




Hyperspectral Image Classification Method Based on 2D–3D CNN and Multibranch Feature Fusion

Zixian Ge , Guo Cao , Xuesong Li , and Peng Fu

Abstract—The emergence of a convolutional neural network (CNN) has greatly promoted the development of hyperspectral image (HSI) classification technology. However, the acquisition of HSI is difficult. The lack of training samples is the primary cause of low classification performance. The traditional CNN-based methods mainly use the 2-D CNN for feature extraction, which makes the interband correlations of HSIs underutilized. The 3-D CNN extracts the joint spectral–spatial information representation, but it depends on a more complex model. Also, too deep or too shallow network cannot extract the image features well. To tackle these issues, we propose an HSI classification method based on the 2D–3D CNN and multibranch feature fusion. We first combine 2-D CNN and 3-D CNN to extract image features. Then, by means of the multibranch neural network, three kinds of features from shallow to deep are extracted and fused in the spectral dimension. Finally, the fused features are passed into several fully connected layers and a softmax layer to obtain the classification results. In addition, our network model utilizes the state-of-the-art activation function Mish to further improve the classification performance. Our experimental results, conducted on four widely used HSI datasets, indicate that the proposed method achieves better performance than the existing alternatives.

Index Terms—Activation function, convolutional neural network (CNN), deep learning, feature fusion, hyperspectral image (HSI) classification.

I. INTRODUCTION

HYPERSPECTRAL remote sensing, also called imaging spectral remote sensing, is a multidimensional information acquisition technology combining imaging technology and spectral technology [1]. The hyperspectral remote sensor can simultaneously obtain the 2-D geometric spatial information and the 1-D spectral information of the target area. Therefore, the hyperspectral data have the form and structure of an “image cube,” which reflects the characteristics of “image spectrum unification.” While imaging the spatial features of the target, each spatial pixel is dispersed to form dozens or even hundreds of narrow bands for continuous spectral coverage. Hyperspectral images (HSIs) combine image information and

spectral information. The image information shows the external characteristics of the sample, such as size, shape, and texture. Meanwhile, the spectral information reflects the differences in the physical structures and chemical compositions of the sample. Therefore, the HSIs reflect the comprehensive characteristics of the image. Different components in the image have different spectral absorption factors, different internal physical structures and chemical compositions lead to different reflection characteristics, it is the basis of HSI classification technology. Hyperspectral remote sensing technology has been widely used in environmental monitoring [2], mineral exploration [3], [4], precision agriculture [5], [6], among others.

Recently, deep learning has become a promising approach to big data analysis. The great breakthrough has been made with the approach in many computer vision tasks. With the development of deep learning technology, research institutions and scholars applied this technology to the HSI classification field and embraced exciting achievements. Deep learning models mainly include Stacked AutoEncoder (SAE) [7]–[9], restricted Boltzmann machine [10], convolutional neural network (CNN) [11]–[14], recurrent neural network (RNN) [15]–[17], and generative adversarial network [18], [19].

Previously, most methods depend on spectral information [20]–[23] or spatial information [24], [25] for classification. However, utilizing spectral or spatial information alone is not enough to extract features with sufficient discrimination. In recent years, researchers tended to combine the spectral and spatial information [26]–[31] to deal with classification tasks, which achieved good performance. We now provide a brief summary of related work for spectral–spatial based methods.

In terms of spectral–spatial based method, one popularly used structure is decoupling the task for two streams, that is, a spectral feature extraction stream and a spatial feature extraction stream. Guo and Zhu [32] proposed a CNN-based spatial feature fusion algorithm. This algorithm first extracts the spectral and spatial features from original HSIs. Then, the spatial information is intelligently fused with the spectral features extracted by the artificial neural network model supervised by center loss, and the classification task is carried out by the fused features. In [33], a unified deep network combined with active transfer learning (TL) was proposed. The model can be well trained with a small number of labeled training data. This method first extracts deep joint spectral–spatial features through a layered and stacked sparse autoencoder (SSAE) network. Then, the corresponding active learning strategy is utilized to select limited

Manuscript received May 13, 2020; revised July 25, 2020 and August 17, 2020; accepted September 10, 2020. Date of publication September 18, 2020; date of current version October 1, 2020. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK2019020735 and in part by the National Natural Science Foundation of China under Grant 61801222 (*Corresponding author: Guo Cao.*)

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zxge727@foxmail.com; caoguo@njjust.edu.cn; cedar_xuesong@163.com; fupeng@njjust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3024841

labeled samples from the source and target domains to fine tune the SSAE network. Shen *et al.* [34] proposed a spectral–spatial domain-specific convolutional deep extreme learning machine in view of large and complex deep model network structure and time-consuming training. The model has a two-branch convolutional learning structure to separately extract spectral features and spatial features. Li *et al.* [35] proposed a two-stream spectral and spatial feature extraction and fusion network based on 2-D CNN. By fusing the features of the spectral feature extraction branch and the spatial feature extraction branch, joint spectral–spatial features are obtained. Moreover, SE-Conv and SE-Res modules based on the squeeze-and-excitation networks are designed to enhance the joint feature extraction capability. Zhou *et al.* [36] presented an HSI classification method using a spectral–spatial long short-term memory (LSTM), which includes spectral LSTM and spatial LSTM. This method first feeds the spectral and spatial features of each pixel into softmax layers for two results and then utilizes a decision fusion method to further obtain the joint classification results. Hang *et al.* [37] explored an attention-aided CNN network for the spectral–spatial classification of HSI. Similarly, this method uses the two-stream strategy to design a spectral attention subnetwork and a spatial subnetwork for joint feature classification. Combined with attention modules, this method aids the network focus on more discriminative channels or positions.

Another candidate structure is feeding a local cube, cropped from HSI, into a CNN module with 2-D or 3-D convolutional kernel for extracting the joint spectral–spatial features [38]. This is the wider used method to directly extract the joint spectral–spatial features for HSI classification. He *et al.* [39] proposed a multiscale 3-D deep convolutional neural network (M3D-DCNN), which jointly learns 2-D multiscale spatial features and 1-D spectral features. Zhong *et al.* [27] designed a supervised 3-D deep learning model for spectral–spatial representation learning and HSI classification. Inspired by the residual network (ResNet), this model contains many jump connections, which allows the network to be composed of more layers to extract the deep features. In [40], a semisupervised deep learning framework based on the ResNet was proposed, which utilizes the limited labeled data supplemented by abundant unlabeled data. This method guides the network to learn from unlabeled data by using the complementary cues of spectral and spatial features. In [41], a hybrid network model was designed by combining 2-D CNN with 3-D CNN, which effectively extracts the spatial and spectral features of HSIs. The model has a simple structure, fast training speed, and good classification performance. But the classification result using a small number of training data is not ideal. Liu *et al.* [42] proposed a deep multigrained cascade forest method called dgcForest. First, the cascade forest is embedded in the multigrained scanning process to obtain the deep representative features with high diversity. Then, these features are put into the pooling layer to reduce the dimensions for cascaded forests. In addition, the training cost of dgcForest is very small and it does not require a lot of computing resources. Compared with DNN, this model saves a lot of time. Gong *et al.* [31] proposed a neural network with multiscale convolution. This model takes the advantage of both

determination-point-process-based diversity-promoting deep metrics and multiscale features for effective HSI classification. Chen *et al.* [43] aimed at the problem that the handcraft network structure cannot adapt well to different datasets; they proposed the automatic CNN models called 1-D auto-CNN and 3-D auto-CNN for HSI classification. First, a search algorithm based on the gradient descent is used to efficiently find the best network structure to evaluate the performance on the validation set. Then, the best CNN architecture is selected for the HSI classification model. In addition, the author designed a new regularization method called “cutput,” which randomly deletes a certain region from the original image. Chen *et al.* [44] proposed a classification model based on the integration of deep learning model and random subspace-based ensemble learning, and used TL strategy to speed up the learning stage. Cui *et al.* [45] presented a multiscale spatial–spectral CNN network to integrate multiple receiving field fusion features and multiscale spatial features at different levels. The fused feature is developed by using the lightweight blocks of multiple reception fields, which contain various types of dilated convolutions. Li *et al.* [46] proposed a seed that uses spatial–spectral feature learning networks to reflect changes in spatial information and to learn robust adaptive features. Instead of connecting independent spatial features and spectral features, this framework combines CNN and SAE to directly extract joint spatial-spectral features from HSIs. In [47], a spatial transformation network (STN) was explored to obtain the best input for CNN-based HSI classification. The introduction of the STN network is used to translate, rotate, and scale the original image, and find optimized inputs for subsequent CNNs. In addition, in order to alleviate overfitting, the regularization technology of DropBlock is introduced to obtain better classification accuracy. Hamida *et al.* [28] proposed a 3-D CNN-based model, which uses different sizes of 3-D convolution simultaneously to process the spatial and spectral components, so as to train the model with fewer parameters. He *et al.* [48] proposed a handcrafted feature extraction method based on the multiscale covariance maps, which has better robustness and classification performance. Li *et al.* [49] proposed a robust 3-D-CapsNet model, which introduces the maximum corentropy criterion to address the noise and outlier problem. Shi and Pun [50] presented a multiscale superpixel-based RNN with SAEs for classification. Sun *et al.* [51] employed an attention mechanism to build an end-to-end spectral–spatial attention network (SSAN). Through this network, they attempted to capture the discriminative spectral–spatial features from the attention areas of HSI cube. Hang *et al.* [52] proposed a coupled CNN model for hyperspectral and LiDAR data fusion, which consists of an HSI network for spectral–spatial feature learning and a LiDAR network for elevation feature learning. Zhu *et al.* [53] presented an improved capsule network called the convolutional capsule (Conv-Capsule). Combined with principal component analysis (PCA) and extended multiattributes profile, this model effectively extracts spectral–spatial features from HSIs.

Among numerous deep learning models, CNN has demonstrated its superiority in the spectral and spatial feature extraction, showing great potential in the HSI classification field. The key technology of CNN is the local receptive field and sharing

of weights, but it also has some problems that can be further optimized. For example, during the process of gradient descent, it is easy to make the results converge to the local minimum, and the pooling layer will lose a lot of useful information. In addition, it is difficult to set the suitable hyperparameters when extracting features from a single branch network.

In CNN-based methods, using 2-D CNN alone will miss interband information. 3-D CNN extracts the joint spectral–spatial feature representation from a series of spectral bands. If different categories in HSIs have a similar texture over many spectral bands, using 3-D CNN alone seems to perform worse. However, connecting 3-D CNN and 2-D CNN with different kernels size makes the network strong enough to extract more abstract spatial representation. In addition, a multibranch network extracts HSI features at different levels, which further enriches the extracted features. Hence, the powerful feature extraction capability of 2D–3D CNN combined with the multibranch feature fusion method makes the network depend on a small number of training data. These are motivations for us to propose an HSI classification method based on the 2D–3D CNN and multibranch feature fusion. Centering on the CNN technology, the proposed method combines 2-D CNN and 3-D CNN with different kernels size to design three branches of the neural network, which extracts features at different levels. Then, the three different features are fused in the spectral dimension. Finally, the fused features are fed into fully connected layers that are added at the end of the model in order to perform the final classification task. Moreover, the new activation function Mish is employed in the network, and 5% of the data are used for training.

The main contributions of this article are summarized as follows.

- 1) A deep network structure based on the 2D–3D CNN and multibranch feature fusion is proposed for HSI classification.
- 2) The proposed model incorporates network outputs at different levels, which effectively improve the classification performance.
- 3) A new activation function Mish is used to replace Relu.

The remaining part of this article is organized as follows. Section II describes the related theoretical basis. Section III introduces the proposed neural network model and the specific implementation steps of the model. The experimental results and analysis are provided in Section IV. Finally, Section V gives the conclusion.

II. RELATED THEORY

A. Convolutional Neural Network

CNN is a feedforward neural network. Its artificial neurons respond to a part of the surrounding cells in the coverage area, and it has excellent performance for large-scale image processing [54], [55]. CNN contains two core design ideas. One is that CNN uses the 2-D structure of the image, that is, the pixels in adjacent areas are usually highly correlated. The other is that its architecture relies on feature sharing, so each channel (that is, the output feature map) is generated by convolution using the same filter in all locations. CNN is mainly composed of four

parts: convolutional layer, activation function, pooling layer, and fully connected layer.

The calculation function of the convolutional layer is shown as

$$y_j^l = \sum_{i=1}^d f(x_i^{l-1} * w_{ij}^l + b_j^l) \quad (1)$$

where the matrix x_i^{l-1} is the i th feature map of the $l-1$ layer, y_j^l is the j th feature map of the l th layer, and d represents the number of the input feature maps. w_{ij}^l and b_j^l are the weight and bias parameters, which are randomly initialized. $f(\cdot)$ indicates a nonlinear activation function, where we utilize Mish in this article. The symbol $*$ represents the convolution operation.

The pooling layer is usually located after the convolutional layer to reduce the spatial dimension of the data. A pooling operation is reducing the number of parameters in the network and effectively preventing overfitting.

The final part of the network usually connects several fully connected layers. Each neuron in a fully connected layer connects with all neurons in the previous layer and sends the output value to the classifier.

All parameters in neural networks are trained by using the backpropagation algorithm. CNN is flexibly combining the convolutional layer, pooling layer, and fully connected layer together so that the network can accurately extract image features.

CNN-based HSI classification methods mainly include 2-D CNN and 3-D CNN. Roy *et al.* [41] used three 3-D CNN and a 2-D CNN to build the network. Hamida *et al.* [28] mainly utilized 3-D CNN in the network. Combining 2-D CNN with 3-D CNN can indeed improve the performance to a certain extent, but selecting the appropriate number of network layers is not easy to control. Because neither too deep nor too shallow networks can well extract image features. This article uses 2-D CNN and 3-D CNN with different kernels size to construct a multibranch network model, which obtains richer and more diverse features.

B. Feature Fusion

In many tasks, feature fusion at different scales is an important means to improve classification performance. Low-level features have large spatial size, which contain more location and detailed information. Due to the use of fewer convolutional layers, there are little semantics and much noise in the features. The high-level features have stronger semantic information, but the features have a small spatial size and poor ability to perceive details. How to effectively integrate the two methods, take the advantage of them, and abandon the bad ones is the key to improve model performance [56].

The feature fusion methods can be divided into early fusion and late fusion according to the order of fusion and prediction. Among them, early fusion methods first fuse multilayer features, then they train a predictor on the fused features. This type of method is also called skip connection, which uses the concat or add operations. Late fusion methods improve the performance by fusing the detection results of different layers. In this article, we use concat for feature fusion.

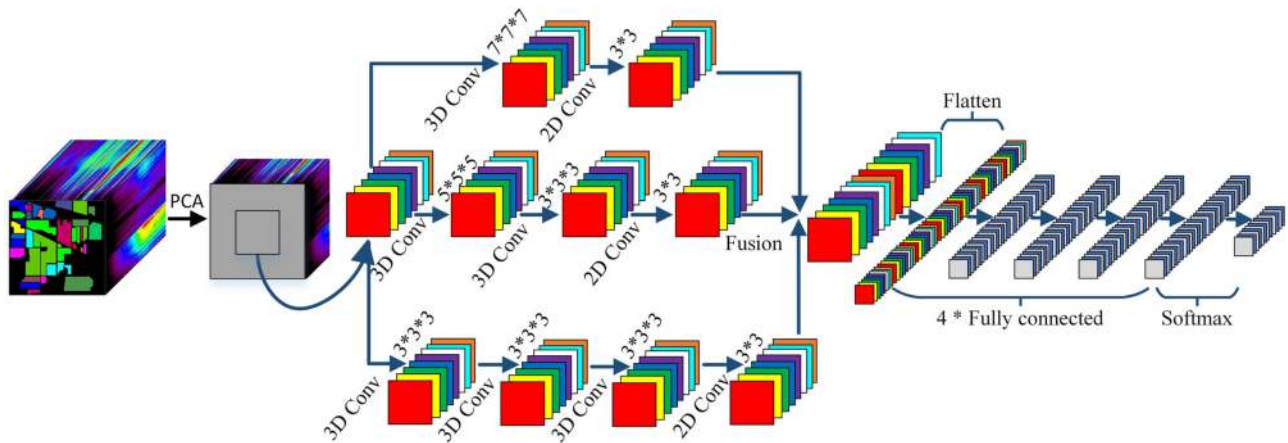


Fig. 1. Flowchart of the proposed HSI classification model.

III. PROPOSED NEURAL NETWORK MODEL

The flowchart in Fig. 1 depicts the proposed model for HSI classification using 2D–3D CNN and multibranch feature fusion. This model first uses PCA to reduce the dimension of the original dataset. And then it inputs the reduced data into the three-branch network to extract features at different levels and fuses the three features by concat. Finally, classification is performed through several fully connected layers and a softmax layer.

A. PCA Processing of HSI Data

For HSI analysis, researchers demonstrated that the redundancy from interband correlation is very high. The data structure in the spectral dimension can be reduced without the significant loss of useful information for subsequent utilization [57]. However, an HSI contains hundreds of spectral bands, which increases the pressure on the network model to process data and also consumes a lot of computing resources.

Recent years, many studies on HSI classification use PCA for data preprocessing [49], [50], [52], [56]. PCA is the most commonly used linear dimensionality reduction method. Its goal is mapping a high-dimensional data to a corresponding low-dimensional data through some linear projection, that is, maximizing the variance. This method reduces the data dimension while retaining more original data features. The core idea of PCA is calculating the similarity between different data features, extracting the main features according to the strength of the correlation, and completing the information fusion [58], [59]. And hence, PCA is applied to the original HSI for dimensionality reduction in the proposed method.

We record the cube data of the original HSI as $H^{m,n,l}$, where H represents the original HSI data, m is the length, n is the width, and l is the number of spectral bands. Each pixel in HSI is composed of a pixel sequence $S = \{s_1, s_2, s_3, \dots, s_l\}$. HSIs contain a large number of spectral bands, which may cause a huge computational burden. Furthermore, there are plenty of redundant data in HSI. In the proposed method, we perform PCA processing on the original HSI. After applying PCA, the

original HSI data $H^{m,n,l}$ is reduced to $H^{m,n,p}$, where the length and width of the image are unchanged, the depth is reduced from l to p .

B. Multibranch CNN

Due to the high correlation among the bands of HSIs, a large number of samples are needed for training in the classification task. However, it is difficult to obtain large numbers of ground reference data because such data collection is expensive and complex. Limited samples in the training set will lead to unreliable training parameters and result in overfitting. Therefore, getting better classification performance with less training data is a necessary aspect we concern in the proposed network model. In addition, shallow neural network extracts the shallow edge information of the image. It has poor ability to express complex function and its generalization ability is also limited. By learning a deep nonlinear network structure, the deep neural network can realize complex function approximation, characterize the distributed representation of input data, and demonstrate a powerful ability to learn the characteristics of dataset from a few training samples.

At present, most classification models only use 2-D CNN or 3-D CNN for feature extraction. Although 2-D CNN has the ability to extract the spatial information, it neglects the rich interband information of HSIs. The excessive use of 3-D CNN will make the network too bloated, which may decrease the classification accuracy. In order to mitigate these issues, 2-D CNN and 3-D CNN are combined to build the multibranch feature fusion network in this article. We design three neural network branches by connecting 2-D CNN to 3-D CNN, where each branch has different numbers of network layers. And the first convolutional layer in three branches has a different kernel size. Through the three branches, we extract three different features and fuse them together to obtain a more discriminative feature representation.

In the proposed network model, a new activation function Mish is used to further improve the performance. Currently, Relu is a widely used activation function in most HSI classification tasks or even in most deep learning fields. However, a new deep

TABLE I
CONFIGURATION OF DEEP NETWORK USED IN FEATURE LEARNING PROCEDURE

Branch	Layer	Input shape	Output shape	Kernel-size	Filters	Connected to
1	Input_1	-	(25,25,30,1)	-	-	-
	conv3d1_1	(25,25,30,1)	(19,19,24,6)	7×7×7	6	Input_1
	reshape1	(19,19,24,6)	(19,19,144)	-	-	conv3d1_1
2	conv2d1	(19,19,144)	(17,17,80)	3×3	80	reshape1
	Input_2	-	(25,25,30,1)	-	-	-
	conv3d2_1	(25,25,30,1)	(21,21,26,6)	5×5×5	6	Input_2
3	conv3d2_2	(21,21,26,6)	(19,19,24,12)	3×3×3	12	conv3d2_1
	reshape2	(19,19,24,12)	(19,19,288)	-	-	conv3d2_2
	conv2d2	(19,19,288)	(17,17,64)	3×3	64	reshape2
3	Input_3	-	(25,25,30,1)	-	-	-
	conv3d3_1	(25,25,30,1)	(23,23,28,8)	3×3×3	8	Input_3
	conv3d3_2	(23,23,28,8)	(21,21,26,16)	3×3×3	16	conv3d3_1
	conv3d3_3	(21,21,26,16)	(19,19,24,32)	3×3×3	32	conv3d3_2
	conv2d3	(19,19,24,32)	(19,19,768)	-	-	conv3d3_3
	conv2d3	(19,19,768)	(17,17,80)	3×3	80	reshape3

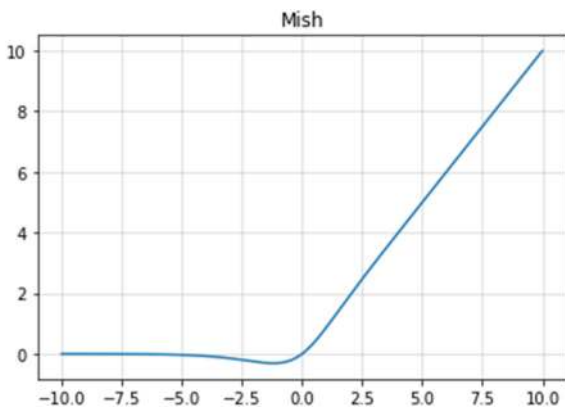


Fig. 2. Image of the activation function Mish.

learning activation function Mish is proposed in [60], which has the accuracy improvement of 1.671% over Relu in the test. The calculation equation of Mish is shown as follows:

$$\text{Mish} = x * \tanh(\ln(1 + e^x)) \quad (2)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3)$$

where x is the input of the function, $\ln(\cdot)$ is a logarithmic calculation, and the calculation equation of $\tanh(\cdot)$ is shown in (3). Fig. 2 displays the image of the Mish function.

Being unbounded above is a desirable attribute of Mish because it avoids saturation, and a slight allowance for negative values, in theory, leads to better gradient flow instead of a strict zero boundary in Relu. Meanwhile, Mish is a continuous function, which guarantees better gradient optimization. The above characteristics ensure that the model using Mish has better classification accuracy and generalization performance.

The detailed distribution and parameters of each layer in the proposed model are described in Table I. As given in Table I, the first branch combines one layer of 3-D CNN and one layer of 2-D CNN to extract shallow features. In this branch, the kernel size of 3-D CNN is $7 \times 7 \times 7$ and the number of kernels is 6, and the kernel of 2-D CNN is 3×3 and the number of kernels is 80. The second branch adopts two layers of 3-D CNN with 6.5×5

$\times 5$ convolution kernels, $12.3 \times 3 \times 3$ convolution kernels, and one layer of 2-D CNN with 64.3×3 convolution kernels. The third branch is designed to extract deeper semantic features. In this branch, we use three layers of 3-D CNN, which have $3 \times 3 \times 3$ convolution kernels with 8, 16, and 32 filters, respectively, and a layer of 2-D CNN with 80.3×3 convolution kernels. The feature cubes extracted from three network branches have the same length and width, only different depths. Based on the length and width of the image, the three features are fused to obtain the mixed feature of the image. Considering that the pooling layer will lose some information, we do not use this layer in the network. In addition, the network structure and hyperparameters we describe above satisfy the four datasets at the same time.

IV. EXPERIMENT AND DISCUSSION

A. Data Description

Four real HSI datasets have been considered in our experiments: The University of Pavia (UP), Indian Pines (IP), Salinas (SA), and Botswana (BOT). We show the false-color composite, ground reference maps, as well as the land cover categories of these datasets in Figs. 3–6.

The first dataset UP, was captured by the reflective optics system imaging spectrometer sensor, acquired in 2001 over the UP in northern Italy. The spatial size of the image is 610×340 with a high resolution of 1.3 m per pixel, and the spectral information consists of 103 bands in the wavelength ranging from 0.43 to $0.86 \mu\text{m}$. The number of different categories contained in the UP scene is 9.

The second dataset IP was captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor in 1992. The spatial size of the scene is 145×145 with a high resolution of 20 m per pixel, and the spectral information consists of 200 bands in the wavelength ranging from 0.4 to $2.5 \mu\text{m}$, after removing 20 water absorption bands. The ground truth of the IP scene contains a total of 16 different categories.

The third dataset SA was captured by the AVIRIS sensor in 1998 over the agricultural area described as SV in California, CA, USA. The spatial size of the data is 512×217 with a high resolution of 3.7 m per pixel, and the spectral information

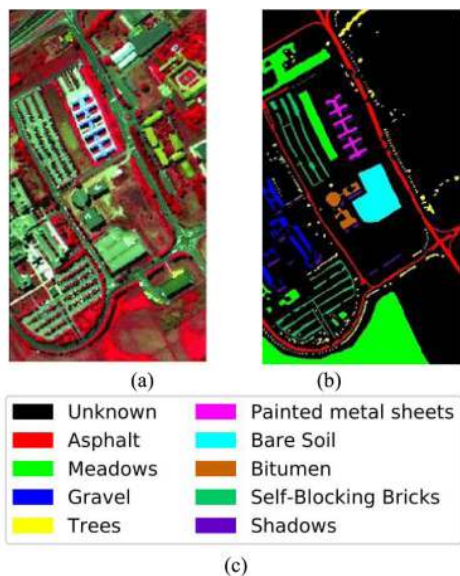


Fig. 3. UP dataset. (a) False-color composite. (b) Ground reference map. (c) Land cover classes in the UP dataset.

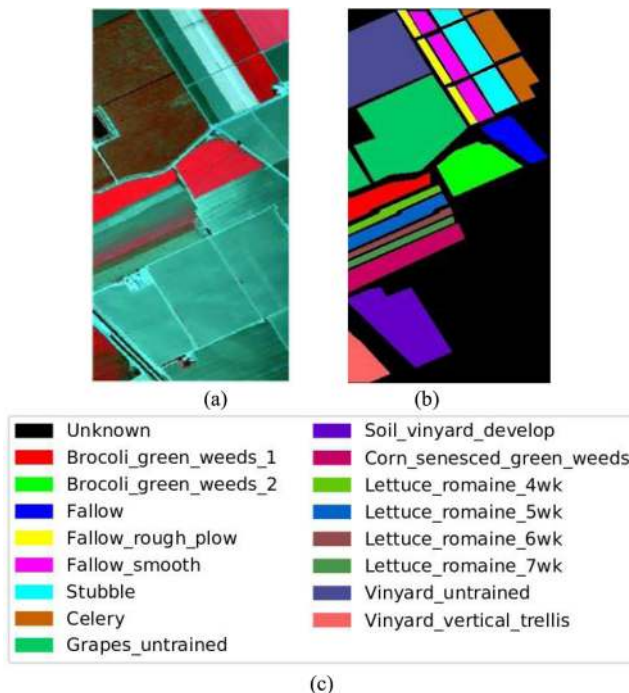


Fig. 5. SA dataset. (a) False-color composite. (b) Ground reference map. (c) Land cover classes in the SA dataset.

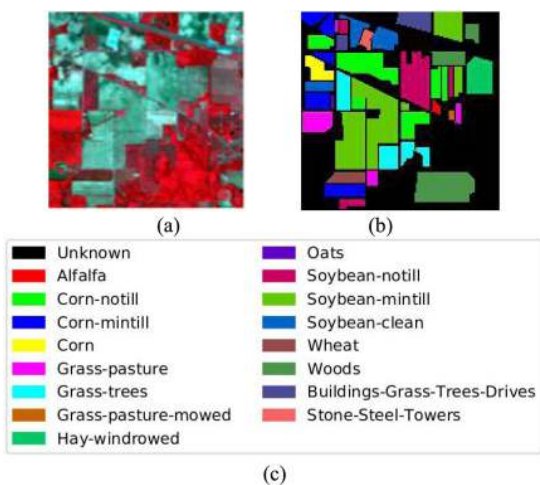


Fig. 4. IP dataset. (a) False-color composite. (b) Ground reference map. (c) Land cover classes in the IP dataset.

consists of 224 bands in the wavelength ranging from 0.4 to 2.5 μm . The available ground truth for the SA contains 16 categories.

The fourth dataset BOT was captured by the NASA EO-1 satellite with the Hyperion sensor over the Okavango Delta, BOT, South Africa, on May 31, 2001. The spatial size of the scene is 1476×256 with a high resolution of 30 m per pixel, and the spectral information consists of 145 bands in the wavelength ranging from 400 to 2500 nm, after removing uncalibrate and noise bands. The ground truth of the BOT scene contains a total of 14 different categories.

We ran out a test on a computer with a ninth-generation Intel Core i7 CPU at 3.0 GHz with the RTX 2080ti graphical processing unit. The operating system is window 10 (64 bit) home, and the experimental platform is python 3.7. We choose the optimal learning rate of 0.0005 based on the classification outcomes.

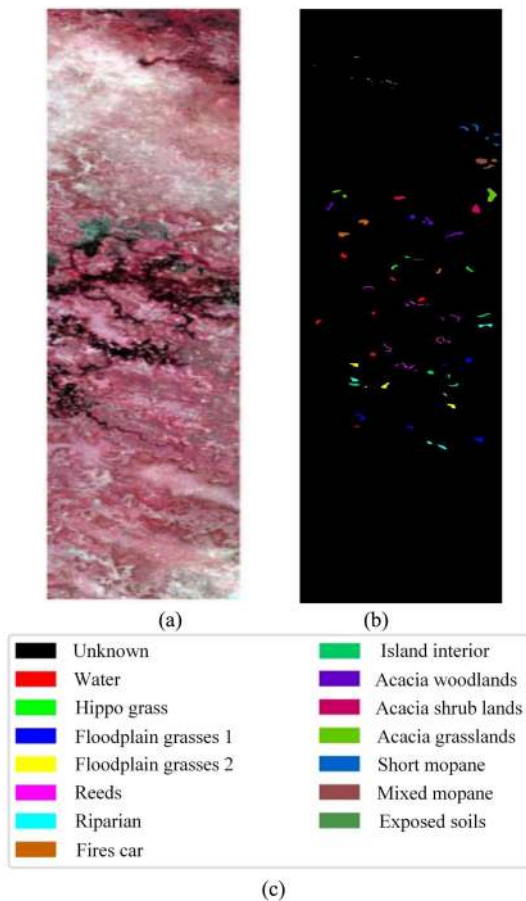


Fig. 6. BOT dataset. (a) False-color composite. (b) Ground reference map. (c) Land cover classes in the BOT dataset.

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY BASED ON 5% SAMPLES FOR TRAINING

Method	University of Pavia			Indian Pines			Salinas			Botswana		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
M3D-DCNN [39]	93.28	93.74	92.01	80.08	71.33	73.79	93.46	93.96	92.05	90.83	89.32	90.06
SSRN [27]	98.99	98.23	99.03	93.46	86.72	92.12	99.37	99.67	99.71	95.79	95.28	95.43
HybirdSN [41]	98.96	97.86	98.63	93.69	87.94	92.80	99.63	99.78	99.59	94.45	94.74	95.81
1-D Auto-CNN [43]	84.86	84.54	82.98	73.28	61.28	66.55	90.26	90.57	89.67	82.44	79.52	80.95
3-D Auto-CNN [43]	94.11	93.93	95.98	92.08	85.68	88.27	96.14	95.62	96.79	93.84	93.49	93.32
3D CNN [28]	92.88	90.70	92.56	71.63	73.39	68.07	92.93	93.50	92.26	85.99	85.94	84.83
MCMs+2DCNN [48]	98.78	96.84	98.38	94.20	95.66	95.05	98.78	96.84	98.38	95.38	94.28	95.43
3D-Conv-Capsule [53]	97.24	96.71	97.05	93.88	93.67	93.31	97.70	96.93	97.25	95.51	94.76	95.34
SVM [61]	93.12	90.75	90.84	62.45	60.30	55.73	91.82	95.43	90.89	85.93	93.29	84.72
Proposed method	99.52	99.17	99.41	96.07	94.14	95.51	99.94	99.92	99.93	96.44	96.80	96.14

For all four datasets, we split the labeled samples into two subsets, i.e., training sets and testing sets. We randomly select 5% of the original dataset for training, and the remaining 95% of the data are used as a test. In the training set, we use threefold cross validation to determine the hyperparameters. In order to assess the results, three widely used quantitative metrics are utilized to evaluate the classification performance.

- 1) Overall accuracy (OA): The percentage of correctly classified pixels.
- 2) Average accuracy (AA): The mean value of the OAs measured over each category.
- 3) Kappa coefficient (Kappa): A statistic measurement over the inter-rate agreement among qualitative items.

B. Classification Results

In this section, we report the quantitative and qualitative results of the proposed method and compare it with the state-of-the-art methods within 5% of training data. The classification results are given in Table II, where we mark the data of methods with the best performance under different conditions in bold.

As displayed in Table II, the proposed method gets better results in OA, AA, and Kappa compared with the listed methods in four different HSIs. Among these datasets, UP has only nine categories, which makes it easier to classify than the other three datasets. SA has a large spatial size and the maximum numbers of spectral bands, so the classification performance of the listed methods on SA is higher. The spatial size of IP is relatively small but with 16 categories, which results in lower classification accuracy. Although BOT has the largest spatial size in all HSIs, it provides the least samples with ground truth. The classification accuracy in BOT is just higher than IP.

In addition to SVM [61], other methods all combine 1-D CNN, 2-D CNN, or 3-D CNN to extract the spectral and spatial features of HSI for classification. Among them, inspired by ResNet, SSRN [27] designs a deeper network model. By introducing skip connections, the method effectively prevents overfitting, and the parameters can get well backpropagation. HybridSN [41] uses several 3-D CNNs to connect a 2-D CNN. The network is simple and effective, and the extracted features have strong discrimination. These two methods make good use of the joint spectral-spatial features of HSI, and the OA of UP, IP, SA, and BOT reaches a high level of 98.9%, 93.4%,

99.3%, and 94.4%, respectively. MCMs+2-D CNN [48] calculates the multiscale covariance map of HSI and inputs it into the 2-D CNN for classification. By using a covariance map, the spatial and spectral information are simultaneously exploited. Considering that this model has a very shallow network, there are no deeper features been extracted, so the classification performance of this method is lower than the previous two. The 3D-Conv-Capsule [53] introduces the idea of capsule network into HSI classification, which combines the advantages of CNN local connections and shared transform matrices with the characteristics of a dynamic routing algorithm in the capsule layer. The OA of the network is close to MCMs+2-D CNN, which are 97.2%, 93.8%, 97.7%, and 95.5% in the four datasets, respectively. The 1-D auto-CNN [43] and 3-D auto-CNN [43] emphasize an automatic idea. These two methods find a group of hyperparameters with the best performance on the validation set from several candidate hyperparameters. Among them, 1-D auto-CNN only uses the spectral information of HSI, and the improved 3-D auto-CNN combines the spectral and spatial information of HSI for classification. We can clearly see that the classification performance of 3-D auto-CNN significantly outperforms 1-D auto-CNN in the four datasets. The methods of M3D-DCNN [39] and 3-D CNN [28] are relatively similar. They both use 3-D CNN as the main component in the network. In comparison, M3D-DCNN uses multiscale convolution kernels to extract more abundant features. Therefore, the classification performance of M3D-DCNN is slightly higher than 3-D CNN. The classification performance gap between the two methods is more obvious in IP, which is 8.45% of OA. These two methods have simple network models and shallow network layers, which cannot fully extract the features with discriminative ability. Therefore, their classification performance is poor compared with the aforementioned methods. As a general comment, the improvement introduced by spectral-spatial models is remarkable. As an illustration, SVM [61] only relies on the spectral features, which results in weak classification performance. Compared with other spectral-spatial methods, the spectral information alone appears not enough discriminative to carry out the accurate classification. As observed in Table II, the classification results of SVM have the greatest number of mistaken pixels.

The limitations of the spectral-based and spatial-based methods can be easily overcome by combining the spectral and spatial

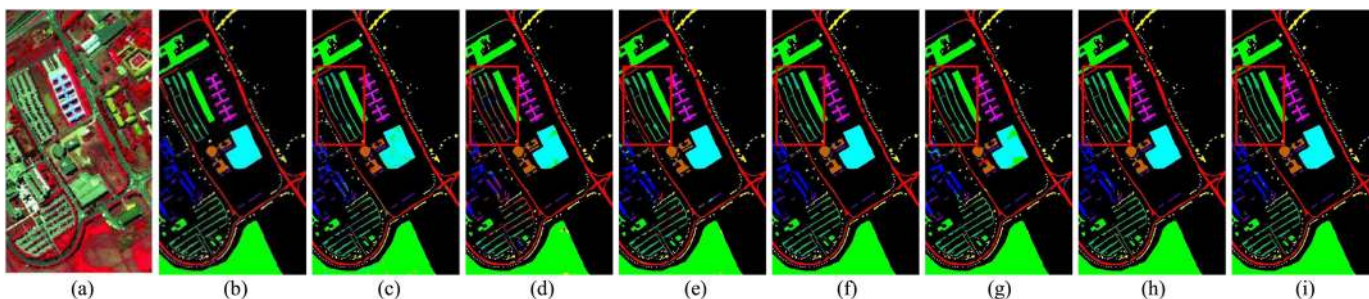


Fig. 7. Classification map of UP. (a) False-color image. (b) Ground truth. (c) SVM. (d) 1-D auto-CNN. (e) 3-D auto-CNN. (f) HybridSN. (g) M3D-DCNN. (h) SSRN. (i) Proposed method.

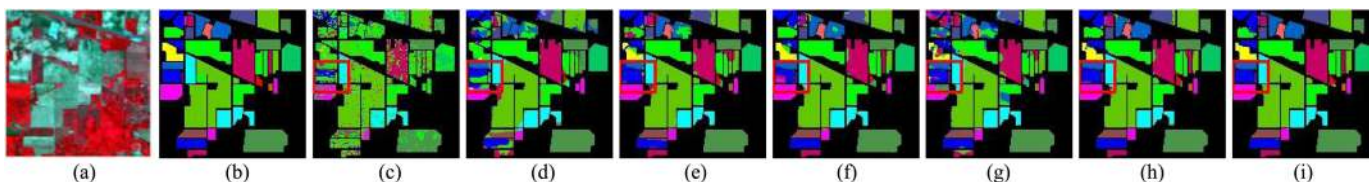


Fig. 8. Classification map of IP. (a) False-color image. (b) Ground truth. (c) SVM. (d) 1-D auto-CNN. (e) 3-D auto-CNN. (f) HybridSN. (g) M3D-DCNN. (h) SSRN. (i) Proposed method.

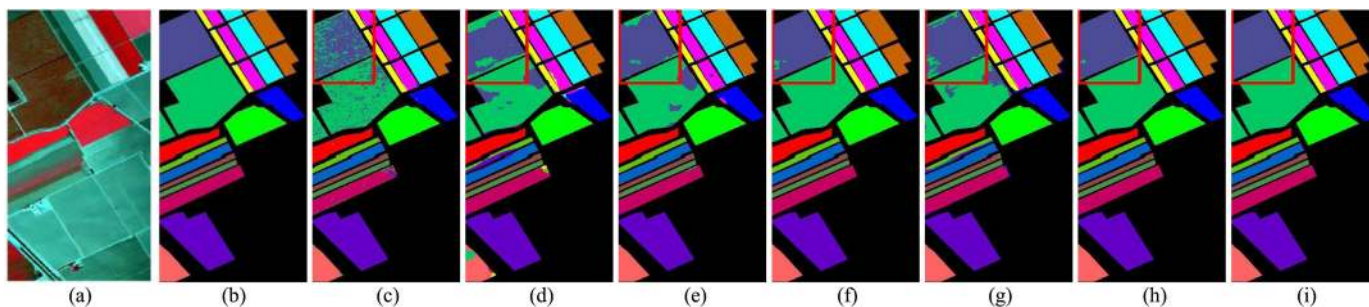


Fig. 9. Classification map of SA. (a) False-color image. (b) Ground truth. (c) SVM. (d) 1-D auto-CNN. (e) 3-D auto-CNN. (f) HybridSN. (g) M3D-DCNN. (h) SSRN. (i) Proposed method.

contextual information, where the combination of 2-D CNN and 3-D CNN is able to significantly reduce the uncertainty and data variability of image pixels, such as HybridSN method, which has the best performance in the whole classification of the four datasets. The proposed method not only extracts the spatial and spectral features of HSI but also obtains different image features from shallow to deep through multiple network branches. Through fusing them together, the network learns the representative and discriminative features, and the extracted features achieve small intraclass differences and large interclass differences. Compared with the best method HybridSN, our proposed method improves the OA by more than 0.56%, 2.38%, 0.31%, and 1.99% in four datasets, respectively. The proposed method generates very similar OA, AA, and Kappa values in all cases, exhibiting better consistency in terms of model performance on the obtained results.

In addition to the quantitative report, we also visualize classification maps of different methods. As we can observe in Figs. 7–10, the contrast areas are marked with red boxes.

The proposed method is able to achieve the best performance in all considered datasets, visually clean classification maps, where the number of mistaken pixels is obviously reduced. The spectral-based methods always result in noisy scatter points in the classification maps [see Figs. 7(c) and (d), 8(c) and (d), 9(c) and (d), and 10(c) and (d)], while the spectral–spatial based methods can well overcome the shortcoming. Obviously, such as 3D-Conv-Capsule, HybridSN, and so on, which directly use the neighbor information, as the model input results in smoother classification maps. Comparing the true ground reference with the classification maps, the proposed method achieves the best performance in all considered datasets. The experimental results demonstrate that our proposed method makes the rich interband information of HSIs better serve the classification task, and the extracted features can distinguish different types of pixel information well.

The comparisons of accuracy and loss convergence for 100 epochs of training and validation sets are portrayed in Fig. 11. It can be seen that the proposed method converges faster

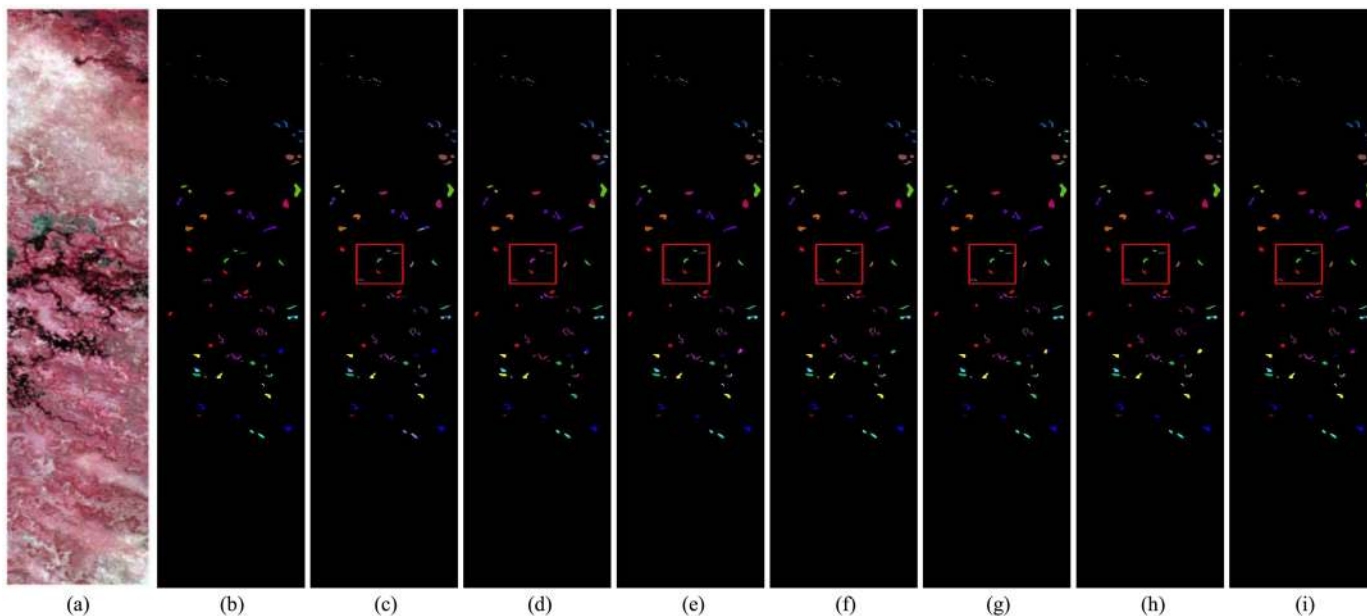


Fig. 10. Classification map of BOT. (a) False-color image. (b) Ground truth. (c) SVM. (d) 1-D auto-CNN. (e) 3-D auto-CNN. (f) HybridSN. (g) M3D-DCNN. (h) SSRN. (i) Proposed method.

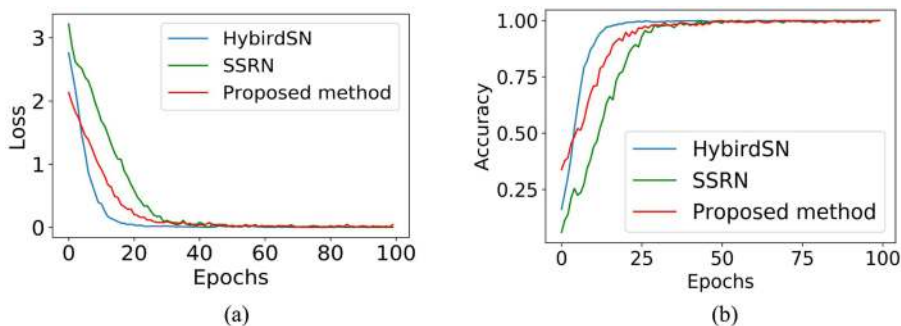


Fig. 11. Comparison of the loss and accuracy of each epoch.

than the SSRN but slower than the HybridSN. The proposed method and SSRN converge at about 50 epochs, while the HybridSN converges at about 30 epochs. HybridSN method obtains the fastest convergence for its simple network structure, which consists of three 3-D convolutional layers and a 2-D convolutional layer. The SSRN method has a deep network to learn a large number of parameters, which leads to the slowest convergence. Although the proposed method does not use a very deep network, the model is composed of multiple network branches. Compared with HybridSN, the proposed method needs to learn more parameters, which causes slower convergence.

The comparison of confusion matrices on IP is presented in Fig. 12 in order to further demonstrate the performance of the proposed method. Although the classification accuracy of the three methods is higher than 90%, they also demonstrate the confusion made by the trained network. For instance, in case of the HybridSN method, the Alfalfa pixels are mixed up with the Hay-windrowed pixels with almost 50% mistaken pixels among all the classified ones. In addition, the Corn-mintill pixels, the

Grass-pasture-mowed pixels, and several other categories have different degrees of confusion. Similarly, the Alfalfa, Corn, Buildings-Grass-Trees-Drives, Stone-Steel-Towers pixels, and other categories have mistaken pixels in the SSAN method. In the proposed method, only the mistaken pixels of the first categories are relatively high (almost 35%), where the Alfalfa category is confused with Grass pasture, the mistaken pixels in other categories are very low. Therefore, the proposed method shows better classification performance.

In order to determine the number of principal components in PCA, we test the classification accuracy on the four datasets after the dimensionality reduction with different principal components. Among them, due to the large spatial scale of UP, SA, and BOT, only 10, 15, and 20 are considered, while seven cases from 10 to 40 are tested in IP. The classification accuracy with different numbers of principal components among the four datasets is shown in Fig. 13.

In Fig. 13, according to the final classification accuracy, we set the number of reduced spectral bands to 15, 30, 15, and 15 for UP, IP, SA, and BOT, respectively.

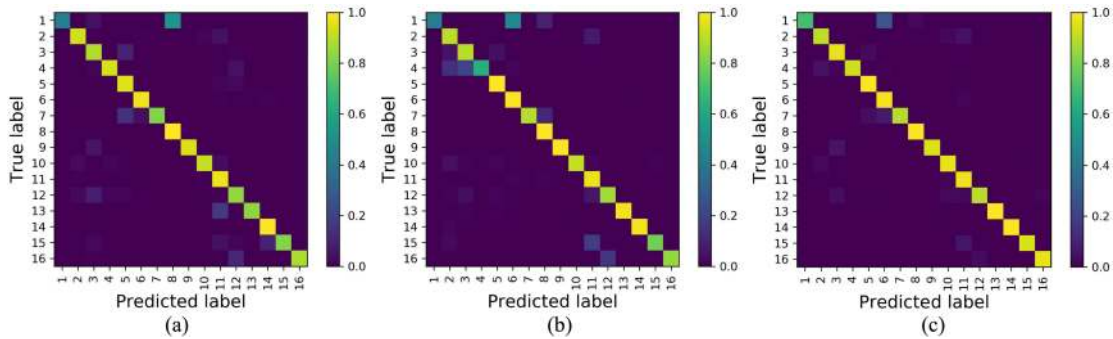


Fig. 12. Comparison of a confusion matrix for classification results on IP dataset. (a) HybridSN. (b) SSRN. (c) Proposed method.

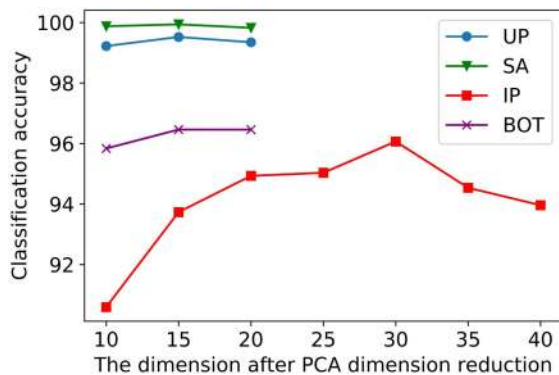


Fig. 13. Classification accuracy versus the number of principal components among three datasets.

C. Comparison of Different Methods With Different Amounts of Training Data

The classification accuracy under different quantities of training data can well reflect the performance of the model's representation ability. We randomly select 3%, 5%, 8%, 10%, 15%, and 20% of labeled pixels for training and test on the rest of the data. The classification results are shown in Fig. 14.

In this figure, it can be seen that as the number of training samples increases, the classification accuracy of almost all methods is improved. Compared with HybridSN, SSRN utilizes skip connection to build a deeper network structure and gets better classification accuracy. The proposed method combines 2D–3D CNN with different kernels size to extract different levels of fusion features from the original HSIs, which performs the best on all training sample proportions and four datasets.

D. Comparison of the Accuracy of the Proposed Model Using 2-D CNN or 3-D CNN Alone

In order to validate the effectiveness of the combination of 2-D CNN and 3-D CNN in the proposed method, two extra models are designed for comparison. We first replace 2-D CNN in three branches of the proposed model with 3-D CNN, that is, 3-D CNN alone. Then, we make the output sizes of the three network

TABLE III
COMPARISON OF OA OF ACTIVATION FUNCTIONS MISH AND RELU ON DIFFERENT DATASETS

Method	University of Pavia	Indian Pines	Salinas	Botswana
Relu	97.6797	93.8477	98.1594	95.0182
Mish	99.5226	96.0666	99.9397	96.4355

branches basically remain the same. Another model using the 2-D CNN alone is simultaneously designed in a similar way. The comparison results are shown in Fig. 15.

In Fig. 15, with different numbers of training data in four datasets, the proposed 2D–3D CNN model always shows the best classification performance. The classification accuracy of 2-D CNN alone and 3-D CNN alone is closer in most cases, but 2-D CNN alone is slightly lower. Compared with 3-D CNN, 2-D CNN ignores the information between the adjacent spectral bands of HSI. The lack of interband correlations is directly reflected in its lower classification accuracy. The experimental results show that the proposed model extracts more discriminative features through the combination of 2-D CNN and 3-D CNN, which significantly improve the final classification performance.

E. Comparison of the Influence of Mish and Relu on Classification Accuracy

This article uses the new activation function Mish in the network, which slightly allows the negative values and has no strict zero boundary. Meanwhile, Relu is the most commonly used activation function in deep learning-based HSI classification tasks. The comparison of the two activation functions is given in Table III.

It can be seen from Table III that after using Mish instead of Relu, the classification accuracy is improved to a certain extent. In UP, IP, SA, and BOT, the accuracy has been increased by 1.8429%, 2.2189%, 1.7803%, and 1.4253%, respectively, with an average increase of 1.8169%. Being a nonmonotonic and overall smoothness activation function, Mish shows its capability in improving the classification accuracy and generalization ability, which is a good choice to replace Relu.

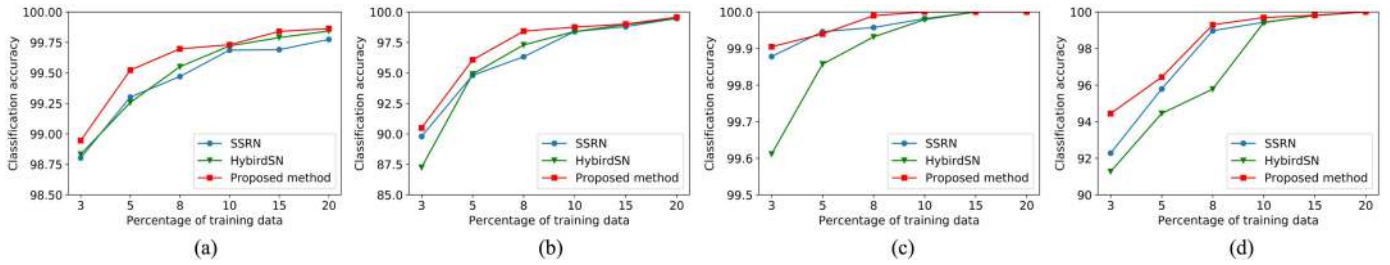


Fig. 14. Classification accuracy under different numbers of training data. (a) UP. (b) IP. (c) SA. (d) BOT.

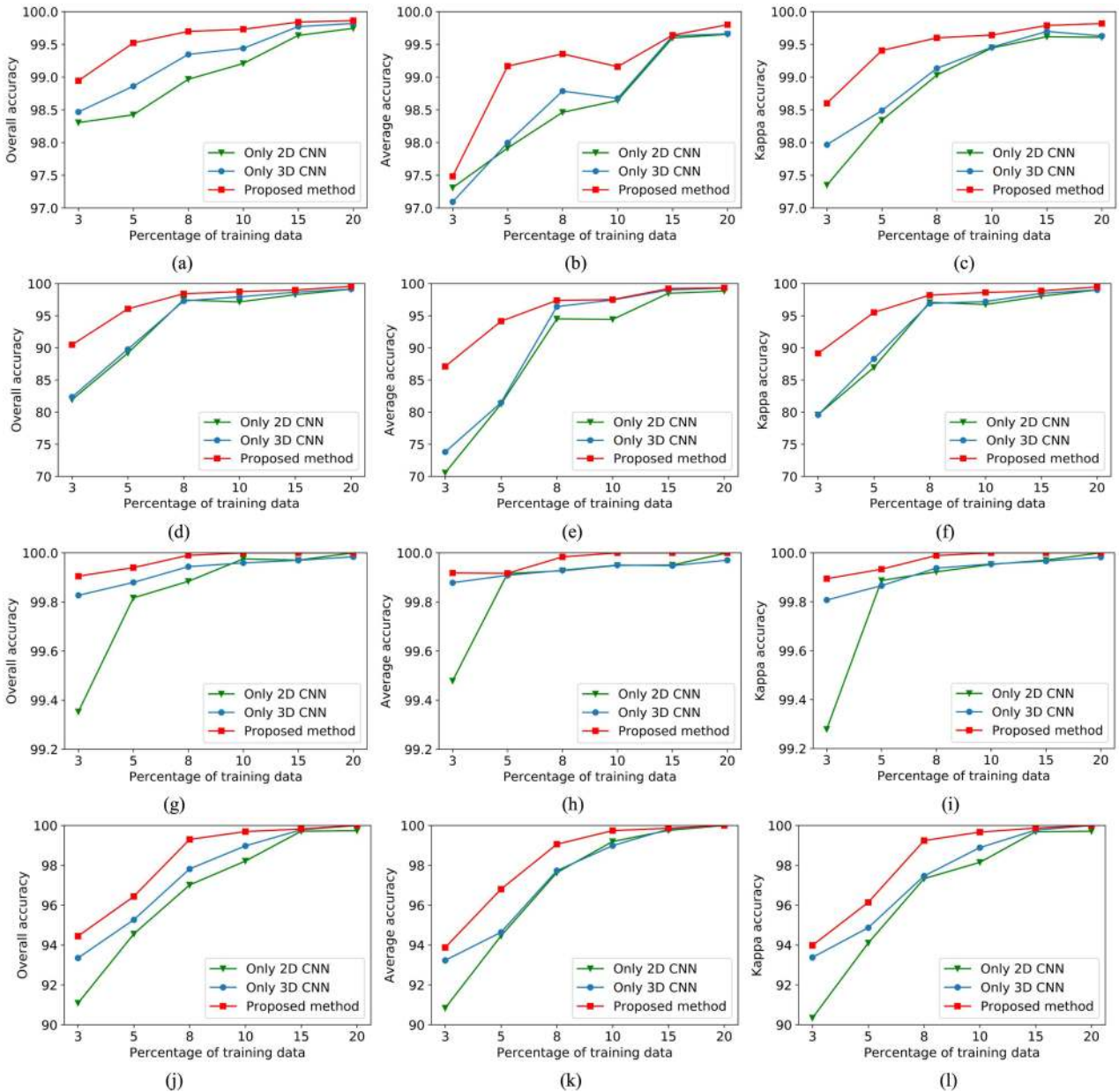


Fig. 15. Comparison of classification performance among the proposed method using 2D-3D CNN, using only 3-D CNN, and using only 2-D CNN. (a) OA comparison of UP. (b) AA comparison of UP. (c) Kappa comparison of UP. (d) OA comparison of IP. (e) AA comparison of IP. (f) Kappa comparison of IP. (g) OA comparison of SA. (h) AA comparison of SA. (i) Kappa comparison of SA. (j) OA comparison of BOT. (k) AA comparison of BOT. (l) Kappa comparison of BOT.

V. CONCLUSION AND FUTURE WORK

In this article, we propose an HSI classification method that combines 2D–3D CNN with multibranch feature fusion. Aiming at the existing problems of HSI classification, the proposed method optimizes the construction of the CNN-based model, establishes a multibranch feature fusion structure, and uses a new activation function Mish. The model first passes the data into three neural network branches after PCA, fuses the obtained features extracted from the three network branches, and then outputs the classification results through several fully connected layers. The experimental results show that the proposed method achieves high performance in OA, AA, and Kappa. It obtains relatively ideal classification results in a small number of training data. In addition, the new activation function Mish has better performance in terms of training stability and accuracy than Relu.

The shortcoming of the proposed method is that the classification accuracy of the model in IP and BOT is slightly lower than the other two datasets. And the fully connected layers we used in the neural network have too much training parameters. Therefore, exploring a more concise neural network model, improving the generalization ability, and obtaining satisfactory classification accuracy on various datasets are our future work.

REFERENCES

- [1] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.
- [2] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, 2019, Art. no. 3071.
- [3] T. A. Carrino, A. P. Crósta, C. L. B. Toledo, and A. M. Silva, "Hyperspectral remote sensing applied to mineral exploration in southern Peru: A multiple data integration approach in the Chapi Chiara gold prospect," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 287–300, Feb. 2018.
- [4] S. Lorenz *et al.*, "Radiometric correction and 3D integration of long-range ground-based hyperspectral imagery for mineral exploration of vertical outcrops," *Remote Sens.*, vol. 10, no. 2, Jan. 2018, Art. no. 176.
- [5] C. Camino, V. González-Dugo, P. Hernández, J. C. Sillero, and P. J. Zarco-Tejada, "Improved nitrogen retrievals with airborne-derived fluorescence and plant traits quantified from VNIR-SWIR hyperspectral imagery in the context of precision agriculture," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 70, pp. 105–117, 2018.
- [6] F. A. Rodrigues, Jr., *et al.*, "Multi-temporal and spectral analysis of high-resolution hyperspectral airborne imagery for precision agriculture: Assessment of wheat grain yield and grain protein content," *Remote Sens.*, vol. 10, no. 6, Jun. 2018, Art. no. 930.
- [7] W. Li *et al.*, "Stacked autoencoder-based deep learning for remote-sensing image classification: A case study of African land-cover mapping," *Int. J. Remote Sens.*, vol. 37, no. 23, pp. 5632–5646, Mar. 2016.
- [8] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, vol. 5, pp. 9021–9031, 2017.
- [9] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.
- [10] M. E. Midhun, S. R. Nair, V. T. N. Prabhakar, and S. S. Kumar, "Deep model for classification of hyperspectral image using restricted Boltzmann machine," in *Proc. Int. Conf. Interdiscip. Adv. Appl. Comput.*, Amritapuri, India, 2014, pp. 1–7.
- [11] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, no. 12, 2015, Art. no. 258619.
- [12] J. Yang, Y. Zhao, J. C.-W. Chan, and C. Yi, "Hyperspectral image classification using two-channel deep convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Beijing, China, 2016, pp. 5079–5082.
- [13] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognit.*, vol. 63, pp. 371–383, Mar. 2017.
- [14] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.
- [15] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
- [16] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, Mar. 2017, Art. no. 298.
- [17] 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, Dec. 2017, Art. no. 1330.
- [18] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
- [19] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "Caps-tripleGAN: GAN-assisted CapsNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7232–7245, Sep. 2019.
- [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. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440–6461, Nov. 2018.
- [22] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.
- [23] 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.
- [24] L. Fang, Z. Liu, and W. Song, "Deep hashing neural networks for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1412–1416, Sep. 2019.
- [25] H. Liang and Q. Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sens.*, vol. 8, no. 2, pp. 99–114, Jan. 2016.
- [26] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.
- [27] 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.
- [28] 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.
- [29] X. Li, Q. Sun, L. Li, Z. Ren, F. Liu, and L. Jiao, "Pixel Dag-recurrent neural network for spectral-spatial hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, 2019, pp. 2718–2721.
- [30] C. Dong, M. Naghedolfeizi, D. Abera, and X. Zeng, "Spectral-spatial discriminant feature learning for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 13, Jun. 2019, Art. no. 1552.
- [31] 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.
- [32] A. J. X. Guo and F. Zhu, "A CNN-based spatial feature fusion algorithm for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7170–7181, Sep. 2019.
- [33] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.
- [34] Y. Shen, L. Xiao, J. Chen, and D. Pan, "A spectral-spatial domain-specific convolutional deep extreme learning machine for supervised hyperspectral image classification," *IEEE Access*, vol. 7, pp. 132240–132252, 2019.

- [35] X. Li, M. Ding, and A. Pizurca, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2615–2629, Apr. 2020.
- [36] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, Feb. 2019.
- [37] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. Bhattacharyya, "Hyperspectral image classification with attention aided CNNs," to be published, doi: [10.1109/TGRS.2020.3007921](https://doi.org/10.1109/TGRS.2020.3007921).
- [38] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (overview and toolbox)," to be published, doi: [10.1109/MGRS.2020.2979764](https://doi.org/10.1109/MGRS.2020.2979764).
- [39] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3904–3908.
- [40] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection," *Remote Sens.*, vol. 10, no. 4, Apr. 2018, Art. no. 574.
- [41] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2019.
- [42] X. Liu, R. Wang, Z. Cai, X. Yin, and Y. Cai, "Deep multigrained cascade forest for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8169–8183, Oct. 2019.
- [43] Y. Chen, K. Zhu, L. Zhu, X. He, J. A. Benediktsson, and P. Ghamisi, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7048–7066, Sep. 2019.
- [44] Y. Chen, Y. Wang, Y. Gu, X. He, P. Ghamisi, and X. Jia, "Deep learning ensemble for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1882–1897, Jun. 2019.
- [45] X. Cui, K. Zheng, L. Gao, B. Zhang, D. Yang, and J. Ren, "Multiscale spatial-spectral convolutional network with image-based framework for hyperspectral imagery classification," *Remote Sens.*, vol. 11, no. 19, Sep. 2019, Art. no. 2220.
- [46] S. Li, X. Zhu, Y. Liu, and J. Bao, "Adaptive spatial-spectral feature learning for hyperspectral image classification," *IEEE Access*, vol. 7, pp. 61534–61547, 2019.
- [47] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019.
- [48] N. He *et al.*, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.
- [49] H.-C. Li, W.-Y. Wang, L. Pan, W. Li, Q. Du, and R. Tao, "Robust capsule network based on maximum coreentropy criterion for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 738–751, Feb. 2020.
- [50] C. Shi and C.-M. Pun, "Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 487–501, Feb. 2020.
- [51] 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.
- [52] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.
- [53] K. Zhu, Y. Chen, P. Ghamisi, X. Jia, and J. A. Benediktsson, "Deep convolutional capsule network for hyperspectral image spectral and spectral-spatial classification," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 223.
- [54] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7178–7186.
- [55] H. J. Yoo, "Deep convolution neural networks in computer vision: A review," *IEEE Trans. Smart Process. Comput.*, vol. 4, no. 1, pp. 35–43, Feb. 2015.
- [56] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4755–4784, Aug. 2017.
- [57] J. A. Richards, *Remote Sensing Digital Image Analysis*. Berlin, Germany, Springer, 2013.
- [58] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, no. 1/3, pp. 37–52, Aug. 1987.
- [59] W. Sun and Q. Du, "Graph-regularized fast and robust principal component analysis for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3185–3195, Jun. 2018.
- [60] D. Misra, "Mish: A self regularized non-monotonic neural activation function," vol. 13, 2019, *arXiv:1908.08681*.
- [61] 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.

Zixian Ge received the M.S. degree in electronics and communications engineering from the Lanzhou University of Technology, Lanzhou, China, in 2019. He is currently working toward the Ph.D. degree in computer science and technology from the Nanjing University of Science and Technology, Nanjing, China.

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

Guo Cao received the Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiao Tong University, Shanghai, China, in 2006.

Since 2007, he has been with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, where he is currently a Full Professor. During January 2012–2013, he was a Visiting Scholar with the Department of Radiology, University of Chicago, Chicago, IL, USA. From August 2016 to 2017, he was a Visiting Scholar with the Department of Geography, University of Florida, Gainesville, FL, USA, where he focused on change detection. His research interests include machine learning, remote sensing image processing, and biometrics.

Xuesong Li received the B.S. degree in computer science and technology in 2014 from the Nanjing University of Science and Technology, Nanjing, China, where he is currently working toward the Ph.D. degree in computer science and technology.

His research interests include image processing and sparse representation.

Peng Fu received the B.E. degree in computer science and technology and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2009 and 2015, respectively.

He is currently a Lecturer and Postdoctoral Fellow with the School of Computer Science and Engineering, NUST. His research interests include image processing, noise estimation, and pattern recognition.