Spectral-spatial self-attention networks for hyperspectral image classification.

ZHANG, X., SUN, G., JIA, X., WU, L., ZHANG, A., REN, J., FU, H. and YAO, Y.

2022

Spectral–Spatial Self-Attention Networks for Hyperspectral Image Classification

Xuming Zhang, Genyun Sun, *Member, IEEE*, Xiuping Jia, *Fellow, IEEE*, Lixin Wu, Aizhu Zhang, *Member, IEEE*, Jinchang Ren, *Senior Member, IEEE*, Hang Fu, and Yanjuan Yao

***Abstract*— This study presents a spectral–spatial self-attention network (SSSAN) for classification of hyperspectral images (HSIs), which can adaptively integrate local features with long-range dependencies related to the pixel to be classified. Specifically, it has two subnetworks. The spatial subnetwork introduces the proposed spatial self-attention module to exploit rich patch-based contextual information related to the center pixel. The spectral subnetwork introduces the proposed spectral self-attention module to exploit the long-range spectral correlation over local spectral features. The extracted spectral and spatial features are then adaptively fused for HSI classification. Experiments conducted on four HSI datasets demonstrate that the proposed network outperforms several state-of-the-art methods.**

***Index Terms*— Convolutional neural network (CNN), deep learning, hyperspectral image (HSI) classification, spatial self-attention module, spectral self-attention module.**

I. Introduction

HYPERSPECTRAL images (HSIs) have hundreds of spectral bands, collecting abundant spectral and spatial information for monitoring the surface of the Earth [1]. Such valuable information enables them to discriminate more land-cover materials under various conditions, facilitating a wide range of applications, including environment observing [2], resources assessment [3], and urban development monitoring [4]. Classification is one of the important tasks for these applications.

Over the past few decades, various HSI classification methods have been developed. Earlier methods are mainly focused on the spectral features, where typical approaches include support vector machines (SVMs) [5], multinomial logistic regression (MLR) [6], and manifold learning [7]. To mitigate the problem of dimensionality inherent in HSI, some dimensionality reduction strategies were proposed based on feature extraction [8] and band selection [9]. Some unmixing models were also proposed to address the spectral mixture issues in HSI, such as the extended linear mixing model (ELMM) [10] and the augmented linear mixing model (ALMM) [11]. Another unmixing model called sparsity-enhanced convolutional decomposition (SeCoDe) was proposed in [12], which uses the convolution operation to learn spatial contextual information to improve its unmixing performance.

An increasing number of methods incorporate spatial features to improve the class representation using spectral features alone. Some works extract spatial features via morphological operators [13], Gabor filters [14], and hypergraph structure [15] or apply Markov random fields (MRFs) [16], among others, and then combine them with spectral features for classification. Others directly extract the joint spectral–spatial features by using 3-D discrete wavelets [17], 3-D scattering wavelets [18], 3-D Gabor filters [19], and so on. Nevertheless, these traditional methods extract the features of the original data in a shallow manner, which is difficult to achieve substantial performance gain.

In recent years, deep learning algorithms have successfully broken the limitations of the traditional feature extraction techniques. It can automatically extract hierarchical features from data, achieving significant progress in computer vision, including object detection [20], semantic segmentation [21], and image classification [22]. Furthermore, various deep learning models have been investigated in HSI classification. Multilayer perceptron (MLP) [23], stacked autoencoder (SAE) [24], and deep belief network (DBN) [25] were used for feature extraction of HSI. In [26], the recurrent neural network (RNN) was used to analyze the hyperspectral sequential data, and then, it was classified via network reasoning. In [27], convolutional neural network (CNN) was used for deep spectral–spatial feature extraction and classification. Hong *et al.* [28] applied the graph convolutional networks (GCNs) to capture large range spatial features as they can model the topological relations between samples through their graph structures.

The aforementioned studies have shown that feature extraction plays a key role in HSI classification, and it goes through

Manuscript received April 19, 2021; revised June 7, 2021; accepted July 4, 2021. This work was supported in part by the National Key Research and Development Program under Grant 2019YFE0126700; and the National Natural Science Foundation of China under Grants (41971292, 41801275, 41871270). *(Corresponding author: Genyun Sun.)*

Xuming Zhang, Genyun Sun, Aizhu Zhang, and Hang Fu are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China, and also with the Laboratory for Marine Mineral Resources, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China (e-mail: genyunsun@163.com).

Xiuping Jia is with the School of Engineering and Information Technology, University of New South Wales at Canberra, Canberra, ACT 2600, Australia.

Lixin Wu is with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China.

Jinchang Ren is with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China, and also with the National Subsea Centre, Robert Gordon University, Aberdeen AB10 7AQ, U.K. (e-mail: jinchang.ren@strath.ac.uk).

Yanjuan Yao is with the Satellite Environment Center (SEC), Ministry of Environment Protection (MEP) of China, Beijing 100094, China (e-mail: yjyao2008@aliyun.com).

an evolution from shallow to deep [29]. Among these deep learning algorithms, CNN generally outperforms others in feature extraction, mainly because its local connections and shared weights characteristics enable it to maintain the original structure while learning spatial features and greatly reduce the number of network parameters [30].

These unique characteristics make the CNN-based methods valuable in spectral–spatial classification of HSI [31]. In [32], CNNs were used to extract deep spatial features. Several end-to-end 2-D CNN models were designed to jointly exploit the spectral–spatial information by using different convolution kernels [33], [34]. More recently, 3-D CNN was used to extract the joint spectral–spatial features for HSI classification [35], [36]. While performance was improved by using the 3-D CNN, the significantly increased parameters may cause overfitting and bring additional computational cost. In [37]–[39], the spatial and spectral features were learned separately by 2-D CNN or other algorithms (e.g., SAE, 1-D CNN, and RNN) and then fused together. This scheme can achieve good performance yet significantly reduce the computational load compared with 3-D CNNs [38]. Considering the insufficient labeled samples of HSIs, we also adopt this scheme in this article to minimize the required parameters for training and avoid overfitting.

In recent years, an increasing number of studies have demonstrated that deeper networks have stronger feature representation ability, but it is difficult to optimize, especially with limited labeled samples [33]. The emergence of the residual network (ResNet) [40] and the dense convolutional network (DenseNet) [41] makes it possible to train deeper networks to boost the performance of HSI classification. In [42], a spectral–spatial residual network (SSRN) was proposed to alleviate the declining-accuracy phenomenon. A fast and dense spectral–spatial convolution (FDSSC) network was proposed to overcome the overfitting problem [43]. To extract spectral–spatial features at multiscales, a cascaded dual-scale crossover network based on SSRN was proposed in [44], where a dual-scale crossover module was designed to capture multiscale features by using different convolution kernels. In addition, a fully dense multiscale fusion network was developed to directly connect feature maps of different layers with different resolutions [45]. Despite these developments, the convolution filters of the CNN-based method still have the limitations of treating the input content equally and only modeling local features. Generally, spectral and spatial features extracted from the input have different contributions to classification.

Recently, the attention mechanism was developed by simulating the human visual system, which can selectively focus on salient parts instead of treating each part equally [1]. Embedding it into the network can promote the representation capacity of the extracted features, which has achieved good performance in computer vision [46]–[48]. Subsequently, many attention mechanisms proposed for scene segmentation of generic natural images in [1], [30], and [49]–[55] have been directly applied into the patch-based CNN for HSI classification. The squeeze-and-excitation (SE) block [55], which uses global pooling to generate the channel attention matrix, was applied to a patch-based CNN to recalibrate their channel-wise feature responses [51], [56]. Subsequently, many similar spectral attention modules [57], [58] were proposed for HSI classification to selectively excite informative channels and suppress useless ones. To make the network adaptively enhance and suppress information in both spectral and spatial dimensions, many spatial–spectral attention modules were proposed. In [1], a new spectral–spatial visual attention-driven module was incorporated into the ResNet to refine the extracted features. The convolutional block attention module (CBAM) proposed for scene segmentation of generic natural images in [48] was adopted in [30] and [49] for HSI classification. Its channel-wise attention module determines the weight of each channel via MaxPooling and AvgPooling layers along the spatial dimension, and the spatial-wise attention module determines the weight of each position in the feature maps via pooling layers along the channel axis. Similarly, a cooperative spectral–spatial attention module for HSI classification was proposed in [53], which generates the spectral and spatial attention maps by using pooling layers to squeeze the spatial and channel dimensions, respectively. In [59], a dual attention network (DANet) was proposed for scene segmentation of generic natural images. Its position self-attention module captures the spatial correlation between any two positions of the feature maps, and the channel self-attention module captures the spectral correlation between any two-channel maps. These self-attention modules [59] were adopted in [52] and [60] for HSI classification and achieved the state-of-the-art performance.

In the aforementioned attention-based methods for HSI classification, some works (e.g., [49], [52], [56]) directly adopted the attention modules [48], [55], [59], which are embedded in pixel-based CNN for scene segmentation of generic natural images, to their patch-based CNN for HSI classification. Although some proposed attention modules are (e.g., [53], [54]) for HSI classification, the way they compute the attention maps is similar to those embedded in pixel-based CNNs for scene segmentation. As seen, none of them specifically design attention modules according to the characteristics of the patch-based CNN. In patch-based CNN, the input patch is used to predict its central pixel, and the neighboring pixels may have different contributions to the classification of the center pixel. Therefore, it is necessary for the patch-based CNN to explore the latent correlations between the center pixel and its neighbors in a global view.

To investigate this opportunity for better HSI classification, we proposed a spectral–spatial self-attention network (SSSAN) with two subnetworks, designed for spectral and spatial feature extraction. Specifically, the spatial subnetwork introduces the proposed spatial self-attention module to capture the spatial feature correlations between the center pixel and its surroundings. Meanwhile, the spectral subnetwork introduces the proposed spectral self-attention module to exploit the long-range correlations over local spectral features. The "score weighted" fusion method [39] is then used to fuse the extracted spatial and spectral features for classification.

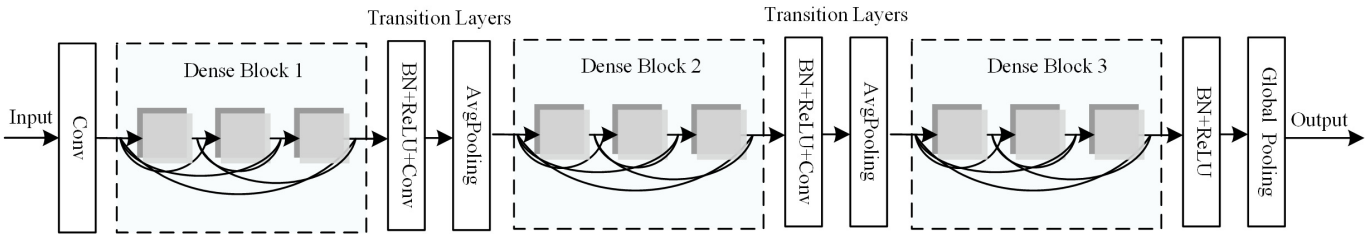


Fig. 1. Architecture of DenseNet.

The main contributions of this article can be summarized as follows.

1) A spatial self-attention module is proposed for the patch-based CNN to exploit the spatial feature correlation between the center pixel and its surroundings, which has improved the spatial feature representation related to the center pixel specifically.
2) A spectral self-attention module is designed for 1-D CNN to capture long-range spectral correlations over local spectral features.
3) The proposed spatial and spectral self-attention modules are designed as add-on blocks so that they can be plugged into any patch-based CNN and 1-D CNN backbone networks, respectively, to generate high-quality discriminant feature. Both modules are lightweight.

The remainder of this article is organized as follows. The related works of the proposed method are presented in Section II. In Section III, we describe the proposed method in detail. The experiments and results are presented and discussed in Section IV. Finally, concluding remarks are provided in Section V.

II. Related Works

In this section, we briefly introduce the basic techniques of the proposed methods, which are the DenseNet and attention mechanisms.

A. Dense Neural Networks

Generally, deeper networks have better performance, but as the network deepens, its parameters will increase, making it harder to train. The emergence of DensNet mitigates this problem. As shown in Fig. 1, the DenseNet framework is mainly composed of dense blocks and transition layers. Relevant details of these components are presented as follows.

As can be seen in each dense block, the input of each layer comes from the outputs of all previous layers of the corresponding block, which can be expressed as

$$\boldsymbol{x}_l = H_l\left(\left[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l-1}\right]\right) \tag{1}$$

where $\left[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l-1}\right]$ denotes that the output feature maps of layers 0 to $l-1$ are concatenated in the channel dimension. $H_l(\cdot)$ represents a composite function, consisting of a batch normalization (BN) layer, ReLU, and a convolutional (Conv) layer with a kernel size of 3×3 (denoted as BN-ReLU-Conv3 $\times$ 3 for short). It should be noted that each Conv layer outputs k feature maps in the dense block, where k is called growth rate in [41]. Assuming that k_0 is the number of channels in the input layer, the lth layer will have $k_0 + k \times (l-1)$ input feature maps. Therefore, as the number of layers increases, the input channels will be very large, though k is set to be small. To alleviate this, a bottleneck layer (i.e., BN-ReLU-Conv1 $\times$ 1) is added before each 3×3 Conv to reduce the number of input channels. Then, $H_l(\cdot)$ changes from BN-ReLU-Conv3 $\times$ 3 to BN-ReLU-Conv1 $\times$ 1-BN-ReLU-Conv3 $\times$ 3.

The layers between the dense blocks are the transition layers, used to reduce the size of feature maps. It consists of a BN layer, ReLU, and a 1×1 Conv layer followed by an average pooling (AvgPooling) layer.

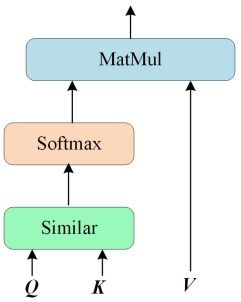


Fig. 2. Main operations of the self-attention mechanism.

B. Attention Mechanism

Attention mechanisms can not only adaptively emphasize or suppress information but also model long-range dependencies of data, which have been widely used in many tasks [59]. Recently, many attention modules have been applied to HSI classification. The attention modules proposed in [48] and [59] are embedded in pixel-based CNN for scene segmentation, which is directly applied to the patch-based CNN for HSI classification [49], [52], [60]. Other attention modules as proposed in [53] and [61] for HSI classification are similar to the modules in [48] and [55], which are designed for scene segmentation. These attention modules can be mainly divided into two categories. The first category is the conventional attention modules [49], [53], [54], [61], which computes the spatial or spectral attention map(s) by using the pooling and FC layers to exploit the inter-spatial or inter-channel relationship of the extracted features. The other category is for the self-attention modules [52], [60], which generates the spatial and spectral self-attention maps by calculating the correlation between features.

A basic structure of self-attention module is shown in Fig. 2. Its inputs consist of three matrices: Query ($\boldsymbol{Q}$),

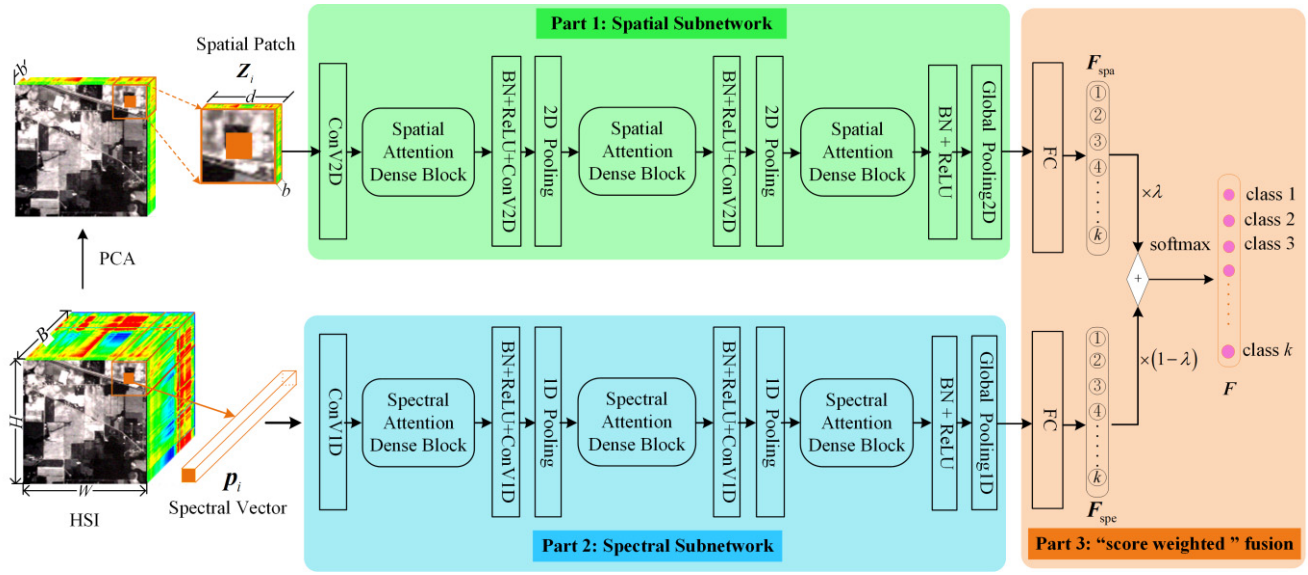


Fig. 3. Overview of the proposed SSSAN.

Key ($\boldsymbol{K}$), and Value ($\boldsymbol{V}$), all of which come from the same input. The similarity is calculated between $\boldsymbol{Q}$ and $\boldsymbol{K}$, and then, the results are normalized by a softmax function, getting the self-attention matrix. Finally, multiply the obtained self-attention matrix by the matrix $\boldsymbol{V}$ to get the output. This operation can be described as

$$\text{Attention}(Q, K, V) = \text{Softmax}[\text{sim}(Q, K)]V. \tag{2}$$

Self-attention mechanisms can effectively strengthen global feature representations by using fewer parameters. These existing ones, however, are mainly designed for the scene segmentation task, usually embedded in the pixel-based CNN. In HSI classification, few attention modules are designed based on the uniqueness of patch-based CNN and 1-D CNN.

III. Method

A. Overview

In this section, a novel self-attention module-based CNN architecture is proposed to optimize the discrimination of the extracted features. As shown in Fig. 3, the proposed network consists of three parts, including the spatial self-attention module-based spatial subnetwork, the spectral self-attention module-based spectral subnetwork, and the "score weighted" fusion and classification. Specifically, denote $\mathbf{H} \in \mathbb{R}^{H\times W\times B}$ as an HSI data cube, where H, W, and B denote, respectively, the length, the width, and the number of bands of $\mathbf{H}$. As PCA has no training parameters, we use PCA to simply reduce the dimension of B into b. After dimension reduction, for a pixel $\boldsymbol{p}_i$ to be classified, a spatial patch $\mathbf{Z}_i \in \mathbb{R}^{d\times d\times b}$ centered at $\boldsymbol{p}_i$ is taken as the spatial subnetwork input. It passes through three spatial attention dense blocks, two 2-D transition layers, and a global 2-D average pooling, and eventually, the 1-D spatial features can be produced. Meanwhile, the spectrum of $\boldsymbol{p}_i$ is taken as the spectral subnetwork input. It passes through three spectral attention dense blocks, two 1-D transition layers, and a global 1-D average pooling, and eventually, the 1-D spectral features can be produced. The extracted spatial and spectral features are fed into the "score weighted" fusion part for classification.

After the network is built, its parameters are initialized with the He normalization [62] and regularized with the L2 weight decay penalty. The network is trained in an end-to-end manner. During the training process, the Adam [63] optimizer is used to update the parameters of the network through backpropagating the gradient of the cross-entropy cost function. In the following, relevant details of the three parts of the proposed network are presented.

B. Spatial Self-Attention Module-Based Spatial Subnetwork

It is essential to explore discriminant spatial feature representations for more effective HSI classification. Over the past few years, many spatial attention modules [48], [54], [59] were proposed to enhance its discriminability. These spatial attention modules encode where to emphasize or suppress by utilizing the inter-spatial relationship of features. However, none of them explore the latent correlation between the center pixel and its surroundings. In the patch-based CNN, the spatial support from the neighbors around class boundary is often invalid as these neighboring pixels sometimes are different from the center pixel's category. During the convolution operation, these neighboring pixels will have a negative effect on feature learning [64]. To resolve this problem, we design a spatial self-attention module for the patch-based CNN. It assigns weights to different features by measuring the similarity between the surrounding features and its central one. Therefore, it can adaptively strengthen the relevantly long-range features to the center pixel while suppressing unnecessary ones for improving the spatial feature representation in predicting the center pixel.

Fig. 4(b) shows the operation of the proposed spatial self-attention module. Let $\boldsymbol{X} \in \mathbb{R}^{w\times w\times c}$ be the input feature maps, where $w \times w$ denotes the spatial size and c denotes the number of channels. Note that w is always an odd number in the

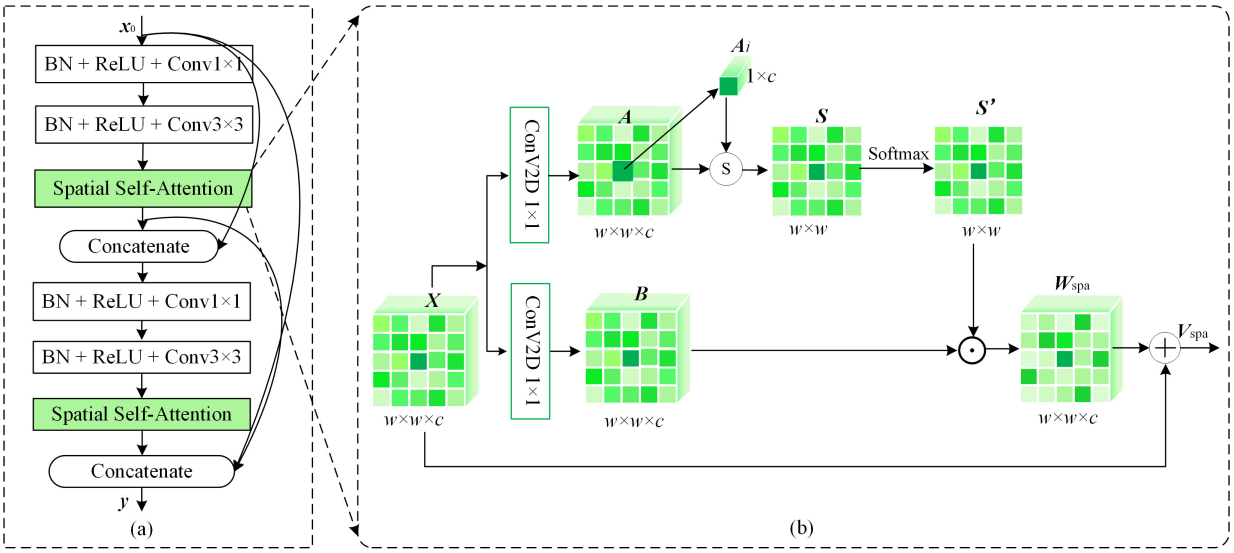


Fig. 4. Spatial attention dense block. (a) Dense block embedded with spatial self-attention module. (b) Proposed spatial self-attention module.

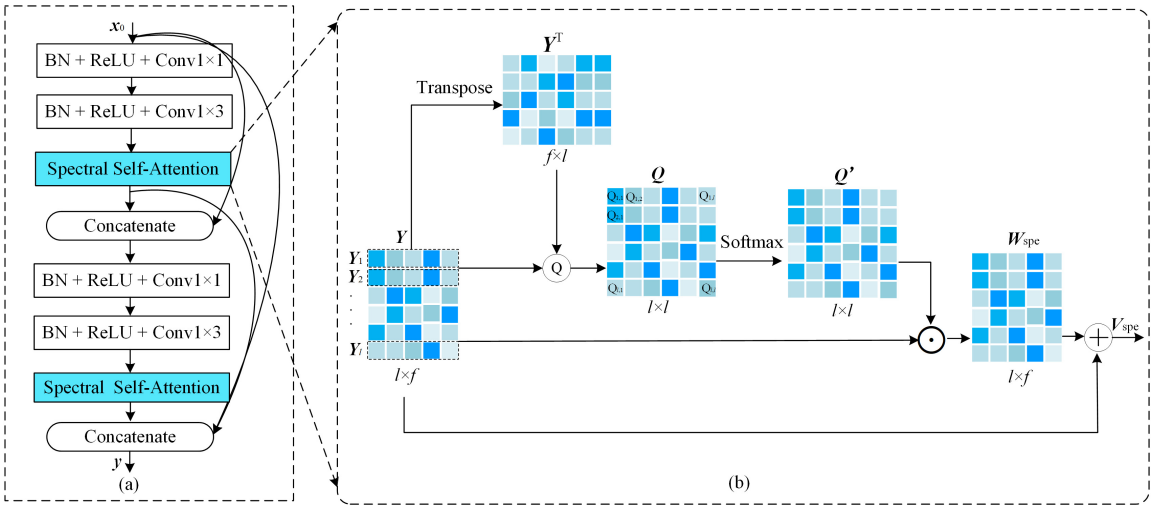


Fig. 5. Spectral attention dense block. (a) Dense block embedded with spectral self-attention module. (b) Proposed spectral self-attention module.

patch-based CNN since the size of the input patch is odd, and the convolution and pooling operations do not change the parity of its inputs. To facilitate the relational operation between spatial features, we first feed $\boldsymbol{X}$ into two parallel 1×1 Conv layers to generate two new feature maps of $\boldsymbol{A}$ and $\boldsymbol{B}$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{w\times w\times c}$. Note that we denote the center vector of $\boldsymbol{A}$ as $\boldsymbol{A}_i \in \mathbb{R}^{1\times c}$ and all its neighbors in $\boldsymbol{A}$ as $\left[\boldsymbol{A}_{i,1}, \boldsymbol{A}_{i,2}, \boldsymbol{A}_{i,3}, \ldots, \boldsymbol{A}_{i,n}\right]$ with $n = w \times w$. The similarity between $\boldsymbol{A}_i$ and its neighbors $\left[\boldsymbol{A}_{i,1}, \boldsymbol{A}_{i,2}, \boldsymbol{A}_{i,3}, \ldots, \boldsymbol{A}_{i,n}\right]$ is evaluated as

$$\boldsymbol{S}_{i,t} = \operatorname{sim}\left(\boldsymbol{A}_i, \boldsymbol{A}_{i,t}\right) = \left(\frac{\boldsymbol{A}_i \boldsymbol{A}_{i,t}^T}{\|\boldsymbol{A}_i\| \left\|\boldsymbol{A}_{i,t}\right\|}\right)^2, \quad t = 1, 2, \ldots, n \tag{3}$$

where $\boldsymbol{S}_{i,t}$ measures the feature correlation between the center vector $\boldsymbol{A}_i$ and its neighborhood vector $\boldsymbol{A}_{i,t}$. The softmax function is used to normalize $\boldsymbol{S} \in \mathbb{R}^{w\times w}$ to obtain the spatial attention map

$$\boldsymbol{S}'_{i,t} = \frac{\exp\left(\boldsymbol{S}_{i,t}\right)}{\sum_{t=1}^{n} \exp\left(\boldsymbol{S}_{i,t}\right)}, \quad t = 1, 2, \ldots, n. \tag{4}$$

A weighted matrix is obtained by

$$\boldsymbol{W}_{\text{spa}} = \boldsymbol{B} \otimes \boldsymbol{S}' \tag{5}$$

where $\otimes$ denotes the element-wise product, in which the spatial attention values are broadcasted along the channel dimension. $\boldsymbol{W}_{\text{spa}} \in \mathbb{R}^{w\times w\times c}$ is the refined output, which focuses on more informative features related to the center pixel, while suppressing unnecessary ones. To generate the residual connection $\boldsymbol{V}_{\text{spa}}$

$$\boldsymbol{V}_{\text{spa}} = \boldsymbol{W}_{\text{spa}} + \boldsymbol{X} \tag{6}$$

where $\boldsymbol{V}_{\text{spa}} \in \mathbb{R}^{w\times w\times c}$. It can be inferred from (6) that the features similar to the central pixel feature are enhanced, while dissimilar ones are suppressed, thus improving the feature representation ability for the center pixel. In other words, it can aggregate patch-based contextual information related to the pixel to be classified according to the spatial attention map $\boldsymbol{S}'$.

By embedding the spatial self-attention module before the concatenate operation, as shown in Fig. 4(a), the spatial

attention dense block can be constructed. It can be expressed as follows:

$$\boldsymbol{y} = [\boldsymbol{x}_0, f(\boldsymbol{x}_0), f([\boldsymbol{x}_0, f(\boldsymbol{x}_0)])] \tag{7}$$

where $\boldsymbol{x}_0$ and $\boldsymbol{y}$ denote the input and output features of the spatial attention dense block, respectively. $[\cdot]$ refers to the concatenation operation, and $f(\cdot)$ denotes the operation of composite function, including BN-ReLU-Conv1 $\times$ 1, BN-ReLU-Conv3 $\times$ 3, and the spatial self-attention module. Note that all the output features of the attention module are passed to subsequent units, which can not only alleviate the vanishing gradient but also strengthen feature representations effectively.

The architecture of the spatial subnetwork is shown in Part 1 of Fig. 3. It consists of a 3 $\times$ 3 Conv layer, three spatial attention dense blocks, two transition layers, and a global average pooling layer. The output spatial features of this subnetwork are fed into Part 3 of Fig. 3 for fusion and data classification.

C. Spectral Self-Attention Module-Based Spectral Subnetwork

With abundant spectral information in HSI, there are inevitably some correlations between spectral bands. Convolution kernels in 1-D CNN can only represent a local cross-channel interaction, i.e., it cannot explore the long-range channel correlation. A few studies [59], [65], [66] used the spectral attention mechanism to encode the long-range dependencies to improve the spectral feature representation, while they were designed for 2-D CNN. In this section, we designed a spectral self-attention module, aiming to capture long-range spectral correlations of 1-D CNN over the local spectral features. It uses the cosine similarity to exploit the interdependencies between channels, improving the spectral feature representation.

The process of the spectral self-attention module is shown in Fig. 5(b). The input spectral feature vectors $\boldsymbol{Y} \in \mathbb{R}^{l\times f}$, where l is the length of the spectral feature vectors and f is the number of channels, equaling to the number of filters in the Conv layer. Considering that the spectral self-attention module needs to calculate the relationship between different channels, we directly performed a similarity calculation between any two channels in $\boldsymbol{Y}$ to maintain this relationship as follows:

$$\boldsymbol{Q}_{u,v} = \mathrm{sim}(\boldsymbol{Y}_u, \boldsymbol{Y}_v) = \left(\frac{\boldsymbol{Y}_u \boldsymbol{Y}_v}{\|\boldsymbol{Y}_u\|\|\boldsymbol{Y}_v\|}\right), \quad u, v = 1, 2, \ldots, l \tag{8}$$

where $\boldsymbol{Q}_{u,v}$ measures the correlation between the uth channel and the vth channel. Then, we use the softmax function to normalize each column of $\boldsymbol{Q} \in \mathbb{R}^{l\times l}$ to obtain the spectral attention probability map $\boldsymbol{Q}$

$$\boldsymbol{Q}'_{u,v} = \frac{\exp(\boldsymbol{Q}_{u,v})}{\sum_{u=1}^{l}\exp(\boldsymbol{Q}_{u,v})}, \quad u, v = 1, 2, \ldots, l. \tag{9}$$

A weighted matrix is obtained by

$$\boldsymbol{W}_{\mathrm{spe}} = \boldsymbol{Y} \otimes \boldsymbol{Q}' \tag{10}$$

where $\boldsymbol{W}_{\mathrm{spe}} \in \mathbb{R}^{l\times f}$ and $\otimes$ denotes the element-wise product. It can be deduced from (10) that the features at each channel are the weighted sum of the features at all channels. Finally, a residual connection is performed to obtain the final output

$$\boldsymbol{V}_{\mathrm{spe}} = \boldsymbol{W}_{\mathrm{spe}} + \boldsymbol{Y} \tag{11}$$

where $\boldsymbol{V}_{\mathrm{spe}} \in \mathbb{R}^{l\times f}$. It can model the long-range spectral correlations between spectral channels, boosting spectral feature discriminability.

Similar to the spatial self-attention module, we insert the spectral self-attention module before the concatenate operation in the spectral attention dense block, as shown in Fig. 5(a). In addition, we can see from Part 2 of Fig. 3 that the setting of the spectral subnetwork is the same as the spatial subnetwork, except that all the Conv and pooling operations in this subnetwork adopt 1-D computation. The output spectral features of this subnetwork are also fed into Part 3 of Fig. 3 for fusion and data classification.

D. Weight Fusion and Classification

Considering that the obtained spatial and spectral features are in two separate domains, we adopt the "score weighted" fusion method in [39] to perform the classification. It can be simply understood that the final score vector is obtained by a weighted sum of the spatial and spectral scores. As shown in Part 3 of Fig. 3, the output features of each subnetwork are fed to an FC layer. Note that the number of neurons in the FC layer is equal to the number of classes, and the value of each neuron can be regarded as a class-specific response. The outputs of FC layers corresponding to the spatial subnetwork and the spectral subnetwork are expressed as $\boldsymbol{F}_{\mathrm{spa}} \in \mathbb{R}^K$ and $\boldsymbol{F}_{\mathrm{spe}} \in \mathbb{R}^K$, respectively, where K is the number of classes. Then, the fused probability in different classes is computed as

$$\boldsymbol{F} = \sigma\left(\lambda \times \boldsymbol{F}_{\mathrm{spa}} + (1-\lambda) \times \boldsymbol{F}_{\mathrm{spe}}\right) \tag{12}$$

where $\sigma(\cdot)$ denotes the softmax function. λ is a weighting parameter in the range of [0, 1], which is initialized to 0.5 and then adaptively and automatically adjusted during the process of the network optimization. Experiments and validations are presented and discussed in Section IV.

IV. Experiment and Discussion

A. Description of the Datasets

We carried out experiments on four datasets: University of Pavia (PU), Salinas (SA), Kennedy Space Center (KSC), and University of Houston (UH). Details of these datasets are given as follows.

The PU dataset was taken over the University of Pavia, Northern Italy, by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. It has 610 $\times$ 340 pixels with a spatial resolution of 1.3 m, composed of 115 bands covering the wavelength from 0.43 to 0.86 μm. After discarding 12 noisy and water absorption bands, only 103 bands were preserved. As summarized in Table I, there are nine land-cover classes, and the training and testing samples with the same settings as [39] are also given.

The SA image was recorded by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the area of Salinas Valley, CA, USA. The image size is 512 $\times$ 217 with

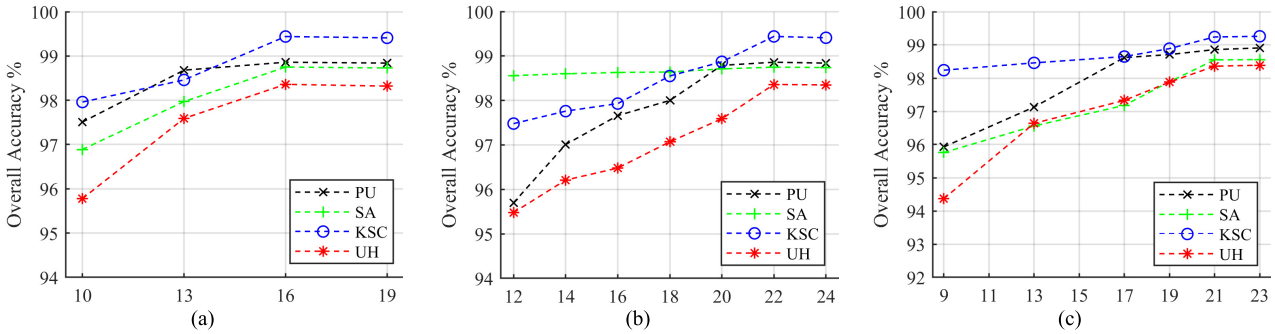


Fig. 6. Effect of (a) number of Conv layers, (b) growth rate k, and (c) patch size on the classification accuracies in the four datasets.

TABLE I
DETAILS OF THE LAND-COVER TYPES AND THE NUMBER OF SAMPLES FOR THE PU DATASET

ID	Color	Land Cover Type	Train	Test	Total
1		Asphalt	100	6531	6631
2		Meadows	100	18549	18649
3		Gravel	100	1999	2099
4		Trees	100	2964	3064
5		Painted metal sheets	100	1245	1345
6		Bare Soil	100	4929	5029
7		Bitumen	100	1230	1330
8		Self-Blocking Bricks	100	3582	3682
9		Shadows	100	847	947
		Total	900	4187	42776

TABLE II
DETAILS OF THE LAND COVER TYPES AND THE NUMBER OF SAMPLES FOR THE SA DATASET

ID	Color	Land Cover Type	Train	Test	Total
1		Brocoli_green_weeds_1	100	1909	2009
2		Brocoli_green_weeds_2	100	3626	3726
3		Fallow	100	1876	1976
4		Fallow_rough_plow	100	1294	1394
5		Fallow_smooth	100	2578	2678
6		Stubble	100	3859	3959
7		Celery	100	3479	3579
8		Grapes_untrained	100	11171	11271
9		Soil_vinyard_develop	100	6103	6203
10		Corn_senesced_green	100	3178	3278
11		Lettuce_romaine_4wk	100	968	1068
12		Lettuce_romaine_5wk	100	1827	1927
13		Lettuce_romaine_6wk	100	816	916
14		Lettuce_romaine_7wk	100	970	1070
15		Vinyard_untrained	100	7168	7268
16		Vinyard_vertical_trellis	100	1707	1807
		Total	1600	52529	54129

a spatial resolution of 3.7 m, composed of 224 bands ranging from 0.36 to 2.5 μm. Before the experiments, 204 valid bands were retained after removing 20 water absorption and noise bands. The ground truth of this scene consists of 16 classes, and 100 samples per class were selected to train the networks and the remaining were used for testing, as listed in Table II.

The UH image was gathered by an airborne sensor, which covers the area of University of Houston. It has 349 $\times$ 1905 pixels with a spatial resolution of 2.5 m and consists of 144 spectral channels ranging from 0.38 to 1.05 μm. These data include 15 land-cover classes. It adopted the standard training and testing sets given by the 2013 GRSS Data Fusion Contest. Details of these classes and the number of training and testing samples in each class are shown in Table III.

The KSC dataset covers an area of KSC, FL, USA, which was also gathered by the AVIRIS sensor. It consists of 512 $\times$ 614 pixels and 176 spectral bands after removing water absorption and low SNR bands. It has a spatial resolution of 18 m and a spectral resolution of 10 nm ranging from 0.4 to 2.5 μm. Details of the land-cover types and the number of training and testing samples in each class are listed in Table IV, which are the same as in [39].

In deep learning, data normalization can unify data magnitude, promote network convergence, and prevent gradient explosion. Therefore, the HSI datasets were normalized to [0, 1] by using the min-max normalization before the training and testing in the following experiments. In addition, these datasets all used the same data argumentation strategy as in [37].

B. Experiment Setting

To verify the performance of the proposed method, we conducted a series of experiments on these four datasets. First, we analyzed the impact of different hyperparameters on classification performance. Second, we evaluated the effects of the proposed spectral self-attention module and spatial self-attention module. We also compared the proposed network with other state-of-the-art CNN-related methods. All the experiments were implemented on Ubuntu 16.04 and a GPU of Nvidia GeForce RTX 2080. The classification performance was measured by four common quantitative metrics: the producer accuracy (PA) of each class, overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa). PA measures the percentage of correctly classified pixels for a certain class, which can be derived for the training dataset or the testing dataset. AA is the average of the PA over all the classes. OA represents the overall percentage of correctly classified pixels for the whole dataset, including all classes,

TABLE III
DETAILS OF THE LAND COVER TYPES AND THE NUMBER OF SAMPLES FOR THE UH DATASET

ID	Color	Land Cover Type	Train	Test	Total
1		Healthy grass	198	1053	1251
2		Stressed grass	190	1064	1254
3		Artificial turf	192	505	697
4		Evergreen trees	188	1056	1244
5		Deciduous trees	186	1056	1242
6		Bare earth	182	143	325
7		Water	196	1072	1268
8		Residential buildings	191	1053	1244
9		Non-residential buildings	193	1059	1252
10		Roads	191	1036	1227
11		Sidewalks	181	1054	1235
12		Crosswalks	192	1041	1233
13		Major thoroughfares	184	285	469
14		Highways	181	247	428
15		Railways	187	473	660
		Total	28332	12197	15029

TABLE IV
DETAILS OF THE LAND COVER TYPES AND THE NUMBER OF SAMPLES FOR THE KSC DATASET

ID	Color	Land Cover Type	Train	Test	Total
1		Scrub	77	684	761
2		willow swamp	25	218	243
3		CP hammock	26	230	256
4		CP/Oak	26	226	252
5		Slash pine	17	144	161
6		Oak/broadleaf	23	206	229
7		Hardwood swam	10	95	105
8		Graminoid marsh	44	387	431
9		patina marsh	52	468	520
10		Cattail marsh	41	363	404
11		Salt marsh	42	377	419
12		Mud flats	51	452	503
13		Water	93	834	927
		Total	527	4684	5211

for either the training or testing dataset. Kappa coefficient is a score that measures the level of agreement between the classification results and the corresponding ground truth (GT). Its value ranges from -1 to 1, and the larger the value, the higher level of agreement. To avoid biased estimation, all experiments were conducted with five independent tests, and the average values were reported for all the evaluation metrics.

C. Parameter Setting

For PCA-based dimensionality reduction, the numbers of the preserved principal components are 3, 4, 2, and 75 for the SA data, the UP data, the UH data, and the KSC data, respectively. This was determined by retaining at least 99% of the total data variation in the original HSI. We trained the network for 25 epochs with a batch size of 100 and a learning rate of 0.0001. The proposed networks were carried out using the Keras framework with TensorFlow as the backend. Besides, we also analyze some key hyperparameters on the classification performance. Details are presented as follows.

1) Effect of the Number of Conv Layers: Fig. 6(a) shows the effect of the number of Conv layers (i.e., the depth of the network) on the OA of the proposed network. Here, the number of Conv layers is calculated within each subnetwork, excluding those in the self-attention modules. Deeper networks generally have more powerful feature representation ability, but too deep networks will cause gradient instability and network degradation. In Fig. 6(a), it is clear that 16 Conv layers lead to the best results on these four datasets. After that, the network performance remains unchanged or decreases slightly. Therefore, in the following experiments, the number of Conv layers is set to 16 for all datasets.

2) Effect of the Growth Rate k: Fig. 6(b) shows the performance on different growth rates k, which determines the width of the network. Increasing the width of the network enables each Conv layer to learn richer features and obtain better performance. However, due to the increased number of parameters, it will increase the possibility of overfitting. From Fig. 6(b), we can see that the OA reaches its peak at 22 on the PU, SA, and UH datasets. Although the OA of the KSC dataset still increases when k exceeds 22, the increase is minor. Therefore, for convenience, the growth rate is uniformly set to 22 for all the datasets.

3) Effect of the Input Patch Size: We also investigated the effect of patch size on the classification performance, which is shown in Fig. 6(c). It can be seen that the OA shows an upward trend, while over the size of 21×21, the rise is minor. This is because larger patches contain more spatial information, which is conducive to classification. However, when the patch size is too large, it may contain some negative information. On the other hand, a large patch will increase the computational load. Therefore, we set the patch size to 21×21 for all the datasets.

D. Contribution of the Self-Attention Modules

In this section, we conducted a series of tests to analyze the contribution of the proposed spatial self-attention module and spectral self-attention module. We separately tested the spatial subnetwork and spectral subnetwork on a different number of training samples, where 50, 100, and 150 labeled samples per class were randomly selected from the PU, SA, and UH datasets, and 5%, 10%, and 15% samples per class were randomly selected from the KSC dataset, respectively.

1) Contribution of the Spatial Self-Attention Module: To verify the effectiveness of the proposed spatial self-attention module, we compared the classification performance of the spatial subnetwork with and without the spatial self-attention module (denoted as Spa-A and Spa, respectively). Fig. 7 shows the results on a different number of training data.

According to Fig. 7, it is clear that employing the spatial self-attention module can consistently improve the performance with lower standard deviation. The spatial self-attention module can promote the discriminant feature learning ability of the network, especially with limited training samples. As shown in Fig. 7, the fewer samples, the more significant the superiority of the Spa-A. This is because the proposed

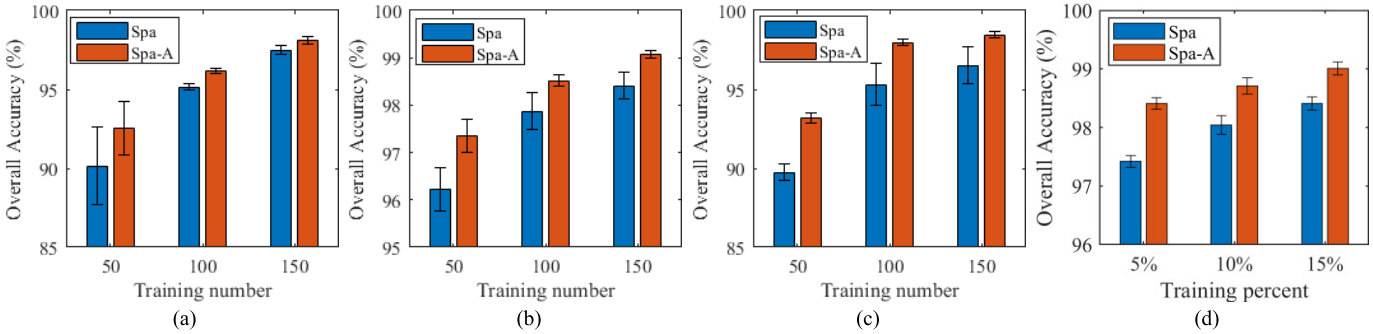


Fig. 7. Effects of the spatial self-attention module on the performance of the proposed spatial subnetwork. (a) PU. (b) SA. (c) UH. (d) KSC.

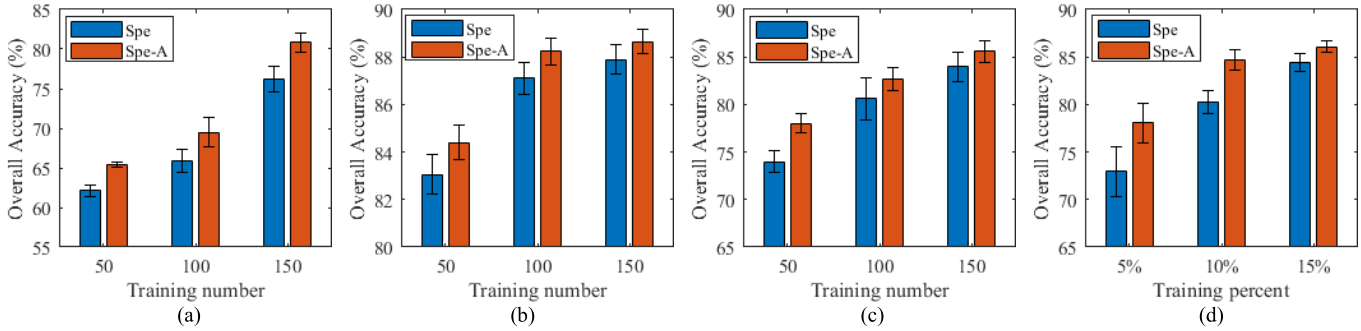


Fig. 8. Effects of the spectral self-attention module on the performance of the proposed spectral subnetwork. (a) PU. (b) SA. (c) UH. (d) KSC.

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT METHODS ON THE PU DATASET

Metrics	CDCNN	SSRN	FDSSC	LSMSC	ASSMN	DBMA	DBDA	SSSAN
OA (%)	91.68 ± 1.81	97.42 ± 1.43	98.27 ± 0.71	98.00 ± 0.91	96.33 ± 1.25	98.05 ± 0.24	98.12 ± 0.36	**98.86** ± 0.18
AA (%)	89.92 ± 2.12	96.55 ± 2.16	98.12 ± 0.80	98.56 ± 0.46	97.22 ± 0.36	97.78 ± 0.36	97.38 ± 0.24	**99.32** ± 0.35
Kappa×100	89.06 ± 2.33	96.58 ± 1.89	97.70 ± 0.94	97.33 ± 0.78	95.11 ± 1.06	97.41 ± 0.31	97.52 ± 0.67	**98.48** ± 0.25
1	96.88	99.80	99.79	98.99	94.12	99.28	98.77	98.15
2	97.52	99.86	99.92	99.92	96.78	99.69	99.79	98.60
3	82.68	83.07	97.07	95.11	93.80	97.77	99.50	100
4	91.63	98.85	98.37	99.04	99.43	96.87	98.77	99.22
5	99.50	100	100	99.75	100	99.92	99.77	100
6	77.20	93.96	91.21	94.76	93.35	91.42	99.17	99.33
7	79.39	99.14	99.78	90.55	100	99.92	97.36	99.76
8	88.85	94.36	97.15	93.82	97.54	96.76	87.72	98.97
9	95.60	99.96	99.76	99.87	100	98.33	95.56	99.88

spatial self-attention module can compensate for information from a small training set by effectively capturing useful spatial information related to the pixels to be classified. It is demonstrated that the proposed spatial self-attention module can enhance the spatial feature representation of the network.

2) Contribution of the Spectral Self-Attention Module: Similarly, to demonstrate the effectiveness of the proposed spectral self-attention module, we tested and compared the spectral subnetwork with and without the spectral self-attention module (denoted as Spe-A and Spe, respectively). Experimental results are shown in Fig. 8.

As can be seen from Fig. 8, using the spectral self-attention module can improve the performance remarkably. In these four datasets, Spe-A has a consistent improvement over the Spe on a different number of training samples. As shown in Fig. 8(c) and (d), Spe-A is able to reach a better OA with lower standard deviation on UH and KSC datasets, especially with fewer training samples. It is demonstrated that using the proposed spectral self-attention module has great benefits to 1-D CNN for spectral feature extraction.

E. Comparison With State-of-the-Art Methods

To evaluate the performance of the proposed method for HSI classification, we compared our method with other existing state-of-the-art CNN-related methods, such as contextual CNN (CCNN) [33], SSRN [42], FDSSC [43], localized spectral features and multiscale spatial features network (LSMSC) [67], adaptive spectral–spatial multiscale network (ASSMN) [39], double-branch multiattention mechanism network (DBMA) [49], and double-branch dual-attention

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT METHODS ON THE SA DATASET

Metrics	CDCNN	SSRN	FDSSC	LSMSC	ASSMN	DBMA	DBDA	SSSAN
OA (%)	86.80 ± 5.73	96.62 ± 0.98	97.73 ± 1.43	98.17 ± 0.43	98.44 ± 0.36	98.34 ± 0.39	98.41 ± 0.45	**98.75** ± 0.28
AA (%)	93.54 ± 2.27	98.49 ± 0.38	99.03 ± 0.25	99.08 ± 0.12	**99.36** ± 0.05	99.26 ± 0.27	98.77 ± 0.07	99.32 ± 0.16
$\kappa \times 100$	85.41 ± 6.22	96.23 ± 1.08	97.51 ± 1.58	97.96 ± 0.42	98.26 ± 0.26	98.15 ± 0.42	98.23 ± 0.51	**98.60** ± 0.19
1	88.61	100	100	100	100.00	100	100	100
2	98.39	99.98	99.96	99.69	100.00	100	100	99.94
3	97.69	100	100	96.26	99.89	100	100	99.89
4	98.77	99.52	99.20	98.01	100	96.12	93.98	99.69
5	97.58	99.63	96.44	99.87	99.30	100	99.92	97.67
6	99.13	100	99.95	100	100	99.97	99.87	99.87
7	99.16	100	100	99.73	100	100	100	99.89
8	87.46	95.57	92.86	98.71	95.51	93.91	99.19	97.72
9	99.79	100	99.88	99.70	100	99.87	100	99.48
10	91.30	97.45	98.19	99.64	99.62	99.97	92.24	99.78
11	88.60	98.24	98.47	91.66	100	100	100	100
12	99.34	99.87	99.90	99.94	100	100	100	100
13	95.29	99.75	100	100	100	99.87	99.88	99.14
14	96.18	99.93	99.79	99.43	100	99.89	99.90	99.90
15	63.25	85.83	97.20	92.81	96.21	98.60	94.33	96.12
16	96.03	100	100	98.34	99.30	100	100	100

TABLE VII
COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT METHODS ON THE UH DATASET

Class	CDCNN	SSRN	FDSSC	LSMSC	ASSMN	DBMA	DBDA	SSSAN
OA (%)	78.34 ± 1.42	82.07 ± 3.22	86.45 ± 1.61	81.19 ± 2.05	67.75 ± 3.42	82.23 ± 2.03	85.02 ± 1.98	**87.47** ± 1.56
AA (%)	81.42 ± 1.64	85.74 ± 2.43	88.60 ± 0.66	83.88 ± 1.26	66.92 ± 3.67	86.08 ± 1.96	86.41 ± 1.38	**88.73** ± 1.76
Kappa×100	76.64 ± 1.55	80.58 ± 3.46	85.36 ± 1.76	79.68 ± 2.12	65.06 ± 3.74	80.79 ± 2.23	83.78 ± 2.12	**86.46** ± 1.23
1	98.14	96.79	98.86	99.42	79.49	99.43	98.81	99.14
2	93.44	90.54	100	99.71	69.77	96.53	100	99.57
3	29.26	36.92	100	100	51.49	30.63	99.74	95.8.3
4	99.72	99.12	99.89	99.86	83.90	98.66	99.31	90.17
5	97.56	100	98.16	100	89.68	97.91	100	97.68
6	30.56	91.74	100	15.66	93.24	67.86	40.24	33.89
7	78.44	89.60	91.02	64.69	80.07	93.99	90.12	98.99
8	87.83	73.37	80.52	64.69	50.30	88.57	92.39	74.45
9	82.61	93.91	46.28	83.81	69.81	70.95	88.01	87.75
10	86.45	63.15	90.89	84.66	35.88	82.41	77.29	90.61
11	67.65	88.69	96.51	97.58	75.27	87.43	51.47	83.06
12	87.64	89.17	96.87	54.05	68.43	85.73	86.17	94.73
13	89.30	97.96	96.89	86.81	37.43	91.62	88.27	97.33
14	96.74	75.20	100	72.49	95.68	100	100	90.82
15	95.93	100	96.05	99.45	23.40	99.45	94.55	99.46

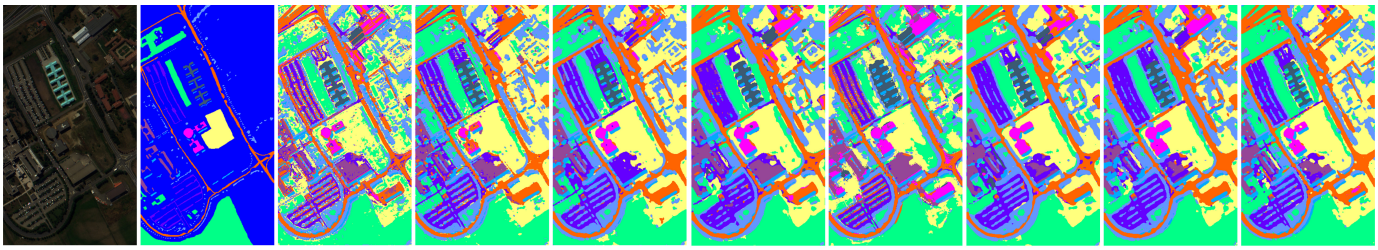
mechanism network (DBDA) [52]. Specifically, CCNN is a traditional spectral–spatial network, and SSRN and FDSSC are spectral–spatial networks based on ResNet and DenseNet, respectively. LSMSC and ASSMN are spectral–spatial multi-scale networks, and DBMA and DBDA are spectral–spatial attention networks. All these methods were implemented using the open-source code with their optimal parameters as described in the corresponding references. Besides, for a fair comparison, all the methods were trained and tested on the same sample sets, as listed in Tables I–IV.

1) Quantitative Evaluation: Quantitative results of OA, AA, Kappa, and PA of each class are listed in Tables V–VIII. We can see that the proposed method achieved higher classification accuracy with lower standard deviation compared with other methods. Taking Table V for example, the proposed method achieved the highest accuracy of 98.86%, which exceeds CCNN, SSRN, FDSSC, LSMSC, ASSMN, DBMA, and DBDA by 7.18%, 1.44%, 0.59%, 0.86%, 2.53%, 0.81%, and 0.74%, respectively. Although the AA of SSSAN is slightly lower than that of ASSMN in Table VI, it achieved higher accuracy in OA and Kappa. Besides, in Tables VII and VIII, the proposed one also achieved the highest accuracy in OA, AA, and Kappa.

It can be seen from Tables VI–VIII that the accuracy for CCNN is lower than those of other methods since it only uses a weak 2-D CNN to extract spectral and spatial features. Compared with CCNN, the accuracy of SSRN is significantly

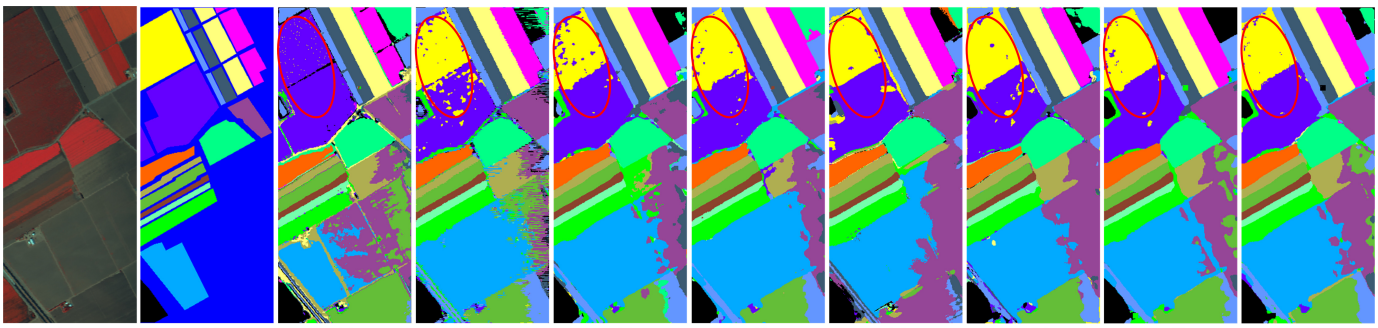
TABLE VIII
COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT METHODS ON THE KSC DATASET

Metrics	CDCNN	SSRN	FDSSC	LSMSC	ASSMN	DBMA	DBDA	SSSAN
OA (%)	87.78 ± 5.58	97.68 ± 2.41	98.40 ± 0.89	98.10 ± 0.51	98.44 ± 0.92	98.23 ± 0.42	98.59 ± 0.53	**99.44** ± 0.13
AA (%)	78.21 ± 9.84	95.79 ± 2.03	97.76 ± 0.87	96.47 ± 0.30	98.00 ± 1.03	96.37 ± 0.56	97.78 ± 0.49	**98.89** ± 0.12
Kappa×100	86.38 ± 6.22	96.72 ± 2.69	98.22 ± 1.0	98.30 ± 0.56	98.27 ± 1.02	98.03 ± 0.46	98.42 ± 0.56	**99.38** ± 0.15
1	97.17	99.12	99.67	97.14	97.09	100	100	100
2	77.70	100	97.20	100	96.91	100	99.78	98.62
3	65.20	100	98.15	98.26	93.88	97.73	96.60	99.57
4	59.37	74.18	86.59	98.62	93.25	86.99	92.86	92.92
5	36.80	100	96.24	77.71	98.15	95.38	82.96	95.14
6	78.89	96.97	98.47	99.51	97.73	97.61	96.89	100
7	49.93	65.07	96.53	83.05	98.85	86.36	96.63	100
8	79.84	97.69	99.32	97.94	98.98	96.18	97.15	99.48
9	92.38	100	99.77	99.76	99.72	100	100	100
10	84.14	100	99.69	100	99.97	100	100	100
11	99.70	99.21	99.51	99.81	100	100	100	100
12	95.90	100	99.74	99.86	99.47	99.10	98.96	100
13	99.74	100	100	100	100	100	100	99.88



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

Fig. 9. Classification maps of different methods on PU dataset. (a) FCM. (b) GT. (c) CDCNN. (d) SSRN. (e) FDSSC. (f) LSMSC. (g) ASSMN. (h) DBMA. (i) DBDA. (j) SSSAN.



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

Fig. 10. Classification maps of different methods on SA dataset. (a) FCM. (b) GT. (c) CDCNN. (d) SSRN. (e) FDSSC. (f) LSMSC. (g) ASSMN. (h) DBMA. (i) DBDA. (j) SSSAN.

improved, due to its use of spectral and spatial residual blocks to consecutively learn spectral and spatial features. FDSSC uses densely connected structures to deeply learn features, obtaining better results than SSRN. To enhance feature learning, LSMSC fuses localized spectral features and multiscale spatial features by considering the correlations between different bands. ASSMN employs a multiscale strategy in spectral and spatial simultaneously. However, FDSSC, LSMSC, and ASSMN cannot achieve good results on some datasets. For example, ASSMN achieved good performance on the SA and KSC datasets, but its accuracy is very low on the PU and UH datasets, especially on the UH dataset. DBMA and DBDA use attention mechanisms, achieving stable results on all these four datasets. Comparatively, DBDA generates better performance compared to DBMA. Furthermore, the proposed method constantly performs better than DBDA on all datasets because it is based on powerful baseline of DenseNet and the proposed self-attention modules. Overall, the proposed method provides better performance on all these four datasets.

2) Qualitative Evaluation: The corresponding classification maps alongside false-color maps (FCMs) and GT are shown in Figs. 9–12. These maps are consistent with the quantitative results listed in Tables V–VIII. The classification maps obtained by our method have the least noise and the clearest

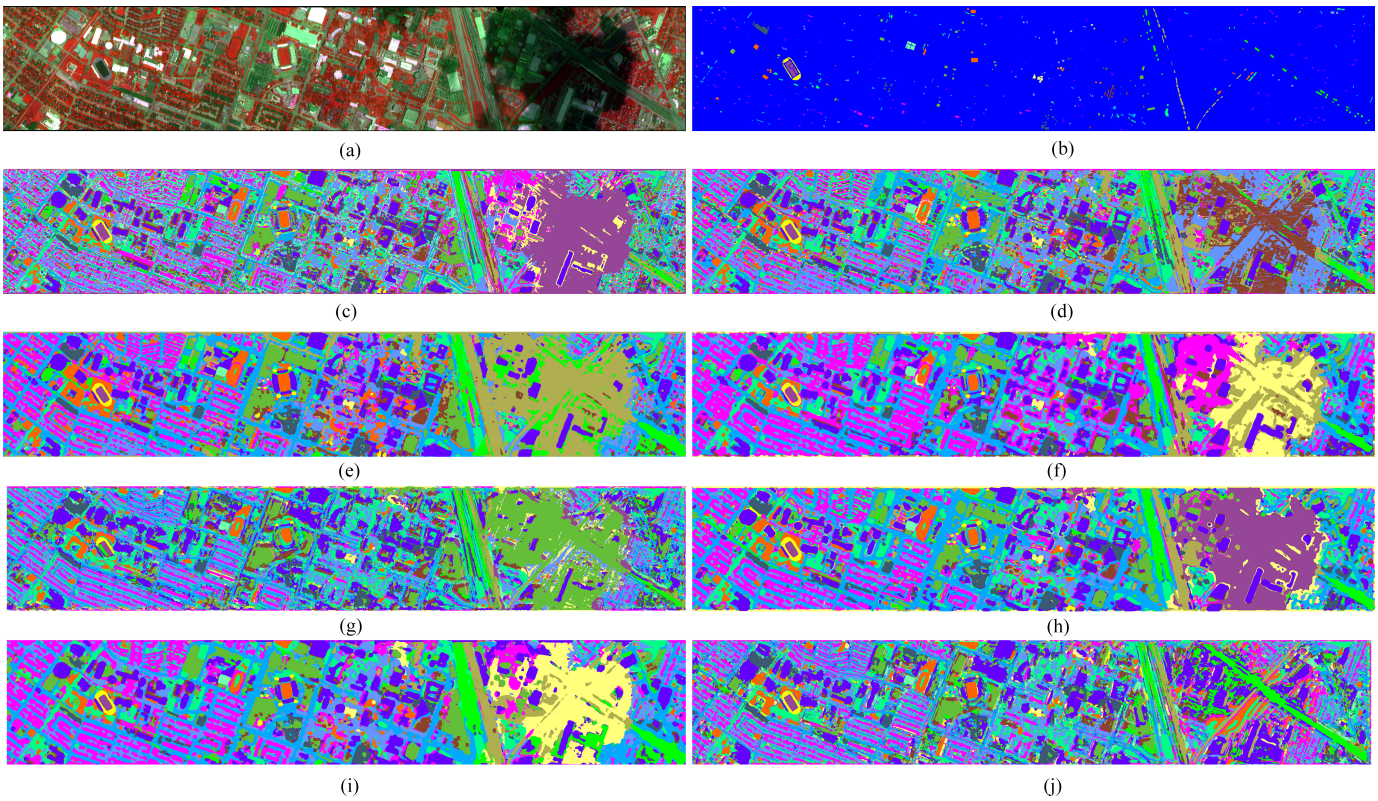


Fig. 11. Classification maps of different methods on UH dataset. (a) FCM. (b) GT. (c) CDCNN. (d) SSRN. (e) FDSSC. (f) LSMSC. (g) ASSMN. (h) DBMA. (i) DBDA. (j) SSSAN.

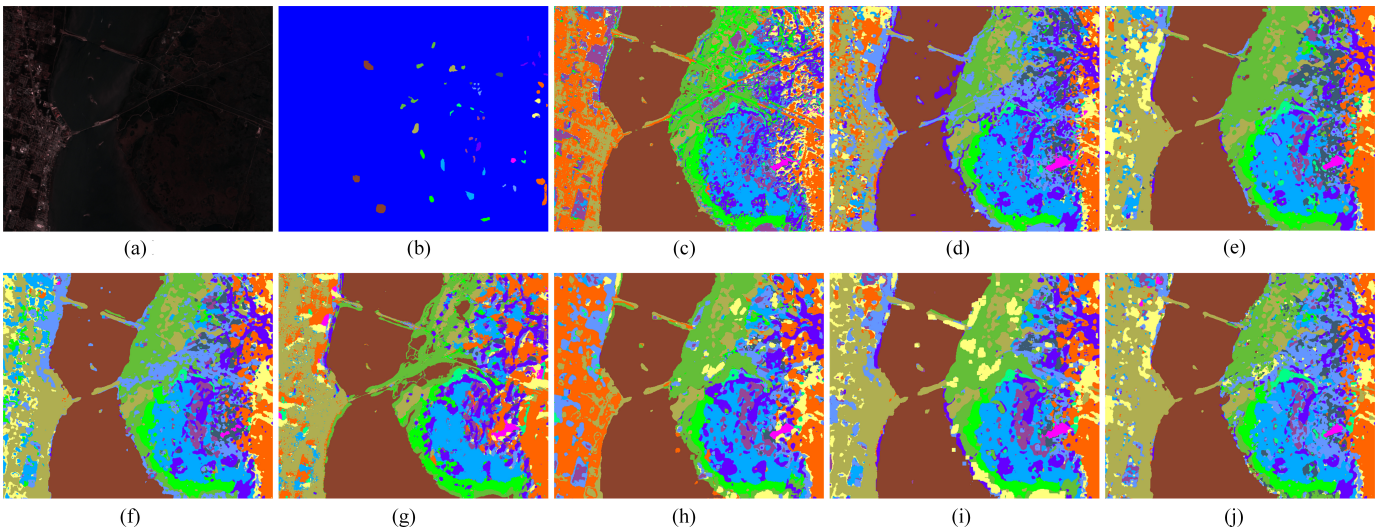


Fig. 12. Classification maps of different methods on KSC dataset. (a) FCM. (b) GT. (c) CDCNN. (d) SSRN. (e) FDSSC. (f) LSMSC. (g) ASSMN. (h) DBMA. (i) DBDA. (j) SSSAN.

object boundary, which is very close to the GT maps. The proposed method can correctly label almost all classes, even some easily confused classes, such as Grapes_untrained and Vinyard_untrained in Fig. 10, which are marked with red circles.

3) Analyses of Running Time: To measure the efficiency of the proposed method, we compared our method with the other seven methods tested in terms of training and test time on PU, SA, UH, and KSC datasets. The results are listed in Table IX. It can be seen that the training time of the proposed method is shorter than SSRN, DBDA, FDSSC, ASSMN, and DBMA. The possible reason is that the proposed methods use the 1-D CNN and 2-D CNN to learn spectral and spatial features, respectively, while others use 3-D CNN with a large number of parameters. On the other hand, the proposed method converged faster than this 3-D CNN-based method (e.g., 25 epochs for the proposed method and about 50–100 epochs for the SSRN, DBDA, FDSSC, and DBMA). Since ASSMN uses the ConvLSTM with multitime step calculation and multibranch architectures, it takes much longer than other methods.

TABLE IX
COMPARISON OF TRAINING AND TESTING TIME OF DIFFERENT METHODS ON THE PU, SA, UH, AND KSC DATASETS

Dataset	Time	CDCNN	SSRN	FDSSC	LSMSC	ASSMN	DBMA	DBDA	SSSAN
PU	T_{train} (s)	81.89	132.33	185.21	112.95	1340.06	186.85	164.27	120.82
	T_{test} (s)	14.08	58.70	19.46	16.69	306.99	24.13	28.22	22.51
SA	T_{train} (s)	95.55	329.56	628.20	175.64	2387.10	797.30	274.12	203.51
	T_{test} (s)	18.09	76.27	30.52	20.89	354.38	51.11	45.68	32.51
UH	T_{train} (s)	156.98	307.17	573.75	221.54	4158.01	321.54	314.35	254.71
	T_{test} (s)	4.10	4.56	4.74	7.64	88.38	7.56	7.52	6.57
KSC	T_{train} (s)	48.14	147.88	239.33	89.20	780.69	155.65	153.95	117.45
	T_{test} (s)	1.72	2.47	3.79	4.76	33.43	4.59	4.16	3.86

CCNN uses the shortest time because it is simple 2-D CNN architecture with less training parameters. LSMSC is the second shortest, which uses the band grouping strategy to reduce the computation burden. Although the running time of CCNN and LSMSC is shorter, their performance is lower than our methods.

V. CONCLUSION

In this article, a novel spectral–spatial self-attention CNN architecture is proposed for HSI classification. First, based on the proposed spatial self-attention module, the spatial subnetwork has significantly enhanced the patch-based relevant long-range contextual information related to the center pixel while suppressing unnecessary one, improving the accuracy for the center pixel recognition. Meanwhile, based on the proposed spectral self-attention module, the spectral subnetwork has successfully extracted more discriminative spectral features by exploiting the long-range spectral correlations over local spectral features.

The weighted fusion of the extracted spectral and spatial features can further improve the classification accuracy. The proposed method is found to outperform a number of state-of-the-art methods, including CCNN, SSRN, LSMSC, ASSMN, DBMA, and DBDA. Future work includes further optimization of the network for fast parameter selection.

REFERENCES

[1] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[2] B. Zhang, D. Wu, L. Zhang, Q. Jiao, and Q. Li, "Application of hyperspectral remote sensing for environment monitoring in mining areas," *Environ. Earth Sci.*, vol. 65, no. 3, pp. 649–658, Feb. 2012.

[3] M. Govender, K. Chetty, and H. Bulcock, "A review of hyperspectral remote sensing and its application in vegetation and water resource studies," *Water SA*, vol. 33, no. 2, pp. 145–152, May 2007.

[4] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[5] J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2020, pp. 813–815.

[6] M. Pal, "Multinomial logistic regression-based feature selection for hyperspectral data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 14, no. 1, pp. 214–220, Feb. 2012.

[7] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[8] C. Rodarmel and J. Shan, "Principal component analysis for hyperspectral image classification," *Surv. Land Inf. Syst.*, vol. 62, no. 2, pp. 115–122, Jun. 2002.

[9] X. Zheng, Y. Yuan, and X. Lu, "Dimensionality reduction by spatial–spectral preservation in selected bands," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5185–5197, Sep. 2017.

[10] L. Drumetz, M.-A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3890–3905, Aug. 2016.

[11] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[12] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, "Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 9, 2021, doi: 10.1109/TGRS.2021.3069845.

[13] J. Plaza, A. Plaza, and C. Barra, "Multi-channel morphological profiles for classification of hyperspectral images using support vector machines," *Sensors*, vol. 9, no. 1, pp. 196–218, Jan. 2009.

[14] M. Shi and G. Healey, "Hyperspectral texture recognition using a multiscale opponent representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1090–1095, May 2003.

[15] R. Ji, Y. Gao, R. Hong, Q. Liu, D. Tao, and X. Li, "Spectral–spatial constraint hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1811–1824, Mar. 2014.

[16] B. Zhang, S. Li, X. Jia, L. Gao, and M. Peng, "Adaptive Markov random field approach for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, Sep. 2011.

[17] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Apr. 2013.

[18] Y. Y. Tang, Y. Lu, and H. Yuan, "Hyperspectral image classification based on three-dimensional scattering wavelet transform," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2467–2480, May 2015.

[19] Z. Zhu, S. Jia, S. He, Y. Sun, Z. Ji, and L. Shen, "Three-dimensional Gabor feature extraction for hyperspectral imagery classification using a memetic framework," *Inf. Sci.*, vol. 298, pp. 274–287, Mar. 2015.

[20] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1951–1959.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2012, pp. 1097–1105.

[23] G. Lokman, H. H. Çelik, and V. Topuz, "Hyperspectral image classification based on multilayer perceptron trained with eigenvalue decay," *Can. J. Remote Sens.*, vol. 46, no. 3, pp. 253–271, May 2020.

[24] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[25] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[26] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[27] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 1–20, Jul. 2016.

[28] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 1–13, Aug. 2020.

[29] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[30] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[31] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7048–7066, Sep. 2019.

[32] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[33] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[34] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.

[35] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, Jan. 2017.

[36] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018.

[37] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[38] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2349–2361, Apr. 2018.

[39] D. Wang, B. Du, L. Zhang, and Y. Xu, "Adaptive spectral–spatial multiscale contextual feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2461–2477, Mar. 2021.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2016, pp. 770–778.

[41] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 2261–2269.

[42] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[43] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, p. 1068, Jul. 2018.

[44] F. Cao and W. Guo, "Cascaded dual-scale crossover network for hyperspectral image classification," *Knowl.-Based Syst.*, vol. 189, Feb. 2020, Art. no. 105122.

[45] Z. Meng, L. Li, L. Jiao, Z. Feng, X. Tang, and M. Liang, "Fully dense multiscale fusion network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 22, p. 2718, Nov. 2019.

[46] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2015, pp. 429–439.

[47] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 121–130, Jan. 2021.

[48] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[49] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, p. 1307, Jun. 2019.

[50] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2647–2650.

[51] H. Huang, C. Pu, Y. Li, and Y. Duan, "Adaptive residual convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2520–2531, May 2020, doi: 10.1109/JSTARS.2020.2995445.

[52] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020.

[53] Z. Dong, Y. Cai, Z. Cai, X. Liu, Z. Yang, and M. Zhuge, "Cooperative spectral–spatial attention dense network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 866–870, May 2021, doi: 10.1109/LGRS.2020.2989437.

[54] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[56] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sens.*, vol. 11, no. 2, p. 159, 2019.

[57] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Sep. 2020.

[58] J. Li *et al.*, "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4304–4318, Jun. 2020.

[59] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[60] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[61] C. Zhang, J. Yue, and Q. Qin, "Global prototypical network for few-shot hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4748–4759, Aug. 2020.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[63] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.

[64] W. Fu, S. Li, L. Fang, X. Kang, and J. A. Benediktsson, "Spectral–spatial hyperspectral classification via shape-adaptive sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2014, pp. 3430–3433.

[65] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 603–612.

[66] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2020, pp. 4002–4011.

[67] G. Sun *et al.*, "Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 91, Sep. 2020, Art. no. 102157.

Xuming Zhang received the B.Sc. and M.Sc. degrees from China University of Petroleum (East China), Qingdao, China, in 2018 and 2021, respectively. She is currently pursuing the Ph.D. degree in geography with Nanjing University, Nanjing, China.

Her research interests include hyperspectral image feature extraction and classification, deep learning, and high-resolution remote sensing.

Genyun Sun (Member, IEEE) received the B.Sc. degree from Wuhan University, Wuhan, China, in 2003, and the Ph.D. degree from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently a Professor with China University of Petroleum (East China), Qingdao, China. His research interests include remote sensing image processing, hyperspectral remote sensing, high-resolution remote sensing, and intelligent optimization algorithm.

Xiuping Jia (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of New South Wales at Sydney, Sydney, NSW, Australia, in 1996.

She has been with the School of Information Technology and Electrical Engineering, University of New South Wales at Canberra, Canberra, ACT, Australia, since 1988, where she is currently a Senior Lecturer. She is also a Guest Professor with Harbin Engineering University, Harbin, China, and an Adjunct Researcher with China National Engineering Research Center for Information Technology in Agriculture, Beijing, China. She has coauthored the remote sensing textbook titled *Remote Sensing Digital Image Analysis* (Springer-Verlag, Third Edition, 1999, and Fourth Edition, 2006). Her research interests include remote sensing and imaging spectrometry.

Dr. Jia is an Editor of the *Annals of GIS* and an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Lixin Wu received the B.S. degree in mining survey from China University of Mining and Technology at Xuzhou, Xuzhou, China, in 1988, and the M.S. and Ph.D. degrees in geomatics from China University of Mining and Technology at Beijing, Beijing, China, in 1991 and 1997, respectively.

He is currently working with the School of Geoscience and Info-Physics, Central South University, Changsha, China, as a Leading Professor of geomatics.

Dr. Wu is an Academician of International Eurasian Academy of Sciences, a WG Member of Global Risk Assessment Framework (GRAF) of United Nations Office for Disaster Risk Reduction (UNDRR), a member of the Infrastructure Implementation Board of Group on Earth Observation and the China National Committee of International Society for Digital Earth (ISDE). He was the former Co-Chair of the User Applications in Remote Sensing Committee and the IEEE Geoscience and Remote Sensing Society. He is also the Chair of WG III-8 of the International Society for Photogrammetry and Remote Sensing (ISPRS) and the Vice Chairman of the Space Observation Committee of China Seismology Society. He is also the Editor-in-Chief of the *Journal of Geography* and *Journal of Geo-Information Science* (Chinese) and an Associate Editor of *Geomatics, Natural Hazards and Risk*.

Aizhu Zhang (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from China University of Petroleum (East China), Qingdao, China, in 2011, 2014, and 2017, respectively.

She is currently a Lecturer with China University of Petroleum (East China). Her research interests include artificial intelligence, pattern recognition, city remote sensing, and wetland remote sensing.

Jinchang Ren (Senior Member, IEEE) received the B.Eng., M.Eng., and D.Eng. degrees from Northwestern Polytechnical University, Xi'an, China, in 1992, 1997, and 2000, respectively, and the Ph.D. degree from the University of Bradford, Bradford, U.K., in 2019.

He is currently a Professor with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K. His research interests include image processing, computer vision, machine learning, and big data analytics.

Dr. Ren acts as an Associate Editor for several international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) and *Journal of the Franklin Institute*.

Hang Fu received the B.Sc. degree from China University of Petroleum (East China), Qingdao, China, in 2019, where he is currently pursuing the M.Sc. degree in geomatics engineering.

His research interests include hyperspectral image feature extraction and classification.

Yanjuan Yao received the B.Sc. degree in geography from Henan University, Kaifeng, China, in 1997, the M.Sc. degree in cartography and remote sensing from Beijing Normal University, Beijing, China, in 2004, and the Ph.D. degree in cartography and remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, in 2007.

She undertook the post-doctoral research at Peking University from 2007 to 2009. She is currently with the Satellite Environment Center (SEC), Ministry of Environmental Protection, Beijing. She is also a Professor. Her research interests include radiation transfer modeling for optical remote sensing, terrestrial parameter inversion from multisource remote sensing data, and quantitative remote sensing application for environmental protection.