	0.9011	**0.0151**	0.9037	6.5897
RFSR [22]	8	37.7712	3.3189	0.8815	0.0166	0.8841	7.0724
Ours	8	**38.6396**	**2.7585**	**0.9033**	**0.0151**	**0.9050**	**6.4101**

band number of each group to 32 and overlap to 8 to reduce the computational cost, i.e., $c = 32$ and $o = 8$. In the SSELM, we set the number of the SSELB as $R = 6$ because of fewer training samples.

Table III shows the average quantitative performance in terms of six evaluation indices on three testing images of six comparing methods. We can notice that our method performs better than other state-of-the-art methods at both two scale factors on the remote sensing hyperspectral dataset. It is worth pointing out that the performance gain on the Chikusei dataset is smaller than that on the CAVE dataset, which is attributed to the following factors. First, the Chikusei dataset is noisier and contains more bands, both of which increase the reconstruction difficulty. Moreover, the available amount of training data of the Chikusei dataset is also fewer. The above factors lead to a comparatively small performance gain. Since less training data are provided, the evaluation results show that our method can better utilize the characteristics of limited hyperspectral data.

Qualitative experiments are also conducted to demonstrate our superiority. In Fig. 8, we visualize the reconstruction results of one testing image from the Chikusei dataset of different competitive methods at scale factors 4 and 8. To be specific, we select the 70th, 100th, and 36th bands of the hyperspectral image in the testing dataset and treat them as the R–G–B channels of composite image for better visualization. From the composite images, we can easily observe that the proposed method can better recover both low- and high-frequency details than other algorithms (please refer to the area marked with red boxes). At the bottom of the visualization result, we also report the PSNR and SSIM values of the composite images. Moreover, we display the mean spectral difference curve of two test scenes at different scale factors from the Chikusei testing dataset in Fig. 9. Among the curve of different super-resolution methods, one of our methods is the lowest, which indicates that our method still achieves the best performance in reducing spectral distortion even when super-resolving remote sensing hyperspectral images with numerous bands.

D. Experimental Results on ICVL Dataset

Captured by the Specim PS Kappa DX4 hyperspectral camera, the ICVL dataset contains 201 images with a size of $1392 \times 1300 \times 31$. We randomly select 60 scenes for training and 15 scenes for testing. For training, we randomly extract patches with $64 \times 64 \times 31$ as the HR image for the scale factor 4, while the patch size is $128 \times 128 \times 31$ when the scale factor is 8. The corresponding LR patches are generated by Bicubic interpolation. Considering the testing efficiency, we crop the central region of each image with a spatial size of 512×512 to form the test dataset. As for hyperparameters of the network, the values of c, o, and R are the same as settings of the CAVE dataset.

Table IV shows the super-resolution performance of all comparing methods. We can observe that our method remains a significant advantage in the reconstruction on a larger dataset. Also, we have conducted qualitative experiments on the ICVL dataset. In Fig. 10, we visualize the composite images of scenes "eve_0331-1602" and "bguCAMP_0514-1718" at the scale factor 8. To be specific, the 6th, 10th, and 18th bands are treated as R–G–B channels of the composite
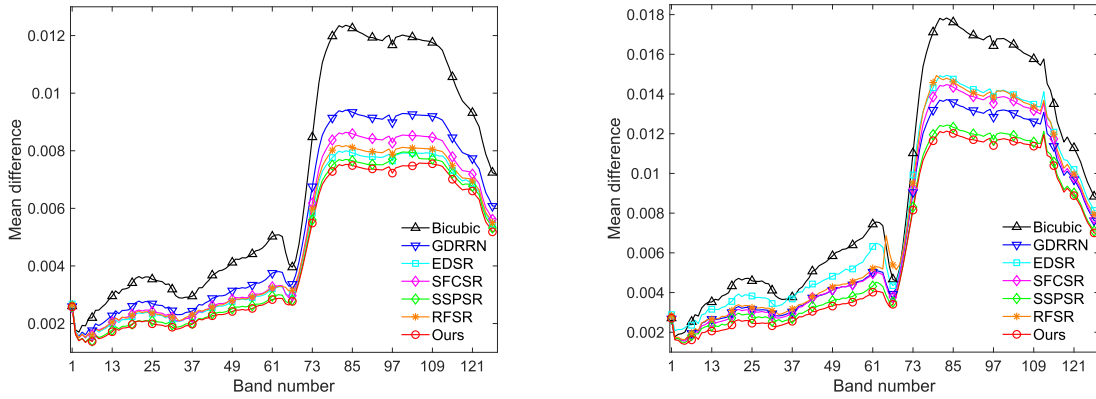
Fig. 9. Mean spectral difference curve of two test hyperspectral images in the Chikusei dataset at the scale factor (Left) 4 and (Right) 8, respectively.

TABLE IV

QUANTITATIVE PERFORMANCE OVER 15 TEST IMAGES ON THE ICVL DATASET. THE BOLD REPRESENTS THE BEST RESULT AND UNDERLINE REPRESENTS THE SECOND BEST

Method	d	PSNR↑	SAM↓	CC↑	RMSE↓	SSIM↑	ERGAS↓
Bicubic	4	43.7590	1.3269	0.9915	0.0090	0.9710	2.1311
EDSR [31]	4	46.2001	1.5090	0.9948	0.0064	0.9825	1.6368
GDRRN [19]	4	46.7422	1.2493	0.9954	0.0063	0.9829	1.5290
SFCSR [16]	4	47.2475	1.2021	0.9959	0.0058	0.9846	1.4661
SSPSR [21]	4	47.3203	1.2490	0.9960	0.0057	0.9853	1.4387
MCNet [18]	4	47.4723	1.1820	0.9960	0.0058	0.9850	1.4193
RFSR [22]	4	47.3863	1.2188	0.9959	0.0058	0.9850	1.4354
MDFL [46]	4	47.3071	1.1818	0.9959	0.0059	0.9848	1.4463
Ours	4	**47.8884**	**1.1445**	**0.9963**	**0.0055**	**0.9861**	**1.3556**
Bicubic	8	37.3165	1.9838	0.9636	0.0192	0.9033	4.3439
EDSR [31]	8	38.7711	2.4062	0.9740	0.0157	0.9216	3.6605
GDRRN [19]	8	38.8679	1.9681	0.9743	0.0159	0.9212	3.6255
SFCSR [16]	8	39.1738	1.9329	0.9765	0.0153	0.9252	3.5272
SSPSR [21]	8	39.5603	1.9499	0.9779	0.0144	0.9298	3.3767
MCNet [18]	8	39.4438	1.9070	0.9773	0.0149	0.9276	3.4327
RFSR [22]	8	39.5394	2.0230	0.9777	0.0146	0.9293	3.4011
MDFL [46]	8	39.3090	1.8893	0.9766	0.0152	0.9258	3.4902
Ours	8	**39.9347**	**1.7906**	**0.9794**	**0.0141**	**0.9331**	**3.2590**

TABLE V

QUANTITATIVE PERFORMANCE OVER SIX TEST IMAGES ON THE PU DATASET. BOLD REPRESENTS THE BEST RESULT AND UNDERLINE REPRESENTS THE SECOND BEST

Method	d	PSNR↑	SAM↓	CC↑	RMSE↓	SSIM↑	ERGAS↓
Bicubic	4	28.5691	5.2791	0.8980	0.0395	0.7597	6.0836
EDSR [31]	4	29.4816	4.9781	0.9169	0.0355	0.8099	5.5198
GDRRN [19]	4	29.7882	4.5563	0.9206	0.0345	0.8202	5.3315
SFCSR [16]	4	29.4515	5.2192	0.9149	0.0359	0.8057	5.5062
SSPSR [21]	4	30.1548	4.5135	0.9273	0.0331	0.8311	5.1032
RFSR [22]	4	29.9788	4.6626	0.9240	0.0338	0.8233	5.2141
Ours	4	**30.5362**	**4.3119**	**0.9339**	**0.0315**	**0.8468**	**4.9056**
Bicubic	8	25.5357	7.4831	0.7884	0.0558	0.5688	8.6066
EDSR [31]	8	26.1892	7.0085	0.8170	0.0520	0.6161	7.9404
GDRRN [19]	8	25.9817	6.6111	0.8077	0.0532	0.6224	8.1473
SFCSR [16]	8	26.1713	7.2966	0.8153	0.0522	0.6147	7.9462
SSPSR [21]	8	26.3244	6.5879	0.8229	0.0512	0.6334	7.8384
RFSR [22]	8	26.4084	7.0082	0.8250	0.0508	0.6281	7.7470
Ours	8	**26.4852**	**6.4670**	**0.8291**	**0.0502**	**0.6384**	**7.7026**

image. It can be observed from the composite images that our method produces sharper details and less blur. The PSNR and SSIM values also demonstrate our superiority. Moreover, we display the mean difference curves of the two scenes at the scale factor 8 in Fig. 10. The proposed method GELIN still achieves the best spectral fidelity.

E. Experimental Results on ROSIS Dataset

The Pavia Center (PC) dataset and Pavia University (PU) dataset were both acquired by the ROSIS sensor. After removing noisy bands, the PU dataset has 103 bands, while the PC dataset has 102 bands. To align the spectral bands, we remove the last band of the PU dataset. For training, the PC dataset is divided into nine images with the size of $360 \times 224 \times 102$. We randomly extract patches with $96 \times 96 \times 102$ pixels as the HR image for the scale factor 4, while the patch size for the scale factor 8 is $192 \times 192 \times 102$. The PU dataset is divided into six images with the size of $200 \times 160 \times 102$ for

testing. For the ROSIS dataset, the values of c, o, and R are the same as the Chikusei dataset settings.

Table V shows the super-resolution performance of six comparing methods. The proposed GELIN method can obtain the best spatial and spectral fidelity at two scales. Meanwhile, two mean difference maps from the PU dataset are shown in Fig. 11. The corresponding PSNR and SAM values are also demonstrated. As can be seen from the error maps, some edge contours do not appear in our results, which shows our methods can better recover the high-frequency details. Moreover, the mean spectral difference curve of two test images is shown in Fig. 12, which demonstrates our advantage in alleviating the spectral distortion of reconstructed images.

F. Ablation Study

1) Effectiveness of SCConv: Different from natural image super-resolution where only spatial information needs to be considered, in hyperspectral image super-resolution, both spatial fidelity and spectral consistency should be well preserved. In our method, an SSELM is designed to exploit both spatial and spectral embedding. The building block of SSELM, the SSELB, consists of an SCConv for spatial feature extraction
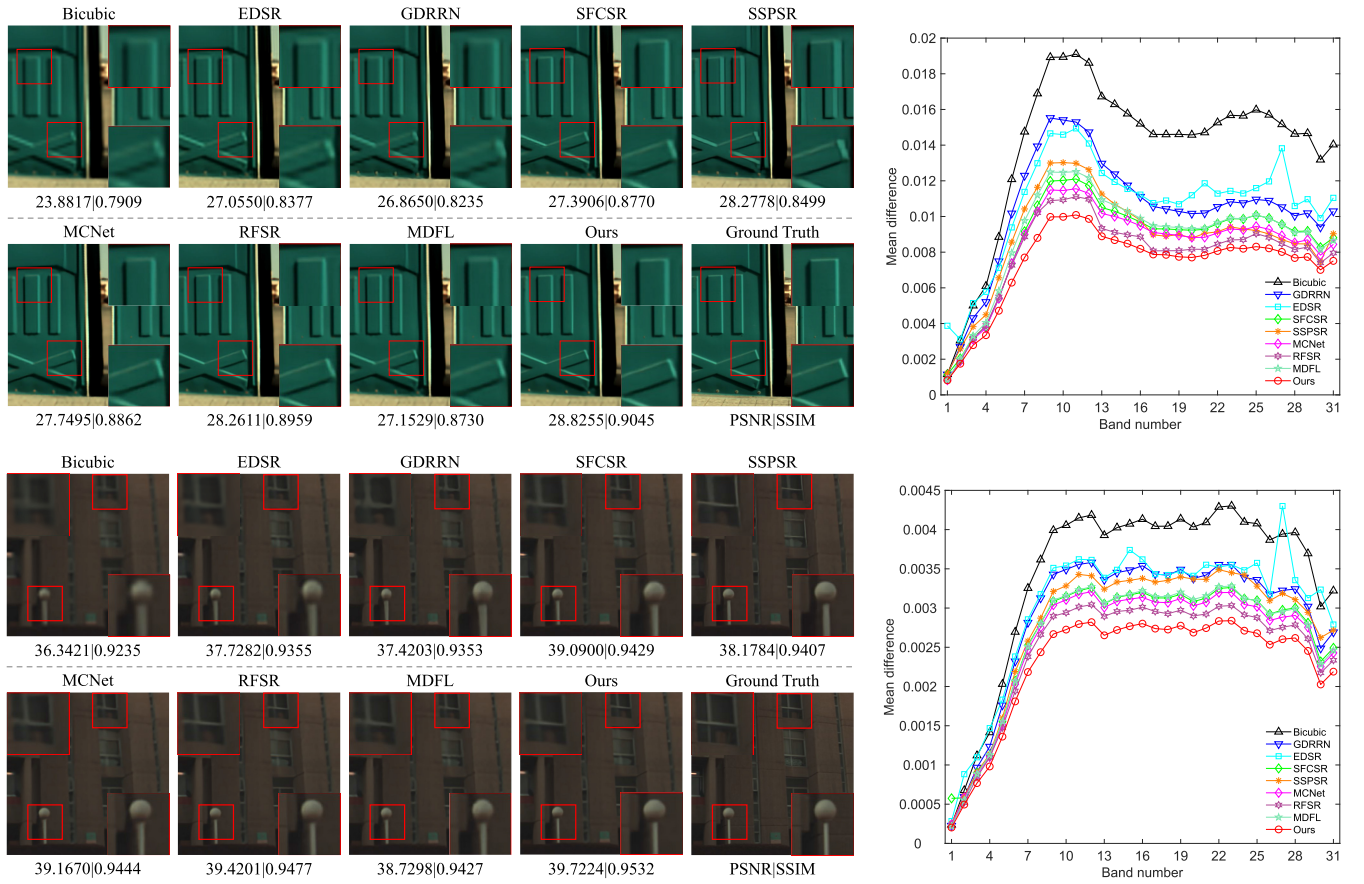
Fig. 10. (Left) Composite images of scenes "eve_0331-1602" and "bguCAMP_0514-1718" at the scale factor 8, the 6th, 10th, and 18th bands are treated as R–G–B channels. The PSNR and SSIM values are reported. (Right) Corresponding mean difference curves of the reconstructed hyperspectral images.
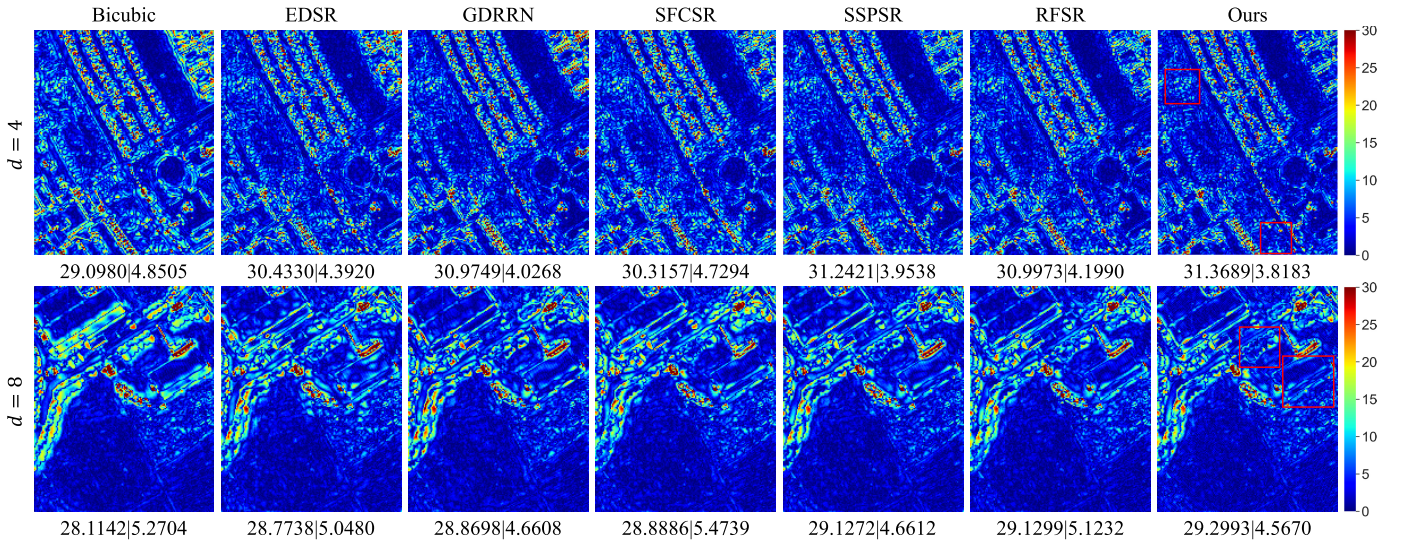

Fig. 11. Mean difference maps of two test hyperspectral images in the PU dataset at the scale factor 4 and 8. The PSNR and SAM values are reported.

and an RCAB for spectral information exploration. Since RCAB has proven to be useful for both natural image super-resolution [38] and hyperspectral image super-resolution [22], we do not investigate it here and just verify the effectiveness of the SCConv. In the ablation study, we simply remove the SCConv and local residual learning shortcut in SSELM.

As shown in Table VI, compared with "Ours w/o SC" where SCConv is removed and SSELB only contains RCAB, our model achieves considerable gains on PSNR and SSIM evaluation indices, which indicates that spatial information exploration is crucial for hyperspectral image super-resolution, and SCConv is capable of learning effective spatial representations.
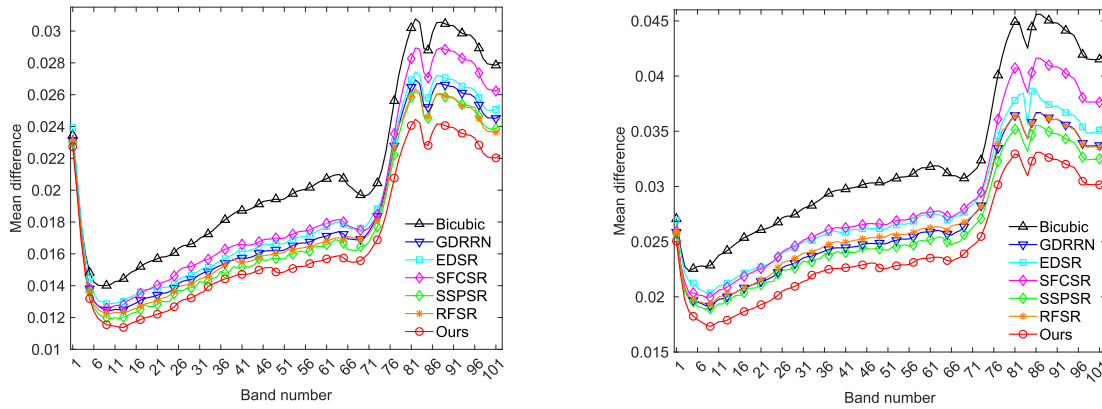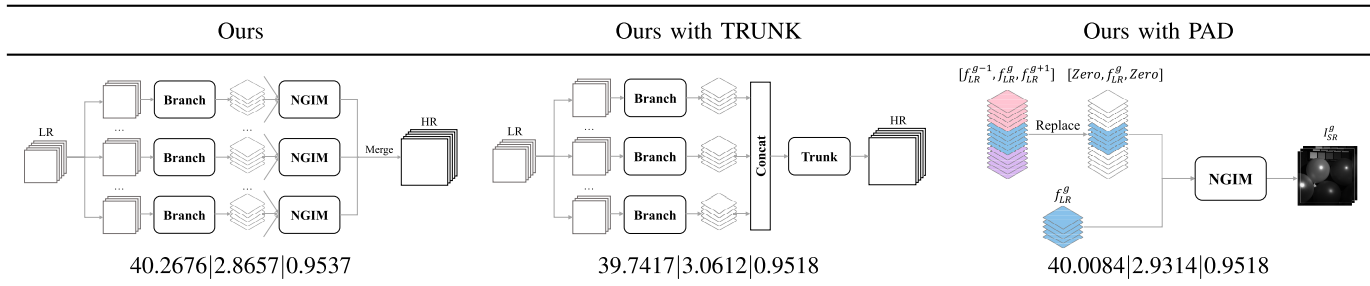
Fig. 12. Mean spectral difference curve of two test hyperspectral images in the PU dataset at the scale factor 4.

TABLE VI
ABLATION STUDY. QUANTITATIVE COMPARISONS AMONG SOME VARIANTS OF THE PROPOSED
METHOD OVER THE TESTING SET OF CAVE DATASET AT THE SCALE FACTOR 4

Variant	Model Size($\times10^6$)	Params($\times10^6$)	FLOPs($\times10^{12}$)	PSNR↑	SAM↓	CC↑	RMSE↓	SSIM↑	ERGAS↓
Ours	95.506	24.4280	12.9796	40.2676	2.8657	0.9888	0.0130	0.9537	3.3599
Ours w/o SC	77.022	19.7054	12.6652	40.0407	2.8871	0.9885	0.0133	0.9529	3.4373
Ours with 2D	95.407	24.4260	13.0519	40.2174	2.8785	0.9886	0.0131	0.9533	3.3813
Ours w/o NGIM	74.225	18.8984	2.0805	39.8447	2.9814	0.9878	0.0136	0.9511	3.5222

TABLE VII
STUDIES ON THE PROPOSED NGIM. THE SKETCHES OF THE PROPOSED GELIN AND TWO VARIANTS ARE SHOWN IN THE TABLE.
THE CORRESPONDING PSNR | SAM | SSIM VALUES ARE REPORTED AT THE BOTTOM OF THE TABLE

Ours	Ours with TRUNK	Ours with PAD
40.2676|2.8657|0.9537	39.7417|3.0612|0.9518	40.0084|2.9314|0.9518

To further demonstrate the advantages of SCConv, we replace it with a normal 2-D convolutional layer with the kernel size of 3 × 3 that is commonly adopted for spatial feature extraction. The variant of the above method is represented as "Ours with 2D" in Table VI. It can be observed from the results that our model equipped with SCConv gets a slight improvement in all evaluation indices with an ignorable increase in the number of parameters and even less computational cost, which shows the efficiency of the proposed method.

2) Effectiveness of NGIM: Captured at the same scene, the different bands of a hyperspectral image look quite similar and contain much spatial redundant information that can be used to recover the missing details. However, existing methods do not make full use of this unique property of hyperspectral images. In this article, we propose an NGIM to exploit the redundant information contained in neighboring image groups to assist in the reconstruction of the target hyperspectral image

group. In order to demonstrate the effect, we remove the NGIM and add one deconvolutional layer at the end of the SSELM to reconstruct HR hyperspectral images group-by-group. As shown in Table VI, "Ours w/o NGIM," where NGIM is removed, shows a significant decrease in all evaluation indices compared to the original model "Ours." This is mainly because simply removing NGIM and super-resolving images in a group-by-group manner not only ignores the correlations among bands but also neglects the rich information redundancy in hyperspectral data.

G. Study on the Variant of NGIM

1) NGIM Versus Concatenate Intermediate Results: As shown in the middle of Table VII, in advanced group-based hyperspectral super-resolution methods [21], [22], [23], the intermediate features or images of each group are usually concatenated together, and then, a trunk network is adopted to conduct deep feature extraction and reconstruction on all
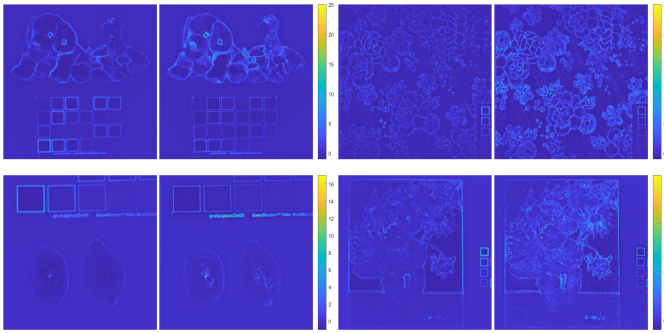
Fig. 13. Feature maps of four scenes from the CAVE dataset in the proposed NGIM. Each image pair consists of (Left) feature map of the target group and (Right) feature map obtained by integrating neighboring image groups.

Fig. 14. Parameters and FLOP studies on the CAVE dataset.

bands simultaneously. Although the trunk network can explore the correlations among bands to some extent, it places great demand on the network size. However, the limited amount of hyperspectral data constrains the size of the trunk network, resulting in insufficient LR-HR mapping learning. Moreover, simply concatenating intermediate results fails to utilize the rich redundant information in different bands, failing in recovering adequate details. To verify the superiority of the proposed NGIM, we recover the channels of the intermediate results to that of the input image groups and then concatenate these results together. As same as the network sketch of "Ours with TRUNK" in Table VII, we replace the NGIM with a trunk network composed of three cascaded SSELBs and one deconvolutional layer to reconstruct the whole image simultaneously. For the sake of fairness, the newly constructed network has a similar number of parameters to the proposed GELIN. As shown at the bottom of Table VII, the variant "Ours with TRUNK" lags behind in all evaluation indices, especially for spatial reconstruction confidence (i.e., PSNR and SSIM), which shows that our method would recover more missing details and obtain higher spatial fidelity.

2) Redundant Information Matters: In the proposed NGIM, we utilize the spatial redundancy contained in neighboring image groups to complement the missing details in the target image group. In order to verify that redundant information indeed contributes to the improvement of reconstruction performance, as shown in the "Our with PAD" of Table VII, we replace the deep embedding of neighboring image groups in NGIM, i.e., f_{LR}^{g-1} and f_{LR}^{g+1}, with zero padding of the same size. After replacement, the variant "Ours with PAD" suffers a notable performance drop (0.25 dB for PSNR). Lacking complementary information, the network is not able to fully utilize the high-similarity characteristic of hyperspectral data to assist the reconstruction process. Furthermore, we visualize the feature maps in NGIM to investigate whether neighboring image groups bring some complementary information to the network. The feature maps are the summation of the output in the channel dimension. As shown in Fig. 13, we demonstrate four different pairs of images in the CAVE dataset. In each image pair, the left one represents the feature map of the target group, while the right one represents the feature map obtained by integrating neighboring image groups. It can be observed
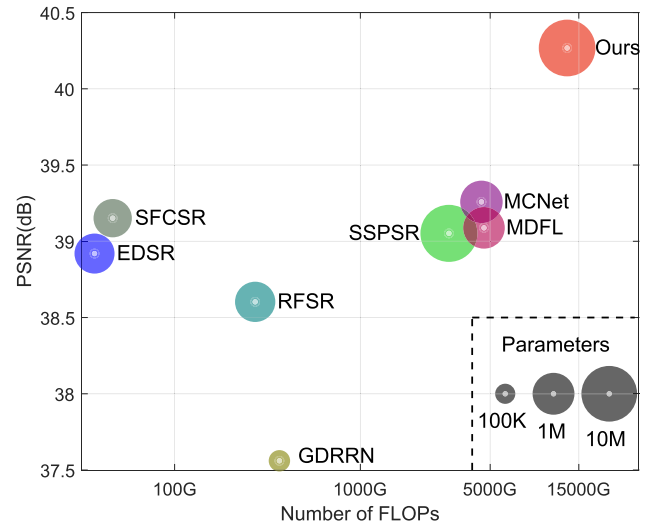
from the visualization results that the right image is more prominent than the left one in some areas, such as the head of the bird toy in the first pair of images. This investigation demonstrates that we can obtain sharper representations and finer grain of some areas by integrating the information of neighboring image groups. Through the proposed NGIM, the rewarding information is further exploited to complement the missing details of the target image group, leading to effective high-frequency detail recovery.

H. Model Parameters and FLOPs

In Fig. 14, we demonstrate the reconstruction performance versus model parameters and floating-point operations (FLOPs) on the CAVE dataset for scale factor 4. It can be noticed that the proposed GELIN obtains the best performance with fewer parameters than the state-of-the-art method SSPSR [21]. The FLOPs of all methods are calculated on the input with the size of $128 \times 128 \times 31$. It is worth mentioning that although the model size of GDRRN [19] is quite small, its FLOPs are not. This is mainly because GDRRN [19] conducts super-resolution in HR space, which greatly increases the computational cost. As can be noticed from the results, the FLOPs of our methods are quite large. This is mainly because the inference process of each image group is running in parallel, which would increase the computational complexity. However, the group-by-group reconstruction manner also contributes to the reconstruction performance, leading to significant improvement on evaluation metrics.

V. DISCUSSION

The superior performance demonstrates that the proposed GELIN could effectively utilize the high-similarity characteristic of hyperspectral data. When super-resolving the target image group, the specifically designed NGIM helps explicitly represent and exploit the complementary information contained in neighboring groups to recover more high-frequency details. We would like to point out that although the group

strategy has been widely adopted in hyperspectral image super-resolution, these methods still reconstruct the whole image at once. On the contrary, the proposed GELIN adopts a group-by-group super-resolution manner, preventing the network from modeling all spectral bands simultaneously. In this way, the difficulty of deep feature extraction and final reconstruction is alleviated.

Similar to many previous works [18], [19], [21], [22], [23], [46], in our experiments, the LR image inputs are generated through conducting bicubic interpolation on original HR images. However, it has been pointed by previous work [50] that an HR image can degrade to many LR images, while the bicubic degradation model only corresponds to one specific situation. The mismatch of the degradation model will deteriorate the final super-resolution results to some extent. It is the limitation of the current single hyperspectral image super-resolution methods that only the bicubic degradation model is considered.

For hyperspectral image super-resolution, a fundamental limitation is that a model trained on one hyperspectral dataset cannot be applied to another dataset. Different hyperspectral datasets are collected by different cameras, which capture images with different spectral response functions and band numbers. As a result, when super-resolving images different hyperspectral datasets, we need to adjust the input-output settings accordingly and retrain the network.

VI. CONCLUSION

In this article, we present a novel hyperspectral image super-resolution network to reconstruct HR images in a group-by-group manner, which mitigates the side effects introduced by the high dimensionality of hyperspectral data and alleviates the difficulty of feature extraction and reconstruction. In the proposed method, we design a spatial–spectral embedding learning network to exploit spatial and spectral information simultaneously, which not only obtains high spatial fidelity but also alleviates spectral distortion. Moreover, an NGIM is proposed to utilize the redundant information contained in neighboring image groups to complement the missing details in the target image group. We have conducted sufficient ablation studies to demonstrate the effectiveness of each proposed module. Extensive experiments on both natural and remote sensing hyperspectral datasets demonstrate the superiority of the proposed method. GELIN not only achieves the best performance in various commonly used image evaluation indices but also generates perceptually satisfactory HR hyperspectral images with more high-frequency details when compared with state-of-the-art algorithms.

REFERENCES

[1] R. O. Green et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, 1998.

[2] F. F. Sabins, "Remote sensing for mineral exploration," *Ore Geol. Rev.*, vol. 14, nos. 3–4, pp. 157–183, Sep. 1999.

[3] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 010901.

[4] A. Lowe, N. Harrison, and A. P. French, "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress," *Plant Methods*, vol. 13, no. 1, p. 80, Dec. 2017.

[5] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[6] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3631–3640.

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

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

[9] 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. (CVPR)*, Jun. 2019, pp. 1585–1594.

[10] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[11] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual Dirichletnet," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

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

[13] H. Irmak, G. B. Akar, and S. E. Yuksel, "A MAP-based approach for hyperspectral imagery super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2942–2951, Jun. 2018.

[14] H. Huang, J. Yu, and W. Sun, "Super-resolution mapping via multi-dictionary based sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3523–3527.

[15] 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.

[16] Q. Wang, Q. Li, and X. Li, "Hyperspectral image superresolution using spectrum and feature context," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11276–11285, Nov. 2020.

[17] 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, p. 1139, 2017.

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

[19] 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 (BigMM)*, Sep. 2018, pp. 1–4.

[20] J. Hu, Y. Li, and W. Xie, "Hyperspectral image super-resolution by spectral difference learning and spatial error correction," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1825–1829, Oct. 2017.

[21] 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.

[22] X. Wang, J. Ma, and J. Jiang, "Hyperspectral image super-resolution via recurrent feedback embedding and spatial–spectral consistency regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5503113.

[23] 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.

[24] Y. Zhang, K. Liu, Y. Dong, K. Wu, and X. Hu, "Semisupervised classification based on SLIC segmentation for hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1440–1444, Aug. 2019.

[25] F. Luo, Z. Zou, J. Liu, and Z. Lin, "Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[26] Y. Zhang, Y. Dong, K. Wu, and T. Chen, "Hyperspectral anomaly detection with Otsu-based isolation forest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9079–9088, 2021.

[27] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10096–10105.

[28] 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. 2015.

[29] 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. (CVPR)*, Jun. 2016, pp. 1646–1654.

[30] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[31] 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 (CVPRW)*, Jul. 2017, pp. 136–144.

[32] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3147–3155.

[33] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[34] X. Wang, J. Ma, and J. Jiang, "Contrastive learning for blind super-resolution via a distortion-specific network," *IEEE/CAA J. Autom. Sinica*, early access, Sep. 6, 2022, doi: 10.1109/JAS.2022.105914.

[35] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2017, pp. 4681–4690.

[36] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[37] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[38] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, Dec. 2018, pp. 286–301.

[39] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11065–11074.

[40] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.*, May 2020, pp. 191–207.

[41] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3517–3526.

[42] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14245–14254.

[43] T. Akgun, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of hyperspectral images," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1860–1875, Nov. 2005.

[44] J. Li, Q. Yuan, H. Shen, X. Meng, and L. Zhang, "Hyperspectral image super-resolution by spectral mixture analysis and spatial–spectral group sparsity," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1250–1254, Sep. 2016.

[45] W. Xie, X. Jia, Y. Li, and J. Lei, "Hyperspectral image super-resolution using deep feature matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6055–6067, Aug. 2019.

[46] Q. Li, Y. Yuan, and Q. Wang, "Hyperspectral image super-resolution via multi-domain feature learning," *Neurocomputing*, vol. 472, pp. 85–94, Feb. 2022.

[47] 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.

[48] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 19–34.

[49] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Laboratory, Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, May 2016.

[50] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.
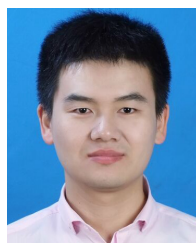
Xinya Wang received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. She is currently pursuing the Ph.D. degree with the Multi-Spectral Vision Processing Laboratory, Wuhan University.

Her research interests include neural networks, machine learning, and image processing.

Qian Hu received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2022, where he is currently pursuing the master's degree.

His research interests include computer vision and image processing.

Junjun Jiang (Senior Member, IEEE) received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014.

He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include image processing and computer vision.

Dr. Jiang won the Finalist of the World's FIRST 10 K Best Paper Award at ICME 2017 and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and the 2015 ACM Wuhan Doctoral Dissertation Award.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. He has authored or coauthored more than 200 refereed journal articles and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, International Journal of Computer Vision (IJCV), CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and pattern recognition.

Dr. Ma has been identified in the 2019–2021 Highly Cited Researcher lists from the Web of Science Group. He is also an Area Editor of *Information Fusion* and an Editorial Board Member of *Neurocomputing*.