HResNetAM: Hierarchical Residual Network With Attention Mechanism for Hyperspectral Image Classification

Zhixiang Xue, Xuchu Yu, Bing Liu, Xiong Tan, and Xiangpo Wei

***Abstract*—This article proposes a novel hierarchical residual network with attention mechanism (HResNetAM) for hyperspectral image (HSI) spectral-spatial classification to improve the performance of conventional deep learning networks. The straightforward convolutional neural network-based models have limitations in exploiting the multiscale spatial and spectral features, and this is the key factor in dealing with the high-dimensional nonlinear characteristics present in HSIs. The proposed hierarchical residual network can extract multiscale spatial and spectral features at a granular level, so the receptive fields range of this network will be increased, which can enhance the feature representation ability of the model. Besides, we utilize the attention mechanism to set adaptive weights for spatial and spectral features of different scales, and this can further improve the discriminative ability of extracted features. Furthermore, the double branch structure is also exploited to extract spectral and spatial features with corresponding convolution kernels in parallel, and the extracted spatial and spectral features of multiple scales are fused for hyperspectral image classification. Four benchmark hyperspectral datasets collected by different sensors and at different acquisition time are employed for classification experiments, and comparative results reveal that the proposed method has competitive advantages in terms of classification performance when compared with other state-of-the-art deep learning models.**

***Index Terms*—Attention mechanism, double branch structure, hierarchical residual network (HResNet), hyperspectral image (HSI), spectral-spatial classification.**

I. Introduction

REMOTE sensing technology is one of the most important components in the field of earth observation (EO), which can perceive and recognize the observed scenes using their different reflection characteristics without making physical contact with the objects. The imaging spectroradiometer can observe the continuous spectrum from visible to short-wave infrared, thus acquired hyperspectral images (HSIs) have hundreds of narrow and approximately continuous spectral bands, and this unique characteristic offers both opportunities and challenges for subsequent information extraction and geoscience applications [1].

Manuscript received December 14, 2020; revised January 14, 2021 and March 7, 2021; accepted March 9, 2021. Date of publication March 17, 2021; date of current version April 7, 2021. *(Corresponding author: Zhixiang Xue.)*

The authors are with the PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China (e-mail: xuegeeker@163.com; xuchu_yu@sina.com; liubing220524@126.com; kjadetx@163.com; 13526635671@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3065987

According to the unique spectral and spatial characteristics, HSI classification aims to determine the ground category of each pixel, which has been widely used in, e.g., environmental monitoring, resource management, urban planning, military, and security applications over the past decade [2].

The intrinsic specificities of HSIs bring several challenges for the classification task, basically, there are three tough problems that need to be solved. 1) The high-dimensional nonlinear characteristic in spectral domain will cause the Hughes phenomenon and affect classification accuracy seriously. 2) The number of annotated samples is often insufficient because labelling samples is expensive and time consuming. 3) Effectively integrating spatial information for spectral-spatial classification to improve pixel-wise classification performance. Aiming to effectively solve above typical problems, lots of classic machine learning models have been exploited for HSI classification [3]. Containing multiple processing layers, deep learning models can learn abstract, intricate, and discriminative features from raw data using backpropagation algorithm, which have brought about striking breakthroughs in many scientific research fields [4]. Deep learning techniques also revolutionize the ways of remote sensing image processing, especially in the HSIs classification field [5]–[8]. According to the feature types employed for classification, HSI classification methods based on deep learning can be generally divided into three categories: Spectral-feature based, spatial-feature based, and spectral-spatial-feature based networks. Due to the fact that both spatial information and spectral information make contributions to HSI classification, the spatial-feature and spectral-spatial-feature-based networks have witnessed more interest in recent years [9].

Because 3D convolutional neural networks (3D-CNNs) can learn spectral–spatial features simultaneously without compressing spectral and spatial information, it is now commonly accepted that 3D-CNNs can be directly utilized for spectral-spatial-feature-based classification without any preprocessing or postprocessing process [10]–[12]. Combining the recurrent network with 3D convolution operators, the recurrent 3D CNN (R-3D-CNN) can exploit both spatial and spectral information for classification [13]. Due to the fact that deeper learning networks can learn more high-level discriminative features, deeper learning models have shown more superiorities in image recognition and classification [14], [15]. But the major problem of very deep networks is the vanishing gradient in the training

process. By introducing identity mapping to the main path of network structure, the residual network (ResNet) framework can ease this training problem in which the underlying error can be propagated through the shortcut [16]. In the contextual deep CNN (CDCNN) initial spectral and spatial information are extracted by multiscale convolutional filter bank, and these joint spatial-spectral features are fed into two residual blocks and fully convolutional network to predict corresponding class label [17]. The spectral-spatial residual network (SSRN) employs spectral and spatial residual blocks to facilitate back propagation of gradients and alleviate the declining-accuracy phenomenon, in which batch normalization is also used to regularize learning process [18]. Aiming to explore the intrinsic complexity of HSI, the deep pyramidal residual networks use pyramidal bottleneck residual blocks to learn high-level spectral–spatial features [19]. To solve the small samples classification of HSI, deep few-shot learning and multiview learning are proposed in the deep residual learning framework recently [20], [21]. Dense network (DenseNet) connects each layer to every other layer in a feed-forward way, which can also alleviate the vanishing-gradient problem [22]. Using densely connected structure in network architecture, the end-to-end fast dense spectral-spatial convolution network (FDSSC) can extract spectral–spatial features for classification, which can lead to extremely accurate classification [23]. Deep and Dense convolutional network introduces two dense blocks to construct deep network and integrate various spectral–spatial features for classification [24].

Visual patterns appear at multiscales in natural scenes. Different objects have different sizes in the same image, and context information of an object may occupy different areas in different images. Therefore, in order to accurately understand objects in image, it is essential to perceive information from different scales. Recently, some CNN-based models try to learn spectral–spatial features of multiple scales for HSI classification. The multilayer fusion dense network (MFDN) uses PCA and 2D dense network to extract spatial features, and the spectral features are extracted by 3D dense blocks, then these features are fused for classification [25]. The CNNs with multiscale convolutions (MS-CNNs) use convolution kernels of different sizes to extract features of different scales, and three types of classification network structures are proposed [26]. The multiscale deep middle-level feature fusion network (MMFN) uses two stages to fuse complementary and related information, the first stage extracts middle-level spectral and spatial features by corresponding scale model, and these middle-scale features are fused using residual blocks in the second stage [27]. The hierarchical multiscale CNN with the auxiliary classifier (HMCNN-AC) extracts multiscale features from image patches of different sizes, and bidirectional long-short-term memory (LSTM) considers these features as sequential data to capture dependence and correlation [28]. In [29], the multiscale residual network (MSRN) utilizes depthwise separable convolution (DSC) to construct multiscale residual block (MRB), and two MRBs are connected by high-level shortcut to aggregate features of different levels.

Inspired by visual perception of the human visual system, the attention mechanism has been employed for HSI classification. In [30], recurrent neural network (RNN) with attention learns the continuous spectrum features, and CNN with attention is designed to extract robust spatial features. Then, the multilayer network uses spectral and spatial features to extract conjoint characteristics. The double-branch multiattention mechanism network (DBMA) and double-branch dual-attention mechanism network (DBDA) use spectral and spatial dense blocks to extract spectral and spatial features, respectively, and the attention modules are utilized to set different weights for extracted features [31], [32]. Aiming to solve the problem that CNNs set the same weight for all spectral bands, the spectral attention module-based convolutional network recalibrates spectral bands so as to strengthen important bands and suppress less useful ones [33]. The end-to-end spectral-spatial squeeze-and-excitation residual bag-of-feature (S3EResBof) model combines the residual block and squeeze-and-excitation block to boost the classification performance, in which batch normalization is also used to regularize the network [34]. In order to suppress the influence of interfering pixels, the spectral-spatial attention network (SSAN) introduces two attention modules to learn more discriminative spectral–spatial features [35]. A series of attention blocks are used in the end-to-end residual spectral–spatial attention network (RSSAN), the first group of attention modules adaptively select spectral bands and spatial pixels, then the second group of attention modules refine the spectral–spatial features, and the residual blocks embedded with attention modules are utilized to optimize the training process [36].

To obtain multiscale representations of objects, feature extractors need to employ different receptive fields to describe objects at different scales [37]. However, the existing CNNs based multiscale extractors can only extract features of fixed receptive fields, which can not extract global and local features at the same time. Current hierarchical features are extracted using the layer wise method, but this method may cause the gradient vanishing phenomenon and need many labeled samples for training. In addition, existing attention-based HSI classification methods only employ single-scale features, which can not make full use of the complex spectral and spatial features of multiple scales. All these factors will affect the HSI classification accuracy to some extent.

Drawing intuition from the success achieved by using the hierarchical residual network (HResNet) to extract multiscale features, the hierarchical residual network with attention mechanism (HResNetAM) is proposed, which not only extracts different scale spectral and spatial features but also employs attention mechanism to promote the discriminative ability of features for HSI classification. Besides, using the residual-like style and batch normalization in the module, the proposed method can also avoid the gradient vanishing problem. Our main contributions in this article can be summarized as follows.

1) First, HResNet block is exploited to extract multiscale spectral and spatial features, and these features can represent the global and local receptive fields of the datasets. And this is the first time to extract spectral and spatial features of multiple scales for HSI classification at a granular level.

2) Second, to take full advantage of the hierarchical spectral and spatial features for classification, the attention mechanism is also employed to adaptively calibrate spectral and spatial features of different scales, which can further promote the discriminability of extracted features for HSI classification.
3) Third, double branch structure for HSI classification is also utilized. In two parallel branches, different sizes of convolution kernels are employed to learn corresponding spectral and spatial features. And the spatial and spectral features of different scales are fused for spectral-spatial classification. In addition, the residual learning and batch normalization can also facilitate the model training.
4) The experimental results, obtained over four benchmark HSI datasets, reveal that the proposed method exhibits potential to learn more discriminative spectral–spatial features, providing competitive performance advantages compared with state-of-the-art deep learning classification models.

The remainder of this article is organized as follows. Section II introduces the proposed HResNet with attention mechanism model in detail. Parameter analysis and comparative HSI classification results are presented in Section III, and Section IV concludes this article.

II. Methodology

The proposed model makes full use of the multiscale feature extraction ability of the HResNet and the weight calibration capability of the attention mechanism. First, drawing intuition from the success achieved by residual network, the hierarchical residual block can not only extract multiscale features from raw data but also avoid the gradient vanishing problem. Then, in order to enhance the discriminative ability of spatial and spectral features with different scales in HSI classification, the spectral attention module and spatial attention module are employed. Finally, the proposed double branch structure which extracts spectral and spatial features separately is described, and detailed model architecture and parameters are also introduced.

A. Residual Learning

Deeper learning models have stronger feature learning and expression capabilities, but the vanishing gradient problem will be exposed in the training process. With the network depth increasing, accuracy get saturated and then degrades rapidly. Unexpectedly, this problem is not caused by overfitting, and adding more layers leads to higher training error. The key idea of residual learning is to introduce identity mapping into the backbone path of network structure. In the training process of deep residual networks, the underlying error can be propagated through the shortcut, which can effectively solve the notorious gradients vanishing problem. The residual learning does not require additional parameters, so it neither adds extra parameter nor increases computational complexity compared with the original network. The deep residual network is composed of many stacked residual units, in which a single residual unit is illustrated in Fig. 1.

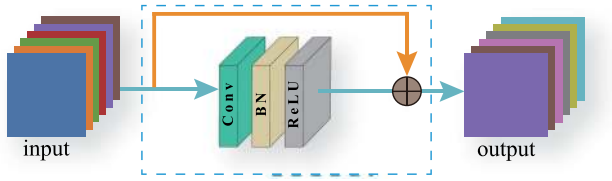


Fig. 1. Illustration of the residual unit.

This residual unit contains one convolutional layer, one batch normalization (BN) layer and one rectified linear unit (ReLU) layer as well as the identity mapping. And the basic form of the residual unit is formulated as

$$\boldsymbol{x}_{i+1} = \boldsymbol{F}(\boldsymbol{x}_i) + \boldsymbol{x}_i \tag{1}$$

in which $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$ are corresponding input and output of the unit, and $\boldsymbol{F}$ refers to the residual function. In order to train the model more efficiently, the batch normalization is implemented after every convolutional layer [38]. Moreover, the rectified linear unit layer is also utilized to extract nonlinear features. Through this skip connection strategy, the residual networks can build very deep network structures without worrying about the gradients vanishing problem. The deep residual networks have been exploited for HSI classification, which can obtain superior classification accuracy than the CNN-based methods [18], [19], [29], [39].

B. Hierarchical Residual Learning

It is critically essential to extract multiscale features for image classification task. Most existing CNNs enhance multiscale representation strength via layer-wise way, while the multiscale representation ability of HResNet refers to the multiple available receptive fields at a granular level. To achieve this goal, the hierarchical residual block divides the input feature maps into several groups, and each subgroup of feature maps is performed with different layers of convolution operators. In the hierarchical residual block, different subgroups of feature maps have different receptive fields, thus the combined feature maps can represent multiscale features, so it can increase the receptive fields of the network [40]. Existing convolutional networks obtain multiscale features by stacking convolutional layers, but these features have relatively fixed receptive fields. The hierarchical residual learning introduces a new scale dimension as an essential factor except existing dimensions of depth, width, and cardinality [41]. In HResNet, the scale dimension means the number of feature groups in a hierarchical residual unit. Fig. 2 shows the hierarchical residual unit with 3 scales, in which $\ominus$ and $\oplus$ mean split operation and concatenation operation, respectively.

We denote input and output of the hierarchical residual unit with $\boldsymbol{x}$ and $\boldsymbol{y}$. First, we split the input feature map $\boldsymbol{x}$ into s feature subsets, and every subset is represented as $\boldsymbol{x}_i$, where $i \in \{1, 2, \ldots, s\}$. The subset $\boldsymbol{x}_i$ has the same spatial size with input $\boldsymbol{x}$, but only $1/s$ channels. Except for $\boldsymbol{x}_1$, every $\boldsymbol{x}_i$ has corresponding convolution operator, denoted by $\boldsymbol{K}_i()$. And we use $\boldsymbol{y}_i$ to denote the output of $\boldsymbol{K}_i()$. To obtain hierarchical

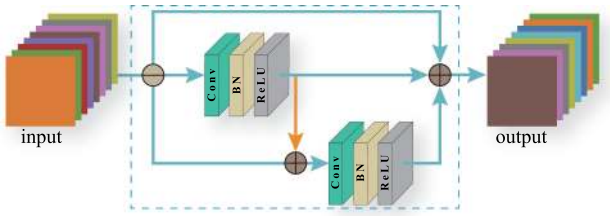


Fig. 2. Illustration of the hierarchical residual unit.

features, we add the output of $\boldsymbol{K}_{i-1}()$ to the feature subset $\boldsymbol{x}_i$, and then they are fed into $\boldsymbol{K}_i()$. Thus, $\boldsymbol{y}_i$ can be generally written as follows:

$$\boldsymbol{y}_i = \begin{cases} \boldsymbol{x}_i & i = 1; \\ \boldsymbol{K}_i(\boldsymbol{x}_i) & i = 2; \\ \boldsymbol{K}_i(\boldsymbol{x}_i + \boldsymbol{y}_{i-1}) & 2 < i \leq s. \end{cases} \tag{2}$$

Through this hierarchical residual structure, each convolution operator $\boldsymbol{K}_i()$ can receive information from subsets $\boldsymbol{x}_j (j \leq i)$, thus the feature split $\boldsymbol{x}_i$ has a larger receptive field than $\boldsymbol{x}_j$. The concatenation operation at the end of hierarchical residual unit combines feature maps of different receptive fields. In addition, the split and concatenation strategy can force the hierarchical residual block process features more efficiently. In the hierarchical residual unit, larger scale factor s allows the unit to learn features with richer receptive field sizes. We also conduct batch normalization and rectified linear unit activation function after every convolutional layer to train the HResNet more effectively. Therefore, the residual-like connections within the hierarchical residual unit could make it capture global and local features at a granular level.

C. Attention Mechanism

Drawing intuition from the human visual system, the attention mechanism can recalibrate channel-wise features by explicitly establishing the relationships between channels [42]. The traditional HSI classification models assign equivalent weights to all pixels and bands in the spatial and spectral domains, respectively. It is a fact that different spatial pixels and spectral bands make unequal discriminative contributions to classification results. For instance, several edge pixels in the HSI block have different labels with the center pixel, and these interfering pixels will weaken the discriminative ability of spectral–spatial features, thereby affecting the classification accuracy. If the weight of these pixels can be suppressed, the discriminability of the spectral–spatial features will be increased. Thus, it is feasible to introduce the attention mechanism to HSI classification, which can focus more on the discriminative and effective spatial and spectral features and weaken information detrimental to classification. Because exploiting spectral and spatial-wise attention is superior to only using channel-wise attention [43], so we adopt spectral attention module as well as spatial attention module simultaneously to recalibrate spectral and spatial features of multiple scales. Two attention modules are introduced in detail as follows.

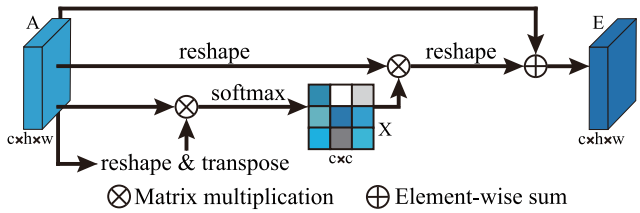


Fig. 3. Structure of the spectral attention module.

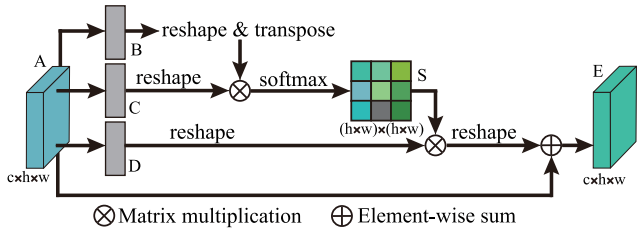


Fig. 4. Structure of the spatial attention module.

1) Spectral Attention Module: The spectral attention module is constructed by modeling the interdependencies between channels, as shown in Fig. 3. The spectral attention map $\boldsymbol{X} \in \mathbb{R}^{c \times c}$ is calculated from the initial input $\boldsymbol{A} \in \mathbb{R}^{c \times h \times w}$, in which $h \times w$ represents the spatial size while c denotes the channels of the original features. Specifically, we first reshape and transpose $\boldsymbol{A} \in \mathbb{R}^{c \times h \times w}$ into $\boldsymbol{A}^T \in \mathbb{R}^{c \times n}$, and conduct a matrix multiplication between $\boldsymbol{A}$ and $\boldsymbol{A}^T$. And the results are fed into a softmax layer to get the attention map $\boldsymbol{X}$

$$x_{ji} = \frac{\exp(A_i \times A_j)}{\sum_{i=1}^{c} \exp(A_i \times A_j)} \tag{3}$$

in which x_{ji} represents the influence of ith channel on the jth channel. In addition, a matrix multiplication is conducted between $\boldsymbol{X}^T$ and $\boldsymbol{A}$, and their results are reshaped into $\mathbb{R}^{c \times h \times w}$. Finally, a scale parameter α is used to weight the results and perform a element-wise sum operator with the input A to obtain the spectral attention map $\boldsymbol{E} \in \mathbb{R}^{c \times h \times w}$

$$E_j = \alpha \sum_{i=1}^{c} (x_{ji} A_j) + A_j \tag{4}$$

where the parameter α is initialized to be 0 and can be optimized gradually in the training process. We can see that the spectral attention feature map E is a weighted combination of all the original channels, which can selectively strengthen informative channels and suppress less useful ones. Therefore, the spectral feature discriminability can be increased through this spectral attention module.

2) Spatial Attention Module: Fig. 4 shows the spatial attention module, the initial input $\boldsymbol{A} \in \mathbb{R}^{c \times h \times w}$ is fed into two different convolution layers to generate two new feature maps $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively, in which $\{B, C\} \in \mathbb{R}^{c \times h \times w}$. And these two feature maps are reshaped into $\mathbb{R}^{c \times n}$, where $n = h \times w$ refers to the number of spatial pixels. Then a matrix multiplication between $\boldsymbol{B}^T$ and $\boldsymbol{C}$ is performed, and the results are fed into a softmax layer to obtain spatial attention map $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ as

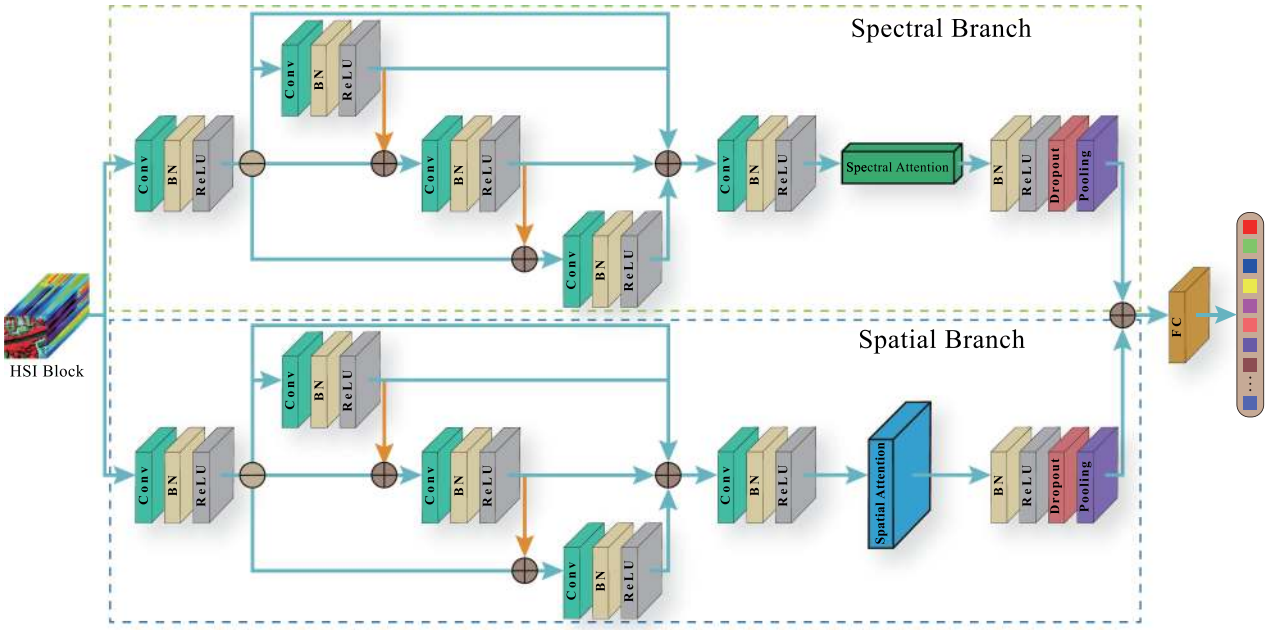


Fig. 5. Framework of our proposed HResNet with attention mechanism (the representative HResNetAM model has 4 scales and 6 kernels).

follows:

$$s_{ji} = \frac{\exp(B_i \times C_j)}{\sum_{i=1}^{n} \exp(B_i \times C_j)} \tag{5}$$

in which s_{ji} measures the ith pixel's influence on the jth pixel. The closer the spatial distance between two pixels, the greater the correlation between them.

A new feature map $\boldsymbol{D} \in \mathbb{R}^{c \times h \times w}$ is also generated from initial input feature $\boldsymbol{A}$ through a convolution layer and reshaped into $\mathbb{R}^{c \times n}$ subsequently. Then a matrix multiplication between $\boldsymbol{D}$ and $\boldsymbol{S}^T$ is performed, and the results are reshaped into $\mathbb{R}^{c \times h \times w}$. Finally, a scale parameter β is utilized to weight the results and perform a element-wise sum operator with the initial input A to get spatial attention map $\boldsymbol{E} \in \mathbb{R}^{c \times h \times w}$ as follows:

$$E_j = \beta \sum_{i=1}^{n} (s_{ji} D_j) + A_j \tag{6}$$

in which the parameter β is initialized to be 0 and can be optimized gradually in the training process. It can be inferred that each position in the spatial attention feature map $\boldsymbol{E}$ is a weighted combination of all the original pixels, which have a global view and selectively emphasize informative positions. Thus, the feature discriminability will be improved in the spatial domain.

D. Framework of the Proposed Model

The whole structure of the HResNetAM model is illustrated in Fig. 5. In order to make the most of the spectral and spatial features of different scales, we adopt the double branch architecture for HSI classification. The upper spectral branch consists of the hierarchical spectral residual network and corresponding spectral attention module. The HResNet containing spectral-based convolution operators are utilized to extract hierarchical spectral features, and the spectral attention module is employed to assign different weights for hierarchical spectral features. The lower spatial branch is composed of hierarchical spatial residual network and corresponding attention block. For the similar purpose, the spatial attention module can recalibrate the spatial features of different scales. The adaptively weighted multiscale spectral and spatial features are fused to conduct the HSI spectral-spatial classification.

TABLE I
DETAILED PARAMETERS OF THE HRESNETAM MODEL

	Layer Name	Output Shape	Filter Size	Padding
Spectral	Input	7×7×102	N/A	N
	Conv1	7×7×49(24)	1×1×5(24)	N
	Conv21	7×7×49(6)	1×1×5(6)	0×0×2
	Conv22	7×7×49(12)	1×1×5(12)	0×0×2
	Conv23	7×7×49(12)	1×1×5(12)	0×0×2
	Conv3	7×7×1(24)	1×1×49(24)	N
Spatial	Input	7×7×102	N/A	N
	Conv1	7×7×1(24)	1×1×102(24)	N
	Conv21	7×7×1(6)	3×3×1(6)	1×1×0
	Conv22	7×7×1(12)	3×3×1(12)	1×1×0
	Conv23	7×7×1(12)	3×3×1(12)	1×1×0
	Conv3	7×7×1(24)	1×1×1(24)	1×1×0
Fusion	Linear	48	N/A	N
	Output	9	N/A	N

The model in Fig. 5 is a HResNetAM network with 4 scales and 6 kernels, and the corresponding detailed parameters of the spatial and spectral feature extraction network in HResNetAM are listed in Table I. In this representative model, we employ the Pavia Centre dataset and the spatial size is set to be 7. And the HSI block serves as the input of the two branch structures. In our proposed model, we employ the convolution kernels with (1, 1,

TABLE II
LAND-COVER CLASSES AND SAMPLES OF THE PAVIA CENTRE DATASET

Class No.	Class Name	Train	Test	Total
1	Water	140	65831	65971
2	Trees	140	7458	7598
3	Meadows	140	2950	3090
4	Bricks	140	2545	2685
5	Soil	140	6444	6584
6	Asphalt	140	9108	9248
7	Bitumen	140	7147	7287
8	Tiles	140	42686	42826
9	Shadow	140	2723	2863
	Total	1260	146892	148152

5) and (3, 3, 1) to extract spectral and spatial features, respectively. Note that the stride of Conv1 in the spectral branch is (1, 1, 2) and stride for other convolution operations in HResNetAM is (1, 1, 1).

III. EXPERIMENTAL RESULTS AND ANALYSIS

In our experiments, all the comparative classification experiments are carried out on a workstation equipped with an Intel Core i9-7900X, an Nvidia Geforce RTX 2080 Ti GPU, and 128 G RAM. The proposed HResNetAM model is implemented using the PyTorch library with Python language. We employ main classification evaluation coefficients, namely, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ) to quantitatively assess the classification performance. And we also exploit classification maps to qualitatively evaluate the experimental results. In order to increase the reliability and credibility of experimental results, we conducted ten trials for each classification experiment with randomly selecting the training samples.

A. Data Description

Four different benchmark hyperspectral datasets collected by different sensors and at different time are utilized to conduct the HSI classification experiments.

Pavia Centre: The Pavia Centre dataset was acquired by the ROSIS sensor over the side of Ticino river, Pavia, northern Italy. The spatial size of this dataset is 1096 $\times$ 715 pixels, and corresponding geometric resolution is 1.3 m. This sensor can acquire 115 bands in total in the wavelength range of 0.43–0.86 μm. After removing the greatly noise-affected channels, the remaining 102 spectral bands are employed for experiments. The corresponding image ground truth differentiates 9 classes, and the detailed land-cover classes, training samples, and test samples are shown in Table II.

Houston 2013: The Houston 2013 dataset was collected by the ITRES CASI-1500 sensor over the University of Houston campus in June 2012, which is provided by the 2013 IEEE GRSS Data Fusion Competition [44]. The spatial size of this image dataset is 349 $\times$ 1905 pixels, and the spatial resolution is 2.5 m. This dataset has 144 spectral bands in the wavelength range of 0.38–1.05 μm. There are 15 land-cover classes within the image coverage, and the detailed land-cover classes, training samples, and test samples are shown in Table III.

TABLE III
LAND-COVER CLASSES AND SAMPLES OF THE HOUSTON 2013 DATASET

Class No.	Class Name	Train	Test	Total
1	Healthy grass	180	1071	1251
2	Stressed grass	180	1074	1254
3	Synthetic grass	180	517	697
4	Trees	180	1064	1244
5	Soil	180	1062	1242
6	Water	100	225	325
7	Residential	180	1088	1268
8	Commercial	180	1064	1244
9	Road	180	1072	1252
10	Highway	180	1047	1227
11	Railway	180	1055	1235
12	Parking lot 1	180	1053	1233
13	Parking lot 2	180	289	469
14	Tennis court	100	328	428
15	Running track	180	480	660
	Total	2540	12489	15029

TABLE IV
LAND-COVER CLASSES AND SAMPLES OF THE DIONI DATASET

Class No.	Class Name	Train	Test	Total
1	Dense Urban Fabric	140	1122	1262
2	Mineral Extraction Sites	100	104	204
3	Non-Irrigated Arable Land	140	474	614
4	Fruit Trees	50	100	150
5	Olive Groves	140	1628	1768
6	Coniferous Forest	140	221	361
7	Dense Sclerophyllous Vegetation	140	4895	5035
8	Sparse Sclerophyllous Vegetation	140	6234	6374
9	Sparsely Vegetated Areas	140	1614	1754
10	Rocks and Sand	140	352	492
11	Water	140	1472	1612
12	Coastal Water	140	258	398
	Total	1550	18474	20024

Dioni: The Dioni dataset is one of the HyRANK benchmark datasets which have been developed in the framework of the ISPRS Scientific Initiatives [45]. The HyRANK benchmark datasets contain two training images (i.e., Dioni and Loukia) along with the corresponding ground truth and two validation images. The spatial size of the Dioni dataset is 250 $\times$ 1376 pixels, which contains 176 spectral channels. There are 16 different land cover classes in the HyRANK benchmark datasets, and the selected Dioni dataset covers 12 classes. The detailed number of training samples and test samples along with the corresponding labels is reported in Table IV.

Houston 2018: The Houston 2018 dataset was gathered by the ITRES CASI-1500 sensor over the University of Houston campus in February 2017, which is provided by the 2018 IEEE GRSS Data Fusion Competition [46]. We only use the training portion of the whole HSI, and the ground truth is resampled to adapt the hyperspectral dataset [47]. The spatial size of this dataset is 601 $\times$ 2384 pixels at 1-m ground sampling distance. There are 48 spectral bands in the wavelength range of 0.38–1.05 μm. And there are 20 urban land-cover classes within image coverage. The detailed number of training samples as well as test samples with corresponding labels is shown in Table V.

TABLE V
LAND-COVER CLASSES AND SAMPLES OF THE HOUSTON 2018 DATASET

Class No.	Class Name	Train	Test	Total
1	Healthy grass	180	9619	9799
2	Stressed grass	180	32322	32502
3	Artificial turf	180	504	684
4	Evergreen trees	180	13408	13588
5	Deciduous trees	180	4868	5048
6	Bare earth	180	4336	4516
7	Water	100	166	266
8	Residential buildings	180	39582	39762
9	Non-residential buildings	180	223504	223684
10	Roads	180	45630	45810
11	Sidewalks	180	33822	34002
12	Crosswalks	180	1336	1516
13	Major thoroughfares	180	46178	46358
14	Highways	180	9669	9849
15	Railways	180	6757	6937
16	Paved parking lots	180	11295	11475
17	Unpaved parking lots	50	99	149
18	Cars	180	6398	6578
19	Trains	180	5185	5365
20	Stadium seats	180	6644	6824
	Total	3390	501322	504712

B. Experimental Setup

To evaluate the performance of proposed HResNetAM, we use several state-of-the-art methods for comparative experiments. These models include the deep learning-based models (i.e., 3DCNN, CDCNN, SSRN, FDSSC, DBMA, and DBDA) as well as the SVM with radial basis function (RBF) kernel. In order to carry out comparative experiments more fairly, we use the same number of training samples in all methods, and 20% of the training samples are set as validation samples. Specifically, the parameters of each method are given separately according to the corresponding articles.

SVM: For SVM with RBF kernel, we employ the cross validation strategy to get the optimal regularization parameter C and kernel parameter γ in the range of $C = \{2^{-2}, 2^{-1}, \ldots, 2^{7}\}$ and $\gamma = \{2^{-2}, 2^{-1}, \ldots, 2^{7}\}$, respectively. And we utilize all spectral bands as input of SVM [48].

3DCNN: This method directly uses 3D convolution operators to extract features of HSI, the architecture of the 3DCNN in [11] contains two convolution layers and the fully connected layer. This model uses 3D image cube as input, and the input size is $5\times5\times$B, where B refers to the spectral bands.

CDCNN: The contextual deep CNN network constructs deeper classification model with residual learning structure, which is composed of the multiscale filter bank and two residual blocks. Then three convolutional layers and one fully connected layer are utilized for HSI classification [17]. The input of CDCNN is $5 \times 5 \times$ B block.

FDSSC: The FDSSC is based on 3D-CNN and dense block, and this model contains two dense blocks and followed by the average pooling, flatten and fully connected layer [23]. And we also use the $9 \times 9 \times$ B image block as input.

SSRN: The SSRN combines residual learning and 3D-CNN, which extracts spectral and spatial features in sequence using corresponding residual blocks, and the following average pooling layer and an fully connected layer are employed for classification. This method also uses $7\times7\times$B image block as input [18].

DAMA: The DAMA is based on attention mechanism and dense block, which contains spectral branch and spatial branch as well as corresponding attention blocks. The convolutions with (1, 1, 7) and (7, 7, 1) kernels are utilized in spectral and spatial branches, respectively, and the size of the input is $7\times7\times$B [31].

DBDA: The architecture of the DBDA is presented in [32], which also contains spectral and spatial dense blocks and corresponding attention blocks. And we use $7\times7\times$B image block as input.

The DBMA [31] and DBDA [32] models utilize dense network to extract spectral and spatial features, and attention modules are employed to recalibrate extracted features. These two methods are used as comparative methods to verify the feature extraction capability of HResNet. In order to conduct the ablation study of attention mechanism, we also design the HResNet model as one comparative method, which has the same network structure with corresponding HResNetAM but without spectral and spatial attention modules.

C. Parameters Analysis and Setting

The parameters in deep learning models can influence the HSI classification to some extent, so we evaluate the main parameters in our proposed model, they are learning rate, spatial size, the number of training samples, as well as the number of scales and kernels. And in our classification experiments, because the HResNetAM with different batch sizes and epoches has relatively stable classification accuracies, so we set the batch size and epochs as 32 and 200, respectively.

1) Learning Rate: The learning rate greatly influences the convergence rate of the network and the HSI classification performance. Referring to the relevant experiments, we analyze the effect of learning rate at $\{0.0001, 0.0002, 0.0003, 0.0008\ 0.001, 0.005, 0.01\}$ on overall accuracies. Fig. 6 shows the ten experimental results on four datasets with different learning rates. In this figure, two independent horizontal lines represent the overall range of the classification results, and the two edges of the box denote upper quartile and lower quartile, respectively. The horizontal line in box refers to median value, and the ♦ denotes abnormal outliers. It can be found that a smaller learning rate has a relatively stable classification accuracy and bigger learning rate will result in larger variance in the classification accuracy. According to the average OA and variance in four groups of HSI classification, we set the learning rate to be 0.0002, 0.0001, 0.0002, and 0.0005 for four benchmark datasets, respectively.

2) Spatial Size: For the purpose of utilizing the spatial information for spectral-spatial classification, we exploit the 3D image cube as input. The spatial size can also influence the HSI classification results, and we set neighborhood size in the range of $\{3, 5, 7, 9, 11, 13\}$. Table VI shows the average overall accuracy and corresponding variance of the proposed method on four hyperspectral datasets with different spatial sizes. Based on

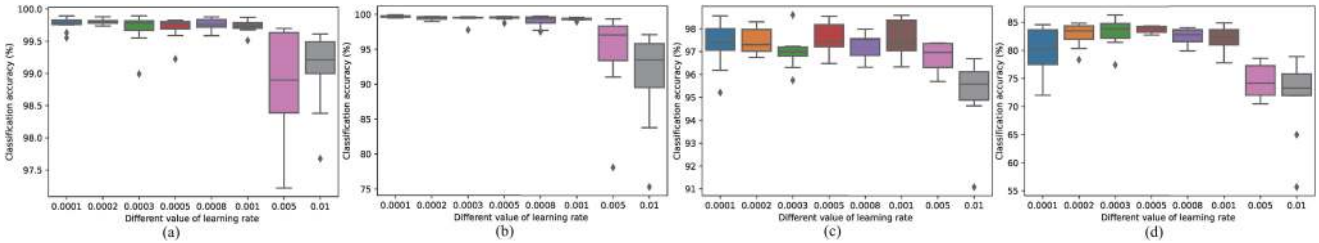


Fig. 6. Box plot of OAs with different learning rates on four different datasets. (a) Pavia Centre dataset. (b) Houston 2013 dataset. (c) Dioni dataset. (d) Houston 2018 dataset.

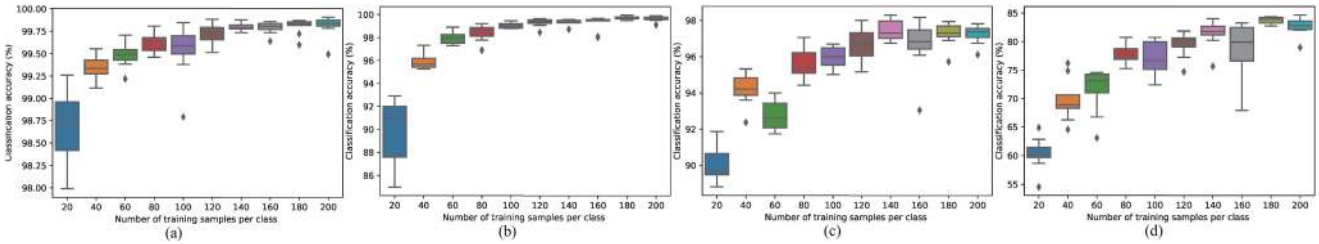


Fig. 7. Box plot of OAs with different number of training samples on four different datasets (a) Pavia Centre dataset. (b) Houston 2013 dataset. (c) Dioni dataset. (d) Houston 2018 dataset.

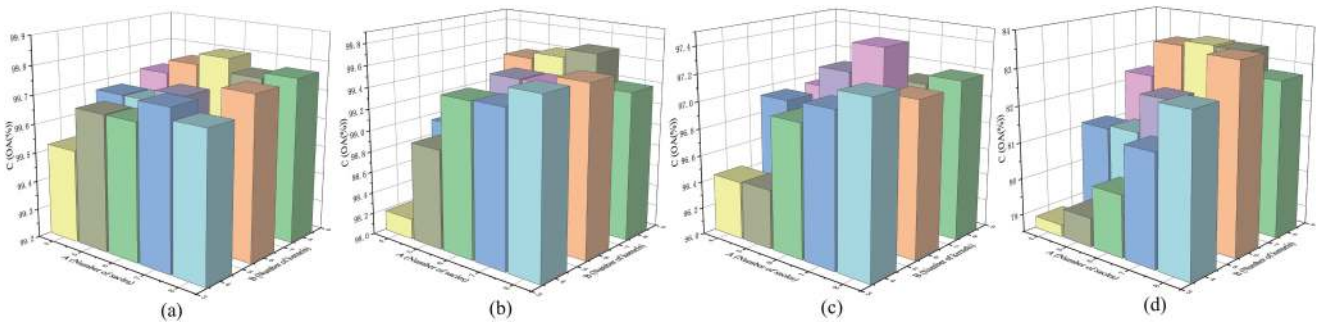


Fig. 8. Bar chart of OAs with different number of scales and kernels on four different datasets. (a) Pavia Centre dataset. (b) Houston 2013 dataset. (c) Dioni dataset. (d) Houston 2018 dataset.

TABLE VI
OVERALL ACCURACY(%) OF THE HRESNETAM METHOD WITH DIFFERENT SPATIAL SIZE ON FOUR DIFFERENT DATASETS

Spatial Size	Pavia Centre	Houston 2013	Dioni	Houston 2018
3×3	99.59 ±0.10	99.26 ±0.12	94.63 ±0.72	75.49 ±1.51
5×5	**99.80** **±0.04**	99.52 ±0.25	96.78 ±0.52	79.62 ±0.94
7×7	99.80 ±0.09	**99.69** **±0.13**	**97.45** **±0.53**	81.23 ±0.14
9×9	99.77 ±0.15	99.47 ±0.24	96.78 ±0.89	83.06 ±1.25
11×11	99.73 ±0.09	97.08 ±1.17	97.15 ±0.77	**83.61** **±0.78**
13×13	99.58 ±0.14	96.41 ±1.03	96.82 ±0.63	83.03 ±2.03

the experimental results, we find that the classification accuracy generally increases and then decreases as the neighborhood increases. Thus the optimal neighborhood sizes of the four datasets are set to be 5, 7, 7, and 11, respectively.

3) Training Samples: The number of training samples also have great influence on HSI classification performance. For the purpose of evaluating the robustness and generalization of HResNetAM model toward different numbers of training samples, we randomly choose $\{20, 40, 60, 80, 100, 120, 140, 160, 180, 200\}$ annotated samples per class for four datasets. Fig. 7 shows the average overall accuracy achieved by HResNetAM with different numbers of training samples on four hyperspectral datasets. From this figure, we can observe that the classification accuracy will quickly reach a relatively stable level with the increase of training samples. According to the overall accuracies of each dataset, we set 140, 180, 140, and 180 per class for training in four datasets, respectively. Since there are fewer labeled samples in some images, we set specific numbers of training samples for those datasets. In the Houston 2013 dataset, we use 100 labeled samples for Water and Tennis Court classes. In the Dioni dataset, we use 100 labeled samples for Mineral Extraction Sites class and 50 labeled samples for Fruit Trees class. And in the Houston

TABLE VII
OA, AA, KAPPA, AND CLASS-SPECIFIC ACCURACY(%) OF DIFFERENT METHODS FOR THE PAVIA CENTRE DATASET (BOLD VALUES REPRESENT THE BEST RESULTS IN THE CORRESPONDING ROWS)

Class No.	SVM [48]	3DCNN [11]	CDCNN [17]	FDSSC [23]	SSRN [18]	DBMA [31]	DBDA [32]	HResNet	HResNetAM
1	99.97	92.91	99.96	99.99	99.99	99.98	99.99	99.99	**100**
2	96.95	97.48	98.55	**99.94**	99.53	98.90	99.34	99.90	99.84
3	76.53	83.36	85.05	92.87	95.46	94.22	92.11	90.64	**96.61**
4	59.11	73.60	86.03	98.17	97.25	97.99	98.63	98.17	**98.99**
5	91.24	93.87	96.73	99.89	99.27	99.56	99.63	99.93	**99.97**
6	86.47	93.15	95.53	**99.08**	98.60	98.73	98.88	98.78	98.64
7	90.72	95.24	96.85	96.56	99.45	99.60	99.86	**99.92**	99.81
8	99.87	99.92	99.94	99.99	99.99	99.96	99.96	99.99	**100**
9	**99.84**	77.59	96.37	99.12	99.55	92.76	94.41	99.57	99.78
OA	96.37 ±0.59	92.97 ±1.37	98.59 ±0.38	99.47 ±0.72	99.67 ±0.17	99.50 ±0.12	99.54 ±0.28	99.59 ±0.49	**99.80 ±0.04**
AA	88.97 ±1.28	89.68 ±0.70	95.00 ±1.53	98.40 ±1.31	98.79 ±0.71	99.50 ±0.11	98.09 ±1.27	98.54 ±1.48	**99.29 ±0.19**
κ	94.88 ±0.83	91.48 ±1.51	97.99 ±0.54	99.24 ±1.02	99.53 ±0.25	99.29 ±0.17	99.35 ±0.41	99.42 ±0.70	**99.72 ±0.06**

TABLE VIII
OA, AA, KAPPA, AND CLASS-SPECIFIC ACCURACY(%) OF DIFFERENT METHODS FOR THE HOUSTON 2013 DATASET (BOLD VALUES REPRESENT THE BEST RESULTS IN THE CORRESPONDING ROWS)

Class No.	SVM [48]	3DCNN [11]	CDCNN [17]	FDSSC [23]	SSRN [18]	DBMA [31]	DBDA [32]	HResNet	HResNetAM
1	95.55	89.80	93.38	99.47	98.25	97.66	99.21	99.20	**99.73**
2	96.98	94.55	88.18	98.00	98.75	98.57	99.90	**99.97**	99.72
3	99.13	71.57	91.65	**100**	99.97	99.95	**100**	**100**	**100**
4	98.14	97.00	97.84	99.83	98.97	98.53	99.95	99.93	**100**
5	95.79	98.01	97.30	99.68	99.85	98.55	99.87	99.95	**99.97**
6	98.49	92.86	96.92	**100**	**100**	99.19	99.91	**100**	99.81
7	87.41	75.35	86.74	**99.71**	98.21	98.49	98.93	99.37	99.58
8	85.49	83.89	92.21	99.50	98.40	99.31	99.60	99.59	**99.78**
9	79.75	85.05	88.64	98.94	98.69	97.20	99.05	98.95	**99.25**
10	85.82	73.09	78.01	99.37	98.19	97.34	98.70	98.60	**99.54**
11	84.40	74.60	84.85	99.39	99.08	99.00	99.39	99.49	**99.54**
12	82.65	75.14	83.82	99.29	97.42	98.07	98.78	98.40	**99.62**
13	60.27	73.42	93.45	98.03	97.59	96.85	96.92	**99.78**	99.18
14	95.41	95.65	96.50	**100**	**100**	99.20	**100**	**100**	**100**
15	99.32	98.91	96.49	99.74	**99.76**	99.21	99.71	**99.76**	**99.76**
OA	89.52 0.87	82.06 4.01	88.72 3.45	99.34 0.26	98.64 0.87	98.33 0.63	99.33 0.25	99.40 0.65	**99.69 0.13**
AA	89.64 1.03	85.26 2.57	91.07 2.22	99.39 0.22	98.88 0.69	98.33 0.63	99.33 0.34	99.53 0.49	**99.70 0.15**
κ	88.66 0.95	80.56 4.30	87.76 3.74	99.28 0.28	98.52 0.95	98.19 0.68	99.28 0.28	99.34 0.70	**99.66 0.14**

TABLE IX
OA, AA, KAPPA, AND CLASS-SPECIFIC ACCURACY(%) OF DIFFERENT METHODS FOR THE DIONI DATASET (BOLD VALUES REPRESENT THE BEST RESULTS IN THE CORRESPONDING ROWS)

Class No.	SVM [48]	3DCNN [11]	CDCNN [17]	FDSSC [23]	SSRN [18]	DBMA [31]	DBDA [32]	HResNet	HResNetAM
1	66.32	71.30	83.66	92.55	91.91	87.52	92.03	**95.48**	95.13
2	76.91	74.48	83.13	86.86	96.27	97.29	96.42	**99.50**	99.04
3	79.41	75.25	82.01	92.81	95.56	91.26	94.62	**97.37**	97.06
4	39.69	8.10	13.59	35.41	37.71	43.58	**58.60**	24.36	54.13
5	87.63	90.02	94.39	97.60	97.72	98.03	97.87	**98.23**	**98.23**
6	85.62	77.13	84.05	93.11	92.17	91.89	**97.12**	93.26	94.56
7	90.88	92.78	93.85	96.02	94.59	95.90	97.07	97.05	**97.28**
8	89.00	93.06	93.74	97.23	97.05	97.57	**97.82**	96.42	97.75
9	77.46	73.66	85.08	92.80	92.15	94.96	96.10	94.94	**96.41**
10	89.10	80.10	91.32	97.84	**98.91**	98.61	95.42	97.80	98.68
11	99.66	99.80	96.94	**100**	99.95	99.98	**100**	**100**	**100**
12	93.59	96.21	87.21	**100**	**100**	98.50	99.91	**100**	**100**
OA	86.24 0.84	87.06 3.39	90.80 2.84	95.73 0.98	95.59 0.87	95.94 1.48	97.04 0.60	96.65 0.43	**97.45 0.53**
AA	81.27 1.74	77.66 5.80	82.41 4.43	90.18 3.41	91.17 2.67	95.94 1.48	93.58 2.28	91.20 1.01	**94.02 2.34**
κ	82.98 1.00	83.82 4.15	88.39 3.56	94.59 1.23	94.42 1.09	94.86 1.86	96.25 0.76	95.75 0.55	**96.76 0.68**

TABLE X
OA, AA, KAPPA, AND CLASS-SPECIFIC ACCURACY(%) OF DIFFERENT METHODS FOR THE HOUSTON 2018 DATASET (BOLD VALUES REPRESENT THE BEST RESULTS IN THE CORRESPONDING ROWS)

Class No.	SVM [48]	3DCNN [11]	CDCNN [17]	FDSSC [23]	SSRN [18]	DBMA [31]	DBDA [32]	HResNet	HResNetAM
1	62.45	63.40	**83.16**	75.75	74.64	74.65	77.21	71.05	72.27
2	84.24	**93.39**	88.78	91.32	93.14	93.05	92.45	92.13	91.94
3	85.19	62.41	98.14	90.55	84.91	89.42	**99.18**	94.07	92.93
4	79.20	85.57	79.26	85.26	87.06	87.80	**89.79**	86.86	86.71
5	21.62	37.21	31.38	46.20	54.68	48.24	53.49	51.41	**57.32**
6	24.93	52.75	57.04	81.85	82.49	83.66	91.35	**92.85**	91.86
7	61.89	25.87	60.01	**98.52**	81.83	78.10	87.83	88.85	88.75
8	52.81	50.35	52.69	60.73	63.67	64.34	65.53	**75.06**	72.81
9	95.07	97.65	97.02	97.66	98.18	98.13	98.45	**98.74**	98.66
10	43.38	53.64	53.62	63.62	70.87	70.03	70.99	70.49	**72.99**
11	44.36	49.65	49.00	52.24	57.67	51.85	**58.83**	56.05	58.56
12	4.85	3.54	5.09	8.34	11.06	7.96	10.17	10.50	**11.58**
13	57.60	61.16	63.18	67.47	73.82	73.41	76.69	80.51	**81.71**
14	32.98	45.92	46.18	52.63	61.74	54.34	60.04	70.31	**75.89**
15	32.37	63.96	64.83	89.66	87.25	89.16	93.02	**94.82**	93.89
16	32.54	45.51	62.74	75.98	81.03	80.34	84.01	**91.05**	86.70
17	95.96	**100**	97.98	**100**	**100**	**100**	**100**	**100**	**100**
18	15.12	25.71	26.44	41.95	54.38	52.59	54.97	68.91	**74.49**
19	29.58	38.37	35.79	47.41	84.09	72.81	81.48	91.15	**92.24**
20	41.89	48.99	61.59	61.78	71.93	67.48	73.02	76.12	**80.76**
OA	52.89	63.25	67.42	73.60	79.51	76.99	80.34	82.37	**83.61**
	5.39	2.60	0.97	3.39	1.59	2.91	1.83	1.86	**0.78**
AA	49.90	55.25	60.69	69.45	73.72	71.86	75.92	78.05	**79.10**
	3.39	2.92	0.66	2.63	1.66	2.91	2.17	2.24	**0.82**
κ	45.81	56.43	60.73	67.62	74.41	71.46	75.41	77.81	**79.27**
	5.14	2.63	1.00	3.73	1.87	3.34	2.12	2.19	**0.94**

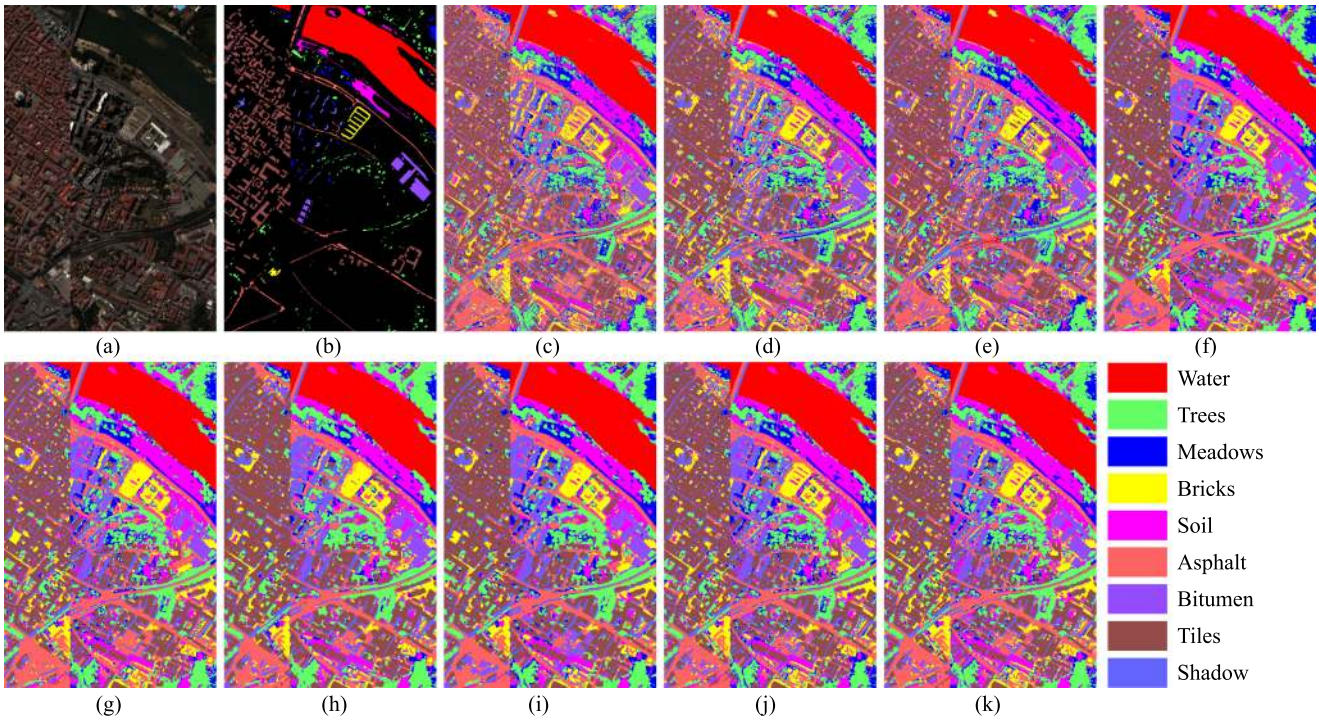


Fig. 9. Classification maps of different methods on the Pavia Centre dataset. (a) Pseudocolor image (bands 93, 53, and 19). (b) Ground-truth map. (c) SVM. (d) 3DCNN. (e) CDCNN. (f) FDSSC. (g) SSRN. (h) DBMA. (i) DBDA. (j) HResNet. (k) HResNetAM.

2018 dataset, 100 labeled samples and 50 labeled samples are used for Water and Unpaved parking lots classes, respectively.

4) Number of Scales and Kernels: In our proposed classification model, different number of scales and kernels can also influence the classification accuracy. For the sake of evaluating the classification ability of HResNetAM model toward different numbers of scales and kernels, we evaluate the influence of these two parameters using cross validation strategy. We set the

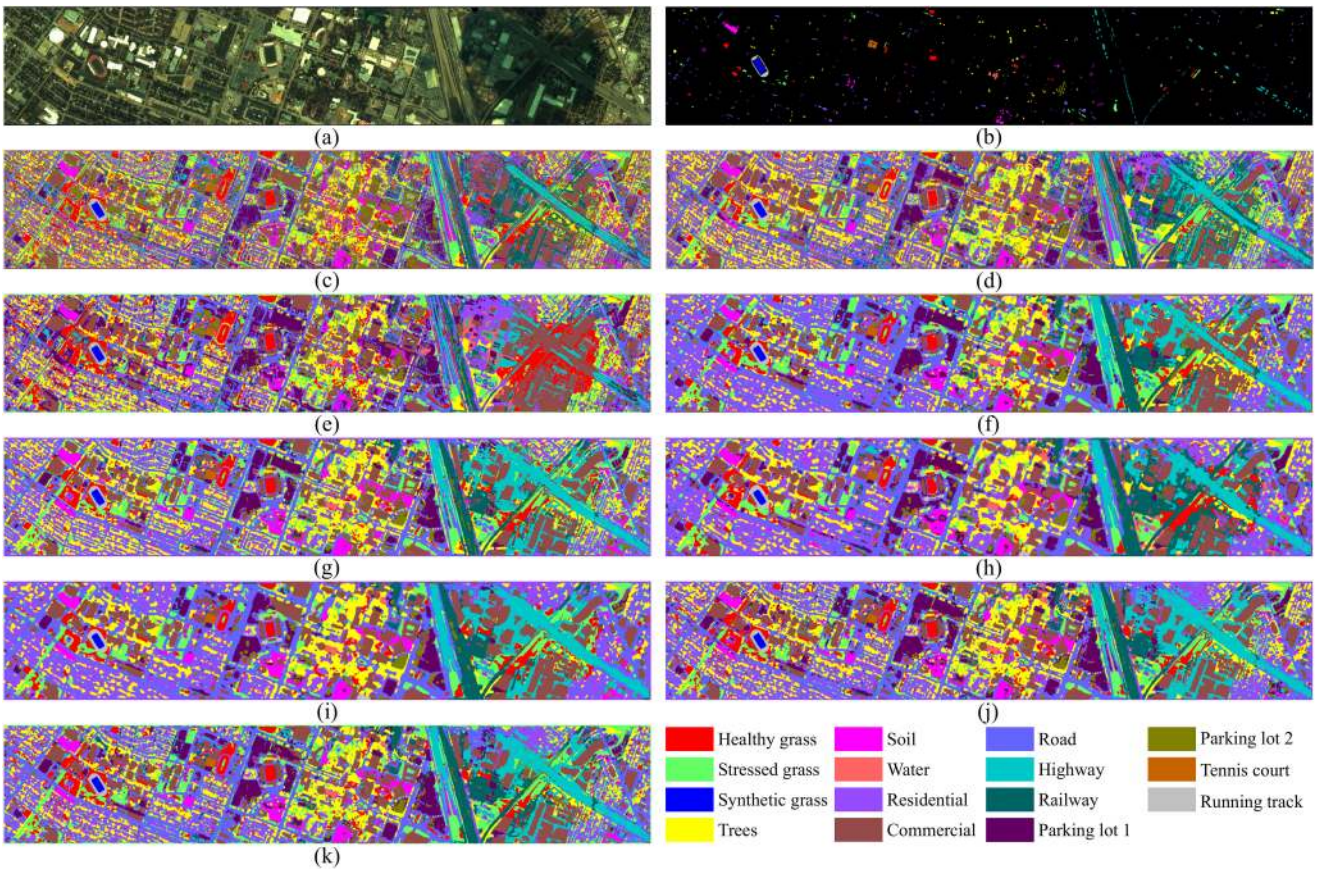


Fig. 10. Classification maps of different methods on the Houston 2013 dataset. (a) Pseudocolor image(bands 70, 30 and 12). (b) Ground-truth map. (c) SVM. (d) 3DCNN. (e) CDCNN. (f) FDSSC. (g) SSRN. (h) DBMA. (i) DBDA. (j) HResNet. (k) HResNetAM.

number of scales and kernels with $\{4, 5, 6, 7, 8\}$ and $\{4, 6, 8\}$, respectively. The average overall classification accuracies using different model structures are shown in Fig. 8. Based on the comparative results, we can learn that the classification accuracy is lower when the numbers of scales and kernels are low, especially for the Dioni and Houston 2018 datasets. Since the proposed method can extract multiscale features, higher scales can effectively improve the classification accuracy especially for more complex images. And the optimal parameter combinations of four HSI datasets are selected as $\{6, 8\}$, $\{7, 8\}$, $\{7, 6\}$, and $\{8, 6\}$, respectively.

D. Comparative Classification Results With State-of-The-Art Methods

1) Quantitative Comparisons: The average OAs, AAs, Kappa coefficients, and corresponding variance as well as average classification accuracy of every land-cover class with different classification methods on four benchmark HSI datasets are listed inTables VII–X. Note that the bold value in these tables represents the optimal value in the corresponding row. From the quantitative comparisons, several conclusions can be drawn, they are listed as follows.

1) First of all, in the case of the same number of training samples, the deeper the model is, the higher the classification accuracy will be. In four classification experiments, the deeper models (i.e., CDCNN, FDSSC, SSRN, DBMA, and DBDA) have higher classification accuracies than the 3DCNN.
2) The attention mechanism will improve classification to a certain degree. In four groups of HSI classification experiments, the DBMA and DBDA with attention mechanism generally have higher classification accuracies than the SSRN and FDSSC. Because the HResNet model has the same network structure with HResNetAM but without attention blocks, the classification results from HResNet can also verify the effectiveness of attention mechanism. And we can see that the HResNetAM has higher classification accuracies than HResNet in four classification experiments.
3) Comparing the overall accuracies of four different datasets, it can be found that the classification performance of four groups of HSIs is different. Among them, the Pavia Centre dataset has the highest classification accuracy and the Houston 2018 dataset has low accuracy. This is mainly due to the different levels of complexity within the

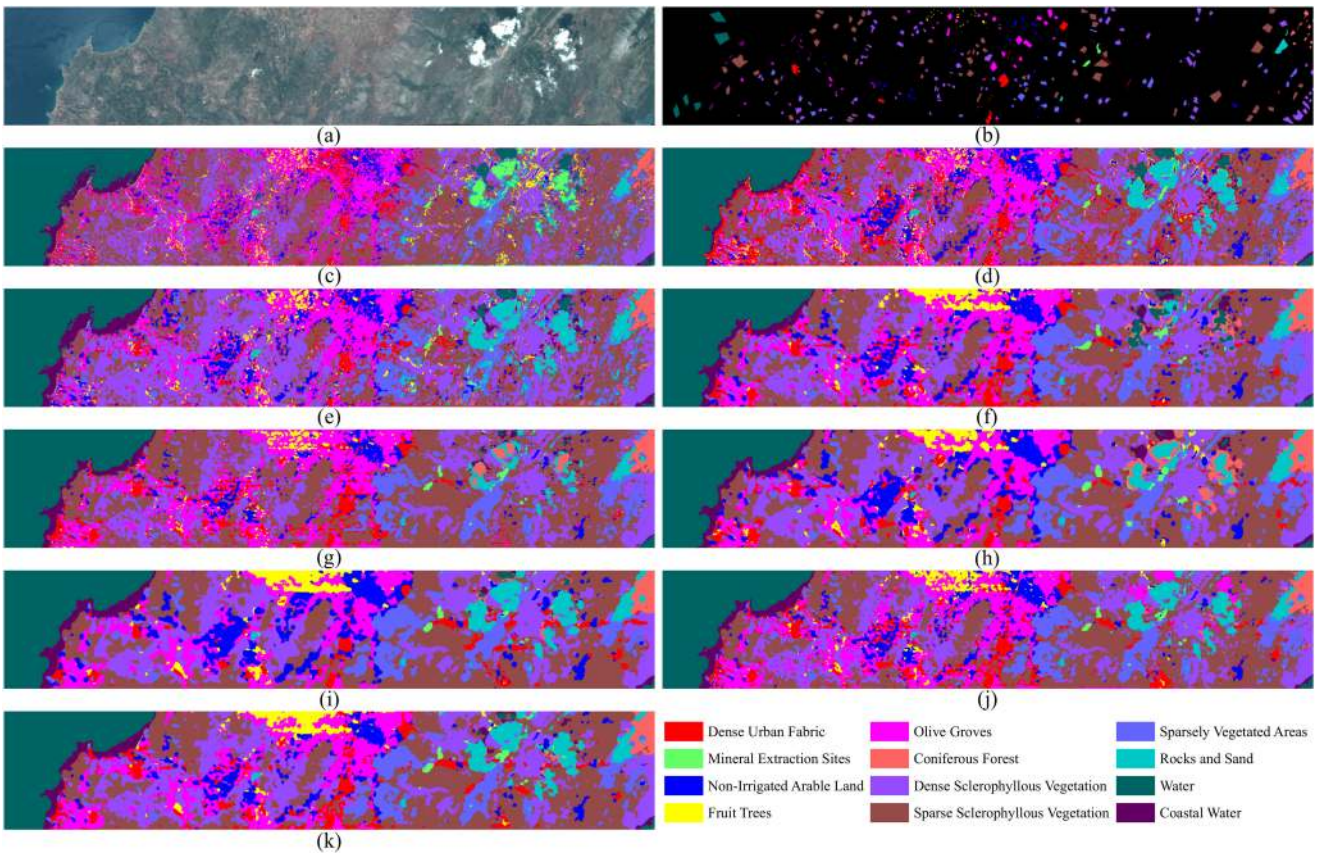


Fig. 11. Classification maps of different methods on the Dioni dataset. (a) Pseudocolor image (bands 23, 11, and 6). (b) Ground-truth map. (c) SVM. (d) 3DCNN. (e) CDCNN. (f) FDSSC. (g) SSRN. (h) DBMA. (i) DBDA. (j) HResNet. (k) HResNetAM.

hyperspectral datasets. However, by introducing the scale factor into HSI classification model, the classification performance can be significantly improved. For example, the average overall classification accuracy of the proposed HResNetAM model is 83.61%, which is 3.27% higher than the DBDA models (i.e., 80.34%). Thus, the scale factor will be helpful for the HSI classification model on complex datasets.

2) Qualitative Comparisons: Except the quantitative evaluation from Tables VII–X, the classification maps obtained by nine different methods are also exploited for qualitative evaluation. Figs. 9–12 show the classification maps on the Pavia Centre dataset, the Houston 2013 dataset, the Dioni dataset, and the Houston 2018 dataset with different classification methods, respectively. In these figures, the pseudocolor image and ground-truth are also displayed, and different land-cover classes are represented by different colors. When comparing the classification maps obtained by different methods in (c)–(k) with the ground-truth map in (b), we can learn that the proposed HResNetAM model can obtain more reasonable classification maps, which can prove the superiority of HResNetAM. In addition, when dealing with more complex images, take the Houston 2018 dataset, for example, traditional deep learning methods have more noise pixels in classification maps. Due to the introduction of scale factor and attention mechanism, the proposed HResNetAM can generate more homogeneous and reasonable classification maps.

E. Discussion

When performing the deep learning models for HSI classification, traditional models often have difficulties in extracting multiscale information at a granular level, which will affect the classification accuracy to some degree. To address this problem, we propose the HResNet with attention mechanism which can learn spectral and spatial features with different scales, and these features are fused for joint classification. The designed HResNetAM model, based on hierarchical residual learning and attention mechanism, can achieve better classification results compared with state-of-the-art deep learning models. The main reasons can be summarized as the following two aspects.

First, the importance of hierarchical features learning ability. The designed model utilizes HResNet to extract spatial and spectral features at different scales for the first time, which can learn characteristics from different receptive fields. And these global and local features can make contributions to the HSI classification results, especially for more complex images, such as the Dioni and Houston 2018 datasets, and the comparative experiments confirm the effectiveness of the scale factor for HSI classification.

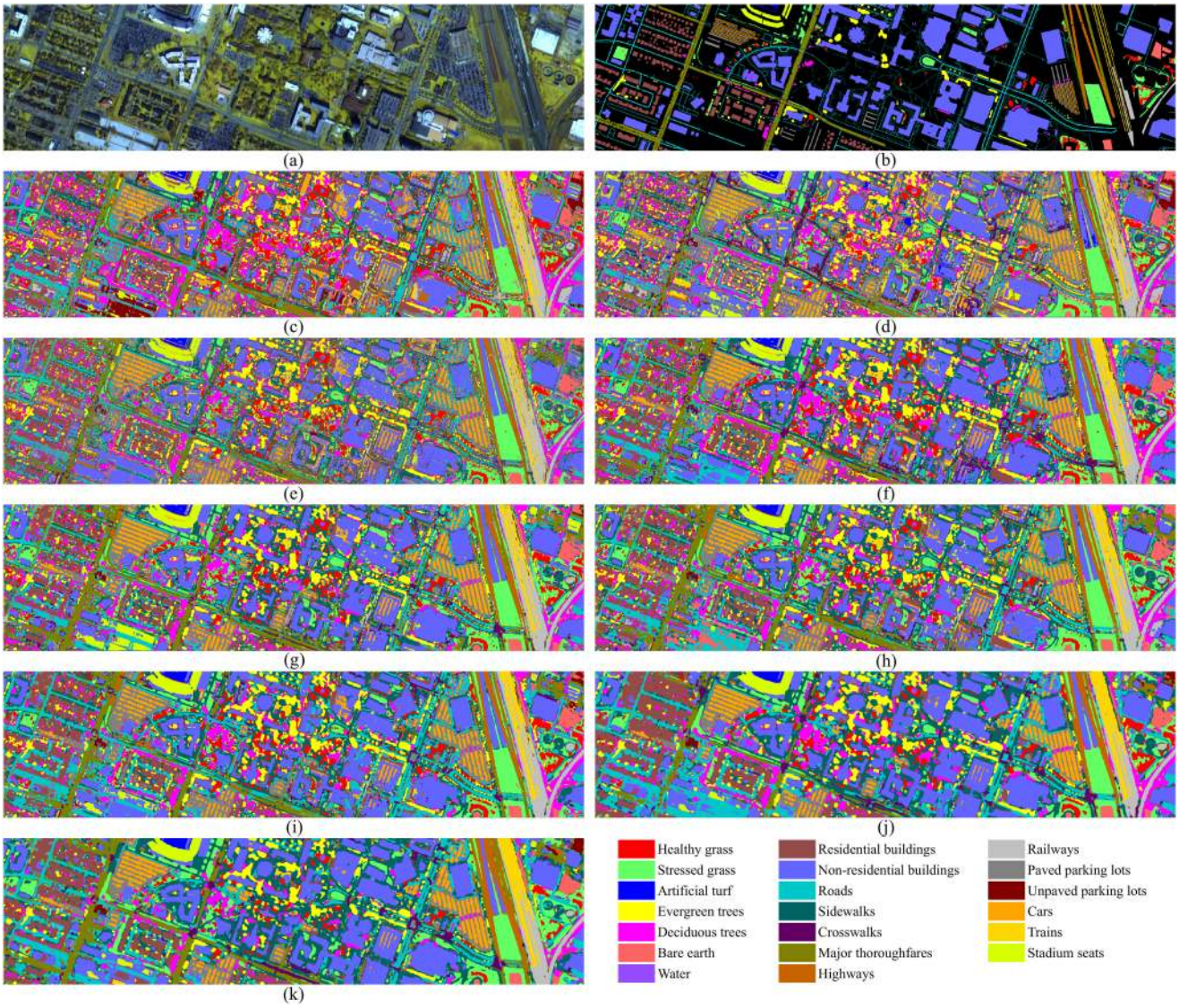


Fig. 12. Classification maps of different methods on the Houston 2018 dataset. (a) Pseudocolor image (bands 23, 11, and 6). (b) Ground-truth map. (c) SVM. (d) 3DCNN. (e) CDCNN. (f) FDSSC. (g) SSRN. (h) DBMA. (i) DBDA. (j) HResNet. (k) HResNetAM.

Second, the attention mechanism can further improve the classification performance to a certain extent. The attention mechanism is orthogonal to HResNet, so it is feasible to combine these two learning strategy for HSI classification. And the experimental results also verify the advantages of combing attention mechanism with HResNet.

IV. Conclusion

In this study, we propose a novel HResNet with attention mechanism model for HSI spectral-spatial classification, which have three advantages. The first one is that the proposed network utilizes hierarchical residual block to extract more discriminative spectral–spatial features of different scales, so as to maintain multiscale information for classification. The second one is that the attention mechanism is employed to calibrate the weights of hierarchical spectral and spatial features, and the third one is the double branch structure has potential in learning the spectral and spatial features separately. Therefore, the novel hierarchical residual network architecture with attention mechanism can extract more complete and discriminative information of HSI data by managing spectral–spatial features at a hierarchical level. And the residual learning structure and batch normalization can further improve the HSI classification efficiency in training process. The performance of HResNetAM has been verified on four benchmark HSIs compared with state-of-the-art models, and the experimental results have confirmed the superiority of proposed method.

Acknowledgment

The authors would like to thank Prof. P. Gamba for providing the Pavia Centre data, ISPRS for providing the Dioni data, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for providing the Houston 2013 and Houston 2018 datasets. The authors also want to thank the editor and reviewers for their careful reading and constructive comments.

REFERENCES

[1] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[2] 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.

[3] U. B. Gewali, S. T. Monteiro, and E. Saber, "Machine learning based hyperspectral image analysis: A survey," 2018, *arXiv:1802.08701*.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[5] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, pp. 195–204, Feb. 2019.

[6] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[7] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[8] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[9] M. Imani and H. Ghassemian, "An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges," *Inf. Fusion*, vol. 59, pp. 59–83, Jul. 2020.

[10] 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.

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

[12] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[13] X. Yang, Y. Ye, X. Li, R. Y. 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.

[14] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, May 2015, pp. 1–14.

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

[17] 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.

[18] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2017.

[19] 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.

[20] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2290–2304, Apr. 2019.

[21] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3034133.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[23] 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, Jul. 2018, Art. no. 1068.

[24] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 1454.

[25] M. Liang, L. Jiao, S. Yang, F. Liu, B. Hou, and H. Chen, "Deep multiscale spectral-spatial feature fusion for hyperspectral images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2911–2924, Aug. 2018.

[26] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, "A CNN with multiscale convolution and diversified metric for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3599–3618, Jun. 2019.

[27] Z. Li, L. Huang, and J. He, "A multiscale deep middle-level feature fusion network for hyperspectral classification," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 695.

[28] X. Zhu and J. Bao, "Hierarchical multi-scale convolutional neural networks for hyperspectral image classification," *Sensors*, vol. 19, no. 7, Apr. 2019, Art. no. 1714.

[29] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, doi: 10.1109/TGRS.2020.3008286.

[30] X. Mei *et al.*, "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, Apr. 2019, Art. no. 963.

[31] 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, Jun. 2019, Art no. 1307.

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

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

[34] S. K. Roy, S. Chatterjee, S. Bhattacharyya, B. B. Chaudhuri, and J. Platoš, "Lightweight spectral-spatial squeeze-and-excitation residual bag-of-features learning for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5277–5290, Aug. 2020.

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

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

[37] R. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, Jan. 2000.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by teducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.

[39] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[40] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[41] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5987–5995.

[42] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[43] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

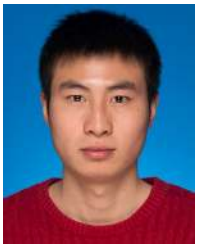
[44] N. Acito, S. Matteoli, A. Rossi, M. Diani, and G. Corsini, "Hyperspectral airborne "Viareggio 2013 Trial" data collection for detection algorithm assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2365–2376, Jun. 2016.

[45] K. Karantzalos, C. Karakizi, Z. Kandylakis, and G. Antoniou, *HyRANK Hyperspectral Satellite Dataset I. (Version V001)* [data set]. Apr. 2018.

[46] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

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

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



Zhixiang Xue received the B.S. degree in measurement and control engineering and M.S. degree in pattern recognition and intelligent system from the Information Engineering University, Zhengzhou, China, in 2014 and 2017, respectively, where he is currently working toward the Ph.D. degree in surveying and mapping science and technology.

His research interests include machine learning and remote sensing image processing.

Xuchu Yu received the Ph.D. degree in photogrammetry and remote sensing from the Institute of Surveying and Mapping, Information Engineering University, Zhengzhou, China, in 1997.

He is currently working with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, as a Professor and a Doctoral Supervisor. His research interests include photogrammetry, remote sensing, and pattern recognition.

Bing Liu received the B.S. degree in measurement and control engineering from the Information Engineering University, Zhengzhou, China, in 2013, the M.S. degree in pattern recognition and intelligent system from the Information Engineering University, Zhengzhou, China, in 2016, and the Ph.D. degree in surveying and mapping science and technology from the Information Engineering University, Zhengzhou, China, in 2019.

He is currently working with the PLA Strategic Support Force Information Engineering University, , Zhengzhou, China, as a Lecturer. His research interests include machine learning, pattern recognition, and signal processing in earth observation.

Dr. Liu is an active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE ACCESS, the *International Journal of Remote Sensing*, the *Remote Sensing Letter*, and the *Journal of Applied Remote Sensing*.

Xiong Tan received the B.S. degree in measurement and control engineering from the Information Engineering University, Zhengzhou, China, in 2008, the M.S. degree in pattern recognition and intelligent system from the Information Engineering University, Zhengzhou, China, in 2011, and the Ph.D. degree in surveying and mapping from the Information Engineering University, Zhengzhou, China, in 2014.

He is currently working with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, as a Lecturer. His research interests include pattern recognition and image processing.

Xiangpo Wei received the B.S. degree in remote sensing science and technology from the Information Engineering University, Zhengzhou, China, in 2012, the M.S. degree in photogrammetric and remote sensing from the Information Engineering University, Zhengzhou, China, in 2015, and the Ph.D. degree in surveying and mapping from the Information Engineering University, Zhengzhou, China, in 2019.

He is currently a Lecturer with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China. His research interests include hyperspectral image processing, pattern recognition, and deep learning.