Hyperspectral Image Classification Based on Dual-Branch Spectral Multiscale Attention Network

Cuiping Shi, *Member, IEEE*, Diling Liao, Yi Xiong, Tianyu Zhang, and Liguo Wang, *Member, IEEE*

***Abstract*—In recent years, convolutional neural networks (CNNs) have been widely used in hyperspectral image classification and have achieved good performance. However, the high dimensions and few samples of hyperspectral remote sensing images tend to be the main factors restricting improvements in classification performance. At present, most advanced classification methods are based on the joint extraction of spatial and spectral features. In this article, an improved dense block based on a multiscale spectral pyramid (MSSP) is proposed. This method uses the idea of multiscale and group convolution of the convolution kernel, which can fully extract spectral information from hyperspectral images. The designed MSSP is the main unit of the spectral dense block (called MSSP Block). Additionally, a short connection with nonlinear transformation is introduced to enhance the representation ability of the model. To demonstrate the effectiveness of the proposed dual-branch multiscale spectral attention network, some experiments are conducted on five commonly used datasets. The experimental results show that, compared with some state-of-the-art methods, the proposed method can provide better classification performance and has strong generalization ability.**

***Index Terms*—Classification, convolutional neural network (CNN), hyperspectral image, multiscale attention, multiscale spectral pyramid (MSSP).**

I. Introduction

IN RECENT years, with the rapid development of imaging technology, remote sensing images have been applied in many fields. Hyperspectral images have high spatial resolution and rich spectral bands [1], which makes them widely used in many fields, such as earth exploration [2], environmental monitoring [3], and ecological science [4].

Hyperspectral image classification is one of the important applications of hyperspectral technology. Hyperspectral images contain rich spatial and spectral information, and fully extracting the spatial and spectral features of images can effectively improve the classification performance of hyperspectral images. Therefore, many methods of extracting spatial and spectral features have been proposed. In the past, some linear-based classification methods were proposed, such as discriminant constraint analysis [5], PCA [6], and balanced local discrimination methods [7]. However, due to the weak representation ability of the linear method, the classification effect is poor when applied to more complex problems. To improve the classification performance, some classification methods based on manifold learning have been proposed, such as the sparse and low rank near-isometric linear embedding method [8], and the semisupervised sparse manifold discriminative analysis method [9].

For image classification, many representative classifiers have been proposed; for example, k-nearest-neighbor classifier based on unsupervised clustering [10], semisupervised logistic regression classifier for high-dimensional data [11], extreme learning classifier with very simple structure [12], sparse-based representation classifier [13], and SVM [14]. Among them, the classifier based on the SVM has obvious advantages in solving small sample size and high-dimensional problem, and it has shown great potential in HSI classification [15].

Hyperspectral images contain abundant information. However, the traditional machine learning methods cannot fully mine the features of hyperspectral images, and only extracted the shallow features of images, resulting in the poor classification effect and weak generalization ability of hyperspectral images. With the rapid development of image processing technology and the improvement in hardware performance, some deep learning methods that can learn deeper features have been proposed. Due to the advanced nature of the deep learning technology, it has been widely used in the field of image processing. In particular, some research works have proved that deep learning technology also has good performance in hyperspectral image classification [16]. To improve the traditional manual spatial–spectral learning method, Tao *et al.* [17] proposed a method based on stacked sparse autoencoders (SAE), which adaptively learns appropriate feature representations from unlabeled data, and finally uses SVM classifier for classification. In [18], a deep belief network (DBN) was proposed to improve the classification accuracy through spatial–spectral localization and classification. However, the SAE and DBN networks have some complete connection layers with a large number of parameters, and the spatial flattening operation also destroys the spatial information of images.

At present, many deep learning methods have been applied to hyperspectral image classification, and have achieved good classification performance. Recurrent neural networks (RNNs)

Manuscript received August 31, 2021; accepted October 7, 2021. Date of publication October 14, 2021; date of current version October 27, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 41701479 and 62071084, in part by the Heilongjiang Science Foundation Project of China under Grant LH2021D022, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 135509136. *(Corresponding author: Cuiping Shi.)*

Cuiping Shi, Diling Liao, Yi Xiong, and Tianyu Zhang are with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: scp1980@126.com; 2020910228@qqhru.edu.cn; 2018132231@qqhru.edu.cn; 2019910178@qqhru.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China (e-mail: wangliguo@hrbeu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3119413

are widely used in image classification because of their good data modeling ability [19]–[21]. However, the feature extraction effect of RNNs is not very good in the case of small samples, which does not make the classification performance ideal. To alleviate this problem, a generative adversarial network is proposed, which can generate high-quality data samples [22]–[29]. Similarly, graph convolutional neural networks, which are modeled by graph structure data, can alleviate the problems caused by small samples in a semisupervision way [30], [31].

Inspired by human vision, a CNN can provide better classification performance for hyperspectral images by using the weight-sharing method of local connection to train the model. In the study of hyperspectral image classification, most methods are based on spatial spectral joint feature extraction [32]. In [33], Zhang *et al.* proposed a dual-channel CNN. One channel uses a 1-D CNN to extract the spectral information of the image, and the other channel uses a 2-D CNN to extract the spatial information of the image. Finally, the spectral information and spatial information extracted by the two channels are fused and classified by a regression classifier. To reduce the number of parameters, Chen *et al.* [34] proposed a 3-D CNN method to extract deep spatial and spectral information at the same time. In [35], Mei *et al.* proposed a new deep learning method C-CNN to explore the feature-learning ability of a five-layer CNN in hyperspectral classification, i.e., integrating spatial context information and spectral information into C-CNN, to improve the representation ability of spatial and spectral information. Although CNN-based methods can effectively extract features, to avoid overfitting, the fine-tuning of parameters usually requires a large number of data samples. Therefore, a densely connection network [36] is proposed, which can improve the generalization ability of the network for hyperspectral images. To improve the learning ability of the deep network and avoid the problems of gradient explosion and gradient dissipation, He *et al.* [37] designed a deep residual network (ResNet), which can make the deep network layer and the shallow network layer perform identity mapping. To jointly learn the spatial and spectral information of hyperspectral images, Zhong *et al.* [38] proposed a supervised residual network (SSRN) based on spatial and spectral residuals, but the training time is long. Wang *et al.* [39] proposed a fast and dense spatial spectral convolution network, which can effectively reduce the data dimension. In [40], Paoletti *et al.* proposed a residual pyramid network (PyResNet), which can gradually increase the feature mapping dimension between layers while balancing the workload of all units. The features extracted from hyperspectral images inevitably contain a lot of redundant information. Inspired by human visual attention, Juan *et al.* [41] proposed a model combining A-ResNet and attention, which can identify the most representative features in the data from the visual perspective. Similarly, Woo *et al.* [42] proposed a convolutional attention module by combining the ResNet network with the attention module of a feedforward CNN, which can retain useful features and discard useless features. Finally, a good classification result of hyperspectral images is obtained. To improve the classification performance of hyperspectral images, the multiscale strategy is also an effective way [43]–[45]. Wu *et al.* [46] proposed a multiscale spatial spectral joint network. Similarly, Pooja *et al.* [47] combined the multiscale strategy with a CNN network to achieve high classification accuracy.

In recent years, attention mechanism is widely used in computer vision and natural language processing [48]–[50]. Wang *et al.* [51] embedded the squeeze and-excitation [52] module into ResNet for HSI classification. To extract more discriminative spatial and spectral features, Ma *et al.* [53] proposed a dual-branch, multiattention network (DBMA), which uses different attention mechanisms to extract the spatial and spectral features of hyperspectral images by dual branches, and then fuse these features for classification. The experimental results show that the DBMA network has a good performance in hyperspectral classification. For further research, Li *et al.* proposed a dual-branch and dual-attention mechanism network (DBDA) [54] based on a new dual attention network [55], [56], which has good classification performance in the case of small number of training samples. Roy *et al.* [57] proposed a Hybrid-SN method, which combines 2-D CNN and 3-D CNN, and 3-D CNN is used to extract the spectral features of the image, whereas 2-D CNN is used to extract the spatial features, and good classification accuracy is obtained. Due to the correlation between noise and spectral band, a CNN with fixed receptive field cannot enable neurons to effectively adjust RF sizes and cross-channel dependencies. Roy *et al.* [58] proposed an attention-based adaptive spectral–spatial kernel improved residual network (A2S2K-ResNet) with spectral attention to capture discriminative spectral and spatial features for HSI classification in an end-to-end training way.

Compared with traditional machine learning methods, the above methods have more advantages in hyperspectral image classification, and have strong generalization ability. However, improving the classification performance of hyperspectral images is still a major challenge in the case of small samples. In the process of hyperspectral image extraction, a large amount of redundant information and the imbalance between different labeled samples greatly reduce the classification performance of hyperspectral images. Therefore, how to obtain more features in the case of limited samples is still worthy of in-depth study.

To obtain more image features with limited samples, a dual-branch multiscale spectral attention network (DBMSA) is proposed, which is based on Dense Net and utilizes multiscale convolution kernels in the spectral branch to extract features of different levels of hyperspectral images. In addition, the attention mechanism is introduced in both the spectral branch and the spatial branch to learn more representative features, so as to enhance the representation ability of specific area of the image.

The main contributions of this article are as follows.

1) Due to the limitations of single-scale convolution kernels, this article proposes a structure of MSSP for the first time. This structure utilizes convolution kernels with different sizes to obtain features of different neighborhoods of the image, which makes the extracted features more comprehensive. Finally, the extracted feature information is fused to help improve the classification performance of hyperspectral images.
2) To strengthen the connection of deep feature information, MSSPs are densely connected, that is, the output of the

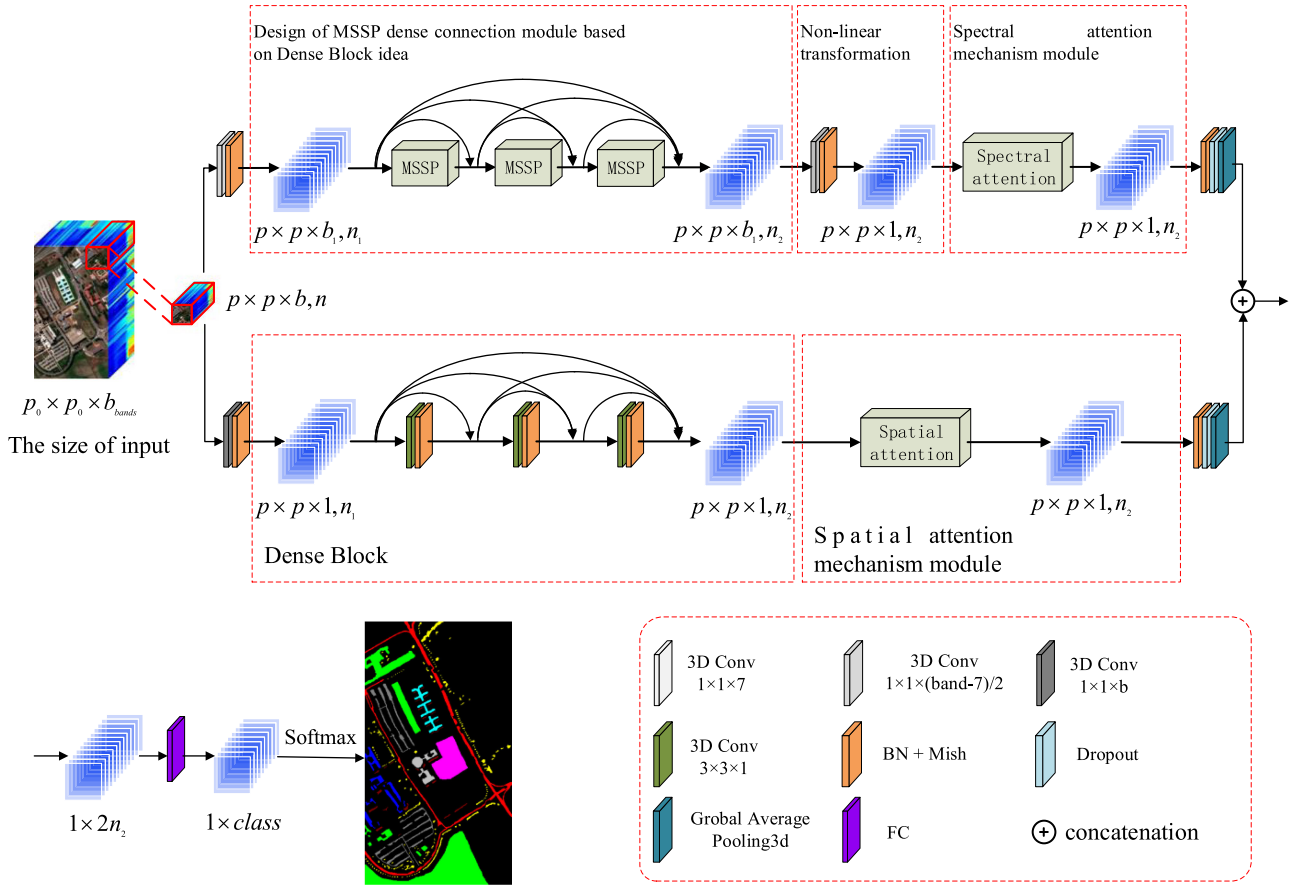


Fig. 1. Overall structure of the proposed DBSMA.

previous layer is used as the input of all subsequent layers. MSSP Block is conducive to a fuller feature extraction of hyperspectral images.

3) To reduce the amount of training parameters, group convolutions with different sizes are used for different branches of the MSSP, which effectively improve the classification performance.
4) The MSSP Block is the first attempt at spectral branching in hyperspectral classification. Experiments show that this method can provide excellent classification performance and has good generalization ability.

The rest of this article are organized as follows. Section II introduces the structure of the DBSMA network in detail. Section III provides the classification results of the DBSMA network on the four common datasets, and compares them with that of some advanced methods. Section IV provides the conclusion.

II. Methodology

For the classification of hyperspectral images, the extraction of the spatial and spectral features is very critical. In this article, a DBMSA network is proposed. For spectral branches, spectral features are extracted from the structure composed of three densely connected MSSPs and a spectral attention mechanism. For spatial branches, a dense block and a spatial attention structure are used to extract spatial features in cooperation. The following four parts will be introduced in detail: the overall structure of DBMSA, spectral feature extraction strategy, spatial feature extraction strategy, and nonlocal feature selection strategy.

A. Structure of DBSMA

The proposed DBMSA model consists of an MSSP dense connection module, a spatial dense connection block, a spectral attention module and a spatial attention module, a fully connected layer, a global average pooling layer, and a classifier. The overall structure is shown in Fig. 1. The size of the input is $P \in R^{p_0 \times p_0 \times b_{bands}}$. To keep the size of the input cube and the output cube unchanged, the zero filling strategy is adopted. To avoid data explosion and gradient disappearance, BN + Mish [59] is used as the normalization and activation function to standardize the input data. In particular, to extract key information as much as possible, spectral attention and spatial attention are utilized to improve the performance of the network. After the output cube of the attention module passes through the dropout layer and the global average pooling layer, it becomes a 1-D vector. Then, the two output vectors of the spectral branch and spatial branch attention are cascaded into a new vector. The activation function is used to process the vector as the sum

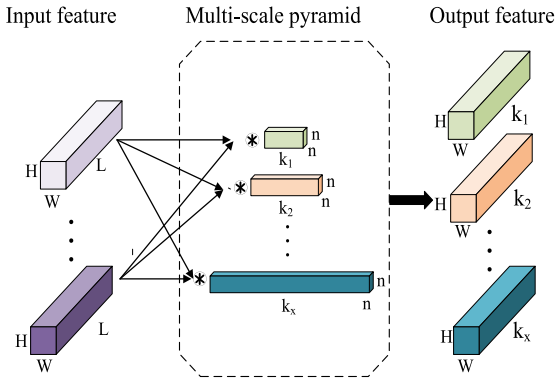


Fig. 2. Structure of pyramid convolution.

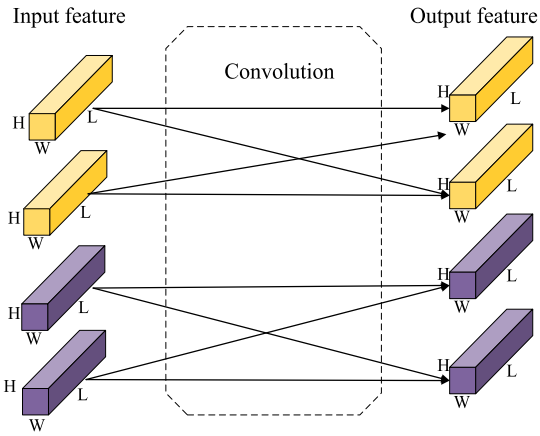


Fig. 3. Structure of grouped convolution.

of the probabilities of all elements is 1, and then it is classified by the classifier.

B. Strategy for Extracting Spectral Features Based on MSSP

1) MSSP Structure: The structure of pyramid convolution is shown in Fig. 2. For pyramid convolution, the size of convolution filter remains unchanged. From the top to the bottom of the pyramid, the depth of the filter is gradually increased. That is, the filter can transition from a smaller receiving field to a larger receiving field to obtain more complementary information. A convolution filter with small scale can obtain detailed information, whereas a filter with large scale can obtain global context information. Therefore, different scale convolution kernels can obtain hierarchical features of the image.

To better extract the spectral features and reduce the computational complexity of the model, randomly shuffled input data are grouped and convolved in MSSP (i.e., the input feature map is grouped in to $1, 2, 4, 8$). Fig. 3 shows the case where the group is equal to 2. Here, the four input feature maps are divided into two groups. Compared with standard convolution, the complexity of grouped convolution [60] is reduced. In particular, there are two situations in grouped convolution: if it is divided into one group (that is, not grouped), the calculation complexity of the convolution is the same as that of the standard convolution; on the contrary, as the number of grouping groups increases, the computational complexity will become lower and lower. Suppose the inputs are N_i feature maps with size $H \times W \times L$, and the size of the filter is $1 \times 1 \times k$. Divide the input feature maps into m groups, then each group of inputs will be N_i/m cubes of size $H \times W \times L$, with N_o/m convolution kernels of size $1 \times 1 \times k$. After grouped convolution, the output will be N_o/m feature maps of size $H \times W \times L$ and the total number of output feature maps is $\frac{N_0}{m} \cdot m$ (where N_i and N_0 are the number of input and output feature maps, and H, W, and L are the height, width, and number of channels, respectively). Among them, the calculation times of standard convolution and grouped convolution are

$$f = k^2 \times L \times W \times H \times l \tag{1}$$

$$F = \left(k^2 \times \frac{L}{m} \times H \times W \times \frac{l}{m}\right) \times m \tag{2}$$

where f represents the number of calculation required for standard convolution, F represents the number of calculation required for grouped convolution, k^2 is the space size of the filter, L represents the number of bands of the input feature map, l represents the number of bands of the output feature map, m is the number of input groups, and H and W are the height and width of the output feature map, respectively. Obviously, $f < F$, that is, the calculation times of grouped convolution is only $1/m$ of that of standard convolution.

Fig. 4 shows the proposed MSSP structure. The input size is $H \times W \times L$. To extract the spectral information effectively, the convolution unit of $1 \times 1 \times 1$ is used to expand the input size. Different sizes of convolution kernels are used for spectral feature extraction. In the branches of different scale convolution kernels, the input is divided into one group, four groups, and eight groups, respectively, for group convolution, and the output features of different branches are fused. However, when the number of network layers increases, network degradation may occur, leading to unsatisfactory model training results. Therefore, after nonlinear convolution, skip connection is utilized to realize residual mapping, so as to avoid gradient disappearance and explosion, that is

$$p(x) = \sigma(x) + q(x) \tag{3}$$

where $\sigma(x)$ is the output of nonlinear residual structure, $q(x)$ is the output of multiscale convolution structure, and $p(x)$ is the output after the model of MSSP.

2) Dense Connection Block Based on MSSP Structure (MSSP Block): To facilitate the flow of information between layers, three MSSPs are further densely connected, as shown in Fig. 5. The input of the ith layer is the sum of the output of the $(i-1)$th previous layer, and the relationship between input and output of MSSP Block can be represented as

$$y_i = h([x_1, x_2, \ldots, x_{i-1}]) \tag{4}$$

where y_i represents the output of the ith MSSP, $h(\cdot)$ represents the function of MSSP, and $[x_1, x_2, \ldots, x_{i-1}]$ represents the output of the previous $(i-1)$ MSSP Block.

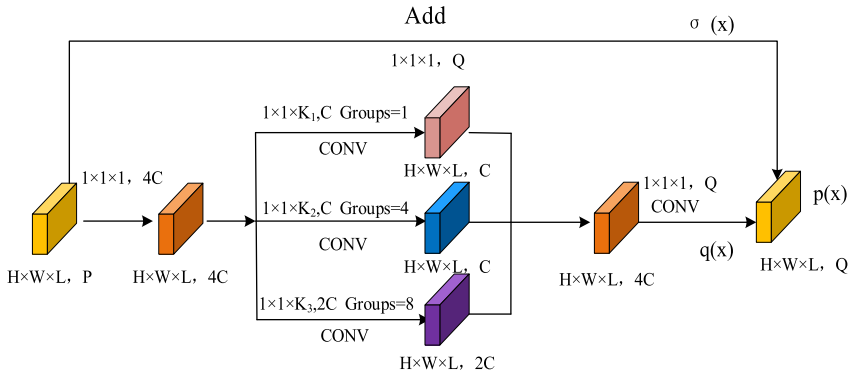


Fig. 4. Proposed MSSP structure.

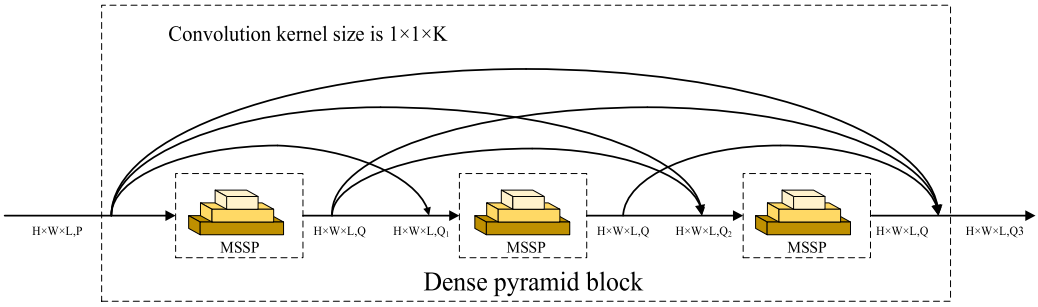


Fig. 5. MSSP Block.

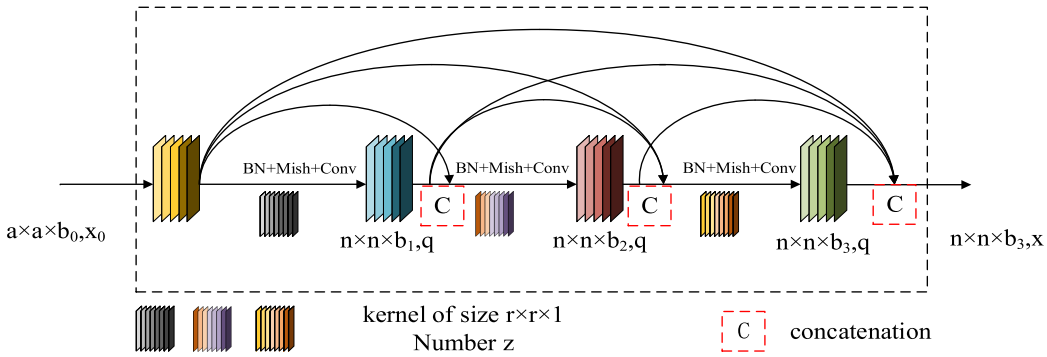


Fig. 6. Spatially densely connected blocks.

Assuming that the input is $P \in R^{H\times W\times L}$, the output after each MSSP is Q feature maps with the same size as the input. After i MSSP Block, the linear relationship between the total number of output feature maps Q_i and the number of output feature maps Q of each MSSP can be represented as

$$Q_i = L + (i-1)Q \tag{5}$$

where Q_i represents the total number of output feature maps after i MSSP Block, L is the number of bands of the input map feature, and Q represents the number of output after each MSSP.

C. Strategy for Extracting Spatial Features

It is difficult to extract the deep spatial features of hyperspectral images by a shallow neural network. To establish the connection relationship between the different layers, shallow and deep layers are connected by skip, so that the layers are densely connected, which can not only facilitate the information flow of information in each layer, but also avoid information loss.

The processing of the dense block in the spatial branch is similar to that of the MSSP Block in the spectral branch. The structure of the spatial branch dense blocks is shown in Fig. 6. The relationship between the input and output of the spatially

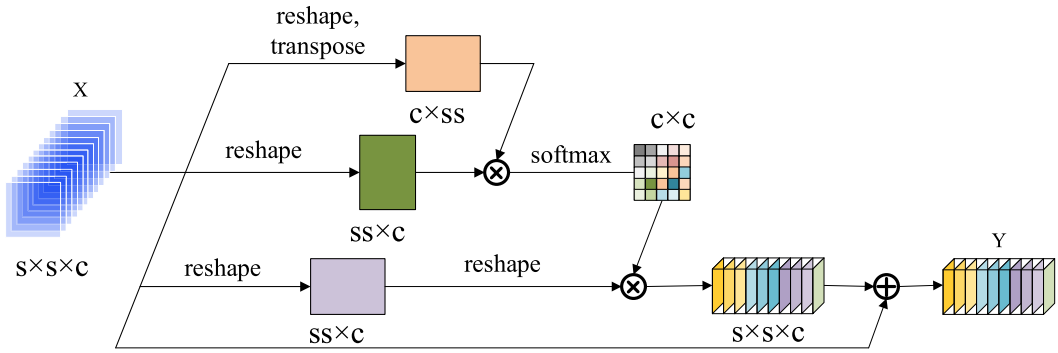


Fig. 7. Spectral attention module.

densely connected block can be represented as

$$x_i = H([x_1, x_2, \ldots, x_{i-1}]) \tag{6}$$

where $H(\cdot)$ is the function of spatially dense connection, and $[x_1, x_2, \ldots, x_{i-1}]$ is the output of previous $(i-1)$ layers. $x_i$is the number of feature maps in the ith layer.

Suppose that the input is x_0 feature maps with size $P \in R^{a\times a\times b_0}$. To avoid the gradient explosion of the input data, BN is used to normalize the data, Mish is the activation function of the input data, the size of the filter is $r \times r \times 1$, and the total number of output feature maps x of the spatially dense block is calculated in the same way as that of multiscale pyramid convolution dense blocks of spectral branches.

D. Strategies for Nonlocal Feature Selection Attention and Fusion Mechanism

The attention mechanism can not only automatically learn important spectral and spatial features, but also suppress useless information in the spectral and spatial. Because it helps to provide good classification effect in image classification, attention mechanism has been widely used in the field of image processing. In DBMSA, the attention mechanism is utilized in the spectral branch and spatial branch, respectively. According to the MSSP Block described in Section II-B and the spatial dense block introduced on Section II-C, the spectral and spatial features of HSI are extracted and fused. The process of attention mechanism in a DBSMA network is described in detail as follows.

The structure of the spectral attention mechanism is shown in Fig. 7. It can be seen that, in the spectral branch, the attention mechanism generates attention maps by understanding the relationship between channels and emphasizing the important parts of the feature map. Assuming that the input size is $P \in R^{s\times s\times c}$ (where $s \times s$ is the space size of input and c is the number of input bands), through matrix multiplication and activation function, the weighted map with channel attention is obtained. On the one hand, the activation function normalizes the data and organizes the attention map into a probability distribution with the weighted sum of each channel being 1. On the other hand, the activation function can be used to highlight the more important parts. Let $X_n(n = 1, 2, \ldots, c)$ be the channel of the input patch, and after passing through activation function layer, the spectral attention map $G \in R^{c\times c}$ is

$$g_{ji} = \frac{\exp(X_i^T \cdot X_j)}{\sum_{\forall j} \exp(X_i^T \cdot X_j)} \tag{7}$$

where g_{ji} is the weight coefficient of the ith channel to the jth channel, that is, the importance of the ith channel to the jth channel. Let α be the attention parameter (if $\alpha = 0$, it means that operation without attention mechanism), then the output of the spectral attention mechanism is

$$Y_j = \alpha \sum\nolimits_{\forall j} g_{ji} X_j + X_j \tag{8}$$

where $Y_n(n = 1, 2, \ldots, c)$ is the n-channel feature map of $Y \in R^{s\times s\times c}$.

The structure of the spatial attention mechanism is shown in Fig. 8. It can be seen that the process of the spatial attention mechanism is similar to that of the spectral attention mechanism. Different from the spectral attention mechanism, the input X is convoluted with the convolution kernel of size $r \times r \times b$, and three new feature maps A, B, and C are obtained, respectively. Here, $\{A, B, C\} \in R^{s\times s\times c}$. Next, A, B, and C are transformed into 2-D matrices with size $ss \times c$ (where ss represents the number of pixels). Then, multiply B and $\mathrm{A^T}$, and obtain the spatial attention map $E \in R^{ss\times ss}$ after the softmax layer, that is

$$e_{ji} = \frac{\exp(A_i \cdot B_j)}{\sum_{\forall j} \exp(A_i \cdot B_j)} \tag{9}$$

where $e_{ji}$is the weight coefficient of the ith pixel to the jth pixel, that is, the importance of the ith pixel to the jth pixel. Then, multiply the matrices C and $\mathrm{E^T}$, and connect the result to the original input X through the residual connection, and the final output is

$$Z_j = \beta \sum\nolimits_{\forall j} e_{ji} C_j + X_j \tag{10}$$

where $Z_n(n = 1, 2, \ldots, ss)$ is the value of the output cube $Z \in R^{s\times s\times c}$ at the spatial position n, and β is the attention parameter.

III. Experimental Results and Analysis

In this section, we first introduce the datasets used in the experiment, then give the hyperparameter settings of the network

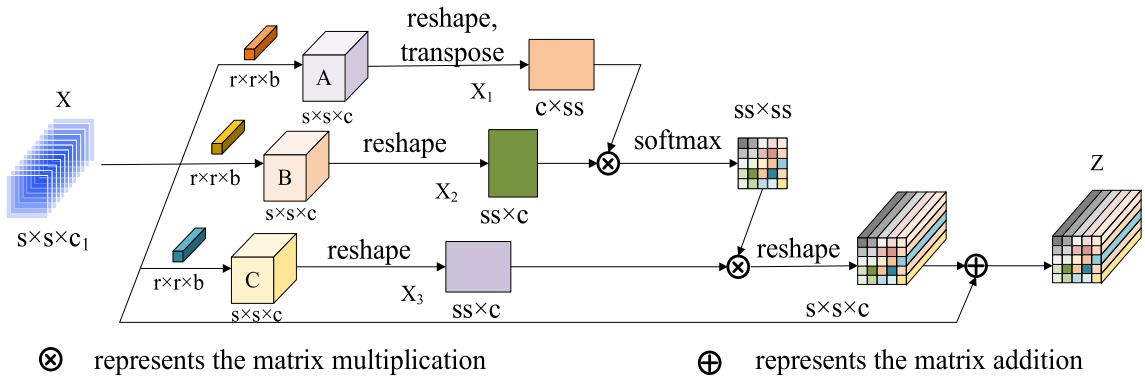


Fig. 8. Spatial attention module.

and detailed analysis of the parameters, and finally analyze the performance of the proposed method and compare it with other advanced methods. To quantitatively analyze the DBMSA, three commonly used quantitative indicators are adopted, namely overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa). To avoid data bias caused by randomness, each experiment is repeated 30 times, and the average of these experimental results is taken as the final result.

A. Hyperspectral Dataset

In this part, we will introduce five datasets in detail, namely Indian Pine (IN), University of Pavia (UP), Kennedy Space Center (KSC), Salinas Valley (SV), and University of Houston (HS). Fig. 9 shows the real image, false color image, and class information of each data in the dataset.

1) IN: The Indian pine dataset is a hyperspectral image acquired by an airborne visible infrared imaging spectrometer in the northwestern part of Indiana, USA. The image spatial size is 145×145, the number of bands is 220, and the wavelength range is 200–2400 nm. The spectral and spatial resolutions are 10 nm and 20 m, respectively. Except for background pixels, there are generally 10 249 spatial pixels used for experiments. There are 16 true types of ground objects, but, because some of them have fewer data labels, only take 9 of the 16 categories. Because 20 are unavailable, the experiment only takes the remaining 200 bands out of the 220 bands for research.
2) UP: This dataset is used for image acquisition through a reflection optical system imaging spectrometer. The size of the image spatial is 610×340, and the spatial resolution is 1.3 m. Among them, the dataset is divided into 9 categories. 115 bands and 12 noise bands are removed, leaving 103 usable bands.
3) KSC: This dataset was obtained by an AVIRIS sensor in Florida in 1996, with a spatial size of 512×614 and a spatial resolution of 18m; in addition, the image consists of 13 feature categories and 176 bands.
4) SV: This dataset is a hyperspectral image obtained through an AVIRIS sensor in the United States. The spatial size of the image is 512×217, and the spatial resolution is 1.7 m. Among them, there are 16 categories of ground objects and 224 bands, but 20 water absorption bands were removed, and the remaining 204 bands were used for hyperspectral image classification experiments.
5) HS: The Houston 2013 (HS) dataset is the competition data of the 2013 GRSS Data Fusion contest, which describes the landscape of Houston University and its surrounding areas. The size of the dataset is 349 × 1905, and the spatial resolution is 2.5 m per pixel. The dataset contains 144 spectral bands and 15 kinds of surface features.

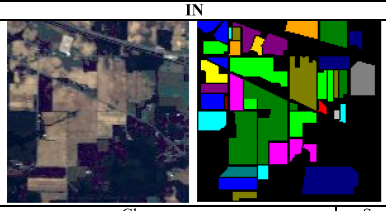
B. Experimental Setup

During the experiment, the learning rate setting ranges are 0.001, 0.005, 0.0001, 0.0005, and 0.00005. Through multiple experiments on each learning rate, the best learning rate in the four datasets is 0.0005; the number of iterations of the experiment is set to 200 and batch size to 16. The hardware platform used in the experiment is Intel(R) Core(TM) i7-9750H CPU, NVIDA GeForce GTX1060 Ti GPU and 8GB memory. The software environment is CUDA 10.0, pytorch 1.2.0 and python 3.7.4. In the experiment, the method in this article is compared with classic classifiers and newer network models in hyperspectral classification, including SVM, SSRN, CDCNN, PyResNet, DBMA, DBDA, Hybrid-SN, and A2S2K-ResNet. In the experiment, OA, AA, and Kappa are used as indicators of model performance, and the average of the results of 30 experiments is taken. In the case of small sample data, the experimental results show that the proposed network model has better classification performance than other advanced methods and has better generalization ability.

C. Parameter Analysis

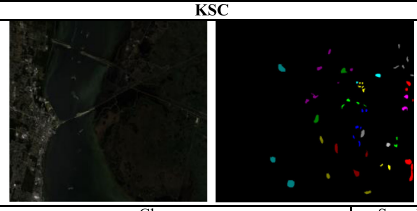
1) For the proposed DBMSA method, the feature extraction methods of spectral branch and spatial branch are different. To avoid the infection of spectral and spatial information, two branches extract spectral and spatial information, respectively. In addition, in the five datasets of IN, UP, KSC, SV, and HS, 3%, 0.5%, 5%, 0.5%, and 2% of the data were randomly selected as training samples, and the remaining data were used as test samples.

IN



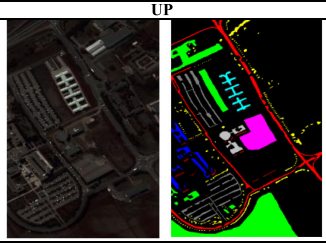
Class			Samples
No.	Color	Name	Numbers
C1		Alfalfa	64
C2		Corn-notill	1428
C3		Corn-mintill	830
C4		Corn	237
C5		Grass-pasture	483
C6		Grass-trees	730
C7		Grass-pasture-mowed	28
C8		Hay-windrowed	478
C9		Oats	20
C10		Soybean-notill	972
C11		Soybean-mintill	2455
C12		Soybean-clean	593
C13		Wheat	205
C14		Woods	1265
C15		Buildings-Grass-Trees	386
C16		Stone-Steel-Towers	93
TOTAL			10267

KSC



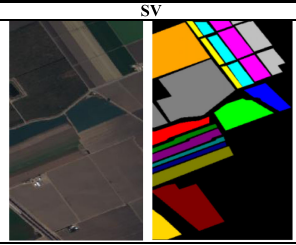
Class			Samples
No.	Color	Name	Numbers
C1		Scrub	761
C2		Willow swamp	243
C3		CP hammock	256
C4		Slash pine	252
C5		Oak/Broadleaf	161
C6		Hardwood	229
C7		Grass-pasture-mowed	105
C8		Graminoid marsh	431
C9		Spartina marsh	520
C10		Cattail marsh	404
C11		Salt marsh	419
C12		Mud flats	503
C13		Water	927
TOTAL			5211

UP



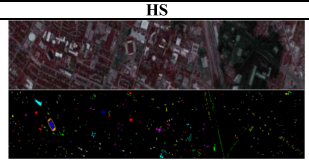
Class			Samples
No.	Color	Name	Numbers
C1		Asphalt	6631
C2		Meadows	18649
C3		Gravel	2099
C4		Trees	3064
C5		Painted metal sheets	1345
C6		Bare Soil	5029
C7		Bitumen	1330
C8		Self-Blocking Bricks	3682
C9		Shadows	947
TOTAL			42776

SV



Class			Samples
No.	Color	Name	Numbers
C1		Brocoil-green-weeds_1	2009
C2		Brocoil-green-weeds_2	3726
C3		Fallow	1976
C4		Fallow-rough-plow	1394
C5		Fallow-smooth	2678
C6		Stubble	3959
C7		Celery	3579
C8		Grapes-untrained	11271
C9		Soil-vin-yard-develop	6203
C10		Corn-senesced-green-weeds	3278
C11		Lettuce-romaine-4wk	1068
C12		Lettuce-romaine-5wk	1927
C13		Lettuce-romaine-6wk	916
C14		Lettuce-romaine-7wk	1070
C15		Vin-yard-untrained	7268
C16		Vin-yard-vertical-trellis	1807
TOTAL			54129

HS



Class			Samples
No.	Color	Name	Numbers
C1		Healthy grass	1251
C2		Stressed grass	1254
C3		Synthetic grass	697
C4		Trees	1244
C5		Soil	1242
C6		Water	325
C7		Residential	1268
C8		Commercial	1244
C9		Road	1252
C10		Highway	1227
C11		Railway	1235
C12		Parking Lot 1	1233
C13		Parking Lot 2	469
C14		Tennis Court	428
C15		Running Track	660
TOTAL			15029

Fig. 9. Real features and false color maps of four common datasets, and the number of available samples.

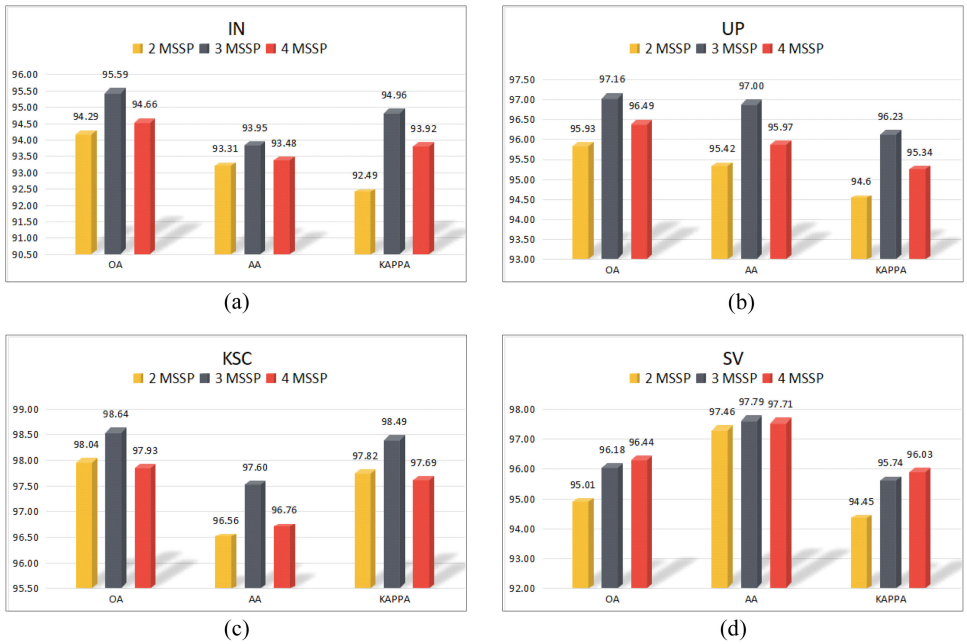


Fig. 10. Classification performance of different numbers of MSSP dense connections. (a) IN. (b) UP. (c) KSC. (d) SV (%).

TABLE I
FOR THE FOUR DATASETS, THE TIME CONSUMED BY TRAINING AND TESTING UNDER DIFFERENT COMBINATIONS OF MSSP NUMBERS (S)

Time(s)	IN		UP		KSC		SV	
	Train	Test	Train	Test	Train	Test	Train	Test
2 MSSP	228	22	88	57	210	9	247	128
3 MSSP	252	40	109	95	308	16	309	231
4 MSSP	461	55	149	133	468	23	466	313

2) The influence of the number of dense connections of MSSP on classification accuracy: In the MSSP Block, the output of the previous MSSP affects the input of the convolution of the next MSSP. Therefore, the classification performance of the network will be affected by the number of MSSP dense connections. When the numbers of MSSP dense connections are 2, 3, and 4, the experimental results are shown in Fig. 10. It can be seen from Fig. 10 that for the IN, UP, and KSC datasets, the OA, AA, and Kappa values obtained by densely connected 2 MSSP Block and densely connected 4 MSSP Block are all lower than those of the densely connected blocks of 3 MSSP Block. Moreover, the classification accuracy of the densely connected blocks of 3 MSSP Block on the four datasets is all above 93.5%. For the SV dataset, although the OA and Kappa values obtained by the dense connection of 4 MSSP Block are 0.26% and 0.29% more than those obtained by the dense connection of 3 MSSP Block, the training time required is more than one-third times, as shown in Table I. According to the above analysis, densely connected blocks consisting of 3 MSSP Block can extract image features more effectively.
3) The effect of the combination of filters in MSSP on classification accuracy: in HSI classification, the size of the filter of CNN is directly related to the size of the receiving field, and the context information and detailed features of the image affect the classification accuracy. To reduce the spatial dimension, the sizes of the convolution filters are usually selected as $1\times1\times3$, $1\times1\times5$, $1\times1\times7$, $1\times1\times9$, and $1\times1\times11$. However, as the size increases, the number of parameters also increases. Therefore, the use of a small-scale filter is relatively widespread. To further explore the influence of the combination of pyramid multiscale filter on the classification performance, the above several convolution kernels are grouped according to the pyramid multiscale principle. Different combinations of multiscale filters are used to obtain different classification accuracy. The experimental results are shown in Table II. Among them, $1\times1\times3$, $1\times1\times5$, and $1\times1\times7$ have the highest classification accuracy in the IN, UP, and KSC datasets. Although this combination method is not the highest in the classification accuracy of the SV dataset, its OA is only 0.24% lower than the highest. In addition, the multiscale combination of $1\times1\times5$, $1\times1\times7$, and $1\times1\times9$ performs poorly in other datasets; that is, their generalization ability is weak. Therefore, the combination of pyramid multiscale filters $1\times1\times3$, $1\times1\times5$, $1\times1\times7$ can provide the best classification performance.

D. Experimental Results and Analysis

To verify the method proposed in this article, according to the parameter settings in Section III-B, the DBMSA is tested on four

TABLE II
INFLUENCE OF THE SIZE COMBINATION OF THE MULTISCALE CONVOLUTION KERNEL IN MSSP ON THE CLASSIFICATION ACCURACY (%)

	IN			UP			KSC			SV		
	OA(%)	AA(%)	K×100	OA(%)	AA(%)	K×100	OA(%)	AA(%)	K×100	OA(%)	AA(%)	K×100
1×1×3 1×1×5 1×1×7	**95.81**	**93.48**	**95.22**	**97.5**	**97.03**	**96.68**	**98.49**	**97.42**	**98.33**	96.28	97.82	95.85
1×1×5 1×1×7 1×1×9	92.16	89.15	91.05	96.95	96.64	95.95	97.72	96.26	97.46	**96.52**	**98.15**	**98.04**
1×1×7 1×1×9 1×1×11	95.52	92.9	94.89	96.59	95.56	95.49	97.89	96.48	97.65	96.19	98.04	95.88

The bold entities means that this method has the best result of the comparison methods.

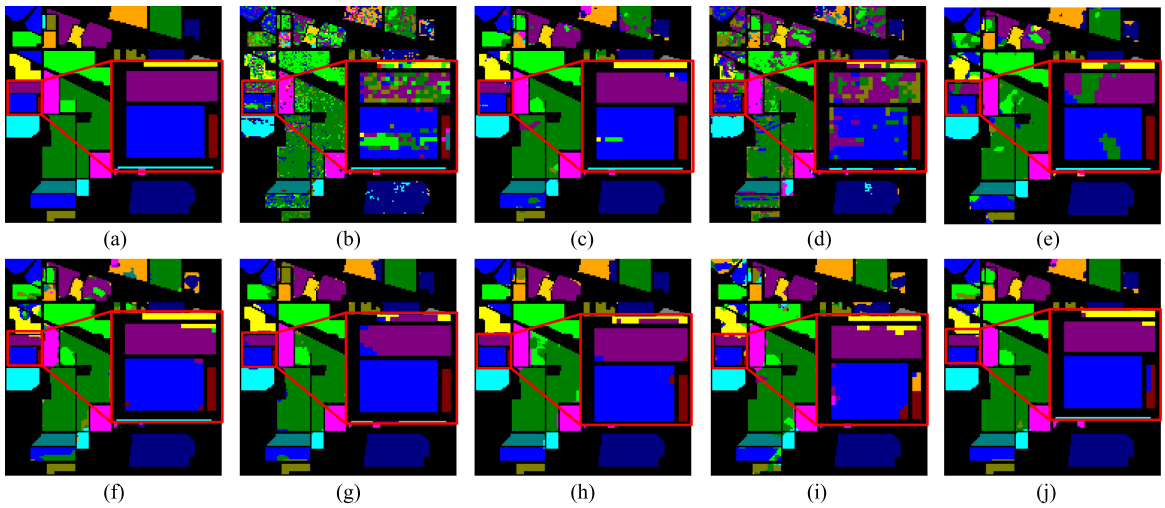


Fig. 11. Classification maps on the IN dataset. (a) Real object map. (b) SVM. (c) SSRN. (d) CDCNN. (e) PyResNet. (f) DBMA. (g) DBDA. (h) Hybrid-SN. (i) A2S2K-ResNet. (j) Proposed.

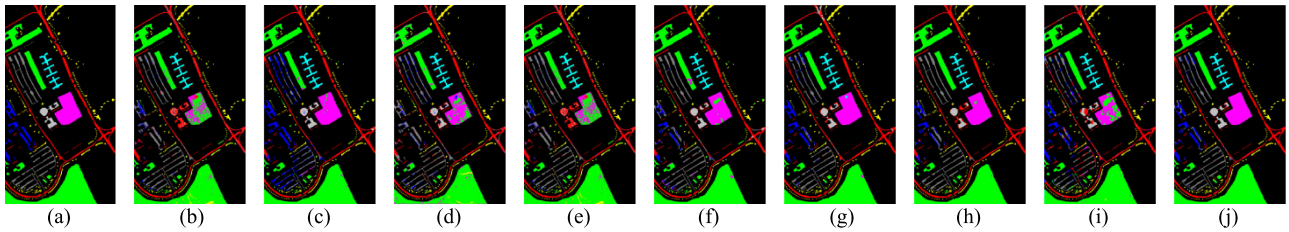


Fig. 12. Classification maps on the UP dataset. (a) Real object map. (b) SVM. (c) SSRN. (d) CDCNN. (e) PyResNet. (f) DBMA. (g) DBDA. (h) Hybrid-SN. (i) A2S2K-ResNet. (j) Proposed.

datasets. The proposed DBMSA method is compared with some classical and state-of-the-art classification methods, i.e., SVM, SSRN, CDCNN, PyResNet, DBMA, DBDA, Hybrid-SN, and A2S2K-ResNet.

1) Experiment 1: Figs. 11 –15 show the comparison of classification results of different methods on five datasets, respectively. It can be seen from Figs. 11–15 that there is a lot of noise in the classification results based on SVM, and the classification effect is not ideal. Compared with the SVM method, CDCNN can provide a better classification performance by exploring the optimal local spatial–spectral context dependence. Compared with the CDCNN method, PyResNet and SSRN extract spatial–spectral features through the deep structure of residual connection, and the classification results are better. To fully extract the spatial–spectral features and avoid the mutual interference of spatial–spectral information, DBMA and DBDA use two branches to extract the spatial–spectral features of hyperspectral images separately, and achieve a good classification effect. The visual images obtained by HybridSN under the end-to-end deep learning framework are relatively smooth and less noisey. By comparison, the visual images obtained by A2S2K-ResNet are coarse. However, the DBMSA not only learns spectral features

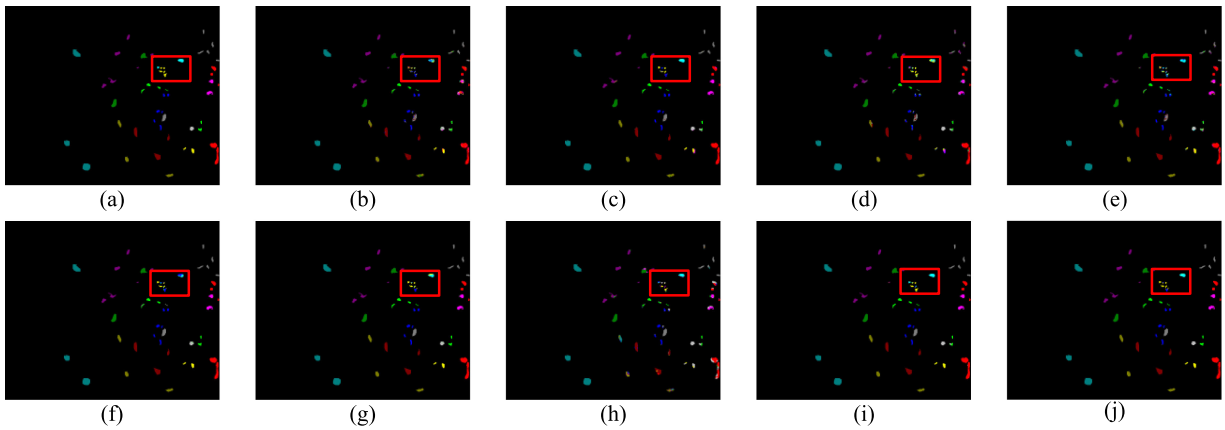


Fig. 13. Classification maps on the KSC dataset. (a) Real object map. (b) SVM. (c) SSRN. (d) CDCNN. (e) PyResNet. (f) DBMA. (g) DBDA. (h) Hybrid-SN. (i) A2S2K-ResNet. (j) Proposed.

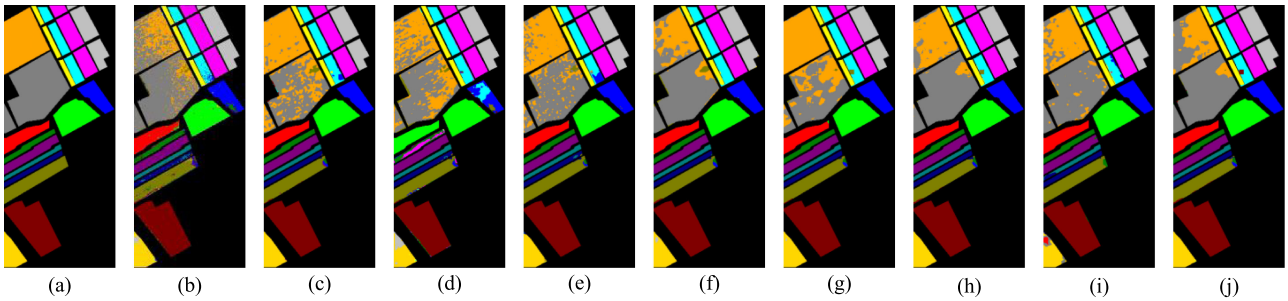


Fig. 14. Classification maps on the SV dataset. (a) Real object map. (b) SVM. (c) SSRN. (d) CDCNN. (e) PyResNet. (f) DBMA. (g) DBDA. (h) Hybrid-SN. (i) A2S2K-ResNet. (j) Proposed.

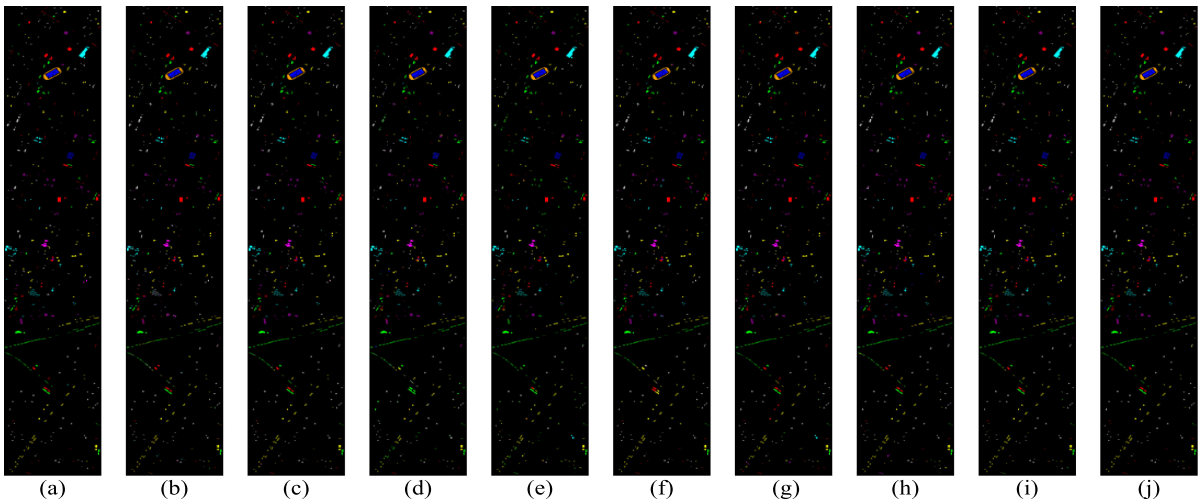


Fig. 15. Classification maps on the HS dataset. (a) Real object map. (b) SVM. (c) SSRN. (d) CDCNN. (e) PyResNet. (f) DBMA. (g) DBDA. (h) Hybrid-SN. (i) A2S2K-ResNet. (j) Proposed.

through convolution kernels with different size in spectral branches, but also improves classification accuracy in the case of small samples through the attention mechanism. Thus, compared with other methods, the obtained classification maps are more accurate and smoother.

The classification results of SVM-based and CNN-based methods are shown in Tables III–VII. It can be seen that, the lowest classification accuracy obtained by SVM, and for the advanced methods, namely SSRN, PyResNet, DBMA, and DBDA methods, the classification accuracy of the DBDA method

TABLE III
CLASSIFICATION RESULTS OF IN DATASET USING 3% TRAINING SAMPLES (VALUE ± STANDARD DEVIATION)

Class	SVM	CDCNN	SSRN	PyResNet	DBMA	DBDA	Hybird-SN	A2S2K-ResNet	Proposed
C1	36.62±0	49.57±7.79	82.54±8.88	26.67±5.88	82.05±5.25	97.49±0.55	81.79±2.93	93.43±0.5	96.92±1.03
C2	55.49±0	65.87±3.87	89.19±1.53	80.92±4.12	85.73±3	93.25±1.85	69.12±6.24	93.01±2.37	95.65±0.11
C3	62.55±0.38	61.2±5.17	87.67±0.88	81.24±8.79	88.44±4.26	92.6±1.07	91±0.81	90.25±0.37	94.82±1.51
C4	42.54±0	53.9±1.68	84.28±1.23	62.17±7.15	87.79±2.27	93.63±1.07	84.87±8.4	89.94±1.7	95.76±0.91
C5	85.05±0	88.36±1.36	97.77±0.37	91.75±1.81	94.85±1.38	98.76±0.27	90.73±2.91	97.78±0.27	98.39±0.01
C6	83.32±0	90.17±2.21	96.43±0.58	94.26±1.31	97.33±0.44	97.85±0.84	88.59±1.95	98.25±1.12	98.02±0.44
C7	59.87±0	56.24±1.22	86.99±2.6	19.75±17.5	50.91±3.85	66.62±3.37	83.62±18.6	81.8±0.97	72.49±1.4
C8	89.67±0	93.93±0.58	96.76±0.61	100±0	98.62±0.41	99.75±0.24	87.24±4.6	99.2±0.3	100±0
C9	39.45±0.29	49.09±7.83	72.15±11.9	69.09±27.11	51.31±0.74	84.42±6.05	60.44±7.11	64.65±4.28	77.8±0.86
C10	62.32±0	63.94±6.05	85.92±3.51	82.96±1.43	84.22±5.24	87.47±0.79	86.25±2.08	89.08±1.02	91.77±0.5
C11	63.73±1.73	68.75±1.81	89.27±1.2	89.59±0.74	87.51±1.68	94.12±1.65	88.95±3.83	90.52±0.93	96.66±0.23
C12	50.55±0	40.3±1.84	86.33±0.88	59.82±2.27	81.18±1.41	92.22±4.95	79.03±2.3	93.66±2.9	93.12±0.49
C13	86.74±0	86.69±5.23	99.14±0.13	80.07±2.03	94.8±1.89	97.69±0.22	93.64±3.99	98.74±0.62	97.49±0.25
C14	88.67±0	86.24±5.61	95.54±0.52	96.31±1.56	95.52±0.75	97.15±0.31	92.65±0.71	95.68±1.34	98.04±0.18
C15	61.82±0	85.63±11.72	89.64±1.67	86.36±4.18	83.19±0.59	93.37±1.19	88.83±3.65	91.86±2.11	94.27±1
C16	98.66±0	92.42±2.48	95.47±1.2	90.37±4.63	93.47±0.51	91.83±0.66	92.23±2.54	94.27±0.45	94.47±2.45
OA(%)	68.76±0	70.43±2.58	90.25±0.42	85.65±1.45	87.95±1.07	93.58±0.55	82.18±1.5	92.55±0.11	**95.81±0**
AA(%)	66.73±0	70.36±1.19	89.69±0.97	75.67±1.27	84.8±0.61	92.17±0.25	84.31±1.61	91.29±0.25	**93.48±0.26**
K×100	63.98±0	66.23±2.75	88.87±0.48	83.6±1.64	86.24±1.21	92.69±0.64	79.85±1.42	91.48±0.12	**95.22±0**
Params	-	1.1225M	364.168k	22.388M	609.791k	382.326k	8.256M	373.184k	498.354k
Runtime(s)	-	24	106	56	222	194	37	40	242

The bold entities means that this method has the best result of the comparison methods.

TABLE IV
CLASSIFICATION RESULTS OF THE UP DATASET USING 0.5% TRAINING SAMPLES (VALUE ± STANDARD DEVIATION)

Class	SVM	CDCNN	SSRN	PyResNet	DBMA	DBDA	Hybird-SN	A2S2K-ResNet	Proposed
C1	81.26±0	86.77±0.47	94.1±2.21	88.11±6.5	89.82±1.38	93.5±0.86	70.33±9.87	81.61±6.34	96.51±0.9
C2	84.53±0	93.72±0.38	96.66±0.79	97.77±1.61	96.08±0.05	99.08±0.16	87.41±6.43	91.26±2.12	99.24±0.28
C3	56.56±0	64.27±0.76	76.75±5.41	30.97±18.97	76.09±6.56	88.85±3.32	64.1±2.01	76.49±8.05	93.59±0.81
C4	94.34±0	95.12±0.83	99.29±0.08	84.79±9.42	95.7±1.5	97.26±0.25	82.4±12.91	99.05±0.38	98.12±0.42
C5	95.38±0	96.52±0.79	99.64±0.2	96.64±4.42	98.45±0.5	98.83±0.32	85.16±11.84	99.3±0.5	98.68±0.06
C6	80.66±0	88.61±6.95	93.85±2.6	54.3±13.16	92.65±1.19	97.46±0.85	81.47±12.65	94±1.37	98.23±0.09
C7	49.13±0	77.29±3.54	86.48±4.29	38.3±25.69	86.72±12.62	91.61±6.65	81.01±17.78	95.99±5.58	99.34±0.33
C8	71.16±0	79.52±0.3	83.71±3.29	75.5±18.22	80.18±2.36	88.42±2.27	72.18±12.61	65.54±0.66	91.37±0.66
C9	99.94±0	91.04±0.57	98.97±0.31	91.15±8.5	94.38±1.41	97.48±0.77	79.58±2.22	92.94±0.62	98.16±0.69
OA(%)	82.06±0	87.94±0.13	92.5±1.33	83.01±1.89	91.8±0.56	96.01±0.03	82.38±4.48	86.81±1.19	**97.5±0.05**
AA(%)	79.22±0	85.32±0.19	92.16±1.32	73.06±3.5	90.01±2.64	94.72±0.59	78.19±9.37	87.96±1.22	**97.03±0.22**
K×100	75.44±0	83.95±0.16	90.89±1.64	76.9±2.64	89.04±0.75	94.71±0.04	73.76±9.36	82.18±1.54	**96.68±0.06**
Params	-	610.6k	216.537k	22.073M	324.376k	202.751k	6.467M	221.976k	318.779k
Runtime(s)	-	42	71	61	96	93	71	182	132

The bold entities means that this method has the best result of the comparison methods.

TABLE V
CLASSIFICATION RESULTS OF THE KSC DATASET USING 5% TRAINING SAMPLES (VALUE ± STANDARD DEVIATION)

Class	SVM	CDCNN	SSRN	PyResNet	DBMA	DBDA	Hybrid-SN	A2S2K-ResNet	Proposed
C1	92.43±0	96.81±0.69	98.4±0.48	99.86±0.14	99.39±0.39	99.67±0.16	88.08±5.95	100±0	99.99±0.02
C2	87.14±0	83.65±1.41	94.52±1.92	92.93±7.05	93.8±2.36	96.58±0.43	76.94±3.25	99.13±0.39	97.55±0.31
C3	72.47±0	83.92±2.96	85.2±5.46	84.22±5.84	80.2±1.62	88.72±2.03	69.65±5.79	87.81±0.62	94.68±3.22
C4	54.45±0	58.61±1.53	74.55±2.39	44.63±15.75	75.31±1.08	80.82±0.59	71.36±7.4	98.53±0.02	91.72±4.05
C5	64.11±0	52.83±3.21	75.13±11.77	72.98±12.98	69.6±6.22	78.14±2.55	83.99±4.44	92.36±0.2	89.67±2.96
C6	65.23±0	77.17±0.29	94.35±0.72	89.91±10.33	95.06±3.41	97.75±1.82	73.62±12.16	99.92±0.11	99.41±0.71
C7	75.5±0	75.34±2.14	84.64±4.05	98.33±1.53	87.08±1.09	95.15±1.22	63.61±14.69	95.85±1.99	95.9±0.66
C8	87.33±0	85.83±0.11	96.97±1.44	94.3±7.86	95.4±1.88	99.08±0.76	76.35±7.53	99.41±0.6	99.74±0.33
C9	87.94±0	91.65±0.29	97.83±0.82	99.87±0.23	96.21±1.07	99.98±0.03	74.55±23.64	99.76±0.05	100±0
C10	96.01±1.73	93.87±0.09	98.84±1	97.05±3.76	96.13±1.85	99.92±0.07	80.07±3.3	100±0	100±0
C11	96.03±0	98.77±0.17	99.14±0.37	98.24±1.65	99.64±0.29	98.92±0.34	94.41±4.86	100±0	98.53±0.4
C12	93.75±0.01	94.08±1.85	99.17±0.28	99.37±0.63	98.19±0.04	98.95±0.18	71.55±0.2	99.64±0.11	99.32±0.03
C13	99.72±0	99.8±0.13	100±0	100±0	100±0	99.97±0.05	91.96±0.11	100±0	99.97±0.05
OA(%)	87.96±0	89.33±0.65	94.52±0.9	93.97±2.44	94.12±0.27	96.76±0.51	79.72±4.31	98.34±0.46	**98.49±0.21**
AA(%)	82.55±0	84.03±0.95	92.15±1.87	90.13±3.65	91.23±0.75	94.9±0.2	78.17±4.24	**97.87±0.08**	97.42±0.25
K×100	86.59±0	88.13±0.73	93.9±1	93.29±2.71	93.45±0.31	96.4±0.57	77.34±4.7	98.24±0.46	**98.33±0.23**
Params	-	563.152k	327.229k	22.309M	539.732k	338.187k	5.122M	335.369k	454.215k
Runtimes(s)	-	18	73	63	174	129	45	296	160

The bold entities means that this method has the best result of the comparison methods.

TABLE VI
CLASSIFICATION RESULTS OF SV DATASET USING 0.5% TRAINING SAMPLES (VALUE ± STANDARD DEVIATION)

Class	SVM	CDCNN	SSRN	PyResNet	DBMA	DBDA	Hybrid-SN	A2S2K-ResNet	Proposed
C1	99.42±0	96.74±3.04	97.18±2.47	88.79±17.63	98.52±2.52	99.73±0.23	95.7±3.37	99.99±0.02	99.53±0.66
C2	98.79±0	96.48±0.56	98.86±1.11	95.17±8.36	99.62±0.32	99.17±0.82	95.51±4.03	99.9±0.03	100±0.01
C3	87.98±0	89.53±1.78	94.25±2.13	85.99±18.68	96.81±0.54	97.47±0.31	99.38±0.31	94.95±3.28	98.67±0.47
C4	97.54±0	95.55±0.06	97.64±0.12	94.15±8.46	92.15±1.52	94.3±0.8	95.14±0.05	98.09±0.21	95.09±0.14
C5	95.06±0.06	96.08±2.51	97.26±1.5	99.13±0.79	96.74±0.94	98.14±1.23	98.72±0.78	98.6±0.1	99.45±0.13
C6	99.9±0	97.34±0.45	99.94±0.04	99.99±0.02	99.32±0.48	99.86±0.16	96.46±2.96	99.9±0.12	99.99±0.01
C7	95.6±0.01	92.89±4.01	99.34±0.35	99.63±0.64	97.68±0.65	98.32±0.27	99.33±0.34	99.97±0.04	98.96±0.21
C8	72.16±0.71	80.44±0.35	85.27±4.58	83.76±10.17	89.38±1.17	91.82±2.63	95.41±1.07	88.07±0.01	93.76±0.6
C9	98.08±0	98.59±0.11	99.38±0.12	99.6±0.34	99.15±0.25	99.07±0.07	99.55±0.13	99.9±0.01	99.16±0.1
C10	85.39±0	86.82±0.84	95.36±0.6	95.07±1.68	93.89±0.86	97.52±0.85	96.99±0.27	97.35±1.44	98.43±0.77
C11	86.98±0	82.65±2.27	95.81±0.26	88.65±10.85	93.62±0.85	95.74±0.26	90.54±4.2	97.33±0.42	96.7±0.4
C12	94.2±0	95.78±0.57	98±0.42	99.93±0.06	97.77±1.6	98.84±0.69	98.24±1.03	98.51±0.23	99.29±0.13
C13	93.43±0	96.88±0.44	98.23±1.07	99.16±1	98.27±0.91	99.49±0.23	87.89±3.54	97.77±2.49	99.84±0.17
C14	92.03±0	92.21±0.18	96.8±1.46	99.34±0.49	95.94±0.54	95.54±0.41	92.52±2.77	95.61±2.31	96.68±0.28
C15	71.02±0	72.84±1.73	82.34±3.5	87.93±5.54	83.02±1.06	83.22±4.71	96.92±2.22	88.44±0.74	89.53±0.37
C16	97.82±0	97.8±0.78	99.54±0.29	94.26±6.17	99.03±0.28	99.98±0.01	99.66±0.19	99.63±0.08	54.96±63.7
OA(%)	86.98±0	88.36±0.28	92.04±0.96	92.73±1.9	92.95±0.33	93.74±0.74	96.06±1.18	95.15±0.31	**96.28±0.14**
AA(%)	91.56±0	91.95±0	95.95±0.21	94.41±0.63	95.68±0.2	96.76±0.17	96.14±0.6	97.13±0.32	**97.82±0.04**
K×100	85.45±0	87.05±0.3	91.14±1.08	91.92±2.09	92.16±0.34	93.05±0.8	95.95±1.31	94.6±0.34	**95.85±0.16**
Params	-	1.8758M	370.312k	21.808M	621.407k	389.622k	5.122M	83.771k	505.650k
Runtime(s)	-	34	129	650	230	225	112	72	265

The bold entities means that this method has the best result of the comparison methods.

TABLE VII
CLASSIFICATION RESULTS OF HS DATASET USING 2% TRAINING SAMPLES (VALUE ± STANDARD DEVIATION)

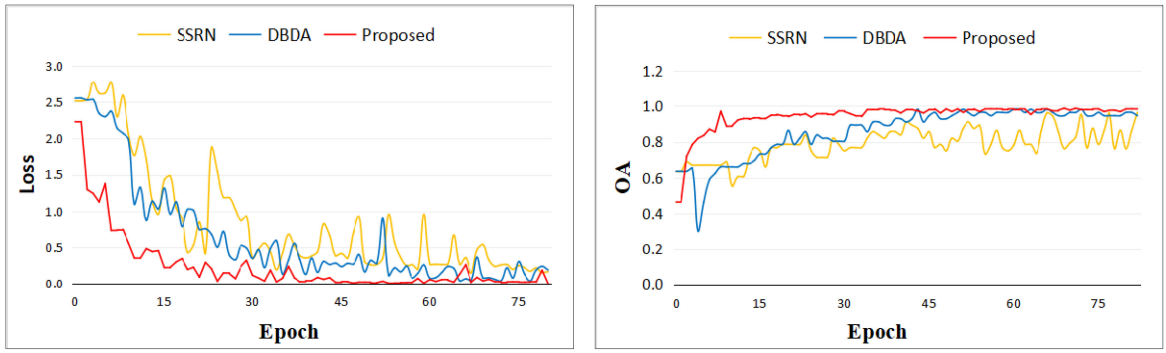
Class	SVM	CDCNN	SSRN	PyResNet	DBMA	DBDA	Hybrid-SN	A2S2K-ResNet	Proposed
C1	92.96±0	77.22±3.37	86.44±6.14	87.92±1	88.51±2.26	89.61±1.7	88.07±2.87	90.72±2.43	91.49±0.68
C2	94.04±0	91.71±4.34	93.87±3.7	91.71±3.5	95.57±1.56	97.12±2.38	95.97±1.97	97.62±1.31	94.69±5.26
C3	99.65±0	72.39±1.36	99.8±0.29	98.02±1.97	100±0	100±0	97.79±0.47	99.63±0.1	100±0
C4	98.58±0	84.75±4.22	96.35±0.86	93.32±1.32	98.51±0.44	98.47±0.37	94.38±2.05	96.51±1.88	99.11±0.2
C5	91.41±0	94.22±2.11	94.5±0.91	91.87±1.01	96.58±1.76	97.74±0.01	94.88±1.28	95.99±0.28	97.71±0.68
C6	99.56±0	79.67±10.22	100±0	95.65±0.49	99.66±0.24	98.83±0.44	96.22±1.88	98.57±1.32	98.08±1.34
C7	75.97±0	81.4±4.37	80.62±1.82	78.25±1.56	85.32±1.57	87.2±2.24	87.03±0.33	93.46±0.18	88.27±0.71
C8	75.86±0	82.26±1.26	86.33±0.66	93.27±3.3	94.41±0.78	95.54±1.9	87.41±2.06	95.25±1.22	93.25±2.91
C9	73.68±0	83.23±2.81	91.02±0.2	73.53±4.19	85.83±0.37	86.86±0.34	81.96±0.47	87.48±0.65	89.05±0.7
C10	74.88±0	64.19±2.14	78.69±1.75	65.26±10.76	90.19±1.33	82.11±1.58	83.04±0.38	78.42±0.66	86.17±1.25
C11	76.63±0	73.88±1.57	84.48±0.33	65.56±7.57	86±0.68	93.95±2.81	87.89±5.56	90.87±1.77	94.69±2.28
C12	73.56±0	81.21±3.57	84.01±5.97	70.12±12.14	88.75±2.38	90.12±1.46	86±0.59	91.47±0.43	91.61±1.68
C13	53.28±0	82.37±1.05	88.35±0.54	93.03±8.68	85.89±0.84	90.66±0.13	93.33±1.41	92.03±2.42	88.56±5.4
C14	88.57±0	82.18±3.8	95.29±4.29	94.41±0.78	98.86±0.11	98.52±0.01	91.43±0.78	97±0.35	97.4±1.57
C15	99.19±0	83.36±2.74	96.76±0.31	94.83±4.04	95.91±0.15	96.15±0.14	96.59±1.79	98.27±1.03	95.84±0.43
OA(%)	84.12±0	79.06±1.94	88.09±2.02	80.09±1.66	90.73±0.95	92.17±0.08	89.31±0.77	92.18±0.74	**92.75±0.05**
AA(%)	84.52±0	80.9±1.58	90.43±1.15	85.79±0.54	92.33±0.65	93.53±0.01	90.87±0.77	93.55±0.41	**93.73±0.48**
K×100	82.81±0	77.2±2.33	87.12±2.18	78.45±1.8	89.98±1.03	91.53±0.09	88.43±0.83	91.55±0.8	**92.15±0.06**
Params	-	812.559k	278127	22.211M	447.046k	280.021k	2.504M	258.199k	396.089k
Runtime(s)	-	29	76	73	253	107	33	154	259

The bold entities means that this method has the best result of the comparison methods.

based on dual branch and dual attention is slightly higher than that of SSRN, PyResNet, and DBMA. It is worth noting that Hybrid-SN performs relatively well only on SV datasets, but poor on other datasets. Similarly, although the AA of the latest A2S2K-ResNet method is slightly higher than that of the proposed method on KSC dataset, its overall performance is always poor on other datasets. Compared with the above methods, the proposed method has the highest classification accuracy. In the four datasets, the OA obtained by the proposed method is 1.81%, 1.01%, 1.73%, and 2.54% higher than the OA obtained by the DBDA method, respectively. In particular, DBMSA achieved 100% classification accuracy in C9 (Spartina marsh) and C10 (Cattail marsh) in the KSC dataset, and C2 (Brocoil_green_weeds_2) in the SV dataset. Figs. 11–15 and Tables III–VII prove the effectiveness of the proposed method.

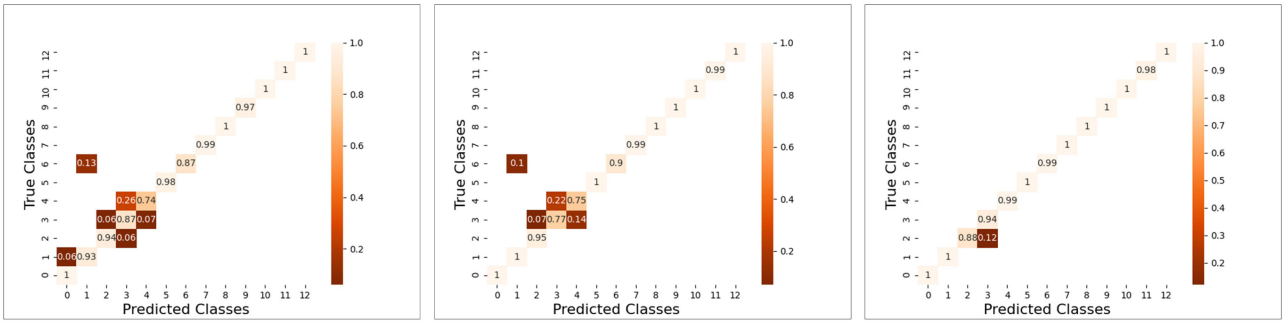
From Tables III–VII, it can be seen that the amount of parameters and running time of the proposed network are moderate. Compared with PyResNet and Hybrid-SN, the amount of parameters of the proposed method is greatly reduced. Compared with those of DBMA and DBDA, the running time is similar, but our method can provide a superior ability of classification performance.

1) Experiment 2: Fig. 16 compares the convergence of verification accuracy and loss on the KSC verification set of



(a) (b)

Fig. 16. Comparison of the loss and verification accuracy curves of each method on KSC dataset. (a) Relationship between verification loss and epochs. (b) Relationship between verification accuracy and epochs.



(a) (b) (c)

Fig. 17. Comparison of confusion matrices of different methods on KSC dataset. (a) SSRN. (b) DBDA. (c) Proposed.

SSRN, DBDA, and the proposed method over 80 epochs. It can be seen that compared with SSRN and DBDA methods, the proposed method converges faster, and it has converged in about 30 generations. Since the SSRN network is deeper, the convergence speed is slower. For DBDA, although the model has fewer parameters, it has a double-branch structure, which makes the convergence of this method slower.

To further verify the effectiveness of the proposed method, the confusion matrices obtained by the above three method on KSC dataset are compared, and the experimental results are shown in Fig. 17. For SSRN method, the classification errors of Slash pin and Oak/Broadleaf are relatively large. Among them, the confusion ratio of true category Slash Pin with CP hammock and Oak/Broadleaf is 6% and 7%, respectively, and the classification error rate of real category Oak/Broadleaf is 26%. For DBDA, CP hammock, Slash pine, and Oak/Broadleaf, all have some confusion, and the classification accuracy of Slash pine and Oak/Broadleaf is poor, with only 77% and 75% accuracy. Compared with the above two methods, the classification accuracy of the proposed method is 100% for most categories, and the classification accuracy of Slash pine and Grass-pasture-mowed can reach more than 94%. This shows that the proposed method still has good classification performance for those easily confused categories.

1) Experiment 3: This experiment compares the classification performance of different methods under different training sample ratios. For the datasets of IN, UP, KSC, and SV, the training ratios of each dataset are set to 1%, 5%, 10%, 15%, and 20%, and SVM, CDCNN, SSRN, PyResNet, DBMA, DBDA, Hybrid-SN, A2S2K-ResNet, and the proposed DBMSA method are tested. The experimental comparison results are shown in Fig. 18. It can be seen that the classification performance of CDCNN and SVM is relatively poor when there are few training samples. For the four datasets, the best overall classification performance is achieved by the proposed DBSMA method. Although the classification accuracy of Hybrid-SN is slightly higher than that of the proposed DBSMA method on SV datasets, the generalization ability of this method is poor. Compared with SSRN, PyResNet, and DBMA, A2S2K-ResNet can achieve relatively good results as a whole, but this method performs poorly in the case of fewer samples. With the increase in the number of samples, each method can achieve higher classification accuracy, but the classification accuracy of the proposed DBSMA method is still the highest. It proves that the proposed method has better generalization ability.
2) Experiment 4: To explore the influence of the input spatial size on the experiment, many experiments with the spatial size of 5×5, 7×7, 9×9, 11×11, and 13×13 have been performed. The experimental results are shown in Table VIII. It is worth noting that the classification accuracy first increases and then decreases with the increase in size.

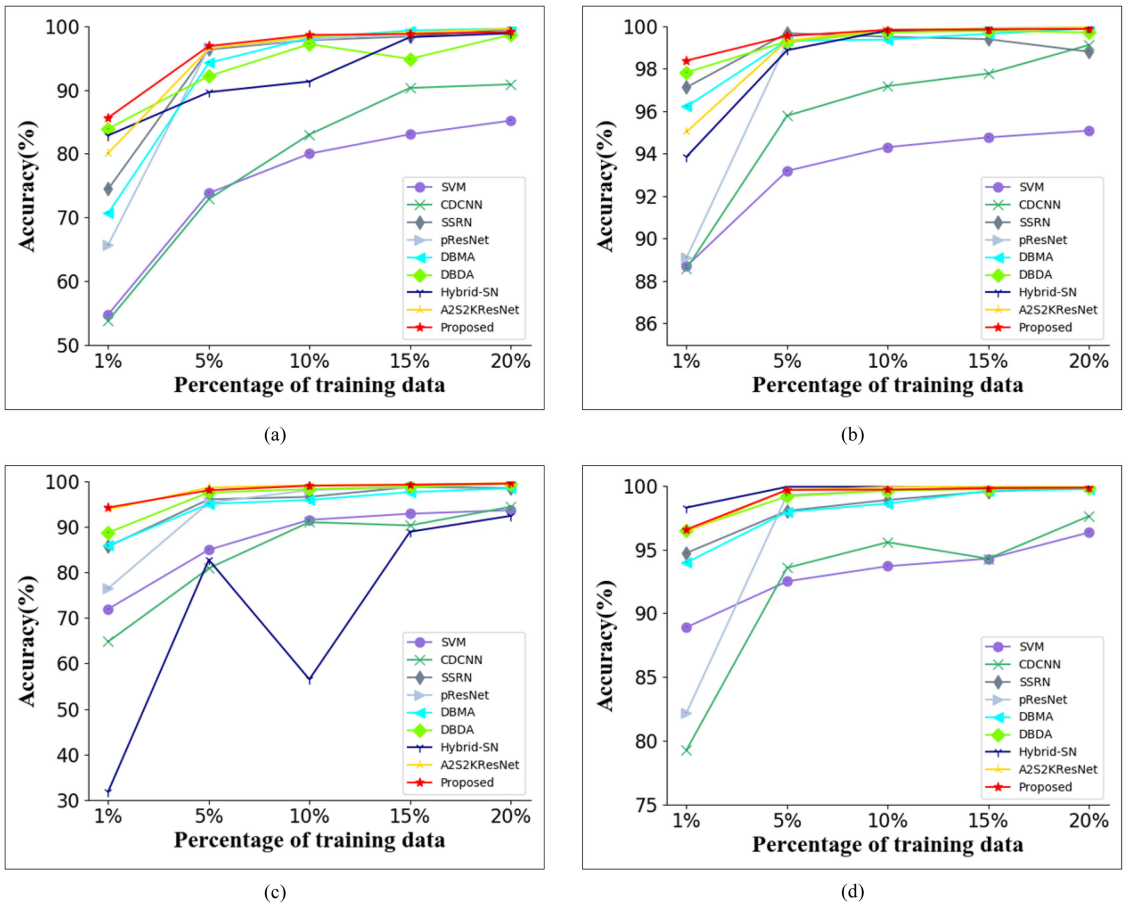


Fig. 18. Comparison results of the classification performance of different methods at different training sample ratios on the IN, UP, KSC, and SV datasets. (a) Classification performance of different methods on the IN dataset. (b) Classification performance of different methods on the UP dataset. (c) Classification performance of different methods on the KSC dataset. (d) Classification performance of different methods on the SV dataset.

When the spatial size is 9×9, the classification accuracy is the best. Therefore, the spatial size of 9×9 is adopted as the input size of the proposed framework.

1) Experiment 5: In addition, we extensively analyzed the different effects of the proposed MSSP block and attention mechanism. In this part, a series of comparative experiments are carried out to illustrate the advantages of MSSP block. Specifically, MSSP blocks are equipped with grouping and without grouping. Table IX shows the classification results of different module combinations on five datasets. It can be observed that the best performance is obtained by combining the grouped MSSP block with the two attention mechanisms, which shows that the scheme has general advantages for all datasets. The classification accuracy of MSSP Block with grouping is improved by 10.37%, 4.61%, 2.89%, 7.22%, and 3.54%, respectively, on IN, UP, KSC, SV, and HS datasets compared with those of other schemes without MSSP Block.

IV. Conclusion

This article proposes a dual-branch spectral multiscale attention network for hyperspectral image classification. It consists of two branches, i.e., spectral branch and spatial branch. In the spectral branch, the structure of the MSSP and the spectral attention mechanism is designed to extract the spectral information. In the spatial branch, the structure of the dense connection block and the spatial attention mechanism is utilized to extract the spatial information. In addition, the features obtained from the two branches are fused and classified. The proposed MSSP of the DBMSA network can obtain the spectral features of different receptive fields, which is beneficial to improve the classification performance of hyperspectral images. The experimental results show that the network model proposed in this article has a good classification performance and strong generalization ability. In future research, we plan to further improve the DBMSA method to more effectively extract the features of hyperspectral images and reduce the running time of it.

Acknowledgment

The authors would like to thank the handling editor and the anonymous reviewers for their careful reading and helpful comments, which are all very valuable for improving the quality of this article. In addition, the authors would like to thank Prof. D. Landgrebe for providing the Indian Pines dataset, Prof. P.

TABLE VIII
CLASSIFICATION ACCURACY ON EACH DATASET WITH DIFFERENT SPATIAL SIZES

		5×5	7×7	9×9	11×11	13×13
IN	**OA(%)**	92.87	94.42	**95.39**	91.96	90.55
	AA(%)	94.23	94.04	**94.42**	87.02	89.52
	Kappa×100	91.88	93.65	**94.74**	90.84	89.96
UP	**OA(%)**	96.28	96.45	**97.02**	96.21	95.38
	AA(%)	95.87	96.16	**96.81**	95.76	94.48
	Kappa×100	95.07	95.29	**96.05**	94.97	93.85
KSC	**OA(%)**	97.22	98.22	**98.49**	97.38	97.19
	AA(%)	96.00	96.94	**97.42**	95.85	95.35
	Kappa×100	96.90	98.02	**98.33**	97.19	96.66
SV	**OA(%)**	95.13	96.08	**96.28**	95.23	95.01
	AA(%)	97.20	97.50	**97.82**	94.68	94.89
	Kappa×100	94.57	**95.97**	**95.85**	95.08	93.55
HS	**OA(%)**	91.50	92.41	**92.75**	91.99	91.89
	AA(%)	92.71	93.50	**93.73**	93.20	92.11
	Kappa×100	90.89	91.88	**92.15**	91.34	91.13

The bold entities means that this method has the best result of the comparison methods.

TABLE IX
ABLATION ANALYSIS OF DIFFERENT MODULES (OA%)

Module							
Module	MSSP Block	No groups		√	√	√	
		Groups					√
	Spectral attention		√	√		√	√
	Spatial attention		√		√	√	√
Data	IN		85.44	74.50	94.07	94.89	**95.81**
	UP		92.89	86.40	95.84	96.65	**97.50**
	KSC		95.60	95.42	96.81	97.24	**98.49**
	SV		89.06	90.23	95.58	95.80	**96.28**
	HS		89.21	90.80	91.69	91.96	**92.75**

The bold entities means that this method has the best result of the comparison methods.

Gamba for providing the UP dataset, Prof. Melba Crawford for providing the KSC dataset, and the Hyperspectral Image Analysis Laboratory, University of Houston, the IEEE GRSS Image Analysis and Data Fusion Technical Committee for providing the University of Houston dataset.

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[2] F. van der Meer, "Analysis of spectral absorption features in hyperspectral imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 5, no. 1, pp. 55–68, Feb. 2004.

[3] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 144–153, Jan. 2015.

[4] A. Ghiyamat and H. Z. Shafri, "A review on hyperspectral remote sensing for homogeneous and heterogeneous forest biodiversity assessment," *Int. J. Remote Sens.*, vol. 31, no. 7, pp. 1837–1856, 2010.

[5] X. Wang, Y. Kong, Y. Gao, and Y. Cheng, "Dimensionality reduction for hyperspectral data based on pairwise constraint discriminative analysis and nonnegative sparse divergence," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1552–1562, Apr. 2017.

[6] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge preserving features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7140–7151, Dec. 2017.

[7] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[8] W. Sun, G. Yang, B. Du, L. Zhang, and L. Zhang, "A sparse and low rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4032–4046, Jul. 2017.

[9] F. Luo, H. Huang, Z. Ma, and J. Liu, "Semi-supervised sparse manifold discriminative analysis for feature extraction of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6197–6211, Oct. 2016.

[10] C. Cariou and K. Chehdi, "A new k-nearest neighbor density-based clustering method and its application to hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 6161–6164.

[11] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[12] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[13] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[14] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[15] J. Feng *et al.*, "Attention multibranch convolutional neural network for hyperspectral image classification based on adaptive region search," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5054–5070, Jun. 2020.

[16] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[17] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.

[18] 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.

[19] 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. 6232–6251, Oct. 2016.

[20] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 298.

[21] 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.
[22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[23] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semi-supervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.
[24] Y. Zhan *et al.*, "Semi-supervised classification of hyperspectral data based on generative adversarial networks and neighborhood majority voting," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 5756–5759.
[25] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.
[26] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[27] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of hyperspectral images based on multi-class spatial–spectral generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5329–5343, Aug. 2019.
[28] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "CVA2E: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5676–5692, Aug. 2020.
[29] J. Feng *et al.*, "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, Apr. 2020, Art. no. 1149.
[30] F. F. Shahraki and S. Prasad, "Graph convolutional neural networks for hyperspectral data classification," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, Nov. 2018, pp. 968–972.
[31] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semi-supervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.
[32] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.
[33] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sens. Lett.*, vol. 8, no. 5, pp. 438–447, May 2017.
[34] 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. 6232–6251, Oct. 2016.
[35] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.
[36] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1454.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 770–778.
[38] 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. 2017.
[39] 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, 2018, Art. no. 1068.
[40] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.
[41] Z. 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.
[42] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vision*, Oct. 2018, pp. 3–19.
[43] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Noise-robust hyperspectral image classification via multi-scale total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1948–1962, Jun. 2019.
[44] S. Fang, D. Quan, S. Wang, L. Zhang, and L. Zhou, "A two-branch network with semi-supervised learning for hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 3860–3863.
[45] B.-S. Liu and W.-I. Zhang, "Multi-Scale convolutional neural networks aggregation for hyperspectral images classification," in *Proc. Symp. Piezoelectrcity, Acoust. Waves Device Appl.*, Jan. 2019, pp. 1–6.
[46] S. Wu, J. Zhang, and C. Zhong, "Multiscale spectral-spatial unified networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul./Aug. 2019, pp. 2706–2709.
[47] K. Pooja, R. R. Nidamanuri, and D. Mishra, "Multi-scale dilated residual convolutional neural network for hyperspectral image classification," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens.*, Sep. 2019, pp. 1–5.
[48] H. Zhu, Y. Miao, and X. Zhang, "Semantic image segmentation with improved position attention and feature fusion," *Neural Process. Lett.*, vol. 52, pp. 329–351, May 2020.
[49] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, vol. 14, no. 11, pp. 2168–2275, Nov. 2019.
[50] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 773–786, Mar. 2019.
[51] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, Apr. 2019, Art. no. 884.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[53] 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, 2019, Art. no. 1307.
[54] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 3146–3154.
[55] 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, Feb. 2020, Art. no. 582.
[56] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.
[57] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
[58] S. K. Roy, S. Manna, and T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2020.
[59] D. Misra, "Mish: A self regularized non-monotonic activation function," ArXiv, vol. abs/1908.08681, Aug. 2019. [Online]. Available: http://arxiv.org/abs/1908.08681
[60] R. Li and C. Duan, "LiteDenseNet: A lightweight network for hyperspectral image classification," ArXiv, vol. abs/2004.08112, Apr. 2020. [Online]. Available: http://arxiv.org/abs/2004.08112

Cuiping Shi (Member, IEEE) received the M.S. degree in signal and information processing from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

From 2017 to 2020, she was a postdoctoral researcher with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China. She is currently an Associate Professor with the Department of Communication Engineering, Qiqihar University, Qiqihar, China. Her main research interests include remote sensing image processing, pattern recognition, and machine learning. She has authored or coauthored two academic books in remote sensing image processing and more than 50 papers in journals and conference proceedings.

Dr. Shi's doctoral dissertation won the Nomination Award of Excellent Doctoral Dissertation of Harbin University of Technology (HIT) in 2016.

Diling Liao received the bachelor's degree in communication engineering from Zhuhai College of Jilin University, Zhuhai, China, in 2019. He is currently working toward the master's degree in information and communication system with Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

Yi Xiong is currently working toward the bachelor's degree in communication engineering with Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

Mr. Xiong's research project won one provincial students awards.

Tianyu Zhang received the bachelor's degree in electronic information science and technology from Qufu Normal University, Qufu, China, in 2019. She is currently working toward the master's degree with Qiqihar University, Qiqihar, China.

Her research interests include hyperspectral image processing and machine learning.

Liguo Wang (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a postdoctoral research position with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China, where he is currently a Professor. From 2020, he has been with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China. His main research interests include remote sensing image processing and machine learning. He has authored or coauthored two books in hyperspectral image processing and more than 130 papers in journals and conference proceedings.