A Novel Hyperspectral Image Classification Model Using Bole Convolution with Three-Directions Attention Mechanism: Small sample and Unbalanced Learning

Weiwei Cai, *Member, IEEE,* Xin Ning, *Senior Member, IEEE*, Guoxiong Zhou, Xiao Bai and Yizhang Jiang, *Senior Member, IEEE,* Wei Li, *Senior Member, IEEE*, Pengjiang Qian, *Senior Member, IEEE*

***Abstract*—Currently, the use of rich spectral and spatial information of hyperspectral images to classify ground objects is a research hotspot. However, the classification ability of existing models is significantly affected by its high data dimensionality and massive information redundancy. Therefore, we focus on the elimination of redundant information and the mining of promising features and propose a novel bole convolution neural network with a tandem three-directions attention mechanism (BTA-Net) for the classification of hyperspectral image. A new bole convolution is proposed for the first time in this algorithm, whose core idea is to enhance effective features and eliminate redundant features through feature punishment and reward strategies. Considering that traditional attention mechanisms often assign weights in a one-direction manner, leading to a loss of the relationship between the spectra, a novel three-directions (horizontal, vertical, and spatial directions) attention mechanism is proposed, and an addition strategy and a maximization strategy are used to jointly assign weights to improve the context sensitivity of spatial spectral features. In addition, we also designed a tandem three-directions attention mechanism module and combined it with a multi-scale bole convolution output to improve classification accuracy and stability even when training samples are small and unbalanced. We conducted scene classification experiments on four commonly used hyperspectral datasets to demonstrate the superiority of the proposed model. The proposed algorithm achieves competitive performance on small samples and unbalanced data, according to the results of comparison and ablation experiments. The source code for BTA-Net can be found at https://github.com/vivitsai/BTA-Net.**

***Index Terms*—Hyperspectral image classification, Bole convolution (BC), Three-directions attention mechanism (TDA), Punishment and reward strategy, Neural networks.**

I. Introduction

With the advancements made in hyperspectral remote sensing technology, its use has recently seen a sharp increase in such fields as agricultural monitoring [1-3], resource detection [4-6], medical diagnosis [7-8] and environmental protection [9]. Hyperspectral images (HSIs) are rich in a host of spectral and spatial information extracted from a large number of ground objects, which can be used to identify and classify ground objects.

The main challenges faced in hyperspectral images (HSIs) classification tasks [10] include the high dimensionality of the images, large volumes of data, feature redundancy, and spectrum correlation. The use of spectral information for HSIs classification emerged as a popular approach in early research on the subject [13,14]; furthermore, dimensionality reduction [15,16] and feature selection have been frequently used to reduce the high dimensionality of spectral data. Many researchers have tried to improve the model's classification performance by adding local spatial links [17,18], with varying degrees of success. However, most of these algorithms rely on data preprocessing and manual feature extraction, which not only rely heavily on prior knowledge but also have limited generalization capability, making it difficult to extract representative discriminative features.

However, the attention mechanism can assign attention weights to features to better extract more discriminative features, so *Xue et al.* [19] proposed an attention-based second-order pooling (A-SOP) operator for Discriminative and representative features are modeled. *Cui et al.* [20] proposed a more concise and efficient spatial attention module to address the issue of a large number of redundant computations in existing spatial attention modules. *Yu et al.* [21] developed a feedback attention module to improve the model's classification ability by enhancing the attention map with semantic information from high-level dense models. They also improved the spatial attention module by considering multi-scale spatial features. By integrating the attention mechanism into ResNet, *Haut et al.* [22] proposed a visual attention - driven HSIs classification model that could better fit the spectral

This work was supported in part by the National Natural Science Foundation of China under Grant 62172192, 62171203, 61901436, and U20A20228. (*Corresponding authors:Pengjiang Qian and Xin Ning.*)

Weiwei Cai, Yizhang Jiang and Pengjiang Qian are with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122,China.(vivitsai@ieee.org,yzjiang@jiangnan.edu.cn;qianpjiang@jiangnan.edu.cn)

Xin Ning is with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China.

Guoxiong Zhou is with School of Computer Information and Engineering, Central South University of Forestry and Technology, Changsha, 102208, China.

Xiao Bai is with School of Computer Science and Engineering, Beihang University, Beijing,100191, China. (baixiao@buaa.edu.cn)

Wei Li is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China. (leewei36@gmail.com)

information contained in hyperspectral data. *Hang et al.* [23] designed a spatial and spectral attention sub-network to assist the CNN classifier focus on more distinguishable channels or positions, achieving higher performance compared with CNNs. In addition, *Sun et al.* [24] considered redundant features to weaken the distinguishing ability of spectral spatial characteristics and decrease the classification performance. The authors were able to capture more discriminative spectral spatial characteristics in the focus area of the HSIs data by introducing an attention mechanism to suppress the influence of interfering pixels. The results of the classification show that the attention module aids in improving classification accuracy. The above attention model is characterized by assigning weights only in a one-direction manner, and not paying enough attention to effective features with rich semantic information, so this drives us to try to assign attention weights from multiple directions.

Although the above algorithms have achieved excellent results for HSIs classification, owing to the high dimensionality, large amount of data, and redundant features of HSIs, these algorithms have a lot of model parameters and are computationally complex, so they take a lot of time and resources. Although some researchers have successfully used PCA to reduce the dimensionality of HSIs [16],[24], PCA cannot effectively eliminate redundant features of HSIs. In addition, *Feng et al.* [25] developed a thermonuclear Euclidean distance affinity matrix to map high-dimensional data to a low-dimensional space, and proposed a graph-based discriminative method with spectral similarity to reduce the dimensionality of HSIs. They also generated reliable low-dimensional features by incorporating low-rank representation and projection learning into the model. *Deng et al.* [26] proposed tensor algebra, a multilinear algebra-based supervised dimensionality reduction method. Even though the methods described above have been proven to be effective, eliminating redundant data from HSIs is still difficult.

Therefore, inspired by the above work, this study considers the characteristics of CNNs and proposes a new bole CNN (BC). Bole in Chinese culture refers to people who are good at identifying the quality of horses; in this study, it is used to refer to neural networks that are good at identifying features. Its core idea is to eliminate redundant features and enhance promising features through a feature punishment and reward strategy. Specifically, for feature maps, this study first uses the sigmoid function to map its weight to the interval (0,1) (for example: 0.1 or 0.5), and then a threshold is set (for example, 0.2) to punish and eliminate features below the threshold and enhance and reward features above it. It is worth explaining in advance that BC only requires a small number of neurons, which results in a better classification performance. This study also considers the limitation that traditional attention mechanisms often assign weights in a single direction, which leads to loss of relationship between the spectra, and proposes a novel three-direction (horizontal, vertical, and spatial directions) attention mechanism module, based on the weight addition and a maximization strategy, to jointly assign weights to improve the context sensitivity of spatial spectral features. In addition, in this study we also design a tandem three-direction attention mechanism module and fuse it with the output of multi-scale BC, which significantly improves classification ability and stability even with small and unbalanced training samples.

The main contributions are as follows:

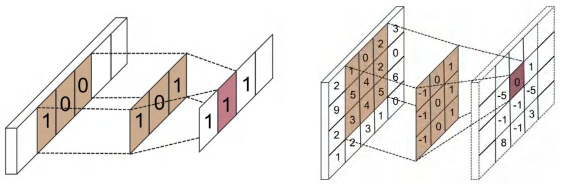
1) A brand-new Bole Convolutional (BC) neural network is proposed, whose core idea is to eliminate redundant features and enhance promising features through a feature punishment and reward strategy. Specifically, we penalize features below the threshold and eliminate them, and reward features above the threshold and enhance them. In addition, the BC requires only a small number of neurons to achieve better classification ability and can significantly reduce the model parameters and computational complexity.
2) Considering the traditional attention mechanism only assigns weights in a single-direction manner, which leads to the loss of the relationship between the spectra, this paper proposes a novel three-direction (horizontal, vertical, and spatial directions) attention mechanism (TDA). Horizontal and vertical attentions are used to obtain the cross weights of the features, whereas spatial direction attention captures the spatial feature weights of the feature map.
3) This paper also proposes an addition strategy and a maximum weight strategy for the weights of the attention mechanism. The addition strategy is based on our previous work [27]. First, the horizontal and vertical weight coefficients are multiplied, and then the spatial attention weight coefficients are added. The maximization strategy is employed to take the largest of three-direction attention weights. Combining the two strategies can improve the model's context sensitivity to spatial spectral features.
4) We evaluate the BTA-NET algorithm on four popular HSI datasets, and the results of comparison and ablation experiments show that the algorithm superior several well-known methods in terms of model performance, number of parameters used, and computational complexity.

The rest of the paper is summarized as follows: The relevant work is discussed in Section II. In Section III, the proposed BTA-Net algorithm is explained in detail. In Section IV, we discuss comparison and ablation experiments. Finally, the conclusions are drawn in Section V.

II. Related Work

A. 1D and 2D Convolutions

The convolutional layer, as a feature extractor, learns the feature representation of the input image. Each neuron, arranged into a feature map in the convolutional layer, has a receptive field in the feature map, which is connected through a set of trainable weights to the neuron neighborhood in the previous layer. The input data and feature map weights are convoluted to create a new feature map, and activated by a nonlinear activation function. The kernel of a 2D-CNN is a matrix, whereas the kernel of a 1D-CNN is a 1-dimensional vector (as shown in Fig. 1). The equations for calculating the



(a) (b)

Fig. 1. 1D and 2D convolution model. (a) 1D convolution model. (b) 2D convolution model.

two convolutions are as follows:

$$h_i^k = f\left(\sum_m \sum_{p=1}^{w} W_i^{p} * x_i^{(k+p)} + b_i\right) \tag{1}$$

$$h_{ij}^{kl} = f\left(\sum_m \sum_{p=1}^{h} \sum_{q=1}^{w} W_{ij}^{pq} * x_{ij}^{(k+q)(l+q)} + b_{ij}\right) \tag{2}$$

where h_i^k and h_{ij}^{kl} are the output at position (k, l) in the j th feature map in the i th layer, f represents the activation function, ($h \times w$) represents the size of the convolution kernel, (p , q) are the indexes of kernel, and m represent the index of the feature maps, and b and W are the biases and weights of the kernel, respectively.

B. Attention Mechanism

The attention mechanism can be used during the processing of HSIs data by the neural network model to focus on the more important set of information among the numerous input data, reduce attention to other information, and improve the accuracy and efficiency of the processing task. First, consider an attention variable $z \in [1,N]$ to represent the index position of the concerned information, that is, $z = i$ to represent the selection of the input i th information, and then calculate the probability a_i of the selection of the input i th information given q and X . X represents the input data, and q represents the query vector. The equation for calculating probability a_i is as follows:

$$a_i = p(z = i | X,q) \tag{3}$$

The probability vector formed by a_i is called the attention distribution. $s(x_i,q)$ is the attention scoring function, and its equation of calculation is as follows:

$$s(x_i,q) = v^T \tanh(Wx_i + Uq) \tag{4}$$

where U , W and v are learnable weight parameters.

The softmax function is then used to convert the attention score to a numerical value. On the one hand, it can be normalised to produce a probability distribution with a sum of all weight coefficients of 1. On the other hand, the softmax function can be used to emphasise the importance of key elements.

$$\begin{aligned} a_i &= soft\max\left(s(x_i,q)\right) \\ &= \frac{\exp\left(s(x_i,q)\right)}{\sum_{j=1}^{N} \exp\left(s(x_i,q)\right)} \end{aligned} \tag{5}$$

Finally, we perform a weighted summation according to the weight coefficient:

$$Attention(X,q) = \sum_{i=1}^{N} a_i x_i \tag{6}$$

In recent studies [22],[23],[27], the attention mechanism algorithm has achieved fruitful results in HSIs classification tasks. We believe that the continuous improvement of the attention mechanism algorithm will help enhance the classification ability of the HSIs classifier.

C. Dimensionality and Parametric Reduction

High dimensionality, which is a prominent feature of HSIs, always affects the performance of classification models. Scholars have attempted to decrease the dimensionality of HSIs and the number of parameters in classification models. To overcome the problem of high data volume, *Reshma et al.* [28] adopted inter-band block correlation coefficient technology and performed QR decomposition and singular value decomposition to decrease the size of HSIs without affecting key information. *Wang et al.* [29] proposed a dimensionality reduction model that couples a thin intrinsic modal function dictionary with a weighted low-rank representation. Compared with the traditional PCA method, it can retain more structural information. *Xu et al.* [30] applied the linear discriminant analysis method of superpixels to capture spatial similarity and proposed a spatial spectrum dimensionality reduction method based on superpixels to solve the limitation of other methods, that is, ignoring spectral similarity. Although these methods have achieved excellent performance, few scholars have studied the elimination of redundant HSIs features from the convolution process.

D. Related methods

Li et al. [37] proposed a positional embedding and importance aggregation BW module to obtain more discriminative features, which obtains remote dependencies in one spatial direction while preserving accurate positional information in the other. To bridge the gap between clear and hazy intrinsic similarity matrices, *Pang et al.* [38] proposed a novel interference suppression approach that extracts interference information at the feature level. *Liu et al.* [39] provide a unified approach for cross-domain classification that minimizes the structural risk of labeling source data, allows statistical adaptation using the Maximum Mean Difference (MMD) criterion, and employs the Nyström approach flexibly to explore domain geometry. Connections are utilized to create adaptable models that perform exceptionally well on classification tests. *Hang et al.* [40] proposed spectral and spatial attention networks for spectral and spatial classification to build a novel attention-assisted CNN model capable of extracting discriminative features. The RSSAN suggested by *Zhu et al.* [41] is capable of stressing relevant bands for classification while suppressing unnecessary bands, and the proposed spatial attention module achieves outstanding classification performance. *Roy et al.* [42] suggested an attention-based adaptive spectrum space kernel improved residual network that extracted spectral space features simultaneously using improved 3-D ResBlocks, greatly increasing the classification model's feature mining capacity.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3201056

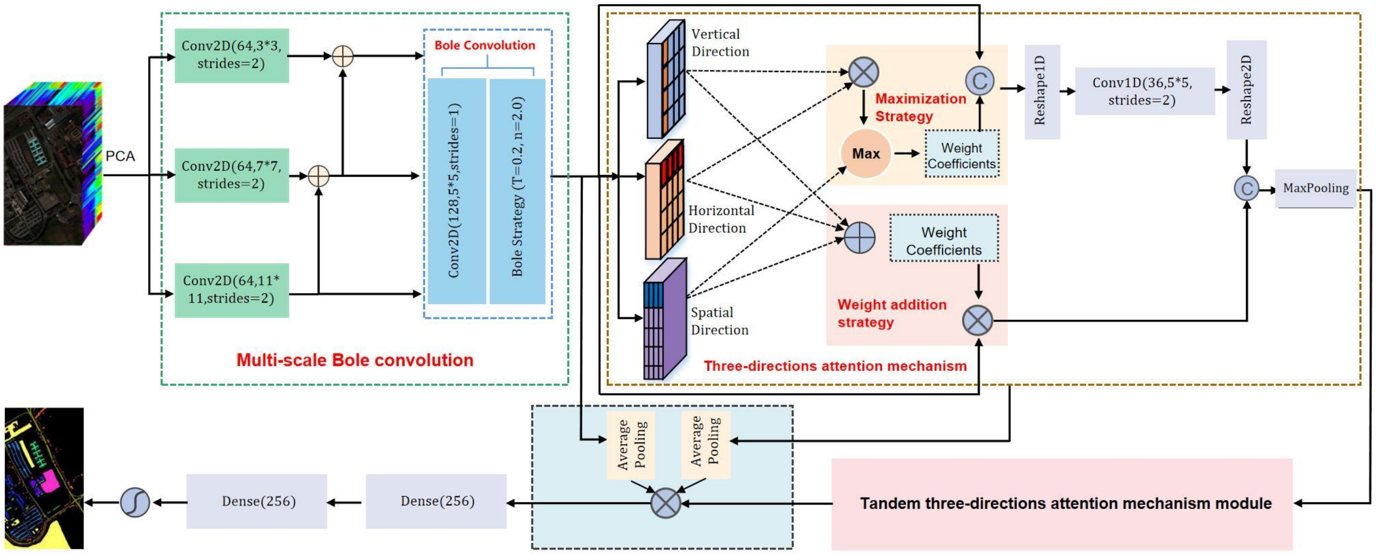

PCA
Conv2D(64,3*3, strides=2)
Conv2D(64,7*7, strides=2)
Conv2D(64,11*11,strides=2)
Bole Convolution
Conv2D(128,5*5,strides=1)
Bole Strategy (T=0.2, n=2.0)
Multi-scale Bole convolution
Vertical Direction
Horizontal Direction
Spatial Direction
Maximization Strategy
Max
Weight Coefficients
Weight Coefficients
Weight addition strategy
Three-directions attention mechanism
C
C
Reshape1D
Conv1D(36,5*5, strides=2)
Reshape2D
MaxPooling
Average Pooling
Average Pooling
Tandem three-directions attention mechanism module
Dense(256)
Dense(256)

Fig. 2. Schematic of the proposed BTA-Net algorithm for hyperspectral image scene classification. The symbol ⊗ denotes the element-wise multiplication operation, the symbol ⊕ represents the element-wise addition operation, and the symbol represents the feature concatenatation operation. Finally, represents the softmax function. Bole convolution performs preliminary brush selection of features, which can provide better feature input for the three-directions attention mechanism.

The attention mechanism is used in all of the above-mentioned HSI classification models based on feature mining and attention mechanism to build a classification model with excellent performance from a single spatial direction or multiple spatial directions, demonstrating that the multi-directional attention is also effective in the HSI classification task.

III. Methodology

The proposed BTA-Net model is presented in Fig. 2. To begin, we use PCA to reduce the HSIs' spectral redundancy. The output features of the multi-scale convolution kernel are then fed into the BC to further eliminate redundant information and improve efficient features. The tandem three-direction attention mechanism is then used to capture more representative deep features and obtain relationships between features. Finally, the BC is combined with the output of the three-direction attention mechanism, and the softmax function is applied to achieve HSIs scene classification.

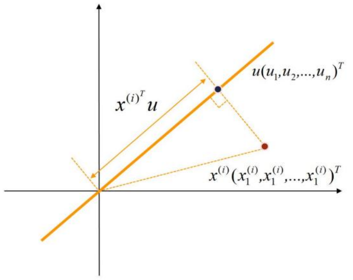


Fig. 3. The data projection for PCA principal component analysis.

A. Principal Components Analysis

Hundreds of continuous spectral bands with a high degree of correlation exist in the original HSIs. This causes the "Hughes" phenomenon as well as a large amount of redundant information in the original data. It is necessary to decrease the dimensionality of the original HSIs data prior to image modeling and analysis. In this paper, the PCA is used to decrease the dimensionality, which not only retains a large amount of feature information in the original HSIs but also reduces the redundancy of the original data. PCA's main goal is to convert m-dimensional features into n-dimensional features, where $n < m$; these n features are linearly independent orthogonal features, where the first principal component vector is the direction of the maximum data variance; the projection of the data is shown in Fig. 3. The equation to calculate covariance, Cov, according to the theory of maximizing variance is as follows:

$$Cov = \frac{1}{m}\sum_{i=1}^{m}\left((x^{(i)})^T u\right)^2 = u^T\left(\frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T\right)u \tag{7}$$

where x^i is a single sample, $u = (u_1, u_2, ..., u_n)^T$ is the projection vector, and n is the number of features. We need to find the eigenvector u corresponding to the eigenvalue with the largest covariance, that is, the first principal component, and so on.

B. Bole Convolution

A difficulty in HSIs classification is that the space features of HSIs have high dimensionality, few samples, high feature redundancy, and long duration of operation. It contains a lot of information about the ground objects' spatial position, structure, and spectral properties, which has a big impact on the classification model's ability to identify and classify them. A slew of redundant features not only degrades the classifier's performance, but also makes feature mining and feature selection more difficult in the neural network model, raising the computational cost. As discussed in the previous section, this study uses PCA to decrease the dimensionality of HSIs data; however, PCA is not suited to deal with redundant information. In addition, feature selection also affects the performance of the classifier. Therefore, to alleviate these issues, we propose a brand-new Bole Convolution. Next, we introduce the structure of the BC and explain feature punishment and reward strategies. In addition, the multi-scale BC convolution is introduced.

1)Overview of the Bole Convolution: Fig. 4 shows the schematic of the Bole convolution used in this study. For the feature map after convolution, we first use the sigmoid function to map each feature point of the convoluted feature map to $(0,1)$. The equation for this calculation is as follows:

$$S(m1_{ij}^{k}) = \frac{1}{1+\exp(-m_{ij}^{k})} \tag{8}$$

where m_{ij} represents the feature point of the i-th row and j-th column of the feature map m, and k is the index of the feature maps. Second, considering that the feature points after the hyperspectral image convolution contain semantic information even if they are negative, this paper does not intend to use the $relu$ function to eliminate the negative values of the feature map. We believe that the semantic features of the negative part of the feature map are not linear. Therefore, the second branch of this paper uses the elu function to process the convolutional feature map, and the equation of calculation is as follows:

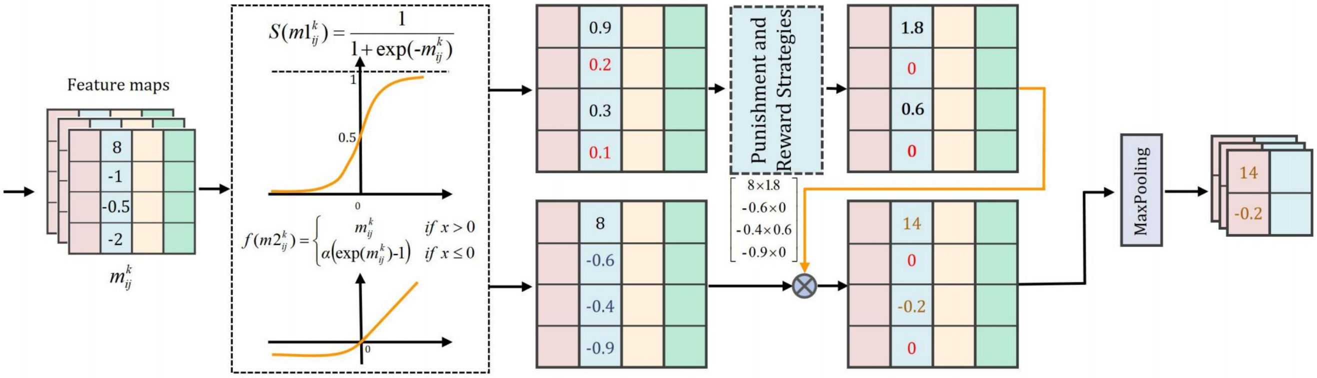

Feature maps
8
-1
-0.5
-2
m_{ij}^{k}
$S(m1_{ij}^{k}) = \frac{1}{1+\exp(-m_{ij}^{k})}$
1
0.5
0
$f(m2_{ij}^{k}) = \begin{cases} m_{ij}^{k} & if\ x>0 \\ \alpha(\exp(m_{ij}^{k})\text{-}1) & if\ x\le 0 \end{cases}$
0
0.9
0.2
0.3
0.1
8
-0.6
-0.4
-0.9
Punishment and Reward Strategies
1.8
0
0.6
0
8×1.8
-0.6×0
-0.4×0.6
-0.9×0
14
0
-0.2
0
MaxPooling
14
-0.2

Fig. 4. Schematic of the proposed Bole convolution.The symbol $\otimes$ represents the element-wise multiplication operation.

$$f(m2_{ij}^{k}) = \begin{cases} m_{ij}^{k} & if\ m_{ij} > 0 \\ \alpha\left(\exp(m_{ij}^{k})\text{-}1\right) & if\ m_{ij} \le 0 \end{cases} \tag{9}$$

where α is an adjustable parameter, which controls when the negative part of elu saturates. Next, we fuse the features of the output feature maps using the above two branches, that is, the Hadamard product of feature maps $m1'$ and $m2'$. The above-mentioned fusion of the output feature maps can be expressed as follows:

$$\left(m1'{*}m2'\right) = m1_{ij}^{k}\, m2_{ij}^{k} \tag{10}$$

where $m1'$ and $m2'$ represent the output feature maps of the first and second branches, and k represents the feature map's index. Finally, maximum pooling is used to obtain the filtered and enhanced features.

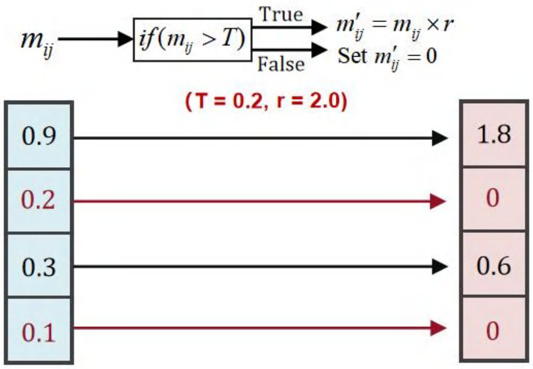


Fig.5. Schematic of the proposed feature punishment and reward strategies.

2)Feature Punishment and Reward Strategies: Punishment and reward is a well-known, efficient, and practical method used in education and management. This paper introduces punishment and reward mechanisms for the processing of features (as shown in Fig.5). The core idea is to set a threshold T, $0 \le T \le 1$, to penalize features below T and reward the features above T. Specifically, for features lower than T, reset to 0 to eliminate the feature, and for features higher than T, reward r times. The equation for this calculation is as follows:

$$bole(m_{ij}^{k}) = \begin{cases} (m_{ij}^{k})' = 0 & if\ \ \frac{1}{1+\exp(-m_{ij}^{k})} < T \\ (m_{ij}^{k})' = m_{ij}^{k} \times r & if\ \ \frac{1}{1+\exp(-m_{ij}^{k})} \ge T \end{cases} \tag{11}$$

where $T \in [0,1]$. r represent the reward coefficient, $r \in (1,+\infty)$. k is the index of the feature map, and m_{ij} represents the feature point in the i-th row and j-th column of the feature map. In the experiment, it is tedious to automatically find the optimal T and n. Considering that the actual situation of HSIs classification task research does not require adaptive critical value, this paper does not conduct research on adaptive critical values T and r. In Section 4, we obtained the optimal parameters of the above critical values through ablation experiments, and proved the effectiveness of the punishment and reward mechanism for the HSIs classification task.

3) *Multi-scale Bole Convolution:* Three multi-scale convolution kernels are used in this study, which have two advantages: First, multi-scale convolution kernels have the advantage of being able to extract HSIs features of various

scales using convolution kernels of various sizes, allowing the filter to extract and learn more features about HSIs. Second, they can learn the filter parameters (weight and offset) overtime to find the best value that is closest to the label. This study employs a multi-scale convolution kernel to create multiple filters in a single convolution layer, allowing for more diverse weight and bias learning, as well as extraction and learning of the semantic characteristics of HSIs data.

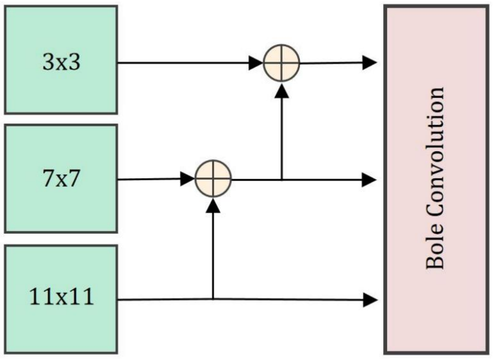


Fig. 6. Schematic of the proposed multi-scale Bole convolution (BC).

As reported by Res2Net [31], it is important to represent features in multiple scales. Fig.6 uses three sizes of 11×11, 7×7 and 3×3 convolution kernels to extract features, and the three extracted feature maps are defined as $m^k_{11\times11}$, $m^k_{7\times7}$ and $m^k_{3\times3}$ respectively. The calculation equation for the output of the multi-scale Bole convolution is as follows:

$$B_{11\times11} = bole\left(m^k_{11\times11}\right) \tag{12}$$

$$B_{7\times7} = bole\left(add[m^k_{11\times11}, m^k_{7\times7}]\right) \tag{13}$$

$$B_{3\times3} = bole\left\{add[m^k_{11\times11}, m^k_{7\times7}, m^k_{3\times3}]\right\} \tag{14}$$

where $bole()$ represents the feature penalty and reward function of Bole convolution, $add()$ represents the element-wise addition operation, and k is the index of the feature map. Finally, the multi-scale Bole convolution's output features are fused, and the calculation formula is as follows:

$$F_{BC} = concatenate[B_{11\times11}, B_{7\times7}, B_{3\times3}] \tag{15}$$

where $concatenate()$ represents feature concatenation and fusion operation.

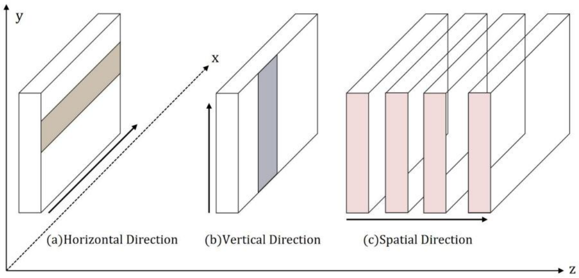


Fig. 7. Schematic of the proposed three-directions attention mechanism (TDA).

C. Three-Directions Attention Mechanism

The attention mechanism has yielded positive results in HSI classification work as a bionic technology based on human visual features. We previously developed a cross-attention mechanism that can create weighted features in both the horizontal and vertical orientations [27], which we successfully applied to hyperspectral image classification tasks, increasing the model's capacity to classify hyperspectral images. Local key areas are frequently localized in a limited area in hyperspectral images with shallow features. When the learning model comes across some mixed pixels or pixels on the classification boundary, it can only classify correctly when the correct pixel features are detected. These features are frequently found in a small area near the input pixel. Due to the presence of some irrelevant regions in the feature vectors, if the CNN model performs feature extraction on all feature vectors, it may produce sub-optimal results. Furthermore, the spectral properties of distinct ground objects in the HSIs are comparable. Within the cluster, the spectral information is largely similar, and the difference between the clusters is minimal. It's also vital to get the spatial aspects of the HSIs data right. To increase classification accuracy, even more, this research offers a novel three-directions attention mechanism (TDA) based on the prior work's horizontal and vertical cross-attention mechanisms, in which we also design spatial direction attention. TDA is depicted schematically in Fig.7.

1)Three-Directions Attention: The proposed three-directions attention mechanism obtains attention weights in the horizontal, vertical and spatial directions. Set the feature map extracted from the previous Conv layer as $m^h_{i,j} \in R^{H\times W}$, where H and W represent the feature map's height and width, and $m^k_{i,j}$ into the horizontal attention module to get the attention weight. The following is the procedure for calculating:

$$Att_h = \frac{\exp(W_h m_{i,j} + b_h)}{\sum_{i,j}\exp(W_h m_{i,j} + b_h)} \tag{16}$$

where W_h and b_h are the dense layer's weight parameters, and Att_h is the horizontal direction's attention coefficient.

For the vertical attention mechanism, we transpose the feature map's matrix to obtain the feature map in the vertical direction. The following is the calculation formula:

$$m^v_{j,i} = (m^v_{i,j})^T \tag{17}$$

where $m^v_{j,i}$ denotes a feature map that has been flipped vertically. Input it into the vertical attention module in the same way to get the vertical attention weight. The weight coefficient is calculated in the following way:

$$Att_v = \frac{\exp(W_v m_{j,i} + b_v)}{\sum_{j,i}\exp(W_v m_{j,i} + b_v)} \tag{18}$$

For the spatial direction attention mechanism, we focus on the weight distribution between feature maps to further obtain spatial features. Assuming that the feature of the input spatial direction attention module is $(m^s_{i,j})^k \in R^{H\times W\times k}$,

$(m_{i,j}^{s})=[(m_{i,j}^{s})^{1},(m_{i,j}^{s})^{2},...,(m_{i,j}^{s})^{k}]$ and k is the feature map index, the following formula is used to calculate the weight coefficient of spatial direction attention:

$$Att_{s}=\frac{\exp(W_{s}(m_{i,j}^{s})^{k}+b_{s})}{\sum_{i,j,k}\exp(W_{s}(m_{i,j}^{s})+b_{s})} \tag{19}$$

where Att_{s} represents the attention coefficient in the spatial direction.

2)Weight Addition and Maximization Strategy:

We proposed two strategies to deal with the weight coefficients of the three directions of attention obtained in the previous step, namely the weight maximization and addition strategy. For the weight maximization strategy, In the previous work, we multiplied the weight coefficients of horizontal and vertical attention and proved its good performance through experiments. Therefore, we proposed a weight addition strategy in the current work, specifically, adding the spatial direction's attention weight coefficient to the previous basis to better obtain the spatial characteristics. The equation to calculate the output coefficient of the weight addition strategy is as follows:

$$Att_{add}=\left(Att_{h}+Att_{v}+Att_{s}\right) \tag{20}$$

where Att_{add} represents the output coefficient of the weight addition strategy. For the maximum weight strategy, we take the largest one of the three attention weight coefficients, and the calculation equation is as follows:

$$Att_{\max}=\max\left[(Att_{h}\otimes Att_{v}),Att_{s}\right] \tag{21}$$

where Att_{add} represents the output coefficient of the weight Maximization strategy. Then let Att_{add} concatenate the input feature map $m_{i,j}^{k}$ to obtain the output feature map of the weight addition strategy, and let $Att_{\max}$ multiply the input feature map $m_{i,j}^{k}$ and through Conv1D to obtain the output feature map of the weight maximization strategy, and finally we concatenated them. Therefore, the output of the overall three-directions attention mechanism is as follows:

$$F_{add}=Conv1D(Att_{add}\otimes m_{i,j}^{k}) \tag{22}$$

$$F_{\max}=concatenate\left[Att_{\max},m_{i,j}^{k}\right] \tag{23}$$

$$F_{TDA}=MP\left\{concatenate[F_{add},F_{\max}]\right\} \tag{24}$$

where F_{TDA} represents the output of the overall three-directions attention. $concatenate()$ represents feature concatenation and fusion operation, and MP is the max pooling.

3)Tandem Three-Directions Attention:

We found that the performance of two tandem three-directions attention modules was better in the experiment than that of one or more modules. To prove this, Section 4 will present the ablation experiment.

D. Feature fusion

In the previous steps, we have obtained the output features F_{BC} and F_{TDA} of the Bole convolution and the three-directions attention mechanism. Next, this paper will use the BC, the TDA, and the Tandem TDA. The output features are fused. In particular, the BC uses the maximum pooling preprocessing, and the three-directions attention mechanism in series uses the average pooling preprocessing, after which the final fused features pass two FC layers and a softmax function to achieve the final HSIs classification. The final fusion is calculated using the following equation:

$$F=concatenate\left[MP(F_{BC}),AP(F_{TDA}),AP\left(F_{TDA}(F_{TDA})\right)\right] \tag{25}$$

where AP represents the average pooling,and MP is the max pooling.

IV. EXPERIMENTS

In this paper, four sets of real-world HSI data were used to assess the BTA-Net algorithm's rationality and effectiveness. Tables I, II, III and IV show the colors and number of training samples corresponding to the ground objects, and Fig. 8 shows the false color images and ground truth of these datasets. First, we conducted comparative experiments with nine well-known algorithms to demonstrate the superiority of the BTA-Net algorithm. Second, we investigate how the algorithm is affected by parameters. We demonstrate the superiority of BC and three-directions attention mechanisms through ablation studies and comparative experiments. The experiments were performed on the same device, with the GPU being a GTX1080, and the software being Keras 2.2.5, TensorFlow 1.14, and Pycharm 2020.2.2. In this paper, Adam is used to optimize the training process, and the learning rate is set to 0.01, β_{1} and β_{2} are set to 0.9 and 0.999, respectively, and the number of iterations is set to 500.

A. Evaluation methods

Average accuracy (AA), overall accuracy (OA), and the Kappa coefficient are the evaluation indicators used in this paper. The number of pixels correctly predicted by the model divided by the total number of pixels on the test set equals OA; AA is the average ratio of the number of pixels correctly predicted for each type to the total number of pixels in each type; and the coefficient is the difference between the model's classification result and the completely random classification result. The calculated result is between -1 and 1, typically between 0 and 1. The higher the Kappa value, the higher the accuracy of the classification. The equation of calculation is as follows:

$$OA=\frac{N_{correct}}{N_{all}} \tag{26}$$

$$AA=\frac{1}{C}\sum_{m=1}^{C}\frac{N_{correct}^{m}}{N_{all}^{m}} \tag{27}$$

$$Kappa=\frac{OA-P_{e}}{1\text{-}P_{e}},P_{e}=\frac{\sum_{m=1}^{C}N_{correct}^{m}\times N_{all}^{m}}{N_{all}\times N_{all}} \tag{28}$$

where $N_{correct}$ represents the correctly classified test sample, N_{all} represents all the test samples, $N_{correct}^{m}$ represents the correctly classified test sample of the m -th type, N_{all}^{m} represents all the test samples of the m -th type, and C represents the total category of the HSIs data.

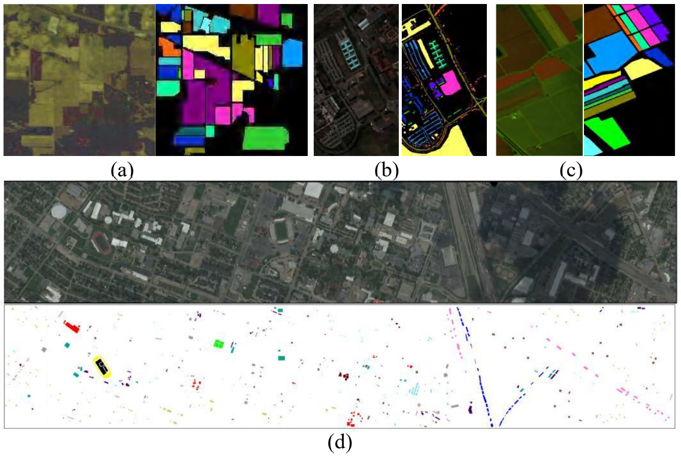


Fig. 8. False color image and GT of the four HSIs datasets. (a)Indian Pines. (b)Pavia University. (c)Salinas. (d)Houston.

TABLE I
NUMBER OF TRAINING SAMPLES FOR THE INDIAN PINES DATASET

ID	Class	Samples	5%	10%	15%
C1	Alfalfa	46	2	4	6
C2	Corn-notill	1428	71	142	214
C3	Corn-mintill	830	41	83	124
C4	Corn	237	11	23	35
C5	Grass-pasture	483	24	48	72
C6	Grass-trees	730	36	73	109
C7	Grass-pasture-mowed	28	1	2	4
C8	Hay-windrowed	478	23	47	71
C9	Oats	20	1	2	3
C10	Soybean-notill	972	48	97	145
C11	Soybean-mintill	2455	122	245	368
C12	Soybean-clean	593	29	59	88
C13	Wheat	205	10	20	30
C14	Woods	1265	63	126	189
C15	Buildings-Grass-Trees-Drives	386	19	38	57
C16	Stone-Steel-Towers	93	4	9	13

TABLE II
NUMBER OF TRAINING SAMPLES FOR THE PAVIA UNIVERSITY DATASET

ID	Class	Samples	0.1%	1%	5%
C1	Asphalt	6631	7	66	331
C2	Meadows	18649	17	186	932
C3	Gravel	2099	2	20	104
C4	Trees	3064	3	30	153
C5	Painted metal sheets	1345	1	13	67
C6	Bare Soil	5029	5	50	251
C7	Bitumen	1330	1	13	66
C8	Self-Blocking Bricks	3682	4	36	184
C9	Shadows	947	1	9	47

TABLE III
NUMBER OF TRAINING SAMPLES FOR THE SALINAS DATASET

ID	Class	Samples	0.1%	1%	5%
C1	Brocoli_green_w eeds_1	2009	2	20	100
C2	Brocoli_green_weeds_2	3726	4	37	186
C3	Fallow	1976	2	19	98
C4	Fallow_rough_plow	1394	1	13	69
C5	Fallow_smooth	2678	3	26	133
C6	Stubble	3959	4	39	197
C7	Celery	3579	4	35	178
C8	Grapes_untrained	11271	11	112	563
C9	Soil_vinyard_develop	6203	6	62	310
C10	Corn_senesced_green_weeds	3278	3	32	163
C11	Lettuce_romaine_4wk	1068	1	10	53
C12	Lettuce_romaine_5wk	1927	2	19	96
C13	Lettuce_romaine_6wk	916	1	9	45
C14	Lettuce_romaine_7wk	1070	1	10	53
C15	Vinyard_untrained	7268	7	72	363
C16	Vinyard_vertical_trellis	1807	2	18	90

TABLE IV
NUMBER OF TRAINING SAMPLES FOR THE HOUSTON DATASET

ID	Class	Samples	0.1%	0.5%	1%
C1	Healthy grass	1251	13	63	125
C2	Stressed grass	1254	13	63	125
C3	Synthetic grass	697	7	35	70
C4	Trees	1244	12	62	124
C5	Soil	1242	12	62	124
C6	Water	325	3	16	33
C7	Residential	1268	13	63	127
C8	Commercial	1244	12	62	124
C9	Road	1252	13	63	125
C10	Highway	1227	12	61	123
C11	Railway	1235	12	62	124
C12	Parking Lot1	1233	12	62	123
C13	Parking Lot2	469	5	23	47
C14	Tennis court	428	4	21	43
C15	Running track	660	7	33	66

B. Datasets

The Indian Pines (IP) dataset was created using the AVIRIS imaging spectrometer in northwestern Indiana, United States. It has a spatial resolution of 20m per pixel, 220 spectral bands, and a wavelength range of 0.4 to 2.45 μ m. The IP dataset has an input size of $145 \times 145 \times 200$ pixels, with a total of 21,025 pixels, 10,776 of which are background pixels. We chose to leave out the background pixels in this study, so only 10,249 pixels were used in the experiments, which included 16 different types of ground objects. In addition, after using PCA, the dimension is 100.

The Pavia University (PU) dataset is an image taken by the University of Pavia's ROSIS imaging spectrometer. It has a spatial dimension of 610×340 pixels. A total of 103 spectral dimension bands were used. The PU dataset's final input size is $610 \times 340 \times 103$ pixels, for a total of 2,207,400 pixels. We excluded the background pixels, as we did with the IP dataset, and thus only 42,776 pixels were used for testing, consisting of nine different types of ground objects. Furthermore, after using PCA, the dimension is 50.

The Salinas (SA) dataset was also acquired by the AVIRIS imaging spectrometer, and its images show the landforms of the Salinas Valley. The spatial resolution of the SA dataset is very high. The SA dataset had an input size of $512 \times 217 \times 204$ pixels. Only 54,129 pixels were used after background pixels were removed, resulting in 16 different types of ground objects. Figure 8 shows the grayscale image and ground truth for the above datasets (c). Furthermore, the dimension after PCA is 29.

The fourth data set was gathered over the University of Houston campus in 2013 and entered into the GRSS Data Fusion Contest. The image size in pixels is 349×1905, with a high spatial resolution of 2.5 m, and the wavelength bands range from 0.38 to 1.05 m. The University of Houston image and the associated ground-truth map are shown in false color in Fig.8(d). The ground-truth map has a total of 15 classes.

TABLE V
CLASSIFICATION PERFORMANCE OF THE DIFFERENT METHODS ON THE IP DATASET USING DIFFERENT TRAINING SAMPLES.

Methods	5%			10%			15%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
3D-CNN	0.5961	0.4037	0.5415	0.7377	0.6235	0.6983	0.7969	0.6897	0.7673
AlexNet	0.6932	0.7051	0.6474	0.7423	0.6571	0.7049	0.8167	0.7955	0.7907
ResNet	0.7274	0.6681	0.6864	0.7804	0.7911	0.7501	0.8300	0.8067	0.8059
3D-DenseNet	0.6770	0.6183	0.6295	0.7684	0.8097	0.7351	0.8142	0.8147	0.7884
DenseNet	0.7316	0.7046	0.7046	0.7892	0.7551	0.7579	0.8428	0.8149	0.8203
SSUN	0.7379	0.7359	0.7006	0.7782	0.7343	0.7459	0.8563	0.8663	0.8439
SAGP	0.7349	0.7658	0.7161	0.7836	0.8089	0.7672	0.8159	0.8650	0.8289
MCNN-CP	0.7769	0.7778	0.7455	0.8356	0.8158	0.8126	0.8596	0.8495	0.8404
CAG	0.7701	0.7827	0.7372	0.8518	0.8496	0.8300	0.8888	0.8913	0.8717
BTA-Net	**0.8162**	**0.8322**	**0.7894**	**0.8713**	**0.8783**	**0.8517**	**0.9184**	**0.9143**	**0.9068**

TABLE VI
CLASSIFICATION PERFORMANCE OF THE DIFFERENT METHODS ON THE PU DATASET USING DIFFERENT TRAINING SAMPLES.

Methods	0.1%			1%			5%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
3D-CNN	0.6589	0.4158	0.6436	0.7877	0.4875	0.7143	0.8089	0.7865	0.7323
AlexNet	0.8269	0.8157	0.8065	0.8791	0.8368	0.8383	0.9290	0.9182	0.9056
ResNet	0.8169	0.8158	0.8011	0.8581	0.8399	0.8105	0.9090	0.9142	0.8785
3D-DenseNet	0.8166	0.8255	0.8179	0.8588	0.8388	0.8081	0.9264	0.9184	0.9021
DenseNet	0.8039	0.8197	0.7989	0.8422	0.8233	0.7887	0.9036	0.8951	0.8719
SSUN	0.8179	0.8217	0.8211	0.8546	0.8768	0.8428	0.9359	0.9247	0.9151
SAGP	0.8288	0.8378	0.8079	0.8597	0.8428	0.8428	0.9109	0.8975	0.8814
MCNN-CP	0.8376	0.8277	0.8286	0.8599	0.8205	0.8116	0.9395	0.9251	0.9197
CAG	0.8469	0.8479	0.8362	0.9030	0.8881	0.8701	0.9528	0.9405	0.9340
BTA-Net	**0.8589**	**0.8557**	**0.8480**	**0.9132**	**0.9074**	**0.8836**	**0.9585**	**0.9495**	**0.9377**

TABLE VII
CLASSIFICATION PERFORMANCE OF THE DIFFERENT METHODS ON THE SA DATASET USING DIFFERENT TRAINING SAMPLES.

Methods	0.1%			1%			5%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
3D-CNN	0.5598	0.5478	0.5631	0.6505	0.4961	0.6138	0.7607	0.6185	0.7328
AlexNet	0.8927	0.8829	0.8711	0.9029	0.9267	0.8918	0.9419	0.9376	0.9353
ResNet	0.8821	0.8945	0.8744	0.8924	0.9220	0.8803	0.9180	0.9529	0.9087
3D-DenseNet	0.8839	0.8867	0.8912	0.8978	0.9356	0.8861	0.8969	0.9315	0.8854
DenseNet	0.8736	0.8817	0.8697	0.8854	0.9131	0.8723	0.9114	0.9447	0.9012
SSUN	0.8765	0.8746	0.8637	0.8959	0.9334	0.8841	0.9033	0.8823	0.8922
SAGP	0.8922	0.9067	0.8799	0.8972	0.9459	0.8856	0.9264	0.9569	0.9181
MCNN-CP	0.8769	0.9017	0.8742	0.8862	0.9177	0.8731	0.9211	0.9462	0.9121
CAG	0.9037	0.9166	0.8816	0.9114	0.9414	0.9015	0.9306	0.9584	0.9357
BTA-Net	**0.9015**	**0.9207**	**0.8891**	**0.9177**	**0.9510**	**0.9078**	**0.9434**	**0.9696**	**0.9369**

TABLE VIII
CLASSIFICATION PERFORMANCE OF THE DIFFERENT METHODS ON THE HOUSTON DATASET USING DIFFERENT TRAINING SAMPLES.

Methods	0.1%			0.5%			1%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
3D-CNN	0.7901	0.8004	0.7801	0.8158	0.8451	0.8217	0.8341	0.8215	0.8109
AlexNet	0.7966	0.7999	0.8023	0.8966	0.8907	0.8946	0.9129	0.8972	0.9001
ResNet	0.8015	0.7904	0.7895	0.8837	0.8879	0.8756	0.9101	0.8976	0.9124
3D-DenseNet	0.8066	0.8109	0.8042	0.9281	0.9408	0.9221	0.9356	0.9152	0.9143
DenseNet	0.7904	0.8105	0.8079	0.9476	0.9555	0.9431	0.9481	0.9249	0.9325
SSUN	0.8125	0.8056	0.8231	0.9435	0.9520	0.9389	0.9845	0.9755	0.9829
SAGP	0.8315	0.8478	0.8265	0.9488	0.9565	0.9387	0.9719	0.9655	0.9698
MCNN-CP	0.8269	0.8397	0.8178	0.9582	0.9627	0.9401	0.9817	0.9623	0.9779
CAG	0.8362	0.8501	0.8269	0.9678	0.9714	0.9625	0.9859	0.9906	0.9912
BTA-Net	**0.8444**	**0.8547**	**0.8318**	**0.9770**	**0.9779**	**0.9751**	**0.9942**	**0.9942**	**0.9938**

C. Comparison Results of Different Methods

This section will compare the quantitative classification results and the classification effect in the case of small samples with six other well-known methods, namely 3D-CNN [34], AlexNet [32], ResNet [33], DenseNet [43], 3D-DenseNet [35], SSUN [12], SAGP [11] ,MCNN-CP [36] and CAG [27]. For a fair comparison, the input and parameter selection of these nine methods were the same as the proposed BTA-Net, and the number of iterations was based on the accuracy of the training sample data converging to 1; for the rest of the settings, refer to the previous section. For four HSIs datasets, the classification performances of different algorithms under the training sample sizes of 0.1%, 1%, 5%, 10%, and 15% were evaluated.

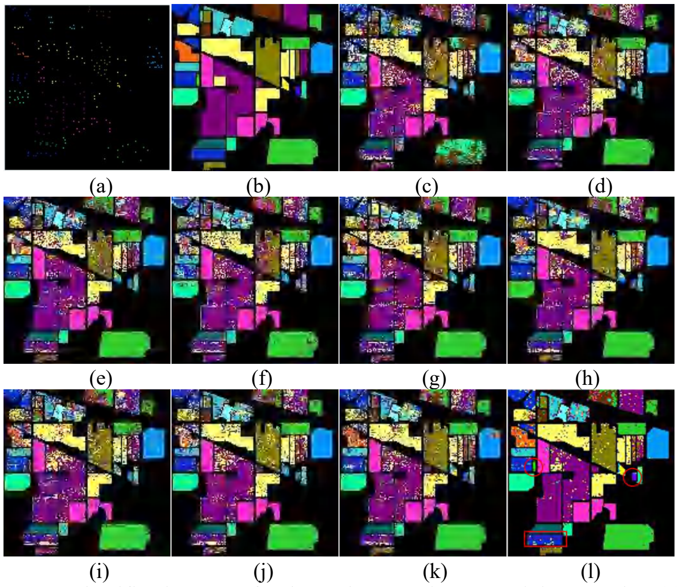


Fig. 9. Classification maps for the IP dataset using 5% training samples. (a) Training set (5%). (b) Ground truth. (c) 3D-CNN. (d) AlexNet. (e) ResNet. (f) 3D-DenseNet. (g) DenseNet. (h) SSUN. (i) SAGP. (j) MCNN-CP. (k) CAG. (l) BTA-Net.

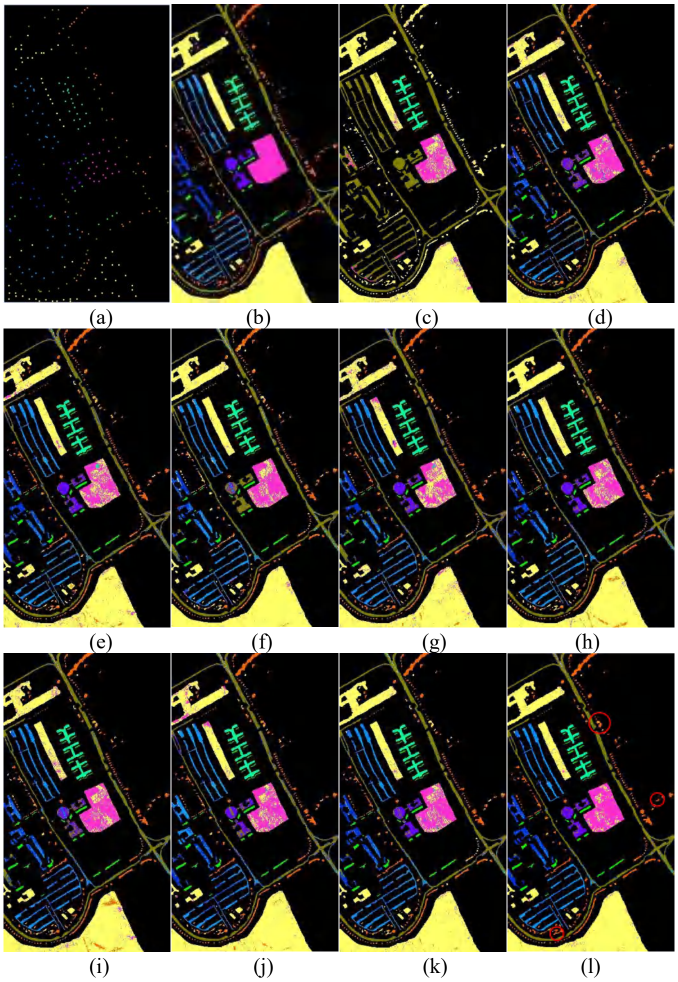


Fig. 10. Classification maps for the PU dataset using 1% training samples. (a) Training set (1%). (b) Ground truth. (c) 3D-CNN. (d) AlexNet. (e) ResNet. (f) 3D-DenseNet. (g) DenseNet. (h) SSUN. (i) SAGP. (j) MCNN-CP. (k) CAG. (l) BTA-Net.

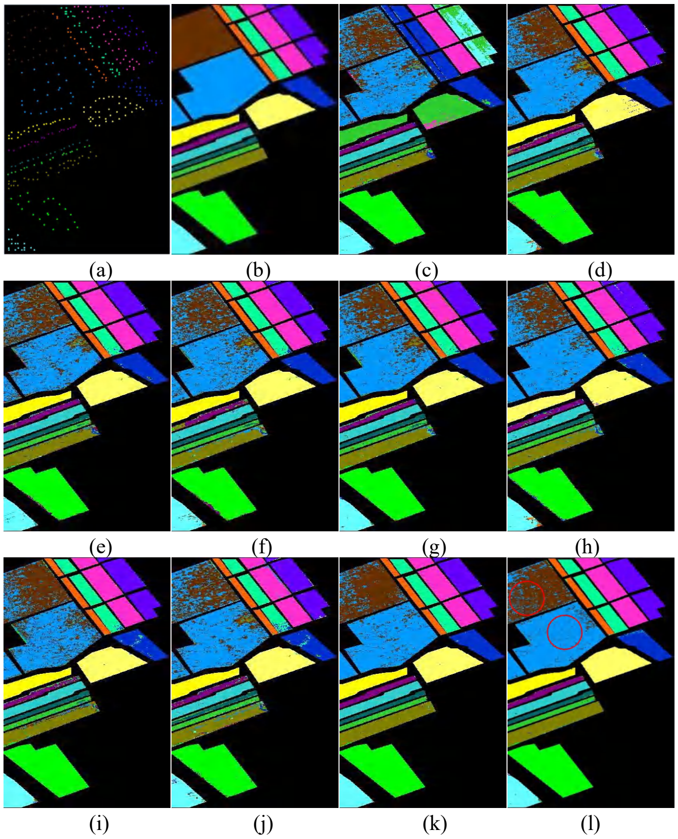


Fig. 11. Classification maps for the SA dataset using 1% training samples. (a) Training set (1%). (b) Ground truth. (c) 3D-CNN. (d) AlexNet. (e) ResNet. (f) 3D-DenseNet. (g) DenseNet. (h) SSUN. (i) SAGP. (j) MCNN-CP. (k) CAG. (l) BTA-Net.

1) *Comparison results on the IP dataset:* The IP dataset sample size is small and the distribution of 16 ground objects is extremely uneven. Compared with the PU and SA datasets using the lowest sample of 1%, this paper chooses 5% of the training samples on the IP dataset. This is because the Grass-pasture-mowed and Oats categories of the IP dataset have only 28 and 20 samples, respectively. Even if 5% of the samples were selected, only one sample was available for training in this experiment (as shown in Table I). Therefore, 5%, 10%, and 15% were used as the training sets for the IP dataset. It can be seen from Fig.9 that AlexNet has the worst performance for classification and has a lot of noise. This is because it is a method of shallow model classification and has poor generalization ability, which is not enough to deal with hyperspectral images' complex spectral spatial distribution.

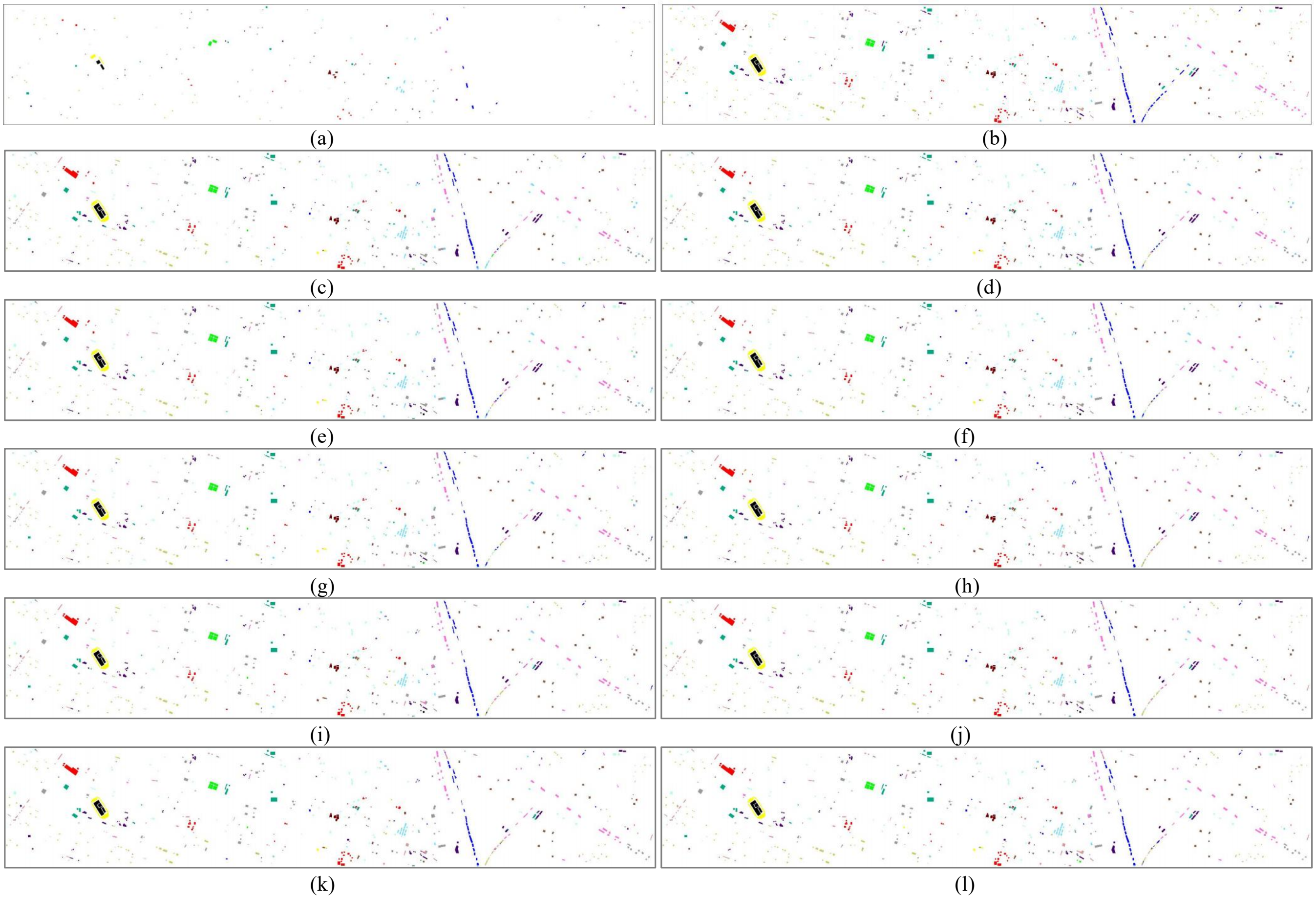
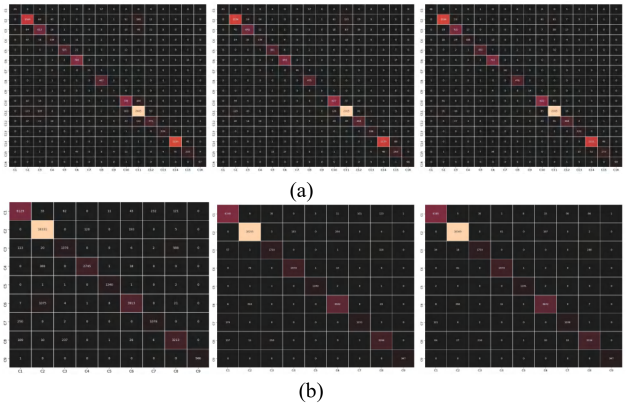
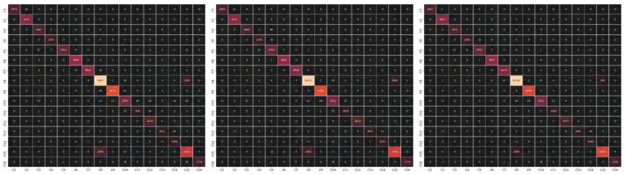

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 12. Classification maps for the Houston dataset using 0.1% training samples. (a) Training set (0.1%). (b) Ground truth. (c) 3D-CNN. (d) AlexNet. (e) ResNet. (f) 3D-DenseNet. (g) DenseNet. (h) SSUN. (i) SAGP. (j) MCNN-CP. (k) CAG. (l) BTA-Net.

CAG has a better visual experience among the remaining five methods than SAGP, which uses increased network model depth to extract more discriminatory features, and the proposed BTA-Net is superior in detail to CAG, and it can classify edge pixels more precisely and get closer to ground truth. Table V shows the quantitative analysis results of various classification models for the IP data set. It can also be seen that the classification accuracy obtained by using the attention mechanism method is significantly better than the accuracy obtained by using the baseline and SSUN methods. This is due to the attention mechanism in the training samples paying more attention to features containing more semantic information. Therefore, the highest classification precision is obtained by BTA-Net using the three-direction attention mechanism. The cross-attention mechanism used in CAG achieved better results compared with the traditional attention mechanism used in SAGP. The performance achieved by SAGP is 0.07–12.5% higher than that of the four methods that do not use the attention mechanism, which demonstrates the effectiveness of attention mechanism in the HSIs classification task. It can be seen from Table V that the residual structure significantly reduces the model parameters; however, the model ’ s performance is improved. This demonstrates that it is possible to improve the classification performance of HSIs by ablating redundant information; it also shows the rationality of BC. In addition, it is also found that the performance of 3DCNN-based methods does not perform well on small training samples, which indicates that 3DCNN needs more training samples. we introduced a new spatial attention to achieve a more robust performance, considering the spatial relationship characteristics of the feature map. OA is 0.07% ~ 12.5% higher than the other six methods, AA is 0.07% ~ 12.5% higher than the other six methods, and Kappa is 0.07% to 12.5% higher than the other six methods. In addition, Fig.9 shows the visualization classification maps of BTA-Net and the other six algorithms on the IP dataset. Fig.13 also shows a visualization of the confusion matrix.


(a)

(b)



(c)

Fig. 13. Visualizations of the confusion matrix of the classification results for the IP, PU, and SA datasets. (a) Training on 5%, 10%, and 15% samples. (b) Training on 1%, 5%, and 10% samples. (c) Training on 1%, 5%, and 10% samples.

2)*Comparison results on the PU dataset:* The dataset of the PU is characterized by scattered and sparse sample locations and uneven distribution. The classification outputs of the various algorithms used on this dataset are shown in Fig.10. The proposed BTA-Net classification result has the least number of misclassified ground objects, the overall is smoother, and there are only a few noise points that are closer to the GT when compared to other algorithms. Table VI also shows that BCTDC has achieved outstanding advantages in all three quantitative indicators. It can be seen that the three-direction attention mechanism, which gives different weights to different features from different directions, selects more effective features, effectively uses more informative features, improves the network's ability to extract features, and improves the model's generalization ability. Because the PU dataset has a larger sample size than the IP dataset, the accuracy of the classification is relatively high, regardless of the method used. Overall, OA, AA, and the Kappa coefficient obtained using BTA-Net were 1.4% – 4.1%, 1.4 – 4.1%, and 1.4 – 4.1% higher than those of the other methods. Fig. 13 also shows a visualization of the confusion matrix.

3)*Comparison results on the SA dataset:* The sample distribution of the SA dataset is more balanced than the IP and PU datasets, and the classification difficulty is lower. It can be seen from Fig. 11 and Table VII that all the methods achieved excellent classification performance on the SA dataset, but the Vinyard_untrained class was affected by spectral characteristics, the performances of the other five groups of methods were not satisfactory, and the three-direction attention can be given from different directions. Assigning different weights to different features can alleviate this limitation. Therefore, the proposed BTA-Net method obtains more robust results. Its OA, AA, and Kappa coefficient are higher than that of other models by 1.4 – 4.1%, 1.4 – 4.1%, and 1.4 – 4.1%, respectively. Fig. 13 shows the visualization results of the confusion matrix.

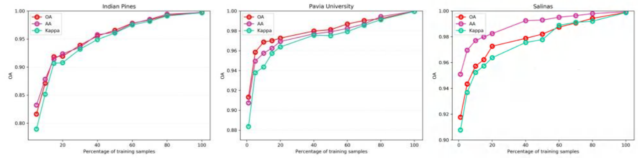
4)*Comparison results on the Houston dataset:* Compared to the previous four data sets, the Houston data set is a much larger and more complex data set with very uneven categories. As can be seen from Figure 12 and Table 8, all the methods achieved good classification performance on the Houston data set, but the railway category was affected by spectral features, and the performance of the other nine groups of methods was not ideal. Three-Directions attention could distribute the attention weight of features from three different directions, thus alleviating this limitation. Therefore, the proposed method achieves more robust results. The OA, AA, and Kappa coefficients are 0.9-6.4%, 0.5-7.5%, and 0.5-6.2% higher than other methods under 0.1% training samples, respectively.

TABLE IX
THE TOTAL PARAMETERS AND TEST TIMES OF THE PROPOSED BTA-NET AND OTHER SIX METHODS ON THE IP, PU AND SA DATASETS

Methods	IP		PU		SA	
	Total params	Times	Total params	Times	Total params	Times
AlexNet	1,360,144	0.69s	3,869,269	5.37s	3,857,160	7.59s
ResNet	920,336	1.34s	903,573	4.98s	932,552	4.35s
DenseNet	2,393,432	1.63s	1,899,613	5.95s	3,022,952	6.52s
SSUN	1,196,795	15.55s	1,210,631	24.11s	1,345,179	39.25s
SAGP	3,828,368	3.72s	3,292,565	10.77s	2,976,072	13.29s
CAG	1,963,296	1.19s	1,945,637	5.43s	1,487,800	5.43s
BTA-Net	**880,325**	**1.03s**	**867,525**	**4.21s**	**862,149**	**4.55s**

D. *Computational Complexity*

For evaluating the proposed classification model, the total parameters and calculation time are important indicators. Table IX shows the total parameters and calculation times for the BTA-Net and six other comparison algorithms. As shown in the table, the proposed model takes slightly longer to calculate on the IP dataset than AlexNet with only eight layers, but it is faster than the other five methods. In addition, the calculation time is reduced by 1.3% compared to a previous work [27]. The proposed BTA-Net also achieved the best computational performance on the PU and SA datasets, reducing the computational cost by 1–3% compared to the other six methods. Furthermore, we found that ResNet has good performance in terms of the total parameters of the model, which is much lower than the other five methods. However, on the four HSIs datasets, the performance of BTA-Net is still 1 – 2% lesser than that of ResNet. The introduction of the BC eliminates redundant parameters and increases the number of effective features. Therefore, to achieve the same performance, BTA-Net does not require too many neurons in the convolutional layer; this enables it to reduce the parameters in the model and effectively decrease computation cost. It is worth noting that it is very popular in practical applications to reduce model parameters and calculation time without loss of classification performance. In addition, the proposed BC can not only be used for the experiments in the present study but can also be improved for other baselines. Further, we replaced BC with one of the convolutional layers in AlexNet to further demonstrate the superiority of the Bole convolution.



(a) (b) (c)

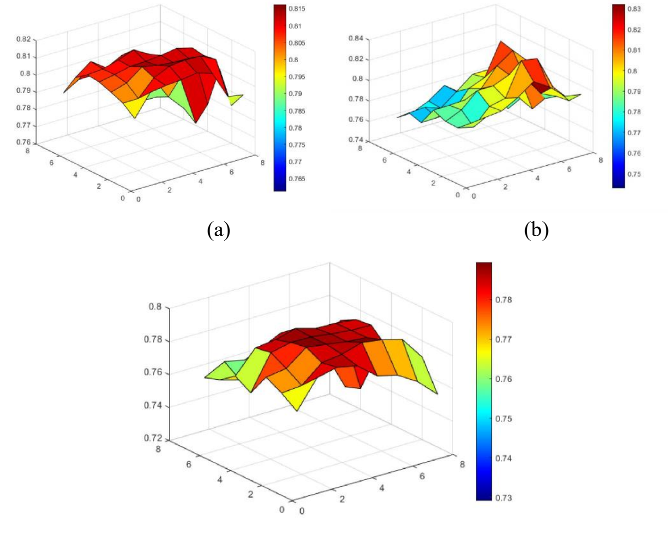
Fig. 14. The change curve of OA, AA and Kappa under different training samples. (a) IP. (b) PU. (c) SA.

E. *Effects of Different Number of Training Samples*

In Section IV.C, it was demonstrated that BTA-Net is superior to several well-known algorithms in the HSIs classification task. To further comprehensively evaluate the algorithm, we divide the datasets into multiple samples for experimentation. We divide the IP dataset into {5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 99%} training samples, and the PU and SA datasets into {1%, 5%, 10%, 15%, 20%, 40%, 50%, 60%, 70%, 80%, 99%} training samples. Using different numbers of training samples from four data sets, the algorithm's classification accuracy is shown in Fig.14. For starters, BTA-Net performed admirably on a smaller set of training samples. Second, the learning model's classification performance improved significantly as the number of training samples increased. Finally, from 1% to 10% of the training samples, the model's classification performance improved qualitatively.

F. Ablation Experiment of Bole convolution

For the threshold, ablation experiments were set up, and the coefficient of BC was rewarded in order to observe the effects on the experimental results. On the IP dataset, we conducted ablation experiments with 5% training samples.



(a) (b)

(c)

Fig. 15. Performance of the Bole convolution with different T and r values on IP dataset. (a) OA. (b) AA. (c) Kappa.

TABLE X
THE AA PERFORMANCE OF THE BOLE CONVOLUTION WITH DIFFERENT T AND R VALUES ON IP DATASET.

Reward→ Threshold↓	1.3	1.5	1.8	2.0	2.5	3.0
0.1	0.7953	0.7971	0.8056	0.8171	0.8075	0.7938
0.2	0.7798	0.7959	0.7994	0.8322	0.8298	0.7973
0.3	0.7911	0.8056	0.8059	0.8124	0.8051	0.7851
0.4	0.7899	0.7908	0.7974	0.8155	0.8262	0.8107
0.5	0.7719	0.7733	0.7918	0.7983	0.8297	0.7974
0.6	0.7765	0.7712	0.7831	0.7957	0.7959	0.7836
0.7	0.7657	0.7725	0.7835	0.7719	0.7729	0.7658

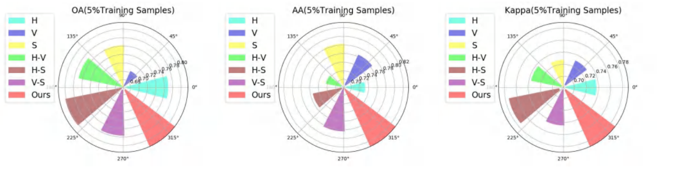
It can be seen from Table X that when the threshold T is set to 0.2 and the reward coefficient r is set to 2.0, the accuracy of each evaluation index is the highest. Redundant features are effectively eliminated, and the classification accuracy of the model is improved. Therefore, this proves that the feature penalty and feature reward strategy of Bole convolution is effective. In addition, it can be clearly seen from Fig. 15 that on the three evaluation indicators of OA, AA, and Kappa coefficient, as the threshold increases, the classification performance of the model gradually increases, and then gradually decreases. Similarly, as the reward coefficient increases, the model classification performance gradually increases, and then gradually decreases. This proves that when the threshold is set to 0.2 and the reward coefficient is set to 2.0, the model achieves the best performance. When T was 0.1, the classification accuracy was low, indicating that the model rewards some redundant information, which affects the classification accuracy.

G. Ablation Experiment of Three-Directions Attention Mechanism

This section conducts ablation experiments for the three-directional attention mechanism to see how they affect the experimental results. Similarly, We conducted ablation experiments using 5% training on the IP dataset. "H" means horizontal attention, "V" means vertical attention, and "S" means spatial attention. In particular, "H-V" represents the cross-attention mechanism of our previous work.

TABLE XI
RESULTS OF THE ABLATION STUDY OF THE THREE-DIRECTIONS ATTENTION MECHANISM USING 5% TRAINING SAMPLES

Methods	IP(5%)		
	OA	AA	Kappa
H	0.7692	0.7318	0.7366
V	0.7042	0.7647	0.7304
S	0.7619	0.7827	0.7272
H-V	0.7697	0.7235	0.7364
H-S	0.8021	0.7543	0.7743
V-S	0.7766	0.7837	0.7449
BTA-Net	**0.8162**	**0.8322**	**0.7894**



(a) (b) (c)

Fig. 16. Limit pie charts using 5% of the training samples on the IP dataset. (a) OA. (b) AA. (c) Kappa.

The performance of spatial attention is the best, as shown in Table XI and Fig.16, whether single or combined attention is used. The HSIs classification model's performance is significantly improved by obtaining spatial features between feature maps. Furthermore, it has been discovered that combined attention is superior to a single attention mechanism, demonstrating that assigning attention weights from different directions is effective.

TABLE XII
ABLATION EXPERIMENT RESULTS OF WEIGHT ASSIGNMENT STRATEGY

Methods	IP(5%)		
	OA	AA	Kappa
Mul+Max	0.7798	0.7564	0.7476
Mul+Add	0.7868	0.7639	0.7560
(Max+Add) **BTA-Net**	**0.8162**	**0.8322**	**0.7894**

Table XII shows that the combination of weight maximization and addition strategy produces the best results, while the combination of weight multiplication and addition strategy outperforms the multiplication and maximization strategy, demonstrating that the weight addition strategy is superior to the weight maximization strategy, and the maximization strategy is better than the multiplication strategy. This also indirectly proves that the feature reward strategy of Bole convolution is effective. We know that because the weight is between (0,1), the multiplication strategy will cause the weight to become smaller, thereby weakening the feature. Lead to a decline in classification performance. Therefore, on the basis of eliminating redundant features, enhancing effective features can improve model classification performance.

TABLE XIII
RESULTS OF THE ABLATION STUDY OF THE BC AND TDA ON THE BASELINE MODELS

Methods	IP(5%)		
	OA	AA	Kappa
AlexNet	0.6932	0.7051	0.6474
AlexNet-BC	0.7411	0.7099	0.7046
AlexNet-TDA	0.7520	0.7665	0.7169
AlexNet-BC-TDA	**0.7750**	**0.7695**	**0.7428**

H. *Effects of BC and TDA on the Baseline Models*

We believe that in the HSIs classification task, the Bole convolution and three-directions attention mechanism proposed in this study are not only effective in this algorithm, but can also be applied to improve other models. Therefore, in this section, we choose the well-known AlexNet to perform ablation experiments, and used BC, TDA and BC+TDA to improve the performance of AlexNet. The experiment was carried out using 5% of the training samples from the IP dataset. Table XIII shows that the BC, TDA, and BC+TDA proposed in this paper significantly improved AlexNet's classification performance. BC application improved OA by 6.9%, TDA application improved OA by 8.5 percent, and BC+TDA application improved OA by 11.8 percent. This adds to the evidence that BC and TDA are effective.

TABLE XIV
ABLATION EXPERIMENT RESULTS OF DIMENSIONALITY REDUCTION

Methods	IP(5%)		PU(1%)		SA(1%)	
	OA	Kappa	OA	Kappa	OA	Kappa
NO_PCA	0.7914	0.7632	0.8956	0.8737	0.9063	0.8862
LDA	0.8105	0.8068	0.9099	**0.8901**	0.9125	0.9063
BTA-Net	**0.8162**	**0.7894**	**0.9132**	0.8836	**0.9177**	**0.9078**

I. *Ablation Experiment of PCA Dimensionality Reduction*

We verified the impact of PCA dimension reduction on classification performance. The experiment used 5% of the training samples from the IP, PU, and SA datasets. As shown in Table XIV, dimensionality reduction using PCA improved classification performance on the four datasets: OA increased by 3.1%, 1.96%, and 1.26%, and Kappa increased by 3.39%, 1.13%, and 2.43%, and PCA is slightly better than LDA.

TABLE XV
ABLATION EXPERIMENT RESULTS OF THE PROPOSED SUBMODULE

Methods	IP(5%)		
	OA	AA	Kappa
BTA-Net-NO-BC	0.7793	0.7788	0.7468
BTA-Net-NO-TDA	0.7659	0.7332	0.7321
BTA-Net	**0.8162**	**0.8322**	**0.7894**

J. *Ablation Experiment of Different Sub-modules*

Ablation studies were conducted on different submodules of BTA-Net. The results reported in Table XV show that the classification performance decreased by 4.73%, 6.85%, and 5.7% in the three indicators when the BC module was removed. However, excluding the TDA module, the classification performance decreased by 6.56%, 13.5%, and 7.82%. This also proves that the TDA module is superior to the BC module. By allocating different attention weights in different directions, more representative features can be obtained, which can significantly improve the performance of the HSIs classification model.

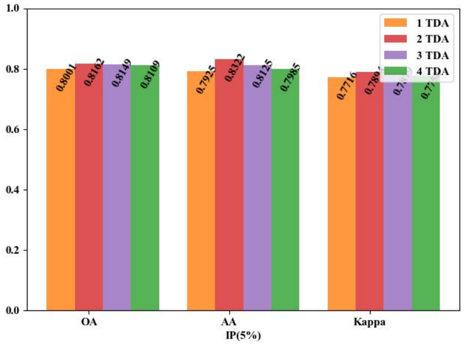


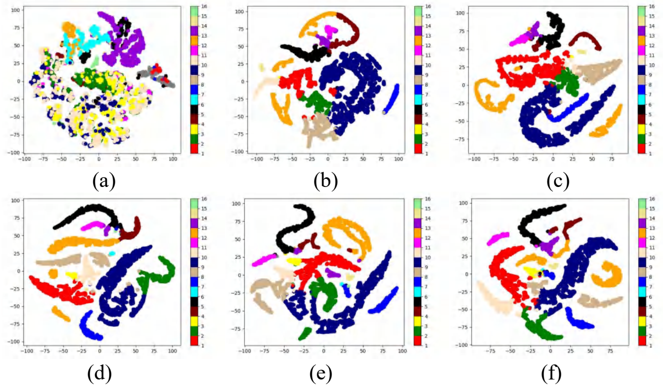
Fig. 17. Visualization results of the ablation study of the different tandems TDA using 5% training samples on IP dataset.

K. *Ablation Experiment of Different Tandems TDA*

In this section, We present an ablation study on various tandem TDAs, which was carried out on the IP dataset with 5% training samples. Figure 16 shows that when two TDA modules are used, the best results are obtained, whereas when one or four TDA modules are used, the worst results are obtained. We know from Section IV.J that the TDA module performs well in feature selection, so more deep feature exploration via two tandem TDAs should help improve the model's classification performance.

L. *t-SNE Data Distributions Visualization*

In this section, we use t-SNE to verify whether the features extracted by BTA-Net are beneficial and discriminative for feature clustering for network training. Figure 18 shows the 2D visualization of BTA-Net and nine comparison models on the IP dataset, which clearly shows that our BTA-Net can distinguish different types of data well. Specifically, the clustering accuracy of 3D-CNN, AlexNet and ResNet models is low, mainly because these models cannot mine discriminative features on small samples and imbalanced data. Furthermore, we can see that the features learned by BTA-Net are both intra-class compact and inter-class separated, mainly thanks to our Bole convolution and three-directions attention mechanism modules.



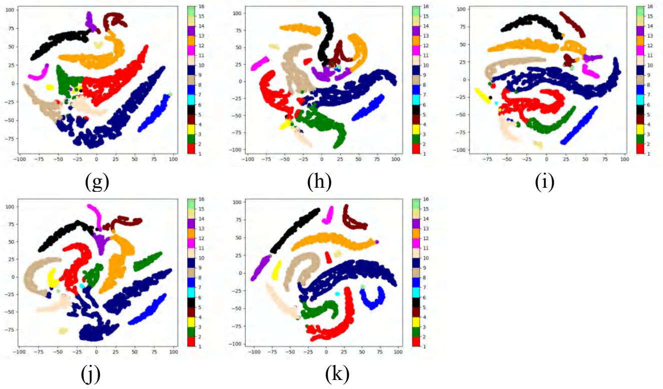


Fig. 18. The 2D t-SNE feature visualization on the Indian Pines dataset Using 5% training samples, showing the data distribution of labeled samples in the feature space, with different colors corresponding to different categories. (a) Original. (b) 3D-CNN. (c) AlexNet. (d) ResNet. (e) 3D-DenseNet. (f) DenseNet. (g) SSUN. (h) SAGP. (i) MCNN-CP. (j) CAG. (k) BTA-Net.

V. CONCLUSION

In this paper, we propose a novel BTA-Net algorithm for hyperspectral image classification, which includes a brand-new Bole convolution and a novel three-directions attention mechanism. In BC, we put forward the features punishment and reward strategy, can efficiently eliminate redundant information and enhance the effective features, reduce the model parameters and to reduce the computational cost. In TDA, we proposed assigning attention weights in three directions (horizontal, vertical and spatial directions), which can capture the best representative features, and propose weight addition and maximization strategies. Through ablation studies, it was demonstrated that spatial attention can most effectively improve the classification performance. Moreover, the addition strategy is better than the maximization strategy than the multiplication strategy. In addition, the BC and TDA proposed in this paper can also be used to improve other classification models, which is an obvious contribution to the research community. A series of comparative experiments and ablation studies have demonstrated the effectiveness and superiority of BTA-Net.

In future works, we will study the adaptive Bole convolutional network and apply it to other classification models and tasks.

REFERENCES

[1] Sahadevan, A. S. (2021). Extraction of spatial-spectral homogeneous patches and fractional abundances for field-scale agriculture monitoring using airborne hyperspectral images. *Computers and Electronics in Agriculture*, 188, 106325.

[2] Lu, B., Dao, P. D., Liu, J., He, Y., & Shang, J. (2020). Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing*, 12(16), 2659.

[3] Zhao, J., Kechasov, D., Rewald, B., Bodner, G., Verheul, M., Clarke, N., & Clarke, J. L. (2020). Deep Learning in Hyperspectral Image Reconstruction from Single RGB images—A Case Study on Tomato Quality Parameters. *Remote Sensing*, 12(19), 3258.

[4] Xue, Q., Qi, M., Li, Z., Yang, B., Li, W., Wang, F., & Li, Q. (2021). Fluorescence hyperspectral imaging system for analysis and visualization of oil sample composition and thickness. *Applied Optics*, 60(27), 8349-8359.

[5] B. Luo, C. Yang, J. Chanussot and L. Zhang, "Crop Yield Estimation Based on Unsupervised Linear Unmixing of Multidate Hyperspectral Imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 162-173, Jan. 2013, doi: 10.1109/TGRS.2012.2198826.

[6] Gerhards, M., Schlerf, M., Mallick, K., & Udelhoven, T. (2019). Challenges and future perspectives of multi-/Hyperspectral thermal infrared remote sensing for crop water-stress detection: A review. *Remote Sensing*, 11(10), 1240.

[7] Lv, M., Li, W., Chen, T., Zhou, J., & Tao, R. (2021). Discriminant tensor-based manifold embedding for medical hyperspectral imagery. *IEEE Journal of Biomedical and Health Informatics*, 25(9), 3517-3528.

[8] Aref, M. H., Youssef, A. B. M., Aboughaleb, I. H., & El-Sharkawy, Y. H. (2021). Characterization of Normal and Malignant Breast Tissues utilizing Hyperspectral Images and Associated Differential Spectrum Algorithm. *Journal of Biomedical Photonics & Engineering*, 7(2), 020302.

[9] Fakhrullin, R., Nigamatzyanova, L., & Fakhrullina, G. (2021). Dark-field/hyperspectral microscopy for detecting nanoscale particles in environmental nanotoxicology research. *Science of The Total Environment*, 772, 145478.

[10] Dou, P., & Zeng, C. (2020). Hyperspectral image classification using feature relations map learning. *Remote Sensing*, 12(18), 2956.

[11] You, H., Tian, S., Yu, L., & Lv, Y. (2019). Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Transactions on Geoscience and Remote Sensing, 58(2), 1281-1293.

[12] Xu, Y., Zhang, L., Du, B., & Zhang, F. (2018). Spectral–spatial unified networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(10), 5893-5909.

[13] Imani, M., & Ghassemian, H. (2020). An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. *Information fusion*, 59, 59-83.

[14] Yu, C., Han, R., Song, M., Liu, C., & Chang, C. I. (2020). A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial – spectral fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 2485-2501.

[15] Otgonbaatar, S., & Datcu, M. (2021). A quantum annealer for subset feature selection and the classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 7057-7065.

[16] Luo, F., Zhang, L., Du, B., & Zhang, L. (2020). Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8), 5336-5353.

[17] Li, Y., Tang, H., Xie, W., & Luo, W. (2021). Multidimensional Local Binary Pattern for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-13.

[18] Hu, J., Shen, X., Yu, H., Shang, X., Guo, Q., & Zhang, B. (2021). Extended Subspace Projection Upon Sample Augmentation Based on Global Spatial and Local Spectral Similarity for Hyperspectral Imagery Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 8653-8664.

[19] Xue, Z., Zhang, M., Liu, Y., & Du, P. (2021). Attention-based second-order pooling network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 59(11), 9600-9615.

[20] Cui, Y., Xia, J., Wang, Z., Gao, S., & Wang, L. (2021). Lightweight Spectral – Spatial Attention Network for Hyperspectral Image Classification. IEEE transactions on geoscience and remote sensing, 60, 1-14.

[21] Yu, C., Han, R., Song, M., Liu, C., & Chang, C. I. (2021). Feedback attention-based dense CNN for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-16.

[22] Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A., & Li, J. (2019). Visual attention-driven hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10), 8065-8080.

[23] Hang, R., Li, Z., Liu, Q., Ghamisi, P., & Bhattacharyya, S. S. (2020). Hyperspectral image classification with attention-aided CNNs. *IEEE Transactions on Geoscience and Remote Sensing*.

[24] Ye, M., Ji, C., Chen, H., Lei, L., Lu, H., & Qian, Y. (2019). Residual deep PCA-based feature extraction for hyperspectral image classification. *Neural Computing and Applications*, 1-14.

[25] Feng, F., Li, W., Du, Q., & Zhang, B. (2017). Dimensionality reduction of hyperspectral image with graph-based discriminant analysis considering spectral similarity. *Remote sensing*, 9(4), 323.

[26] Deng, Y. J., Li, H. C., Fu, K., Du, Q., & Emery, W. J. (2018). Tensor low-rank discriminant embedding for hyperspectral image dimensionality reduction. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12), 7183-7194.

[27] W. Cai and Z. Wei, "Remote Sensing Image Classification Based on a Cross-Attention Mechanism and Graph Convolution," in *IEEE Geoscience and Remote Sensing Letters*.

[28] Reshma, R., Sowmya, V., & Soman, K. P. (2018). Effect of Legendre–Fenchel denoising and SVD-based dimensionality reduction algorithm on hyperspectral image classification. *Neural Computing and Applications*, 29(8), 301-310.

[29] Wang, X., & Liu, F. (2017). Weighted low-rank representation-based dimension reduction for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 14(11), 1938-1942.

[30] Xu, H., Zhang, H., He, W., & Zhang, L. (2019). Superpixel-based spatial-spectral dimension reduction for hyperspectral imagery classification. *Neurocomputing*, 360, 138-150.

[31] Gao, S., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. H. (2019). Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*.

[32] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.

[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[34] Li, Y., Zhang, H., & Shen, Q. (2017). Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing*, 9(1), 67.

[35] Zhang, C., Li, G., Du, S., Tan, W., & Gao, F. (2019). Three-dimensional densely connected convolutional network for hyperspectral remote sensing image classification. Journal of Applied Remote Sensing, 13(1), 016519.

[36] Zheng, J., Feng, Y., Bai, C., & Zhang, J. (2020). Hyperspectral image classification using mixed convolutions and covariance pooling. IEEE Transactions on Geoscience and Remote Sensing, 59(1), 522-534.

[37] X. Li, B. Liu, K. Zhang and W. Liu, "Location Soft-Aggregation-Based Band Weighting for Hyperspectral Image Classification," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022, Art no. 5509005, doi: 10.1109/LGRS.2021.3109484.

[38] J. Pang, D. Zhang, H. Li, W. Liu and Z. Yu, "Hazy Re-ID: An Interference Suppression Model for Domain Adaptation Person Re-Identification Under Inclement Weather Condition," 2021 *IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428462.

[39] Liu, W., Li, J., Liu, B., Guan, W., Zhou, Y., & Xu, C. (2021). Unified cross-domain classification via geometric and statistical adaptations. *Pattern Recognition*, 110, 107658.

[40] Hang, R., Li, Z., Liu, Q., Ghamisi, P., & Bhattacharyya, S. S. (2020). Hyperspectral image classification with attention-aided CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3), 2281-2293.

[41] Zhu, M., Jiao, L., Liu, F., Yang, S., & Wang, J. (2020). Residual spectral–spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1), 449-462.

[42] Roy, S. K., Manna, S., Song, T., & Bruzzone, L. (2020). Attention-based adaptive spectral – spatial kernel ResNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9), 7831-7843.

[43] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Weiwei Cai (*Member, IEEE*) is with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China and also with the Graduate School, Northern Arizona University, Flagstaff, AZ 86011, USA. Prior to that, he worked with IT industry for more than ten years in the roles of an System Architect and the Program Manager. He is an *Associate Editor* of Journal Of Cloud Computing, an *Editorial Board Member* of Wireless Communications and Mobile Computing. He has also served as a *Guest Editor* of the CMES-Computer Modeling in Engineering & Sciences and Information. His research interests include machine learning, deep learning, and computer vision.

Xin Ning (*Senior Member, IEEE*) received the B.S. degree in software engineering in 2012, and the Ph.D. degree in electronic circuit and system from university of Chinese Academy of Sciences, in 2017. He is currently an Associate Professor with the Laboratory of Artificial Neural Networks and High Speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences. His current research interests include pattern recognition, computer vision, and image processing.

Guoxiong Zhou received his B. S. degree from the Hunan Agricultural University in 2002 and his M.S. degree from Central South University in China in 2006. He received his Ph. D. degree from Central South University in 2010 and is an associate professor at the Central South University of Forestry and Technology. His main research interests include machine learning and pattern recognition.

Xiao Bai received the B.Eng. degree in computer science from Beihang University of China, Beijing, China, in 2001, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2006.

He was a Research Officer (Fellow, Scientist) with the Computer Science Department, University of Bath, until 2008. He is currently a Full Professor with the School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, China. He has authored or co-authored more than 100 papers in journals and refereed conferences. His current research interests include pattern recognition, image processing, and remote sensing image analysis. He is an Associate Editor for *Pattern Recognition* and *Signal Processing*.

Yizhang Jiang (*Senior Member, IEEE*) received the B.S. degree from the Nanjing University of Science and Technology, China, in 2010 and the M.S. and Ph.D. degrees from the Jiangnan University, Wuxi, China, in 2012 and 2015, respectively. He is currently an Associate Professor with the School of Digital Media, Jiangnan University since Jan. 2016. He has been a research assistant in the Department of Computing, Hong Kong Polytechnic University, for almost two years. His research interests include pattern recognition, intelligent computation and their applications in medicine. He is an *Associate Editor* of IEEE Access (SCIE, 2019-), an *Associate Editor* of Frontiers in Medical Technology (2020-), an *Associate Editor* of Frontiers in Psychology (SSCI, 2021-), an *Editorial Board Member* of Technology and Health Care (SCIE, 2020-), an *Editorial Board Member* of Journal of Organizational and End User Computing (SSCI, SCIE, EI 2020-), and an *Editorial Review Board Member* of International Journal of Healthcare Information Systems and Informatics (ESCI & EI，2019-).

Wei Li (*Senior Member, IEEE*) received the B.Sc. degree in telecommunications engineering from Xid ian University, Xi ’an, China, in 2007, the M.Sc. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engi neering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent 1 year as a Postdoctoral Researcher at the University of California, Davis, CA, USA. He was a Professor with the College of Information Science and Technology at Beijing University of Chemical Technology from 2013 to 2019. He is currently a professor with the School of Information and Electronics, Beijing Institute of Technology. His research interests include hyperspectral image analysis, pattern recognition, and data compression.

He is currently serving as an Associate Editor for the IEEE Signal Processing Letters and the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS). He has served as Guest Editor for special issue of Journal of Real-Time Image Processing, Remote Sensing, and IEEE JSTARS. He received the 2015 Best Reviewer award from IEEE Geoscience and Remote Sensing Society (GRSS).

Pengjiang Qian (*Senior Member, IEEE*) received the Ph.D. degree from Jiangnan University, Wuxi, China, in March 2011. He is currently a Full Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University. He has authored or coauthored more than 80 papers published in international/national journals and conferences, such as IEEE Transactions on Medical Imaging, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, IEEE Transactions on Cybernetics, IEEE Transactions on Fuzzy Systems, Information Fusion, Pattern Recognition, Information Sciences, and Knowledge-Based Systems. His research interests include data mining, pattern recognition, bioinformatics, and their applications, such as analysis and processing for medical imaging, intelligent traffic dispatching, and advanced business intelligence in logistics.