Bidirectional 3D Quasi-Recurrent Neural Network for Hyperspectral Image Super-Resolution

Ying Fu, *Member, IEEE*, Zhiyuan Liang, and Shaodi You, *Member, IEEE*

Abstract—**Hyperspectral imaging is unable to acquire images with high resolution in both spatial and spectral dimensions yet, due to physical hardware limitations. It can only produce low spatial resolution images in most cases and thus hyperspectral image (HSI) spatial super-resolution is important. Recently, deep learning-based methods for HSI spatial super-resolution have been actively exploited. However, existing methods do not focus on structural spatial-spectral correlation and global correlation along spectra, which cannot fully exploit useful information for super-resolution. Also, some of the methods are straightforward extension of RGB super-resolution methods, which have fixed number of spectral channels and cannot be generally applied to hyperspectral images whose number of channels varies. Furthermore, unlike RGB images, existing HSI datasets are small and limit the performance of learning-based methods. In this article, we design a bidirectional 3D quasi-recurrent neural network for HSI super-resolution with arbitrary number of bands. Specifically, we introduce a core unit that contains a 3D convolutional module and a bidirectional quasi-recurrent pooling module to effectively extract structural spatial-spectral correlation and global correlation along spectra, respectively. By combining domain knowledge of HSI with a novel pretraining strategy, our method can be well generalized to remote sensing HSI datasets with limited number of training data. Extensive evaluations and comparisons on HSI super-resolution demonstrate improvements over state-of-the-art methods, in terms of both restoration accuracy and visual quality.**

Index Terms—**Bidirectional 3D quasi-recurrent neural network, global correlation along spectra, hyperspectral image super-resolution, structural spatial-spectral correlation.**

I. Introduction

HYPERSPECRAL image (HSI) is generally regarded as a data cube, which provides abundant spectral information, spatial information, and radiation information. In view of this advantage, HSIs are widely applied to agriculture [1], [2], environmental monitoring [3], [4], target detection [5], [6], and more. However, there exists a trade-off between the spatial resolution and spectral resolution physically. It usually maintains the high spectral resolution of HSI at the expense of spatial resolution. Thus, spatial super-resolution is essential for HSI, especially in remote sensing field [7]. To achieve this, the spatial-spectral correlation is usually exploited for HSI super-resolution [8]–[14].

Traditional methods based on sparse and dictionary learning [15]–[17] or low-rank approximation [18] are developed. The performance of these super-resolution methods is often determined by how well the prior knowledge of intrinsic characteristics of HSI is modeled. Besides, these methods are time-consuming, because they are usually formulated as an optimization problem and must be solved iteratively.

Recently, convolutional neural network (CNN)-based HSI super-resolution methods are presented [14], [19]–[22]. Yuan *et al.* [23] and Dong *et al.* [24] extend RGB-based super-resolution networks for HSI super-resolution, but these methods usually lack the capability to extract important information from different spectral channels, because they are mostly implemented with 2D CNNs. To explore both spatial context and spectral correlation, a 3-D full CNN framework (3D-FCNN) is introduced by Mei *et al.* [9]. However, it does not take the global correlation along spectra into consideration, which may limit the performance. Li *et al.* [20] propose a vanilla unidirectional recurrent structure to model HSIs. In this structure, the hidden state propagates unidirectionally and the hidden state only depends on the previous states, resulting in causal dependency which is unreasonable in nontemporal sequence modeling, e.g., HSI. In addition, most existing neural networks have fixed number of channels, and they cannot be used for HSIs with arbitrary number of spectral bands. Learning an effective network for single HSI super-resolution is also challenging, especially for remotely sensed HSI, due to the limited training samples [25]–[28].

In this article, we present a bidirectional 3D quasi-recurrent neural network (Bi-3DQRNN) for single HSI super-resolution, which can well exploit the domain knowledge of the HSI—structural spatial-spectral correlation and global correlation along spectra. And we utilize 3D convolution so that it is able to generalize to HSIs with various number of bands. We design a core unit as the basic building block of network, namely bidirectional 3D quasi-recurrent unit (Bi-3DQRU). This unit is composed of a 3D convolutional module and a bidirectional quasi-recurrent pooling module. The former one is utilized to extract structural spatial-spectral correlation of HSI, while the latter one is responsible for extracting global correlation along spectra. Besides, we design the quasi-recurrent pooling layer in Bi-3DQRU as a bidirectional structure to eliminate the causal dependency. In the bidirectional structure, each layer contains

Manuscript received August 31, 2020; revised November 25, 2020 and January 24, 2021; accepted January 30, 2021. Date of publication February 9, 2021; date of current version March 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61827901 and Grant 62088101. *(Corresponding author: Ying Fu.)*

Ying Fu and Zhiyuan Liang are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: fuying@bit.edu.cn; zhiyuan_liang@bit.edu.cn).

Shaodi You is with the Computer Vision Research Group, Insititute of Informatics, University of Amsterdam, 1000 Amsterdam, The Netherlands (e-mail: s.you@uva.nl).

Digital Object Identifier 10.1109/JSTARS.2021.3057936

two sublayers whose hidden states are propagated with in the inverse directions, and the results are generated by adding the outputs of sublayers. By combining domain knowledge of HSI with 3D deep learning, our Bi-3DQRNN can be capable and flexible.

Experiments are performed on a natural ICVL hyperspectral dataset [29] and five remotely sensed imagery datasets, including Pavia Centre [25], Pavia University [25], Salinas Valley [30], Indians Pines [31], and Urban [26]. The results on the ICVL dataset show that our method outperforms state-of-the-art extended RGB image super-resolution methods and single HSI super-resolution methods, indicating the capability of our model. In the experiments on remotely sensed datasets, we first divide the remote sensing imagery into training and testing regions, then train all super-resolution models on training set and make comparisons between the proposed method and competing methods. Next, we adopt a novel pretraining strategy to boost the performance of methods based on 3D deep learning. Specifically, we apply the model that pretraining on ICVL dataset to real-world remote sensing images with various number of bands. Although HSIs in ICVL dataset are in the visible range for the natural scenes, which are quite different to the remotely sensed dataset. The model training on remote sensing HSI dataset performs worse than the pretrained model, since the number of training samples is extremely small. The quantitative and visual results turn out that our pretrained model performs better, and beats most of models that trained from scratch on remotely sensed imagery dataset. To illustrate the robustness of our proposed method, we fine-tune our pretrained model from the 31-band HSI data of ICVL dataset, and achieve better performance. These experiments on remotely sensed datasets indicate the flexibility of our model. Extensive comparisons on super-resolution accuracy and visual quality from HSI and remotely sensed datasets are discussed in Section IV.

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

1) We present a CNN-based single HSI super-resolution method, which can make full use of the structural spatial-spectral correlation and global correlation along spectra of the HSI.
2) We introduce a bidirectional structure embedded in each layer of our network to take both forward and backward spectral dependency of HSI into account.
3) We effectively solve the problem of insufficient remotely sensed HSI training data by introducing the HSI in the visible range for the natural scenes with a novel pretraining strategy, which shows the robustness and flexibility of our method on the remotely sensed HSI.

II. Related Work

HSI super-resolution aims to reconstruct high-resolution HSI from degraded low-resolution HSI. It can be divided into two categories, i.e., fusion-based HSI super-resolution and single HSI super-resolution. The former one asks for a higher spatial resolution auxiliary image, e.g., panchromatic [32], RGB [33], or multispectral image [34]. The latter one does not require additional information, and only utilizes the information in the low-resolution input, which is more practical and has attracted widespread interest in recent years. In this work, we mainly focus on the single HSI super-resolution task. In the following, we review the related work from two respects, including traditional and deep learning-based methods.

A. Traditional Methods

There are several methods based on dictionary learning and low-rank approximation for HSI super-resolution. Gou *et al.* [15] introduced nonlocal self-similarity and local kernel constraint regularization terms into the HSI optimization process, and proposed a nonlocal pairwise dictionary learning model with local and nonlocal priors. Han *et al.* [16] presented an alternative directional approach of multipliers to estimate sparse codes of the high-resolution HSI. To explore the relationship among the sparse coefficients, Tang *et al.* [17] learned a spectral dictionary that could estimate a suitable size, and introduced double ℓ_1 regularized sparse representation to obtain a high-resolution HSI. Wang *et al.* [18] regarded HSI super-resolution task as a noncovex optimization problem, and applied the noncovex tensor penalty and 3D total variation term to model the intrinsic characteristics of HSIs. To improve the performance, these hand-crafted approaches become more and more complex and cannot match the demand on running time. Besides, these methods need to design the priors carefully and they may not represent the data well. Recently, deep learning techniques are wildly explored in HSI super-resolution tasks, which can well parallel process data and avoid hand-crafted prior.

B. Deep Learning-Based Methods

Here, we introduce three kinds of deep learning-based methods; they are RGB image super-resolution methods, HSI super-resolution methods with 2D convolution, and HSI super-resolution methods with 3D convolution.

We know RGB image super-resolution methods based on deep learning can be utilized to HSIs by resetting the input number of bands. We first introduce some popular RGB image deep learning-based super-resolution approaches. Kim *et al.* [35] proposed a very deep convolutional network (VDSR) with residual-learning and gradient clipping, which allowed adding depth to the network. Lim *et al.* [36] developed an enhanced deep super-resolution network (EDSR) by removing batch normalization layer, and won the NTIRE2017 super-resolution challenge. In order to achieve different upscaling factors in a single model, they also proposed multiscale deep super-resolution (MDSR). To make networks be deeper and wider, Zhang *et al.* [37] proposed a very deep residual channel attention networks (RCAN) combining residual-learning and channel attention mechanism.

In recent years, deep learning techniques for HSI super-resolution have been widely explored and achieved remarkable success. Li *et al.* [38] proposed a deep spectral difference convolutional neural network (SDCNN) with spatial constraint strategy, using five convolutional layers to conduct HSI super-resolution task. He and Liu [39] presented a deep Laplacian pyramid network (LPN), which progressively reconstructed the high spatial resolution HSIs in a coarse-to-fine way with multiple pyramid levels. These methods mainly focus on the spectral dimension. Jiang *et al.* [11] made a step forward by investigating how to adapt state-of-the-art residual learning-based single gray/RGB image super-resolution approaches for computationally efficient single HSI super-resolution, and proposed a spatial-spectral prior network (SSPSR) to exploit the spatial information and the correlation between the spectra of the HSI. Xie *et al.* [40] proposed a deep feature matrix factorization method by applying deep neural network and coupled nonnegative matrix factorization on the key bands in each subset of HSI, to generate super-resolved HSI. Hu *et al.* [14] introduced an intrafusion network (IFN) for HSI super-resolution. It first calculates spectral difference between two adjacent bands [38], and then feeds the two adjacent bands and their spectral difference band to a super-resolution network namely SRCNN [41]. Finally, the outputs from SRCNN are combined in a pixel-wise manner.

Since 3D convolution can extract spatial and spectral information simultaneously, more research has begun to utilize 3D convolution for single HSI super-resolution. Mei *et al.* [9] proposed 3D full convolution neural network (3D-FCNN) to exploit both the spatial dimension information of neighboring pixels and spectral dimension information of neighboring bands. Yang *et al.* [12] adopted wavelet decomposition to capture textures and structures in HSI, and proposed wavelet 3D CNN for multi super-resolution scales. Wang *et al.* [42] presented a three-branch frequency-separated 3D CNN, which inhibited spectral distortion. Li *et al.* [13] held a view that it is unsuitable to pay more attention to the mining of HSI spatial information, when the spectral information is not sufficiently exploited. Therefore, they proposed a mixed convolutional network (MCNet) for HSI super-resolution. Hu *et al.* [22] designed a multiple feature fusion and aggregation network with 3D convolution (MFFA-3D) equipped with multiscale connections and two-step multiscale strategy to obtain the high-resolution HSI. These 3D deep learning-based methods have achieved effective reconstruction results. These works mainly focus on exploiting the adjacent bands, and do not well explore the global correlation along spectra of HSI.

In this work, we consider the structural spatial-spectral correlation and global correlation along spectra of HSI, and model them in bidirectional 3D quasi-recurrent neural network, which makes our approach both capable and robust.

III. Bidirectional 3D Quasi-Recurrent Neural Network

In this section, we first briefly review the problem formulation and our motivation. Then, we introduce the overall architecture of proposed method, i.e., Bi-3DQRNN, which can well extract structural spatial-spectral correlation and global correlation along spectra. Finally, the basic building block of Bi-3DQRNN is illustrated in detail, and we describe the bidirectional structure embedded in this building block that can well eliminate causal dependency.

A. Motivation and Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{rH \times rW \times C}$ and $\mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ denote the ground truth HSI and low-resolution HSI input, where r, H, W, C indicate the scale factor of super-resolution, spatial height, spatial width, and number of spectral bands, respectively. In general, the low-resolution HSI $\mathbf{Y}$ can be modeled as a degradation model on the original high-resolution HSI $\mathbf{X}$ like

$$\mathbf{Y} = D\left(\mathbf{X}; \sigma\right) \tag{1}$$

where the nonlinear mapping D is represented as the degradation manipulation, and σ denotes the parameters in the degradation model.

Single HSI super-resolution problem based on supervised learning can be described as

$$\hat{\mathbf{X}} = \mathbf{G}\left(\mathbf{Y}; \theta\right) \tag{2}$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{rH \times rW \times C}$ is the predicted high-resolution HSI. $\mathbf{G}$ denotes the super-resolution model, and θ denotes corresponding parameters to be learned.

To learn the model in (2), the objective function can be expressed as

$$\hat{\theta} = \min_{\theta} L\left(\hat{\mathbf{X}}, \mathbf{X}\right) + \lambda \Phi\left(\theta\right) \tag{3}$$

where L represents loss function, λ denotes penalty coefficient, and $\Phi(\theta)$ indicates regularization terms. Learning the super-resolution model is equivalent to find the right parameter θ minimizing the loss function L.

Previous works [9], [12], [22], [42] have modeled HSIs with 3D CNN, which significantly improve the quality of reconstructed HSIs and make model itself more flexible on the number of bands. Besides, 3D CNN can well model the spatial-spectral correlation. Accordingly, we design a 3D CNN-based single HSI super-resolution model, which can effectively extract the structural spatial-spectral correlation and global correlation along spectra of an HSI with arbitrary number of bands.

Besides, some works [11], [14], [38], [39] have attempted to exploit spatial-spectral prior, but still pay little attention to the global correlation along spectra of HSI. Thus, we equip 3D convolutions with quasi-recurrent pooling function to exploit the global correlation along spectra. In addition, we insert a bidirectional structure into the core unit of our network to overcome the problem of causal dependency.

The global information in spectral dimension can be extracted effectively by dynamically merging the previous states [43]. However, the vanilla recurrent structure only allows hidden states propagate forward, resulting in causal dependency. The alternative directional structure in [44] propagates hidden states forward or backward each layer, which introduces the bias due to the asymmetrical structure.

B. Network Architecture

The overall architecture of our bidirectional 3D quasi-recurrent neural network is shown in Fig. 1. We adopt preupsampling framework to learn the end-to-end relationship between

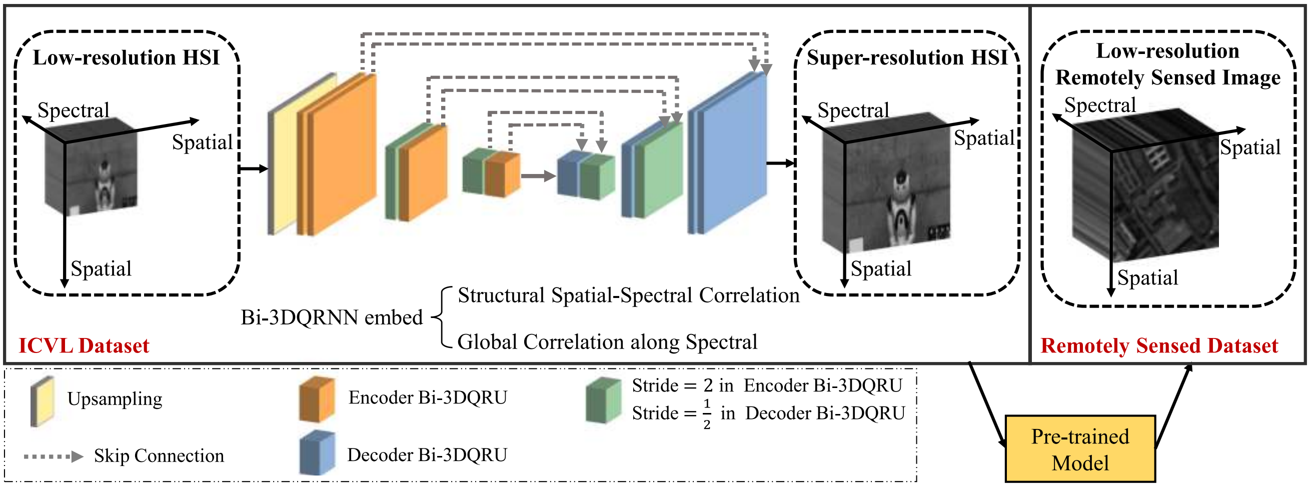


Fig. 1. Overall architecture of our method. The network is a residual encoder–decoder framework, with six pairs of Bi-3DQRUs. The bidirectional structure is equipped in all Bi-3DQRUs, and skip connections are added in each pair. We use “Stride $= \frac{1}{2}$” to represent the combination of a trilinear interpolation and a convolution, which is similar to transpose convolution from the perspective of effect. Our pretrained model on ICVL dataset can be applied to remotely sensed dataset directly, which performs better than competing methods. The details are shown in Section IV.

TABLE I
NETWORK CONFIGURATION OF OUR BI-3DQRNN FOR HSI SUPER-RESOLUTION

	Layer name	Feature extractor	Output size
Encoder	conv3d$_1$	$3 \times 3 \times 3, 16$	$H \times W \times C$
	conv3d$_2$	$3 \times 3 \times 3, 16$	$H \times W \times C$
	conv3d$_3$	$3 \times 3 \times 3, 32$	$\frac{H}{2} \times \frac{W}{2} \times C$
	conv3d$_4$	$3 \times 3 \times 3, 32$	$\frac{H}{2} \times \frac{W}{2} \times C$
	conv3d$_5$	$3 \times 3 \times 3, 64$	$\frac{H}{4} \times \frac{W}{4} \times C$
	conv3d$_6$	$3 \times 3 \times 3, 64$	$\frac{H}{4} \times \frac{W}{4} \times C$
Decoder	convTrans3d$_1$	$3 \times 3 \times 3, 64$	$\frac{H}{4} \times \frac{W}{4} \times C$
	upsample$_1$	$3 \times 3 \times 3, 32$	$\frac{H}{2} \times \frac{W}{2} \times C$
	conv3d$_7$	$3 \times 3 \times 3, 32$	$\frac{H}{2} \times \frac{W}{2} \times C$
	convTrans3d$_2$	$3 \times 3 \times 3, 32$	$\frac{H}{2} \times \frac{W}{2} \times C$
	upsample$_2$	$3 \times 3 \times 3, 16$	$H \times W \times C$
	conv3d$_8$	$3 \times 3 \times 3, 16$	$H \times W \times C$
	convTrans3d$_3$	$3 \times 3 \times 3, 16$	$H \times W \times C$
	convTrans3d$_4$	$3 \times 3 \times 3, 1$	$H \times W \times C$

We refer to ‘$3 \times 3 \times 3, 16$’ as 3×3 Kernel size and 16 output feature maps.

low-resolution HSI and high-resolution one. After an upsampling layer, the low-resolution HSI is fed into the proposed network and the corresponding super-resolved HSI is obtained. Besides, we apply the model that pretrained on ICVL dataset to remotely sensed imagery dataset to show the super-resolution effect of introducing the natural HSI data.

The proposed network Bi-3DQRNN is a bidirectional residual encoder–decoder with 12 layers. Each layer is a convolutional or deconvolutional Bi-3DQRU, which is symmetric. Table I illustrates the detailed network configuration. In the encoder part, the strides are set to 2 in the second and fourth layers to half the spatial size of output feature maps, and they are symmetrically set in the decoder part. It is worth noticing that we replace commonly used transpose convolution with a combination of trilinear interpolation and convolution to double the spatial size of output feature maps. Performing downsample and upsample can help us expand the receptive field, which can well exploit the context information in larger image region.

C. Bidirectional 3D Quasi-Recurrent Unit

The Bi-3DQRU is the basic building block of our network, as shown in Fig. 2. It consists of two basic components, a 3D convolution module and a bidirectional quasi-recurrent pooling module.

3D Convolutional Module: We perform two dependency 3D convolutions on the feature maps that come from previous layer, and then pass the results through different activation functions to generate an intermediate tensor $\mathbf{M}$ and a gate tensor $\mathbf{G}$. This process can be formally described as

$$\begin{aligned} \mathbf{M} &= \tanh\left(\mathbf{W}_m * \mathbf{I}\right) \\ \mathbf{G} &= \varphi\left(\mathbf{W}_g * \mathbf{I}\right) \end{aligned} \tag{4}$$

where $\mathbf{I} \in \mathbb{R}^{C_{in} \times rH \times rW \times C}$ is the input feature map from previous layer. $\mathbf{M}$ and $\mathbf{G} \in \mathbb{R}^{C_{out} \times rH \times rW \times C}$ are intermediate tensor and gate tensor, respectively. $\mathbf{W}_m$ and $\mathbf{W}_g \in \mathbb{R}^{C_{out} \times C_{in} \times 3 \times 3 \times 3}$ are both 3D convolutional filter kernels, and $*$ denotes the 3D convolutional operator. In our 3D convolutional module, φ means the sigmoid activation function.

The 3D convolutional module takes advantage of its 3D kernel to not only extract the information in spatial domain like a 2D convolution, but also extract the information in adjacent spectra. Hence, the 3D convolutional module is able to exploit the structural correlation of HSIs. Besides, 3D convolution makes Bi-3DQRNN able to handle HSIs in any number of bands.

Bidirectional Quasi-Recurrent Pooling Module: We apply the bidirectional quasi-recurrent pooling after 3D convolutional module to exploit the global correlation along spectra in HSIs. The visualization of the model is in Fig. 2.

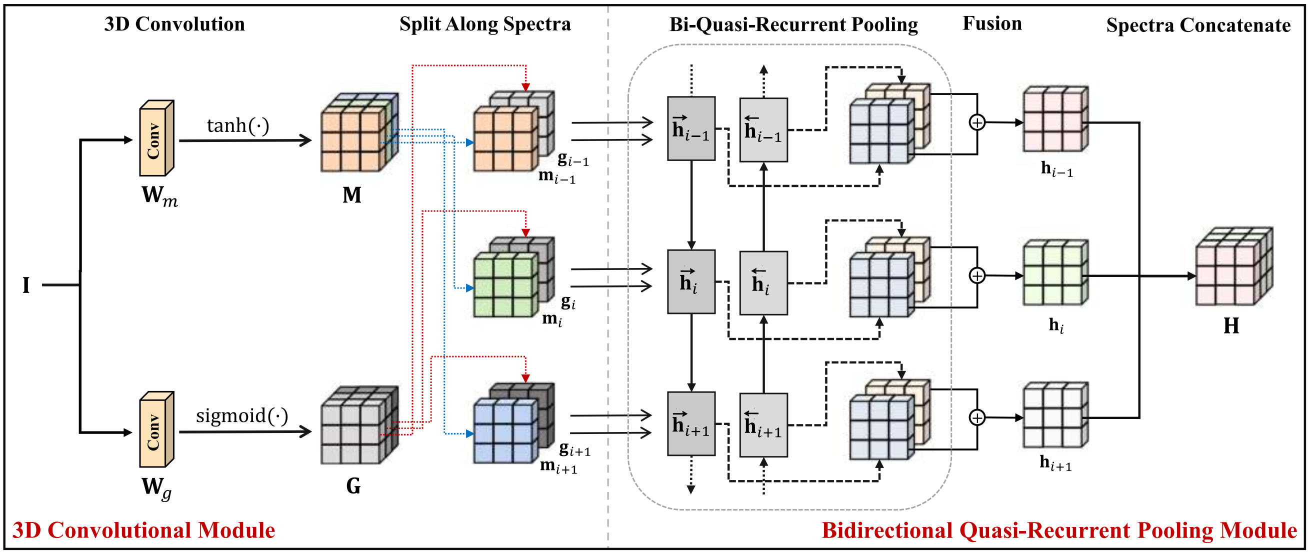


Fig. 2. Architecture of Bi-3DQRU. The full computational graph can be described in five steps. First, two 3D convolutions with different activation functions are performed on the input feature map $\mathbf{I}$, which produce an intermediate tensor $\mathbf{M}$ and a gate tensor $\mathbf{G}$. Second, tensors $\mathbf{M}$ and $\mathbf{G}$ are split along spectra, obtaining two sequences $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_C\}$ and $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_C\}$, respectively. Third, a bidirectional quasi-recurrent pooling (bi-quasi-recurrent pooling) function is applied on the intermediate tensor $\mathbf{m}_i$ that is controlled by the gate tensor $\mathbf{g}_i$ (the exact working mechanism of this bidirectional structure is illustrated in Fig. 3). Fourth, each pair of forward and backward hidden states in the previous recurrent network are combined with an element-wise addition, generating a new hidden state $\mathbf{h}_i$. Finally, all the spatial planes along spectra are concatenated into output feature map $\mathbf{H}$.

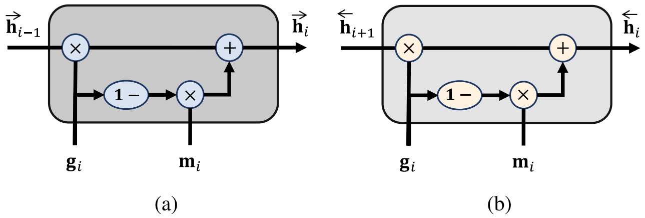

$\overrightarrow{\mathbf{h}}_{i-1}$
×
+
$\overrightarrow{\mathbf{h}}_i$
1 −
×
$\mathbf{g}_i$
$\mathbf{m}_i$
(a)
$\overleftarrow{\mathbf{h}}_{i+1}$
×
+
$\overleftarrow{\mathbf{h}}_i$
1 −
×
$\mathbf{g}_i$
$\mathbf{m}_i$
(b)

Fig. 3. Details of our bidirectional qausi-recurrent pooling structure which is embedded in Bi-3DQRU. Take the forward case for example, $\overrightarrow{\mathbf{h}}_i$ is obtained by its previous hidden state $\overrightarrow{\mathbf{h}}_{i-1}$, intermediate tensor $\mathbf{m}_i$, and gate tensor $\mathbf{g}_i$. (a) Forward quasi-recurrent pooling. (b) Backward quasi-recurrent pooling.

In this module, we combine pooling operation with dynamic gating mechanism in a bidirectional structure. First, we split the intermediate tensor $\mathbf{M}$ and gate tensor $\mathbf{G}$ along spectra to produce sequences $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_C\}$ and $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_C\}$ at the end of last module. Second, these states $\mathbf{m}_i$ and $\mathbf{g}_i$ are fed into the dynamic gating function, generating the forward hidden states $\overrightarrow{\mathbf{h}}_i$ and backward hidden states $\overleftarrow{\mathbf{h}}_i$ as follows:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_i &= \mathbf{g}_i \odot \overrightarrow{\mathbf{h}}_{i-1} + (1-\mathbf{g}_i) \odot \mathbf{m}_i, \ \forall i \in [1, C]\\ \overleftarrow{\mathbf{h}}_i &= \mathbf{g}_i \odot \overleftarrow{\mathbf{h}}_{i+1} + (1-\mathbf{g}_i) \odot \mathbf{m}_i, \ \forall i \in [1, C]\end{aligned} \tag{5}$$

where $\overrightarrow{\mathbf{h}}_{i-1}$ is the hidden state that merges all the previous states, and also means the output of the $(i-1)$th spectral band, and we set $\overrightarrow{\mathbf{h}}_0 = 0$. As for the backward propagation, $\overleftarrow{\mathbf{h}}_{i+1}$ is the hidden state that merges all the posterior states, and we set $\overleftarrow{\mathbf{h}}_{C+1} = 0$. $\odot$ denotes an element-wise multiplication.

Third, the forward and backward hidden states are combined, producing the hidden state $\mathbf{h}_i$, which is formulated as

$$\mathbf{h}_i = \overrightarrow{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i \, . \tag{6}$$

Finally, the output of Bi-3DQRU is obtained by concatenating the hidden states along spectral dimension. In the following, some details about bidirectional structure are explored. The bidirectional structure is introduced to eliminate the causal dependency caused by unidirectional 3DQRU, and there is other way to achieve the same effect, such as the alternative structure introduced by Wei *et al.* [44]. However, the alternative structure only exploits forward or backward information along spectra in each layer, which limits its performance. The ablation study in Section IV-D verifies the capability of our bidirectional structure.

The gate tensor $\mathbf{g}_i$ is utilized to control the weights of the previous memory $\overrightarrow{\mathbf{h}}_{i-1}$ (or posterior memory $\overleftarrow{\mathbf{h}}_{i+1}$) and the current intermediate tensor $\mathbf{m}_i$. It is worth noticing that the value of $\mathbf{g}_i$ only depends on the current input feature maps $\mathbf{I}$, which makes the gate tensor tend to learn more from the input image itself, rather than the parameters learned during training.

In summary, our Bi-3DQRU takes the advantage of 3D kernels in the 3D convolutional module to correlate the structural information of HSIs not only in spatial domain but also in adjacent spectra. By recurrently weighting and merging the intermediate tensor from 3D convolutional module, with the help of bidirectional quasi-recurrent pooling module, the global correlation along spectra of HSI can be effectively exploited as well.

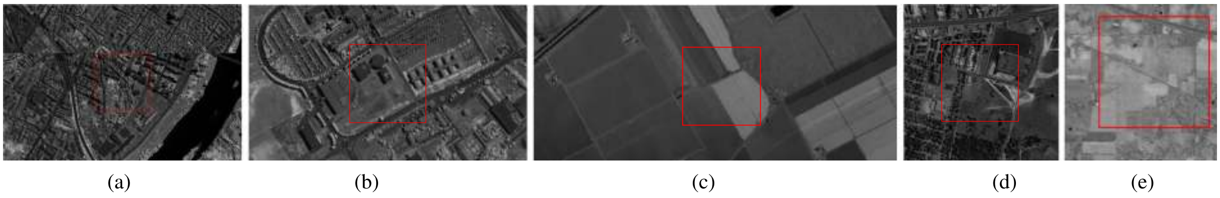


Fig. 4. Visual examples of the remotely sensed datasets. In each remotely sensed imagery, we use a red frame to box the testing region, while the other region is for training. (a) Pavia Centre. (b) Pavia Univeristy. (c) Salinas Valley. (d) Urban. (e) Indian Pines.

IV. Experiments

We conduct a series of experiments on a natural HSI dataset and five remotely sensed HSI datasets to evaluate the performance of our network. In this section, we first introduce the experimental datasets and evaluation metrics. Then, the implementation details and competing methods are listed. After that, we provide the quantitative and qualitative results on all datasets. Finally, the ablation studies are performed to analyze the proposed modules.

A. Datasets and Settings

Datasets: We conduct experiments on two types of datasets, including natural HSIs (i.e., ICVL dataset) and remotely sensed images (i.e., Pavia Centre, Pavia University, Salinas Valley, Urban, and Indian Pines). The remotely sensed images are illustrated in Fig. 4.

1) *ICVL:* It consists of 201 HSIs collected by a Specim PS Kappa DX4 hyperspectral camera at 1392 × 1300 spatial resolution. It has 31 spectral bands in wavelength ranging from 400 to 700 nm at 10 nm intervals.
2) *Pavia Centre:* It is a scene from Pavia, Northern Italy, acquired by the ROSIS sensor during a flight campaign. The spatial resolution of Pavia Centre is 1096 × 715, and the number of spectral bands is 102. The geometric resolution is 1.3 m.
3) *Pavia University.* Pavia University is similar to Pavia Centre, which is also acquired by the ROSIS sensor over Pavia. Its spatial resolution is 610 × 340, with 103 bands, and its geometric resolution is 1.3 m.
4) *Salinas Valley:* This scene is gathered by the AVIRIS sensor from Salinas Valley, California. The spatial resolution is 512 × 217, with 224 spectral bands.
5) *Urban:* The spatial resolution is 307 × 307, with 201 bands in the wavelength from 400 to 2500 nm, each of which corresponds to a 2 × 2 m^2 area. Some bands are seriously polluted because of the dense water vapor and atmospheric effects, there are 162 channels remained.
6) *Indian Pines:* This scene is acquired by AVIRIS sensor over the Indian Pines test site, North-western Indiana. The spatial resolution is 145 × 145, and the number of spectral band is 224, whose wavelength is ranged from 400 to 2500 nm.

Evaluation Metrics: Two sets of quantitative quality metrics are adopted, where PSNR and SSIM [45] are used to evaluate spatial fidelity, and SAM [46] is employed to measure the spectral similarity. It is worth mentioning that we calculate PSNR and SSIM in a band-wise manner, i.e., calculating PSNR and SSIM band by band for each HSI and averaging them all afterwards. Larger values of PSNR and SSIM suggest better performance, while a smaller value of SAM implies better performance.

Network Learning: Our Bi-3DQRNN is learned by minimizing the mean square error (MSE) between the predicted high-resolution HSI $\hat{\mathbf{X}}$ and the ground truth $\mathbf{X}$ during the training phase. Adam optimizer [47] is adopted. The learning rate is initialized as 1×10^{-4}, and gradually decays tothe minimum 5×10^{-5}. Our method is implemented by the deep learning framework PyTorch with NVIDIA GTX 1080Ti GPU.

Degradation Model: Regarding the HSI from the above six datasets as ground truth, we use a 8 × 8 Gaussian filter $(\sigma = 3)$ to smooth each band of HSI, and then downsample each band by a scale factor with bicubic interpolation [48], like [49], to obtain the corresponding low-resolution HSIs.

Competing Methods: We adopt bicubic interpolation as the baseline and compare our method against both RGB image super-resolution methods and HSI super-resolution methods.

For the RGB image super-resolution methods, we compare our method with several state-of-the-art methods, including VDSR [35], band-wise VDSR [35], MDSR [36], and RCAN [37]. The VDSR framework applies its model to the luminance components. In our experiments, we extend VDSR/MDSR/RCAN to HSIs super-resolution task by setting the input channel equals to the number of spectral bands of HSI. To conduct band-wise VDSR (BWVDSR), we set the input channel equals to 1, so that the VDSR model processes the input features one band each time. MDSR is a multiscale EDSR [36], which can reconstruct high-resolution images of different upscaling factors in a single model.

For the HSI super-resolution methods, we compare with four state-of-the-art methods, including 3D-FCNN [9], MCNet [13], SSPSR [11], and IFN [14]. We carefully adjust the hyperparameters of these compared methods to achieve their best performance.

B. Experiments on Natural HSI Dataset

For the experiments on ICVL dataset, we randomly select 100 HSIs for training, and the rest are for testing. The training set is built from multiple overlapped cubes cropped from each HSI, and each cropped cube has a spatial size of 64 × 64 and a full spectral size of 31 for the purpose of preserving the

TABLE II
QUANTITATIVE EVALUATION OF COMPETING SUPER-RESOLUTION METHODS BY AVERAGE PSNR/SSIM/SAM FOR DIFFERENT SCALE FACTORS ON ICVL DATASET

Scale Factor	Metrics	Methods									
		Bicubic [48]	VDSR [35]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours
×2	PSNR ↑	34.923	41.729	39.675	42.032	42.280	40.456	43.530	43.612	37.984	**44.917**
	SSIM ↑	0.9628	0.9894	0.9859	0.9910	0.9918	0.9853	0.9926	0.9931	0.9771	**0.9932**
	SAM ↓	0.0678	0.0365	0.0648	0.0358	0.0331	0.0431	0.0261	0.0279	0.0543	**0.0252**
×4	PSNR ↑	34.573	38.957	36.980	39.029	39.182	38.911	37.443	39.095	36.283	**40.387**
	SSIM ↑	0.9592	0.9791	0.9745	0.9788	0.9782	0.9780	0.9745	0.9789	0.9695	**0.9817**
	SAM ↓	0.0682	0.0366	0.0668	0.0368	0.0362	0.0438	0.0347	0.0345	0.0527	**0.0294**
×8	PSNR ↑	31.720	32.405	32.218	32.261	32.421	33.275	32.867	33.401	31.980	**33.655**
	SSIM ↑	0.9277	0.9322	0.9344	0.9348	0.9350	0.9369	0.9101	0.9311	0.9283	**0.9394**
	SAM ↓	0.0741	0.0714	0.0719	0.0701	0.0671	0.0502	0.0584	0.0557	0.0723	**0.0381**

The bold indicates the best performance.

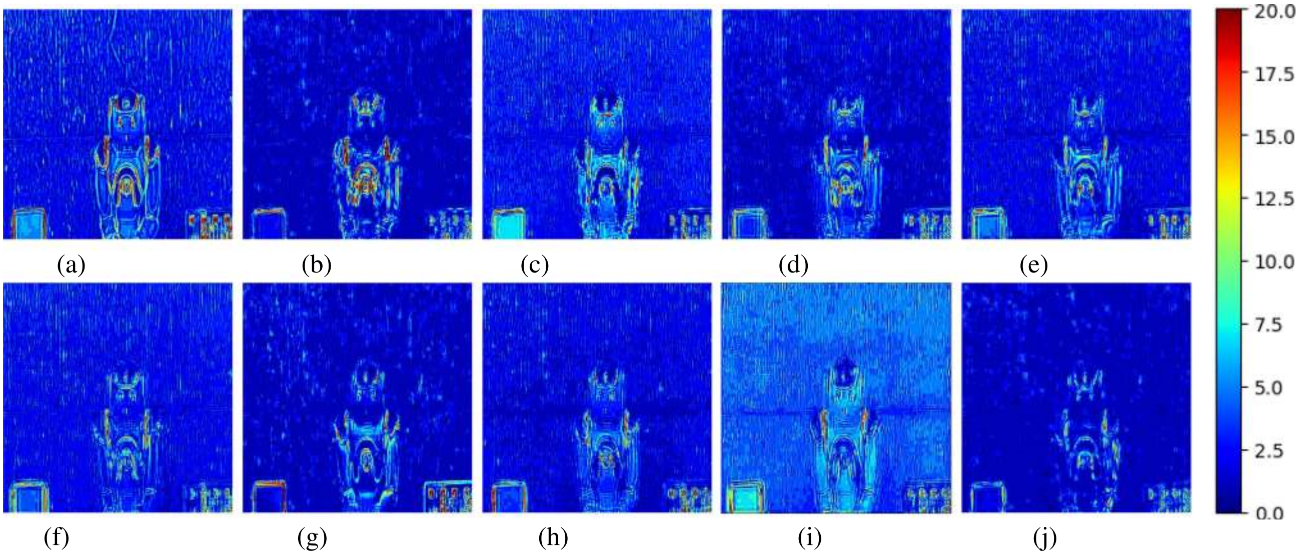
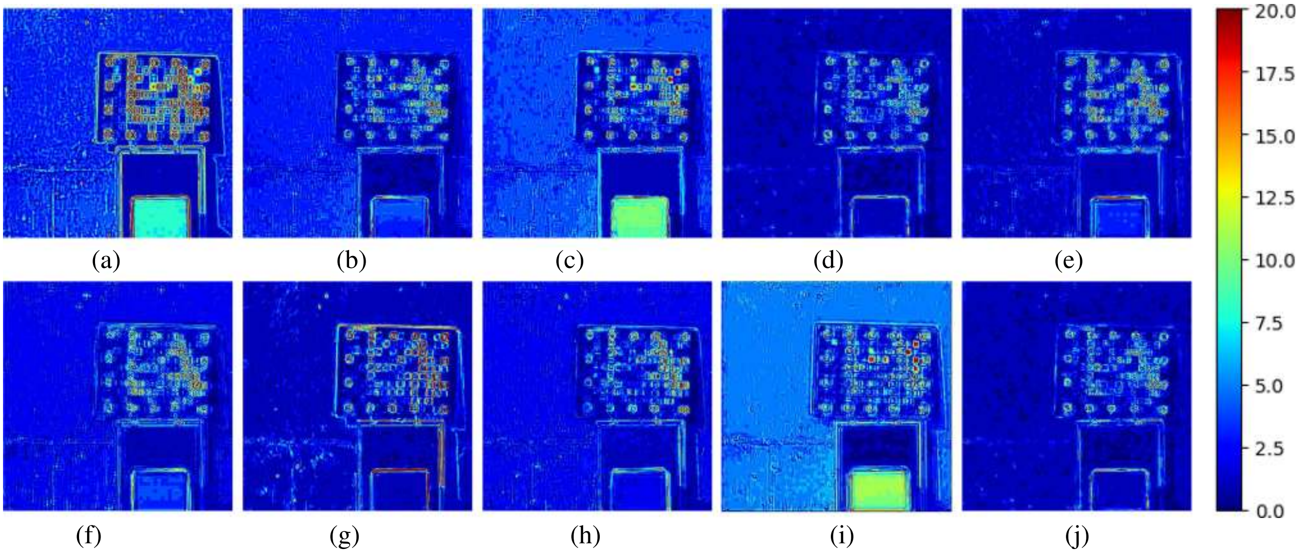

(a) (b) (c) (d) (e)
(f) (g) (h) (i) (j)

Fig. 5. Error maps at the 20th band of HSI with scale factor ×4 on ICVL dataset. (a) Bicubic. (b) VDSR. (c) BWVDSR. (d) MDSR. (e) RCAN. (f) 3D-FCNN. (g) MCNet. (h) SSPSR. (i) IFN. (j) Ours.

complete spectra. In addition, a set of data augmentation techniques, such as rotation and scaling, are employed to generate roughly 50 k training samples in total. As for testing, we crop the centre region of each image with size of $512 \times 512 \times 31$ like [50].

Table II summarizes the quantitative evaluation of state-of-the-art super-resolution algorithms by average PSNR, SSIM, and SAM for different scale factors. As shown in Table II, our approach achieves better results in comparison with other algorithms on the ICVL dataset. Specifically, bicubic interpolation is the simplest but the worst method. For the RGB image super-resolution algorithms, VDSR performs better than BWVDSR, indicating the exploration of spectral correlation benefits super-resolution performance. MDSR and RCAN provide better results than VDSR, which means effective network design helps improve the performance. Our method surpasses all the competing state-of-the-art RGB image methods in all ×2, ×4 and ×8 cases. For the HSI super-resolution competing algorithms, the performance of IFN is the worst in all super-resolve scale factor cases. 3D-FCNN performs poor in both ×2, ×4 cases. MCNet surpasses 3D-FCNN a lot in ×2 case, while the performance gets worse as the scale factor increases. SSPSR achieves slightly higher performance in comparison with MCNet in some cases. In ×8 case, the improvement of all methods is not insignificant, compared with bicubic. Our Bi-3DQRNN still achieves better performance compared with these competing methods.

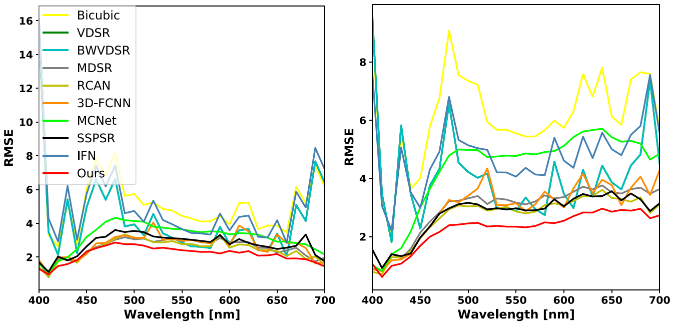
Figs. 5 and 6 show the images of ground truth and error maps at the 20th band on two HSIs with scale factor ×4. The error maps are the absolution errors between the ground truth and the restored results. We can see that the absolute error results obtained by our method are very low, and our method produces shallow edge or no edge in some areas. This result implies that our method can restore more information than the competing methods, which is consistent with the results in Table II. Besides, we provide the root mean square error (RMSE) results along



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 6. Error maps at the 20th band of HSI with scale factor ×4 on ICVL dataset. (a) Bicubic. (b) VDSR. (c) BWVDSR. (d) MDSR. (e) RCAN. (f) 3D-FCNN. (g) MCNet. (h) SSPSR. (i) IFN. (j) Ours.



(a) (b)

Fig. 7. RMSE results along spectra for all methods from the above HSIs with scale factor ×4 on ICVL dataset. (a) Corresponding to HSI in Fig. 5. (b) Corresponding to HSI in Fig. 6.

spectra for all methods in Fig. 7 corresponding to the HSIs in Figs. 5 and 6. The smaller value of RMSE implies more realistic visual results, and the RMSE curve shows that the results of our method are closer to the ground truth, which indicates embedding more domain knowledge helps achieve higher performance.

Furthermore, we make a comparison on FLOPs, number of parameters, and running time among all deep learning based methods. The quantitative results are listed in Table III and the relationships between running time and PSNR are presented in Fig. 8. As shown in Table III, the quantitative results of VDSR and BWVDSR are quite similar, and they have the smallest model sizes and the fastest running speeds, which is consistent with intuition. MDSR and 3D-FCNN algorithms have fewer FLOPs than Bi-3DQRNN, while others are far more complex than our method. As for running time, except for VDSR, BWVDSR, 3D-FCNN, and IFN, others are far more time-consuming than our approach. In conclusion, our network

TABLE III
QUANTITATIVE EVALUATION OF DEEP LEARNING-BASED SUPER-RESOLUTION METHODS BY FLOPS, NUMBER OF PARAMETERS (PARAMS), TIME, AND PSNR FOR ×2 SCALE FACTOR ON ICVL DATASET

	FLOPs	Params	Time (ms)	PSNR (dB)
VDSR	2.67G	669.26k	11.22	41.729
BWVDSR	2.54G	664.70k	11.70	39.675
MDSR	25.15G	6538.78k	110.60	42.032
RCAN	58.85G	15476.93k	476.71	42.280
3D-FCNN	26.59G	224.83k	23.23	40.456
MCNet	297.95G	1928.32k	431.05	43.530
SSPSR	192.35G	18454.76k	216.13	43.612
IFN	0.66G	57.28k	10.67	37.984
Ours	27.28G	1288.23k	57.83	44.917

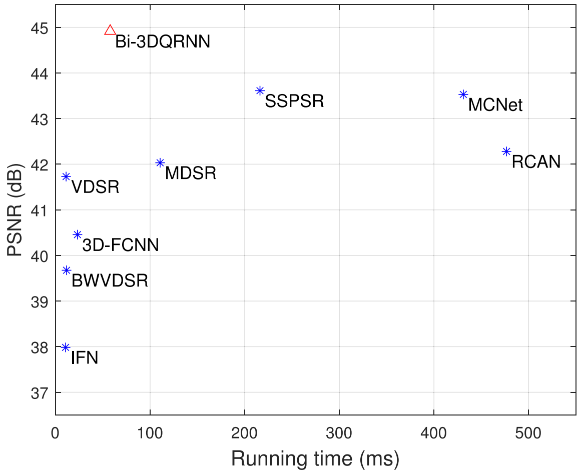


Fig. 8. PSNR vs. running time on ICVL dataset in ×2 scale factor.

TABLE IV
DETAILS OF TESTING REGION SIZE AND POSITION IN EACH REMOTELY SENSED IMAGERY

Dataset	Size	Position
Pavia Centre	256 × 256	(420, 230) & (675, 485)
Pavia University	256 × 256	(117, 42) & (432, 297)
Salinas Valley	128 × 128	(192, 45) & (319, 172)
Urban	120 × 120	(93, 93) & (212, 212)
Indian Pines	100 × 100	(33, 13) & (132, 112)

In the *Position* column, we use the two coordinates to represent the top left and bottom right vertexes, respectively.

is at the middle-level complexity but the best effect among these methods.

C. Experiments on Remotely Sensed Images

To verify the effectiveness of our Bi-3DQRNN, we design two types of experiments with different experimental settings.

Experimental Setting I: In this setting, we conduct ×2, ×3, and ×4 scale factors in Indian Pines, and ×2, ×4, and ×8 in others. The reason is that the spatial size of Indian Pines is too small (145 × 145), resulting in difficulties to conduct large super-resolution scales. Since there is only one image in each remotely sensed dataset, we crop a certain region of the image to obtain a subimage as the testing data, while the rest is for training.[1]

The detailed size and position of testing region are shown in Table IV. The position is represented by the coordinate of top-left and bottom-right vertexes in each remotely sensed imagery, which is corresponding to the red frame in Fig. 4. The training set is the region that is out of the red frame in each remotely sensed imagery.

There are three experiments on the five remotely sensed datasets. First, we train all methods from scratch in the remotely sensed training set. We use Ours-S to represent our approach that is trained from scratch. Second, we select competing methods that can handle HSIs with arbitrary number of bands, and then apply the models pretrained on ICVL dataset to remotely sensed datasets directly. We use -P to represent the models that are pretrained on ICVL dataset. Third, we fine-tune our model which is pretrained on ICVL dataset to verify the robustness of our method. It is denoted as Ours-F. These quantitative results are shown in Tables V–IX.

The results on Pavia Centre are shown in Table V. From the left part of Table V, we can see that our method achieves the best performance in all super-resolution scales. On the right of vertical line, we use -P and -F to represent the pretrained model on ICVL dataset and the fine-tuned model, respectively. Among all competing methods, only 3D-FCNN is able to handle HSIs with arbitrary number of bands, and we apply the models pretrained on ICVL dataset to remotely sensed datasets, denoted as 3D-FCNN-P. We can see that Ours-P is much better than 3D-FCNN-P. Besides, Ours-P is better than Ours-S. It means the result of training from scratch on Pavia Centre is worse than pretrained on ICVL dataset, which demonstrates our method can effectively overcome the problem of insufficient training data. Furthermore, by fine-tuning our Bi-3DQRNN model with small remotely sensed training data, our method can achieve a better performance. It can be noticed that insufficient training samples may encounter unsatisfactory results, and our method can effectively reduce the dependency on the remotely sensed training samples.

Tables VI–IX provide the results on Pavia University, Salinas Valley, Urban, and Indian Pines datasets, respectively. From Tables V to IX, we can see that our fine-tuned model performs the best, which proves the ability of solving the problem of insufficient training data.

Figs. 9 and 10 are the visual comparisons under experimental setting I, respectively. In addition, Fig. 11 shows the RMSE results in each remotely sensed imagery. As is indicated in Figs. 9 and 10, Ours-S recovers more details and sharper edges than the competing methods. Meanwhile, our method by pre-training (Ours-P) and fine-tuning (Ours-F) produce better results than Ours-S.

Experimental Setting II: To further verify the flexibility of our pretraining strategy, we apply this strategy on datasets that are collected by the same sensor and different sensors. We test super-resolution experiment on Pavia University dataset. In the same sensor case, we set Pavia Centre as the training data, given Pavia Centre and Pavia Centre captured by ROSIS sensor. In Table X, we use -Pavia to represent the model pretrained on Pavia Centre. However, it is common that there are few training samples in remotely sensed datasets, leading to the problem of insufficient training data that cannot be fundamentally solved. By contrast, we utilize the large-scale spectral data of natural scenes from ICVL dataset, which is corresponded to our pretraining strategy. We use -ICVL to represent the model pretrained on ICVL dataset. These two cases are quite different, since the training and testing sets are from different sensors and wavelengths.

Table X shows the quantitative results in Pavia University. Our strategy of training (-ICVL) is much better than the previous training strategy (-Pavia). Besides, our training strategy can also boost the performance of 3D-FCNN as well. The corresponding visual results of these methods are illustrated in Fig. 12.

The experimental results of remotely sensed images indicate that a small training set generated directly from Pavia Centre dataset limits the performance of the learning-based model. In addition, the pretraining strategy makes it possible to utilize the natural HSIs to improve the remote sensing HSIs, and establishes the bridge of natural HSIs and remote sensing HSIs. Fine-tuning the model directly pretrained on ICVL dataset to predict the image of Pavia University has significantly improved the results. It demonstrates that the proposed method is robust and can be well generalized to data that has not been seen during training and has different spectral numbers. First, due to the introduction of 3D convolution, our network handles HSIs with any number of spectra as input. Besides, our network can properly process HSIs through well-designed architecture, thereby improving the generalization ability of the network.

[1] Since the performance of VDSR is similar to BWVDSR, we compare with all mentioned competing algorithms except VDSR to save the space.

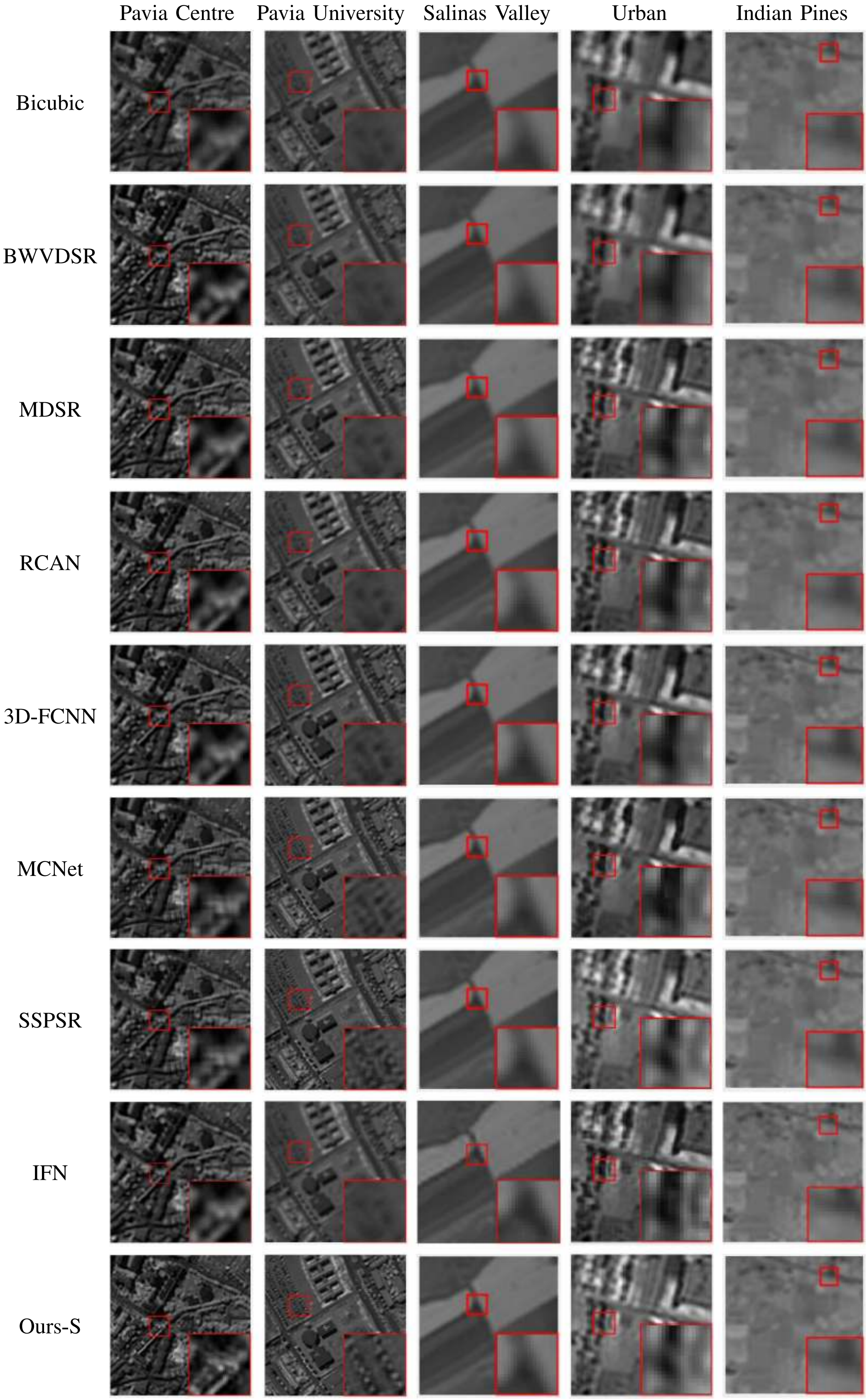


Fig. 9. Super-resolution results with scale factor ×2 on all remotely sensed datasets. Comparisons are between all training from scratch methods. We select the 88th band of Pavia Centre and Pavia University, the 68th band of Salinas Valley and Indian Pines, and the 100th band of Urban to visualize the effects.

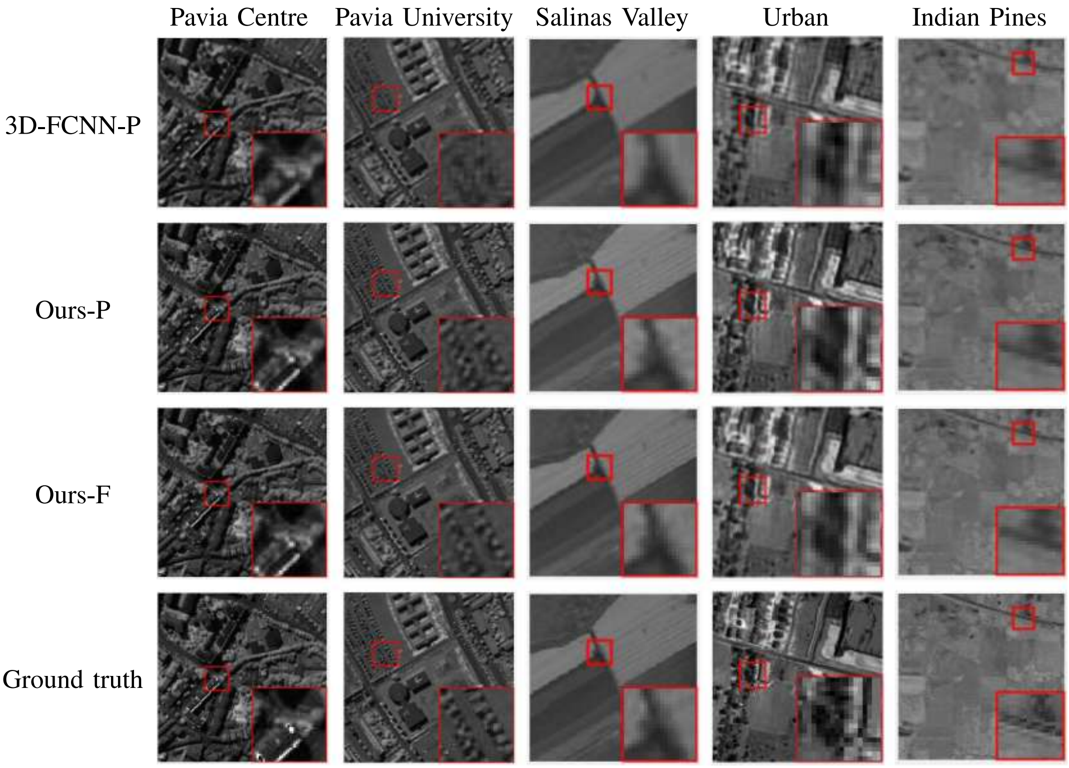
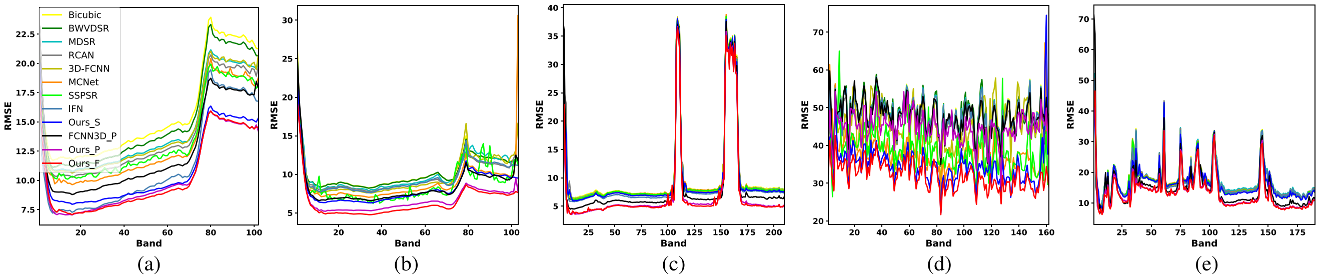
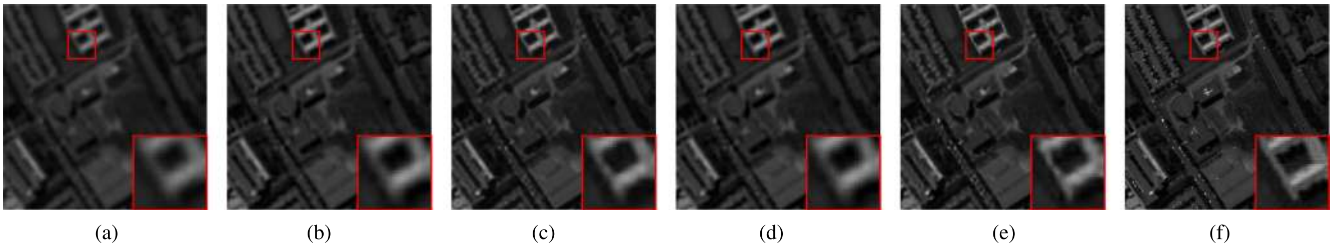


Fig. 10. Super-resolution results with scale factor ×2 on all remotely sensed datasets. Comparisons are between all pretrained or fine-tuned methods. We select the 88th band of Pavia Centre and Pavia University, the 68th band of Salinas Valley and Indian Pines, and the 100th band of Urban to visualize the effects.



(a) (b) (c) (d) (e)

Fig. 11. RMSE results along spectra for all methods with scale factor ×2 on the five remotely sensed datasets. (a) Pavia Centre. (b) Pavia University. (c) Salinas Valley. (d) Urban. (e) Indian Pines.



(a) (b) (c) (d) (e) (f)

Fig. 12. Super-resolution results at the 37th band of Pavia University scene with scale factor ×4 on Pavia University dataset in Experimental setting II. (a) Bicubic. (b) 3D-FCNN-Pavia. (c) Ours-Pavia. (d) 3D-FCNN-ICVL. (e) Ours-ICVL. (f) Ground truth.

TABLE V

Quantitative Evaluation of Competing Super-Resolution Methods by Average PSNR/SSIM/SAM for Different Scale Factors on Pavia Centre Dataset

Scale Factor	Metrics	Methods											
		Bicubic [48]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours-S	3D-FCNN-P [9]	Ours-P	Ours-F
×2	PSNR ↑	24.409	25.141	25.227	25.549	25.295	26.000	25.792	27.468	27.707	26.571	28.199	**28.267**
	SSIM ↑	0.7692	0.8174	0.8123	0.8217	0.8172	0.8358	0.8288	0.8627	0.8841	0.8574	0.8929	**0.8963**
	SAM ↓	0.1149	0.1090	0.1079	0.1044	0.1073	0.1031	0.1107	0.0992	0.0959	0.1051	0.0918	**0.0892**
×4	PSNR ↑	24.227	24.891	24.972	25.271	25.070	25.131	25.271	25.221	25.489	25.489	25.686	**25.708**
	SSIM ↑	0.7581	0.8016	0.7974	0.8058	0.8032	0.8115	0.8127	0.7982	0.8156	0.8158	0.8200	**0.8204**
	SAM ↓	0.1170	0.1116	0.1108	0.1077	0.1100	0.1124	0.1106	0.1101	0.1088	0.1117	0.1051	**0.1036**
×8	PSNR ↑	21.950	21.717	21.875	21.921	21.865	21.992	22.003	21.987	22.008	21.928	22.107	**22.110**
	SSIM ↑	0.6497	0.6517	0.6534	0.6529	0.6538	0.6492	0.6374	0.6499	0.6559	0.6549	**0.6630**	0.6617
	SAM ↓	0.1412	0.1438	0.1426	0.1407	0.1431	0.1412	0.1954	0.1409	0.1402	0.1436	0.1377	**0.1371**

The bold indicates the best performance.

TABLE VI

Quantitative Evaluation of Competing Super-Resolution Methods by Average PSNR/SSIM/SAM for Different Scale Factors on Pavia University Dataset

Scale Factor	Metrics	Methods											
		Bicubic [48]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours-S	3D-FCNN-P [9]	Ours-P	Ours-F
×2	PSNR ↑	28.099	28.121	28.313	28.584	28.571	29.353	29.660	29.918	30.101	29.931	31.794	**32.396**
	SSIM ↑	0.8630	0.8638	0.8762	0.8785	0.8861	0.8990	0.9069	0.9048	0.9084	0.9052	0.9251	**0.9304**
	SAM ↓	0.0794	0.0793	0.0790	0.0779	0.0822	0.0843	0.0951	0.0788	0.0719	0.0757	0.0644	**0.0621**
×4	PSNR ↑	27.962	27.982	28.138	28.395	28.422	28.467	28.483	28.012	28.601	28.973	29.193	**29.356**
	SSIM ↑	0.8565	0.8571	0.8678	0.8699	0.8751	0.8765	0.8724	0.8621	0.8770	0.8824	0.8846	**0.8894**
	SAM ↓	0.0805	0.0804	0.0804	0.0793	0.0838	0.0873	0.0939	0.0798	0.0765	0.0793	0.0716	**0.0696**
×8	PSNR ↑	25.600	26.130	25.938	25.779	25.898	26.103	26.330	25.912	26.126	26.145	26.269	**26.344**
	SSIM ↑	0.7913	0.7973	0.7967	0.7678	0.7939	0.7981	0.8029	0.7938	0.7983	0.8001	0.8018	**0.8038**
	SAM ↓	0.1108	0.0942	0.0963	0.0997	0.0995	0.0982	0.0977	0.1089	0.0926	0.0964	0.0902	**0.0899**

The bold indicates the best performance.

TABLE VII

Quantitative Evaluation of Competing Super-Resolution Methods by Average PSNR/SSIM/SAM for Different Scale Factors on Salinas Valley Dataset

Scale Factor	Metrics	Methods											
		Bicubic [48]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours-S	3D-FCNN-P [9]	Ours-P	Ours-F
×2	PSNR ↑	29.075	29.081	29.168	29.282	29.168	29.165	29.205	29.382	29.492	31.005	32.423	**32.542**
	SSIM ↑	0.8745	0.8761	0.8775	0.8797	0.8792	0.8819	0.8817	0.8840	0.8849	0.8943	0.9002	**0.9037**
	SAM ↓	0.1170	0.1077	0.1072	0.1082	0.1230	0.1120	0.1119	0.1169	0.1066	0.1058	0.0985	**0.0984**
×4	PSNR ↑	29.030	29.129	29.104	29.208	29.135	29.133	29.177	29.112	29.493	29.677	30.042	**30.130**
	SSIM ↑	0.8721	0.8802	0.8747	0.8767	0.8760	0.8791	0.8790	0.8797	0.8820	0.8829	0.8876	**0.8888**
	SAM ↓	0.1181	0.1079	0.1073	0.1083	0.1231	0.1119	0.1118	0.1179	0.1065	0.1073	**0.1007**	0.1008
×8	PSNR ↑	27.803	27.905	27.860	27.843	27.706	27.702	27.741	27.831	27.861	27.834	27.707	**27.901**
	SSIM ↑	0.8388	0.8482	0.8486	0.8486	0.8636	0.8485	0.8489	0.8401	0.8496	0.8494	0.8449	**0.8492**
	SAM ↓	0.1194	0.1115	0.1099	0.1110	0.1256	0.1147	0.1146	0.1190	0.1096	0.1109	**0.1049**	0.1052

The bold indicates the best performance.

D. Ablation Study

In this section, we provide comprehensive ablation study to verify the effectiveness of our proposed method. We focus on the implementation of each component especially associated with HSI domain knowledge embedding.

Modules Investigation: To show the effectiveness of 3D convolution and quasi-recurrent pooling modules in our Bi-3DQRU, the ablation study is performed and the corresponding results on ICVL dataset are provided in Table XI. In the experiment, two variants of our Bi-3DQRU are tested, i.e., 3D convolutional (C3D) [51] and Bi-2DQRU. C3D is

TABLE VIII
QUANTITATIVE EVALUATION OF COMPETING SUPER-RESOLUTION METHODS BY AVERAGE PSNR/SSIM/SAM FOR DIFFERENT SCALE FACTORS ON URBAN DATASET

Scale Factor	Metrics	Methods											
		Bicubic [48]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours-S	3D-FCNN-P [9]	Ours-P	Ours-F
×2	PSNR ↑	22.891	23.087	24.081	24.200	24.380	25.044	25.547	24.279	25.925	24.390	26.157	**26.728**
	SSIM ↑	0.7031	0.7483	0.7766	0.7831	0.7919	0.8169	0.8378	0.7899	0.8513	0.8025	0.8646	**0.8771**
	SAM ↓	0.1320	0.1301	0.1154	0.1136	0.1171	0.1129	0.1125	0.1173	0.1034	0.1128	0.0976	**0.0883**
×4	PSNR ↑	20.183	20.677	20.351	20.312	20.634	20.980	20.752	20.229	20.472	20.557	20.842	**20.915**
	SSIM ↑	0.6648	0.6657	0.6649	0.6652	0.6663	0.6665	0.6672	0.7001	0.7054	0.6642	0.6991	**0.7104**
	SAM ↓	0.1610	0.1613	0.1611	0.1608	0.1605	0.1605	0.1602	0.1598	0.1556	0.1531	0.1455	**0.1227**
×8	PSNR ↑	20.669	20.672	20.673	20.675	20.670	20.691	20.696	20.680	20.679	20.680	20.713	**20.798**
	SSIM ↑	0.5707	0.5711	0.5725	0.5729	0.5790	0.5731	0.5782	0.5745	0.5787	0.5778	0.5774	**0.5787**
	SAM ↓	0.1618	0.1614	0.1609	0.1600	0.1571	0.1572	0.1569	0.1617	0.1495	0.1573	0.1604	**0.1495**

The bold indicates the best performance.

TABLE IX
QUANTITATIVE EVALUATION OF COMPETING SUPER-RESOLUTION METHODS BY AVERAGE PSNR/SSIM/SAM FOR DIFFERENT SCALE FACTORS ON INDIAN PINES DATASET

Scale Factor	Metrics	Methods											
		Bicubic [48]	BWVDSR [35]	MDSR [36]	RCAN [37]	3D-FCNN [9]	MCNet [13]	SSPSR [11]	IFN [14]	Ours-S	3D-FCNN-P [9]	Ours-P	Ours-F
×2	PSNR ↑	23.670	23.678	23.556	23.714	23.918	23.710	23.791	23.880	24.005	25.303	26.014	**26.148**
	SSIM ↑	0.6485	0.6491	0.6466	0.6552	0.6977	0.6876	0.6918	0.6501	0.6893	0.7612	0.7938	**0.7976**
	SAM ↓	0.1368	0.1368	0.1370	0.1352	0.1367	0.1409	0.1390	0.1302	0.1296	0.1283	0.1107	**0.1104**
×3	PSNR ↑	23.694	23.701	23.575	23.733	23.932	23.714	23.800	23.718	24.055	24.921	25.357	**25.433**
	SSIM ↑	0.6489	0.6495	0.6468	0.6554	0.6977	0.6874	0.6913	0.6521	0.6902	0.7418	0.7629	**0.7673**
	SAM ↓	0.1366	0.1367	0.1369	0.1351	0.1367	0.1404	0.1391	0.1332	0.1294	0.1302	0.1139	**0.1140**
×4	PSNR ↑	23.237	23.240	23.155	23.279	23.350	23.181	23.2467	23.240	23.422	23.899	23.996	**24.171**
	SSIM ↑	0.6055	0.6060	0.6062	0.6131	0.6429	0.6359	0.6387	0.6078	0.6323	0.6653	0.6635	**0.6763**
	SAM ↓	0.1413	0.1413	0.1412	0.1396	0.1422	0.1459	0.1443	0.1401	0.1356	0.1385	0.1232	**0.1236**

The bold indicates the best performance.

TABLE X
QUANTITATIVE EVALUATION OF COMPETING SUPER-RESOLUTION METHODS BY AVERAGE PSNR/SSIM/SAM FOR DIFFERENT SCALE FACTORS ON PAVIA UNIVERSITY SCENE

Scale Factor	Metrics	Methods				
		Bicubic [48]	3D-FCNN -Pavia [9]	Ours -Pavia	3D-FCNN -ICVL [9]	Ours -ICVL
×2	PSNR ↑	28.099	29.068	30.653	29.931	**31.794**
	SSIM ↑	0.8630	0.8944	0.9249	0.9052	**0.9251**
	SAM ↓	0.0794	0.0774	0.0765	0.0757	**0.0644**
×4	PSNR ↑	27.962	28.096	28.130	28.973	**29.193**
	SSIM ↑	0.8565	0.8648	0.8664	0.8824	**0.8846**
	SAM ↓	0.0805	0.0811	0.0809	0.0793	**0.0716**
×8	PSNR ↑	25.600	25.624	25.908	26.145	**26.269**
	SSIM ↑	0.7913	0.7930	0.7943	0.8001	**0.8018**
	SAM ↓	0.1108	0.1121	0.1153	0.0964	**0.0902**

We use '-Pavia' to represent the pretrained model on Pavia Centre dataset, and '-ICVL' is to denote the pretrained model on ICVL dataset.
The bold indicates the best performance.

TABLE XI
MODULES INVESTIGATION ON ICVL DATASET

Scale Factor	Metrics	C3D	Bi-2DQRU	Bi-3DQRU
x2	PSNR ↑	40.338	42.216	**44.917**
	SSIM ↑	0.9854	0.9902	**0.9932**
	SAM ↓	0.0461	0.0442	**0.0252**
x4	PSNR ↑	38.478	39.628	**40.387**
	SSIM ↑	0.9766	0.9805	**0.9817**
	SAM ↓	0.0464	0.0442	**0.0294**
x8	PSNR ↑	33.059	33.612	**33.655**
	SSIM ↑	0.9354	0.9391	**0.9394**
	SAM ↓	0.0517	0.0531	**0.0381**
	Time (s)	0.56	0.61	0.72

The bold indicates the best performance.

formed by removing the quasi-recurrent pool module. Bi-2DQRU is constructed by replacing the 3D convolution with 2D convolution.

We can see that the performance of C3D is worse than our Bi-3DQRU (−4.6 dB in ×2 case, −1.9 dB in ×4 case, and −0.6 dB in ×8 case). C3D has removed quasi-recurrent pool function compared with our Bi-3DQRU, and cannot well model the global correlation along spectra. Meanwhile, if we keep the quasi-recurrent pool module but replace 3D convolution with 2D convolution, the PSNR is decreased by 2.7 dB in ×2 case, 0.8 dB in ×4 case, and 0.04 dB in ×8 case.

TABLE XII
Number of Feature Maps (Feature Maps), FLOPS, and Number of Parameters (Params) of Unidirectional Structure, Alternative Directional Structure, and Bidirectional Structure.

	Unidirectional		Alternative Directional		Bidirectional
Feature Maps	16	20	16	20	16
FLOPS	18.2889G	28.4807G	18.1868G	28.3530G	27.2805G
Params	859.680k	1342.440k	858.816k	1341.360k	1288.227k

TABLE XIII
Ablation Study on the Direction of Network on ICVL Dataset

Scale Factor	Metrics	Unidirectional (20 feature maps)	Alternative Directional (20 feature maps)	Bidirectional (16 feature maps)
x2	PSNR ↑	43.436	43.507	**44.917**
	SSIM ↑	0.9920	0.9919	**0.9923**
	SAM ↓	0.0295	0.0272	**0.0252**
x4	PSNR ↑	40.317	40.293	**40.387**
	SSIM ↑	0.9818	0.9813	**0.9817**
	SAM ↓	0.0309	0.0298	**0.0294**
x8	PSNR ↑	33.604	33.603	**33.655**
	SSIM ↑	0.9383	0.9388	**0.9394**
	SAM ↓	0.0383	0.0387	**0.0381**

The bold indicates the best performance.

Direction of Network: To investigate the effect of our bidirectional structure, we make a comprehensive comparison between unidirectional structure, alternative directional structure, and our bidirectional structure. The computation overload of networks is illustrated in Table XII. We can see that our bidirectional structure with 16 feature maps has approximately 1.5 times parameters compared to the other two structures. Then, we increase the number of feature maps of unidirectional structure and alternative directional structure. Thereby, we set the number of feature maps unidirectional structure and alternative directional structure as 20, since $16 \times \sqrt{1.5} \approx 20$. Note that increasing the number of feature maps implies increasing the number of input channels and output channels simultaneously. The corresponding results on ICVL dataset are provided in Table XIII. Unidirectional structure only considers the forward spectral dependency. On the contrary, our bidirectional structure considers both forward and backward dependencies, thereby improving the performance. Although the alternative directional structure saves the number of parameters, its performance is limited by the bias caused by the asymmetrical structure of the network. Compared with unidirectional and alternative directional structure network, the bidirectional one performs better in all super-resolution scale factors.

V. Conclusion

In this article, we presented Bi-3DQRNN, a single HSI super-resolution method that makes full use of the structural spatial-spectral correlation and global correlation along spectra simultaneously. To introduce the backward spectral dependency of HSI, we adopt a bidirectional structure. The well-designed Bi-3DQRNN can effectively deal with the insufficiency of remotely sensed training samples. Evaluations and comparisons on a natural HSI dataset and five remotely sensed datasets show that our Bi-3DQRNN outperforms state-of-the-art HSI super-resolution methods and demonstrate that the introduction of natural HSIs can effectively improve the performance on remote sensing HSI.

In future, it is worth investigating how to compress our network, especially the quasi-recurrent pooling module, and make it more lightweight to extend its application fields.

References

[1] C. Atzberger, "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs," *Remote Sens.*, vol. 5, no. 2, pp. 949–981, Feb. 2013.

[2] I. Aneece and P. Thenkabail, "Accuracies achieved in classifying five leading world crop types and their growth stages using optimal Earth observing-1 hyperion hyperspectral narrowbands on Google Earth Engine," *Remote Sens.*, vol. 10, no. 12, Dec. 2018, Art. no. 2027.

[3] M. Moroni, E. Lupo, E. Marra, and A. Cenedese, "Hyperspectral image analysis in environmental monitoring: Setup new tunable filter platform," *Procedia Environ. Sci.*, vol. 19, pp. 885–894, 2013.

[4] C. G. Diniz *et al.*, "DETER-B: The new Amazon near real-time deforestation detection system," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3619–3628, Jul. 2015.

[5] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Electron.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[6] D. Zhu, B. Du, and L. Zhang, "Target dictionary construction-based sparse representation hyperspectral target detection methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1254–1264, Apr. 2019.

[7] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[8] F. Ling, X. Li, Y. Du, and F. Xiao, "Super-resolution land cover mapping with spatial-temporal dependence by integrating a former fine resolution map," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1816–1825, May 2014.

[9] 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.

[10] J. Hu, M. Zhao, and Y. Li, "Hyperspectral image super-resolution by deep spatial-spectral exploitation," *Remote Sens.*, vol. 11, no. 10, 2019, Art. no. 1229.

[11] 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.

[12] J. Yang, Y. Zhao, J. C. Chan, and L. Xiao, "A multi-scale wavelet 3D-CNN for hyperspectral image super-resolution," *Remote Sens.*, vol. 11, no. 13, 2019, Art. no. 1557.

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

[14] 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.

[15] S. Gou, S. Liu, S. Yang, and L. Jiao, "Remote sensing image super-resolution reconstruction based on nonlocal pairwise dictionaries and double regularization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4784–4792, Dec. 2014.

[16] X. Han, J. Yu, and W. Sun, "Hyperspectral image super-resolution based on non-factorization sparse representation and dictionary learning," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 963–966.

[17] S. Tang, Y. Xu, L. Huang, and L. Sun, "Hyperspectral image super-resolution via adaptive dictionary learning and double ℓ_1 constraint," *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2809.

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

[19] X. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with CNN for hyperspectral image super-resolution," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 2506–2510.

[20] 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.

[21] K. Zheng *et al.*, "Separable-spectral convolution and inception network for hyperspectral image super-resolution," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 10, pp. 2593–2607, 2019.

[22] 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, 2020.

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

[24] W. Dong, H. Wang, F. Wu, G. Shi, and X. Li, "Deep spatial-spectral representation learning for hyperspectral image denoising," *IEEE Trans. Comput. Imag.*, vol. 5, no. 4, pp. 635–648, Dec. 2019.

[25] P. Gamba, "A collection of data for urban area characterization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, pp. 69–72.

[26] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.

[27] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," Dept. Comput. Sci., Columbia Univ. CUCS-061-08, Tech. Rep., Nov. 2008.

[28] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 193–200.

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

[30] 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.

[31] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ, USA: John Wiley and Sons, 2005.

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

[33] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11661–11670.

[34] Q. Wei, J. M. Bioucasdias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Electron.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[35] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[36] 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.*, 2017, pp. 1132–1140.

[37] 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.*, 2018, pp. 294–310.

[38] 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.

[39] Z. He and L. Liu, "Hyperspectral image super-resolution inspired by deep Laplacian pyramid network," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1939.

[40] 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.

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

[42] L. Wang, T. Bi, and Y. Shi, "A frequency-separated 3D-CNN for hyperspectral image super-resolution," *IEEE Access*, vol. 8, pp. 86 367–86379, 2020.

[43] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[44] K. Wei, Y. Fu, and H. Huang, "3-D quasi-recurrent neural network for hyperspectral image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 363–375, Jan. 2021.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[46] R. H. Yuhas, J. W. Boardman, and A. F. Goetz, "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," in *Proc. Summaries 4th Annu. JPL Airborne Geosci. Workshop*, 1993, pp. 205–208.

[47] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent*., May 2015.

[48] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[49] 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.

[50] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Joint camera spectral sensitivity selection and hyperspectral image recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 788–804.

[51] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. Int. Conf. Comput*. Vis., 2015, pp. 4597–4605.

Ying Fu received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2009, the M.S. degree in automation from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree in information science and technology from the University of Tokyo, Japan, in 2015.

She is currently a Professor with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. Her research interests include computer vision, image and video processing, and computational photography.

Zhiyuan Liang received the bachelor's degree in Internet of Things engineering from the Heifei University of Technology, Hefei, China, in 2020. She is currently working toward the master degree from the Beijing Institute of Technology, Beijing, China.

Her research interests include computer vision, deep learning, and their applications on remote sensing image processing.

Shaodi You received the B.S. degree in electronic information engineering from Tsinghua University, China, in 2009, the M.E. degree in automation and Ph.D. degree in information science and technology from the University of Tokyo, Japan, in 2012 and 2015, respectively.

He is an Assistant Professor with the Computer Vision Research Group, Institute of Informatics, University of Amsterdam, The Netherlands. His research interests include physics-based vision, perception-based vision and learning, and 3D geometry. He is best known for his work on visual enhancement for rainy days.

Dr. You is the Program Chair for ICCV2017, ICCV2019 Workshop on Physics-Based Vision meets Deep Learning. He is the Program Chair for ICCV2019 Workshop on e-Heritage. He is the General Chair of DICTA2018 and the Workshop Chair for ACCV2018.