IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 February 2023, accepted 22 February 2023, date of publication 28 February 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250447

RESEARCH ARTICLE

An Improved 3D-2D Convolutional Neural Network Based on Feature Optimization for Hyperspectral Image Classification

YAMEI MA[1], SHUANGTING WANG[1,2,3], WEIBING DU[1,2,3], AND XIAOQIAN CHENG[1]

[1]School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China
[2]Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Zhengzhou, Henan 450000, China
[3]Key Laboratory of Spatiotemporal Perception and Intelligent Processing, Ministry of Natural Resources, Zhengzhou 450000, China

Corresponding author: Weibing Du (dwb@hpu.edu.cn)

This work was supported in part by the Joint Fund of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province; in part by the Key Laboratory of Spatiotemporal Perception and Intelligent processing, Ministry of Natural Resources, under Grant 211102; in part by the Principal Investigator (PI) Project of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province, under Grant 2020C002; in part by the Fundamental Research Funds for the Universities of Henan Province under Grant NSFRF220424; in part by the Project of Provincial Key Technologies Research and Development Program of Henan under Grant 222102320306; and in part by the Doctoral Fund of Henan Polytechnic University under Grant B2022-8.

ABSTRACT As a new technology in the field of remote sensing, hyperspectral remote sensing has been widely used in land classification, mineral exploration, environmental monitoring, and other areas. In recent years, deep learning has achieved outstanding results in hyperspectral image classification tasks. However, problems such as low classification accuracy for small sample classes in unbalanced datasets and lack of robustness of the models usually lead to unstable classification performance of hyperspectral images. Therefore, from the perspective of feature optimization, we propose an improved hybrid convolutional neural network for hyperspectral image feature extraction and classification. Different from the current simple multi-scale feature extraction, we first optimize the features of each scale, and then perform multi-scale feature fusion. To this end, we use 3D dilated convolution to design a multi-level feature extraction block (MFB), which can be used to extract features with different correlation strengths at a fixed scale. Then, we construct a spatial multi-scale interactive attention (SMIA) module in the spatial feature enhancement phase, which can refine the multi-scale features through the attention weights of multi-scale feature interaction, and further improve the quality of spatial features. Finally, experiments were performed on different datasets, including balanced and unbalanced samples. The results show that the proposed model is more accurate and the extracted features are more robust.

INDEX TERMS Hybrid convolutional neural network; feature extraction; attention mechanism; spectral-spatial classification; unbalanced dataset.

I. INTRODUCTION

As a new branch of remote sensing, hyperspectral remote sensing has developed rapidly in recent years. At present, hyperspectral images (HSI) have been widely used in many fields, such as mineral exploration and environmental monitoring [1], [2], [3], [4]. Using hyperspectral images to classify land use has also become an important research direction of hyperspectral applications [5]. However, there is a high correlation between the spectral bands of hyperspectral data, which brings some redundant data to the classification process [6]. In addition, due to the high dimensionality of the data, the demand for samples in the process of using deep learning for classification also increases. However, it is difficult to obtain samples of hyperspectral images, which requires a lot of manpower and material resources [7]. Therefore, the current research mainly focuses on how to extract more discriminant features from fewer samples [8], [9], [10].

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

In recent years, with the revival of artificial intelligence, hyperspectral image classification methods based on deep learning have developed rapidly [11], [12], [13], [14], [15]. Chen et al. studied the classification of hyperspectral data using three typical deep learning network architectures, and the results show that the convolutional neural network (CNN) has great advantages in the feature extraction process of hyperspectral image classification. Based on this, the researchers proposed some classical network models. For example, Lee and Kwon [16] learned multi-scale features through multi-scale convolutional filter banks similar to GoogleNet [17]. Roy et al. [18] constructed a hybrid convolutional neural network by combining two-dimensional and three-dimensional convolution. Li et al. [19] used two branches to extract spectral and spatial features respectively, and classified the fusion features. However, when constructing a deep learning framework for hyperspectral image classification, gradient vanishing is inevitable, which will degrade the performance of the model. In this regard, Zhong et al. [20] constructed a spectral-spatial residual network (SSRN) by using the residual connection. The network establishes a deep CNN classification model by connecting two spectral feature extraction blocks and two spatial feature extraction blocks in series, and achieves excellent performance. Inspired by SSRN, Wang et al. [21] designed a fast dense spectral spatial convolution framework (FDSSC) using dense connections, which can further improve the accuracy. These two works show that the use of shortcut connection [22], [23], [24] in hyperspectral image classification can effectively alleviate the gradient dispersion phenomenon.

Multi-scale feature extraction is an important idea in target detection. In recent years, it has achieved good results in image classification. In the process of hyperspectral image classification, the large convolution kernel ensures the scale invariance of feature extraction to a certain extent, but it will damage the feature extraction of smaller targets. Therefore, it is of great significance to apply the multi-scale fusion strategy to hyperspectral image classification. He et al. [25] constructed a five-layer multi-scale network and proposed an end-to-end way to learn spatial-spectral features. The results show that multi-scale feature extraction can significantly reduce misclassifications. In addition, Li et al. [26] designed a multi-scale deep middle-level feature fusion network with multi-scale input. The model inputs cubes of different scales into multiple models for individual training and fine-tuning, and then extracts the middle-level features of each model for fusion and classification. The results achieved high classification accuracy. However, large convolution kernels in multi-scale feature extraction usually lead to a sharp increase in the number of parameters, which is very unfriendly to hyperspectral image classification. In this regard, dilated convolution [27] can increase the size of the convolution kernel while ensuring that the parameters remain unchanged. Gao et al. [28] introduced dilated convolution into multi-scale feature extraction module to classify hyperspectral images. They found that dilated convolutions can further improve classification performance. However, due to the excellent performance of multi-scale features, most networks ignore the feature optimization problem on a single scale.

The phenomenon of small samples is an inherent characteristic of the mapping of the real world to the digital world, and hyperspectral data processing also faces such a challenge all the time. For the lack of samples, there are currently two main solutions: 1) data augmentation (sample expansion) [29], [30], [31], [32]; 2) adaptation to small sample learning [33], [34], [35]. Generally speaking, the larger the number of training samples in deep learning, the more representative the extracted data features, the smaller the sample size, and the less general the feature expression. However, the last two years have seen numerous proposed processing solutions for small sample data, with encouraging results. He et al. [36] designed a semi-supervised classification model by using generative adversarial networks. Mei et al. [37] propose an unsupervised three-dimensional (3D) convolutional autoencoder network, that extract features and reconstruct them by convolution and deconvolution. This approach produces better results than traditional unsupervised algorithms. Xie et al. [38] introduced transfer learning to solve the problem of limited samples, and the results show that transfer learning can greatly improve the operational efficiency of the network model.

In addition, solving the feature learning problem of small sample categories by optimizing the model structure is also an important research direction of deep learning. In the study of hyperspectral image classification, hybrid convolutional neural networks show outstanding advantages. Hyperspectral image is a three-dimensional cube data containing spectral and spatial information. Using 3D convolution to extract features is more in line with the structural pattern of the data. However, 3D convolution has a more complex operation process, and some ground object categories in the spectral dimension have noise. Therefore, scholars have proposed to use 2D convolution to further extract spatial features to improve feature quality [18]. In this regard, Feng et al. [10] used 3D convolution and 2D separable convolution to construct dense blocks, and extracted spatial spectrum and spatial features in turn. Zhang et al. [8] optimized the 3D-2D hybrid convolutional neural network model, which uses spectral and spatial attention mechanisms to refine the extracted features, and achieved excellent results in a small number of samples. Attention mechanism is an important deep learning technology to solve pixel interference and simulate human eye focusing. Ma et al. [39] and Li et al. [40] introduced the attention mechanism with different structures into FDSSC [21] for small sample experiments. The results show that the attention mechanism has a significant effect on improving the classification performance of small samples. However, the structure of the hybrid model is single, and the extracted features lack discrimination. By observing the experimental results of the above network model, it can be found that the

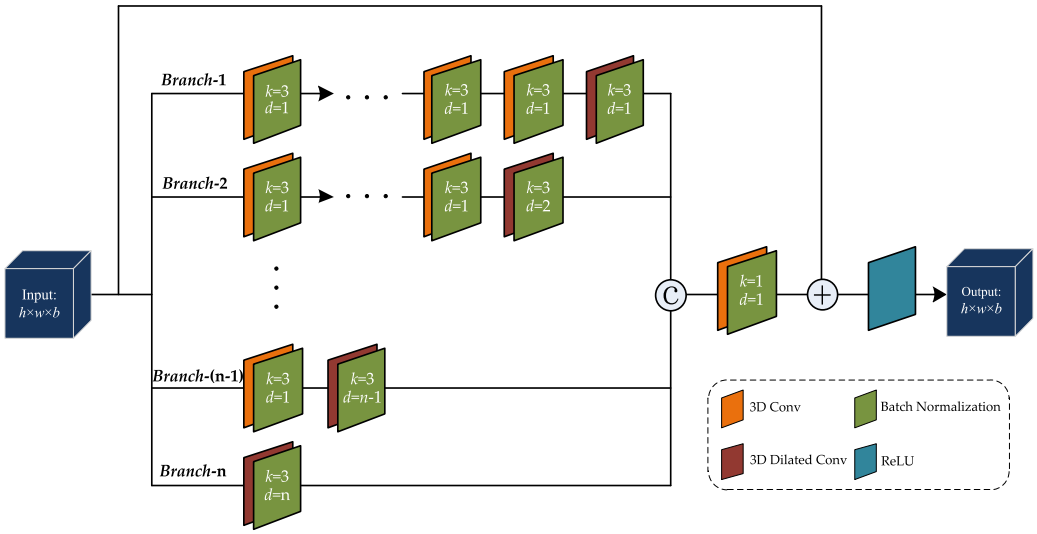

FIGURE 1. Multi-level feature extraction block (MFB).

features extracted on the unbalanced datasets still produce low classification accuracy in the small sample category.

Hyperspectral data contain significant information, and how to effectively extract more discriminative features has always been an important research direction in hyperspectral image classification. Inspired by the above research, we propose an improved 3D-2D hybrid convolutional neural network to improve the feature quality of small sample classes of hyperspectral images and improve the robustness of the model for feature extraction of different classes. The network uses multi-level feature extraction blocks (MFB) to optimize single-scale features before multi-scale feature extraction to improve the robustness of spatial-spectral joint features. Then, when further extracting spatial information, a spatial multi-scale feature interactive attention (SMIA) module is introduced to refine features and enhance important feature responses in multi-scale space. Finally, the global average pooling is used instead of the fully connected layer to input the refined features into softmax for classification.

The main contributions of this study are summarized as follows:

1) From the perspective of feature optimization, we propose an improved 3D-2D hybrid convolutional neural network model, which optimizes 3D features and 2D spatial features respectively to improve information utilization efficiency.
2) We design a multi-level feature extraction block (MFB) to capture features with different correlation strengths between each pixel and the center pixel at the same scale, and fuse MFBs of different scales at different depths to obtain multi-correlation multi-scale spatial-spectral joint features.
3) We construct a spatial multi-scale interactive attention (SMIA) module to further optimize spatial features. To control the number of parameters, depthwise separable convolution is used instead of regular convolution to extract multi-scale features.
4) The proposed method is compared with other methods, and the results show that the proposed method can more effectively improve the robustness of the model to different classes of feature extraction.

The rest of this paper is organized as follows. Section II describes MFB, SMIA, and the proposed overall network architecture. Section III conducts experiments and discussions on different datasets, including parameter analysis, method comparison, sample size analysis, and ablation experiments. Finally, the corresponding conclusions are given in section IV.

II. METHODOLOGY

A. MULTI-LEVEL FEATURE EXTRACTION BLOCK

Traditional multi-scale feature extraction usually uses convolutional kernels of different sizes to perform parallel operations on the input features, followed by feature aggregation. However, it assumes that using only one set of convolutional kernels at each scale branch is sufficient to extract the important features at that scale. In fact, the correlation of features extracted at a single scale using different convolutional structures varies greatly. In order to more fully extract the discriminative features of hyperspectral images, we provide a new idea, namely intra-block single-scale fusion and inter-block multi-scale fusion. Hence, we combined dilated convolution to design a multi-level feature extraction block at a single scale. Dilated convolution [27] expands the receptive field by inserting a certain proportion of nulls into the convolution kernel, which is called the dilation rate. When the dilation rate is 1, it is equivalent to regular convolution. The main advantage of dilated convolution is that it can directly capture highly correlated features over long distances without downsampling or adding additional parameters. The MFB mainly

extracts features with different correlation strengths between each pixel and the central pixel by introducing dilated convolution in each branch.

As shown in Figure 1, the scale of the MFB block is defined as s, the size of the convolution kernel in the block is k, and the number of branches is n. Then there is a correspondence $s = (n-1) \times (k-1) + k$, and the dilation rate of the dilated convolution of each branch in the block is $d = \{1, 2, \ldots, n\}$. Specifically, the first branch consists of n layers of regular convolutions (containing a dilated convolution with a dilation rate of 1) to extract high-level semantic features within the MFB receptive field. The second branch combines the kernel size of the last two convolutional layers in the first branch to obtain $(n-2)$ layers of regular convolution and a $d = 2$ dilated convolutional layer. The third branch converts the last three convolution layers in the first branch into a dilated convolution layer with a kernel size of $d = 3$. By analogy, the dilated convolutions with increasing dilation rates are added to each branch one by one. The last branch consists of a layer of dilated convolutions of $d = n$. The structure aims to consider the correlation feature extraction between pixels with different distances and central pixels, which can improve the classification performance of the same class pixels with far distances to a certain extent. The lower the branch, the more compact the feature extraction between neighboring pixels, and the weaker the correlation between central and distant pixels. In contrast, for higher branches, the neighborhood feature extraction is more sparse, and the long-distance feature dependence is stronger. In addition, with the decrease of branches, the abstraction level of features is gradually increasing. After the feature extraction of each branch, '*concat*' is used to connect the different features extracted, and the features of different branches are fused through a $1 \times 1 \times 1$ convolutional layer while adjusting the channel dimension. Finally, identical residual connections are added to the block to mitigate gradient disappearance. In addition, to accelerate the convergence of the network and improve the nonlinear expression, there is a batch normalization layer behind each convolutional layer in the block, and a nonlinear activation function layer is added after the residual connection.

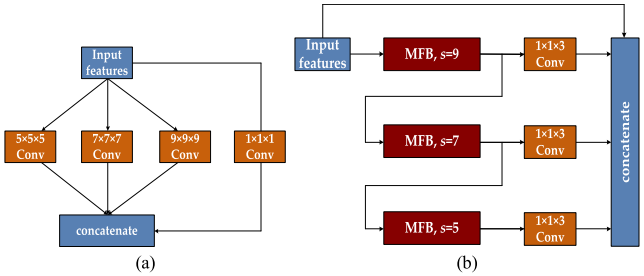

FIGURE 2. Multi-scale feature fusion structure. (a) Traditional multi-scale feature fusion structure; (b) Improved multi-scale feature fusion structure optimized by MFB.

B. IMPROVED MULTI-SCALE FEATURE FUSION

A large number of studies have shown that multi-scale feature extraction is necessary for hyperspectral image classification, but few people consider whether the feature extraction at each scale is sufficient. Figure 2(a) is the traditional multi-scale feature extraction module in the neural network. It can be found that the module directly inputs features into different sizes of kernels for multi-scale feature extraction and fusion. As shown in Figure 2(b), to optimize the multi-scale feature extraction structure, we introduce the MFB into the multi-scale feature extraction module and improve it. Specifically, we use MFB to extract optimized features of different scales on the backbone network step by step, and then cascade the features of different scales to obtain optimized multi-scale features.

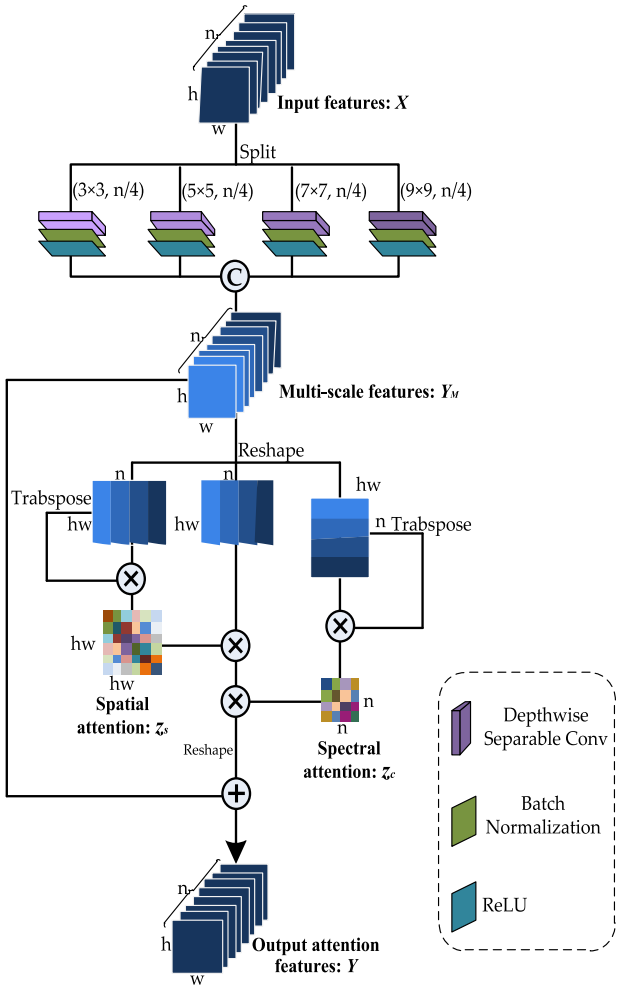

FIGURE 3. The spatial multi-scale interactive attention (SMIA) module.

C. SPATIAL MULTI-SCALE INTERACTIVE ATTENTION MODULE

To simulate the autofocus function of the human visual system, computer vision proposes to use attention mechanism to improve the response of the region of interest. It has been successfully applied in scene segmentation, image classification and text translation [41], [42], [43]. In the process of hyperspectral image feature extraction, there will inevitably

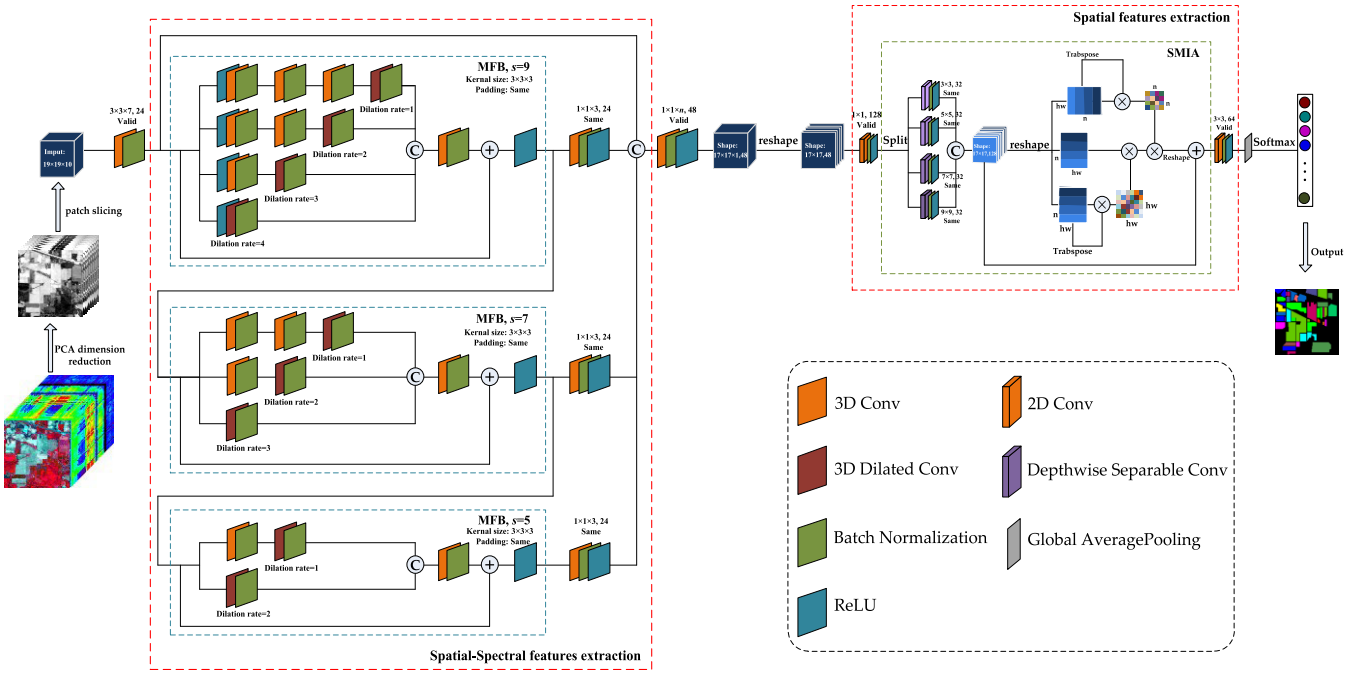

FIGURE 4. Illustration of complete architecture of network model.

be interference terms, and weighting through the attention mechanism can effectively suppress noise and enhance the representation of important information. To further extract 2D spatial information and improve feature quality, we propose a spatial multi-scale interactive attention module in this paper. As shown in Figure 3, the module is divided into two main parts, namely multi-scale spatial feature extraction and feature interaction refinement.

In the multi-scale feature extraction stage, the input feature is defined as $\boldsymbol{X} \in \mathbf{R}^{(w\times h\times n)}$, where $w \times h$ is the space size and n is the number of channels. Firstly, the input data is divided into four groups of features with the same size by channel grouping, which is expressed as $\boldsymbol{X}_i \in \mathbf{R}^{(w\times h\times n/4)}$, $i = 1, 2, 3, 4$. This process can reduce the number of subsequent convolution parameters to 1/4 of the conventional multi-scale feature extraction blocks. Then, the same number of 2D convolution kernels with different sizes are used for feature extraction, and cascaded to obtain a multi-scale feature $\boldsymbol{Y}_M$, which has the same size as $\boldsymbol{X}$. In addition, to further control the increase in the number of parameters, we replace the regular 2D convolution with depthwise separable convolution (DSC), which has been applied in some studies [44], [45]. DSC divides the regular 2D convolution into two processes: depthwise convolution and pointwise convolution. The specific multi-scale feature calculation process is as follows:

$$F_{k\times k}\left(\boldsymbol{X}_i\right) = \delta\left(\boldsymbol{W}_k * \boldsymbol{X}_i + \boldsymbol{b}_k\right) \tag{1}$$

$$\boldsymbol{Y}_M = \left[F_{3\times 3}\left(\boldsymbol{X}_1\right), F_{5\times 5}\left(\boldsymbol{X}_2\right), F_{7\times 7}\left(\boldsymbol{X}_3\right), F_{9\times 9}\left(\boldsymbol{X}_4\right)\right] \tag{2}$$

where $\boldsymbol{W}_k$ and $\boldsymbol{b}_k$ are the weight and bias parameters corresponding to a convolution of scale k, $k = \{3, 5, 7, 9\}$, respectively. δ denotes the ReLU activation function and $[\cdot]$ denotes the concatenate operation.

In the process of feature refinement, '*reshape*' is performed on the size of $\boldsymbol{Y}_M$ to obtain $\boldsymbol{Y}_M \in \mathbf{R}^{(z\times n)}$, where z is equal to $w \times h$. Then the spatial attention map $\boldsymbol{Z}_s$ and the channel attention map $\boldsymbol{Z}_c$ are obtained by multiplying $\boldsymbol{Y}_M$ with its transpose. The channel dimension contains the spatial features of different scales, so the channel attention map is the correlation weight of the interaction of different scale features. The specific calculation process is as follows:

$$\boldsymbol{Z}_s = \sigma\left(\boldsymbol{Y}_M \boldsymbol{Y}_M^T\right) \tag{3}$$

$$\boldsymbol{Z}_c = \sigma\left(\boldsymbol{Y}_M^T \boldsymbol{Y}_M\right) \tag{4}$$

where $\boldsymbol{Z}_s \in \mathbf{R}^{(z\times z)}$ represents the correlation between spatial pixels within each scale, $\boldsymbol{Z}_c \in \mathbf{R}^{(n\times n)}$ represents the correlation between multi-scale features, and σ is the sigmoid activation function. Afterward, the module combines the interaction between different scale channels with the interaction between spatial pixels in each channel to obtain a spatially multi-scale interactive thinning feature map $\boldsymbol{Y}$. Finally, skip connections are added to facilitate algorithm convergence. $\boldsymbol{Y}$ is calculated as follows:

$$\boldsymbol{Y} = \boldsymbol{Z}_s \boldsymbol{X} \boldsymbol{Z}_c + \boldsymbol{X} \tag{5}$$

D. HSI CLASSIFICATION BASED ON THE PROPOSED METHOD

The overall structure of the proposed model is shown in Figure 4. Noting that hyperspectral images have a very high spectral dimension, we first perform PCA dimensionality

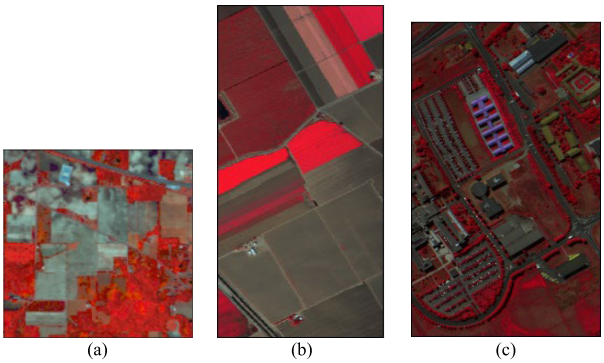

FIGURE 5. Band composite images. (a) IP dataset; (b) SV dataset; (c) PU dataset.

reduction on the entire hyperspectral data. Principal component analysis (PCA) is one of the important methods for dimensionality reduction of hyperspectral images. It mainly removes the correlation between spectra by K-L transform, and then determines the important spectral features according to the contribution of terrain attribute information. After data dimensionality reduction, standardiz-ation is performed channel by channel, which can avoid errors caused by data differences between channels. In addition, the whole image input network will increase the burden of network training. Therefore, we use the patch as the input of the network. After determining the patch size, we cut the patch around each pixel in the original image. The background elements are then removed to reduce the network running burden.

The feature extraction process of the model can be divided into two stages: the spatial-spectral feature extraction stage and the spatial enhancement stage. First, a $3 \times 3 \times 7$ convolution kernel is used to extract the shallow features of the input features. Then, in the spatial-spectral feature extraction stage, the multi-scale feature extraction structure optimized by MFB is used to extract multi-scale features. This stage can improve the robustness of the model to different classes of features. In the spatial enhancement stage, the 3D features are first converted to 2D features by '*reshape*' and the channel dimension is extended using a 2D convolution with a kernel size of 1×1. Then the extended features are input into the spatial multi-scale interactive attention module to obtain multi-scale spatial refinement features. Finally, the fully connected layer is replaced by global pooling, and the features are input into softmax for classification. In addition, to reduce the impact of overfitting, dropout is added before the classification layer.

III. EXPERIMENT AND ANALYSIS

A. EXPERIMENTAL DATA

The proposed network model was evaluated and analyzed by using three commonly used publicly available datasets: the Indian Pines (IP) dataset, the Salinas Valley (SV) dataset, and the Pavia University (PU) dataset. The three datasets have the following differences: 1) IP and SV datasets have strong spatial homogeneity and are mostly vegetation classes, while PU datasets have more building classes and more discrete spatial distribution; 2) The IP dataset has significantly different sample sizes for different classes, which can verify the robustness of the model, and the PU and SV datasets have relatively balanced sample sizes with no class of very

TABLE 1. Sample size per class and training set partitioning on IP datasets.

No.	Class	Training	Testing	Total
1	Alfalfa	2	44	46
2	Corn-notill	71	1357	1428
3	Corn-mintill	41	789	830
4	Corn	11	226	237
5	Grass-pasture	24	459	483
6	Grass-trees	36	694	730
7	Grass-pasture-mowed	1	27	28
8	Hay-windrowed	23	455	478
9	Oats	1	19	20
10	Soybean-notill	48	924	972
11	Soybean-mintill	122	2333	2455
12	Soybean-clean	29	564	593
13	Wheat	10	195	205
14	Woods	63	1202	1265
15	Buildings-Grass-Trees-Drives	19	367	386
16	Stone-Steel-Towers	4	89	93
	Total	505	9744	10249

TABLE 2. Sample size per class and training set partitioning on SV datasets.

No.	Class	Training	Testing	Total
1	Brocoli_green_weeds_1	20	1989	2009
2	Brocoli_green_weeds_2	37	3689	3726
3	Fallow	19	1957	1976
4	Fallow_rough_plow	13	1381	1394
5	Fallow_smooth	26	2652	2678
6	Stubble	39	3920	3959
7	Celery	35	3544	3579
8	Grapes_untrained	112	11159	11271
9	Soil_vinyard_develop	62	6141	6203
10	Corn_senesced_green_weeds	32	3246	3278
11	Lettuce_romaine_4wk	10	1058	1068
12	Lettuce_romaine_5wk	19	1908	1927
13	Lettuce_romaine_6wk	9	907	916
14	Lettuce_romaine_7wk	10	1060	1070
15	Vinyard_untrained	72	7196	7268
16	Vinyard_vertical_trellis	18	1789	1807
	Total	533	53596	54129

TABLE 3. Sample size per class and training set partitioning on PU datasets.

No.	Class	Training	Testing	Total
1	Asphalt	66	6565	6631
2	Meadows	186	18463	18649
3	Gravel	20	2079	2099
4	Trees	30	3034	3064
5	Painted metal sheets	13	1332	1345
6	Bare Soil	50	4979	5029
7	Bitumen	13	1317	1330
8	Self-Blocking Bricks	36	3646	3682
9	Shadows	9	938	947
	Total	423	42353	42776

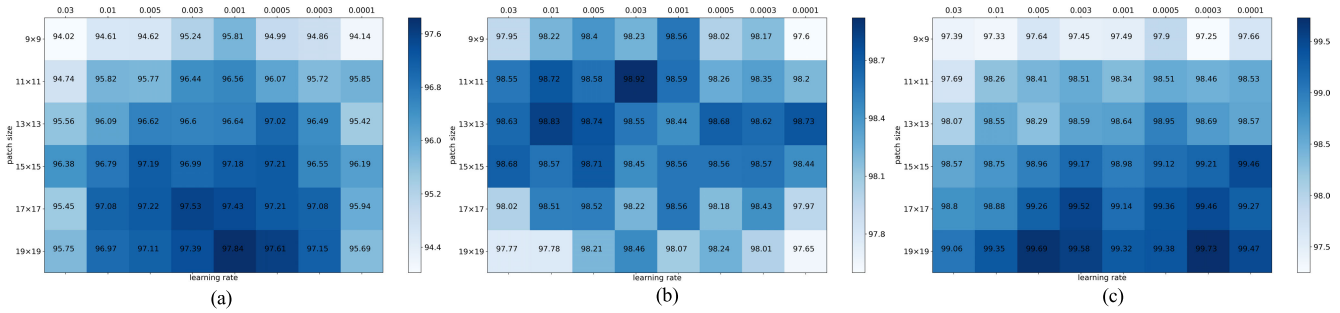

FIGURE 6. Parameter heat maps on three datasets. (a) IP dataset; (b) PU dataset; (c) SV dataset.

few samples, thus emphasizing the accuracy of the model. Figure 5 shows the band composite images for the three datasets. Tables 1-3 show the total number of samples in each class and the number of training and testing samples on the three datasets.

The IP dataset was collected in 1992 by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) airborne sensor at the Indiana Pine Forest Experimental Area in northwest Indiana, USA. The data have a wavelength range from 0.4 to 2.5 μm and consist of 224 spectral bands and 145 $\times$ 145 pixels with a spatial resolution of 20 meters. The quality of spectral images strongly affects feature extraction, so the bands with low signal-to-noise ratio and that are affected by water vapor are removed, leaving only 200 spectral bands for experimental analysis.

The SV dataset is also obtained by the AVIRIS sensor over the Salinas Valley, California. The data wavelength range is the same as the IP dataset (0.4-2.5μm), and there are 224 spectral bands. Unlike the IP dataset, SV data retains 204 bands for experiments. In addition, the data consists of 512 $\times$ 217 pixels with a high spatial resolution of 3.7 m / pixel.

The PU dataset was acquired by the Reflective Optics System Imaging Spectrometer sensor at the University of Pavia in northern Italy. The data contains 115 spectral bands with a wavelength range of 0.43-0.86μm, and 103 spectral bands are retained after removing noise interference bands. The dimension of the data is 610 $\times$ 340 pixels with a spatial resolution of 1.3 m.

B. EXPERIMENTAL SETUP AND EVALUATION INDEX

In the whole experiment, the training samples are randomly selected. In the study of hyperspectral image classification, there is no clear definition of small sample learning. Therefore, this paper refers to the recent research work for training sample division and experimental analysis [46], [47], [48], [49]. In addition, to ensure the effective flow of information in the propagation process, the robust neural network parameter initialization method proposed by He et al. [50] is used. The optimizer is Adam [51], which calculates its update step size by considering the mean and non-central variance of the gradient in the backpropagation of the neural network. The batch size of training is 16, and the maximum epoch is set to 150. All experiments were conducted on a Windows 10 laptop equipped with i7-11800 CPU, NVIDIA GeForce RTX 3060 GPU and 16 GB RAM using the Python 3.6 Tensorflow deep learning framework. In this paper, the network model is quantitatively analyzed and evaluated by five metrics: precision, recall, F1-score, average classification accuracy (AA), and the kappa coefficient (Kappa). Precision describes how many of the samples predicted as a specific category are correct prediction results. Recall reflects the quantitative relationship between the actual samples of each category and the correctly classified samples. F1-score weighs the precision and recall for comprehensive evaluation. AA evaluates the average classification accuracy for each class. The Kappa coefficient is a consistency metric used to evaluate the consistency between the model classification results and the actual classes. In the above metrics, precision, recall, and F1-score are weighted, so the recall is equivalent to the overall classification accuracy (OA).

C. PARAMETER ANALYSIS

Hyperparameters strongly affect model performance, so it is important to select the appropriate hyperparameters. Learning rate is an important hyperparameter in deep learning. Too large learning rate will lead to model divergence, too small learning rate will reduce the efficiency of the model. To ensure better network performance, search for the best learning rate from {0.03, 0.01, 0.005, 0.003, 0.001, 0.0005, 0.0003, 0.0001} based on prior knowledge. In addition, inappropriate patch size may lead to insufficient feature extraction or increase the computational burden of the model. Therefore, to give full play to the performance of the proposed model and ensure the efficiency of the network, this paper sets the PCA of the three data to 10, and then optimizes the initial learning rate and patch size.

The experimental results of IP, PU and SV datasets under different initial learning rates and patch sizes are shown in Figure 6. Different columns represent different learning rates, and the lateral right learning rate gradually decreases. Different rows represent different patch sizes, and the size increases vertically down. The color depth of each grid in the matrix represents the quality of the classification results (as shown

in the right color bar). Intuitively, the matrices in Figure 6 (a) and Figure 6 (c) show darker color representations in the lower half, indicating that the two datasets can produce higher accuracy when the patch size is larger. In contrast, the results shown in Figure 6 (b) show that the PU dataset can produce more high precision at the appropriate patch size, and the performance on the larger patch size is reduced. This is mainly due to the PU dataset in the larger patch size will introduce different classes of interference pixels. In addition, it can be found that with the decrease of the initial learning rate, the three datasets have no obvious trend. However, the optimal results under different initial learning rates are not generated on the same patch size, which proves the importance of parameter optimization to some extent. Finally, according to Figure 6, we determine the learning rates of the IP, PU and SV datasets are 0.001,0.003 and 0.0003, respectively, and the patch sizes are 19×19,11×11 and 19×19.

D. COMPARATIVE ANALYSIS WITH OTHER METHODS

In order to evaluate the performance of the proposed model in the case of small samples, this section will compare and analyze the proposed method with several deep learning methods developed in recent years, which are DFFN [52], SSRN [20], MSDN [53], BAM-CM [54], HybridSN [18], MSRN-A [55], MSR-3DCNN [56], and MDAN [49]. The following is a brief introduction to each method in order of publication time.

(1) DFFN: This model was proposed by Song et al. in 2018. DFFN uses the residual block constructed by 2D convolution as the basic unit of feature extraction to design a deep feature extraction network, and fuses low, medium and high-level features for classification.
(2) SSRN: SSRN was proposed by Zhong et al. in 2018. The model uses 3D convolution and residual connection to extract spectral and spatial features, and the extracted features are pooled and classified.
(3) MSDN: Zhang et al. proposed a multi-scale dense network in 2019 and described it in detail. MSDN combines 3D stride convolution to extract spatial-spectral joint features at different sampling levels, and uses dense connections to aggregate features at different levels.
(4) BAM-CM: Dong et al. proposed a band selection attention module in 2019, which implements adaptive band weighting processing operations.
(5) HybridSN: HybridSN proposed by Roy et al. in 2020 contains three 3D convolutions and one 2D convolution. It extracts spatial-spectral and spatial features successively and classifies them after passing through two fully connected layers.
(6) MSRN-A: Zhang et al. proposed MSRN-A in 2020. The model takes 3D-2D hybrid convolutional neural network as the basic framework, and refines 3D and 2D features respectively by using spatial-spectral attention and spatial attention.
(7) MSR-3DCNN: This architecture was proposed by Xu et al. in 2021. MSR-3DCNN introduces dilation rate in the spectral dimension of 3D convolution, and extracts the spatial-spectral joint features of multi-spectral resolution.
(8) MDAN: MDAN is a multi-dimensional feature extraction architecture proposed by Liu et al. in 2022. Based on the 3D-2D model, the architecture further integrates the 1D feature extraction structure and uses the improved CBAM for feature refinement.
(9) Proposed: We combine the feature extraction advantages of hybrid convolutional neural networks, and use MFB and SMIA to optimize the spatial-spectral and spatial features to improve the robustness of the model features and reduce the difference in the accuracy of each class in the unbalanced dataset.

To ensure fair testing, the network settings of different methods (such as patch size and PCA number, etc.) are the same as those in the original study, which ensures the best results for each method. In addition, all experimental methods are carried out on the Tensorflow deep learning framework.

1) INDIAN PINES DATASET

As shown in Table 4, the quantitative classification results of different algorithms on IP datasets are shown. Obviously, the proposed model produced the highest classification accuracy, with OA (Recall), F1-score, AA and Kappa of 97.84%, 97.86%, 96.92% and 97.54, respectively. MSRN-A and DFFN are slightly inferior to the proposed model, with OA of 96.76% and 95.79%, respectively. The results for SSRN, HybridSN, and MSR-3DCNN are similar, while the F1-score of BAM-CM, MDAN and MSDN were all lower than 90%. Specifically, compared with MSRN-A and DFFN, the results of the proposed model are significantly improved in classes with fewer samples (such as Class 1, Class 7 and Class 9). Especially in the 7th and 9th classes (only one sample was used for training), several models produced very low accuracy. These show that the proposed model can significantly improve the classification accuracy of small sample classes. In addition, the proposed models yield accuracies above 90% for all classes (the lowest being 92.47% for class 7), which indicates that the proposed model can guarantee the accuracy balance of various classes on imbalanced datasets to a certain extent. Furthermore, it can be found that the overall accuracy of SSRN, BAM-CM, MDAN and MSDN is significantly higher than the average accuracy, which is also an important manifestation of the low accuracy of small sample classes.

Figure 7 shows the classification maps of different methods on the IP dataset. It can be found that the visual effect of these classification maps is consistent with the results in Table 4. MSDN and MDAN show a lot of salt and pepper noise in some classes. In contrast, the classification results of other methods are more uniform, but boundary and internal misclassification still exist. In this regard, although the proposed model still has misclassification at the boundary, it eliminates

TABLE 4. Classification results of algorithms applied to IP dataset.

No.	DFFN	SSRN	MSDN	BAM-CM	HybridSN	MSRN-A	MSR-3DCNN	MDAN	Proposed
1	79.73	13.18	24.55	34.44	57.62	88.57	74.42	37.62	**94.88**
2	96.57	96.76	61.90	89.17	94.12	96.84	92.33	83.34	**97.74**
3	92.99	95.59	55.67	88.42	92.61	95.42	89.57	79.17	**97.21**
4	94.30	77.70	23.63	71.92	67.48	**97.10**	82.93	62.15	96.86
5	**96.71**	96.34	69.80	89.95	88.32	91.40	92.27	87.63	93.66
6	98.17	97.12	93.14	95.94	**99.27**	97.90	99.11	94.70	97.61
7	68.15	0	11.11	11.99	46.92	63.85	87.69	0	**92.47**
8	**100**	**100**	99.47	97.71	99.95	**100**	99.52	99.67	**100**
9	64.21	1.05	17.89	34.03	57.78	64.44	44.21	0	**96.67**
10	93.56	94.03	64.72	86.90	91.91	94.81	87.19	90.79	**97.62**
11	96.02	94.22	86.85	93.61	97.12	97.61	96.86	91.86	**98.43**
12	90.95	93.16	50.92	76.60	78.24	94.68	85.29	70.07	**95.66**
13	98.55	99.79	89.44	93.47	96.54	98.59	97.01	95.57	**100**
14	98.89	99.25	96.92	97.62	98.70	99.74	98.91	97.81	**99.76**
15	97.47	97.06	54.93	90.69	89.60	**97.76**	88.14	69.71	97.52
16	84.14	94.61	30.11	76.66	65.24	93.09	77.73	39.29	**94.71**
Precision(%)	96.07±0.52	94.54±0.75	75.04±1.74	90.39±0.77	93.24±1.06	96.84±0.26	93.47±0.29	86.40±1.70	**97.93±0.31**
Recall/OA(%)	95.79±0.59	94.83±0.89	74.39±1.96	90.21±0.74	93.15±1.00	96.76±0.26	93.33±0.31	86.77±1.36	**97.84±0.35**
F1-score(%)	95.80±0.58	94.44±0.91	73.19±2.20	89.97±0.84	92.95±1.00	96.75±0.26	93.24±0.31	86.34±1.53	**97.86±0.33**
AA(%)	90.65±3.02	78.12±1.56	58.19±1.81	76.81±2.51	82.59±1.79	91.99±1.03	87.07±2.21	68.71±2.32	**96.92±1.14**
Kappa×100	95.21±0.66	94.11±1.01	70.36±2.34	88.81±0.87	92.17±1.13	96.31±0.29	92.37±0.36	84.91±1.54	**97.54±0.40**

TABLE 5. Classification results of algorithms applied to PU dataset.

No.	DFFN	SSRN	MSDN	BAM-CM	HybridSN	MSRN-A	MSR-3DCNN	MDAN	Proposed
1	95.90	**99.03**	92.28	93.71	97.89	95.86	85.64	90.37	98.47
2	99.44	99.80	98.59	99.38	**99.98**	99.87	98.81	99.66	**99.98**
3	**95.11**	84.33	52.92	72.70	76.13	84.13	78.17	75.76	91.46
4	83.31	96.68	92.60	93.61	90.74	96.49	81.32	89.22	**96.89**
5	98.25	99.94	97.76	98.48	98.45	99.71	**100**	98.85	99.88
6	98.93	96.46	58.58	93.39	97.35	99.50	85.39	98.41	**99.83**
7	94.98	92.71	71.51	75.22	95.72	91.80	96.52	88.40	**99.60**
8	96.89	93.96	80.06	88.30	88.88	98.37	74.33	68.88	**98.59**
9	71.45	**99.55**	93.80	94.34	68.91	95.76	72.15	56.25	98.26
Precision(%)	96.45±0.53	97.69±0.39	87.53±0.90	94.31±0.56	95.74±0.93	97.81±0.42	90.51±2.54	92.47±0.85	**98.95±0.07**
Recall/OA(%)	96.45±0.53	97.58±0.37	87.67±0.99	94.23±0.67	95.69±0.95	97.71±0.42	90.20±2.81	92.17±0.79	**98.92±0.06**
F1-score(%)	96.40±0.52	97.55±0.37	87.00±1.14	94.20±0.64	95.56±0.97	97.70±0.43	89.99±2.89	91.98±0.85	**98.92±0.06**
AA(%)	92.70±1.05	95.83±0.55	82.01±1.32	89.90±1.63	90.45±2.24	95.72±1.18	85.81±4.30	85.09±1.92	**98.11±0.04**
Kappa×100	95.29±0.70	96.79±0.49	83.31±1.40	92.33±0.89	94.26±1.27	96.97±0.56	86.88±3.77	89.57±1.07	**98.56±0.08**

TABLE 6. Classification results of algorithms applied to SV dataset.

No.	DFFN	SSRN	MSDN	BAM-CM	HybridSN	MSRN-A	MSR-3DCNN	MDAN	Proposed
1	99.89	99.77	94.09	99.44	99.71	99.71	98.16	99.94	**100**
2	99.95	99.99	98.56	99.52	**100**	**100**	99.98	99.99	**100**
3	98.97	99.94	95.99	99.34	99.93	**100**	99.99	99.91	**100**
4	99.13	**99.74**	98.16	97.57	98.81	98.89	99.20	98.18	98.94
5	98.72	95.93	94.57	97.13	**99.91**	96.76	96.82	99.04	99.08
6	99.77	**100**	99.77	99.98	99.61	99.98	99.89	99.65	**100**
7	98.70	99.97	99.09	99.03	99.95	99.87	99.96	**99.98**	99.97
8	98.11	92.76	78.89	92.25	96.92	97.62	95.35	96.79	**99.41**
9	99.93	99.98	98.88	99.74	99.99	**100**	99.62	99.84	**100**
10	99.68	98.84	93.44	97.76	99.03	99.11	98.34	97.40	**99.83**
11	96.65	97.83	91.40	94.57	95.26	**100**	99.77	97.73	99.79
12	99.04	**100**	99.23	99.68	99.33	99.95	96.99	99.31	99.85
13	99.05	99.34	96.67	94.83	97.13	**99.55**	92.56	99.26	98.02
14	93.60	95.64	92.94	99.10	93.04	98.84	98.79	**99.27**	99.11
15	96.55	89.00	44.71	92.25	98.11	99.72	88.09	94.43	**99.98**
16	99.34	97.70	95.79	95.83	99.31	98.44	97.85	98.80	**100**
Precision(%)	98.66±0.49	96.66±0.42	86.26±1.18	96.53±0.64	98.68±0.61	99.14±0.18	96.76±0.57	98.19±0.50	**99.74±0.25**
Recall/OA(%)	98.62±0.51	96.50±0.45	86.26±1.24	96.47±0.66	98.64±0.62	99.12±0.19	96.71±0.60	98.14±0.46	**99.73±0.26**
F1-score(%)	98.61±0.51	96.49±0.50	85.73±1.27	96.48±0.66	98.64±0.62	99.12±0.19	96.69±0.61	98.14±0.47	**99.73±0.26**
AA(%)	98.57±0.63	97.90±0.47	92.01±1.11	97.38±0.26	98.50±0.63	99.28±0.19	97.61±0.81	98.72±0.33	**99.62±0.28**
Kappa×100	98.46±0.56	96.11±0.51	84.67±1.37	96.07±0.73	98.49±0.69	99.02±0.21	96.33±0.66	97.93±0.51	**99.70±0.29**

the internal noise of each class area. The classification result map is visually closest to the ground truth value.

2) PAVIA UNIVERSITY DATASET

Table 5 shows the classification accuracy of each method on the PU dataset. It can be seen that the proposed model exhibits the best performance, and the overall accuracy and F1-score are equal to 98.92%. Compared with the proposed model, the OA of MSRN-A and SSRN decreased by 1.21% and 1.34%, respectively, and DFFN decreased by 2.47%. Compared with the other remaining methods, the performance of the proposed model is significantly improved. Observing the data in the table, it can be found that the accuracy of all classes

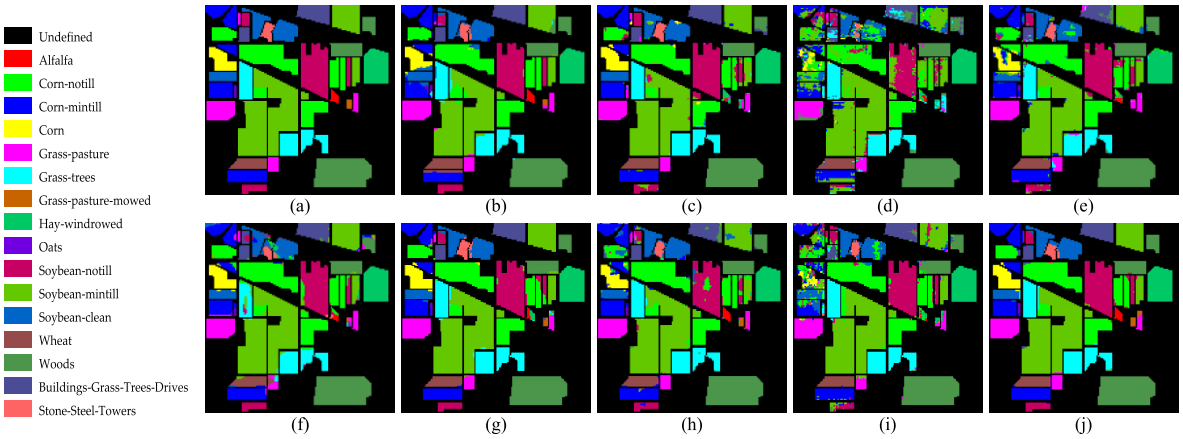

FIGURE 7. Classification map of each algorithm applied to IP dataset. (a) Ground Truth; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

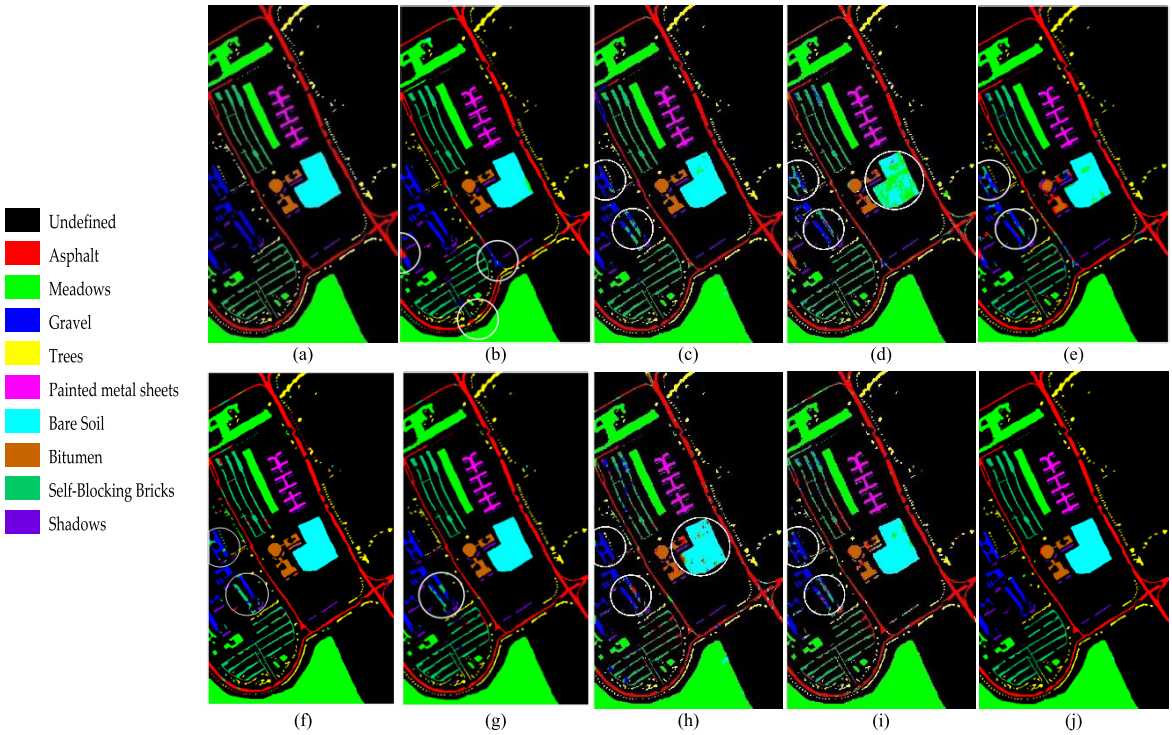

FIGURE 8. Classification map of each algorithm applied to PU dataset. (a) Ground Truth; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

on the PU dataset is above 90%, while other methods always have the accuracy of one or some classes below 90%. This again proves the stability of the proposed model. In addition, the gap between OA and AA of each method on the PU dataset is significantly reduced, mainly because there is no class with few samples on the PU dataset. It should be noted that the accuracy of different methods on the PU dataset is generally improved relative to the IP dataset, but the accuracy of the MSR-3DCNN (−3.13%) model is reduced. This indicates that the application stability of MSR-3DCNN for different data still needs to be further improved. Figure 8 is the classification maps of each method on the PU dataset. As can be seen from the annotations in the figure, MSR-3DCNN and MSDN have obvious misclassifications, while MSRN-A and others have misclassification on "Gravel" and "Self-Blocking Bricks". DFFN performs well, but there are still misclassifications on "Asphalt". In contrast, the proposed model shows better visual effects in the above regions.

3) SALINAS VALLEY DATASET

Table 6 shows the quantitative results on the SV dataset. It can be seen that each method on the SV dataset has achieved high

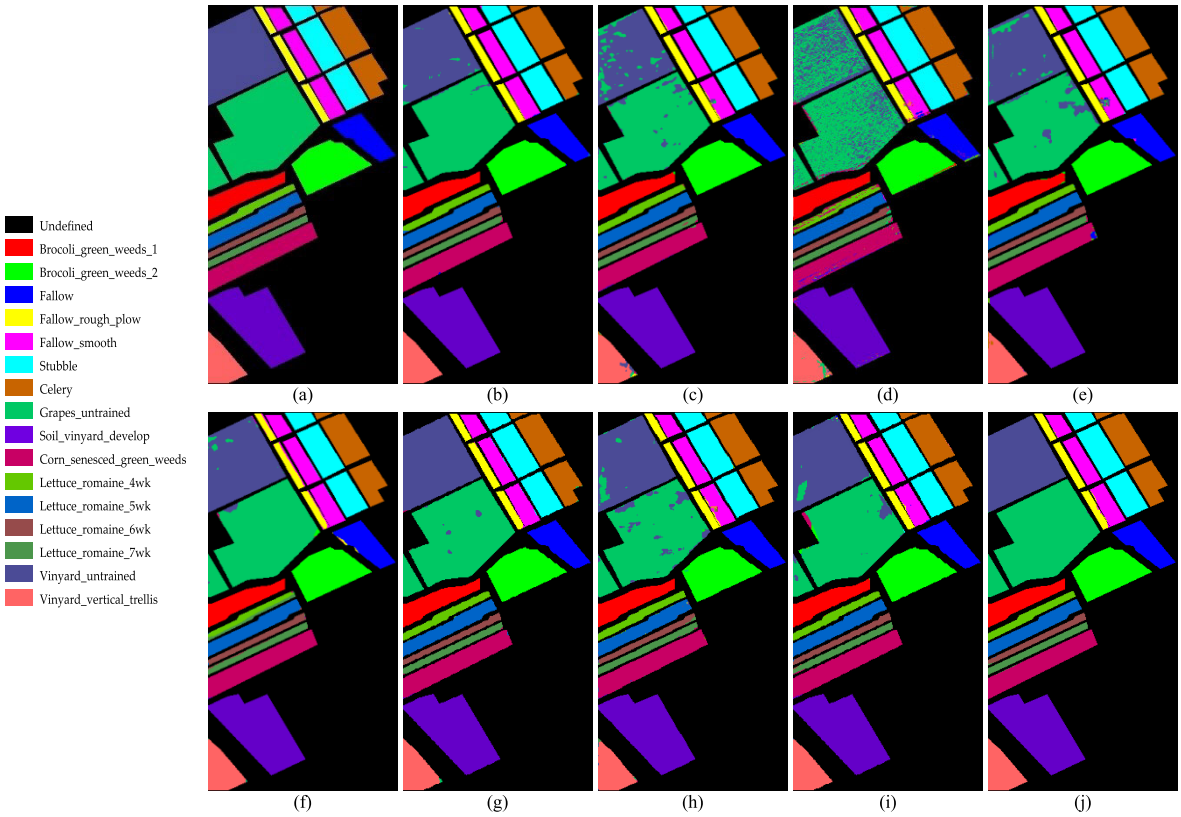

FIGURE 9. Classification map of each algorithm applied to SV dataset. (a) Ground Truth; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

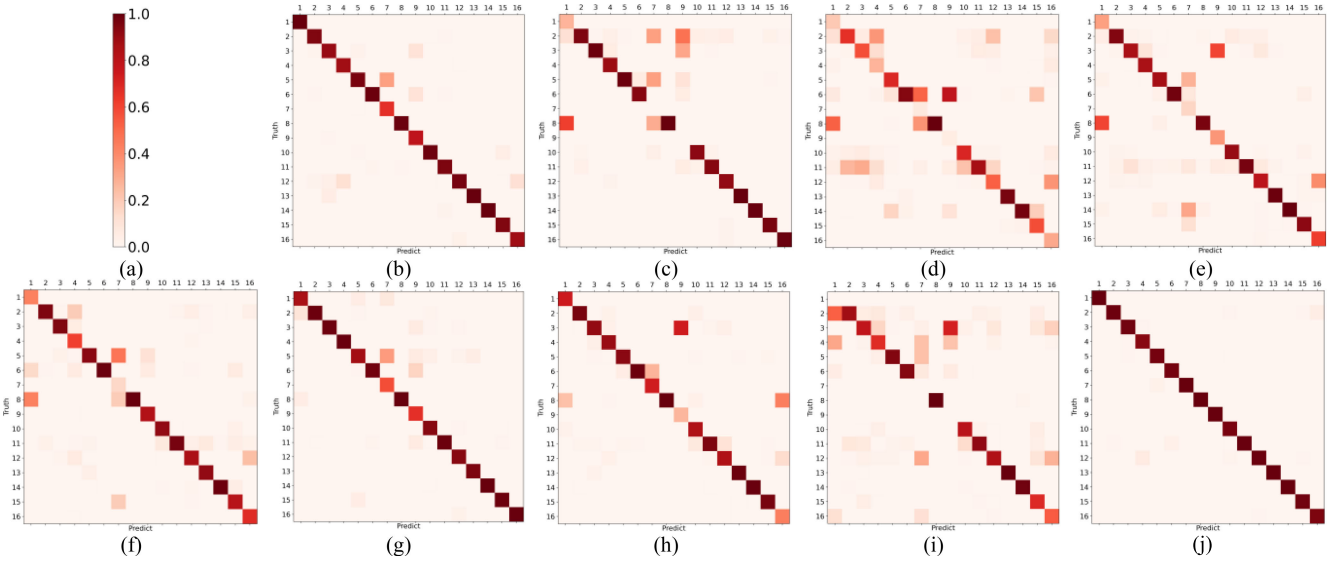

FIGURE 10. Visualization of different algorithm confusion matrices on IP datasets. (a) Chroma bar; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

accuracy, which is due to the relatively good sample quality of the SV data (compared with the IP dataset, the samples of the SV data are more balanced, and compared with the PU dataset, the SV data has higher spatial homogeneity). But in terms of overall accuracy, the proposed model showed the best results, with an OA of 99.73 %. Although the accuracy of some classes was slightly lower (<99%), the overall trend was still the highest, and the accuracy of each class remained

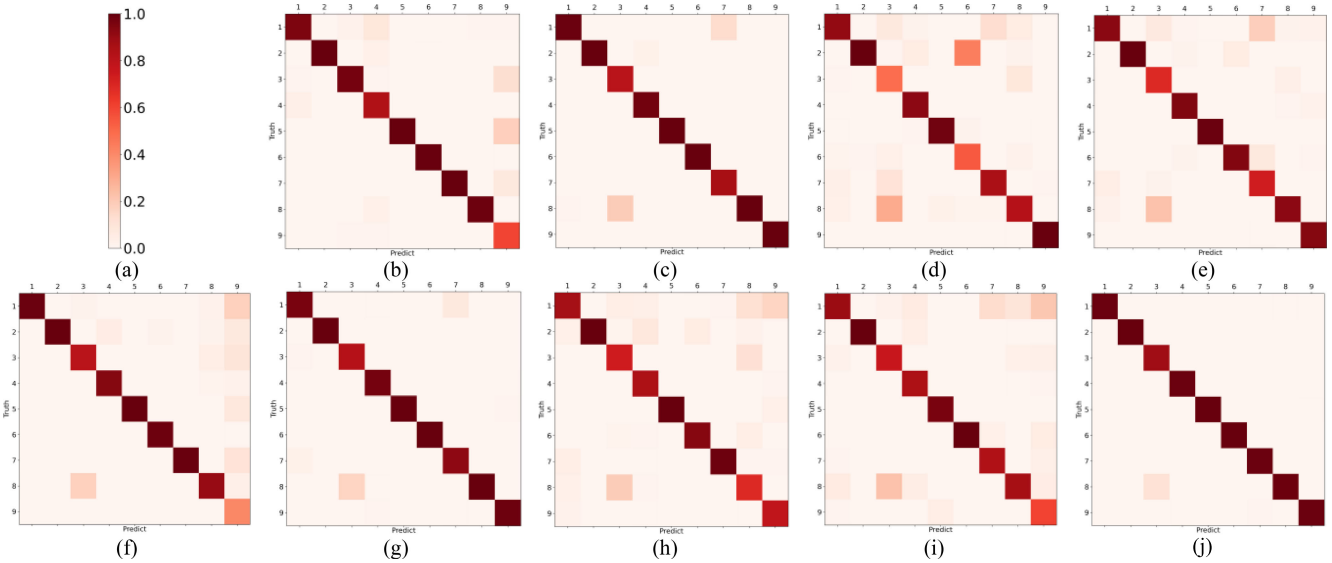

FIGURE 11. Visualization of different algorithm confusion matrices on PU datasets. (a) Chroma bar; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

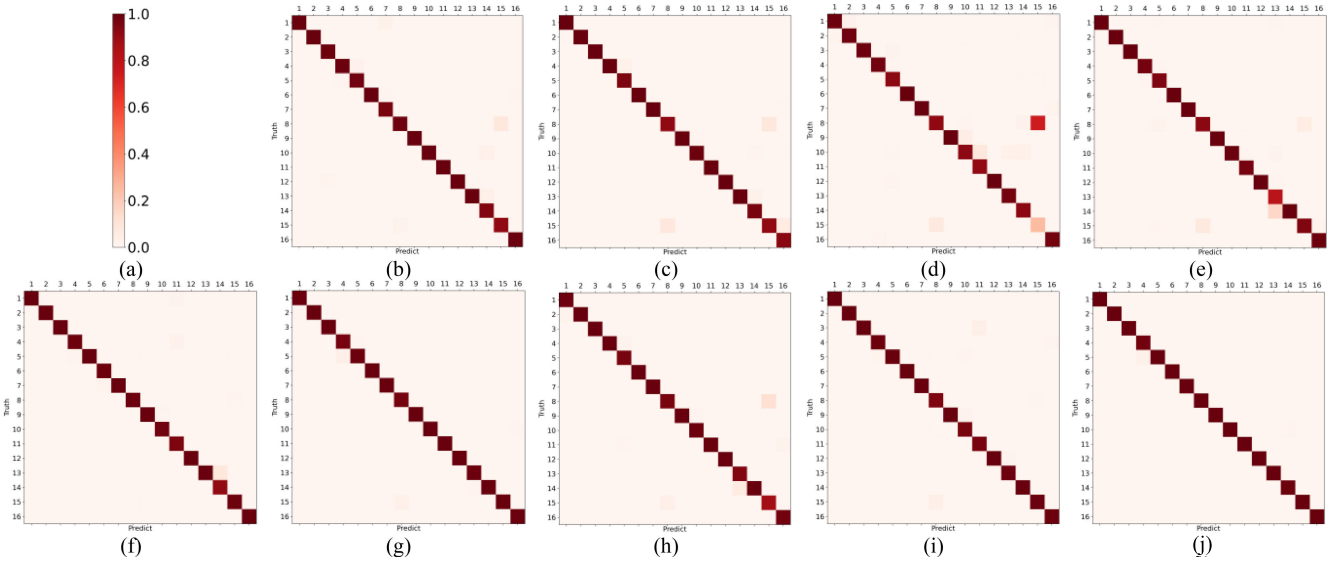

FIGURE 12. Visualization of different algorithm confusion matrices on SV datasets. (a) Chroma bar; (b) DFFN; (c) SSRN; (d) MSDN; (e) BAM-CM; (f) HybridSN; (g) MSRN-A; (h) MSR-3DCNN; (i) MDAN; (j) Proposed.

relatively stable at over 98%. Figure 9 shows the classification maps for each method on the SV dataset. The proposed method is obviously superior to other models, especially in the misclassification between ''Grapes_untrained'' and ''Vinyard_untrained''. The misclassification of each class of the proposed model is basically eliminated, and the boundary is smoother, which is closest to the ground truth image as a whole.

Figure 10-12 shows the confusion matrix visualization results of different methods on three datasets. It can be found that the main diagonal color of the proposed model is more consistent than other methods, which indicates that the classification accuracy of the model is higher. Overall, the quantitative and qualitative results on the three datasets lead to several conclusions: 1) The proposed model produces the highest accuracy on all three datasets, and the accuracy for each class is above 90%. This shows that the feature optimization strategy in the model can not only improve the accuracy of small sample classes, but also ensure the stability of other classes. 2) In addition to the method proposed in this paper, MSRN-A also shows good results. It can be noted that both MSRN-A and the proposed model use 3D-2D hybrid network to extract features, and introduce multi-scale and attention structure, which may be the key to extracting discriminative

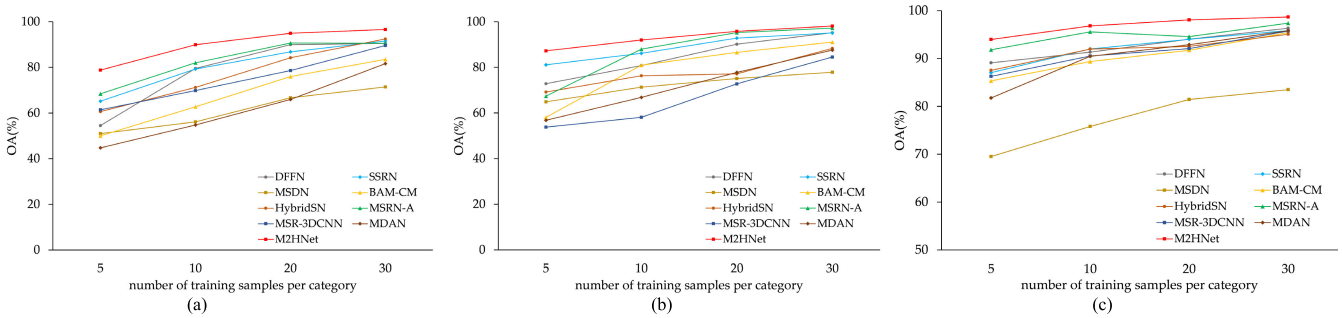

FIGURE 13. Overall classification accuracy of each method for different numbers of balanced samples. (a) IP dataset; (b) PU dataset; (c) SV dataset.

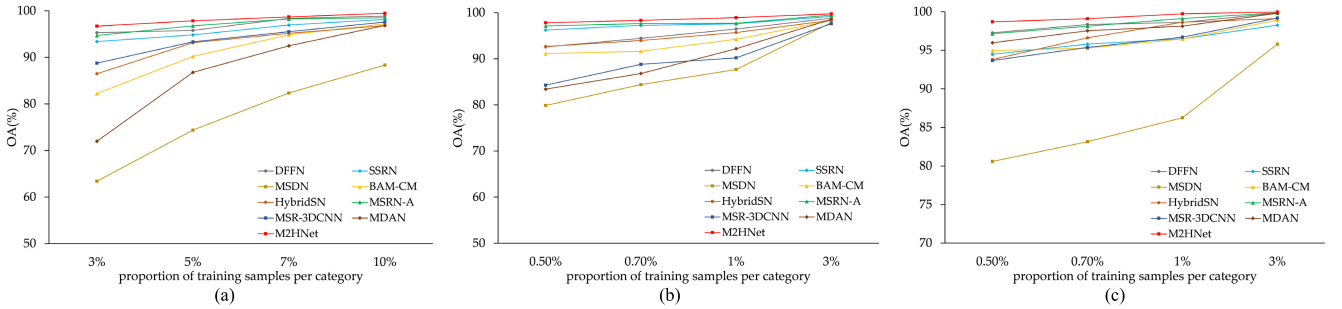

FIGURE 14. Overall classification accuracy of each method for different numbers of unbalanced samples. (a) IP dataset; (b) PU dataset; (c) SV dataset.

features in image classification. 3) By observing the results of the remaining other methods (such as MSR-3DCNN, MDAN, etc.) on three datasets, it can be found that the results of each method on different datasets are unstable and slightly lack the generalization of different data.

E. SAMPLE SIZE ANALYSIS

Evaluating the performance of models under different sample sizes is an important manifestation of model quality. Deep learning usually requires a large number of training samples to enable the model to obtain advanced learning capabilities. However, it is not easy to obtain a large number of samples in real life. Figure 10 and Figure 11 are small sample tests under balanced and unbalanced training sets, respectively. The balanced sample selection rule is the same across all three datasets, with 5, 10, 20, and 30 training samples randomly selected from each class for analysis. When the sample size is greater than or equal to half of the class, half of the class is used as the training sample. For unbalanced training sets, the training proportion of IP data is 3%, 5%, 7% and 10%, and each class retains at least one training sample. PU and SV datasets select 0.5%, 0.7%, 1% and 3% as training samples, respectively.

From Figures 10 and 11, it can be seen that the overall accuracy of all methods increases as the number of training samples increases, indicating that it makes more sense to ensure that different models are evaluated at the same sample size. In addition, the proposed model has obvious advantages in the balanced training set state, and the gap with other methods is small in the unbalanced training set state. When testing on balanced training sets, classes with fewer total samples are easier to learn, while classes with more total samples are not easy to learn because of the smaller training set. In the unbalanced training set test, different classes determine the number of training samples in the same proportion. At this time, the classes with more total sample size have sufficient training sets, while the classes with less total sample size will produce small samples or even very few samples. Therefore, the two experiments reflect the robustness of the model to different classes of feature extraction from different aspects. As shown in Figure 10, in the case of balanced training set, the proposed model is significantly higher than other methods, which indicates that when each class selects the same number of samples for training, the proposed model has the best ability to extract different classes of features. The performance of MSRN-A is slightly inferior to the proposed method, but it lacks robustness to different data and different training sample sizes. Other methods are significantly affected by sample size on IP and PU datasets, but perform better on SV datasets. MSDN performed poorly, which may be due to the small sample size in the experiment of the balanced training set. In the unbalanced sample state, as shown in Figure 11, the proposed model has obvious advantages when the proportion of training samples is the smallest. As the proportion increases, the gap between the

TABLE 7. The training and testing time of each method on three datasets.

		DFFN	SSRN	MSDN	BAM-CM	HybridSN	MSRN-A	MSR-3DCNN	MDAN	Proposed
	Parameters	424K	353K	2581K	834K	5122K	298K	3088K	460K	209K
	FLOPS	522M	112M	5069M	499M	495M	4845M	1568M	48M	274M
IP	Training Time(s)	238.98	235.80	460.74	149.85	106.44	531.84	415.01	70.99	406.84
	Test Time(s)	3.66	7.97	20.50	2.27	3.72	14.86	22.46	7.47	14.27
PU	Training Time(s)	265.50	221.31	591.99	106.33	103.36	272.81	355.77	60.11	272.54
	Test Time(s)	10.37	22.51	194.73	7.62	21.92	75.36	96.82	34.29	45.69
SV	Training Time(s)	251.33	266.36	1187.47	169.00	139.49	425.21	437.42	94.60	346.92
	Test Time(s)	13.14	34.08	257.52	14.82	30.43	92.94	119.54	44.18	77.36

results of different methods decreases rapidly. In particular, when the proportion of training samples on the PU dataset is the highest, all methods produce excellent results. In summary, the experimental results under the two training sample selection schemes show that the features extracted by the proposed model have stronger robustness and the resulting results have higher accuracy.

F. EFFICIENCY ANALYSIS

Table 7 lists the number of parameters and floating point operations (FLOPs) of each algorithm model, and the training and testing time on three datasets. From Table 7, The number of model parameters and FLOPs is not directly related to the time consumption in the model operation process. This is because convolution grouping, depth separability and computer equipment are all important factors affecting the running speed. Therefore, the parameters and FLOP of the proposed model are lower than those of SSRN and HybridSN, but the time consumption is higher than that of the two. In addition, MDAN always takes the least time. On the IP dataset, MSRN-A has the longest training time. On PU and SV datasets, MSDN takes the most time. In contrast, the proposed model performs moderately in training time on three datasets. The proposed model increases the network width and depth, and introduces the multi-scale interactive attention mechanism, which is the main reason for the increase in time consumption. But correspondingly, it is these structures that extract more robust features. Compared with the results of other methods on different data, the model in this paper performs best and most stable on three data. Therefore, considering the efficiency, accuracy and generalization, the time consumption of the proposed model is acceptable.

G. ABLATION EXPERIMENTS

In this section, different ablation experiments are carried out to further analyze the effectiveness of MFB and SMIA modules. Firstly, the proposed model is denoted as "MFB-SMIA" for ease of observation; secondly, to verify the effectiveness of multi-level feature extraction, the branch with dilated convolution in the multi-level feature extraction block is removed, and only the branch that can extract high-level abstract features is retained, denoted as "SFB-SMIA". In addition, to demonstrate the effectiveness of dilated convolution, the dilated convolution kernel in the MFB is replaced by a regular convolution kernel with the same receptive field

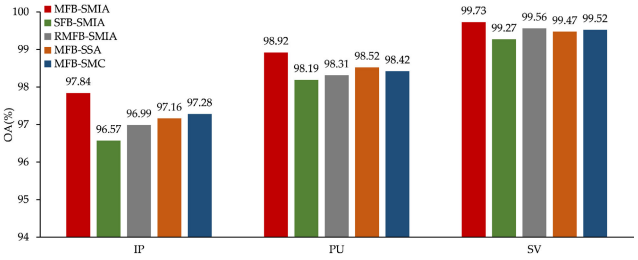

FIGURE 15. Effectiveness analysis of MFB and SMIA modules on these datasets.

size, denoted as "RMFB-SMIA". After that, to verify the effectiveness of the spatial multi-scale interactive attention module, the multi-scale and attention mechanism in SMIA are ablated respectively. The former transforms the spatial multi-scale interactive attention into the spatial single-scale attention, which is expressed as 'MFB-SSA'. The latter deletes the attention interaction mechanism and retains the spatial multi-scale feature extraction, which is expressed as 'MFB-SMC'.

Figure 12 shows the overall accuracy of ablation experiments on three datasets. It can be found from the analysis of MFB that the three datasets have a unified trend, which shows that MFB-SMIA has the best performance, followed by RMFB-SMIA, and SFB-SMIA has the lowest accuracy. Therefore, two conclusions can be drawn: 1) Feature optimization of single-scale features can significantly improve classification accuracy; 2) Dilated convolution also has a beneficial effect on improving classification performance. Analysis of the impact of SMIA reveals that the spatial multi-scale interactive attention module leads to significant performance gains on all three datasets. In contrast, single-scale attention without scale interaction (MFB-SSA) and multi-scale features without attention mechanism (MFB-SMC) perform slightly worse. This suggests that refinement of interaction features between spatial multi-scales contributes to improved classification performance. Furthermore, the module can be arbitrarily inserted into different convolutional networks without incurring a significant increase in the number of parameters, and is therefore extremely generalizable.

H. EXPERIMENTAL RESULTS OF OTHER DATASETS

To verify the generalization of the proposed model to different datasets, in this section, we selected three datasets obtained

TABLE 8. The training and testing time of each method on three datasets.

	WHU-Hi-LongKou		MUUFL		Trento	
	AA	OA	AA	OA	AA	OA
DFFN	97.51	99.08	75.48	89.69	96.47	99.08
SSRN	96.88	99.24	82.00	93.52	97.85	98.39
MSDN	97.92	99.22	74.72	90.38	95.33	97.95
BAM-CM	94.48	98.44	78.36	92.96	95.16	98.36
HybridSN	95.92	98.90	81.36	94.16	94.32	98.67
MSRN-A	97.49	99.14	81.97	94.33	97.35	98.68
MSR-3DCNN	97.86	99.18	83.50	93.51	96.41	97.93
MDAN	96.97	98.69	75.21	91.78	95.01	97.95
Proposed	98.68	99.50	87.38	95.37	98.36	99.25

by other different sensors for supplementary verification. Specifically, the WHU-Hi-LongKou dataset was collected by the Headwall Nano-Hyperspec sensor in Longkou Town, Hubei Province, China. The MUUFL dataset was collected by the (CASI)-1500 sensor over the University of Southern Mississippi Bay Park campus in Long Beach, Mississippi. The Trento dataset was taken by the AISA Eagle sensor in a rural area in southern Trento, Italy. Table 8 shows the average classification accuracy and overall classification accuracy of the three datasets on each algorithm. Obviously, the proposed model still produces the best results, with the highest accuracy on all three datasets. However, the sub-optimal results on the three datasets are shown in different algorithms, which also proves that the proposed model has a more robust feature extraction ability and strong generalization for different data.

IV. CONCLUSION

To improve the feature quality of small sample classes of hyperspectral images and improve the robustness of the model for feature extraction of different classes, we propose a hyperspectral image classification model with multi-correlation and spatial multi-scale interactive feature refinement. The model takes 3D-2D hybrid convolutional neural network as the basic framework. In the 3D structure, multi-scale features are extracted by multi-level feature extraction blocks, and a spatial multi-scale interactive attention module is designed in the 2D structure to refine multi-scale features. The proposed model can effectively extract the features of various classes on different data and solve the problem of difficult classification of unbalanced sample data. Experiments were conducted on three commonly used publicly available datasets, and the overall accuracies of 97.84%, 98.92% and 99.73% were obtained using 5%(IP), 1%(PU) and 1%(SV) training samples. Compared with other methods, the proposed model can adapt well to small sample image classification and has more stable and more accurate classification results. In addition, ablation experiments on MFB and SMIA show that they are of great significance to improve the performance of the model. This paper fully considers the spatial and spatial-spectral information, but ignores the influence of spectral dimension information. The next step will fully consider the multi-dimensional information of hyperspectral images for learning and classification, and combine the adaptive feature extraction module to optimize the information of multiple dimensions. In addition, lightweight network models have become the trend of the future, so designing highly versatile and pervasive lightweight models is the next priority.

REFERENCES

[1] F. van der Meer, "Analysis of spectral absorption features in hyperspectral imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 5, no. 1, pp. 55–68, Feb. 2004, doi: 10.1016/j.jag.2003.09.001.

[2] J. Ardouin, J. Levesque, and T. A. Rea, "A demonstration of hyperspectral image exploitation for military applications," in *Proc. 10th Int. Conf. Inf. Fusion*, Jul. 2007, pp. 1–8.

[3] I. C. C. Acosta, M. Khodadadzadeh, L. Tusa, P. Ghamisi, and R. Gloaguen, "A machine learning framework for drill-core mineral mapping using hyperspectral and high-resolution mineralogical data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4829–4842, Dec. 2019.

[4] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, p. 2659, 2020.

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

[6] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, 2015.

[7] D. Hong, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021, doi: 10.1109/MGRS.2021.3064051.

[8] J. Zhang, F. Wei, F. Feng, and C. Wang, "Spatial–spectral feature refinement for hyperspectral image classification based on attention-dense 3D-2D-CNN," *Sensors*, vol. 20, no. 18, p. 5191, Sep. 2020, doi: 10.3390/s20185191.

[9] C. Li, Z. Qiu, X. Cao, Z. Chen, H. Gao, and Z. Hua, "Hybrid dilated convolution with multi-scale residual fusion network for hyperspectral image classification," *Micromachines*, vol. 12, no. 5, p. 545, May 2021.

[10] F. Feng, Y. Zhang, J. Zhang, and B. Liu, "Small sample hyperspectral image classification based on cascade fusion of mixed spatial–spectral features and second-order pooling," *Remote Sens.*, vol. 14, no. 3, p. 505, Jan. 2022, doi: 10.3390/rs14030505.

[11] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jul. 2015, Art. no. 258619, doi: 10.1155/2015/258619.

[12] 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, doi: 10.1109/JSTARS.2014.2329330.

[13] 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, doi: 10.1109/JSTARS.2015.2388577.

[14] 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.

[15] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018, doi: 10.1016/j.isprsjprs.2017.11.021.

[16] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017, doi: 10.1109/TIP.2017.2725580.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 1–9.

[18] 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, Jun. 2020, doi: 10.1109/LGRS.2019.2918719.

[19] Z. Li, T. Wang, W. Li, Q. Du, C. Wang, C. Liu, and X. Shi, "Deep multilayer fusion dense network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1258–1270, 2020.
[20] 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, doi: 10.1109/TGRS.2017.2755542.
[21] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, p. 1068, Jul. 2018, doi: 10.3390/rs10071068.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV*. Cham, Switzerland: Springer, 2016, pp. 630–645, doi: 10.1007/978-3-319-46493-0_38.
[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[25] 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. (ICIP)*, Sep. 2017, pp. 3904–3908.
[26] Z. Li, L. Huang, and J. He, "A multiscale deep middle-level feature fusion network for hyperspectral classification," *Remote Sens.*, vol. 11, no. 6, p. 695, 2019, doi: 10.3390/rs11060695.
[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn, Represent.*, San Juan, PR, USA, 2016, pp. 1–13.
[28] H. Gao, Z. Chen, and C. Li, "Hierarchical shrinkage multiscale network for hyperspectral image classification with hierarchical feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5760–5772, May 2021.
[29] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Nov. 2017, doi: 10.1109/TGRS.2016.2616355.
[30] J. Feng, J. Chen, L. Liu, X. Cao, X. Zhang, L. Jiao, and T. Yu, "CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019, doi: 10.1109/JSTARS.2019.2900705.
[31] W. Li, C. Chen, M. Zhang, H. Li, and Q. Du, "Data augmentation for hyperspectral image classification with deep CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 593–597, Apr. 2019, doi: 10.1109/LGRS.2018.2878773.
[32] W. Wang, X. Liu, and X. Mou, "Data augmentation and spectral structure features for limited samples hyperspectral classification," *Remote Sens.*, vol. 13, no. 4, p. 547, Feb. 2021.
[33] J. Li, Q. Du, Y. Li, and W. Li, "Hyperspectral image classification with imbalanced data based on orthogonal complement subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3838–3851, Jul. 2018, doi: 10.1109/TGRS.2018.2813366.
[34] L. Fang, W. Zhao, N. He, and J. Zhu, "Multiscale CNNs ensemble based self-learning for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1593–1597, Sep. 2020, doi: 10.1109/LGRS.2019.2950441.
[35] L. Huang, Y. Chen, and X. He, "Weakly supervised classification of hyperspectral image based on complementary learning," *Remote Sens.*, vol. 13, no. 24, p. 5009, Dec. 2021, doi: 10.3390/rs13245009.
[36] Z. He, H. Liu, Y. Wang, and J. Hu, "Generative adversarial networks-based semi-supervised learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 10, p. 1042, Oct. 2017, doi: 10.3390/rs9101042.
[37] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Apr. 2019, doi: 10.1109/TGRS.2019.2908756.
[38] F. Xie, Q. Gao, C. Jin, and F. Zhao, "Hyperspectral image classification based on superpixel pooling convolutional neural network with transfer learning," *Remote Sens.*, vol. 13, no. 5, p. 930, Mar. 2021, doi: 10.3390/rs13050930.
[39] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, p. 1307, 2019.
[40] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020, doi: 10.3390/rs12030582.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–11.
[42] Y. Chen, Y. Kalantidis, J. Li, and S. Y. Feng, "A2-Nets: Double attention networks," in *Proc. NeurIPS*, vol. 31, 2018, pp. 1–10.
[43] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[44] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021, doi: 10.1109/TGRS.2020.3008286.
[45] Y. Jiang, Y. Li, and H. Zhang, "Hyperspectral image classification based on 3-D separable ResNet and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1949–1953, Dec. 2019.
[46] F. Feng, S. Wang, C. Wang, and J. Zhang, "Learning deep hierarchical spatial–spectral features for hyperspectral image classification based on residual 3D-2D CNN," *Sensors*, vol. 19, no. 23, p. 5276, Nov. 2019.
[47] Q. Liu, L. Xiao, F. Liu, and J. Huan, "SSCDenseNet: A spectral–spatial convolutional dense network for hyperspectral image classification," *Acta Electronica Sinica*, vol. 48, no. 4, pp. 751–762, Nov. 2020.
[48] H. Gong, Q. Li, C. Li, H. Dai, Z. He, W. Wang, H. Li, F. Han, A. Tuniyazi, and T. Mu, "Multiscale information fusion for hyperspectral image classification based on hybrid 2D-3D CNN," *Remote Sens.*, vol. 13, no. 12, p. 2268, Jun. 2021, doi: 10.3390/rs13122268.
[49] J. Liu, K. Zhang, S. Wu, H. Shi, Y. Zhao, Y. Sun, H. Zhuang, and E. Fu, "An investigation of a multidimensional CNN combined with an attention mechanism model to resolve small-sample problems in hyperspectral image classification," *Remote Sens.*, vol. 14, no. 3, p. 785, Feb. 2022, doi: 10.3390/rs14030785.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1026–1034.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[52] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
[53] 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, Aug. 2019, doi: 10.1109/TGRS.2019.2925615.
[54] H. Dong, L. Zhang, and B. Zou, "Band attention convolutional networks for hyperspectral image classification," 2019, *arXiv:1906.04379*.
[55] X. Zhang, T. Wang, and Y. Yang, "Hyperspectral images classification based on multi-scale residual network," 2020, *arXiv:2004.12381*.
[56] H. Xu, W. Yao, L. Cheng, and B. Li, "Multiple spectral resolution 3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 7, p. 1248, Mar. 2021, doi: 10.3390/rs13071248.

YAMEI MA was born in Zhoukou, Henan, China, in 1999. She received the B.E. degree from Henan Polytechnic University, Jiaozuo, China, in 2020, where she is currently pursuing the M.S. degree. Her research interests include deep learning and hyper-spectral image classification.

SHUANGTING WANG was born in Xingyang, Henan, China, in 1962. He received the B.S. and M.S. degrees from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 1983 and 1990, respectively. He is currently a Professor with Henan Polytechnic University. He is mainly engaged in teaching and research of photogrammetry and remote sensing. He has presided over or participated in more than 60 research projects and published more than 70 articles. His main research interests include photogrammetry, aerial and space photography engineering, remote sensing, and image processing and recognition.

WEIBING DU was born in Zhengzhou, Henan, China, in 1985. He received the B.S. degree in surveying and mapping engineering from the Henan University of Urban Construction, Pingdingshan, in 2004, and the Ph.D. degree in surveying and mapping science and technology from Henan Polytechnic University, Jiaozuo, China, in 2014. He is currently an Associate Professor with Henan Polytechnic University. He is mainly engaged in teaching and scientific research of photogrammetry and remote sensing. His research interests include multi-source remote sensing information processing, UAV photogrammetry and remote sensing, and remote sensing monitoring of glacier spatial and temporal changes.

XIAOQIAN CHENG was born in Shijiazhuang, Hebei, China, in 1984. She received the B.S. degree from the China University of Petroleum, Qingdao, in 2007, the M.S. degree from Wuhan University, Wuhan, in 2009, and the Ph.D. degree from Henan Polytechnic University, Jiaozuo, in 2021. She is mainly engaged in teaching and research of photogrammetry and remote sensing. Her research interests include urban remote sensing and spatiotemporal data analysis and application.