

Pyramidal Multiscale Convolutional Network With Polarized Self-Attention for Pixel-Wise Hyperspectral Image Classification

Haimiao Ge¹, Ligu Wang¹, Moqi Liu¹, *Graduate Student Member, IEEE*, Xiaoyu Zhao,
Yuexia Zhu, Haizhu Pan², and Yanzhong Liu

Abstract—In recent years, pixel-wise hyperspectral image (HSI) classification has received growing attention in the field of remote sensing. Plenty of spectral-spatial convolutional neural network (CNN) methods with diverse attention mechanisms have been proposed for HSI classification due to the attention mechanisms being able to provide more flexibility over standard convolutional blocks. However, it remains a challenge to effectively extract multiscale features of high-resolution HSI in a real-world complex environment. In this article, we propose a pyramidal multiscale spectral-spatial convolutional network with polarized self-attention for pixel-wise HSI classification. It contains three stages: channel-wise feature extraction network, spatial-wise feature extraction network, and classification network, which are used to extract spectral features, extract spatial features, and generate classification results, respectively. Pyramidal convolutional blocks and polarized attention blocks are combined to extract spectral and spatial features of HSI. Furthermore, residual aggregation and one-shot aggregation are employed to better converge the network. The experimental results on several public HSI datasets demonstrate that the proposed network outperforms other related methods.

Index Terms—Convolutional neural network (CNN), hyperspectral image (HSI) classification, multiscale feature extraction, polarized self-attention (PSA) mechanism.

I. INTRODUCTION

HYPERSPECTRAL image (HSI) is obtained by the remote sensor and contains hundreds of continuous and narrow spectral bands ranging from visible to short-wave infrared. HSI can effectively characterize interesting land cover objects [1] and has been widely used in many research fields, such as urban planning [2], environmental monitoring [3], fine agriculture [4], mineral exploration [5], [6], and military

Manuscript received 22 September 2022; revised 28 November 2022 and 4 January 2023; accepted 27 January 2023. Date of publication 14 February 2023; date of current version 24 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071084, in part by the Leading Talents Project of the State Ethnic Affairs Commission, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities under Grant 145109218. (*Corresponding author: Ligu Wang.*)

Haimiao Ge, Moqi Liu, Xiaoyu Zhao, Yuexia Zhu, Haizhu Pan, and Yanzhong Liu are with the College of Computer and Control Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: 01557@qqhru.edu.cn; 2020935653@qqhru.edu.cn; 2021935730@qqhru.edu.cn; zyxyee@163.com; panhaizhu@qqhru.edu.cn; 01514@qqhru.edu.cn).

Ligu Wang is with the College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China (e-mail: wangliguo@hrbeu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3244805

targets [7]. With the rapid development of remote sensing technology and hyperspectral imaging technology, it has been easier to acquire HSI datasets. However, the analysis and process of the HSI datasets remain insufficient [8].

The pixel-wise classification of HSI, which appears as an important issue of HSI processing technology, achieves a phenomenal interest of researchers and has been studied by many scholars in recent years [9], [10]. The purpose of the pixel-wise classification is to assign a unique category label to each pixel of the HSI dataset. Traditional machine learning HSI classification approaches use handcrafted features to train the classifier, such as local binary patterns (LBPs) [11], histogram of oriented gradients (HOG) [12], global image scale-invariant (GIST) [13], K-nearest neighbors (KNN) [14], extreme learning machine (ELM) [15], and support vector machine (SVM) [16]. Although these handcrafted features can effectively represent various shallow attributes of HSI, the robustness and discriminability of the methods are difficult to be maintained in complex real-world remote sensing environments. Furthermore, the parameter setting and domain knowledge also limit the usage of handcrafted features in HSI classification tasks. In contrast, deep learning methods can automatically learn the shallow features and deep semantic information from HSI dataset in a hierarchical manner, which has shown great potential for feature representation in HSI classification tasks [17], [18].

In recent years, many deep learning-based frameworks have been proposed, such as recurrent neural networks (RNNs) [19], convolutional neural networks (CNNs) [20], graph convolutional neural networks (GCNNs) [21], and generative adversarial neural networks (GANNs) [22]. Among these frameworks, the CNN framework, which has been widely used in RGB image processing, is applied to pixel-wise HSI classification for its excellent performance. CNN employs spatial weight sharing of the convolutional kernel to reduce the computational complexity and uses activation functions to add nonlinearities to the network. According to the extracted features, the CNN-based frameworks can be divided into three types: spectral CNN, spatial CNN, and spectral-spatial CNN [23]. The spectral CNNs take advantage of the abundant spectral signature of HSI and exploit the spectral features (1-D vector) to improve the classification accuracy. For example, Hu et al. [24] proposed deep CNNs to classify HSIs directly

in the spectral domain. Five layers are implemented on each spectral signature to discriminate against others. The experimental results show that the proposed method can obtain better accuracy than some traditional methods. In [25], each 1-D spectral vector of a pixel is transformed into a 2-D spectral feature matrix to get rid of the bondage of strong correlation among bands for HSI classification. The 1×1 and 3×3 convolutional layers are implemented in the CNN framework to better deal with HSI information and accomplish feature reuse. Jin et al. [26] propose a deep neural network classification model for the pixels of wheat HSI to accurately discern the disease areas. In this model, the pixel spectra data are reshaped into a 2-D data structure. A hybrid network framework with a convolutional layer and bidirectional recurrent layer is reconstructed to improve the generalization of the model. In [27], comparisons are conducted among KNN, SVM, and CNN models in the spectral dimension of HSIs over four rice seed varieties. The result shows that the CNN model performs better than the corresponding KNN and SVM in most cases. Although spectral CNNs achieve better results than traditional classification methods, the CNNs are constrained to extract the spectral signatures of HSI, while the spatial information is insufficiently utilized. In contrast, spatial CNN models employ a spatial map (2-D matrix) as the input data to extract the spatial information from the HSI dataset. For example, Li et al. [28] use principal component analysis (PCA) to extract the first principal component (PC) with refined spatial information and propose a full CNN with convolution, deconvolution, and pooling layers to enhance the deep features. After the feature enhancement, the optimized ELM is utilized for classification. Xu et al. [29] propose a random patches network for HSI classification, which uses 2-D convolutional kernels in the CNN framework. In [30], Gabor filters are employed to combine with the 2-D convolutional filters for HSI classification to mitigate the problem of overfitting. The classification results show that the proposed model provides competitive results. In [31], a spatial CNN framework is proposed for HSI classification embedded with an extracted hashing feature. The proposed CNN achieves a powerful distinguishing ability from different classes. Although spatial CNNs can effectively extract the spatial information of HSI pixels to improve the classification accuracy, CNNs inevitably lose a large amount of spectral information. To avoid this problem, spectral-spatial CNN is naturally implemented for pixel-wise HSI classification, which can jointly extract spectral and spatial information from the HSI dataset. The input data of the spectral-spatial CNN is 3-D cube data, which is always a square HSI data cube cropped centered on the corresponding pixel. The spectral-spatial CNN has greatly improved the classification accuracy and is the dominant research of the HSI classification. For example, Li et al. [32] proposed a 3-D CNN framework to extract the deep spectral-spatial combined features of the HSI dataset. The experimental results show that the proposed 3-D CNN method outperforms the stacked autoencoder, deep brief network, and 2-D CNN network. Zhong et al. [33] design an end-to-end spectral-spatial residual network (SSRN) that takes raw 3-D cubes as input data for HSI classification. The residual blocks of the network consec-

tively learn discriminative features from spectral signatures and spatial contexts in HSI. The experimental results show that the proposed network achieves competitive HSI classification accuracy in agricultural, rural-urban, and urban datasets. Zhang et al. [34] propose a 3-D lightweight CNN for limited-samples-based HSI classification. Two learning strategies are proposed to further alleviate the small sample problem, which are the cross-sensor strategy and the cross-modal strategy. Experiments demonstrate that the proposed network achieves competitive performance for HSI classification. Roy et al. [35] propose a bilinear fusion mechanism for HSI classification. The excitation operation is performed using the fused output of the squeeze operation. The experimental results confirm the superiority of the proposed method. Jia et al. [36] also propose a lightweight CNN. The spatial-spectral Schrodinger eigenmaps feature extraction is first adopted to obtain the joint spatial-spectral information. A dual-scale convolutional module is designed to address the spatial-spectral features and obtain the hierarchical structure description of the dataset. The features are addressed by a bichannel fusion module and are imported into a global average pooling classifier to achieve the classification results.

Although the spectral-spatial CNN has significantly improved the accuracy of HSI classification, there are still some problems to be solved, such as convergence of deep network, limitation of labeled samples, and extraction of complex land cover objects [37]. To address these issues, scholars strive to optimize existing spectral-spatial CNN frameworks. To solve the problem of deep network convergence, residual blocks and densely connected structures are introduced to improve the CNN frameworks. For instance, Wang et al. [38] propose a fast dense spectral-spatial convolutional framework for HSI classification. Different convolutional kernel sizes are used to extract spectral and spatial features separately. Densely connected structures are used for deep learning of features. Paoletti et al. [39] propose a residual-based CNN approach, which is grouped in pyramidal bottleneck residual blocks, to involve more locations as the network depth increases to preserve the time complexity per layer. Meanwhile, the multiscale strategy is implemented to construct the CNNs to better use the limited samples and extract features of complex land cover objects. For example, Liu et al. [40] propose a 2-D-3-D CNN with spectral-spatial multiscale feature fusion for HSI classification. The network employs two diverse backbone modules for feature representation. A hierarchical feature extraction module is used to capture multiscale spectral features, and a multilevel fusion structure is used to extract multistage spatial features. In [41], a multiscale self-looping CNN is proposed for HSI classification. Each layer in a self-looping block contains both forward and backward connections, which can efficiently fuse the shallow and deep features extracted by different layers. Furthermore, the dual-branch strategy is introduced to the spectral-spatial CNN framework. Wang et al. [42] propose a dual-branch dense residual network for HSI classification. One branch is based on 1-D convolution, which is used to extract spectral features. Another branch is based on 2-D convolution, which is used to extract spatial features. Residual units and dense structures

are introduced to fuse the information of different convolutional layers. The experimental results show that the proposed method achieves superior classification performance compared with the state-of-the-art methods. At the same time, attention mechanisms are employed to combine with convolutional layers to make the network more flexible. Li et al. [43] propose a spectral–spatial network with channel and position global context attention (SSGC) for HSI classification. Pan et al. [1] propose a one-shot dense network (OSDN) with polarized attention for HSI classification. The one-shot units are used to maintain the information of different layers, and the polarized attention is used to extract the high internal resolution spectral and spatial information. It is worth noting that the aggressive improvements effectively enhance the performance of spectral–spatial CNN frameworks, and the improvements of spectral–spatial CNNs are not limited to the abovementioned methods.

In this article, we propose a pyramidal multiscale spectral–spatial CNN (PMCN) with polarized attention for pixel-wise HSI classification. The proposed network contains three stages: channel-wise feature extraction network, spatial-wise feature extraction network, and classification network. The channel-wise feature extraction network is used to extract the spectral features of the HSI dataset, and the spatial-wise feature extraction network is used to extract the spatial features. The classification network is used to obtain classification results. Pyramidal multiscale convolutional blocks and polarized self-attention (PSA) blocks are combined to extract complex spectral and spatial features with high resolution. Batch normalization (BN) [44], parametric rectified linear unit (PReLU) [45], and Mish [46] are implemented to maintain the stability and nonlinearity of the network. Furthermore, residual aggregation and one-shot aggregation are introduced to better converge the network. Finally, the classification network is used to fuse the features and obtains the classification results. The main contributions are summarized as follows.

- 1) We improve the traditional pyramidal multiscale convolutional block that uses the pseudo-3-D multiscale spectral convolutions and spatial convolutions to construct spectral feature extraction blocks and spatial feature extraction blocks, respectively. This approach can reduce the complexity of the proposed network without reducing the classification accuracy and make the network easier to be trained.
- 2) The residual aggregation and one-shot aggregation are jointly employed in the proposed network. This approach can effectively maintain the shallow feature of the low-level layers so that the network can adequately integrate the features of different layers for better convergence and improve the efficiency of the proposed network.
- 3) The polarized attention mechanism is used to help the multiscale convolutional blocks to extract spectral and spatial features. This approach can effectively extract the segment that needs to be noticed based on the characteristics of the input feature map and is an attractive complement to standard multiscale convolutional blocks at high internal resolution.

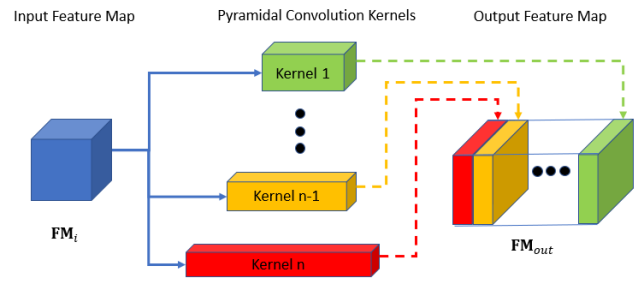


Fig. 1. Structure of PyConv.

The rest of this article is organized as follows. Section II introduces the related work, such as the cube-based HSI classification framework, pyramidal convolution (PyConv), attention mechanism, and aggregation methods. The details of the proposed network are given in Section III. Section IV lists the experimental results, and Section V makes some discussions. Section VI gives the conclusion of this article and discusses future work.

II. RELATED WORK

A. Cube-Based HSI Classification Framework

To extract the spectral–spatial features of the HSI, the cube-based method is introduced to pixel-wise HSI classification [47]. In this method, a square HSI data cube is cropped and centered on the corresponding pixel, which is utilized as the input data of the network. The land cover label of the 3-D cube is determined by its central pixel. To be specific, giving an HSI dataset $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ and the land cover label of the i th pixel $y_i \in \{1, 2, \dots, m\}$, where D is the number of channels (spectral dimensions), $H \times W$ is the spatial size of the HSI dataset, m is the number of land cover categories. The HSI data cube of the i th pixel can be described as $\mathbf{x}_i \in \mathbb{R}^{D \times h \times w}$, which is centered on the i th pixel in spatial dimension and the spatial size is $h \times w$. In general, we can denote the i th labeled pixel as (\mathbf{x}_i, y_i) .

B. Pyramidal Convolution

The PyConv [37] is a multiscale 3-D convolutional network architecture that uses a local multiscale context aggregation module and a global multiscale context aggregation block to parse the input feature map. Different from the standard convolution, PyConv enlarges the receptive field of the kernel and applies different types of kernels with different spatial and spectral resolutions in parallel. The structure of the PyConv is illustrated in Fig. 1. Given the feature map $\mathbf{FM}_i \in \mathbb{R}^{C \times h \times w}$, where C is the number of channels and $h \times w$ is the spatial size, PyConv uses different types of 3-D kernels in a pyramid that produces a series of outputs and aggregates the outputs into an output feature map $\mathbf{FM}_{out} \in \mathbb{R}^{C \times h \times w}$. In general, the size of the 3-D kernels can be varied into two directions: spatial-wise and channel-wise. As can be seen from Fig. 1, the spatial size of the kernels increases from the bottom of the pyramid to the top, and the channel size of the kernels simultaneously decreases. The pyramidal structure provides a pool of combinations with different types and sizes of

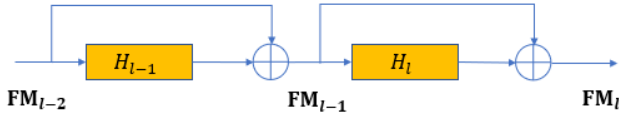


Fig. 2. Illustration of the residual aggregation.

kernels. The architecture can possess the ability to acquire complementary information so that the smaller receptive fields can focus on small objects and the larger receptive fields can dedicate feature maps to the larger objects and the contextual information.

C. Attention Mechanism

Benefiting from the human perception process, the attention mechanism is designed to focus more on the informative areas and takes less into account nonessential areas [48]. It obtains linear weights to represent the contributions to extract features based on the correlations between objects, which can be interpreted as a method of feature transformation. The attention mechanism, which is used to address the weakness of standard convolutions [49], has shown excellent performance in various tasks, such as image categorization, image caption, text-to-image synthesis, and scene segmentation [50].

Self-attention [51], [52] is a kind of attention model that uses an input tensor to compute the attention weights and reweights the input tensor by these weights. In general, it works as a standard component to capture long-range interactions. As a result, self-attention models are always inserted after convolutional blocks to augment the network to handle both short- and long-range dependence. In this article, a powerful self-attention for pixel-wise regression, named PSA [53], is introduced to the proposed network. It keeps high internal resolution and fuses SoftMax-sigmoid composition in both channel-only and spatial-only attention blocks. The detailed implementation is described in Section III-C.

D. Residual Aggregation, Dense Aggregation, and One-Shot Aggregation

In general, deep neural networks have a powerful ability to extract abstract information from input datasets that can provide effective support for downstream tasks. However, as the neural network deepens, the gradient dispersion/explosion phenomenon and network degradation phenomenon often prevent the network to be successfully converged. To address these issues, residual aggregation, also known as residual connection or identity mapping, is proposed [54]. As shown in Fig. 2, we can see that a skip connection is added to the basic traditional deep neural network. H is the hidden layer that represents several convolutional layers with BN layers and activation layers, and \oplus is a summation operator. The skip connection allows the input feature map to be passed directly to the subsequent layers in a summative way. The output feature map of the l th hidden layer can be expressed as

$$\mathbf{FM}_l = H_l(\mathbf{FM}_{l-1}) + \mathbf{FM}_{l-1}. \quad (1)$$

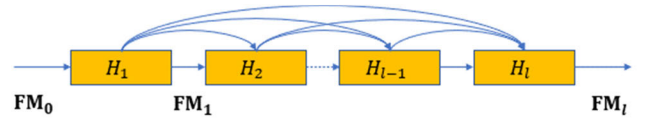


Fig. 3. Illustration of the dense aggregation.

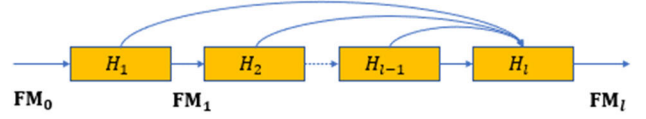


Fig. 4. Illustration of the one-shot aggregation.

As mentioned by Zhu et al. [55], information carried by early feature maps would be washed out as it is summed with others. To better maintain the early information, dense aggregation is proposed [56]. Different from residual aggregation, dense aggregation utilizes the concatenation operator to converge the feature maps that preserve information in its original form. As shown in Fig. 3, all previous feature maps of the early layers can be used to compute the output of the l th layer

$$\mathbf{FM}_l = H_l[\mathbf{FM}_0, \mathbf{FM}_1, \dots, \mathbf{FM}_{l-1}]. \quad (2)$$

We can see that if the hidden layer H_l produces k feature maps, the input of the H_{l+1} will be $k_0 + k \times l$ input feature maps, where k_0 is the size of the input dataset, while the output of H_{l+1} will still be k feature maps.

However, we find in the experiment that networks with dense aggregation spend more energy and time than those with residual aggregation. To improve the dense aggregation to be more efficient, one-shot aggregation [57] is proposed which can preserve the benefit of concatenative aggregation for feature extraction. As shown in Fig. 4, one-shot aggregation aggregates intermediate features at once. Experiments show that one-shot aggregation provides great benefits to computation efficiency while preserving the advantage of dense aggregation.

III. METHODOLOGY

In this section, we first introduce the framework of the proposed network in detail. Second, channel-wise and spatial-wise pyramidal convolutional blocks are described. Finally, the implementation of PSA blocks is discussed.

A. Framework of the PMCN

The structure of the PMCN is shown in Fig. 5. We can see that the proposed network can be divided into three parts: channel-wise feature extraction network, spatial-wise feature extraction network, and classification network. The channel-wise feature extraction network is composed of three channel-wise pyramidal convolutional blocks, one channel-only block of PSA, and four convolutional layers. Residual aggregation and one-shot aggregation are utilized to preserve early information. The spatial-wise feature extraction network layouts after the channel-wise feature extraction network. Similar

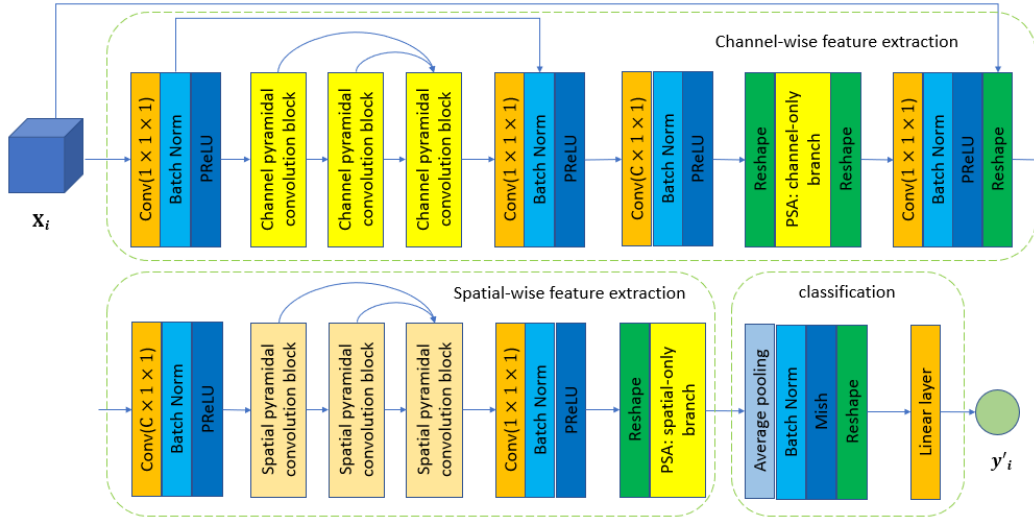


Fig. 5. Structure of the proposed network.

to the channel-wise feature extraction network, the spatial-wise feature extraction network is composed of three spatial-wise pyramidal convolutional blocks, one spatial-only block of PSA, and two convolutional layers. One-shot aggregation is implemented among the spatial pyramidal convolutional blocks. BN and PReLU are arranged in appropriate locations to maintain the stability and nonlinearity of the network. Finally, the classification network is assigned to provide the classification result, which contains an average pooling layer, BN layer, Mish, and linear layer. The average pooling layer is used to concentrate features from extracted feature maps. The BN layer is applied to stabilize the network and make the network easier to be converged. The Mish activation function is employed to provide a wider range of values for the output data. The linear layer is implemented to provide final classification results. Assuming the input data are $\mathbf{x}_i \in \mathbb{R}^{D \times h \times w}$, where \mathbf{x}_i is the cube-based HSI data of i th pixel, D is the number of channels, and $h \times w$ is the spatial size of the data, the output of the network is $y'_i \in \mathbb{R}^{1 \times m}$, where m is the number of land cover categories. To be specific, we take the input dataset $\mathbf{x}_i \in \mathbb{R}^{103 \times 15 \times 15}$ as an example to specify the data flow of the network. The detailed steps of the proposed network are shown in Table I. Cross-entropy loss is used to train the proposed network, which can be expressed as

$$L_i = -[y_i \log y'_i + (1 - y_i) \log(1 - y'_i)] \quad (3)$$

where y_i is the land cover label of the i th pixel, L_i is the cross-entropy loss of the i th pixel. In addition, early stopping and dynamic learning rate [48] technologies are also implemented to reduce the training time and provide better network convergence.

B. Channel-Wise and Spatial-Wise Pyramidal Convolutional Blocks

In the proposed network, pyramidal convolutional blocks are introduced to extract multiscale information from feature maps. Different from the traditional PyConv in which the

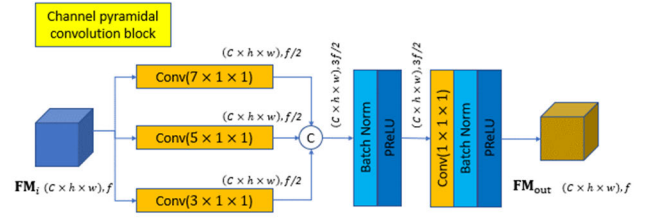


Fig. 6. Structure of the channel-wise pyramidal convolutional block.

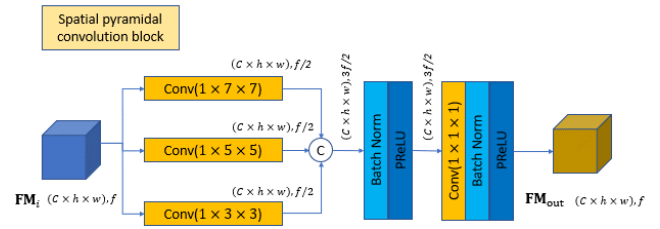


Fig. 7. Structure of the spatial-wise pyramidal convolutional block.

channel and spatial size of 3-D convolutional kernels vary jointly, we clearly separate the kernels into channel-wise kernels and spatial-wise kernels. The size of the multiscale kernels only varies in the channel or spatial dimension, which can effectively reduce the computation complexity of the network. As a result, two kinds of pyramidal convolutional blocks are conducted: channel-wise pyramidal convolutional blocks and spatial-wise pyramidal convolutional blocks and are used in the channel-wise feature extraction network and spatial-wise feature extraction network, respectively. In addition, instead of segmenting the input data as the traditional PyConv does, we use the complete input data directly for feature extraction to maintain the integrity of the feature maps.

To be specific, the structures of the channel-wise and spatial-wise pyramidal convolutional blocks are illustrated in Figs. 6 and 7. Assuming the input data are \mathbf{FM}_i , we can see that the channel-wise pyramidal convolutional block contains three convolutional layers with $(7 \times 1 \times 1)$, $(5 \times 1 \times 1)$,

TABLE I
DETAILED STEPS OF THE PROPOSED NETWORK

Input size	Layer name	Kernel Size	Filters	Output size
(103 × 15 × 15,1)	Conv-BN-PReLU	(1 × 1 × 1)	24	(103 × 15 × 15,24)
(103 × 15 × 15,24)	Channel Pyramidal	/	/	(103 × 15 × 15,24)
(103 × 15 × 15,24)	Channel Pyramidal	/	/	(103 × 15 × 15,24)
(103 × 15 × 15,24)	Channel Pyramidal	/	/	(103 × 15 × 15,24)
(103 × 15 × 15,72)	Conv-BN-PReLU	(1 × 1 × 1)	24	(103 × 15 × 15,24)
(103 × 15 × 15,24)	Conv-BN-PReLU	(103 × 1 × 1)	24	(1 × 15 × 15,24)
(1 × 15 × 15,24)	Reshape	/	/	(24 × 15 × 15)
(24 × 15 × 15)	PSA: channel-only	/	/	(24 × 15 × 15)
(24 × 15 × 15)	Reshape	/	/	(1 × 15 × 15,24)
(1 × 15 × 15,24)	Conv-BN-PReLU	(1 × 1 × 1)	103	(1 × 15 × 15,103)
(1 × 15 × 15,103)	Reshape	/	/	(103 × 15 × 15,1)
(103 × 15 × 15,1)	Conv-BN-PReLU	(103 × 1 × 1)	24	(1 × 15 × 15,24)
(1 × 15 × 15,24)	Spatial Pyramidal	/	/	(1 × 15 × 15,24)
(1 × 15 × 15,24)	Spatial Pyramidal	/	/	(1 × 15 × 15,24)
(1 × 15 × 15,24)	Spatial Pyramidal	/	/	(1 × 15 × 15,24)
(1 × 15 × 15,72)	Conv-BN-PReLU	(1 × 1 × 1)	60	(1 × 15 × 15,60)
(1 × 15 × 15,60)	Reshape	/	/	(60 × 15 × 15)
(60 × 15 × 15)	PSA: spatial-only	/	/	(60 × 15 × 15)
(60 × 15 × 15)	AvgPool-BN-Mish	/	/	(60 × 1 × 1)
(60 × 1 × 1)	Reshape	/	/	(60 × 1)
(60 × 1)	Linear Layer	(1 × 1)	9	(9 × 1)

and $(3 \times 1 \times 1)$ kernels to extract multiscale features. After that, the concatenation operator is conducted to converge the features. BN and PReLU are used to provide stability and nonlinearity for the network. Finally, convolutional layers with BN and PReLU are used to reduce the dimension of the feature maps and provide the output (\mathbf{FM}_{out}). The spatial-wise pyramidal convolutional block contains three convolutional layers with $(1 \times 7 \times 7)$, $(1 \times 5 \times 5)$, and $(1 \times 3 \times 3)$ kernels to extract multiscale spatial features. Similar to the channel-wise pyramidal convolutional block, the concatenation operator is conducted to generate the feature maps. After that, the convolutional layer, BN, and PReLU are used to provide the final output.

C. PSA Blocks: Channel-Only Block and Spatial-Only Block

PSA is a kind of self-attention mechanism, which designs for high-resolution pixel-wise regression. It can maintain high internal resolution in the computation of the channel and the spatial attention while fully collapsing the input tensors along the corresponding dimensions and composing nonlinearity to fit the output distribution of typical fine-grained regression. To be specific, two kinds of PSA blocks are introduced: channel-only block and spatial-only block. Given the input feature map \mathbf{FM}_i , the channel-wise attention weight $A^{ch}(\mathbf{FM}_i) \in \mathbb{R}^{C \times 1 \times 1}$ can be expressed as

$$A^{ch}(\mathbf{FM}_i) = F_{SG} \left[\mathbf{W}_z \left(\sigma_1(\mathbf{W}_v(\mathbf{FM}_i)) \times F_{SM}(\sigma_2(\mathbf{W}_q(\mathbf{FM}_i))) \right) \right] \quad (4)$$

where \mathbf{W}_q , \mathbf{W}_v , and \mathbf{W}_z are the 1×1 convolutional layers, σ_1 and σ_2 are the tensor reshape operators, $F_{SM}(\cdot)$ is a SoftMax operator, “ \times ” is the matrix dot-product operation, and $F_{SG}(\cdot)$ is a sigmoid operator. The output of the channel-only block is \mathbf{FM}_{out}^{ch} , and can be expressed as

$$\mathbf{FM}_{out}^{ch} = A^{ch}(\mathbf{FM}_i) \odot^{ch} \mathbf{FM}_i \quad (5)$$

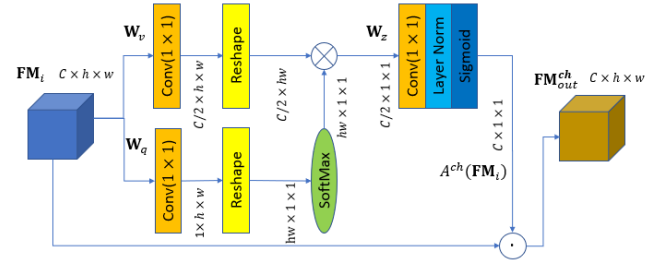


Fig. 8. Structure of the channel-only block of PSA.

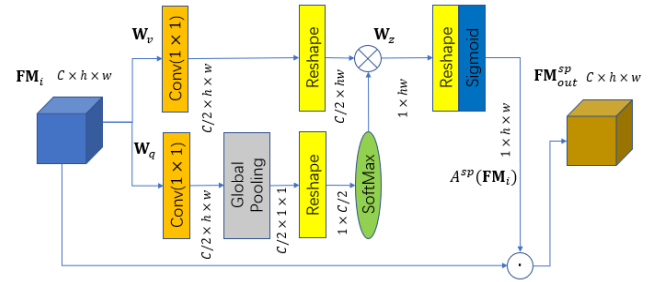


Fig. 9. Structure of the spatial-only block of PSA.

where \odot^{ch} is a channel-wise multiplication operator. The structure of the channel-only block of PSA is shown in Fig. 8.










The structure of the spatial-only block is shown in Fig. 9. The $A^{sp}(\mathbf{FM}_i) \in \mathbb{R}^{1 \times h \times w}$ can be expressed as

$$A^{sp}(\mathbf{FM}_i) = F_{SG} \left[\sigma_3 \left(F_{SM} \left(\sigma_1 \left(F_{GP}(\mathbf{W}_q(\mathbf{FM}_i)) \right) \right) \times \sigma_2(\mathbf{W}_v(\mathbf{FM}_i)) \right) \right] \quad (6)$$

where \mathbf{W}_q and \mathbf{W}_v are the standard 1×1 convolutional layers, σ_1 , σ_2 , and σ_3 are the tensor reshape operators, and F_{GP} is a global pooling operator. The output of the spatial-only block

TABLE II

CLASSES, COLORS, AND NUMBER OF SAMPLES OF THE UP DATASET

DATASET					
Class	Color	Total	Train	Validation	Test
C1		6631	66	66	6499
C2		18649	186	186	18277
C3		2099	20	20	2059
C4		3064	30	30	3004
C5		1345	13	13	1319
C6		5029	50	50	4929
C7		1330	13	13	1304
C8		3682	36	36	3610
C9		947	9	9	929
Total		42776	423	423	41930

is $\mathbf{FM}_{\text{out}}^{\text{SP}}$, which can be expressed as

$$\mathbf{FM}_{\text{out}}^{\text{SP}} = A^{\text{SP}}(\mathbf{FM}_i) \odot^{\text{SP}} \mathbf{FM}_i \quad (7)$$

where \odot^{SP} is a spatial-wise multiplication operator.

IV. EXPERIMENT

A. Hyperspectral Dataset Description

In the experiment, five well-known HSI datasets with different land covers and resolutions are used to evaluate the effectiveness of the proposed network, including the University of Pavia dataset (UP), the WHU-Hi-HongHu dataset (HH) [58], the Forest Farm of Gaofeng dataset (GF) [59], the GF-5 advanced HSI dataset (AH) [60], and the Houston University dataset (HU) [61]. The brief views of the five HSIs are described as follows.

1) *University of Pavia Dataset*: The UP dataset was collected by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the University of Pavia, Pavia, Italy, in 2003. The spatial size is 610×340 , and the spatial resolution is about 1.3 m per pixel. After dropping 12 noise-contaminated spectral bands, the UP dataset contains 103 bands with a spectral wavelength ranging from 430 to 860 nm. About 21% of pixels are labeled into nine categories, including asphalt, meadows, gravel, trees, metal sheets, bare soil, bitumen, bricks, and shadows. We randomly select 1% of labeled samples as training samples and validation samples, respectively. The remaining labeled samples are used as testing samples. The detailed classes, colors, and the number of samples of the UP dataset are shown in Table II.

2) *WHU-Hi-Honghu Dataset*: The HH dataset was acquired by the unmanned aerial vehicle (UAV) platform, which is over a complex agricultural area in Honghu City, Hubei Province, China. The spatial size is 940×475 . The spatial resolution is about 0.043 m per pixel. It contains 270 spectral bands ranging from 400 to 1000 nm. An intercepted area with 16 categories is introduced to our experiment, including Red roof, Road, Bare soil, Cotton, Rape, Chinese cabbage, Pakchoi, Cabbage, Tuber mustard, Brassica parachinensis, Small brassica chinensis, Lactuca sativa, Celtnce, Romaine lettuce, White radish, and Garlic sprout. The spatial size is 240×330 ranging in rows (701, 940) and columns (1, 330). We randomly select 1% of labeled samples as training samples and validation samples,

TABLE III

CLASSES, COLORS, AND NUMBER OF SAMPLES OF THE HH DATASET

























Class	Color	Total	Train	Validation	Test
C1		3320	33	33	3254
C2		1482	14	14	1454
C3		18725	187	187	18351
C4		1792	17	17	1758
C5		14939	149	149	14641
C6		5808	58	58	5692
C7		4054	40	40	3974
C8		2375	23	23	2329
C9		939	9	9	921
C10		2584	25	25	2532
C11		3979	39	39	3901
C12		4307	43	43	4221
C13		1002	10	10	982
C14		563	5	5	553
C15		973	9	9	955
C16		2037	20	20	1997
Total		68879	681	681	67517

TABLE IV

CLASSES, COLORS, AND NUMBER OF SAMPLES OF THE GF DATASET

Class	Color	Total	Train	Validation	Test
C1		1812	18	18	1776
C2		4411	44	44	4323
C3		4175	41	41	4093
C4		27476	274	274	26928
C5		3000	30	30	2940
C6		5801	58	58	5685
C7		6673	66	66	6541
C8		3162	31	31	3100
Total		56510	562	562	55386

respectively. The remaining labeled samples are used as testing samples. The detailed information is listed in Table III.

3) *Forest Farm of Gaofeng Dataset*: The GF dataset was acquired by the AISA Eagle II diffraction grating push-broom hyperspectral imager in 2018 over the Jiepai branch of Gaofeng State Owned Forest Farm, Nanning, Guangxi Province, China. The spatial size is 572×906 . The spatial resolution is about 1.0 m per pixel. The dataset covers the spectral range of 400–1000 nm with 125 bands. An intercepted area with eight categories is introduced to our experiment, including *Cunninghamia lanceolata*, *Pinus massoniana*, *Pinus elliottii*, *Eucalyptus urophylla*, *Mytilaria laosensis*, *Camellia oleifera*, Road, and Cutting bland. The spatial size is 400×400 , which ranges in rows of (1, 400) and columns of (1, 400). We randomly select 1% of labeled samples as training samples and validation samples, respectively. The remaining labeled samples are used as testing samples. The detailed information is displayed in Table IV.

4) *GF-5 Advanced HSI Dataset*: The AH dataset was obtained by the GF-5 satellite over the Jiangxia District, Wuhan City, Hubei Province, and covers an area of 109.4 km². It is a mixed landscape with mining and agriculture areas, and the types of surface objects are complex. The spatial size is 218×561 . The spatial resolution is about 30 m. Its spectral range extends from 400 to 2500 nm with 120 bands. The land covers are classified into six categories, including Surface-mined area, Road, Water, Crop land, Forest land, and Construction land. We randomly select 5% of labeled samples

TABLE V






















CLASSES, COLORS, AND NUMBER OF SAMPLES OF THE AH DATASET					
Class	Color	Total	Train	Validation	Test
C1		4838	241	241	4356
C2		486	24	24	438
C3		1026	51	51	924
C4		924	46	46	832
C5		1516	75	75	1366
C6		549	27	27	495
Total		9339	464	464	8411

TABLE VI

CLASSES, COLORS, AND NUMBER OF SAMPLES OF THE HU DATASET					
Class	Color	Total	Train	Validation	Test
C1		1251	12	12	1227
C2		1254	12	12	1230
C3		697	6	6	685
C4		1244	12	12	1220
C5		1242	12	12	1218
C6		325	3	3	319
C7		1268	12	12	1244
C8		1244	12	12	1220
C9		1252	12	12	1228
C10		1227	12	12	1203
C11		1235	12	12	1211
C12		1233	12	12	1209
C13		469	4	4	461
C14		428	4	4	420
C15		660	6	6	648
Total		15029	143	143	14743

as training samples and validation samples, respectively. The remaining labeled samples are used as testing samples. The classes, colors, and the number of samples for each class are exhaustively provided in Table V.

5) *Houston University Dataset*: The HU dataset was acquired over the University of Houston campus, Houston, TX, USA, and the neighboring urban area in 2012. The spatial size of the dataset is 349×1905 . The spatial resolution is 2.5 m per pixel. It has 144 spectral bands in the 380–1050-nm region. The land covers are classified into 15 categories, including Healthy grass, Stressed grass, Synthetic grass, Trees, Soil, Water, Residential, Commercial, Road, Highway, Railway, Parking Lot 1, Parking Lot 2, Tennis Court, and Running track. Due to the memory capacity limitation, we downscale the HU dataset to 30-D by PCA. We randomly select 1% of labeled samples as the training samples and validation samples. The remaining labeled samples are used as testing samples. The detailed information is listed in Table VI.

B. Experimental Setup and Assessment Indices

To evaluate the performance of the proposed network, five different types of HSIs, including four airborne datasets and one satellite dataset with different resolution and land cover types, are introduced to our experiment. Nine representative methods are selected for comparison, including SVM, HYbrid Spectral convolutional neural Network (HYSN) [62], SSRN [33], enhanced multiscale feature fusion network (EMFFN) [63], double-branch multi-attention mechanism network (DBMA) [54], double-branch dual-attention mechanism network (DBDA) [48], pyramidal convolution and iterative attention network (PCIA) [37], SSGC [43], and OSDN [1].

To be specific, the SVM with radial basis function (RBF) kernel is employed as a representative of the traditional method for HSI classification. The HYSN is employed as a representative of the traditional convolutional network. The SSRN is used to represent the traditional convolutional network with residual aggregation. The EMFFN is accepted to represent the multiscale convolutional network. The DBMA and DBDA represent the two-branch convolutional network with attention blocks. The PCIA is employed to represent the pyramidal multiscale convolutional network with attention blocks. The SSGC and OSDN are used to represent the state-of-the-art convolutional network. The competitors are described in detail as follows.

- 1) *SVM*: The SVM with RBF kernel is employed in the experiment. The raw spectral vectors of the pixels are fed into the SVM as the input data. The penalty parameter C and the RBF kernel width σ of SVM are selected by Grid SearchCV, both in the range of $(10^{-2}, 10^2)$.
- 2) *HYSN*: The HYSN is a spectral-spatial 3-D-CNN followed by spatial 2-D-CNN. Three multiscale 3-D convolutional layers with $7 \times 3 \times 3$, $5 \times 3 \times 3$, and $3 \times 3 \times 3$ kernels are used in the method to extract joint spectral-spatial features. One 2-D convolutional layer with a 3×3 kernel is used to learn more abstract level spatial features. Two fully connected layers are implemented after 3-D and 2-D layers to provide the final classification results.
- 3) *SSRN*: In the SSRN, the spectral and spatial residual blocks are introduced to learn discriminative features from spectral signatures and spatial contexts in HSI. Two kinds of 3-D kernels with $7 \times 1 \times 1$ and $1 \times 3 \times 3$ window sizes are used in the network to extract spectral information and spatial information, respectively. BN and rectified linear unit (ReLU) operators are added after each convolutional layer.
- 4) *EMFFN*: The EMFFN is an enhanced multiscale feature fusion network, which consists of two networks named spectral cascaded dilated convolutional network (CDCN) and parallel multipath network (PMN). The features collected from the two subnetworks are combined into EMFFN using the designed consolidated loss function. In the CDCN, four dilated 2-D convolutional layers with kernel size 6×1 are used to extract the spectral information. The dilation rate $d = 2^i$ ($i = 0, 1, 2, 3$) is designed for the blocks. A channel attention module is implemented after the dilated convolutional layers to further extract the long-range information. In the PMN, the input data are downsampled to 5-D by PCA. Multiscale 2-D convolutional layers with 7×7 , 5×5 , and 3×3 kernels are introduced to extract multilevel spatial information. Three parallel paths are used to fuse the multiscale features to leverage both shallow and deep features.
- 5) *DBMA*: The DBMA is a double-branch multiattention mechanism network for HSI classification. Two branches networks are used to extract spectral and spatial features, respectively. Two types of attention mechanisms are applied in the two branches. The sizes of the 3-D kernels

TABLE VII
CLASSIFICATION OA (%), AA (%), AND KAPPA WITH SD AND THE TRAINING TIME (S) OF THE UP DATASET

Class	SVM	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
C1	89.86±2.81	82.37±1.43	98.45±1.92	83.27±2.47	95.46±1.00	90.06±3.81	89.38±5.22	87.76±2.40	96.38±1.95	97.27±7.83
C2	96.90±0.39	94.32±0.19	93.07±1.55	89.34±4.04	98.69±0.09	98.99±0.18	99.41±0.31	98.55±0.98	99.55±0.37	99.09±1.41
C3	54.31±11.1	62.03±1.38	52.69±7.19	68.53±10.52	82.96±6.06	93.65±3.98	83.78±16.48	96.33±3.72	95.53±4.35	99.60±0.61
C4	85.32±5.44	99.31±0.32	99.77±0.15	87.20±2.86	95.66±0.53	96.45±0.44	97.67±0.95	97.66±0.14	96.93±0.86	99.82±0.10
C5	99.05±0.24	98.43±1.34	86.60±25.45	98.15±1.61	99.06±0.19	99.26±0.13	99.49±0.06	99.62±0.11	97.02±1.77	98.72±0.75
C6	75.18±2.23	83.26±1.24	97.83±0.67	89.68±1.80	99.48±0.25	96.95±0.72	99.31±0.46	99.99±0.02	99.79±0.06	98.28±2.06
C7	65.70±12.47	88.39±1.95	34.07±20.44	73.45±7.97	98.16±0.91	99.91±0.19	100.00±0.00	100.00±0.00	100.00±0.00	90.02±6.69
C8	88.77±5.10	75.38±1.25	90.55±6.06	75.58±4.81	84.75±1.94	84.49±5.65	87.80±3.06	93.08±5.98	90.37±3.55	92.82±9.84
C9	99.72±0.05	97.18±0.87	98.60±2.53	99.74±0.20	97.83±0.37	95.85±0.80	98.79±0.29	97.37±0.65	98.79±0.69	98.59±1.14
OA	88.81±0.90	88.21±0.04	84.03±4.00	86.14±2.48	96.01±0.31	95.33±1.19	95.43±2.44	96.12±0.90	97.73±0.58	97.88±3.20
AA	83.88±2.44	86.74±0.22	83.52±5.20	84.99±1.94	94.67±0.65	95.07±0.94	95.07±2.49	96.71±0.52	97.15±0.39	97.14±1.54
Kappa	0.8498	0.8428	0.7879	0.8125	0.9470	0.9378	0.9393	0.9482	0.9698	0.9719
	±0.0123	±0.0006	±0.0519	±0.0365	±0.0041	±0.016	±0.0324	±0.0122	±0.0078	± 0.0431
Time	-	8.54	18.40	50.06	20.77	22.24	16.60	8.66	7.52	75.40

are $7 \times 1 \times 1$ and $1 \times 3 \times 3$. Dense aggregation is introduced to mention the multilevel features.

- 6) *DBDA*: The DBDA is a double-branch multiattention mechanism network, which is the same as DBMA. The channel attention block and spatial attention block are different from DBMA. Moreover, Mish is adopted as the activation function.
- 7) *PCIA*: The PCIA is a double-branch network, which is the same as the DBMA and DBDA. The PyConv is applied to extract multiscale spectral and spatial information in the two branches. The sizes of the 3-D convolutional layers are $7 \times 1 \times 1$, $5 \times 1 \times 1$, $3 \times 1 \times 1$, $1 \times 7 \times 7$, $1 \times 5 \times 5$, and $1 \times 3 \times 3$. An iterative attention mechanism is introduced in the PCIA.
- 8) *SSGC*: The channel and position global context attention blocks are proposed in the SSGC. The rest of the network architecture is the same as the DBMA and DBDA.
- 9) *OSDN*: The one-shot aggregation and polarized attention blocks are introduced in the OSDN. The rest of the network architecture is the same as the DBMA and DBDA.

For all the competitive networks, the spatial size of the HSI patch cube is set to 11×11 . The batch size is set to 32. The epoch is set to 200, and the initial learning rate is set to 0.0005. The Adam optimizer is adopted with an attenuation rate of (0.9, 0.999) and a fuzzy factor of 10^{-8} . The learning rate is dynamically adjusted every 15 epochs by cosine annealing [64]. Moreover, the early stopping technique is employed in the training process. If the loss on the validation dataset does not change within 20 epochs, the training process will move to the test session. Furthermore, the dropout technique with 0.5 probability is applied to enhance the generalization capability of the model.

To quantitatively measure the performance of the competitors, the overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) are implemented in the experiments. All experiments are repeated five times independently. The average values of the experimental results are reported as the final results. The experimental hardware environment is a deep learning workstation with an Intel Xeon E5-2680v4

processor 2.4 GHz and NVIDIA GeForce RTX 2080Ti GPU. The software environment is CUDA v11.2, PyTorch 1.10, and Python 3.8.

C. Experimental Results

We first assess the performance and training time of the various methods on the UP dataset. The classification results are given in Table VII. The best OA, AA, Kappa, and the largest training time are highlighted in bold. We can see that the proposed PMCN achieves competitive classification results in each category, OA, AA, and Kappa in most cases. Comparing the OAs of the competitors, PMCN achieves 9.07%, 9.67%, 13.85%, 11.74%, 1.87%, 2.55%, 2.45%, 1.76%, and 0.15% of the OA more than that of SVM, HYSN, SSRN, EMFFN, DBMA, DBDA, PCIA, SSGC, and OSDN, respectively. It is because we use the pyramidal multiscale convolutional blocks and PSA blocks to jointly extract spectral and spatial information. Furthermore, we use residual aggregation and one-shot aggregation to maintain the multilevel features of the network, which allows the network to be designed deeper. The OA of SVM is lower than that of deep convolutional networks in most cases, except HYSN, SSRN, and EMFFN. It is because convolutional networks implicitly use the spatial information of the pixels and can be considered the spatial-spectral-based classification method. By obtaining more available information on pixels, deep convolutional networks can achieve better classification results than SVM. Comparing the deep convolutional networks, we can see that the HYSN, SSRN, and EMFFN provide lower OAs than the later networks. It indicates that effective extraction of discriminative spectral and spatial features of the UP dataset is difficult for traditional 3-D and 2-D CNNs. The two-branch networks (DBMA and DBDA) outperform the traditional deep convolutional networks (HYSN, SSRN, and EMFFN). The pyramidal multiscale network (PCIA) provides an OA of 95.43%, which is better than DBDA and less than DBMA. Moreover, the networks using more techniques (SSGC, OSDN, and PMCN), such as two-branch structure, multiscale convolution, attention mechanism, dense aggregation, and one-shot aggregation, achieve better results than those of the former networks. The SSRN and PMCN provide a relatively high standard deviation (SD)

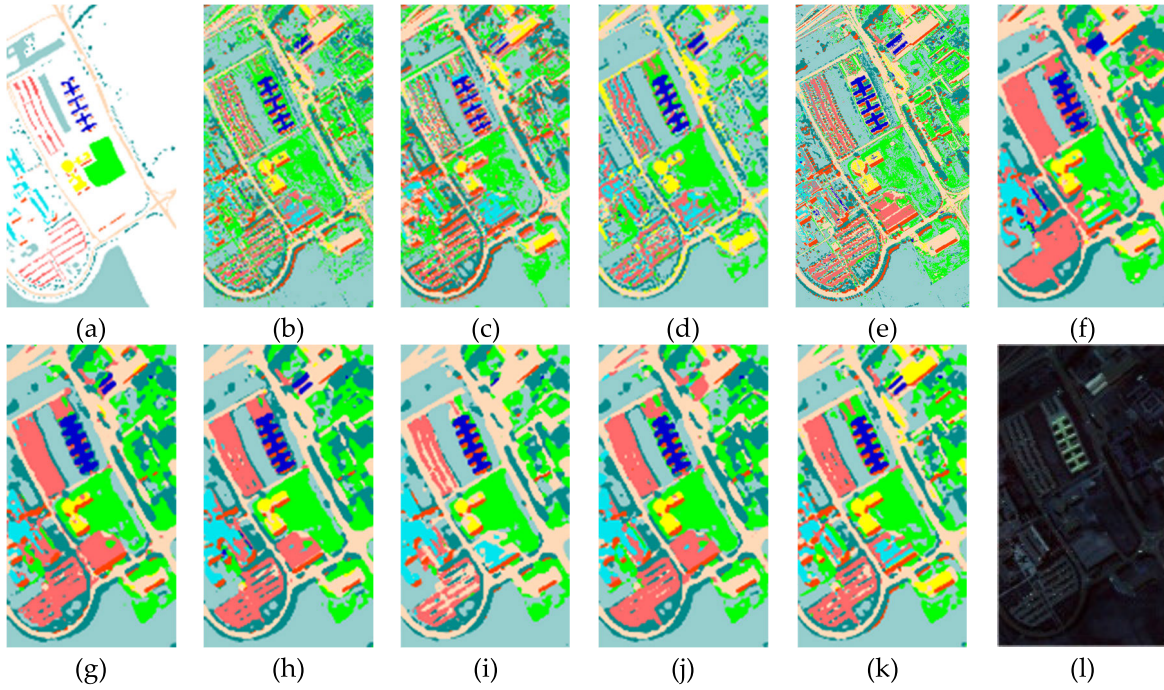


Fig. 10. Full-factor classification maps for the UP dataset: (a) ground truth; (b) SVM; (c) HYSN; (d) SSRN; (e) EMFFN; (f) DBMA; (g) DBDA; (h) PCIA; (i) SSGC; (j) OSDN; (k) PMCN; and (l) false-color image.

TABLE VIII
CLASSIFICATION OA (%), AA (%), AND KAPPA WITH SD AND THE TRAINING TIME (S) OF THE HH DATASET

Class	SVM	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
C1	86.19±2.99	88.54±2.37	91.77±33.86	81.11±3.77	97.81±0.77	97.42±1.99	98.09±0.47	97.01±2.97	97.84±1.05	99.39±1.00
C2	69.74±5.59	74.79±3.08	94.25±35.48	72.35±2.35	77.70±4.23	72.21±1.27	88.86±1.11	84.12±5.11	91.44±5.12	87.82±2.53
C3	93.21±1.08	92.34±0.85	86.41±8.23	84.09±0.72	97.39±1.04	96.27±0.71	96.18±0.61	97.69±3.45	96.58±0.66	98.53±3.13
C4	83.64±7.96	86.24±2.81	68.92±28.57	73.76±2.03	95.35±2.08	99.03±0.59	95.69±2.21	98.33±5.71	95.82±1.31	97.61±1.16
C5	95.92±0.78	95.31±1.40	89.17±7.36	86.86±1.54	99.09±0.08	98.23±1.43	98.24±0.83	99.83±1.91	96.05±0.73	99.42±2.86
C6	84.66±1.44	86.55±0.93	66.35±16.29	69.23±0.61	97.49±0.84	96.74±0.56	93.08±4.18	96.65±1.30	97.64±4.13	91.00±4.79
C7	47.45±3.38	72.51±5.99	94.22±39.12	51.79±4.11	91.46±4.12	94.88±1.80	83.94±2.81	85.95±4.35	88.22±1.35	78.58±4.49
C8	95.81±1.33	96.32±0.72	97.66±39.18	95.48±1.18	99.99±0.02	98.31±0.54	99.93±0.08	100.00±0.00	97.98±0.70	99.23±0.38
C9	26.04±8.79	57.95±6.08	0.00±0.00	29.95±6.42	93.33±1.25	97.42±1.95	92.07±1.62	86.04±6.85	98.48±0.65	98.65±4.03
C10	49.83±5.98	77.76±0.61	84.14±33.33	60.58±1.04	89.84±3.31	83.96±2.71	88.05±2.77	96.03±5.67	73.56±3.17	98.26±5.10
C11	62.23±3.06	87.86±1.49	89.64±37.32	61.12±2.43	95.44±2.03	92.15±2.92	91.62±2.48	96.90±2.93	98.51±6.03	96.30±3.21
C12	60.10±2.78	86.25±2.07	99.32±0.27	81.45±2.53	97.43±0.49	98.60±0.86	98.43±1.55	99.32±7.78	94.83±2.09	99.95±4.35
C13	67.02±8.19	81.96±2.48	50.00±22.61	83.56±3.43	89.77±2.36	95.98±1.44	98.01±0.92	88.82±5.46	98.33±0.45	98.22±5.08
C14	70.22±5.98	49.26±6.60	87.67±35.07	66.34±5.59	82.63±3.28	90.12±4.89	82.82±1.82	98.55±5.08	93.16±2.61	78.02±2.21
C15	29.24±3.31	61.63±4.75	99.43±48.65	32.58±5.58	72.15±3.87	65.81±3.31	93.36±4.49	84.75±3.29	95.03±3.61	98.63±0.45
C16	55.95±7.79	83.83±2.11	97.83±43.12	55.76±3.33	91.46±4.88	92.77±4.45	96.49±0.86	71.43±1.02	94.61±1.80	92.41±0.30
OA	80.40±0.24	88.01±0.42	85.70±8.67	78.02±0.41	95.64±0.91	94.90±0.56	95.06±1.19	95.58±2.28	94.83±0.73	96.05±1.17
AA	67.33±0.68	79.94±0.87	81.05±21.71	67.88±1.60	91.77±1.02	91.87±0.40	93.43±1.17	92.59±4.88	94.25±0.91	94.50±1.10
Kappa	0.7684	0.8591	0.8293	0.7388	0.9490	0.9403	0.9420	0.9483	0.9393	0.9537
	±0.0026	±0.0049	±0.1091	±0.0046	±0.0106	±0.0066	±0.0141	±0.0271	±0.0086	±0.0161
Time	-	35.66	257.69	60.79	89.20	206.56	122.37	89.34	69.49	112.81

of OAs than other methods, which shows that the robustness of the SSRN and PMCN is not strong. PMCN requires the most training time (75.40 s) to train the network, which is discussed in Section V-C. The full-factor classification maps of the competitors are shown in Fig. 10. We can see that the salt-pepper noise appears in the classification map of SVM. In contrast, the classification maps of the convolutional networks are smooth. It shows that convolutional networks can improve the smoothness of the classification maps by extracting spatial features of HSI datasets.

To further evaluate the performance of the proposed method, experiments are implemented on a high spatial resolution HSI dataset, which is the HH dataset (0.043 m per pixel). From Table VIII, we can see that the spectral-based classification

method (SVM) achieves the lowest OA (80.4%) except for the EMFFN. It indicates that it is difficult to classify the land cover objects using only spectral signatures on the HH dataset. EMFFN obtains the lowest OA of 78.02%. HYSN and SSRN achieve higher OAs (88.01%, 85.70%) than SVM and EMFFN. Observing the classification accuracy of various categories, we can see that some categories are still hard to be classified for SVM, HYSN, SSRN, and EMFFN such as C2, C4, C6, C9, C10, C11, C13, C14, C15, and C16. Especially, the C9 failed to be classified by SVM (26.04%), SSRN (0.00%), and EMFFN (29.95%). In contrast, DBMA and DBDA obtain better classification accuracies (95.64%, 94.90%) than those of the former methods. The PCIA consistently achieves competitive results (95.06%), which indicates

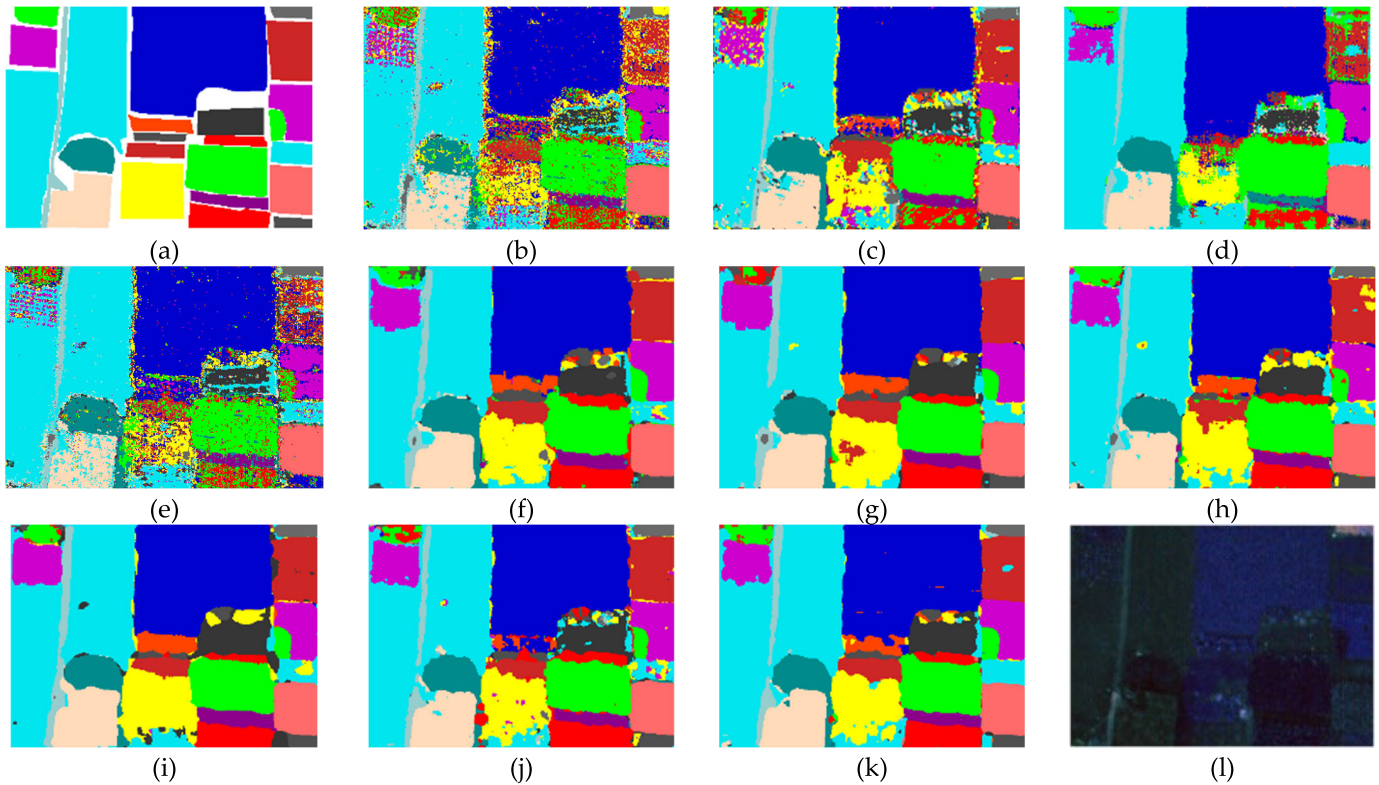


Fig. 11. Full-factor classification maps for the HH dataset: (a) ground truth; (b) SVM; (c) HYSN; (d) SSRN; (e) EMFFN; (f) DBMA; (g) DBDA; (h) PCIA; (i) SSGC; (j) OSDN; (k) PMCN; and (l) false-color image.

TABLE IX
CLASSIFICATION OA (%), AA (%), AND KAPPA WITH SD AND THE TRAINING TIME (S) OF THE GF DATASET

Class	SVM	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
C1	14.68±6.03	75.17±18.53	54.71±45.54	26.06±1.28	99.97±0.05	96.71±1.35	99.96±0.07	99.17±1.54	99.94±0.11	98.12±0.81
C2	47.34±6.43	80.33±7.09	72.39±28.15	51.19±4.78	96.03±0.49	97.76±0.70	99.40±0.64	97.98±1.11	99.76±0.15	99.53±0.51
C3	36.59±5.92	73.08±6.35	66.40±8.77	42.29±1.17	99.89±0.07	96.77±4.91	99.55±0.65	100.00±0.00	99.82±0.21	99.12±0.42
C4	93.44±0.50	96.72±0.73	90.81±1.80	82.92±0.95	99.90±0.06	99.85±0.04	99.60±0.47	99.74±0.12	99.90±0.07	99.92±0.97
C5	23.05±1.45	88.71±3.95	78.31±39.25	52.71±3.35	98.88±0.34	96.67±0.108	97.44±1.22	99.99±0.02	94.20±2.85	99.93±0.84
C6	60.48±2.67	93.57±4.01	82.45±14.57	74.18±3.46	98.71±0.07	99.98±0.02	99.97±0.05	99.94±0.06	99.98±0.02	99.89±0.05
C7	98.64±0.29	99.74±0.16	86.69±6.73	99.96±0.04	99.91±0.03	99.96±0.02	100.00±0.00	99.88±0.11	99.82±0.07	99.35±0.42
C8	87.26±10.30	99.09±0.52	94.64±5.03	82.25±1.84	99.65±0.05	97.80±0.81	98.92±0.59	99.32±0.44	99.84±0.09	99.54±0.21
OA	76.26±0.50	92.41±1.50	86.18±3.73	76.91±0.42	99.39±0.05	99.06±0.44	99.52±0.25	99.62±0.08	99.56±0.14	99.70±0.23
AA	57.68±0.72	88.30±3.08	78.30±15.37	63.94±0.71	99.12±0.05	98.19±0.65	99.35±0.25	99.50±0.21	99.16±0.35	99.48±0.25
Kappa	0.6564	0.8943	0.8018	0.6685	0.9917	0.9849	0.9933	0.9929	0.9939	0.9959
	±0.0067	±0.0206	±0.0539	±0.0070	±0.0007	±0.0061	±0.0035	±0.0012	±0.0020	±0.0032
Time	-	9.38	35.29	69.02	21.99	21.02	55.30	91.05	57.54	62.89

that the pyramidal convolutional network can provide high discriminatory ability for HSI classification tasks. SSGC and OSDN obtain OAs of 95.58% and 94.83%. PMCN achieves the highest OA (96.05%), AA (94.50%), and Kappa (0.9537) among all the competitors. The SSRN provides the highest SD among the classification frameworks. From Fig. 11, we can see that the C9 is classified to be C5 by SVM, SSRN, and EMFFN. There are some salt-pepper noises in classification maps of SVM, HYSN, SSRN, and EMFFN. DBMA, DBDA, SSGC, and OSDN provide better classification maps. However, there are still some ambiguities and misclassifications in C2, C3, and C7. PMCN obtained more clear and smooth classification maps in most categories.

The GF dataset is a forest farm that is applied to forestry tree species classification. The spectral responses of different plants of the same family and genus are very close to each other, and the classification results of most existing spectral-based

methods tend to be reduced. As shown in Table IX, the OA of SVM is 76.26%. For some specific classes, such as C1, C2, C3, and C5, the accuracy is less than 50%. HYSN, SSRN, and EMFFN provide better classification accuracies than SVM. However, the accuracies of C1 (75.17%, 54.71%, 26.06%) and C3 (73.08%, 66.40%, 42.29%) are still insufficient. Conversely, DBMA, DBDA, PCIA, SSGC, OSDN, and PMCN provide satisfactory classification accuracies, especially for C1, C3, and C5. PMCN achieves competitive results in most cases. The full-factor classification maps are shown in Fig. 12, and the classification map by the PMCN is almost the same as the ground truth.

Furthermore, the AH dataset is applied to evaluate the performance of the methods. It is a satellite dataset with mining and agriculture areas. In particular, the labeled samples of the AH dataset are disjointly marked. It is a challenge to effectively extract the spatial feature of a pixel. As shown

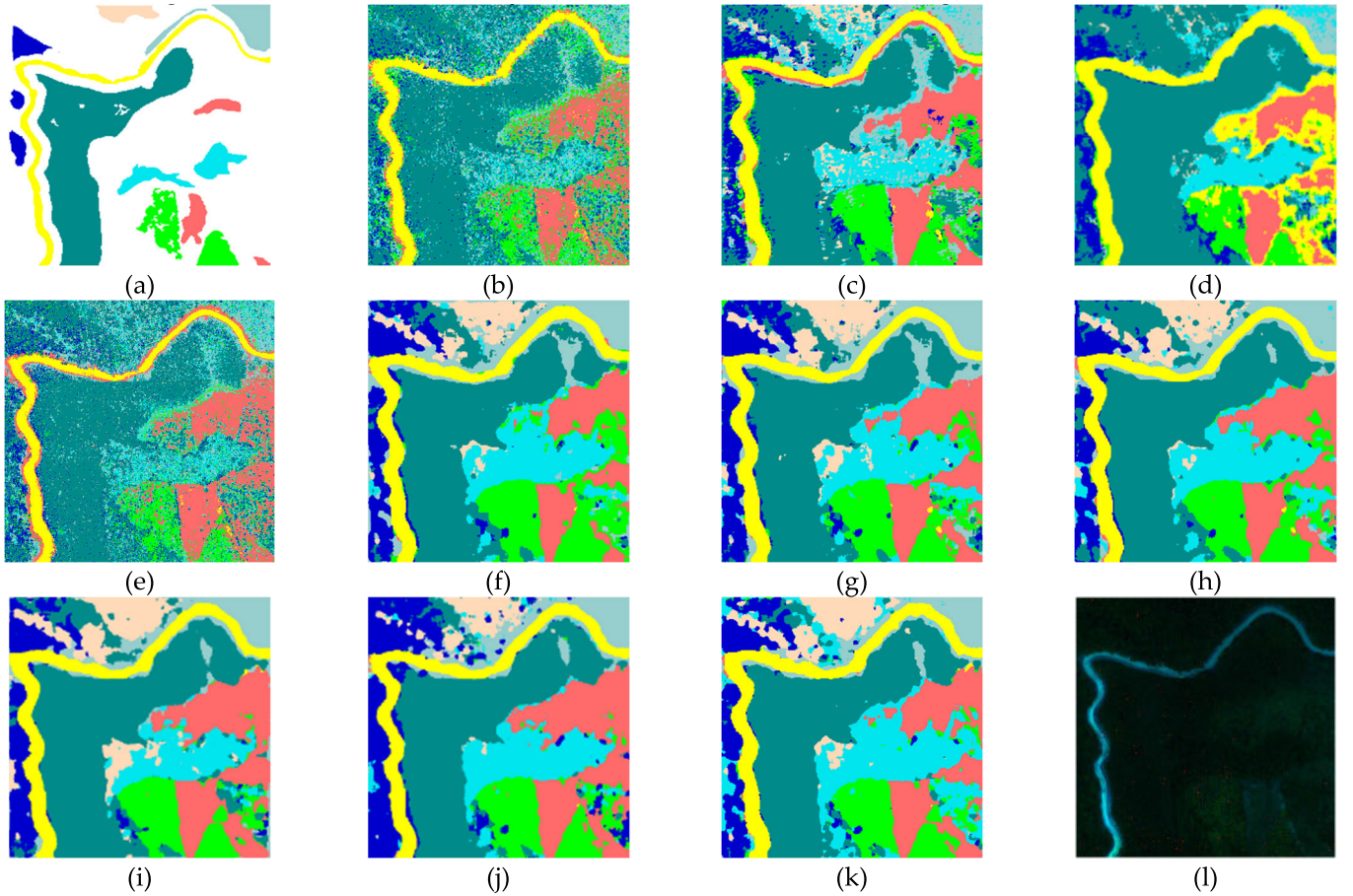


Fig. 12. Full-factor classification maps for the GF dataset: (a) ground truth; (b) SVM; (c) HYSN; (d) SSRN; (e) EMFFN; (f) DBMA; (g) DBDA; (h) PCIA; (i) SSGC; (j) OSDN; (k) PMCN; and (l) false-color image.

TABLE X
CLASSIFICATION OA (%), AA (%), AND KAPPA WITH SD AND THE TRAINING TIME (S) OF THE AH DATASET

Class	SVM	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
C1	90.73±1.78	88.52±4.99	88.44±2.89	84.51±0.35	92.20±1.42	85.24±4.99	92.37±2.47	89.96±6.53	85.67±2.41	92.97±2.93
C2	47.03±6.62	38.53±18.28	38.00±15.05	66.96±1.97	58.54±1.91	47.26±6.04	62.79±2.04	57.57±4.71	64.14±7.69	67.45±2.22
C3	62.21±6.99	68.93±10.49	53.42±10.58	79.42±0.74	80.11±5.93	85.17±1.44	89.73±2.21	77.78±3.77	81.21±2.86	75.71±9.34
C4	38.29±8.02	44.84±16.95	44.02±11.70	51.43±0.70	56.17±3.92	51.87±1.30	44.89±2.46	44.58±5.44	56.76±6.74	70.79±6.99
C5	68.01±3.69	65.68±10.45	82.93±5.50	67.24±0.88	60.09±4.37	69.12±1.96	65.15±3.99	86.15±8.78	75.73±7.44	60.32±7.72
C6	35.63±5.48	50.74±19.04	89.41±37.73	67.75±3.35	79.33±4.87	72.89±8.24	78.78±1.32	74.46±7.99	80.52±2.37	78.59±5.13
OA	73.15±1.32	74.03±4.30	75.97±5.73	76.82±0.36	78.86±0.86	76.92±2.57	78.13±0.13	79.15±2.95	79.72±1.53	80.73±2.58
AA	56.98±1.94	59.54±12.23	66.04±11.23	69.55±0.98	71.07±0.93	68.59±0.70	72.28±0.85	71.75±2.19	74.01±1.98	74.36±2.46
Kappa	0.5870	0.6131	0.6334	0.6456	0.6837	0.6406	0.6750	0.6844	0.6848	0.7140
	±0.0169	±0.0874	±0.0944	±0.0050	±0.0153	±0.0516	±0.0031	±0.0607	±0.0204	±0.0436
Time	-	9.45	34.52	56.53	33.50	26.16	46.32	29.21	15.32	64.38

in Table X, the spatial-spectral-based deep convolutional networks (HYSN, SSRN, EMFFN, DBMA, DBDA, PCIA, SSGC, OSDN, and PMCN) achieve limited improvement than the spectral-based method (SVM), which ranges from 0.88% to 7.58%. The reason is that the disjointly marked samples restrict the ability of the cube-based approach to extract spatial information. Under the condition of restricted spatial information, the discrimination capability of convolutional networks cannot be sufficiently exploited. Benefiting from the multiscale property of PyConv, PMCN obtains the highest classification accuracy (80.73%) among the competitors. The full-factor classification maps for the AH dataset are shown in Fig. 13. We can see that PMCN yields a finer-grained classification map than that of DBMA, DBDA, PCIA, SSGC, and OSDN.

This may be due to the ability of polarized attention blocks to extract detailed spatial and spectral features of pixels.

Finally, the HU dataset is employed to evaluate the performance of the methods under limited labeled sample condition. In the experiment, the number of training samples of different categories ranges from 3 to 12. It is difficult to learn discriminative information effectively from such a small number of training samples. As shown in Table XI, SSRN achieves the lowest OA, which indicates that the CNN cannot effectively extract useful features by simply stacking the 3-D and 2-D convolutional layers under limited labeled sample condition. The SVM, HYSN, and EMFFN obtain similar classification accuracies (79.59%, 79.13%, 76.37%). In contrast, DBMA, DBDA, PCIA, SSGC, OSDN, and PMCN

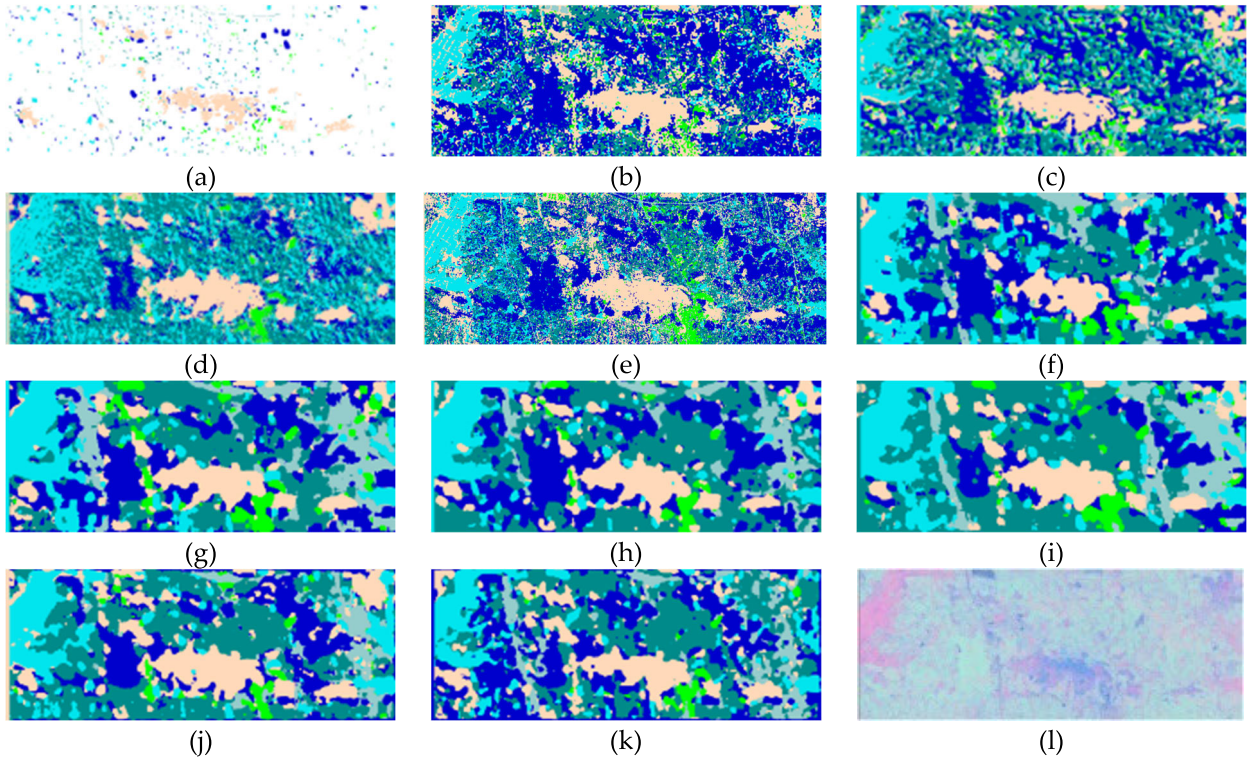


Fig. 13. Full-factor classification maps for the AH dataset: (a) ground truth; (b) SVM; (c) HYSN; (d) SSRN; (e) EMFFN; (f) DBMA; (g) DBDA; (h) PCIA; (i) SSGC; (j) OSDN; (k) PMCN; and (l) false-color image.

TABLE XI
CLASSIFICATION OA (%), AA (%), AND KAPPA WITH SD AND THE TRAINING TIME (S) OF THE HU DATASET

Class	SVM	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
C1	92.67±5.06	82.06±2.71	68.37±18.32	85.97±3.46	89.57±0.59	86.73±0.82	80.15±0.38	87.60±1.36	86.54±1.02	88.11±1.80
C2	90.99±4.43	79.75±6.43	93.18±22.63	92.52±0.39	87.35±1.25	93.37±3.77	93.95±0.34	84.16±6.59	87.34±1.58	89.46±0.84
C3	98.99±0.69	98.32±1.10	56.72±18.60	99.85±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	99.77±0.33	97.74±0.34
C4	93.24±3.40	77.90±4.33	68.80±12.22	93.06±0.95	81.08±1.05	80.91±2.53	78.64±3.00	84.61±1.85	91.93±0.91	97.07±0.91
C5	97.45±1.81	91.35±2.47	92.03±24.39	89.64±1.48	99.95±0.04	94.92±0.64	96.08±0.79	93.98±2.30	92.12±3.40	92.32±0.34
C6	83.73±5.62	99.42±0.84	87.27±11.17	98.58±0.41	92.14±0.06	100.00±0.00	100.00±0.00	100.00±0.00	99.93±0.14	96.36±0.44
C7	72.95±5.54	85.47±4.09	58.65±31.54	82.19±2.30	84.98±1.71	77.69±5.37	69.93±0.88	95.93±1.45	68.85±10.32	78.32±6.51
C8	56.82±7.89	87.18±4.05	62.56±22.42	82.36±0.99	99.47±0.06	82.52±3.21	89.35±0.73	91.29±7.68	99.60±0.81	96.92±4.13
C9	79.21±7.80	74.58±7.37	61.76±14.87	55.76±2.46	77.92±2.89	75.61±5.11	80.44±8.82	73.43±7.04	84.65±1.39	81.93±1.10
C10	76.02±13.56	59.08±2.52	51.39±32.69	52.68±2.44	73.90±1.35	88.89±6.32	70.70±9.45	73.32±6.87	87.21±1.79	89.06±1.46
C11	69.67±12.50	82.11±5.39	98.93±37.96	70.05±5.20	75.93±4.23	83.62±3.86	89.75±9.80	81.13±9.29	89.38±4.99	81.74±2.65
C12	62.53±8.90	59.89±6.06	47.89±18.75	45.37±1.30	79.23±1.47	71.93±7.80	71.61±4.67	87.31±12.99	66.26±5.68	85.04±1.90
C13	22.28±4.44	96.60±1.39	95.40±45.02	61.52±1.86	93.31±1.72	84.00±1.80	93.86±2.86	84.55±7.13	95.89±3.46	75.97±0.71
C14	89.56±12.59	96.00±3.81	94.74±12.79	92.79±0.88	100.00±0.00	100.00±0.00	97.00±3.67	92.92±3.07	90.49±2.16	90.41±1.99
C15	99.11±0.41	89.11±2.95	85.04±37.64	99.63±0.27	90.61±0.35	94.32±0.22	96.26±0.23	78.07±1.36	94.38±1.20	95.07±2.02
OA	79.59±1.32	79.13±1.59	66.32±12.48	76.37±0.37	85.59±0.81	85.42±2.92	83.33±2.93	85.19±4.43	85.68±2.89	87.98±0.52
AA	79.02±1.69	83.86±1.42	74.85±16.72	80.14±0.35	88.36±0.44	87.63±2.36	87.18±1.89	87.22±2.03	88.96±1.43	89.03±0.27
Kappa	0.7791	0.7744	0.6351	0.7443	0.8441	0.8423	0.8196	0.8399	0.8451	0.8701
	±0.0144	±0.017	±0.1358	±0.0040	±0.0088	±0.0317	±0.0318	±0.0480	±0.0314	±0.0056
Time	-	2.11	3.71	30.79	12.29	9.91	6.63	9.12	6.47	14.79

improve the OAs significantly ranging from 3.74% to 8.39%. From Fig. 14, we can see that the SVM, HYSN, SSRN, and EMFFN achieve finer-grained classification maps than that of the later methods. It shows that these methods prefer to adopt spectral signatures to classify the pixels. In contrast, the DBMA, DBDA, PCIA, SSGC, OSDN, and PMCN make more use of spatial contextual information to extract discriminative features and obtain spatially smoother classification maps.

V. DISCUSSION

A. Comparison of Different Spatial Patch Sizes

In this section, we will focus on the issue of patch size, which is a hyperparameter of the cube-based convolutional

TABLE XII
IMPACT OF PATCH SIZE ON THE OA OF THE PMCN IN THE FIVE HSI DATASETS

Patch size	UP (%)	HH (%)	GF (%)	AH (%)	HU (%)
7 × 7	96.34	94.39	98.05	82.67	85.84
9 × 9	95.70	92.94	97.98	81.62	86.90
11 × 11	97.88	96.05	99.70	80.73	87.98
13 × 13	97.30	96.00	98.58	77.78	85.77
15 × 15	97.11	95.90	98.45	77.53	82.32

network. In general, an appropriate patch size can help the network to extract effective spatial information. Small or large patch sizes may affect the discriminative ability of the network

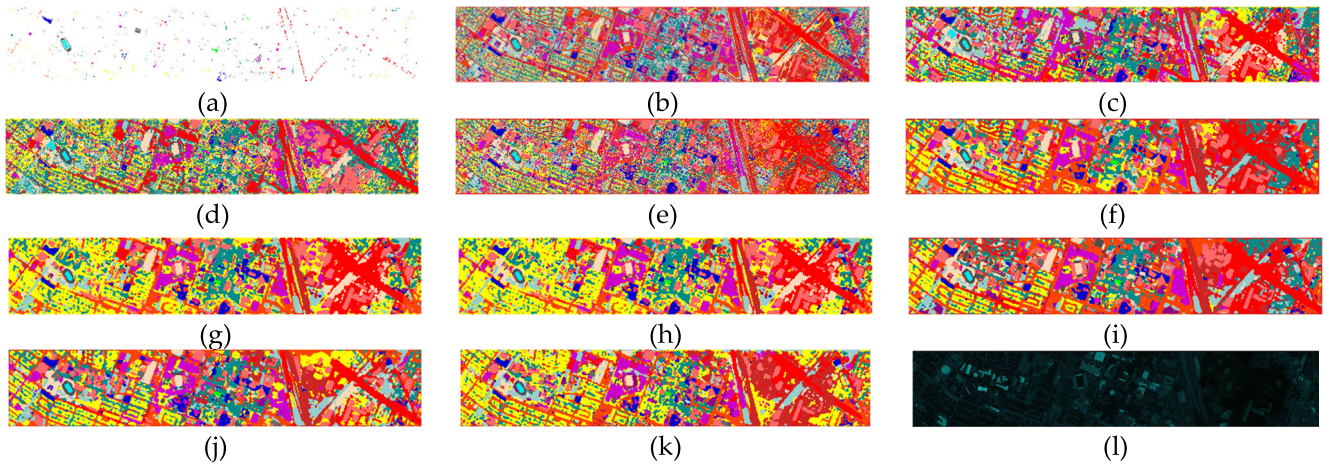


Fig. 14. Full-factor classification maps for the Houston dataset: (a) ground truth; (b) SVM; (c) HYSN; (d) SSRN; (e) EMFFN; (f) DBMA; (g) DBDA; (h) PCIA; (i) SSGC; (j) OSDN; (k) PMCN; and (l) false-color image.

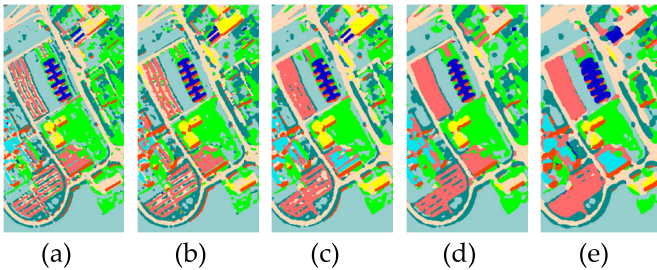


Fig. 15. Classification maps with different patch sizes of the UP dataset: (a) 7×7 ; (b) 9×9 ; (c) 11×11 ; (d) 13×13 ; and (e) 15×15 .

by providing insufficient or excessive spatial information. As shown in Table XII, we report the OAs with different spatial patch sizes ranging from 7×7 to 15×15 with a 2-pixel interval. We can see that the classification accuracies vary with the patch sizes. The best OA is acquired when the patch size is 11×11 in the UP, HH, GF, and HU datasets, which is as expected. The classification maps of UP with different patch sizes are shown in Fig. 15 as an example. However, the best OA is acquired when the patch size is 7×7 in the AH dataset. It is understandable that the labeled samples in the AH dataset are disjointly marked, which is different from the other datasets. The spatial neighborhood area of pixels is restricted in the AH dataset. As a result, a larger patch size cannot effectively provide more spatial information for the network, but rather affect the discriminability of the pixels. In practice, we recommend using smaller patch sizes for datasets that provide disjoint labeled samples. In our experiment, we consistently choose 11×11 as the value of the patch size for the four datasets to keep consistency.

B. Comparison of Different Training Sample Proportions

In this section, we will discuss the performance of the competitors under different proportional training sample conditions in the five HSI datasets. It is an important analysis that the supervised learning methods are data-driven-based and the percentage (number) of the training samples plays a leading role in the learning process of the models. In order

to comprehensively analyze the performance of the proposed PMCN under different proportional training sample conditions, we randomly select 0.5%, 1%, 1.5%, 2%, 3%, 4%, and 5% of labeled samples for UP, HH, GF, and HU datasets, and 5%, 6%, 7%, 8%, 9%, and 10% of labeled samples for AH dataset as the training samples. In general, a larger proportion of training samples can provide more discriminative information for the data-driven-based classification methods, thus improving the classification accuracy of the models. The classification results are reported in Fig. 16. It can be seen clearly that the classification accuracies of the methods increase with the growth of the training sample proportion as expected. With a smaller percentage of training samples (0.5% for the UP, HH, GF, and HU datasets and 5% for the AH dataset), the classification accuracies of the competitors are subsequently reduced. Comparing the classification methods, the classification accuracies of SSGC, OSDN, and PMCN decrease less than those of other methods. It indicates that these classification methods are more capable of extracting discriminable features with limited labeled samples. PMCN achieves consistently competitive results with the increase of the training sample proportion. Specifically, we can see in Fig. 16(d) and (e) that PMCN obtains higher classification accuracies than other methods for the AH and HU datasets. It indicates that PMCN has the best ability to effectively extract discriminable features under the condition of limited spatial context information and limited training samples. The experimental results demonstrate again the utility and effectiveness of the combination of pyramidal multiscale convolutional block and polarized attention block for HSI classification tasks and provide thoughts for researchers to design high-performance networks.

C. Comparison of Computational Cost and Complexity

In the following, we will discuss the computational cost and complexity of the proposed PMCN. Table XIII shows the comparison of the number of parameters and floating-point operations (FLOPs) of different methods on five datasets, which are calculated with one batch size. The number of

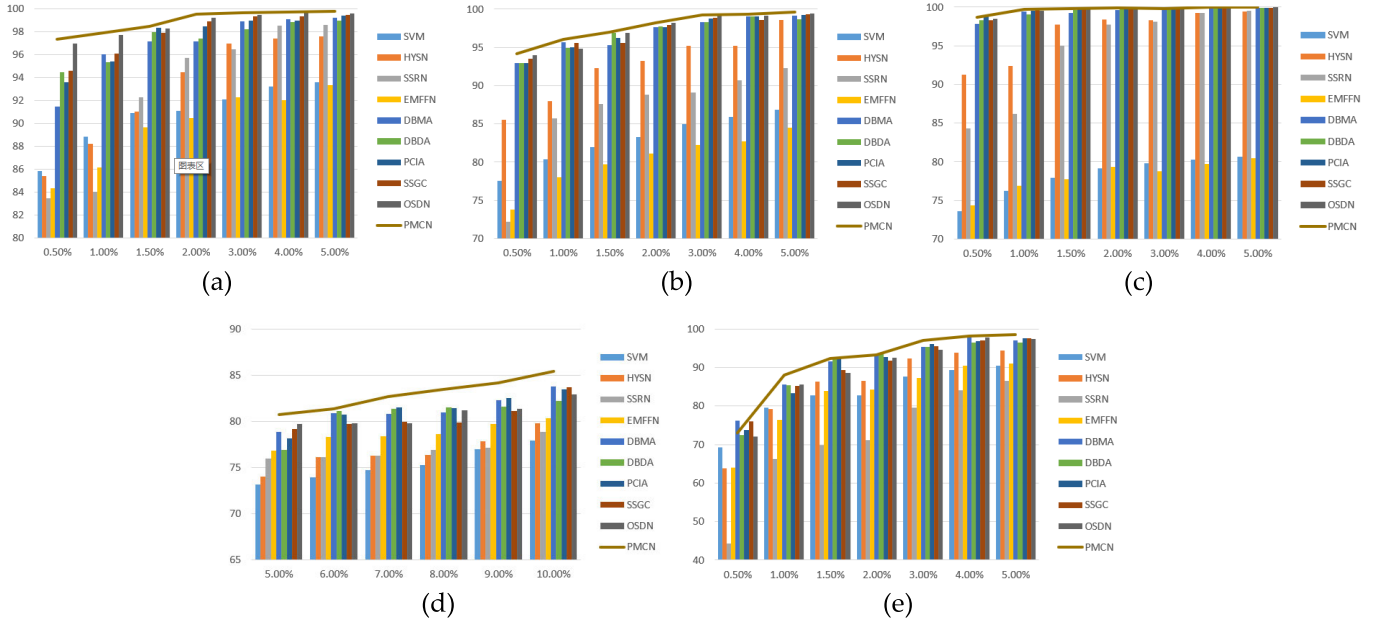


Fig. 16. Comparison of OA using different training sample proportions: (a) UP; (b) HH; (c) GF; (d) AH; and (e) HU.

TABLE XIII
NUMBER OF PARAMETERS (P) AND FLOPS (F) OF THE COMPETITORS

Dataset	Metrics	HYSN	SSRN	EMFFN	DBMA	DBDA	PCIA	SSGC	OSDN	PMCN
PU	P (M)	2.28	0.22	1.24	0.21	0.21	0.23	0.20	0.05	0.11
	F (MMac)	194.34	121.27	21.2	81.28	80.79	65.45	81.31	52.31	276.77
HH	P (M)	3.82	0.47	2.53	0.51	0.51	0.54	0.51	0.10	0.21
	F (MMac)	521.21	320.61	35.93	216.38	214.20	167.05	216.41	138.72	708.45
GF	P (M)	2.49	0.25	1.41	0.25	0.25	0.27	0.24	0.06	0.12
	F (MMac)	237.66	147.69	23.14	99.18	98.47	78.91	99.22	63.76	333.64
AH	P (M)	2.43	0.24	1.38	0.23	0.23	0.26	0.23	0.06	0.12
	F (MMac)	225.85	140.49	22.70	94.30	93.65	75.24	94.34	60.64	320.71
HU	P (M)	1.6	0.10	0.69	0.07	0.07	0.10	0.07	0.03	0.06
	F (MMac)	48.63	32.41	14.76	21.06	21.32	20.16	21.09	13.79	88.07

parameters and FLOPs of PMCN are highlighted in bold. We can see that the values of parameters and FLOPs vary with the size of the datasets and methods. In general, a larger dataset size leads to larger values of parameters and FLOPs. Checking the values of parameters, HYSN contains the highest number of parameters. It is because HYSN uses cascading stacked 3-D convolutional layers to jointly extract the spatial and spectral features. The EMFFN provides the second-highest number of parameters. It is due to the multiscale convolutional layers of the network. SSRN, DBMA, DBDA, PCIA, and SSGC contain a similar number of parameters, which are significantly lower than that of HYSN and EMFFN. It is because these methods improve the traditional cascading stacked 3-D convolutional layers to specialized 3-D convolutional blocks and divide the feature extraction module into the spatial branch and spectral branch individually. PMCN and OSDN contain a lower number of parameters than the former methods. It benefits from the use of lightweight feature extraction modules and the one-shot aggregation mechanism, which enables the extracted features to be finely fused in the convolutional networks. PMCN contains a larger number of parameters than OSDN due to its pyramidal multiscale convolutional blocks.

Observing the FLOPs of the methods, PMCN obtains the highest value of FLOPs. It is because PMCN processes the raw input data without reducing the dimensions. As a result, it is considered to use a dimension reduction algorithm to process the raw dataset to reduce the FLOPs of PMCN. In addition, the multiscale pyramid blocks also increase the FLOPs of PMCN. HYSN obtains higher FLOPs than those of other methods, except for PMCN. SSRN, DBMA, DBDA, PCIA, and SSGC obtain similar FLOPs. OSDN obtains lower FLOPs than that of the other methods except for EMFFN as expected. EMFFN obtains the lowest FLOPs in most cases for fewer convolutional layers conducted in the framework.

D. Ablation Analyses

In this section, we design four ablation experiments to analyze the effectiveness of the technologies applied in the proposed network, including the attention mechanism, the one-shot aggregation, the PyConv, and the Mish activation function. First, we perform an ablation experiment on the effectiveness of the attention mechanism. In the PMCN, two polarized attention blocks are implemented: channel-only attention block and spatial-only attention block. The

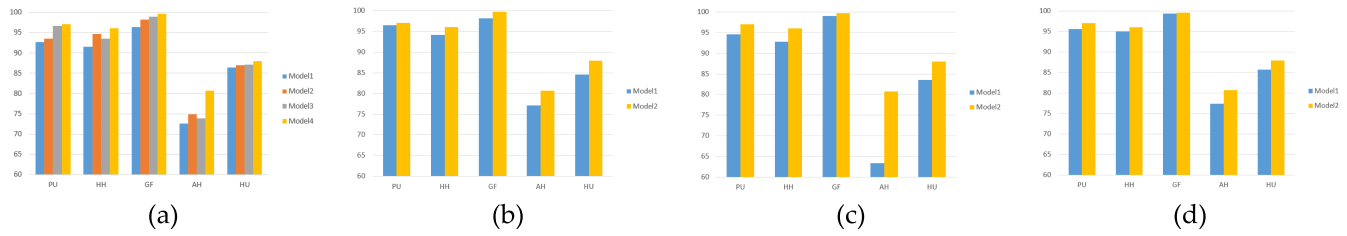


Fig. 17. Ablation experiments of PMCN on five HSI datasets: (a) ablation experiment for attention mechanism; (b) ablation experiment for one-shot aggregation; (c) ablation experiment for PyConv; and (d) ablation experiment for Mish activation function.

classification results of PMCN with different attention blocks are shown in Fig. 17(a). Model 1 denotes that no attention mechanism is applied in the PMCN. Model 2 denotes that only channel-only PSA block is applied in the PMCN. Model 3 denotes that only spatial-only PSA block is applied in the PMCN. Model 4 denotes that both channel-only and spatial-only blocks are used in the PMCN. Taking the UP dataset as an example, the baseline OA of PMCN is 92.64% when polarized attention blocks are not applied. Both channel-only attention block alone and spatial-only attention block alone improve the classification accuracy on the basis of baseline (0.88%, 3.99%). Comparing the improvement of classification accuracy of five HSI datasets by channel-only attention block and spatial-only attention block, it is found that the boost on network discrimination is variable in different datasets. It shows that the validity of the channel-only attention block and spatial-only attention block is determined by the characteristics of the dataset, which is not invariable. Finally, as expected, PMCN obtains the highest classification accuracy by using both channel-only and spatial-only PSA blocks. Second, we perform an ablation experiment on the effectiveness of the one-shot aggregation. The OAs are shown in Fig. 17(b). Model 1 denotes that one-shot aggregation is not applied in the PMCN, while Model 2 denotes that one-shot aggregation is applied. We can see that there is a slight improvement in classification accuracy when using one-shot aggregation, which ranges from 0.54% to 3.63%. The experimental results convincingly demonstrate the effectiveness of one-shot aggregation. Third, we conduct an ablation experiment on the effectiveness of the PyConv. The OAs are shown in Fig. 17(c). Model 1 denotes that only single-scale convolutional layers with $5 \times 1 \times 1$ and $1 \times 5 \times 5$ window sizes are applied in the two branches of PMCN. Model 2 denotes that pyramidal convolutional blocks are applied in the PMCN. We can see a significant improvement in classification accuracy when using PyConv, especially on the AH dataset (63.35%, 80.73%). It indicates that PyConv is superior to single-scale convolution in its ability to extract discriminative features from spectral signatures and spatial context information. Finally, we conduct the ablation experiment on the effectiveness of the Mish activation function. The classification results are shown in Fig. 17(d). Model 1 denotes that PReLU is applied in the classification subsection network of PMCN. Model 2 denotes that Mish is applied. We can see consistent improvements in classification accuracy on the five HSI datasets, which range from 0.31% to 3.39%. The results confirm the validity of the Mish activation function.

VI. CONCLUSION

In this article, a pyramidal multiscale convolutional network with PSA is proposed for pixel-wise HSI classification. The proposed PMCN mainly contains three stages: channel-wise feature extraction network, spatial-wise feature extraction network, and classification network. Pyramidal convolutional blocks and polarized attention blocks are converted to extract spectral and spatial features, respectively. The pyramidal convolutional blocks are used to extract multiscale features, and the polarized attention blocks are used to provide more flexibility. Compared to the previous attention mechanisms used in HSI classification methods, polarized attention can better process HSI with high internal resolution. Furthermore, residual aggregation and one-shot aggregation are employed to fuse feature maps of different layers. Finally, a classification network is used to obtain the classification results. Five different types of HSIs are introduced to evaluate the performance of the proposed PMCN. Nine representative methods are employed for our comparison. The experimental results show that the proposed method provides competitive performance among the related methods. In addition, the spatial patch size, training sample proportion, computational cost, and ablation analyses are discussed. In the future, we will combine PSA mechanism with other convolutional networks and apply these models to other HSI datasets.

REFERENCES

- [1] H. Pan, M. Liu, H. Ge, and L. Wang, "One-shot dense network with polarized attention for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 9, p. 2265, May 2022.
- [2] Y. Zhong, Q. Cao, J. Zhao, A. Ma, B. Zhao, and L. Zhang, "Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and LiDAR data," *Remote Sens.*, vol. 9, no. 8, p. 868, Aug. 2017.
- [3] M. B. Stuart, M. Davies, M. J. Hobbs, T. D. Pering, A. J. S. McGonigle, and J. R. Willmott, "High-resolution hyperspectral imaging using low-cost components: Application within environmental monitoring scenarios," *Sensors*, vol. 22, no. 12, p. 4652, Jun. 2022.
- [4] B. Lu, P. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, p. 2659, Aug. 2020.
- [5] S. Lorenz et al., "Radiometric correction and 3D integration of long-range ground-based hyperspectral imagery for mineral exploration of vertical outcrops," *Remote Sens.*, vol. 10, no. 2, p. 176, Jan. 2018.
- [6] B. Chung et al., "Detection of magnesite and associated gangue minerals using hyperspectral remote sensing—A laboratory approach," *Remote Sens.*, vol. 12, no. 8, p. 1325, Apr. 2020.
- [7] M. Shimon, R. Haelterman, and C. Perneel, "Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

- [8] H. Ge, L. Wang, H. Pan, Y. Zhu, X. Zhao, and M. Liu, "Affinity propagation based on structural similarity index and local outlier factor for hyperspectral image clustering," *Remote Sens.*, vol. 14, no. 5, p. 1195, Feb. 2022.
- [9] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, Aug. 2022, Art. no. 102926.
- [10] X. Li et al., "Multi-view learning for hyperspectral image classification: An overview," *Neurocomputing*, vol. 500, pp. 499–517, Aug. 2022.
- [11] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and Fisher vectors," *Remote Sens.*, vol. 8, no. 6, p. 483, Jun. 2016.
- [12] A. Satpathy, X. Jiang, and H.-L. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 287–297, Jan. 2014.
- [13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
- [14] X. Huang, Y. Ye, and H. Zhang, "Extensions of kmeans-type algorithms: A new clustering framework by integrating intracluster compactness and intercluster separation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 46–1433, Aug. 2014.
- [15] A. Samat, P. J. Du, S. C. Liu, J. Li, and L. Cheng, "(ELMs)-L-2: Ensemble extreme learning machines for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1060–1069, Apr. 2014.
- [16] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [17] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [18] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.
- [19] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, p. 298, Mar. 2017.
- [20] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.
- [21] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.
- [22] 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.
- [23] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.
- [24] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, 2015.
- [25] H. Gao, Y. Yang, C. Li, H. Zhou, and X. Qu, "Joint alternate small convolution and feature reuse for hyperspectral image classification," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 9, p. 349, Aug. 2018.
- [26] X. Jin, L. Jie, S. Wang, H. Qi, and S. Li, "Classifying wheat hyperspectral pixels of healthy heads and fusarium head blight disease using a deep neural network in the wild field," *Remote Sens.*, vol. 10, no. 3, p. 395, Mar. 2018.
- [27] Z. Qiu, J. Chen, Y. Zhao, S. Zhu, Y. He, and C. Zhang, "Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network," *Appl. Sci.*, vol. 8, no. 2, p. 212, Jan. 2018.
- [28] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, "Classification of hyperspectral imagery using a new fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 292–296, Feb. 2018.
- [29] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 344–357, Aug. 2018.
- [30] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with Gabor filtering and convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2355–2359, Dec. 2017.
- [31] C. Yu et al., "Hyperspectral image classification method based on CNN architecture embedding with hashing semantic feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1866–1881, Jun. 2019.
- [32] 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.
- [33] 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.
- [34] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.
- [35] S. K. Roy, S. R. Dubey, S. Chatterjee, and B. B. Chaudhuri, "FuSENet: Fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification," *IET Image Process.*, vol. 14, no. 8, pp. 1653–1661, 2020.
- [36] S. Jia et al., "A lightweight convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4150–4163, May 2021.
- [37] H. Shi, G. Cao, Z. Ge, Y. Zhang, and P. Fu, "Double-branch network with pyramidal convolution and iterative attention for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 7, p. 1403, Apr. 2021.
- [38] 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.
- [39] 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.
- [40] D. Liu et al., "A novel 2D-3D CNN with spectral-spatial multi-scale feature fusion for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 22, p. 4621, Nov. 2021.
- [41] S. Pande and B. Banerjee, "HyperLoopNet: Hyperspectral image classification using multiscale self-looping convolutional networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 422–438, Jan. 2022.
- [42] Y. Wang, B. Liang, M. Ding, and J. Li, "Dual-branch dense residual network for hyperspectral imagery classification," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2581–2602, Apr. 2020.
- [43] Z. Li, X. Cui, L. Wang, H. Zhang, X. Zhu, and Y. Zhang, "Spectral and spatial global context attention for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 4, p. 771, Feb. 2021.
- [44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–9.
- [45] R. S. Thakur, R. N. Yadav, and L. Gupta, "PRELU and edge-aware filter-based image denoiser using convolutional neural network," *IET Image Process.*, vol. 14, no. 15, pp. 3869–3879, Dec. 2020.
- [46] X. Hu, W. Yang, H. Wen, Y. Liu, and Y. Peng, "A lightweight 1-D convolution augmented transformer with metric learning for hyperspectral image classification," *Sensors*, vol. 21, no. 5, p. 1751, Mar. 2021.
- [47] W. Wei, J. Zhang, L. Zhang, C. Tian, and Y. Zhang, "Deep cube-pair network for hyperspectral imagery classification," *Remote Sens.*, vol. 10, no. 5, p. 783, May 2018.
- [48] 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.
- [49] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–10.
- [50] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
- [51] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 2, 2018, pp. 1–5.

- [52] J. B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.
- [53] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise mapping," *Neurocomputing*, vol. 506, pp. 158–167, Sep. 2022.
- [54] 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.
- [55] L. Zhu, R. Deng, M. Maire, Z. Deng, G. Mori, and A. P. Tan, "Sparsely aggregated convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 186–201.
- [56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.
- [57] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.
- [58] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112012.
- [59] B. Zhang, L. Zhao, and X. Zhang, "Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111938.
- [60] W. Chen, S. Ouyang, J. Yang, X. Li, G. Zhou, and L. Wang, "JAGAN: A framework for complex land cover classification using Gaofen-5 AHSI images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1591–1603, 2022.
- [61] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.
- [62] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3D-2D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
- [63] J. Yang, C. Wu, B. Du, and L. Zhang, "Enhanced multiscale feature fusion network for HSI classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10328–10347, Jan. 2021.
- [64] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with restarts," in *Proc. ICLR*, 2017, pp. 1–16.



Moqi Liu (Graduate Student Member, IEEE) received the B.E. degree in Internet of Things engineering from Zhoukou Normal University, Zhoukou, China, in 2020. He is currently pursuing the M.E. degree in electronic information with Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image classification and spectral reconstruction.



Xiaoyu Zhao received the B.E. degree from the Shandong University of Technology of Computer Science and Technology, Zibo, Shandong, China, in 2021. She is currently pursuing the master's degree in computer technology with the School of Computer and Control Engineering, Qiqihar University, Qiqihar, Heilongjiang, China.

Her research interests include deep learning and hyperspectral image classification.



Yuexia Zhu received the B.E. degree in computer science and technology from Henan Polytechnic University, Jiaozuo, China, in 2020. She is currently pursuing the M.E. degree in electronic information with Qiqihar University, Qiqihar, China.

Her research interests include hyperspectral image classification and deep learning.



Haimiao Ge received the M.S. degree from the College of Computer and Control Engineering, Qiqihar University, Qiqihar, China, in 2013.

He is currently an Associate Professor with the College of Computer and Control Engineering, Qiqihar University. His research interests include machine learning and hyperspectral image processing.



Haizhu Pan received the M.S. degree from the College of Computer and Control Engineering, Qiqihar University, Qiqihar, China, in 2009, and the Ph.D. degree from the College of Aerospace and Civil Engineering, Harbin Engineering University, Harbin, China, in 2018.

She is currently a Full Professor with the College of Computer and Control Engineering, Qiqihar University. Her research interests include numerical analysis, intelligent computing, machine learning, and hyperspectral image processing.



Ligu Wang received the M.S. degree in natural science and the Ph.D. degree in engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2002 and 2006, respectively.

He held a post-doctoral research position at the College of Information and Communications Engineering, Harbin Engineering University, Harbin, from 2006 to 2008. He is currently a Professor with the College of Information and Communication Engineering, Dalian Minzu University, Dalian, China. He has authored four books and more than

250 articles in journals and conference proceedings and holds 35 patents. His research interests are remote sensing image processing and machine learning.



Yanzhong Liu received the M.A. degree from the College of Chemical Engineering, Qiqihar University, Qiqihar, China, in 2012.

He is currently an Associate Professor with Qiqihar University. His research interests include pattern recognition and deep learning.