IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 7, 2019, accepted November 29, 2019, date of publication December 2, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957163

Hyperspectral Image Classification With Pre-Activation Residual Attention Network

HONGMIN GAO[1,2], YAO YANG[1], DAN YAO[1], AND CHENMING LI[1]
[1]College of Computer and Information, Hohai University, Nanjing 211100, China
[2]Nantong Ocean and Coastal Engineering Research Institute, Hohai University, Nantong 226300, China

Corresponding author: Chenming Li (lcm@hhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701166, in part by the National Key Research and Development Program of China under Grant 2018YFC1508106, in part by the Fundamental Research Funds for the Central Universities under Grant 2018B16314, in part by the Nantong Science and Technology Project under Grant MS12017026-2, in part by the China Postdoctoral Science Foundation under Grant 2018M632215, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2017QNRC001, in part by the National Science Foundation for Young Scientists of China under Grant 51709271, and in part by the Science Fund for Distinguished Young Scholars of Jiangxi Province under Grant under Grant 2018ACB21029.

ABSTRACT Recently, convolutional neural networks (CNNs) have been introduced for hyperspectral image (HSI) classification and shown considerable classification performance. However, the previous CNNs designed for spectral-spatial HSI classification lay stress on the learning for the spatial correlation of HSI data and neglect the channel responses of feature maps. Furthermore, the lack of training samples remains the major challenge for CNN-based HSI classification methods to achieve better performance. To address the aforementioned issues, this paper proposes a new end-to-end pre-activation residual attention network (PRAN) for HSI classification. The pre-activation mechanism and attention mechanism are introduced into the proposed network, and a pre-activation residual attention block (PRAB) is designed, which allows the proposed network to carry adaptively feature recalibration of channel responses and learn more robust spectral-spatial joint feature representations. The proposed PRAN is equipped with two PRABs and several convolutional layers with different kernel sizes, which enables the PRAN to extract high-level discriminative features. Experimental results on three benchmark HSI datasets reveal that the proposed method is provided with competitive performance over several state-of-the-art HSI classification methods, especially when the training set size is relatively small.

INDEX TERMS Hyperspectral image classification, convolutional neural network, pre-activation mechanism, attention mechanism.

I. INTRODUCTION

Hyperspectral images (HSIs) are composed of hundreds of continuous spectral channels with spectral resolution of nanometer order. Compared with ordinary remote sensing images, HSIs contain more abundant spectral and spatial information, which makes the accurate identification of ground materials possible [1]. Therefore, hyperspectral remote sensing technology has been widely used in many fields, including agriculture [2], environmental earth sciences [3], military surveillance [4]. Furthermore, HSI classification has become a very hot research topic in the remote sensing analysis field.

Most traditional methods only incorporate spectral information to achieve HSI classification, such as k-nearest neighbor [5], support vector machine (SVM) [6], [7], multinomial logical regression [8], [9], extreme learning machine [10] and so on. Although those methods can make full use of spectral information, the final classification accuracy is unsatisfactory due to obvious intra-class differences and unobvious inter-class differences of hyperspectral data on the spectral domain. Besides, the curse of dimensionality, namely the Hughes phenomenon [11], makes it a challenge for those methods to achieve better classification performance.

In order to enhance the classification performance, many spectral-spatial classification methods have been proposed, which can extract both spectral and spatial features of hyperspectral data. For instance, Benediktsson *et al.* [12] adopted multiple morphological operations to design spectral-spatial classifier. Yu *et al.* [13] integrated the subspace-based SVM classification method with an adaptive Markov

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

random field (MRF) approach to model the spectral and spatial information. In [14], [15], sparse representation was introduced to analyze and process HSI. Zhou *et al.* [16] developed a spectral-spatial feature learning method, which exploited spectral and spatial features in a hierarchical fashion and adopted kernel-based extreme learning machine to classify image pixels. In [17], the 3-dimensional (3-D) discrete wavelet transform was combined with MRF for HSI classification. A new discriminative low-rank Gabor filter method [18] was proposed to classify hyperspectral data and provided with excellent performance in terms of both accuracy and computation time.

The above-mentioned methods only extract hand-crafted features and highly depend on domain knowledge, though they can improve the accuracy of HSI classification. On the contrary, deep learning methods can automatically learn hierarchical feature representation from the raw data in an end-to-end manner, thus avoiding the process of hand-crafted features extraction. In recent years, deep learning has attracted increased attention for its remarkable performance in many fields, such as image classification [19], [20], target detection [21], [22], and natural language processing [23]. Motivated by these successful applications, many efforts have been made to classify hyperspectral data based on the deep learning. Chen *et al.* [24] first introduced the stacked autoencoder, a deep learning framework, to extract spectral and spatial features of the HSI. After that, Liu *et al.* [25] combined the stacked denoising autoencoders and superpixel-based spatial constraints to improve the HSI classification performance. In [26], a stacked sparse autoencoder was proposed to adaptively construct features from unlabeled data by learning a feature mapping function. Moreover, the reference [27] proposed a compact and discriminative stacked autoencoder for HSI classification. In [28] and [29], the deep belief network was also introduced for HSI classification. Li *et al.* [30] adopted 1-D convolution layers and proposed an adaptive spatial-spectral feature learning network. Although the aforementioned deep models [24]–[30] can extract deep hierarchical features, the input sample must be flattened into a 1-D vector in order to satisfy the input requirement, which results in that they cannot make full use of the spatial information of HSIs. Moreover, limited labeled samples of the HSI make those deep learning models be plagued by small sample size problem, which causes great challenges for HSI classification.

To solve the above problems, many researchers designed 2-D CNN model to extract discriminative spatial features from 3-D image cubes [31]–[38]. For instance, to learn the joint spectral-spatial features from HSI, Yang *et al.* [33] proposed a two-branch CNN and trained the model through transfer learning. Lee *et al.* [35] proposed a contextual deep CNN (DCNN), where a multi-scale filter bank was utilized to achieve the joint exploitation of the spatial-spectral information. The [36] combined CNN with MRF to classify HSIs. Song *et al.* [37] adopted residual connection and proposed a deep feature fusion network (DFFN), which can fuse the outputs from different hierarchical layers. To learn the spectral-spatial features, Ma *et al.* [38] designed a deep deconvolution network with skip architecture. Though those CNN-based HSI classification methods can utilize the spatial context information, they only convolve feature maps on the spatial dimension and neglect the spectral correlations, which are very important for HSI classification. For the reason that all the convolutional layers in their architectures applied 2-D convolutional operations.

Considering the limitation of 2-D convolution layers, some 3-D CNN models were proposed to classify hyperspectral data [39]–[41]. The 3-D convolutional operations can convolve feature maps on both spatial dimension and spectral dimension simultaneously, and then enables the 3-D CNN extract spectral correlation and joint spectral-spatial correlation information. Paoletti *et al.* [42] proposed a deep 3-D CNN architecture and obtained high classification accuracies. In [43], a spectral-spatial residual network (SSRN) was developed and the SSRN can consecutively learn discriminative features from abundant spectral signatures and spatial contexts in an HSI. A pyramidal residual network [44] were also developed to capture the spectral and spatial features simultaneously. Wang *et al.* [45] proposed a fast dense spectral-spatial convolution framework, which extracted spectral and spatial features separately by designing dense spectral block, dense spatial block and reducing dimension layer. Furthermore, a multiscale deep middle-level feature fusion network [46] was proposed to extract more discriminative features by fusing multiscale deep middle-level features. Very recently, Chen *et al.* [47] explored the automatic design of CNN for HSI classification for the first time and developed a 3D Auto-CNN model. Those CNN-based methods effectively improve the classification accuracy of HSIs and perform well on small training set. However, they attach importance to learn the spatial correlation of HSI data and neglect the channel responses of features, which are also crucial for the HSI classification. Moreover, to deal with the gradient vanishing/explosion problem and mitigate the overfitting problem caused by limited training samples, residual connection is widely used in many existing CNN-based HSI classification methods such as the DCNN [35], DFFN [37] and SSRN [43]. However, the residual blocks in their network adopt post-activation mechanism, which means the activation function ReLU is after convolutional operation. The ReLU will forcibly convert the signal to 0 if the signal is negative, which may cause the loss of some informative residual features.

To address the above issues, this paper builds a novel residual network with attention mechanism for spectral-spatial HSI classification. The main contributions of this paper can be summarized as follows.

1) To deal with the gradient vanishing/explosion problem and enhance the classification performance of proposed network, residual connections and batch normalization (BN) are adopted in the proposed network. Different from the previous networks used in

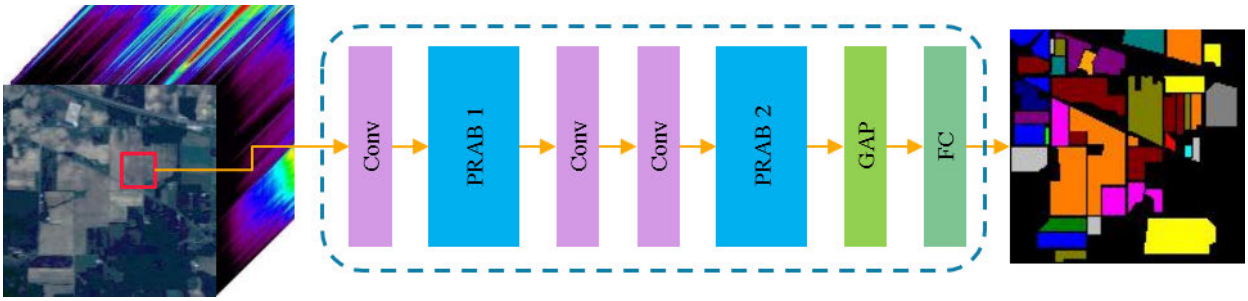


FIGURE 1. Framework of proposed HSI classification method.

HSI classification, we introduce the pre-activation mechanism into the residual block to learn more robust spectral and spatial feature representation, thus achieve better generalization performance.

2) To learn more robust spectral-spatial feature representations from input image patches, we introduce attention mechanism into the residual block and construct a pre-activation residual attention block (PRAB), which can adaptively recalibrate channel feature responses by explicitly modelling interdependencies between channels.
3) A pre-activation residual attention network (PRAN) is proposed to improve HSI classification outcomes on small training sample size. The proposed PRAN contains two PRABs, which allow the network to better learn hierarchical features. Note that the experimental results on three real HSIs demonstrate the competitive advantage of proposed PRAN in terms of accuracy over several state-of-the-art HSI classification methods.

The remainder of this paper is organized as follows. Section 2 introduces the proposed HSI classification method in detail. In section 3, the performance of proposed method is evaluated by carrying comparisons with several state-of-the-art HSI classification method, and the experimental results on three benchmark HSI datasets are analyzed and discussed. Finally, section 4 concludes this paper and suggests some future works.

II. PROPOSED METHOD

The framework of proposed method is showed in Figure 1. As can be observed, the PRAN includes three convolutional layers, two PRABs, a global average pooling (GAP) layer and a fully connection (FC) layer.

HSI dataset can be denoted as $D \in R^{H\times W\times B}$, where H, W and B denote the height, width and band number of the HSI, respectively. In order to extract spectral-spatial features, we adopt 3-D image patches centered on labeled pixels as the input samples of the proposed PRAN, and the label of image patch is the label of corresponding center pixel. The size of the image patch is $S \times S \times B$, where $S \times S$ denotes the neighborhood spatial size. Suppose the HSI dataset contains N labeled pixels, then the image patch set can be denoted as $X = \{x_1, x_2, \ldots, x_N\} \in R^{S\times S\times B}$, where x_i is the ith image patch. The corresponding ground-truth label set can be denoted as $Y = \{y_1, y_2, \ldots, y_N\}$, where $y_i \in \{1, 2, \ldots, Q\}$ is the label of x_i and Q is the number of land-cover classes. The patch set X is divided into training set, validation set and test set. Correspondingly, the Y is divided into three groups. Before training the PRAN, hyperparameters (such as learning rate, batch size, and patch size) are configured. The PRAN is trained for 200 epochs. In each epoch, the training set is divided into some mini-batches and the mini-batch data is fed into the network one by one. In the training process, the prediction label vectors of training set are obtained through forward propagation of the model, then cross entropy loss function is adopted to compute the difference between predicted label vectors and the corresponding one-hot label vectors which converted by the ground-truth labels. After that, the learned parameters of the PRAN are updated through back propagation algorithm. In addition, during training stage, the validation set is classified and the classification accuracy is computed every few epochs, so as to monitor the model performance. In this way, we can select the trained model with the highest accuracy. Finally, the test set is adopted for evaluating the performance of trained PRAN.

A. RESIDUAL CONNECTION AND PRE-ACTIVATION MECHANISM

The residual block is adopted as the key component of proposed PRAN, the architecture of which is shown in Figure 2(a). As can be seen, the residual block is composed of two convolutional layers and a residual connection (also known as skip connection). Through skip connection, the low-level features and high-level features can be aggregated in an addition manner. In this way, the residual block can mitigate the gradient vanishing/explosion problem which usually exist in deep network. Each residual block can be calculated as follows:

$$H(x) = f\left(F(x) + x\right) \tag{1}$$

where x and $H(\cdot)$ denote the input and output of residual block, respectively, F refers to a residual learning function and $F(x)$ denotes the output of convolutional layer before summation operation, f denotes the activation function.

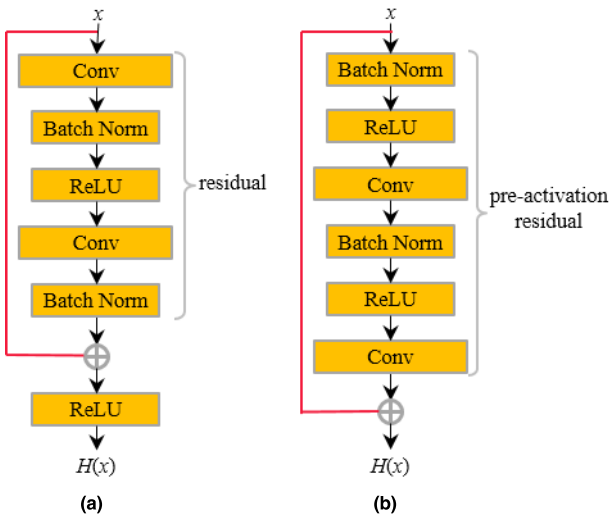


FIGURE 2. Architecture of normal residual block (a) and pre-activation residual block (b).

To obtain better performance, we apply BN and pre-activation mechanism in the residual block of proposed network. As shown in Figure 2(b), the pre-activation architecture is implemented by moving BN and ReLU activation function before convolution operation. The pre-activation residual block can be calculated as

$$H(x) = F(x) + x \tag{2}$$

As shown in Figure 2(a), the activation function f in (1) is ReLU, which means

$$f(x) = \max(0, x) \tag{3}$$

The ReLU will forcibly converts the signal to 0 if the signal is negative, which may cause the loss of some informative residual features in normal residual block. If make the f an identity mapping, the (1) will be equivalent to (2). The identity mapping enables signals be propagated directly between any two units, which means the features learned by the residual learning function will not be lost. In this way, the pre-activation mechanism makes it easier to train the network, and enhances the generalization performance of the network.

B. PRE-ACTIVATION RESIDUAL ATTENTION BLOCK

Due to that the data with all spectral bands are directly used as the inputs of proposed network, it is inevitable to carry redundant information which may degrade the classification accuracy. To address this issue, we adopt the Squeeze-and-Excitation (SE) block [48] to adaptively recalibrate channel feature responses by explicitly modelling interdependencies between channels, thus it can be regarded as a channel attention mechanism. We add the attention mechanism into the pre-activation residual block and propose a pre-activation residual attention block (PRAB).

The details of PRAB are depicted in Figure 3. The attention mechanism is added after convolution operation, but before summation operation. It allows the PRAB to perform feature recalibration, thus selectively emphasizes informative

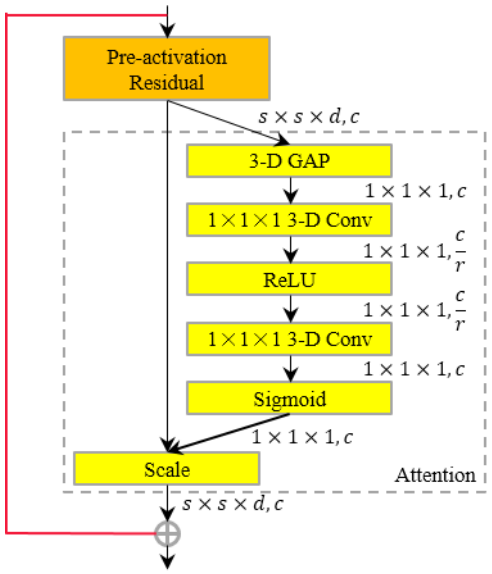


FIGURE 3. Architecture of pre-activation residual attention block.

features and suppress the less important features. Assume the size and number of the input feature maps of attention mechanism is $s \times s \times d$ and c, where c and d denote the size of channel dimension and depth dimension, respectively. Each feature map is first processed by a 3-D global average pooling (GAP) layer to squeeze global spatial information, thus c $1 \times 1 \times 1$ channel feature tensors are generated. Then, the feature tensors are input into a $1 \times 1 \times 1$ 3-D convolutional layer to reduce the channel dimensionality. Specifically, after convolution operation, the channel dimension of the feature tensor becomes c/r, where r is a reduction ratio. In the proposed network, r is set to 4. Next, a ReLU function is applied to improve nonlinearity of channel responses and another $1 \times 1 \times 1$ 3-D convolutional layer is adopted to increase the channel dimension and generates c feature tensors. Lastly, a sigmoid function is employed, and the output is multiplied with the feature maps from pre-activation residual to rescale the final output of attention mechanism to $cs \times s \times d$ feature maps. In this way, channel weights are assigned to each feature map, thus achieve adaptively recalibrating features. Furthermore, an attention mechanism is provided with $2 * c^2/r$ parameters, which are derived from the two 3-D convolutional layers within it. Note that the proposed PRAN only contains two attention blocks, which cause increasing very few parameters for the network.

C. ARCHITECTURE OF PROPOSED NETWORK

Taking Indian Pines dataset as an example and the $7 \times 7 \times 200$ image patches are used as the input samples, the details of proposed PRAN are shown in Figure 4. Each convolutional layer is followed with BN and ReLU except that in the PRAB. Referring to SSRN, the proposed PRAN first puts particular emphasis on learning spectral features from raw input data, then puts particular emphasis on learning spatial features, thus extracts discriminative spectral-spatial joint features.

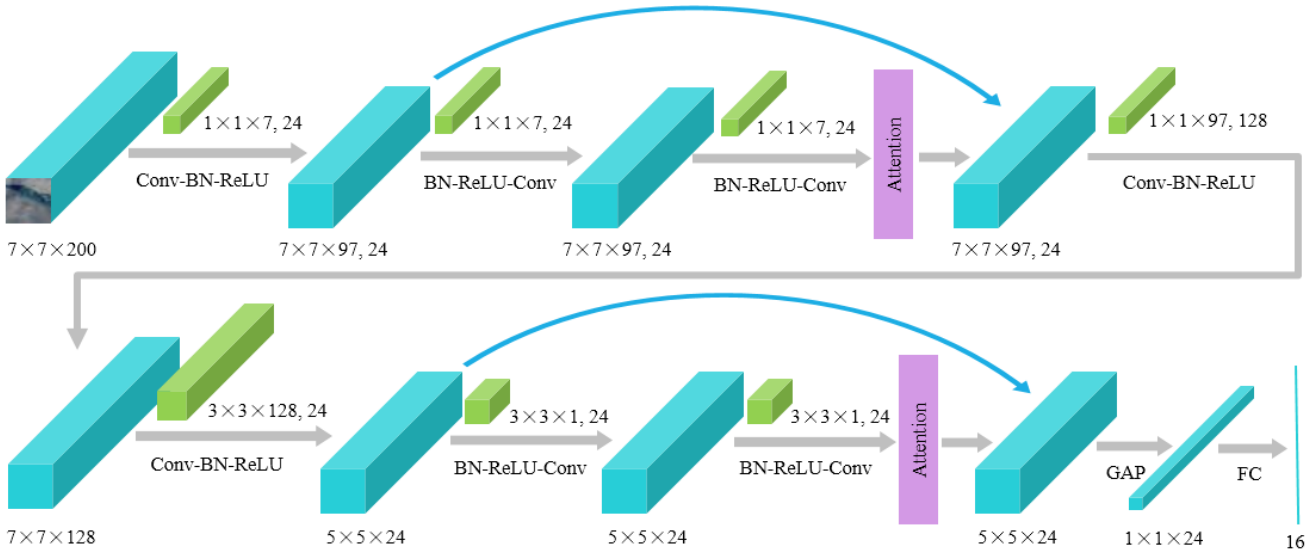


FIGURE 4. Architecture of proposed pre-activation residual attention network. The cuboids in mainstream refer to features, other cuboids refer to convolution kernels. "1 × 1 × 7, 24" means 24 convolution kernels with size 1 × 1 × 7, and "7 × 7 × 97, 24" means 24 feature cuboids with size 7 × 7 × 97. Other parameters have similar meanings, and no further elaboration is needed.

Finally, the joint features are processed by GAP and FC operation. The FC operation can adaptively generate feature vector, the length of which is equal to the number of land-cover classes in the HSI data. Because there are 16 land-cover classes in Indian Pines dataset, the length of output vector is 16 in Figure 4. In addition, it is noted that the stride of the first convolutional layer is (1, 1, 2), so the channel dimension of the input samples is reduced from 200 to 97. All the other convolutional layers in the proposed PRAN is equipped with the stride of (1, 1, 1). In PRAB, all convolutional layers use padding to keep the sizes of feature cuboids unchanged. Due to without using padding, the spatial size or channel dimension is reduced when feature cuboids are processed by the convolutional layers outside the PRAB.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATA SETS

1) INDIAN PINES

This image was captured by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the Indian Pines test site in North-western Indiana in 1992. It contains 16 classes and 145 × 145 pixels with the spatial resolution of 20m per pixel. There are 224 spectral bands in the wavelength range from 400 to 2500 nm. After discarding 20 water absorption bands, the remaining 200 bands are adopted for classification. Figure 5 shows the pseudo color image and ground-truth image of this data. As reported in Table 1, 20%, 10%, and 70% of labeled samples are randomly selected for training, validation (val), and test sets, respectively.

2) PAVIA UNIVERSITY

This image was captured by Reflective Optics System Imaging Spectrometer in Northern Italy in 2001. It contains

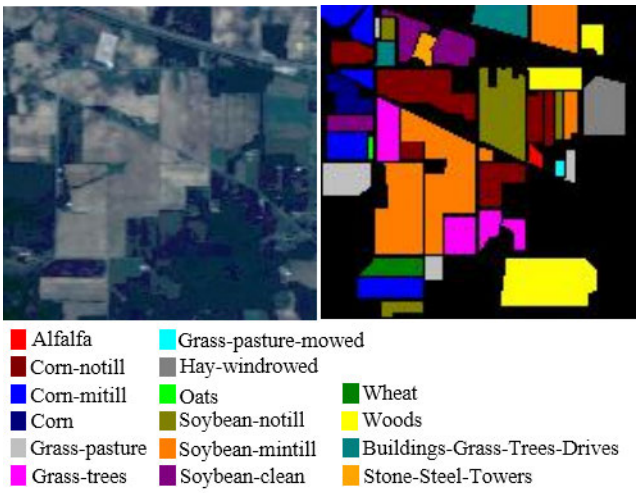


FIGURE 5. Pseudo color image and ground-truth map of Indian Pines data.

9 land-cover classes and 610 × 340 pixels with the spatial resolution of 1.3m per pixel. After discarding the noisy bands, the remaining 103 bands are adopted for experiments, which covers the wavelength range from 430 to 860 nm. Figure 6 shows the pseudo color image and ground-truth image of this data. As reported in Table 2, 10%, 10%, and 80% of labeled samples are randomly selected for training, validation, and test sets, respectively.

3) SALINAS

This image was collected by the AVIRIS sensor over Salinas Valley, California. It contains 16 land-cover classes and 512 × 217 pixels with the spatial resolution of 3.7m per pixel. We discarded the 20 water absorption bands and only 204 bands are persevered for experiments. The pseudo color

TABLE 1. Number of training, validation and test samples in Indian Pines dataset.

No.	Class	Training	Val	Test
1	Alfalfa	9	4	33
2	Corn-notill	285	142	1001
3	Corn-mitill	166	83	581
4	Corn	47	23	167
5	Grass-pasture	96	48	339
6	Grass-trees	146	73	511
7	Grass-pasture-mowed	5	2	21
8	Hay-windrowed	95	47	336
9	Oats	4	2	14
10	Soybean-notill	194	97	681
11	Soybean-mintill	491	245	1719
12	Soybean-clean	118	59	416
13	Wheat	41	20	144
14	Woods	253	126	886
15	Buildings-Grass-Trees-Drives	77	38	271
16	Stone-Steel-Towers	18	9	66
	Total	2045	1018	7186

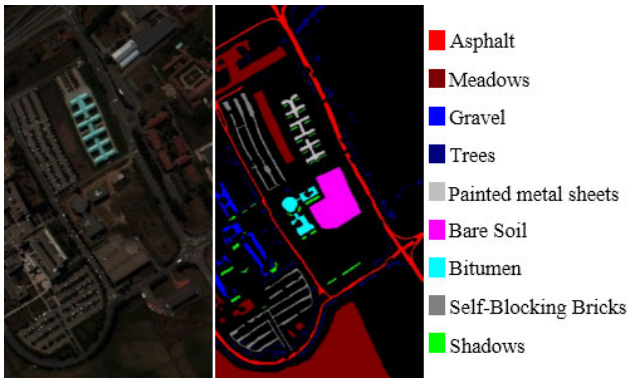


FIGURE 6. Pseudo color image and ground-truth map of Pavia University data.

TABLE 2. Number of training, validation and test samples in Pavia University dataset.

No.	Class	Training	Val	Test
1	Asphalt	663	663	5305
2	Meadows	1864	1864	14921
3	Gravel	209	209	1681
4	Trees	306	306	2452
5	Painted metal sheets	134	134	1077
6	Bare Soil	502	502	4025
7	Bitumen	133	133	1064
8	Self-Blocking Bricks	368	368	2946
9	Shadows	94	94	759
	Total	4273	4273	34230

image and ground-truth image are shown in Figure 7. For this dataset, the ratio of training samples, validation samples, and test samples is 1:1:8, the details are reported in Table 3.

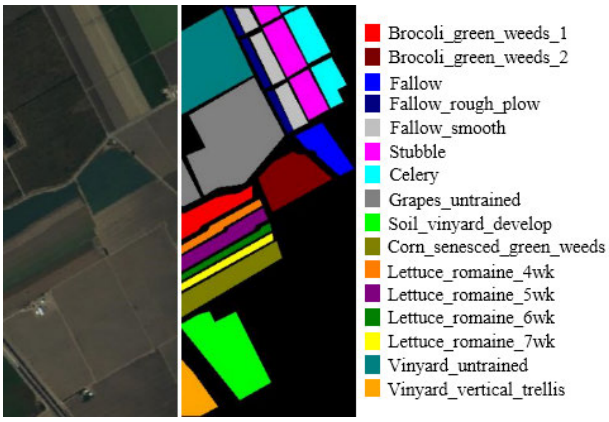


FIGURE 7. Pseudo color image and ground-truth map of Salinas data.

TABLE 3. Number of training, validation and test samples in Salinas dataset.

No.	Class	Training	Val	Test
1	Brocoli_green_weeds_1	200	200	1609
2	Brocoli_green_weeds_2	372	372	2982
3	Fallow	197	197	1582
4	Fallow_rough_plow	139	139	1116
5	Fallow_smooth	267	267	2144
6	Stubble	395	395	3169
7	Celery	357	357	2865
8	Grapes_untrained	1127	1127	9017
9	Soil_vinyard_develop	620	620	4963
10	Corn_senesced_green_weeds	327	327	2624
11	Lettuce_romaine_4wk	106	106	856
12	Lettuce_romaine_5wk	192	192	1543
13	Lettuce_romaine_6wk	91	91	734
14	Lettuce_romaine_7wk	107	107	856
15	Vinyard_untrained	726	726	5816
16	Vinyard_vertical_trellis	180	180	1447
	Total	5403	5403	43323

B. EXPERIMENTAL SETUP

The overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) are used to evaluate the classification performance of proposed method. Among them, OA denotes the ratio of the number of samples correctly classified to the total number of all labeled samples. AA denotes the average of classification accuracy of all classes. The kappa coefficient is used to assess the agreement of classification for all the classes. The greater the κ value is, the better the overall classification effect is.

We repeat all experiments for 10 times with randomly selected training samples so as to obtain the mean and standard deviation of OA, AA, and κ. In the training process, for all three datasets, learning rate, batch size and total epochs are 0.0003, 32 and 200, respectively. The RMSProp optimizer is adopted to optimal the learnable parameters of proposed network. All experiments are conducted on a computer with

TABLE 4. Classification results (OA) of proposed method with different patch size.

Patch size	Indian pines	Pavia University	Salinas
3 × 3	95.95±0.65	98.35±0.13	96.23±0.23
5 × 5	99.45±0.18	99.79±0.05	99.45±0.13
7 × 7	99.67±0.14	99.92±0.03	99.93±0.01
9 × 9	99.55±0.14	99.91±0.04	99.98±0.01
11 × 11	99.43±0.15	99.90±0.04	99.94±0.01

RAM 8G and NVIDIA GeForce GTX 1050Ti GPU (4G of ROM). The experimental results are divided into three parts. First, we analyze the effect of size of input image patches on the performance of proposed method. Second, the effectiveness of PRAB is verified. Finally, the performance of proposed method is evaluated by comparing with other classification methods.

C. EFFECT OF PATCH SIZE

To analyze the effect of patch size for the performance of proposed method, we carried the experiments on all three datasets and compared the classification results of proposed method with different patch size, as shown in Table 4. It can be observed that, as the patch size increases, the OA value first increases rapidly and then decreases slightly for Indian Pines dataset. And for the other two datasets, the OA value first increases rapidly and then becomes stable. The reason is that small patch size (3×3) makes the spatial information is not fully utilized, thus causes unsatisfactory classification accuracies. And larger patch size enables the proposed method to extract more discriminative features and achieve better classification results. However, when the image patch exceeds a certain size, it will lead to redundant information or noise, which cannot cause the improvement on accuracy and may even degrade the accuracy. The optimal patch size is 7×7, 7×7 and 9×9 for Indian Pines, Pavia University and Salinas datasets, respectively. Considering the larger patch size leads to higher computational cost, the patch size is to 7×7 for all the three datasets.

D. EFFECT OF PRE-ACTIVATION RESIDUAL ATTENTION BLOCK

To verify the effectiveness of the attention mechanism and PRAB, we compare the proposed PRAN with the deep residual network (DRN) and pre-activation residual network (PRN). Among them, the DRN is obtained by replacing the PRABs in PRAN with normal residual blocks, and the difference between PRN and PRAN is that the PRN does not adopt attention mechanism, while the PRAN is provided with attention mechanism. Table 5 reports the classification results of the DRN, PRN and PRAN on all three datasets. It is obvious that proposed PRAN achieved better classification results than the DRN for all three datasets, which demonstrates the superiority of the PRAB. Furthermore, compared with the PRN, the PRAN improves the classification accuracies of all three datasets, because the attention mechanism selectively strengthens informative channels and suppresses less useful channels, thus results in the proposed PRAN can learn discriminative spectral and spatial features simultaneously. For the reason that the accuracy is very high (higher than 99%), the improvements of accuracy caused by the PRAN are not so obvious. Note that all the accuracies are the averaged results over 10 repeated experiments with randomly selected training samples, small improvements demonstrate the effectiveness of PRAB to some extent.

TABLE 5. Effect of prab on all three datasets.

Datasets	Metrics	DRN	PRN	PRAN
Indian Pines	OA(%)	99.28±0.19	99.56±0.16	99.67±0.14
	AA(%)	99.01±0.76	99.25±0.64	99.37±0.57
	κ×100	99.13±0.19	99.50±0.16	99.62±0.16
Pavia University	OA(%)	99.47±0.04	99.83±0.04	99.92±0.03
	AA(%)	99.39±0.05	99.75±0.06	99.87±0.06
	κ×100	99.42±0.04	99.81±0.05	99.90±0.04
Salinas	OA(%)	99.52±0.05	99.82±0.04	99.90±0.04
	AA(%)	99.49±0.02	99.81±0.02	99.93±0.01
	κ×100	99.46±0.04	99.78±0.04	99.89±0.04

E. COMPARISON OF DIFFERENT CLASSIFICATION METHODS

We compare our method with the SVM [7] and several state-of-the-art CNN-based methods, including DCNN [35], DFFN [37] and SSRN [43]. For SVM-based method, only single RBF kernel is adopted, the optimal kernel parameter γ and the penalty parameter C are tuned by grid search method. Additionally, the original data is processed by principal component analysis (PCA), then training patches (patch size is 25×25) centered with labeled pixels are extracted. The patches are transformed into one-dimension data to training the SVM. The DCNN and DFFN are 2-D CNN, and the SSRN is a 3-D CNN. All of them used residual connections to design deep architectures and improve their performance in HSI classification. The architectures of those three CNN models are deeper than PRAN. Among them, the number of layers with weights in DCNN and SSRN is 10 and 12, respectively. In DFFN, there is more than 20 layers with weights. In addition, both the DCNN and SSRN adopt 3D image patches extracted from original HSI as the inputs. As for the DFFN, PCA is applied over the hyperspectral data to reduce the dimensions and obtain major spectral information, then input image patches are extracted from the dimension-reduced data. The optimal hyperparameters of DCNN, DFFN and SSRN are set as corresponding references. For fair comparison, the local response normalization in DCNN is replaced by BN. The division of datasets is according to Tables 1-3.

Tables 6-8 report the classification results of different methods on three datasets. As we can see, the accuracies obtained by SVM classifier are the lowest for all three

TABLE 6. Classification results of different methods on Indian Pines dataset.

Class	SVM	DCNN	DFFN	SSRN	PRAN
1	72.16	95.15	97.56	98.57	**98.78**
2	98.97	96.88	97.99	**99.45**	99.42
3	91.14	98.00	98.65	99.14	**99.52**
4	51.05	95.92	97.06	99.11	**99.55**
5	93.17	98.11	98.40	98.90	**99.16**
6	97.65	99.82	99.27	**99.88**	99.86
7	39.56	97.61	96.66	97.85	**98.22**
8	88.01	99.97	99.88	99.95	**100.0**
9	27.50	87.85	93.78	98.40	**98.57**
10	92.45	97.22	97.94	**99.51**	99.42
11	95.16	97.86	98.23	99.13	**99.71**
12	63.34	97.66	98.32	99.26	**99.68**
13	91.89	99.58	99.61	99.30	**99.85**
14	92.61	98.93	**99.89**	99.28	99.71
15	71.00	96.49	97.19	**99.86**	99.85
16	0.00	97.13	96.45	97.85	**98.62**
OA(%)	89.44±0.61	97.94±0.28	98.23±0.28	99.41±0.17	**99.67±0.14**
AA(%)	72.85±0.14	97.14±0.99	97.93±0.87	99.09±0.79	**99.37±0.57**
$\kappa\times100$	87.87±0.70	97.66±0.32	97.87±0.32	99.33±0.19	**99.62±0.16**

TABLE 7. Classification results of different methods on Pavia University dataset.

Class	SVM	DCNN	DFFN	SSRN	PRAN
1	97.34	98.37	99.57	99.89	**99.96**
2	99.63	99.88	99.95	99.98	**99.99**
3	89.94	94.14	99.29	99.09	**99.67**
4	95.52	98.33	95.34	99.64	**99.76**
5	**100.0**	99.99	99.42	**100.0**	99.95
6	97.77	99.25	99.95	**100.0**	**100.0**
7	92.65	95.99	99.28	**99.94**	**99.94**
8	95.42	96.66	98.54	99.61	**99.67**
9	75.82	99.94	93.51	**100.0**	99.86
OA(%)	97.19±0.23	98.78±0.19	99.23±0.08	99.87±0.04	**99.92±0.03**
AA(%)	93.79±0.53	98.06±0.33	98.32±0.20	99.79±0.06	**99.87±0.06**
$\kappa\times100$	96.28±0.31	98.39±0.25	98.98±0.11	99.82±0.06	**99.90±0.04**

datasets. Because it requires 1-D input data, which causes the loss of spatial information. In addition, it cannot extract deep hierarchical features due to its shallow structure. All CNN-based comparison methods achieve high classification accuracies. The main reason is that they are equipped with deep architecture, which enables them to learn high-level discriminative features of HSIs. In addition, the residual connection adopted in those methods effectively alleviates the overfitting of deep architecture. However, most convolutional layers in the DCNN are composed of 1×1 convolutional kernels, which leads to the limited ability to extract spatial correlation features. Instead, many 3×3 convolutional layers are stacked in the DFFN, SSRN and PRAN, thus these methods can extract more informative spatial correlation features and achieve higher accuracies.

Compared with the DFFN, the PRAN consistently provides excellent performance for all three datasets. For example, the PRAN achieves 1.44% and 0.69% increase of mean OA for Indian Pines and Pavia University data, respectively. For Salinas data, the OA/AA/κ obtained by the DFFN are slightly higher than those obtained by the PRAN, but the gap is extremely small (only 0.03% in mean OA). Note that samples of the classes Asphalt, Grass-pasture-mowed, Oats, Wheat in Indian Pines datasets are very few, the PRAN performs obviously better in these classes than the DFFN. It demonstrates that the PRAN can extract discriminative features more robustly than the DFFN. As for the SSRN, the PRAN performs marginally better than it for all three datasets, but the improvements are not that clear, simply because the classification accuracy is very high (higher than

TABLE 8. Classification results of different methods on Salinas dataset.

Class	SVM	DCNN	DFFN	SSRN	PRAN
1	99.98	99.74	**100.0**	99.99	**100.0**
2	99.96	99.99	99.96	**100.0**	**100.0**
3	99.87	99.89	**100.0**	**100.0**	**100.0**
4	99.88	99.86	99.87	**99.91**	**99.91**
5	99.78	99.06	99.51	99.52	**99.80**
6	99.98	99.99	99.94	**100.0**	**100.0**
7	99.99	99.86	99.94	**100.0**	**100.0**
8	75.32	95.37	**99.99**	99.84	99.77
9	**100.0**	99.97	**100.0**	**100.0**	**100.0**
10	99.93	98.31	99.94	99.92	**99.95**
11	99.35	99.36	99.75	99.95	**99.97**
12	99.99	99.99	**100.0**	**100.0**	**100.0**
13	**100.0**	99.72	99.80	99.98	99.97
14	**99.96**	98.57	99.76	99.90	99.88
15	94.59	92.79	**99.94**	99.62	99.76
16	99.81	99.23	**100.0**	99.85	99.84
OA(%)	94.09±0.28	97.82±0.18	**99.93±0.02**	99.87±0.04	99.90±0.03
AA(%)	98.02±0.84	98.86±0.11	99.90±0.03	99.90±0.02	**99.93±0.01**
$\kappa\times100$	93.45±0.31	97.57±0.20	**99.92±0.02**	99.86±0.04	99.89±0.03

FIGURE 8. Classification maps of different methods on Indian Pines data: (a) SVM. (b) DCNN. (c) DFFN. (d) SSRN. (e) PRAN.

99.8% in both OA and AA). Moreover, the SSRN is equipped with deeper architecture than PRAN, which means the computational cost of the former is higher than that of the latter. From another point of view, though the architecture of PRAN is shallower, it does not cause the PRAN to possess worse generalization performance than the SSRN. Not only that, it makes the PRAN easier to train and faster in HSI classification.

Figures 8-10 visualize the classification results of different methods which close to the corresponding mean OA on all three datasets. For all three datasets, there exist many misclassified pixels in the classification maps generated by the SVM and DCNN. And the SVM causes more misclassified pixels, which is consistent with the above quantitative results. The DFFN, SSRN and our proposed methods bring about very little noise in the corresponding classification maps especially for Pavia University and Salinas datasets.

In order to further evaluate the robustness and generalization ability of proposed method, the classification results obtained by proposed method are compared with those obtained by comparison methods under different training set size. Figure 11 displays the OA obtained by different methods on Indian Pines, Pavia University, and Salinas datasets, respectively. Note that the percentages of training samples are reported in Figure 11, and 10% of all samples are used for validation set, the rest samples are used for test set. All results are the average over 10 repeated experiments with randomly selected training samples. As we can see, the accuracies of all methods increase as the numbers of training samples increase. Moreover, the proposed PRAN consistently provides competitive performances over the other compared methods under all different training set size. In particular, the smaller the training set is, the more obvious the superiority of proposed method over all compared methods is.

Furthermore, we select 4/8 training samples per class for each dataset and classify all three datasets with the PRAN, DFFN and SSRN. It should be noted that 4 training samples per class means only 64, 36, and 64 samples are used for

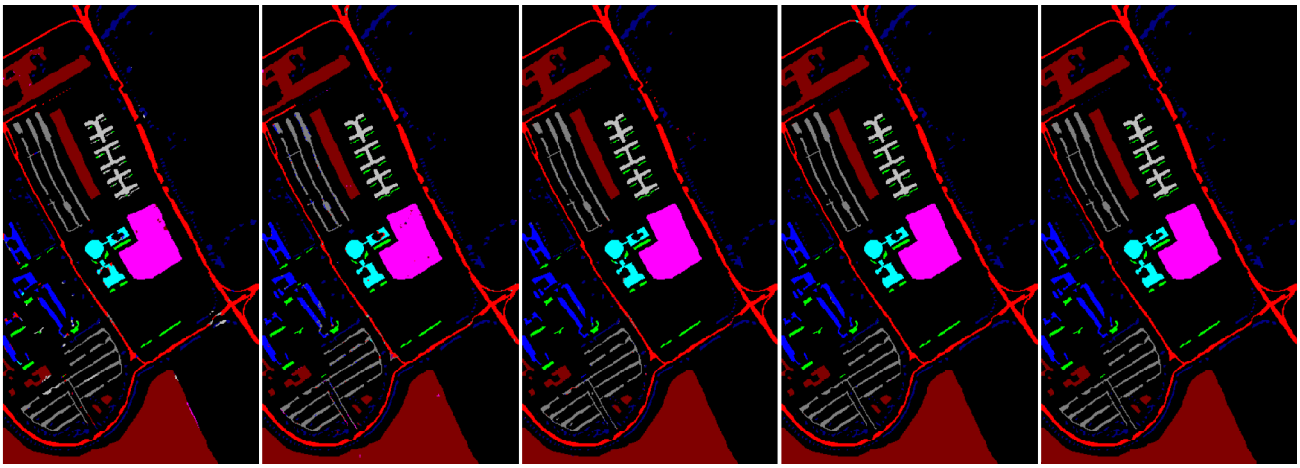


FIGURE 9. Classification maps obtained by different methods for Pavia University data: (a) SVM. (b) DCNN. (c) DFFN. (d) SSRN. (e) PRAN.

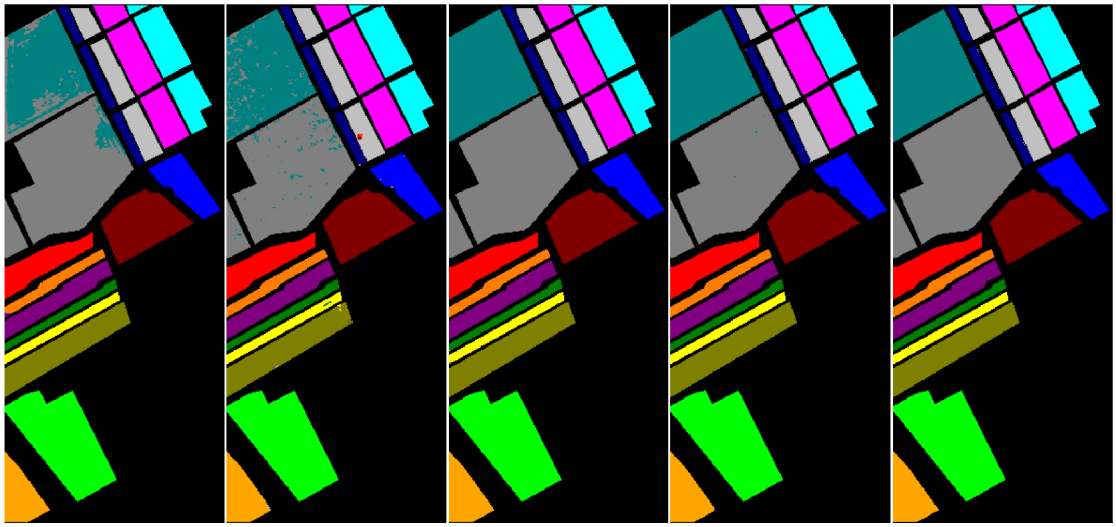


FIGURE 10. Classification maps obtained by different methods for Salinas data: (a) SVM. (b) DCNN. (c) DFFN. (d) SSRN. (e) PRAN.

TABLE 9. Classification results of 4 training samples per class.

Datasets	Indian Pines			Pavia University			Salinas		
Metrics	OA	AA	$\kappa\times100$	OA	AA	$\kappa\times100$	OA	AA	$\kappa\times100$
DFFN	63.57±1.54	54.88±1.37	59.44±1.71	60.97±8.13	60.69±4.52	53.11±9.65	83.24±2.32	87.93±1.43	81.82±2.57
SSRN	61.30±0.99	60.45±0.96	56.33±1.11	61.71±3.13	66.98±3.12	53.52±2.90	82.06±1.40	89.05±1.51	80.18±1.52
PRAN	66.25±1.20	67.19±3.06	62.23±1.16	71.08±5.91	73.70±3.71	63.47±6.49	84.37±1.00	89.65±0.83	82.67±1.06

training for Indian Pines, Pavia University and Salinas datasets, respectively. In other words, the number of training samples is less than 1% of the total number of all labeled samples. In the same way, the percentage of validation set is 10% and the rest samples are adopted to evaluate the model performance. Tables 9 and Table 10 display the corresponding classification results in detail. Here, we take the classification results (OA) of Pavia University as an example. When 4 samples per class are selected to train the network, compared with DFFN and SSRN, the PRAN improves 10.11% and 9.37% in OA (see Table 9), respectively. And in Table 10, when 8 samples per class are used for training, the OA obtained by PRAN

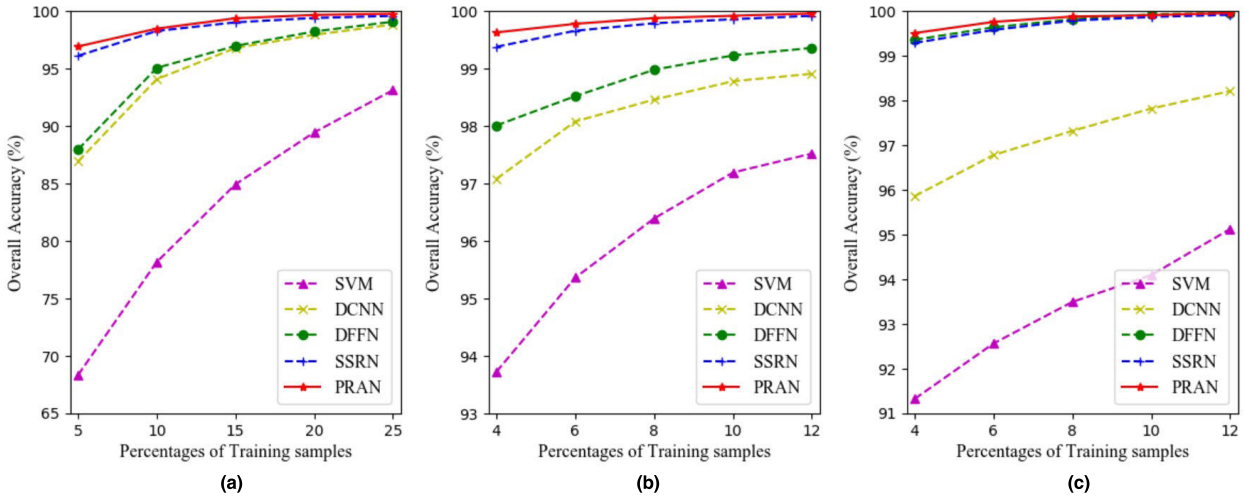


FIGURE 11. Classification results (OA) obtained by different methods under different training set sizes on: (a) Indian Pines image; (b) Pavia University image; (c) Salinas image.

TABLE 10. Classification results of 8 training samples per class.

Datasets	Indian Pines			Pavia University			Salinas		
Metrics	OA	AA	$\kappa\times100$	OA	AA	$\kappa\times100$	OA	AA	$\kappa\times100$
DFFN	72.41±3.13	62.51±1.17	69.13±3.38	72.25±4.66	73.20±3.52	67.06±5.61	87.04±2.32	91.93±1.43	86.82±2.57
SSRN	70.68±1.40	67.75±2.20	66.99±1.68	67.88±2.57	74.03±2.01	60.75±2.69	86.06±0.95	90.30±1.10	84.51±1.05
PRAN	74.81±1.18	76.75±2.59	71.69±1.39	74.55±3.47	78.83±2.19	69.29±3.96	88.28±1.17	92.02±0.85	87.89±1.29

is 2.30% and 6.67% higher than that obtained by DFFN and SSRN, respectively. In the other two datasets, compared with the DFFN and SSRN, the PRAN also obviously improves the classification results of both 4 training samples per class and 8 training samples per class. The above experimental results further demonstrate the superiority of the proposed method under small training sample size.

Table 11 reports the training time (s) of proposed PRAN and comparison methods. As can be observed, both the DCNN and DFFN take less time for training than SSRN and PRAN. It is because the DCNN and DFFN adopt the 2-D convolutional layer as the basic element while the SSRN and PRAN adopt the 3-D convolutional layer as the basic element. Although the computational cost of the DFFN is lower than the PRAN, the DFFN requires input image patches with large size (such as 25×25 for Indian Pines dataset), otherwise the classification will degrade. Larger patch size means more noise may appear in the image patches, thus causes worse classification performance. Therefore, it may face challenges to adopt DFFN for HSI classification, especially when the spatial distribution of land cover is complicated and confused. Fortunately, the PRAN is almost free from this constraint. Despite the classification accuracy of the DFFN is pretty close to PRAN when training set is relatively large, the superiority of PRAN gradually increases as the training samples decreases (see Figure 11). The training time of SSRN is roughly 2 times longer than the PRAN due to its deeper architecture. Therefore, the PRAN is evidently faster than SSRN when used for HSI classification.

TABLE 11. Training time of different methods on all three datasets.

Methods	SVM	DCNN	DFFN	SSRN	PRAN
Indian Pines	43.56	540.8	692.9	2551.3	1135.2
Pavia University	136.1	962.5	1126.8	4380.3	1958.0
Salinas	176.1	1354.7	1446.2	6350.7	2872.6

To sum up this section, in terms of classification accuracy and classification speed, the PRAN is able to provide competitive performance over these compared state-of-the-art methods.

IV. CONCLUSION

In this paper, we propose a pre-activation residual attention network, that incorporates both spectral and spatial information, for hyperspectral image classification. Specifically, different from previous CNN-based HSI classification methods, the proposed method adopts pre-activation mechanism to enhance the generalization performance of the network. Moreover, to extract more robust spectral-spatial features, attention mechanism is introduced to build a pre-activation residual block, which allows the proposed network to adaptively recalibrate channel feature responses and effectively exploit discriminative features. Experimental

results on three benchmark HSI datasets demonstrate that the competitive advantage of proposed method when compared with SVM and several state-of-the-art methods (including DCNN, DFFN and SSRN), especially under small training set.

Despite the superiority of the proposed method, it has large number of parameters needed to be learned due to the use of 3-D convolution kernels, thus results in high computational cost. Therefore, the future research will try to develop new approach, such as replacing 3-D convolution with octave convolution, to decrease the computational cost without degrading the classification accuracy. Furthermore, on account of insufficient training samples in HSIs, we will combine the proposed method with advanced data augmentation technique to further improve classification performance.

ACKNOWLEDGMENT

(Hongmin Gao and Yao Yang are co-first authors.)

REFERENCES

[1] L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li, "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.

[2] B. Luo, C. Yang, J. Chanussot, and L. Zhang, "Crop yield estimation based on unsupervised linear unmixing of multidate hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 162–173, Jan. 2013.

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

[4] A. Ertürk, M.-D. Iordache, and A. Plaza, "Sparse unmixing with dictionary pruning for hyperspectral change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 321–330, Jan. 2017.

[5] B. Tu, S. Huang, L. Fang, G. Zhang, J. Wang, and B. Zheng, "Hyperspectral image classification via weighted joint nearest neighbor and sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4063–4075, Nov. 2018.

[6] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.

[7] B. Waske, S. van der Linden, J. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2880–2889, Jul. 2010.

[8] Z. Wu, Q. Wang, A. Plaza, J. Li, L. Sun, and Z. Wei, "Real-time implementation of the sparse multinomial logistic regression for hyperspectral image classification on GPUs," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1456–1460, Jul. 2015.

[9] M. Khodadadzadeh, J. Li, A. Plaza, and J. M. Bioucas-Dias, "A subspace-based multinomial logistic regression for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2105–2109, Dec. 2014.

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

[11] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[12] J. A. Benediktsson, M. Pesaresi, and K. Amason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[13] H. Yu, L. Gao, J. Li, S. S. Li, B. Zhang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields," *Remote Sens.*, vol. 8, no. 4, p. 355, Apr. 2016.

[14] H. Yu, L. Gao, and B. Zhang, "Union of random subspace-based group sparse representation for hyperspectral imagery classification," *Remote Sens. Lett.*, vol. 9, no. 6, pp. 534–540, Mar. 2018.

[15] L. Gan, J. Xia, P. Du, and Z. Xu, "Dissimilarity-weighted sparse representation for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 1968–1972, Nov. 2017.

[16] Y. Zhou and Y. Wei, "Learning hierarchical spectral–spatial features for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1667–1678, Jul. 2016.

[17] X. Cao, L. Xu, D. Meng, Q. Zhao, and Z. Xu, "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification," *Neurocomputing*, vol. 226, pp. 90–100, Feb. 2017.

[18] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.

[19] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," 2017, *arXiv:1704.06904*. [Online]. Available: https://arxiv.org/abs/1704.06904

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[23] Y. Xu and J. Liu, "Implicitly incorporating morphological information into word embedding," 2017, *arXiv:1701.02481*. [Online]. Available: https://arxiv.org/abs/1701.02481

[24] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[25] Y. Liu, G. Cao, Q. Shen, and M. Siegel, "Hyperspectral classification via deep networks and superpixel segmentation," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3459–3482, Jul. 2015.

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

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

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

[29] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[30] S. Li, X. Zhu, Y. Liu, and J. Bao, "Adaptive spatial-spectral feature learning for hyperspectral image classification," *IEEE Access*, vol. 7, pp. 61534–61547, 2019, doi: 10.1109/ACCESS.2019.2916095.

[31] H. Gao, Y. Yang, S. Lei, C. Li, H. Zhou, and X. Qu, "Multi-branch fusion network for hyperspectral image classification," *Knowl.-Based Syst.*, vol. 167, pp. 11–25, Mar. 2019.

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

[33] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[34] B. Pan, Z. Shi, and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 108–119, Nov. 2018.

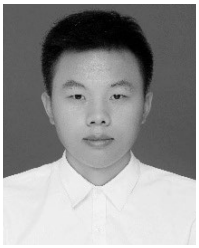
[35] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[36] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[37] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[38] X. Ma, A. Fu, J. Wang, H. Wang, and B. Yin, "Hyperspectral image classification based on deep deconvolution network with skip architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4781–4791, Aug. 2018.
[39] 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, 2017.
[40] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.
[41] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[42] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018, doi: 10.1016/j.isprsjprs.2017.11.021.
[43] 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.
[44] 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.
[45] 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, 2018.
[46] Z. Li, L. Huang, and J. He, "A multiscale deep middle-level feature fusion network for hyperspectral classification," *Remote Sens.*, vol. 11, no. 6, p. 695, 2019.
[47] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7048–7066, Sep. 2019.
[48] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2913372.

HONGMIN GAO received the Ph.D. degree from Hohai University, in 2014. He is currently an Associate Professor with the College of Computer and Information, Hohai University, Nanjing, China. His research interests include deep learning, information fusion, and image processing in remote sensing.



YAO YANG received the B.S. degree in communication engineering from Hohai University, Nanjing, China, in 2018, where he is currently pursuing the degree with the College of Computer and Information. His research interests include deep learning and image processing.

DAN YAO received the B.S. degree in communication engineering from Hohai University, Nanjing, China, in 2018, where he is currently pursuing the degree with the College of Computer and Information. His research interests include machine learning, neural networks, and image processing.

CHENMING LI received the B.S., M.S., and Ph.D. degrees in computer application technology from Hohai University, Nanjing, China, in 1993, 2003, and 2010, respectively. He is currently a Professor and the Deputy Dean of the College of Computer and Information, Hohai University. His current research interests include information processing systems and applications, system modeling and simulation, multisensor systems, and information processing. He is a Senior Member of China Computer Federation and Chinese Institute of Electronic.