Multiscale DenseNet Meets With Bi-RNN for Hyperspectral Image Classification

Lianhui Liang, *Student Member, IEEE*, Shaoquan Zhang, and Jun Li, *Fellow, IEEE*

***Abstract*—Convolutional neural network (CNN) has been successfully introduced to hyperspectral image (HSI) classification and achieved effective performance. With the depth of the CNN increases, it may cause the gradient to become zero, and the structure lacks the utilization of the correlated spatial feature information between different convolutional layers. At the same time, this single-scale convolution kernel is insufficient in expressing the complex spatial structure information of HSI. In addition, the CNN-based methods treat the HSIs spectral band data as a disordered vector in the process of feature extraction, which abandons the exploitation of its internal spectral correlations. To address these issues, we propose a novel spectral–spatial network classification framework based on multiscale dense connected convolutional network (DenseNet) and bidirection recurrent neural network (Bi-RNN) with attention mechanism network (MDRN). For the proposed MDRN, in terms of spatial feature extraction, a multiscale DenseNet is exploited to combine shallow and deep convolution features to extract the multiscale and complex spatial structure features at each layer. In the aspects of spectral feature extraction, Bi-RNN with attention mechanism is used to capture the inner spectral correlations within a continuous spectrum. Three standard real hyperspectral datasets were used to verify the effectiveness of the proposed MDRN approach. Experimental results indicate that the proposed MDRN method can make full use of the spectral and spatial information of the image, and it has better performance than some advanced algorithms in HSI classification. Finally, in the application of hyperspectral data captured by Gaofen-5 satellite, the practicability of the proposed MDRN method is also superior to other methods.**

***Index Terms*—Attention mechanism, bidirection recurrent neural network (Bi-RNN), convolutional neural network (CNN), dense connected convolutional network (DenseNet), hyperspectral image (HSI) classification.**

Manuscript received 20 January 2022; revised 13 May 2022; accepted 23 June 2022. Date of publication 28 June 2022; date of current version 18 July 2022. This work was supported in part by the Key Research and Development Program of Hunan Province under Grant 2019SK2102, in part by the National Natural Science Foundation of China under Grant 61901208, Grant 61771496, and Grant 62141105, and in part by China Postdoctoral Science Foundation under Grant 2020M672483. *(Corresponding author: Shaoquan Zhang.)*

Lianhui Liang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: lianglh@hnu.edu.cn).

Shaoquan Zhang is with the Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing, School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China, and with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: zhangshaoquan1@163.com).

Jun Li is with the Hubei Key Laboratory of Intelligent Geo-Information Processing, School of Computer Science, China University of Geosciences, Wuhan 430078, China (e-mail: lijuncug@cug.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3187009

I. Introduction

HYPERSPECTRAL image (HSI) consists of hundreds of continuous and narrow spectral bands of surface objects from the visible to infrared wavelength range [1], [2], thereby containing plentiful spectral signatures and spatial information, which makes it possible to accurately distinguish geographic objects and classify objects. In recent years, HSI are widely applied in many filed, such as urban planning [3], target detection [4], fine agriculture [5], among others [6]. HSI classification aims at assigning a certain unique label to per pixel in the HSI, which is an important HSI processing task. At present, how to effectively classify the HSI has become a hot research topic [7]–[9].

With the great breakthrough and rapid development of the deep learning (DL) methods in the domain of computer vision [10], [11], it has brought many inspirations to remote sensing image processing [7], [8], [12], [13]. In contrast with traditional manual feature-based methods, such as support vector machine (SVM) [14], morphological profiles [15], and k-nearest neighbor [16], DL-based methods have powerful feature extraction capabilities, that is, they can learn more discriminative and high-level semantic information from the abundant spectral signatures and complex spatial features of HSI [8], [12], [17], so it has been widely used in HSI classification. In [18], a deep stacked autoencoder network was used to learn spectral features from HSI for classification. Chen *et al.* [19] applied the singular restricted Boltzmann machine and a multiple layer deep network to extract the spectral signatures of HSI. However, the aforementioned methods only used spectral information and ignored the role of spatial structure information in enhancing the feature representation. Recent spectral–spatial feature extracting method illustrated that complex spatial structure can provide extra contextual information, which is favorable for enhancing classification performance. Chen *et al.* [20] adopted a 3-D convolutional neural network (CNN)-based feature extraction model with combined regularization to obtain effective spectral–spatial features. Yang *et al.* [21] proposed a double-branch spectral–spatial feature extraction model. It utilized 1-D CNN to extract spectral features and 2-D CNN to extract spatial features. Connect the learned spectral features and spatial features, and then fed them to the fully connected (FC) layer to extract the spectral–spatial features for subsequent classification. By using multiview deep autoencoder model to fuse spectral and spatial features, a new method for HSI classification based on multiview deep neural network is proposed [22]. In [23], a simplified 2D-3D CNN framework was introduced to achieve the spatial and

spectral feature fusion, thereby effectively extracting refined features and improving classification performance. Among them, a 2-D CNN was primary used to extract spatial feature, and 3-D CNN mainly focused on employing the band corelation data by exploiting a reduced kernel. Despite the deeper convolution neural networks can capture finer features and present satisfactory results in classification, they cannot make full use of features at different levels, which also may result in gradient disappearance and overfitting. The dense connected convolutional network (DenseNet) can effectively alleviate those problems, which uses concatenation for feature fusion and dense connection. In detail, it ensures the maximum flow of HSI information between all layers in the network by directly connecting all layers to each other [24]. Li *et al.* [25] proposed a deep multilayer fusion dense network based on 2-D dense block and 3-D dense block dual branch network, which effectively alleviates the problem of gradient disappearance and strengthens the use of features. In [26], a fast dense spectral–spatial convolution framework (FDSSC) was proposed, which uses two different dense blocks to deepen the network and strengthen the direct use of feature information between different layers.

However, since the abovementioned methods are only based on a single-scale convolution kernel, it cannot effectively learn the feature information reflecting the complex spatial structure, which results in limited classification quality. Therefore, it is necessary to improve the classification performance by combining the multiscale features of different convolution layers. In [27], the multiscale and multilevel spectral–spatial feature fusion network was proposed for HSI classification. In this model, three neighborhood blocks of different scales were used as the input of the network to fully extract the feature information of different scales of HSI. Furthermore, this network also utilizes 2D-3D alternating residual blocks to combine the spatial features extracted by 2-D CNN with the spectral features extracted by 3-D CNN to achieve the fusion of spatial features and spectral features. In [28], a method based on multiscale dense network was proposed, which uses different scale information in the entire network structure to achieve deep feature extraction and multiscale feature fusion. In [29], a multiscale DenseNet framework was proposed, which can utilize the feature information among different layers through dense block, and can also apply multiple dense blocks to fuse the spatial feature information of different scales among different layers.

The spectral band data of HSIs are essentially a sequence-based data structure. However, a large number of advanced spectral classifiers, such as random forest, SVM, and CNN-based classifiers, belong to vector-based methods. They regard it as a disordered high-dimensional vector for feature extraction, which does not conform to the characteristics of HSI spectral band data, which may cause information loss [30]. Aiming at this problem, a recurrent neural network (RNN) framework for HSI classification is proposed, which first considers the intrinsic sequential data structure of hyperspectral pixels to train a spectral classifier, and it is shown that this method has statistically higher accuracy than SVM and CNN. [30]. A bidirectional convolutional long short-term memory network [31] was proposed to learn spectral–spatial features from HSI, which processes the spectral bands as sequence data. In terms of experimental results, this network has achieved better performance than the CNN. In [32], RNN was employed to simulate the dependencies among different spectral bands to extract contextual information in the data. A cascaded RNN model [33] was proposed, which exploits two RNN layers to learn the complementary information from nonadjacent spectral bands and eliminate redundant information between adjacent spectrums. In [34], a spectral–spatial attention network (SSAN) was proposed, which uses bidirection recurrent neural network (Bi-RNN) to learn the internal spectral correlation within the continuous spectral bands, and uses CNN to obtain the spatial correlation between adjacent pixels. This strategy greatly improves the classification performance.

More recently, the attention mechanism has been widely used in HSI classification and has shown great potential in suppressing redundant information in the feature map and capturing significant features. Mei *et al.* [35] proposed a double-branch multiattention mechanism network (DBMA) based on DenseNet block and convolution block attention module, which utilizes two kinds of attention mechanisms to obtain more discriminative spectral and spatial features, respectively. Inspired by DBMA, a double-branch dual-attention mechanism network (DBDA) [36] was introduced, which exploits the channelwise and spatialwise attention mechanism to optimize the extracted feature maps. Hang *et al.* [37] proposed an attention-aided CNN model. This model adds attention modules to each convolutional layer of CNN to help the network pay attention to more saliency channels or spatial locations. In [38], a framework of 3-D octave (3DOC) convolution with the SSAN was proposed to extract saliency spectral–spatial information. The attention networks of spectral and spatial dimensions were introduced to emphasize the spectral bands and spatial information that greatly contributes to the HSI classification results.

Although the SSAN method mentioned previously treats the spectral data as a sequence structure and improves the extraction ability of spectral features, there is still a large room for improvement because of the proposed CNN-based method is insufficient for the utilization of the complex spatial structure features. As the depth of the CNN increases, it may also lead to the phenomenon that the gradient becomes zero. Second, although the abovementioned DBDA and DBMA methods both use the DenseNet structure, they can make full use of the feature information of different convolutional layers, alleviate the phenomenon of gradient explosion, and obtain good classification results. However, they are still CNN-based vector methods on spectral features, which lead to loss of information when representing hyperspectral pixels [30].

In order to alleviate the abovementioned problems and make full use of spectral and spatial features, inspired by DenseNet and the state-of-the-art SSAN method, a novel HSI classification framework based on the joint network of the multiscale DenseNet and Bi-RNN is proposed. It consists of two subnetworks, a multiscale DenseNet, and a Bi-RNN with attention mechanism network, which are used as spatial and spectral feature extractors, respectively. In the spatial subnetwork, using multiscale DenseNet instead of traditional CNN can not only

alleviate the phenomenon of gradient descent and combine the shallow and deep convolution layers features, but also utilize the different sizes of the convolution kernel in each dense block to extract multiscale spatial features. In the spectral subnetwork, we use the Bi-RNN attention network to extract the spectral features instead of the aforementioned CNN-based vector method. In addition, a new self-regularized and nonmonotonic Mish activation function is used to replace the Rule activation function in our proposed method to speed up convergence and further enhance the performance of the proposed model. The main contributions of this study are presented in the following:

1) In order to make full use of the complex spatial information, multiscale DenseNet is exploited to combine the shallow and deep convolution layers features, and extract spatial feature information of different scales in each layer.
2) A Bi-RNN with attention mechanism network is utilized to extract spectral feature information, which assign greater weights to the important spectral bands via adding additional spectral attention parameters, to improve the correlation between adjacent spectral bands.
3) To extract the integrated spectral–spatial features, we connect the Bi-RNN and multiscale DenseNet to form a new FC layer, and feed it into the softmax layer to predict the probability distribution of each class. In the network, a new self-regularized and nonmonotonic Mish activation function is used to replace Rule to further enhance the performance of the proposed model.

The rest of this article is organized as follows. Section II briefly introduces the related work. Section III presents the overall framework of the proposed MDRN model and the specific implementation process of the model. Section IV shows the classification results of MDRN, and discusses the performance of classification by comparing it with other widely used classification methods on three hyperspectral datasets. Section V shows the practical application and analysis results of MDRN in the hyperspectral data captured by Gaofen-5 (GF-5) satellite. Finally, Section VI concludes this article.

II. Related Works

A. Single-Scale DenseNet

The traditional CNN network only transfers the feature maps from one convolutional layer to the next layer in a single forward manner. It lacks the use of information from different layers in the CNN model to train the network. In general, DenseNet directly connects each layer with other layers, and combines features by concatenating them in the channel dimension to ensure that the information flow between each layer is maximized. Each convolution layer accepts information from the previous layer as input, and then conveys its feature map to the later layer [24]. Therefore, DenseNet can both sufficiently utilize the different layers of feature information and alleviate the phenomenon of gradient disappearance. Since the dense block is the basic unit in DenseNet, assuming that the output feature maps of the lth layer are z_l in DenseNet, the output of the lth layer dense block

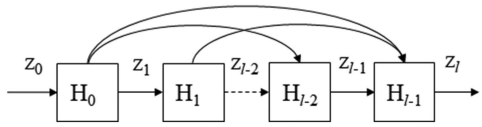

Fig. 1. Architecture of the dense convolution network.

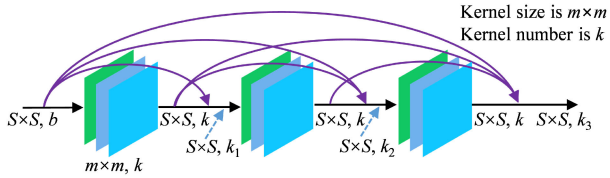

Fig. 2. Structure of the dense block used in our framework.

can be expressed as

$$z_l = H_l[z_0, z_1, \ldots, z_{l-1}] \tag{1}$$

where $H_l(\cdot)$ represents a functional module, which includes batch normalization layers, convolution layers, and Mish activation layers. $z_0, z_1, \ldots, z_{l-1}$ are the output feature maps of the previous corresponding convolutional layer. The architecture of the dense convolutional network is shown in Fig. 1. First, in order to fuse the spatial features of all bands and suppress irrelevant interference information, principal component analysis (PCA) is utilized to reduce the dimensionality of HSI to a low-dimensional subspace. After PCA, we can get a small spatial patch that contains almost all the feature information of the HSI. Assume we get the spatial size of the output features is $S \times S$ with b channels. Then, we regard the output features as the input of the dense block. The architecture of the dense block used in our framework is presented in Fig. 2. Specifically, the spatial size of the dense blocks input features is $S \times S$ with b channels. The number of convolution kernel in each convolutional layer is k, and the size of kernel is $m \times m$ ($m = 1, 3, 5, \ldots$). In addition, the padding method of 2-D convolution is set to "SAME" in each convolutional layer. Since the number of feature maps is equal to the number of convolution kernels, the shape of feature maps produced by each convolutional layer is $S \times S$ with k channels. Simultaneously, the number of channels in each layer is linearly related to the number of convolutional layers. The number of channels k_l in the dense block of the lth layers can be computed as follows:

$$k_l = b + (l-1) * k \tag{2}$$

where b expresses the initial channel's number in the input features. By means of using the dense block, the complex spatial features information of the HSI can be efficiently obtained from the shallow and deep convolutional layers.

III. Proposed Method

The overall structure of the proposed multiscale DenseNet and Bi-RNN joint network is shown in Fig. 3. As can be seen, it is mainly consisted of three parts: spectral feature extraction network, spatial feature extraction network, and spectral–spatial

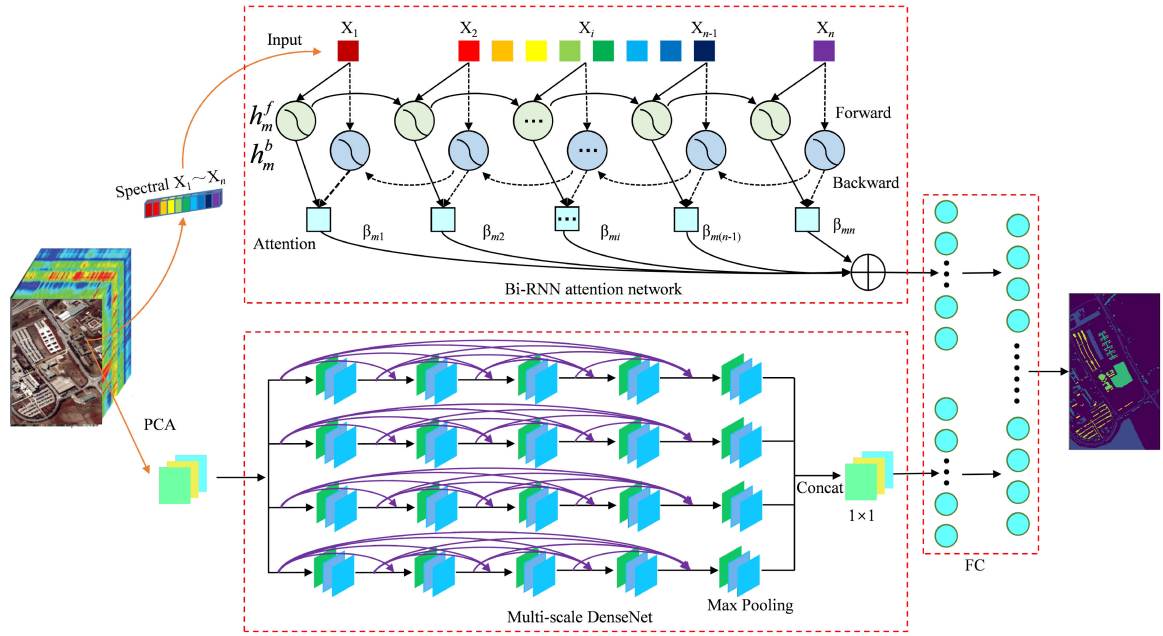

Fig. 3. Overall structure of the proposed multiscale DenseNet and Bi-RNN joint network.

features fusion network. Generally, the HSI data can be expressed as $I \in R^{H\times W\times B}$, where H, W, and B represent that the HSI data contain $H \times W$ pixels, and B spectral bands, respectively.

In the proposed MDRN model, in the aspect of spatial feature information extraction, a multiscale convolution kernel DenseNet is designed as the spatial feature extraction network substructure to extract the plentiful spatial information from the low-dimensional 3-D HSI data obtained by PCA. Then, the maximum pooling layer is used to further reduce the size of the spatial feature map after each single-scale convolution kernel DenseNet. In addition, after the fusion spatial feature map is obtained by connecting the outputs of the four maximum pooling layers, 1×1 convolution is applied to reduce the dimensionality of the feature channel. In terms of spectral feature information extraction, according to the magnitude of the classification contribution, the Bi-RNN attention network substructure is exploited to assign appropriate weights to each spectral band of the original HSI data to fully extract the spectral feature information of the HSI. Subsequently, in order to utilize better the spectral–spatial information to classification, an FC layer is exploited to fuse spectral and spatial features.

A. *Multiscale DenseNet for Spatial Feature*

Despite DenseNet can fully extract detailed features from different convolutional layers, the abovementioned is only performed under a fixed-size single-scale convolution kernel. Generally speaking, since the fact that large homogeneous regions need large-scale kernels, the detailed structural information requires small-scale kernels. Therefore, a single-scale convolution kernel cannot obtain the complex structure spatial features of the HSI well. This feature not only contains large homogeneous regions, but also has rich detailed structural information [26].

In this context, we develop a multiscale convolution kernel DenseNet method to address the aforementioned problems. Multiscale DenseNet consists of four dense blocks, four maximum pooling blocks, and a reducing dimension block. In this article, the experimental parameters setting example of the multiscale convolution kernel DenseNet on the UP dataset is shown in Fig. 4.

Specifically, after PCA, the input pixel $m \times m$ of this model is set to 27×27, and the number of channels is 4. In order to fully exploit the multiscale spatial information, the size of the convolution kernel $m \times m$ in each dense block is set to 1×1, 3×3, 7×7, and 19×19, respectively (detail in Section IV). Meanwhile, the corresponding number of the convolutional kernel is set to 16, 32, 32, and 32, respectively. After addressing via the first convolution layer, we can obtain 20 spatial feature maps with the size of 27×27. Next, we can get 52 fixed spatial feature maps through the second convolutional layer. The feature maps obtained from the third and fourth layers can be deduced by analogy. Then, after the fourth convolutional layer, the maximum pooling is used to further reduce the dimension of the spatial feature maps size. After the maximum pooling operation, we get 116 feature maps with a spatial size of 14×14. Finally, assuming that the output feature of the single-scale dense block is S_i, the spatial feature extraction process of multiscale dense blocks D_i can be expressed as

$$D_i = \text{Concat}[S_1, S_2, \ldots, S_i] \tag{3}$$

where $S_1, S_2, \ldots, S_i$, respectively, represent the output of the maximum pooling operation after the single-scale convolution kernel dense block. Concat$(\cdot)$ denotes the concatenation operation, and the value of i is the number of dense blocks. In addition,

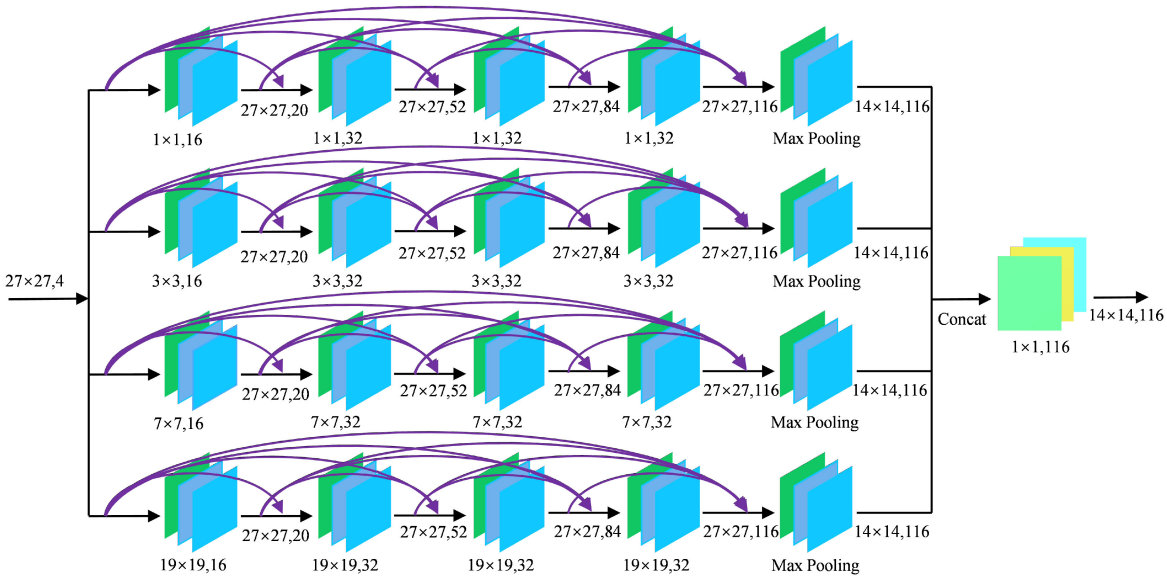

Fig. 4. The flowchart of multiscale convolution kernel DenseNet.

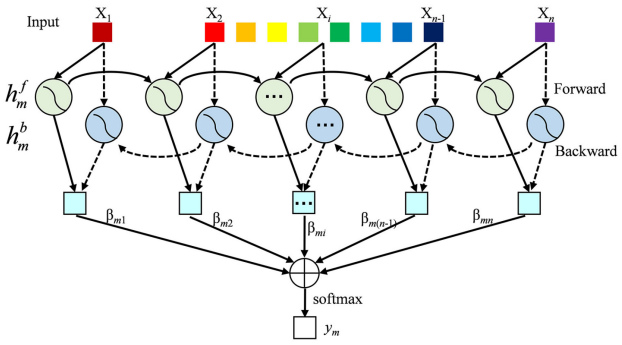

Fig. 5. Structure of Bi-RNN with attention mechanism network.

in order to reduce the dimensionality of the channel and reduce the network parameters to shorten model training and testing time, 1×1 convolution is utilized in spatial feature extraction. After this reducing dimension block, the 116 feature maps with the spatial size of 14×14 are obtained.

B. Bi-RNN With Attention Mechanism for Spectral Feature

The basic units of Bi-RNN model is composed of input layer, hidden layer, and output layer. The basic conception of Bi-RNN is to connect two forward and backward hidden layers, and its structure can make full use of the contextual information corresponding to sequential data [34]. Since hyperspectral pixels can essentially be regarded as a sequence-based data structure in spectral space, Bi-RNN model is utilized to obtain the spectral feature of hyperspectral sequential data in this article. The flowchart of Bi-RNN with attention mechanism network is shown in Fig. 5. The model processes the forward and backward input information to the same output layer g_m with two separate hidden layers, so that the output of the bidirectional hidden layer is calculated as follows:

$$g_m = \text{concat}\left(h_m^f, h_m^b\right) \tag{4}$$

where h_m^f and h_m^b are the output of the forward hidden layer and the backward hidden layer, respectively. Assuming the input is a spectral vector of hyperspectral pixel sequence data $X = (x_1, x_2, \ldots, x_n)$, then h_m^f and h_m^b can be, respectively, represented as follows:

$$h_m^f = \delta(W_{xf} x_m + W_{hf} h_{m-1} + b_f) \tag{5}$$

$$h_m^b = \delta(W_{xb} x_m + W_{hb} h_{m+1} + b_b) \tag{6}$$

where $\delta(\cdot)$ is a nonlinear activation function, x_m represents the m_{th} spectral band, and the value of m ranges from 1 to n. W_{xf} and W_{hf} are the input weight matrix of the current step and the weight matrix of the recurrent activation hidden unit h_{m-1} at the previous step, respectively. W_{xb} and W_{hb} are also the input weight matrix of the present step, and the weight matrix of the recurrent activation hidden unit h_{m+1} at the succeeding step, respectively. The b_f and b_b are bias coefficients.

The spectral vectors of the HSI are input to the Bi-RNN model one by one to obtain continuous spectrum features information with forward and backward directions. Since the spectrum is not a constant straight line, but a continuous curve with peaks and valleys, this means that each spectral channel contributes differently to the spectral characteristics. Therefore, in order to better learn the spectral relationships, it is necessary to assign an appropriate weight to each spectral channel according to the magnitude of the classification contribution. The attention mechanism used in the Bi-RNN model is a good way to assign an appropriate weight to each spectral channel.

As illustrated in [34], an attention layer is added to decode different spectral band information to extract more spectral features. The output weights of the attention layer can be computed

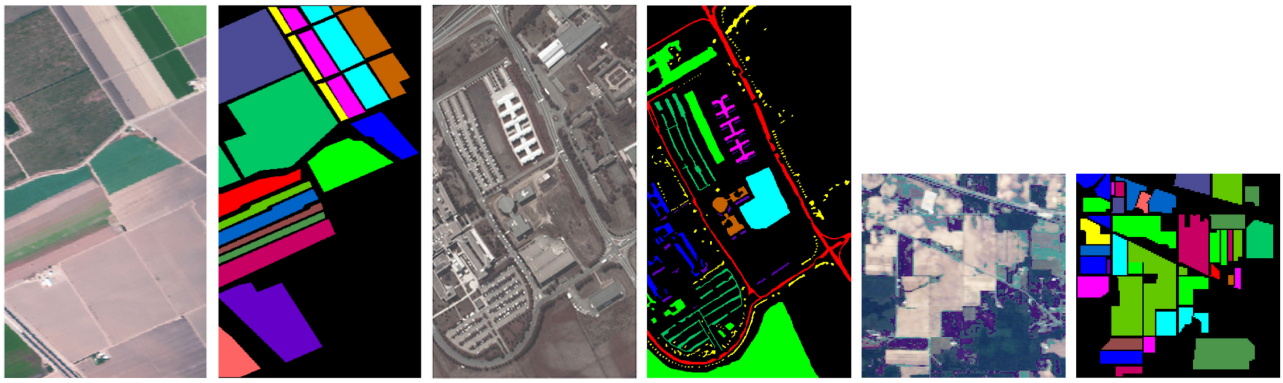

Fig. 6. False-color image and ground truth map for SV, UP, and IP dataset. (a) False-color image of the SV dataset. (b) Ground truth map of the SV dataset. (c) False-color image of the UP dataset. (d) Ground truth map of the UP dataset. (e) False-color image of the IP dataset. (f) Ground truth map of the IP dataset.

as follows:

$$\beta_{im} = f(W_i^o \varphi(W_i g_m + b_i) + b_i^o) \tag{7}$$

where W_i and W_i^o are weight matrices, b_i and b_i^o are bias coefficients, and φ and $f(\cdot)$ are tanh and softmax activation functions, respectively. Finally, the outputs of the bidirectional hidden layer g_m are multiplied by the corresponding attention weights β_{im}, and a new spectral feature vector y_m is obtained by adding them together.

C. Spectral–Spatial Feature Fusion

The whole structure of the multiscale DenseNet and Bi-RNN with attention mechanism network is shown in Fig. 3. The main idea is to extract the rich spatial feature information and spectral features by using the spatial branching network of multiscale DenseNet and the spectral branching network of Bi-RNN with attention mechanism. After that, the last FC layer in the spatial branch network and the last FC layer in the spectral branch network are merged to form a new spectral–spatial feature FC layer. In the end, the classification results are obtained through the FC layer and the softmax layer.

In addition, since the new activation function Mish has been proven to perform better than Relu in terms of model accuracy and convergence speed [39], [40], the Mish function can be expressed as follows:

$$\text{Mish}(x) = x * \varphi(\ln(e^x + 1)) \tag{8}$$

where φ expresses tanh activation function. The Mish activation function is used instead of Relu activation function in the proposed network to further enhance the performance of the proposed model.

IV. Experiments and Analysis

In this section, three HSI datasets are employed to evaluate the effectiveness of the proposed MDRN, including Salinas Valley (SV), University of Pavia (UP), and Indian Pines (IP).

A. Datasets Description

Salinas Valley: The SV dataset was acquired by the AVIRIS over the Salinas Valley of California. The SV dataset contains 512×217 pixels with a spatial resolution of 3.7 m/pixel, and the spectral wavelength cover range 0.36 to 2.5 μm. It has 204 spectral bands after discarding 20 water absorption spectral bands and consists of 16 classes of ground objects with 54 129 labeled pixels. The false-color image and ground truth classification map are shown in Fig. 6(a) and (b).

University of Pavia: The UP dataset was captured by the ROSIS over the city of Pavia, Italy, in 2002. The UP dataset contains 610×340 pixels with a 1.3 m/pixel spatial resolution, and the spectral wavelength cover range 0.43 to 8.6 μm. It consists of 103 spectral bands after discarding 12 noisy bands and includes nine classes of ground objects with 42 776 labeled pixels. Fig. 6(c) and (d) exhibits the false-color image and corresponding ground truth map.

Indian Pines: The IP dataset was captured by the AVIRIS sensor over northwestern Indiana, USA. It consists of 145×145 pixels with a spatial resolution of 20 m/pixel and the spectral band cover range is 0.4 to 2.5 μm. It contains a total of 220 spectral bands, which includes 20 water absorption spectral bands. There are 16 classes of ground objects with 10 249 labeled pixels for this scene. As shown in Fig. 6(e) and (f), the false-color image and corresponding ground truth map are demonstrated.

B. Experimental Configuration and Evaluation Indicators

In the respect of experimental conditions, all experiments were conducted on a computer with Intel Core i9-9900K@3.60 GHz CPU, NVIDIA Geforce RTX 2080Ti 11 GB, RAM 64 GB, the software platform was python 3.6.0 and TensorFlow 1.9.0 framework.

In order to validate the effectiveness of the proposed MDRN approach, MDRN is compared with several advanced algorithms, including SVM [14], FDSSC [26], DBDA [36], DBMA [35], spectral–spatial residual network (SSRN) [41], SSAN [34], ARNN [34], and 3DOC [38]. Subsequently, these methods will be introduced, respectively.

SVM: SVM with nonlinear discriminant function.

FDSSC: The framework of the FDSSC method is based on two different densely 3-D CNN blocks to extract the spatial and spectral information without any attention mechanism.

DBDA: The framework of the DBDA method is a dual-attention mechanism network, based on two densely 3-D CNN blocks, which applies a self-attention mechanism to better obtain both spectral features and spatial features.

DBMA: The framework of the DBMA method is a densely connected 3-D CNN-based double-branch multiattention mechanism network, which applies both channelwise attention to focus on spectral features and spatialwise attention to emphasize spatial features.

SSRN: The framework of the SSRN method is a SSRN based on 3-D CNN.

SSAN: The framework of the SSAN method is an SSAN, which is based on CNN with attention and Bi-RNN with attention.

ARNN: The framework of the ARNN method is based on CNN and Bi-RNN with attention, where the CNN branch has no attention.

3DOC: The framework of the 3DOC method is a 3DOC convolution with the SSAN.

In addition, in order to quantitatively evaluate the performance of the methods in this article, three standard evaluation metrics are exploited, including overall classification accuracy (OA), average classification accuracy (AA), and Kappa coefficient (Kappa) [42].

Meantime, to reduce the random errors, each experimental samples are randomly selected, and all experiments were executed 20 times. The final classification results of all methods are the average and standard deviation. The whole experimental sample data are divided into training set, validation set, and test set. For the three datasets SV, UP, and IP, we randomly choose 1.1%, 2%, and 12% labeled samples of per class as the training and validation samples for each dataset, respectively. Meanwhile, the remaining samplings are selected as testing samples.

C. Parameters Analysis

It plays an important role for the performance of the model in the design of the network framework and the parameter settings, such as learning rate, dropout, input spatial size, kernel numbers of convolutional layers, and batch size. Therefore, the effect of these hyperparameters will be specifically analyzed.

1) Learning Rate: The learning rate controls the step size of the gradient descent when the model is trained each time. The convergence and speed of the training process largely depend on the setting of the learning rate. In some cases, when the learning rate is too small, it may cause the model to fail to converge; otherwise, it may cause the model to oscillate periodically. Thus, a suitable learning rate is extremely significant for the model training. We conducted lots of experiments on each dataset to select the optimal learning rate from [0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005]. In order to accelerate the training speed and avoid converging a local minimum, the optimal learning rate for the SV, UP, and IP datasets was all set to 0.005 in the first 5000 training epochs, and then set to 0.0005.

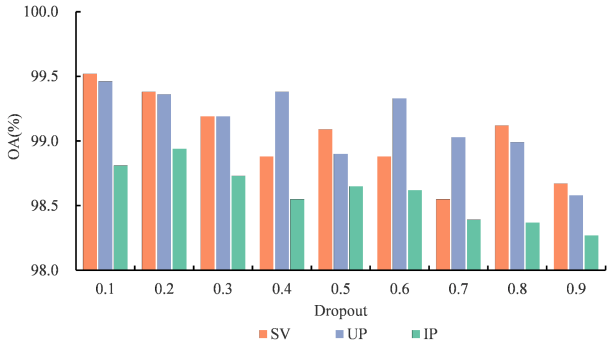

Fig. 7. OA of MDRN method with different dropout ratios.

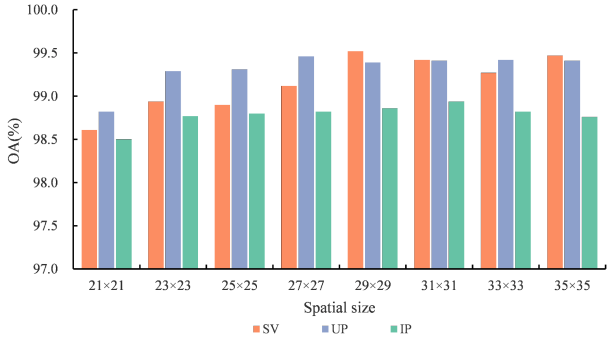

Fig. 8. OA of MDRN method with different spatial size.

2) Dropout: When dropout is training a neural network model, it can effectively alleviate the occurrence of overfitting and achieve the effect of regularization to a certain extent. The experimental results of choosing different dropout ratios on the SV, UP, and IP datasets are shown in Fig. 7. The results illustrate that the best classification accuracy is obtained when the dropout of the SV, UP, and IP datasets are 0.1, 0.1, and 0.2, respectively.

3) Spatial Size: How much complex spatial information around a pixel is used for spatial features learning depends on the size of the spatial neighbor region. In order to evaluate the influence of the spatial input sizes on classification performance, we tested the MDRN method on a large set of spatial sizes 21, 23, 25, 27, 29, 31, 33, and 35. The OA of the MDRN method with the different spatial sizes is shown in Fig. 8, where the MDRN achieved the optimal performance when the batch size is 29×29, 31×31, and 27×27 on the SV, IP, and UP datasets, respectively.

4) Kernel Size: As mentioned in [43], a small convolutional kernel can reflect the detailed spatial structural information on HSI, whereas a large convolutional kernel can express the large homogeneous region. The effects of the single-scale convolutional kernel of different sizes on the classification accuracy of the proposed method are analyzed on three HSI datasets. As shown in Fig. 9, the classification accuracy varies in the situation when the MDRN method considers convolutional kernel

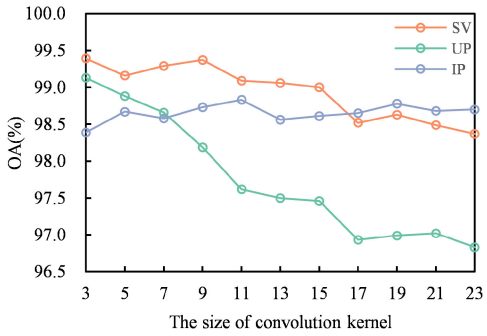

Fig. 9. Experimental results of different single scale convolution kernel on three datasets.

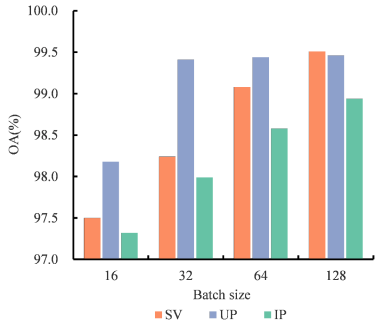

Fig. 10. OA of MDRN method with different batch size.

of different sizes as input, and the smaller or larger convolutional kernel cannot ensure better performance. In our proposed method, four convolutional kernels of different sizes are used, which is set to 1, 3, 7, and 19, separately.

5) Batch Size: The values of batch size are in the range from 16 to 128 for researching its influence, and the OA results of the MDRN method with different batch size are shown in Fig. 10. It is obvious that the accuracy of MDRN method improves as the batch size value varies from 16 to 128 on three datasets. As the batch size value increases, the memory requirements also increase. Thus, selecting an appropriate batch size value can not only improve the convergence accuracy of the model, but also sufficiently enhance the memory utilization. Since the MDRN method obtains the highest accuracy when the batch size value is 128 on three HSI datasets, we set the batch size value to 128 after considering the memory capacity.

D. Classification Result and Maps

For a fair comparison, the experimental parameters of all methods are carried out by utilizing the optimal parameters in the reference paper. The classification results of all methods on SV, UP, and IP datasets are given in Tables I–III, where the optimal classification accuracy is given in bold, and the corresponding classification maps for those methods are shown in Figs. 11–13.

From Table I, we can see that the proposed MDRN method obtains the best classification accuracy, with 99.51% for OA, 99.57% for AA, and 99.44% for Kappa. For SVM, it obtains the worst accuracy with only 89.34% OA. Compare with the SVM method depended on hand-crafted feature extraction, other methods achieve better classification accuracy, because those methods based on DL can extract hierarchical and nonlinear features, and they also consider the spatial information and spectral information for classification. In addition, the OA obtained by the MDRN method is higher than ARNN 4.89% OA and SSAN 3.61% OA. The reasons can be summarized as follows: first, the multiscale densely connected CNN is deeper than the traditional CNN in structure, and can make full use of the features of the shallow and deep convolution layers. Second, the multiscale strategy can utilize more complex spatial information than CNN. These two advantages guarantee MDRN can obtain more discriminative spatial features. The classification results of the DBMA, FDSSC and DBDA methods based on densely connected CNN are worse than those of MDRN method, which illustrates that the Bi-RNN with attention mechanism has stronger ability in extracting spectral features.

From Fig. 11, it can be observed that compared with other classification maps based on spectral–spatial features fusion, the classification maps of SVM contain a lot of salt-and-pepper noise. Among these methods, the classification map of the proposed MDRN method presents less noise and smoother features in homogenous regions, which indicates that our method has better performance in extracting complex spatial structure features and spectral features. Furthermore, compared with ARNN, it can be seen that the classification maps of MDRN exhibit relatively the least noise, especially in large homogeneous regions, which further illustrates the advantages of multiscale convolutional kernels based on dense connections in spatial features extraction.

The Table II presents that the classification accuracy based on DL method is better than SVM. For example, the OA obtained by 3DOC is 97.92%, which is around 6.17% higher than SVM. Although the classification results of the MDRN method are worse than other methods in some classes, the accuracy per class utilizing the proposed MDRN method exceeds 97.99%, and the OA, AA, and Kappa values are the highest among all methods. It demonstrates that the proposed method can effectively obtain the distinguishing features between different classes. Specifically, the classification accuracy of the MDRN method is 99.46% in terms of OA, while the OA of other competitive methods is less than 98.90%.

Similarly, it can be obviously seen from Fig. 12 that the classification maps of SSRN, DBDA, FDSSC, DBMA, SSAN, and 3DOC methods are better than SVM, but those methods still have a great deal of mislabeled noise in the blue part of the corresponding classification maps. In addition, although the ARNN and MDRN methods have relatively less noise in the blue part, the ARNN method has more obvious mislabeling noise than the MDRN method in the middle light blue-part of the classification map. As a result, compared with other methods, MDRN has the best classification result on the UP dataset.

It can be seen from Table III that compared with other spectral–spatial methods based on DL, the proposed MRDN method still obtains the optimal classification accuracy on the

TABLE I
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS ON SV DATASET

Class	SVM	FDSSC	DBDA	DBMA	SSRN	SSAN	ARNN	3DOC	MDRN
C1	98.98	**100.00**	97.62	**100.00**	**100.00**	99.03	83.60	**100.00**	**100.00**
C2	99.37	**100.00**	99.93	**100.00**	99.98	93.83	96.71	99.84	99.97
C3	90.73	99.85	98.27	99.52	98.98	98.60	99.33	**100.00**	99.95
C4	99.19	98.08	97.30	95.13	98.53	99.63	99.93	95.37	**99.93**
C5	94.65	99.25	99.72	98.81	**99.80**	99.54	99.35	98.47	99.47
C6	99.54	99.99	99.94	99.79	99.92	**100.00**	**100.00**	99.46	**100.00**
C7	99.60	**100.00**	99.82	99.93	99.81	99.20	95.34	99.60	99.69
C8	84.69	97.46	95.55	98.29	92.70	92.10	87.29	96.31	**98.66**
C9	99.52	99.87	99.73	99.75	99.81	98.68	99.39	98.29	**99.74**
C10	91.04	99.36	99.13	98.31	98.78	99.13	99.31	95.63	**100.00**
C11	95.21	97.90	98.09	96.51	96.63	99.52	99.71	**100.00**	99.81
C12	97.66	99.88	99.01	99.07	98.95	99.79	99.95	**99.95**	99.89
C13	97.99	99.91	99.95	99.60	98.75	88.59	88.37	**100.00**	99.66
C14	88.91	**99.20**	98.65	98.71	99.17	90.63	92.64	96.65	97.61
C15	59.09	94.21	91.02	95.55	91.84	91.45	92.39	94.25	**99.61**
C16	94.68	**100.00**	99.90	99.78	99.65	95.93	98.59	99.60	99.21
OA/%	89.34 (0.20)	98.26 (1.45)	97.27 (0.46)	97.90 (0.56)	96.90 (0.55)	95.87 (0.37)	94.62 (0.35)	97.64 (0.43)	**99.51** (0.18)
AA/%	88.11 (0.22)	99.06 (0.59)	98.35 (0.28)	98.48 (0.42)	98.33 (0.14)	96.61 (0.19)	95.74 (0.91)	98.34 (0.31)	**99.57** (0.11)
Kappa/%	93.18 (0.45)	98.07 (1.61)	96.96 (0.52)	97.66 (0.63)	96.55 (0.61)	95.41 (0.31)	94.02 (0.78)	97.38 (0.52)	**99.44** (0.19)
Training/s	28.05	358.56	153.78	259.93	206.04	2391.28	2384.14	1849.12	2625.62
Testing/s	3.64	33.63	42.19	42.69	26.34	50.57	50.01	81.35	54.85

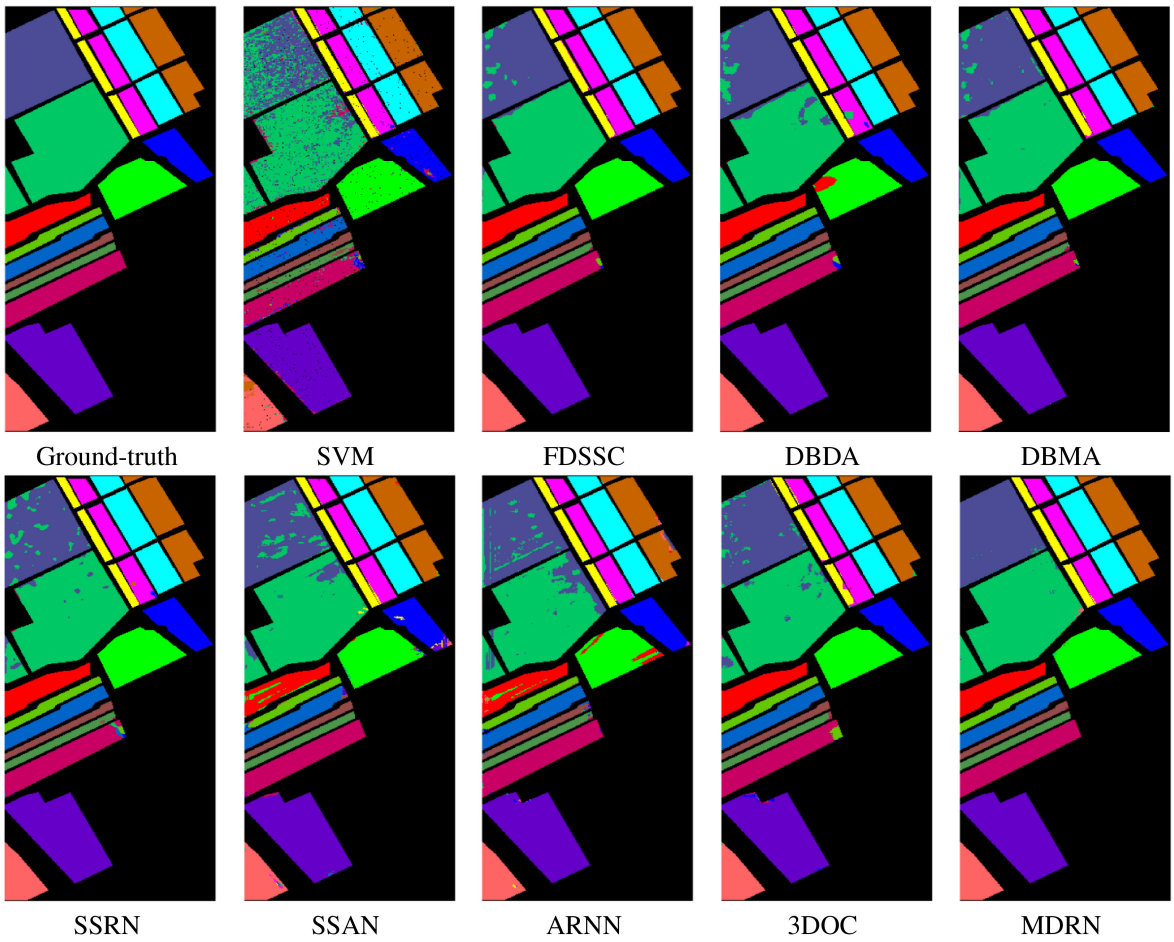

Fig. 11. Classification maps obtained by different methods on SV dataset.

TABLE II
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS ON UP DATASET

Class	SVM	FDSSC	DBDA	DBMA	SSRN	SSAN	ARNN	3DOC	MDRN
C1	89.63	98.99	99.03	99.19	99.03	93.50	91.86	97.49	**99.20**
C2	98.52	99.52	99.75	99.05	99.62	97.97	99.25	99.79	**99.91**
C3	79.35	98.64	96.99	92.83	94.56	79.21	79.21	89.43	**98.91**
C4	91.29	99.32	98.28	96.01	**99.64**	94.90	96.84	95.27	97.99
C5	99.07	99.71	99.39	99.06	99.98	99.92	99.69	**100.00**	**100.00**
C6	77.32	99.32	99.21	99.44	98.75	98.65	90.66	98.96	**100.00**
C7	79.62	**100.00**	97.64	99.05	98.80	87.46	84.95	93.57	99.61
C8	88.00	94.63	93.65	94.08	93.05	90.83	96.12	95.81	**98.16**
C9	**100.00**	98.16	98.53	97.74	96.69	94.39	95.93	97.47	99.78
OA/%	91.75 (0.35)	98.89 (0.37)	98.64 (0.52)	98.05 (0.38)	98.45 (0.91)	95.45 (0.20)	95.13 (0.10)	97.92 (0.17)	**99.46** (0.06)
AA/%	88.95 (0.46)	98.70 (0.33)	98.05 (0.80)	97.38 (0.52)	97.79 (1.52)	96.21 (0.18)	92.58 (0.12)	96.42 (0.24)	**99.28** (0.15)
Kappa/%	89.20 (0.75)	98.53 (0.49)	96.20 (0.69)	97.42 (0.51)	97.93 (1.44)	94.93 (0.25)	93.56 (0.12)	97.24 (0.22)	**99.29** (0.09)
Training/s	21.17	237.24	150.53	158.74	155.38	1369.03	1243.65	733.88	3337.01
Testing/s	1.56	18.21	25.41	22.14	17.62	21.87	21.80	44.32	27.38

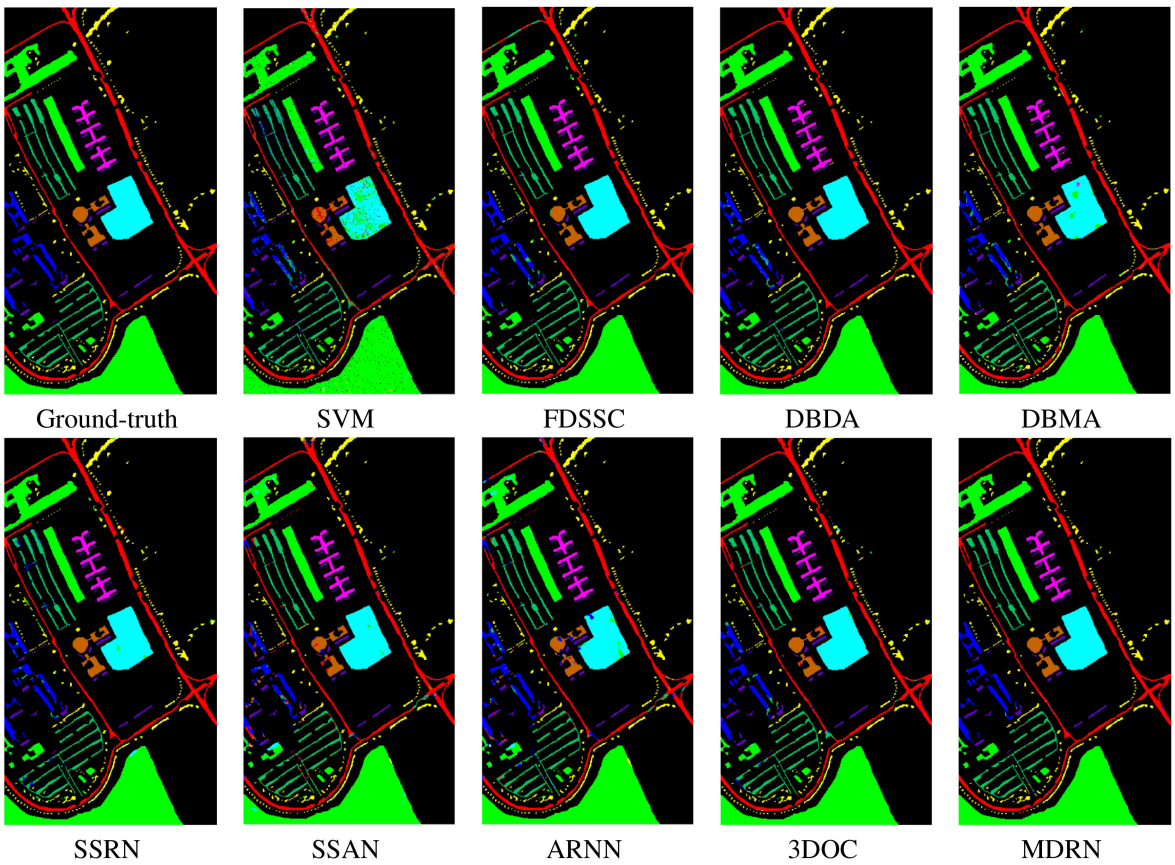

Fig. 12. Classification maps obtained by different methods on UP dataset.

three indicators of OA, AA, and Kappa. In addition, the classification accuracy of all methods is not higher than 98.17% in the twelfth class, whose results are not ideal. The reason is that the within-class variability is high, and the limited training samples cannot be used to express completeness.

As shown in Fig. 13, compared with other methods based on spectral–spatial features fusion, the classification maps of the MDRN method show fewer misclassified pixels on the right-hand side of the classification map.

E. Investigation of the Proportion of Training Samples

The performance of DL largely depends on the training samples. Generally, the more samples used for training, the

TABLE III
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS ON IP DATASET

Class	SVM	FDSSC	DBDA	DBMA	SSRN	SSAN	ARNN	3DOC	MDRN
C1	11.76	99.02	98.85	92.74	**100.00**	**100.00**	97.06	97.06	**100.00**
C2	67.34	97.45	98.15	97.29	97.69	**99.63**	96.77	98.62	99.35
C3	58.73	98.51	98.66	**99.24**	98.64	98.41	97.78	97.78	97.62
C4	45.25	99.05	95.87	97.72	95.94	96.65	98.32	**100.00**	98.88
C5	82.29	98.83	99.00	97.39	**99.02**	98.64	90.19	91.01	98.37
C6	97.47	98.77	**99.68**	98.90	99.04	99.10	99.10	99.28	97.83
C7	50.00	83.37	90.61	89.10	99.09	95.00	**100.00**	**100.00**	95.00
C8	99.45	99.82	99.89	99.65	98.65	**100.00**	**100.00**	**100.00**	**100.00**
C9	14.29	93.33	95.67	84.31	**100.00**	78.57	85.71	92.86	**100.00**
C10	71.41	97.52	97.62	96.44	96.09	96.88	97.56	96.88	**98.10**
C11	81.07	99.23	98.76	98.22	98.53	98.50	99.03	98.45	**99.62**
C12	49.89	**98.17**	97.78	93.35	97.87	95.55	95.32	97.12	97.10
C13	98.71	96.82	97.59	99.35	98.98	98.71	**100.00**	98.71	99.35
C14	95.11	98.58	98.54	98.72	99.20	98.86	**100.00**	99.69	99.90
C15	48.97	98.80	98.64	95.23	98.51	98.29	95.89	**99.66**	99.32
C16	75.36	95.01	90.64	93.00	90.57	**100.00**	92.75	**100.00**	98.55
OA/%	76.21 (0.54)	98.41 (0.15)	98.36 (0.36)	97.56 (0.66)	98.15 (0.50)	98.41 (0.18)	97.82 (0.27)	98.21 (0.29)	**98.93** (0.14)
AA/%	72.68 (0.66)	97.02 (0.87)	97.25 (1.63)	95.68 (1.47)	97.99 (0.57)	97.05 (0.79)	96.59 (0.75)	97.94 (0.95)	**98.69** (0.51)
Kappa/%	65.44 (1.01)	98.19 (0.17)	98.13 (0.41)	97.98 (0.76)	97.89 (0.56)	98.20 (0.20)	97.52 (0.31)	97.96 (0.24)	**98.79** (0.16)
Training/s	166.21	618.34	368.14	523.25	415.69	2832.77	2572.35	5240.06	5960.82
Testing/s	1.96	5.21	6.41	6.57	3.99	10.02	10.01	21.24	11.89

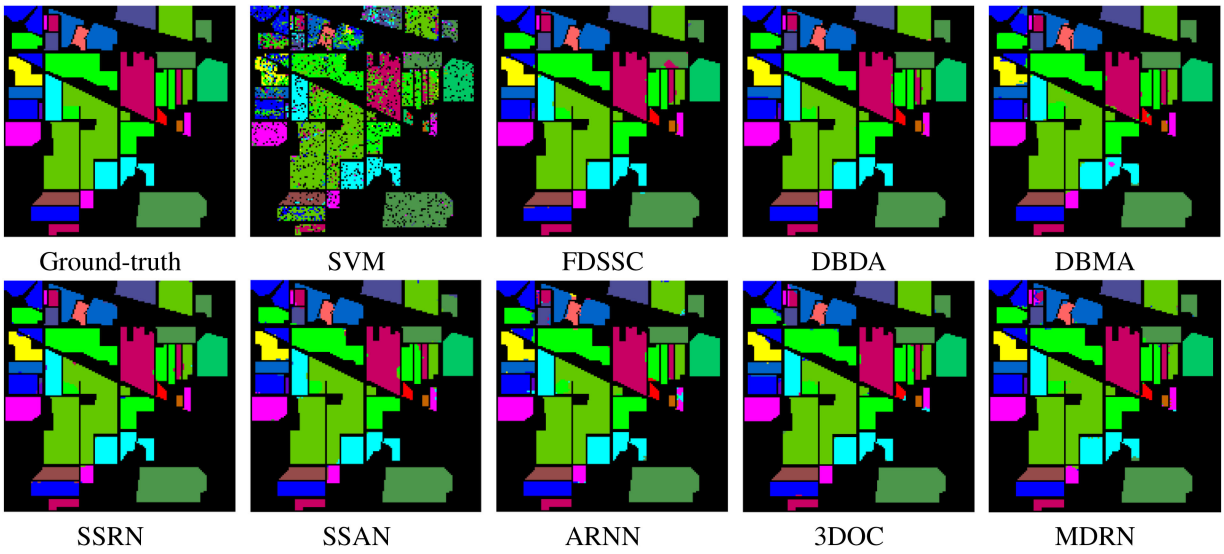

Fig. 13. Classification maps obtained by different methods on IP dataset.

better the classification performance obtained. Therefore, in order to test the dependence of the MDRN training results on the training samples, we randomly select different percentages of samples for experimentation. The experimental results on three datasets with different percentages of training samples are shown in Fig. 14. The results indicate that the proposed MDRN still performs better than other methods in a situation where the number of training samples is limited. On the IP dataset with different percentages of training samples, the classification accuracy of SVM is too low and less than 86%, so it does not appear in Fig. 14(a). Meanwhile, due to the high cost of labeling datasets, the proposed MDRN method can also save a lot of labor costs.

F. Computer Cost

As given in Tables I–V, we presented the training time cost and test time cost of different comparison methods in these experiments. It can be seen that not only the classification accuracy but also the time computation cost of MDRN is the highest among all methods. It is determined by the network framework designed by the MDRN method. Since the multiscale DenseNet proposed

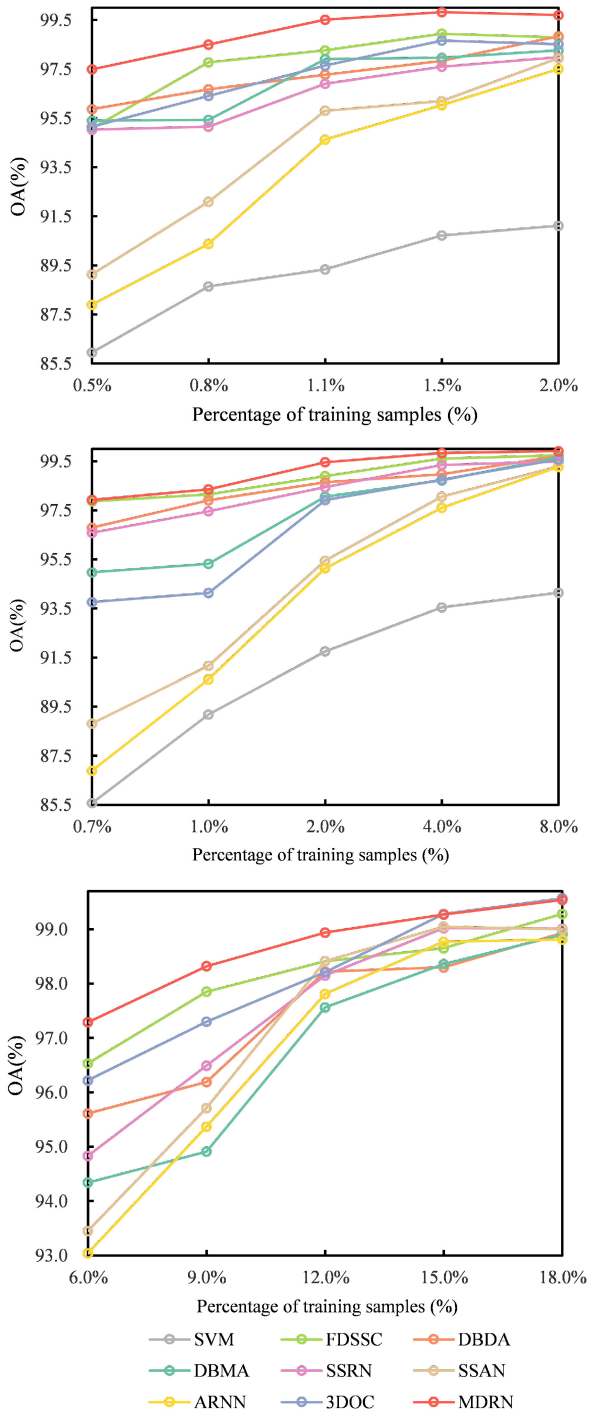

Fig. 14. Experimental results on three datasets with different percentages of training samples. (a) On SV dataset. (b) On UP dataset. (c) On IP dataset.

in this article is structurally deeper than the traditional CNN used in the SSAN method, it can fully utilize the features of shallow and deep convolutional layers in the network. Then, the network composed of dense blocks of multiscale convolution kernel can exploit more complex spatial information than other CNN-based comparison methods. These two characteristics ensure that the MDRN method can obtain more discriminative spatial features, but the deeper and wider spatial branch network of the SSAN method leads to huge training costs.

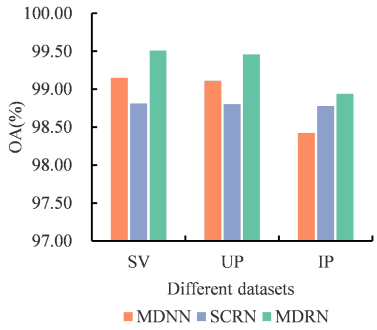

Fig. 15. Experimental results of different methods on three datasets.

Fig. 16. (a) False-color image and (b) ground-truth map for DTL dataset.

Furthermore, it is worth noting that although the test time of the MDRN method is faster than that of the octave convolution-based 3DOC method on all datasets, the former consumes more training time. This is because the number of times the MDRN method model is trained is larger.

G. Ablation Study

Two ablation experiments are conducted on three HSI datasets to verify the effectiveness of the proposed multiscale DenseNet block and Bi-RNN block. In this article, the SCRN model means that the model is based on the Bi-RNN and single-scale convolution kernel DenseNet without multiscale dense block structure. In addition, the MDNN model means that the model is based on the multiscale DenseNet block without Bi-RNN block. Fig. 15 exhibits the classification results achieved by the proposed MDRN model and this model without the corresponding component for the same training data. The experimental results of different methods on three datasets are shown in Fig. 15.

It can be clearly seen from Fig. 15 that without the corresponding component, the classification performance of the proposed MDRN model is worse than that of the proposed MDRN method. The reasons for those phenomena can be summarized into two respects. On the one hand, the introduction of multiscale DenseNet block can fully utilize complex spatial structural information. On the other hand, the introduction of Bi-RNN with attention block can adaptively assign appropriate weights to the spectral bands with different contributes, and obtain the correlation of adjacent spectral bands to better characterize spectral features of the HSI.

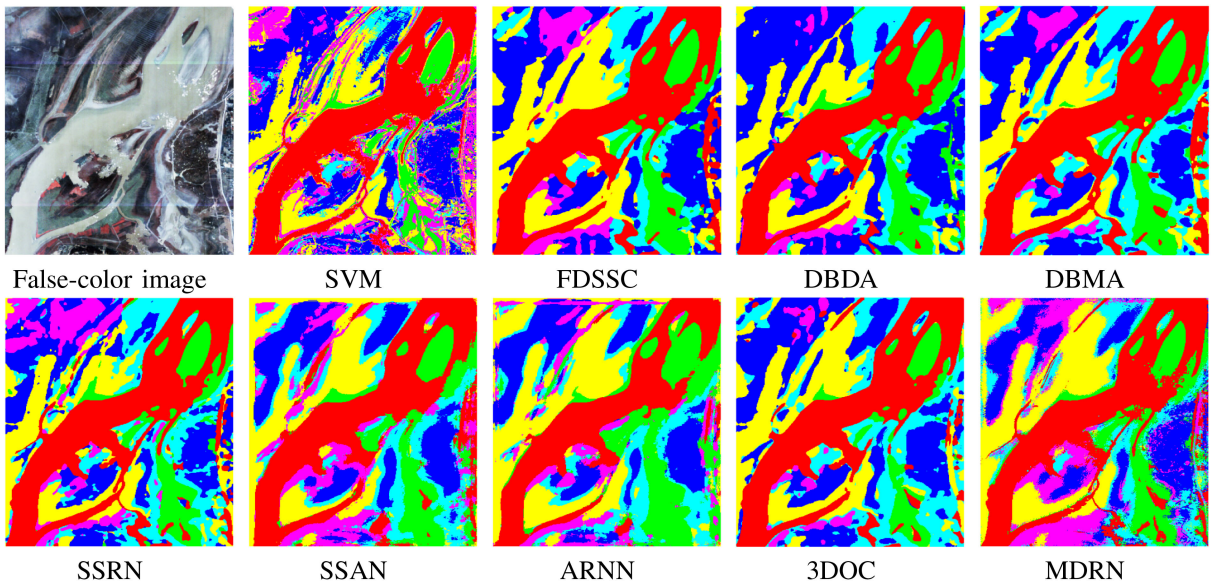

Fig. 17. Classification maps obtained by different methods on DTL dataset.

In addition, it can be seen from Tables I–III that the classification accuracy of ARNN on the three datasets of SV, UP, and IP is much lower than that of SCRN, which further shows that the dense network structure can fully utilize the feature information from different convolutional layers. It has stronger feature extraction ability than traditional CNN (without dense block structure).

V. Practical Application and Analysis

In this section, we further verify the practicability of the proposed MDRN method, by using the HSI data of a scene in the Dongting Lake basin, acquired by GF-5 satellite, over the Dongting Lake basin, on January 22, 2019, to conduct a test experiment. The GF-5 satellite is the latest satellite launched by China in 2018 to achieve hyperspectral resolution earth observation. It has six payloads to acquire remote sensing data at high spectral resolution in the ultraviolet to long-wave infrared bands [44], [45]. In this article, we named this HSI data Dongting Lake basin (DTL) dataset. The DTL dataset contains 456×352 pixels with a spatial resolution of 30 m/pixel, and the spectral wavelength cover ranges from 0.4 to 2.5 μm. The number of original spectral bands is 330. After eliminating 20 water absorption spectral bands whose noisy bands include: [193–199, 248–260], 310 spectral bands have remained for the classification. It consists of six classes of ground objects with 4816 labeled samples. Regarding the categories of labeled samples, we currently mark small-scale, high-confidence samples as training samples for model training based on the visual interpretation of satellite remote sensing images provided by Google Earth. Fig. 16 shows the false-color image and the ground truth classification map. In this experiment, we randomly choose 3% labeled samples of each class as training and validation samples. The samples for each class of training, validation, and testing for DTL dataset are listed in Table IV.

TABLE IV
Samples for Each Class of Train, Validation, and Test for DTL Dataset

Class	Total	Train	Validation	Test
C1	1120	34	34	1052
C2	736	23	23	690
C3	877	27	27	823
C4	1006	31	31	944
C5	644	20	20	604
C6	433	13	13	407
Total	4816	148	148	4520

From Table V, we can observe that our proposed MDRN method still achieves the optimal classification accuracy in our practical application scenes, with 99.53% for OA, 99.62% for AA, and 99.43% for Kappa. The classification maps using different methods on the DTL dataset is presented in Fig. 17. The classification map of the SVM method has a lot of salt-and-pepper noise and misclassification points. Relatively speaking, the classification map based on DL has better effect and better homogeneity. Compared with other DL-based method, SSAN and MDRN methods present higher classification accuracy. From the perspective of OA, compared with the SSAN, our proposed MDRN can achieve an improvement of 0.64%. From the perspective of classification map, the SSAN method based on the single-scale convolution kernel not only has some misclassified noise labels at the bottom of the map, but also is not as good as MDRN in the extraction of complex spatial details. Compared with other DL-based methods, the proposed MDRN method is not oversmoothed in details, and is closer to the original image. It further illustrates that the structure based on the multiscale convolution kernel DenseNet and Bi-RNN with attention mechanism can obtain more discriminative spatial and spectral features, and achieve superior results.

TABLE V
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS ON DTL DATASET

Class	SVM	FDSSC	DBDA	DBMA	SSRN	SSAN	ARNN	3DOC	MDRN
C1	99.43	**100.00**	**100.00**	99.46	**100.00**	**100.00**	96.96	97.53	99.72
C2	99.71	97.91	96.39	96.73	95.88	**100.00**	95.65	**100.00**	**100.00**
C3	95.87	97.23	97.72	97.96	97.27	**100.00**	99.88	95.38	99.64
C4	95.55	**100.00**	99.94	99.24	98.84	95.44	99.36	97.67	98.41
C5	90.07	98.99	98.55	98.02	97.46	98.84	99.17	**100.00**	**100.00**
C6	87.22	99.58	99.90	98.72	98.69	**100.00**	99.51	96.56	**100.00**
OA/%	95.66 (0.91)	98.96 (0.61)	98.76 (0.74)	98.55 (0.79)	98.10 (0.60)	98.89 (0.36)	98.31 (0.17)	97.79 (0.41)	**99.53** (0.20)
AA/%	94.71 (1.10)	98.95 (0.73)	98.75 (0.75)	98.39 (1.01)	98.02 (0.81)	99.04 (0.57)	98.42 (0.53)	97.86 (0.43)	**99.62** (0.26)
Kappa/%	94.64 (1.09)	98.74 (0.74)	98.48 (0.90)	98.24 (0.96)	97.68 (0.73)	98.65 (0.43)	97.95 (0.36)	97.30 (0.50)	**99.43** (0.25)
Training/s	3.02	232.40	80.19	94.08	84.40	3108.81	2957.17	1013.62	2913.31
Testing/s	0.14	4.22	4.88	5.39	3.27	8.81	8.75	18.95	9.58

VI. CONCLUSION

In this article, we propose a novel method based on multiscale DenseNet and Bi-RNN with attention mechanism for HSI classification. The proposed MDRN method includes two subnetworks, including a 2-D densely connected network with convolution kernels of different sizes and a Bi-RNN with attention mechanism, which extract complex spatial features and spectral features, respectively. Then, the extracted spatial feature information and spectral feature information are fused through the FC layer to make better use of the HSI feature information for classification tasks. Experiments conducted on three real datasets and a DTL dataset captured by GF-5 satellite demonstrate that the proposed MDRN approach can achieve excellent performance compared with other methods.

In the spatial feature extraction network of the MDRN model, the input of the multiscale DenseNet is some principal components obtained after preprocessing through the PCA method, which may both ignore the correlations among spectral bands and increase the complexity of the model. In future work, we will focus on exploring a more concise network structure while enhancing the classification accuracy of various datasets. Since the graph convolutional network can perform flexible convolution on arbitrary irregular image regions, it has received extensive attention, which is also one of our main research works in the future [46], [47]. In addition, due to the limited number of training samples, obtaining labeled sample data is usually very time-consuming, we will also explore semisupervised or unsupervised DL methods for HSI classification.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their outstanding comments and suggestions, which greatly helped to improve the technical quality and presentation of this article.

REFERENCES

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

[2] L. Ma, M. M. Crawford, L. Zhu, and Y. Liu, "Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2305–2323, Apr. 2019.

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

[4] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[5] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017.

[6] Y. Zhong *et al.*, "Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 4, pp. 46–62, Dec. 2018.

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

[8] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[9] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, early access, 2022, doi: 10.1109/MGRS.2022.3145854.

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

[11] 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.

[12] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[13] H. Patel and K. P. Upla, "A shallow network for hyperspectral image classification using an autoencoder with convolutional neural network," *Multimedia Tools Appl. Volume*, vol. 81, no. 1, pp. 695–714, 2021.

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

[15] J. Li, P. Marpu, A. Plaza, J. Bioucas-Dias, and J. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[16] W. Song, S. Li, X. Kang, and K. Huang, "Hyperspectral image classification based on KNN sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 2411–2414.

[17] X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[18] 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.
[19] 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.
[20] 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.
[21] J. Yang, Y. Zhao, and J. C. 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.
[22] A. Sellami and S. Tabbone, "Deep neural networks-based relevant latent representation learning for hyperspectral image classification," *Pattern Recognit.*, vol. 121, 2022, Art. no. 108224.
[23] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial-spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, Apr. 2020.
[24] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 2261–2269.
[25] Z. Li *et al.*, "Deep multilayer fusion dense network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1258–1270, Mar. 2020.
[26] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral-spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1068.
[27] C. Mu, Z. Guo, and Y. Liu, "A multi-scale and multi-level spectral-spatial feature fusion network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 125.
[28] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.
[29] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 246–259, Jan. 2021.
[30] L. Mou, P. Ghamis, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
[31] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1330.
[32] H. Wu and S. Prasad, "Convolutional recurrent neural networks forhyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 298.
[33] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.
[34] X. Mei *et al.*, "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 963.
[35] 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, 2019, Art. no. 1307.
[36] 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, 2020, Art. no. 582.
[37] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.
[38] X. Tang *et al.*, "Hyperspectral image classification based on 3-D octave convolution with spatial-spectral attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2430–2447, Mar. 2021.
[39] D. Misra, "Mish: A self regularized non-monotonic neural activation function," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1908.08681
[40] Z. Ge, G. Cao, X. Li, and P. Fu, "Hyperspectral image classification method based on 2D-3D CNN and multibranch feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5776–5788, 2020.
[41] 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.
[42] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.
[43] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.
[44] H. Ren, X. Ye, R. Liu, J. Dong, and Q. Qin, "Improving land surface temperature and emissivity retrieval from the chinese Gaofen-5 satellite using a hybrid algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1080–1090, Feb. 2018.
[45] H. Shi, Z. Li, H. Ye, H. Luo, W. Xiong, and X. Wang, "First level 1 product results of the greenhouse gas monitoring instrument on the Gaofen-5 satellite," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 899–914, Feb. 2021.
[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016. [Online]. Available: http://arxiv.org/abs/1609.02907
[47] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

Lianhui Liang (Student Member, IEEE) received the B.S. degree in measurement and control technology and instrument from Hunan University of Technology, Zhuzhou, China, in 2016, and the M.S. degree in instrumentation and meter engineering from Hunan University, Changsha, China, in 2019.

He is currently pursuing Ph.D. degree in control science and engineering from the College of electrical and Information Engineering, Hunan University. His research interests include hyperspectral image processing, signal processing, and deep learning.

Shaoquan Zhang received the B.S. degree in communication engineering and the M.E. degree in power engineering from the Nanchang Institute of Technology, Nanchang, China, in 2012 and 2015, respectively, and the Ph.D. degree in cartography and geography information system from Sun Yat-sen University, Guangzhou, China, in 2018.

He is currently an Associate Professor with the Nanchang Institute of Technology. His research interests include hyperspectral unmixing, sparse representation, and machine learning.

Jun Li (Fellow, IEEE) received the B.S. degree in geographic information systems from Hunan Normal University, Changsha, China, in 2004, the M.E. degree in remote sensing from Peking University, Beijing, China, in 2007, and the Ph.D. degree in electrical engineering from the Instituto de Telecomunicações, Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Lisbon, Portugal, in 2011.

She is currently a Full Professor with the China University of Geosciences, Wuhan, China. Her main research interests include remotely sensed hyperspectral image analysis, signal processing, supervised/semisupervised learning, and active learning.

Prof. Li is the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. She has been a Guest Editor of several journals, including the PROCEEDINGS OF THE IEEE and the *ISPRS Journal of Photogrammetry and Remote Sensing*.