Nonlocal Band Attention Network for Hyperspectral Image Band Selection

Tiancong Li, Yaoming Cai, Zhihua Cai, Xiaobo Liu, and Qiubo Hu

***Abstract*—Band selection (BS) is a foundational problem for the analysis of high-dimensional hyperspectral image (HSI) cubes. Recent developments in the visual attention mechanism allow for specifically modeling the complex relationship among different components. Inspired by this, this article proposes a novel band selection network, termed as nonlocal band attention network (NBAN), based on using a nonlocal band attention reconstruction network to adaptively calculate band weights. The framework consists of a band attention module, which aims to extract the long-range attention and reweight the original spectral bands, and a reconstruction network which is used to restore the reweighted data, resulting in a flexible architecture. The resulting BS network is able to capture the nonlinear and the long-range dependencies between spectral bands, making it more effective and robust to select the informative bands automatically. Finally, we compare the result of NBAN with six popular existing band selection methods on three hyperspectral datasets, the result showing that the long-range relationship is helpful for band selection processing. Besides, the classification performance shows that the advantage of NBAN is particularly obvious when the size of the selected band subset is small. Extensive experiments strongly evidence that the proposed NBAN method outperforms many current models on three popular HSI images consistently.**

***Index Terms*—Attention mechanism, band selection, global relationship, hyperspectral image, spectral reconstruction.**

I. Introduction

HYPERSPECTRAL images (HSIs) have been widely applied in various fields, such as agriculture [1], [2], land management [3], medical imaging [4], [5], and forensics [6]. Hundreds of bands in hyperspectral images not only contain rich spectral and spatial information but also bring a great challenge for hyperspectral data processing. Due to the imaging characteristics of HSIs, there is a high correlation between adjacent bands [7], thus leading to huge data redundancy. The high dimensional and redundant HSIs data will result in huge expenditure and extravagant computing resources. On the other hand, it often suffers from the so-called curse of dimensionality [8], [9], which will impair the classification ability of classifiers.

Feature extraction and band selection (BS) are the two most common methods to transform the high-dimensional HSI data to a lower one [10]. Feature extraction methods are widely used in HSIs data processing [11]– [13]. The core idea of these methods is to find a mapping from high-dimensional space to low-dimensional space. However, feature extraction changes the original feature space and causes the loss of the physical characteristics of HSI data [10]. The basic idea of BS is to select the most representative band from the original data. Compared with feature extraction [14], BS preserves the main physical attributes of the data to a great extent and protects the information of the original data as much as possible [15].

BS methods can be classed into supervised and unsupervised methods. Since no prior knowledge is needed and its better robustness, unsupervised BS methods have attracted a great deal of attention. Over the past decade, many unsupervised BS methods have been proposed [16]. Some of BS methods view Band selection as a combinational optimization problem and use a heuristic searching method to optimize it, such as multiobjective optimization-based band selection (MOBS) [17]–[19]. Some of them are the cluster-based methods which cluster the spectral bands and select the target bands, such as subspace clustering (ISSC) [7], [20]. These methods consider the similarity between spectral bands and achieved good results in recent [7], [20]. Other BS methods are based on band-ranking which assign a rank for each spectral band by assessing their score, e.g., maximum-variance principal component analysis (MVPCA) [21], sparse representation (SpaBS) [22], [23], and geometry-based band selection (OPBS) [24].

Many existing BS methods commonly view every single spectral band as an independent feature. However, there is a nonlinear relationship exists between each band [7], [25]. Cai *et al.* [25] proposed an end to end framework (BS-net), which uses a convolution layer and an attention module to reconstruct the original data and to find the connection of bands. However, due to the limitation of the convolution kernel BS-Net can not explore the nonlinear relationship between bands over a long distance.

Recently, deep neural network (DNN) [26], [27] has attracted increasing attention in HSI processing. Due to its ability to find the nonlinear relationship between the features, DNN has been

Manuscript received January 6, 2021; revised February 4, 2021; accepted March 7, 2021. Date of publication March 12, 2021; date of current version April 5, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61773355, Grant 61973285, and Grant 61603355, in part by the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) under Grant CUGL17022, and Grant 1910491T06, and in part by the National Nature Science Foundation of Hubei Province under Grant 2018CFB528. *(Corresponding author: Zhihua Cai.)*

Tiancong Li, Yaoming Cai, Zhihua Cai, and Qiubo Hu are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: litiancong0331@outlook.com; caiyaom@cug.edu.cn; zhcai@cug.edu.cn; 1208966760@qq.com).

Xiaobo Liu is with the School of Automation and also with the Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, China University of Geosciences Wuhan 430074, China (e-mail: xbliu@cug.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3065687

widely applied in HSI classification and feature extraction [25]. With the development of the DNN, convolution neural network (CNN) [28] a variant of DNN, has been proven powerful to extract spatial relationships between images and it has become one of the most popular models for HSI processing [29]. CNN is widely used in many neural network models and architectures. For example, the auto-encoders of CNN version are always used to deal with image reconstruction. In addition, attention mechanism [30] have attracted increasing attention for the image classification problem, due to its ability to make the whole framework focus on salient information. Many channel weighting methods are equipped with attention modules, for example, Residual attention network [31] and spatial transformer networks [32]. Due to the ability of the different version for DNN, it also can be used in extract the correlation between the spectral, i.e., [25]. BS-net consider the nonlinear correlation between the spectral bands, that is why BS-net performances better than other exists BS method. Our proposed framework continues the idea of BS-net and uses DNN to find the relationship between bands, which we discussed in detail in Section III.

In this article, we develop a band selection network framework that considers the global relationship of all bands called nonlocal band attention network (NBAN). Specifically, we assume that there is a long-range relationship that can help an informative band subset to restore the complete spectral band set. Instead of evaluating the connection between adjacent bands, our framework extracts the long-range relationship by an attention score matrix that is generated with an attention module. Finally, NBAN is end-to-end trainable which makes it can be viewed as a unified framework and combined with many popular networks.

To sum up, the main contributions of this work are as follows.

1) By assuming a long-range relationship exists between the spectral bands, we propose a novel method for HSI band selection called NBAN. Our proposed method measures the significance of each band by calculating the restore contribution of the target band to other bands and an attention score matrix is used to extract the long-range relationship between the spectral bands. Finally, the attention score matrix is applied to hyperspectral band selection directly, which attempts to provide a new idea for unsupervised band selection.
2) We introduce nonlocal attention into the module of BS-Net [25] by considering long-range relationship which means that we have a receptive field in the process of band selection. The long-range relationship makes NBAN have global metrics rather than only consider a small range of band relationships. That enables NBAN to achieve a better performance when select a small size of band subset.
3) We show that the proposed method can better shield the noise bands and achieves a good result on three HSI datasets. The final experiment results show that our proposed method achieves the best performance not only on the classification performance but also on the correlation between the selected band subset. At the end of the experiment, we analyze why our framework can better avoid selecting those noise bands and achieve better classification performance than other BS methods by combining with the characteristics of information entropy of three datasets.

The rest of the article is structured as follows. In Section II, we first describe motivation and review the related work. Second, we define the notations and show the details of our proposed method in Section III. Next, we design experiments to compare with existing BS methods and discuss their results in Section IV. Finally, we conclude with a summary and final remarks in Section V.

II. Related Work and Motivation

A. Attention Mechanism

The inspiration for attention mechanisms mainly comes from human beings. The core idea of attention mechanism is to make the modules ignore extraneous information and focus on key information. Attention mechanism is widely applied in natural language processing [33], [34] and image processing [29], [32], [35], [36]. In this article, we mainly focus on its application in image processing. Attention module can be considered as a function f which measures the significance of the features and formulates an attention map. The attention map can be taken as a reference to reweight the raw data. In image processing, the task of attention module can be defined as follows:

$$\mathbf{H} = \mathbf{Z} \otimes \boldsymbol{a}. \tag{1}$$

Here, $\boldsymbol{a}$ is a score vector of features $\boldsymbol{a} \in \mathbb{R}^b$ that generated by attention module, Z denotes a feature map $Z \in \mathbb{R}^{mn \times b}$, and H is the resulted feature map $H \in \mathbb{R}^{mn \times b}$. Attention module f is widely achieved by a neural network, that makes $\boldsymbol{a}$ can extract the nonlinear relationship from the original feature map. The network focuses the key information from the whole training process, and generate an attention map. By combining the attention map $\boldsymbol{a}$ and original feature map Z, H will focus on the key information and give less attention to the extraneous information.

Due to the different objects of concern, attention modules can be classed into spatial attention, channel attention [35], and joint attention [37], [38]. The spatial attention is utilized to learn the relationship between the spatial pixels. In practice, convolution kernel is widely used in spatial attention modules, due to its powerful ability to extract the information between adjacent pixels. Meanwhile, the convolution operation is also applied in the channel attention mechanism. For example, Hu *et al.* [35] proposed a simple network branch that uses an average pooling layer and convolution layer to squeeze the spatial information and get channel attention. On the other hand, due to the limitation of kernel size, most of the spatial attention modules of using convolution kernels cannot consider the long-range relationship between elements. Although the focus of spatial attention mechanism and channel attention mechanism is different, the shortcomings of convolution kernels leads to the limited ability to extract contextual relations from spatial attention and channel attention. To solve this problem, Wang *et al.* [39] proposed a network that calculates the similarity between pixels and learn the long-range relationship from the data. They calculated the

similarity between all the pixels and obtain a more comprehensive relationship between all pixels.

To sum up, attention mechanism has great potential in feature selection. In this article, we not only use the traditional channel attention but also use some concepts of spatial to find the long-range relationship between bands. In the following section, we will show the attention modules of our proposed framework and discuss how it works.

B. Auto-Encoder

As a structure of DNN, auto-encoder is widely applied in neural language processing [40], [41], and image processing [42], [43]. With the development of CNN, an auto-encoder of convolutional version can better extract the information from the images than the original one. In this article, we mainly focus on the auto-encoder of convolutional version. In practice, We define an auto-encoder as a function f, which takes a tensor $\mathbf{X}$ as input and outputs a resulted tensor $\mathbf{Y}$. Then the auto-encoder can be defined as follows:

$$\mathbf{Y} = f(\mathbf{X}; \Theta) \tag{2}$$

where Θ denotes the trainable parameters in the auto-encoder. The training process of auto-encoder can be defined as two stages: feedforward and backward. In the feedforward process, the auto-encoder transforms the input tensor $\mathbf{X}$ into a latent space by its encoder layer. In order to extract the information, the encode layer always performs convolution operation in image processing. Then the decode layer tries to restore the data and produces a certain output $\mathbf{Y}$. The encode layer and decode layer are composed of multiple convolution kernels of different sizes and after convolution operation, there is an elementwise function between convolution kernels.

The second stage is called backward. After the stage of feedforward, the auto-encoder needs to update the parameters by using the method of gradient descent. A cost function is used to calculate the cost between the original tensor $\mathbf{X}$ and the result tensor $\mathbf{Y}$. Then a method such as mean square error (MSE) is utilized to minimize the cost. Finally, cost function can be defined as

$$L(\Theta) = \mathrm{Cost}(\mathbf{X} - \mathbf{Y}; \Theta) \tag{3}$$

where Θ denotes the parameters of the auto-encoder and Θ is updated by

$$\Theta = \Theta - \eta \frac{\partial L}{\partial \Theta}. \tag{4}$$

Here, η is learning rate and ∂ denotes the partial derivative operation.

C. Motivation

The purpose of BS is to select some representative bands to improve computational efficiency. This article purpose based on the assumption, i.e., select the band set with global characteristics will perform better than those band only consider the local relationships. However, most of BS methods divide the whole band set into several categories or just evaluate each band as an independent feature [21]–[24]. These methods limits the expression ability of the selected band set, and make the result fall into a trivial solution. A method to solve this problem is to enlarge the receptive field of the network, such as extract the long-range relationship of the whole band set. By assuming a band can be jointly represented by the others bands, the data of original band set can be written as $\mathbf{XC} = \mathbf{X}$, where $\mathbf{C}$ is a score matrix to reveal the significance of each band to other bands. Moreover, the score matrix $\mathbf{C}$ can be used to select the most informative bands as an important reference. As a deep learning method, BS-Net takes convolutional neural networks as band attention module and reconstruction network which makes it more advantageous to other BS methods. However, it is also face some shortcomings such as the following. 1) The expression ability of the selected band subset is limited, especially when the size of the subset is small. 2) The score matrix C cannot extract enough information from the whole band set due to the limitations of the convolution kernel size. Hence, this article attempts to establish a new nonlocal evaluation framework to select the more global bands by extracting the long-range relationship between the spectral bands.

III. Proposed Network

We denote an HSI dataset consisting of b spectral bands and $n \times m$ pixels as $U \in \mathbb{R}^{n \times m \times b}$. For convenience, we regard U as $B = \{B_i\}_{i=1}^{b}$. Our goal is to find a function $\psi : \Omega = \psi(B)$ which can produce a subset contains the most representative bands. In this section we lay out an end-to-end trainable framework for BS, then describe how it works. To begin with, we summarize the structure of the model and then the details of each module are shown in the following sections.

A. Architecture of NBAN

The core idea of NBAN is to rank the significance of the bands in the process of sparse band reconstruction with a nonlocal way. We try to restore the whole band set by only using a few informative bands. In the process of reconstruction, those bands that can represent the vast majority of bands should achieve more attention. To this end, in order to select the most influential bands we proposed a framework consists of a band attention module and a reconstruction network.

The schema of NBAN is shown in Fig. 1. Aiming to rank the significance of the bands, we first consider the long-rage relationship between the bands and design a band attention module. The input data is first extracted the correlation by the attention module, and generate an attention score matrix. The band attention module is a branch network that contains the characteristics of spatial attention and channel attention. In this module, we use a matrix $\mathbf{C}$ called the attention score matrix to collect the long-range relationship between bands. Then $\mathbf{C}$ will help to reweight the original data. The original data are reweighted by matrix operation with reference to $\mathbf{C}$. The details of the attention module are shown in part B. Next, a reconstruction network is to restore the original spectral bands from the reweighted bands. The reweighted data are restored by the reconstruction module. The details of the reconstruction net

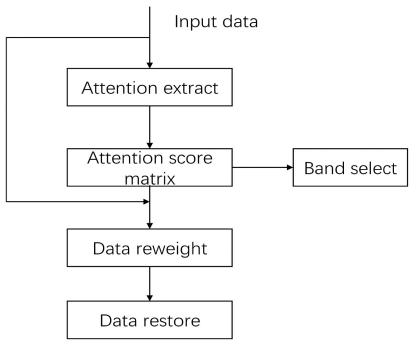


Fig. 1. Schema of NBAN. An attention score matrix is to compute the similarity between the spectral bands during the process of spectral reconstruction and the representative bands are selected by the final attention score matrix. The final selected spectral bands is selected by the attention score matrix after training.

are shown in part C. In the process of the reconstruction, the attention module adjusts the weight of bands to reconstruction and measure the significance of each band. After training, the final attention score matrix can be utilized to select the representative bands.

B. Nonlocal Attention Module

A nonlocal attention module is an embedded unit which can reweight the original data to a new feature map $\mathbf{O}$. To make the selected spectral bands more global, a reweight operation is to calculate the correlation between each spectral band in a nonlocal way. This allows us to measure the relationship with a bigger receptive field. In Fig. 2, we show the details of the nonlocal channel attention module. A feature map $\mathbf{X} \in I$ which consists of $d \times d$ pixels and b spectral bands is given as an input. The reweighted dataset O can be defined as

$$\mathbf{O} = \mathbf{X} \otimes \mathbf{C} \tag{5}$$

where $\mathbf{X} \in \mathbb{R}^{d^2 \times b}$, and $\otimes$ denotes reweight operation. $\mathbf{C}$ is an attention score matrix that used to extract the relationships between the spectral bands.

1) Attention Score Matrix: Comparing with the attention module in BS-net, we employ an attention score matrix to record the relationship we extract from the spectral bands so that the framework can learn more information from the original feature map.

We obtain the attention score matrix by attending to all pixels in each band and taking their weighted average in embedding space, this follows the design of [39], [30] and the nonlocal operation can be written as (6). We simplify the embedded Gaussian nonlocal module [39] and use the improved version on band attention. The similarity between each spectral band is calculated in a embedded Gaussian way and the attention score matrix measure the significance of each spectral band to others. Specifically, two (1×1) kernels with a stride of (1×1) are to learn the correlation between the spectral bands. Then to standardize the data, we follow it with a sigmoid function. The similarity function $f(x_i, x_j)$ can be defined as (7)

$$\mathbf{C_i} = \frac{1}{H(\mathbf{X})} \sum_{\forall j} f(x_i, x_j). \tag{6}$$

$$f(x_i, x_j) = e^{\sigma(\mathbf{X})^T \phi(\mathbf{X})}. \tag{7}$$

where $\frac{1}{H(\mathbf{X})}$ denotes a normalization function, and i is the index of the target position, and j is the index of enumerates all other positions. f denotes the similarity between two pixels.

In order to reweight the band set, we view each band as a combination of all bands and calculate the restore weights of bands to each other. Specifically, the greater the similarity between the bands, the greater the reconstruction weight. To ensure the standardization of the generated data, we set the sum of bands weight to 1 by calculating with a softmax function along with the column of the attention score matrix, which means that we can regard each column of the matrix as the reconstruction cost of the corresponding band and each line can represent the reconstruction weight of the corresponding band to other bands. The last attention score matrix can be written as

$$\mathbf{C} = \mathrm{softmax}(\mathbf{X}^T \mathbf{W}_\sigma^T \mathbf{W}_\phi \mathbf{X}) \quad s.t. \sum_{i=1}^{b} \mathbf{C}_{ij} = 1. \tag{8}$$

Here, $\sigma(\mathbf{X}_i) = \mathbf{W}_\sigma \mathbf{X}$ and $\phi(\mathbf{X}) = \mathbf{W}_\phi \mathbf{X}$. $\mathbf{W}_\sigma$ and $\mathbf{W}_\phi$ are the learning parameter of the convolution layer. The attention score matrix represents the relationship between pixels and the values in the matrix are all positive.

2) Band Reweighting: Next we describe the reweight operation. To reweight the data, we take $\otimes$ as a reweight operator and use the attention score matrix as a reference. Each element of the reweighted data is a combination of the elements of the same position in other bands, and use $\mathbf{C}_{ij}$ to denote the constituent weight, where $\mathbf{C}_{ij}$ refers to the element of the ith row and jth column in the attention score matrix. Then we can write the outputs element $\mathbf{O}_{ij}$ as (9). Finally, the reweighted data $\mathbf{O}$ can be calculated by (10)

$$\mathbf{O}_{ij} = \sum_{j=1}^{b} \mathbf{X}_{ij} \mathbf{C}_{ji}. \tag{9}$$

$$\mathbf{O} = \mathbf{X}\mathrm{softmax}(\mathbf{X}^T \mathbf{W}_\sigma^T \mathbf{W}_\phi \mathbf{X}). \tag{10}$$

Here, $\mathbf{O}_{ij}$ is the element of ith row and jth column on the reconstructed dataset.

C. Reconstruction Net

Following the attention operation, we employ a reconstruction net (RN) to restore the reweighted spectral bands. The RN can be defined as a function f which takes the reweighted data O as input data and outputs a restored dataset $\hat{X}$ as

$$\hat{\mathbf{X}} = f(\mathbf{O}; \Theta_c). \tag{11}$$

Here, Θ_c is the trainable parameters involved in RN.

The MSE is used as the cost function to help recover the data. We define the cost function $\mathcal{L}$ as follows:

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{S} \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_2^2 \tag{12}$$

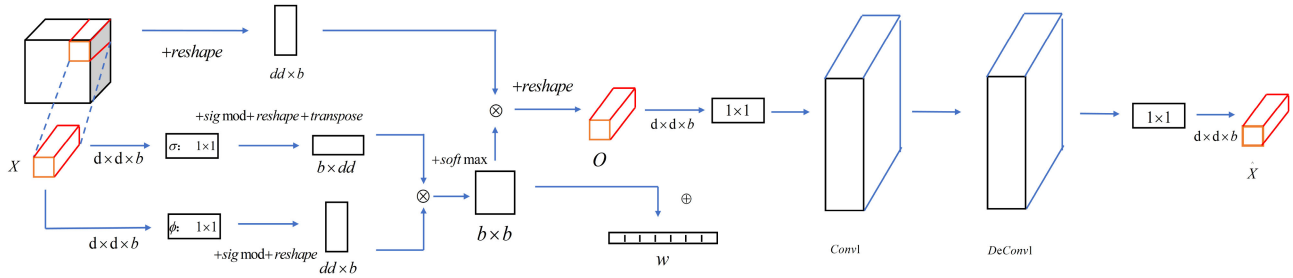


Fig. 2. Overall network structure of NBAN. The framework consists of a nonlocal attention module and a reconstruction network. $\otimes$ is an operator, which denotes the operation of matrix multiplication. O is the data that reweighted by the attention score matrix and has the same shape as X. The differences between X and $\hat{X}$ are calculated to feedback the framework and update the parameters.

where X denotes the original feature map and S is the number of training samples. Equation (12) can be optimized by using a gradient descent method, such as stochastic gradient descent (SGD) and adaptive moment estimation (Adam). The details of RN can be seen in Fig. 1. First, the reweighted data $\mathbf{O}$ processed by a (1×1) convolution kernel with a strids of (1×1). In order to restore the data, we consider the vanishing gradient and simplify the auto-encoder by only using one convolutional encoder(Conv1) and one deconvolutional decoder(Deconv1) to up-samples feature maps. We employ the reconstruction error between the prediction results and the original data to feedback adjustment and form the final attention score matrix $\mathbf{C}$, then $\mathbf{C}$ is a reference to select the bands.

Algorithm 1: NBAN.

Input: Band Set: $U \in \mathbb{R}^{N \times M \times b}$, the number of selected bands: n

Output: The selected band subset

```
Preprocess the band set U;
Generate the training samples: X from U;
while training iteration reaches the maximum iteration do
    Extract long-range relationship to generate the attention score matrix: C;
    Reweight original data: O = X ⊗ C;
    Restore spectral bands: X̂ = f(O; Θ_c);
    Calculate the reconstruction cost and update the parameters;
end
Get weight vector ω from final attention score matrix;
Select top n bands from the weight vector: ω;
```

D. Informative Band Subset Selection

In this step, our goal is to measure the significance of each spectral bands. In order to select the informative band subset, we evaluate the importance of each band by a vector $\omega = [\omega_1, \omega_2, \ldots, \omega_i]$, where ω_i refers to the importance of the ith band. As we mentioned in part A, we view each line in $\mathbf{C}$ as the reconstruction weight of the corresponding band to other bands. In other words, the greater the reconstruction weight, the more important the band is to other bands. With this solution, we assess ω_i as

$$\omega_i = \sum_{j=1}^{b} \mathbf{C}_{ij} \tag{13}$$

where ω_i denotes the evaluation weight vector of all bands, i is the line number and j is the column number of the attention score matrix. And then we sort ω and select the informative band set. The pseudocodes of NBAN are shown in Algorithm 1.

IV. Experiment and Discussion

In this section, we explore the use of NBAN and discuss how it works on real three datasets. In part A, we begin with introduce three datasets, training details and evaluation criteria. Then, we test NBAN with a classifier, analysis the convergence of the framework, and compare the performance with six popular BS methods in part B. Finally, we investigate the reasons why NBAN performs better from the selected band subsets in part C.

A. Dataset and Training Details

To evaluate the influence of NBAN, we employ Indian Pines, Pavia University, and Salinas as testbed for exploring the performance of NBAN. Indian Pines consists of 145×145 pixels and 200 spectral bands. The dataset includes 16 kinds of different categories and in the wavelength range $0.4 - 2.5(\times 10^{-6})$ meters. Pavia University consists of 610×340 pixels and 103 spectral bands. The dataset is divided into nine classes and in the wavelength range of $430 - 860$ nm. Salinas consists of 512×217 pixels with 204 spectral bands in the wavelength range of $0.36 - 2.5(\times 10^{-6})$ meters and contains 16 classes.

For better evaluating the performance of the selected band subsets, support vector machine (SVM) is utilized [44]–[46] as the classifier. We randomly select 5% labeled samples from three datasets as the training set and set the optimal window size as 7×7 for each dataset. We train the network for 80 epochs on Pavia University, and 100 epochs on Indian Pines and Salinas. The kernel size of Conv1 and Deconv1 are $3 \times 3 \times 128$. The optimum learning rate used for three datasets is 0.00001. Overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) are calculated by NBAN for 20 independent runs.

To analyze the selected band set, information entropy and mean spectral angle (MSA) [15], [47], [48] are calculated as an

TABLE I
HYPERPARAMETERS OF THE BS METHODS

Methods	hyper-parameters
NBAN	$\lambda = 1e5, maxiter = 100$
BS-Net-Conv	$\lambda = 1e-2, \eta = 2e-3, maxiter = 100$
ISSC	$\lambda = 1e5$
SpaBS	$\lambda = 1e2$
MVPCA	-
MOBS	$maxiter = 100, NP = 100$
OPBS	-

important evaluation criteria. Entropy is calculated to represent the amount of information in one band about another band. The entropy can be defined as follows:

$$H(B_i) = -\sum_{y} P(y) log P(y) \tag{14}$$

where $P(i)$ denotes the grey level of histogram bins of B_i. The larger entropy is, the greater the amount of information exists.

MSA is an average unit of measurement to indicate the degree of data matching for a band set. The MSA for band subset B can be written as

$$MSA(B) = \frac{2}{k(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha(i,j) \tag{15}$$

where $\alpha(i,j)$ denotes the spectral angle between the ith band and jth band. $\alpha(i,j)$ can be calculated as

$$\alpha(i,j) = arccos(\frac{B_i^T B_j}{||B_i||\ ||B_j||}). \tag{16}$$

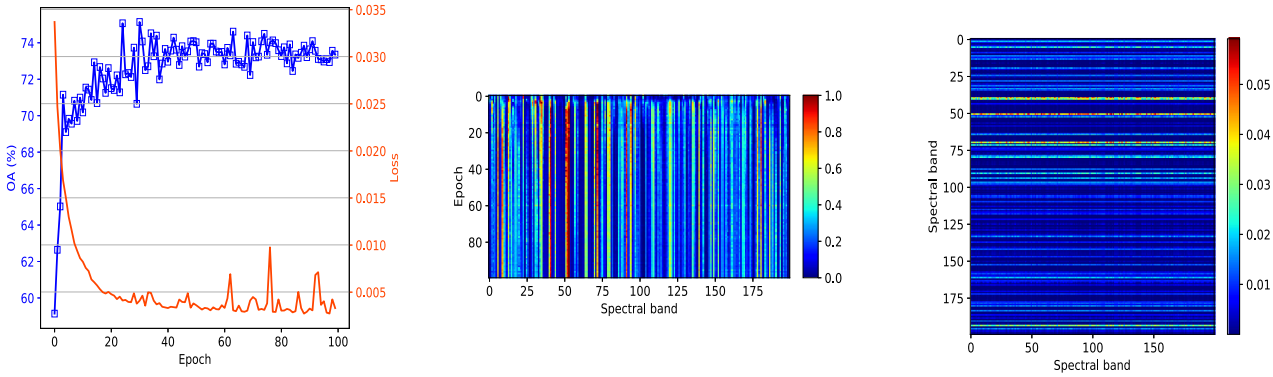
The larger MSA is, the less redundancy is contained between the band subset. The methods are evaluated with Python 3.5 running on an Intel Xeon E5-2620 2.10 GHz CPU with 32 GB RAM [25]. We implement all methods with TensorFlow-GPU 1.6.1 and accelerate them on an NVIDIA RTX-2080TI GPU with 11 GB graphic memory and the hyperparameters of the contrast methods are listed in Table I.

B. Experiment Results

In this part, we design the experiments to prove the effectiveness of NBAN. We employ ISSC, SpaBS, MVPCA, MOBS, OPB, and BS-Net-Conv as our comparative methods. After that, the performance of using all bands is also compared with our method as an important reference.

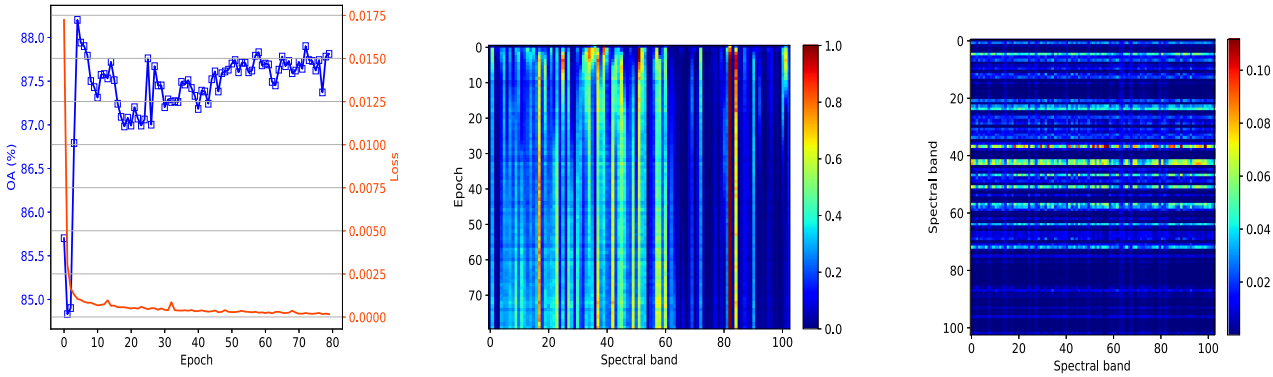
1) Analysis of Convergence of NBAN: In this part, we discuss the convergence of NBAN on different HSIs, with visualize the loss curves and the classification accuracy. For Indian Pines, the curves of loss and classification are shown in Fig. 3(a). It can be seen that when we train NBAN on Indian Pines the reconstruction errors decrease and the accuracy of SVM increases at the same time. The loss values of NBAN close to 0.002 after 20 iterations, and there has been a huge improvement in accuracy at the same time. Finally, the value of accuracy stabilizes around 73% after 40 interactions when we train our method on Indian Pines. There is nearly a 13% improvement in accuracy, which means that our method has a good effect on the band selection. The use of long-range attention mechanism will make loss converge faster than other BS methods, it is another advantage of our method. Similar to Indian Pines, the loss curves of Pavia University and Salinas are shown in Figs. 4(a) and 5(a). Fig. 3(b) represents the selection process of NBAN on Indian Pines. For convenience, we scale the bands' weight into range [0,1] and find that the importance of the band changed with the training iteration. Almost all the spectral bands' weights are same at first, but with the increase of the training iterations, we can observe that some of them become prominent compared with other bands which means that we can view this phenomenon as a process of band selection. Furthermore, to explore the relationship between the correlation matrix and the selected bands, we further visualize the correlation matrix of the trained network on Indian Pines. As we can see in Fig. 3(c), the informative bands and the trivial bands are distinguished by the lines in the matrix. The horizontal lines in the graph of the attention score matrix mean our method enhanced the weight of some specific bands successfully instead of randomly increasing the weight of the matrix. In Fig. 4(b) and Fig. 5(b) and (c), we can see the selection process and the final attention score matrix of Pavia University and Salinas. For Pavia University, the selected bands mainly concentrated before the 80th band. However, the selected bands are evenly distributed on Salinas. The specific band distribution will be discussed in detail in part C.

2) Performance Comparison: To show the classification performance of our method, we compare the classification results of different BS methods under different sizes of band subset. For Indian Pines, we can see from Fig. 8 that NBAN achieves the best OA when the band subset size is over 5, followed by BS-Net-Conv, MOBS, and others BS methods. It is observed that two deep learning methods NBAN and BS-Net-Conv perform better than other BS methods in most cases and the OA of NBAN increases larger than 70% when the subset size is only 15. Then we find a counter-intuitive phenomenon from these curves. The classification performance is not always increased by selecting more bands in some BS methods. We find that the OA curve of SpaBS shows a downward trend when the subset size larger than 17 and the classification performance of BS-Net-Conv start decrease when the subset size larger than 23, it is the so-called Hughes phenomenon [9], [8]. Furthermore, we can observe that the OA curve of NBAN rises continuously throughout the whole curve which means that NBAN can select more informative bands on Indian Pines. For Pavia University, it can be seen from Fig. 8 that NBAN achieves better OA than other methods when the subsets are smaller than 15. When the subset size larger than 20, NBAN, MOBS, BS-Net-Conv, and ISSC achieve close OA. Although MOBS achieves a better performance when the size of the subsets larger than 17, our proposed method is still comparable to it. Because of the Huges phenomenon, the classification accuracy of NBAN increases first and then decreases with the selected bands. Then for Salinas, we can see from Fig. 8(c) that NBAN achieves the best OA when the size of the subset larger than 20. The OA of most BS methods no longer increases unless NBAN when the subset size larger than 21, which means that our method can choose more informative bands. Moreover, when the subset size larger than 19, two deep



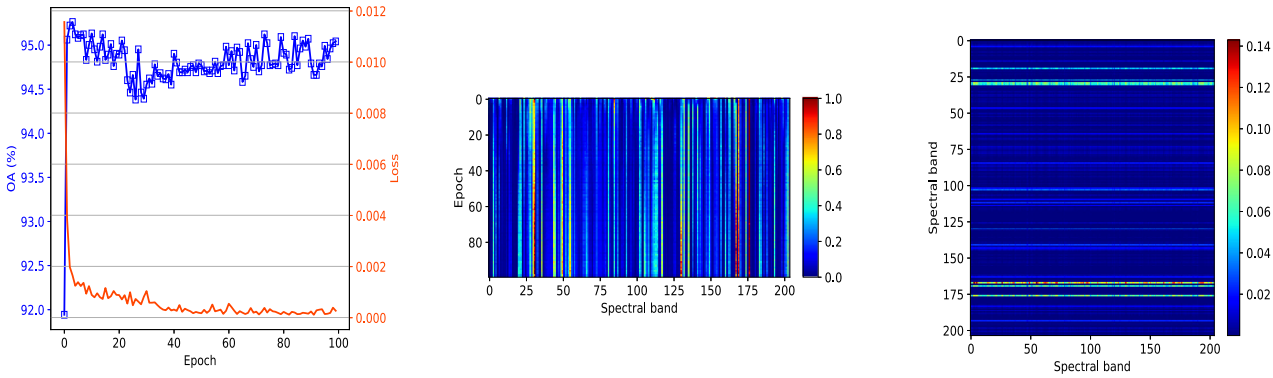
(a) (b) (c)

Fig. 3. Analysis of the convergence of NBAN on Indian Pines. (a) Curve of the loss and accuracy under different iterations. (b) Band weights under varying iterations. (c) Final attention score matrix.



(a) (b) (c)

Fig. 4. Analysis of the convergence of NBAN on Salinas. (a) Curve of the loss and accuracy under different iterations. (b) Band weights under varying iterations. (c) Final attention score matrix.



(a) (b) (c)

Fig. 5. Analysis of the convergence of NBAN on Salinas. (a) Curve of the loss and accuracy under different iterations. (b) Band weights under varying iterations. (c) Final attention score matrix.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS USING 25 BANDS ON INDIAN PINES DATASET

	BS-Net-Conv	ISSC	SpaBS	MVPCA	MOBS	OPBS	NBAN	All band
OA	70.61±2.56	54.95±2.01	56.37±3.12	49.75±2.00	64.22±3.46	55.59±2.96	**73.34±2.21**	62.78±2.82
AA	75.17±1.17	63.06±1.07	63.16±1.23	55.47±1.04	70.67±1.23	62.41±1.12	**75.21±0.79**	73.8±0.86
Kappa	0.717±0.013	0.579±0.012	0.580±0.014	0.494±0.011	0.666±0.014	0.572±0.013	**0.718±0.009**	0.700±0.010
1	59.83±28.50	26.28±16.72	36.56±20.05	19.77±17.26	45.65±22.02	37.26±24.51	**61.82±18.28**	26.93±16.07
2	**74.34±3.07**	63.28±3.97	59.11±5.08	53.26±2.69	70.50±3.49	59.71±4.41	73.74±2.85	69.15±3.27
3	62.38±4.49	49.26±6.91	48.64±2.93	37.32±3.95	61.95±5.00	48.14±4.00	**65.61±4.3**	55.5±6.72
4	54.48±14.08	34.81±8.84	38.56±9.78	13.48±3.38	43.32±13.31	28.46±11.24	**57.78±9.93**	37.94±11.93
5	83.17±5.56	65.56±8.52	82.34±4.01	64.29±6.32	78.88±7.18	78.16±6.24	**89.92±3.9**	82.31±4.7
6	94.63±2.51	83.58±3.14	82.67±4.96	81.44±4.17	86.36±3.76	87.62±4.66	**95.70±1.67**	92.86±2.82
7	52.18±40.35	21.09±19.21	26.44±27.63	33.91±24.45	36.35±33.38	30.94±31.70	**62.04± 20.58**	35.46±29.78
8	96.32±2.41	87.10±4.10	86.45±7.78	86.60±8.29	95.96±1.91	89.09±7.15	96.5±2.20	**96.90±2.1**
9	28.35±26.93	7.24±9.77	4.64±6.76	2.66±3.96	16.86±21.22	13.75±15.83	**49.74±17.89**	16.36±16.69
10	**69.05±5.38**	54.38±7.40	52.00±5.81	43.33±5.70	63.55±6.34	49.44±4.85	61.89±4.28	68.59±4.76
11	71.39±3.76	62.30±3.02	58.06±3.05	49.13±3.53	65.44±3.41	60.99±3.36	70.10±2.33	**77.44±2.83**
12	**64.99±9.31**	35.06±5.41	40.61±7.02	32.35±8.25	60.10±10.04	30.29±5.88	64.17±5.27	47.64±9.07
13	95.24±2.94	87.69±8.86	89.42±6.71	80.21±13.33	90.86±6.93	79.20±9.12	**97.9±1.56**	92.19±4.66
14	89.03±3.89	83.56±4.74	88.61±3.71	85.26±4.20	89.88±3.68	86.74±5.04	90.59±2.69	**93.62±2.18**
15	49.17±7.54	33.29±9.82	41.14±8.54	32.10±6.37	34.38±6.19	30.76±6.73	**50.63±8.57**	46.81±6.72
16	85.16±5.15	84.70±4.74	66.69±19.80	81.94±6.44	87.47±5.03	78.89±16.62	**85.34±8.37**	64.62±17.49

The significance of bold entities means that the best performing parts of each category.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS USING 13 BANDS ON PAVIA UNIVERSITY AND 25 BANDS ON SALINAS DATASET

		BS-Net-Conv	ISSC	SpaBS	MVPCA	MOBS	OPBS	NBAN	All band
PAVIAU	OA	85.25±0.65	84.53±0.42	80.7±0.57	73.44±2.17	84.53±0.72	85.35±0.63	87.94±0.46	**90.10±0.49**
	AA	87.72±0.25	87.42±0.33	84.59±0.25	81.91±0.21	88.48±0.23	85.35±0.63	90.25±0.27	**92.09±0.33**
	Kappa	0.835±0.03	0.830±0.05	0.719±0.004	0.756±0.003	0.847±0.003	0.844±0.003	0.870±0.004	**0.895±0.004**
SL	OA	95.46±0.24	94.31±0.18	94.46±0.27	89.41±0.37	95.48±0.18	95.18±0.34	**95.89±0.16**	95.84±0.26
	AA	91.71±0.16	90.18±0.22	90.69±0.21	86.01±0.25	91.97±0.20	91.60±0.20	**92.33±0.17**	91.84±0.36
	Kappa	0.907±0.002	0.890±0.002	0.896±0.002	0.844±0.003	0.911±0.002	0.906±0.002	**0.914±0.002**	0.909±0.002

The significance of bold entities means that the best performing parts of each category.

learning methods achieve better classification performance than the performance of using all bands.

In order to observe the performance of each BS method, we show the details of performance for Indian Pines in Table II and the results of Pavia University and Salinas in Table III. For Indian Pines, it is observed that NBAN achieves the best OA(73.34%), AA(75.21%), and Kappa(0.718). NBAN achieves the best score in 10 classes and the method of using all bands wins in No.8, No.11, and No.14 class. BS-Net-Conv gets better performance in No.2, No.10, and No.12 class. Then for Pavia University, we can see that the method of using all bands achieves the best result, NBAN is worse than the method of using all bands but better than other methods when the subset size is set to 13. For Salinas, NBAN achieves the best performance on this dataset when the subset size is set to 21. MOBS and BS-net-Conv achieve very close results, in this subset size. The performances for the three dataset show that NBAN performs better than other BS methods, followed by BS-Net-Conv. BS-Net-Conv takes into account the nonlinear relationship between the spectral bands and achieves good results. Compared with BS-net, NBAN calculates the long-range relationship on this basis so that it gets the best performances.

In the process of comparison, we find that two deep learning methods perform more stable than other BS methods and the Hughes phenomenon always appears later. For Indian Pines and Salinas, NBAN and BS-Net-Conv achieve better OA than the OA of using all bands which means that the data redundancy between all bands affects the accuracy of the classification and prove BS methods is beneficial to data processing. Comparing with NBAN and BS-Net-Conv, we notice from the curves that NBAN is more advantageous when the subset size is smaller than 11 and two methods achieve close OA when the subset size larger than 13. This occurs because NBAN considers the long-range relationship between the whole band set. However, the size of convolution kernel limits the receptive field range of BS-Net-Conv. Therefore when we only choose a few bands from the subset of BS-Net-Conv, the bands with the highest score only can represent the bands within a limited area. Compared with convolution operation in BS-Net-Conv, nonlocal attention extracts more relationships from the global band set. When the subset size is small, the bands in the subset which contains long-range relationship can better represent the whole band set.

C. Analysis of the Selected Band Subset

To verify the selected band subset by NBAN is more informative, we visualize the selected band subset and the informative entropy. The subsets of selected bands for three dataset are shown in Table V. For the sake of fairness, we avoid the Hughes phenomenon and size of the band subsets for three datasets are $15, 15$, and 20.

1) Indian Pines: The distribution of selected bands for Indian Pines is shown in Fig. 6. To observe the characteristics of the

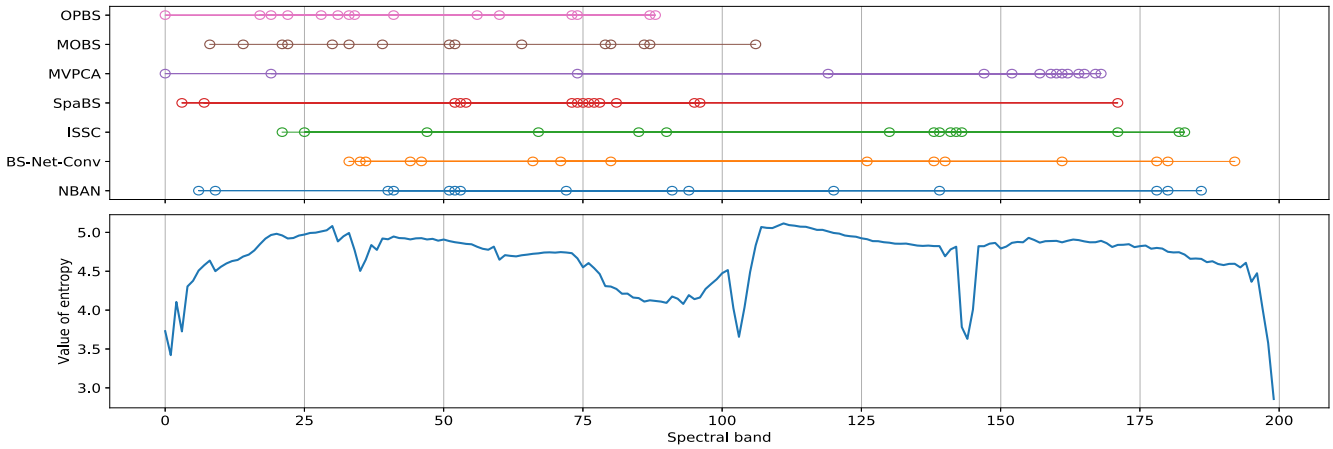


Fig. 6. Band distributions of the best 15 bands selected by different BS methods and the entropy value for Indian Pines.

TABLE IV
MSA OF THE BEST 15 BANDS FOR INDIAN PINES AND PAVIA UNIVERSITY. THE MSA OF THE BEST 20 BANDS FOR SALINAS

Method	Indian Pines	Pavia University	Salinas
NBAN	0.618	0.804	0.346
BS-Net-Conv	0.618	0.747	0.329
ISSC	0.631	0.8	0.434
SpaBS	0.562	0.809	0.309
MVPCA	0.782	0.813	0.345
MOBS	0.676	0.788	0.316
OPBS	0.835	0.755	0.321

bands, we visualized the information entropy of each band. For Indian Pines, we can see that there are some bands with low entropy in the band set from Fig. 6. These bands are called noise bands which lead to a negative impact on data processing and BS methods should avoid selecting these bands. As shown in Fig. 6, the band distribution selected by our proposed method is relatively uniform. The band subset selected by NBAN avoids those bands with low information entropy such as 0-3103-112 and 217-220. As we know that there are huge differences between the noise bands and the normal bands, NBAN has a long-range receptive field when selecting the band subset so that our method can better avoid the noise bands. In Table IV, we find OPBS and MVPCA achieve better MSA than other BS methods. However, our method does not perform very well. This happens because the band subsets of OPBS and MVPCA exist some noise bands. The noise band may cause an increase in MSA [15] because of its difference from other bands. We also find that the band distributions of OPBS and MVPCA are concentrated. However, the information gap between adjacent bands is always small which means that noise bands have a great influence on MSA. Compared with other BS methods, the classification performance of OPBS and MVPCA is poor due to selecting the noise bands.

2) Pavia University: Fig. 7 show the distributions of selected bands for Pavia university. The entropy curve of Pavia University is smoother than the curve of Indian Pines and shows an upward trend. According to the information entropy, we can divide the band set into three parts. The first part is consists of the bands before 20th with low entropy. The second one has the largest number of bands, distributes between bands 20th–80th. The last part distributes after the 80th which has the highest entropy and there is a drop near the 70th band. We can see that the selected bands of NBAN mainly distribute in the middle position and there are two bands distribute after the 80th bands. That happened because we consider the long-range relationship so that NBAN can choose a subset of bands that match the overall band as much as possible. So the selected bands of NBAN mainly distribute between 20th and 60th. Meanwhile, NBAN also avoids selecting those bands with very low entropy. Specifically, as shown in Fig. 8(b) when the subset is small NBAN has an advantage because the selected bands are more representative of most bands. However, when the size of subset getting bigger the bands with higher entropy may have more advantages, but NBAN is still comparable because there are some bands also distribute in the part of high information entropy. To further analyze the correlation between the selected bands, we show their MSA for Pavia University in Table IV. MVPCA achieves the best MSA in this part, but NBAN is still comparable. Although the MSA of MVPCA is high, it ignores the bands with higher information entropy in the process of selection. So the classification performance of MVPCA is worse than other BS methods. To sum up, the selected subset of NBAN achieves good results on classification performance and correlation performance when the subset size is small.

3) Salinas: We show the distributions of selected bands for Salinas in Fig. 9. From the entropy curve we observe that there are some sharply decreasing regions, i.e., 105–107, 146–147, and 200–203. Different with Pavia University, the value of entropy for Salinas is stable at about 4.5. As we can see from the bands distribution, NBAN avoid the sharply decreasing regions and distribute in the conventional bands. Meanwhile, NBAN also ignores the bands in 0–25 with the low entropy. The MSA of each BS methods are given in Table IV. ISSC achieves the best result and NBAN is in the second place. However, as shown in Fig. 8, ISSC achieve worse classification performance than most other BS methods. That happens because ISSC chooses too many continuous bands and cause information redundancy. In addition, ISSC also choose too many noise bands which lead to a negative

TABLE V
BEST 15 BANDS SELECTED BY DIFFERENT BS METHODS FOR INDIAN PINES AND PAVIA UNIVERSITY. THE BEST 20 BANDS SELECTED BY DIFFERENT BS METHODS FOR SALINAS

Data set	Methods	Selected Band Subset
Indian Pines	NBAN	[53, 41, 52, 91, 72, 9, 6, 139, 180, 178, 94, 186, 51, 120, 40]
	BS-Net-Conv	[46, 33, 140, 161, 80, 35, 178, 44, 126, 36, 138, 71, 180, 66, 192]
	ISSC	[171, 130, 67, 85, 182, 183, 47, 143, 138, 90, 139, 141, 25, 142, 21]
	SpaBS	[7, 96, 52, 171, 53, 3, 76, 75, 74, 95, 77, 73, 78, 54, 81]
	MVPCA	[167, 74, 168, 0, 147, 165, 161, 162, 152, 19, 160, 119, 164, 159, 157]
	MOBS	[8, 14, 21, 22, 30, 33, 39, 51, 52, 64, 79, 80, 86, 87, 106]
	OPBS	[28, 41, 60, 0, 74, 34, 88, 19, 17, 33, 56, 87, 22, 31, 73]
Pavia University	NBAN	[82, 46 ,56 ,37 ,84, 40, 23, 51, 43, 58, 34, 27, 57, 21, 45]
	BS-Net-Conv	[90, 42, 16, 48, 71, 3, 78, 38, 80, 53, 7, 31, 4, 99, 98]
	ISSC	[51, 76, 7, 64, 31, 8, 0, 24, 40, 30, 5, 3, 6, 27, 2]
	SpaBS	[50, 48, 16, 22, 4, 102, 21, 25, 23, 47, 24, 20, 31, 26, 42]
	MVPCA	[48, 22, 51, 16, 52, 21, 65, 17, 20, 53, 18, 54, 19, 55, 76]
	MOBS	[4, 15, 23, 25, 33, 35, 42, 53, 58, 61, 62, 64, 67, 73, 101]
	OPBS	[90, 62, 14, 0, 2, 72, 102, 4, 33, 1, 6, 84, 45, 82, 8]
Salinas	NBAN	[167, 176, 169 ,31 ,138 , 50 ,55 , 30 , 28 ,130 , 49 ,36 ,117 , 44 ,132 , 81 ,105 ,190, 174 ,193]
	BS-Net-Conv	[116, 153, 19, 189, 97, 179, 171, 141, 95, 144, 142, 46, 104, 203, 91, 18, 176, 108, 150, 194]
	ISSC	[141, 182, 106, 147, 107, 146, 108, 202, 203, 109, 145, 148, 112, 201, 110, 113, 144, 149, 105, 154]
	SpaBS	[0, 79, 166, 80, 203, 78, 77, 76, 55, 81, 97, 5, 23, 75, 2, 82, 56, 74, 143, 85]
	MVPCA	[169, 67, 168, 63, 68, 78, 167, 166, 165, 69, 164, 163, 77, 162, 70, 62, 160, 161, 76, 158]
	MOBS	[16, 17, 19, 45, 52, 58, 62, 75, 81, 97, 99, 131, 135, 136, 139, 142, 173, 174, 176, 181]
	OPBS	[44, 31, 37, 66, 11, 1, 164, 2, 18, 0, 3, 40, 4, 54, 33, 96, 5, 36, 6, 151]

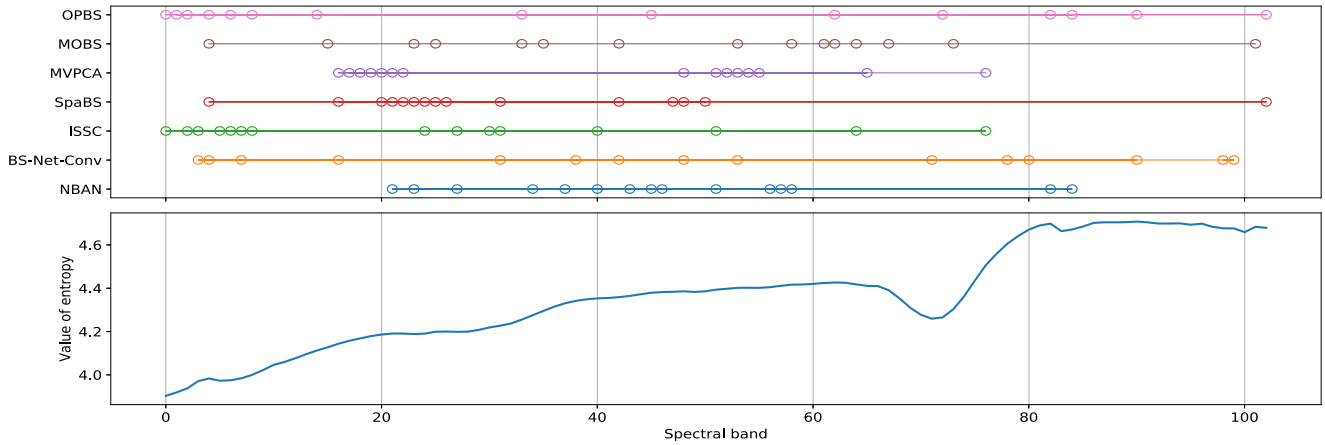


Fig. 7. Band distributions of the best 15 bands selected by different BS methods and the entropy value for Pavia University.

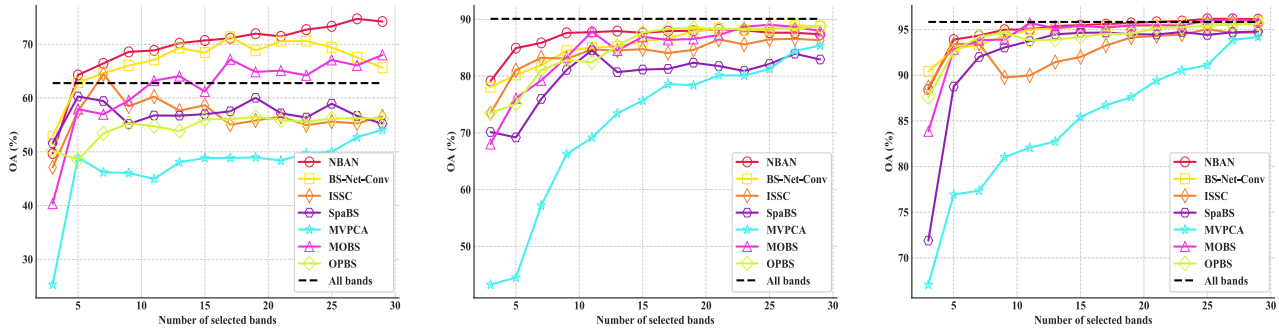


Fig. 8. Classification performance of using different band subset size on three dataset. (a) Performance on Indian Pines. (b) Performance on Pavia University. (c) Result of Salinas.

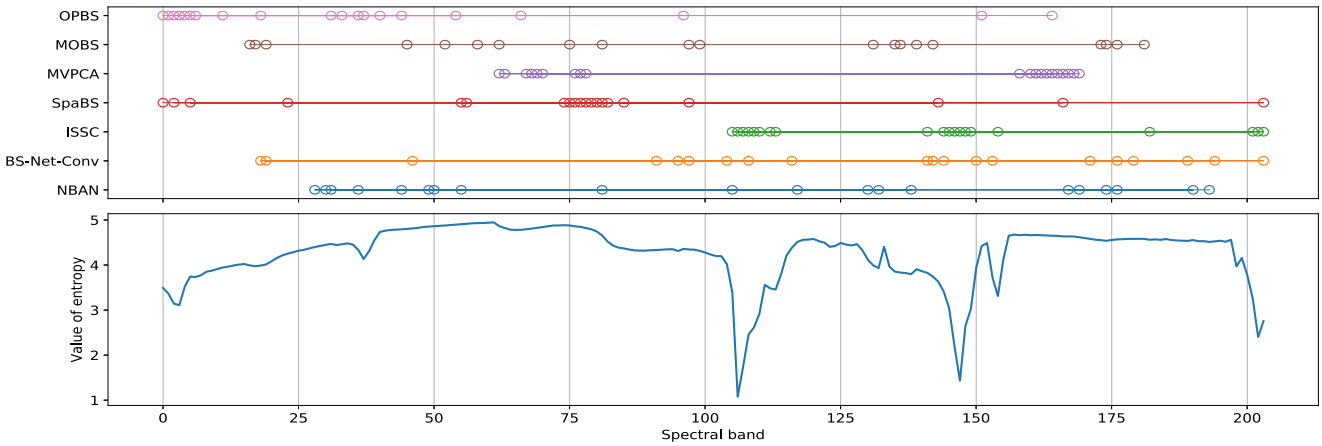


Fig. 9. Band distributions of the best 20 bands selected by different BS methods and the entropy value for Salinas.

impact. The distribution of MOBS is similar to that of NBAN, so it also achieves a good performance on classification. However, the bands in the position of 0–20 with low entropy make the classification performance of MOBS worse than NBAN. There are some noise bands in the selected band subset of BS-NET-Conv lead to a negative impact on classification performance. In a nutshell, NBAN can select the representative bands with low correlation and avoids the noise bands.

4) Discussion: From observing the distribution of selected band subset we find that the value of entropy has an effect on classification performance. However, the classification performance of selecting all the bands with high information entropy without considering the whole band classification will not achieve the best result. In addition, noise band has a negative effect on the classification performance, but it will reduce the correlation of the band subset. Since the similarity between the noise bands and the normal bands is always low and our method measures the significance of each band by considering the reconstruction contribution for the whole band set, NBAN can better avoid selecting those noise bands than other BS methods. Meanwhile, both from the classification performance and the distribution of the selected spectral bands we can observe that BS-Net-Conv and NBAN achieve better result than other BS methods. This phenomenon proves that it is important for BS to capture nonlinear relationships. For BS-Net-Conv, the use of CNN and fully connected neural network enables the framework to find a more comprehensive relationship between the spectral bands [25]. However, the antinoise ability and the interpretability of BS-Net-Conv is also limited. Compared with BS-Net, the attention module of NBAN uses matrix operations so that we can more easily interpret the effects of NBAN. Meanwhile, by calculating the global relationship, the selection result of NBAN can better avoid the interference of noise.

V. Conclusion

In this article, we propose a framework called NBAN with a no-local attention module to consider the long-range relationship from the whole dataset. The main idea of the framework is to restore the HSI data by using the correlation between the whole band set so that we can extract the long-range relationship and increase the receptive field of the network. The framework consists of two modules, nonlocal attention module and reconstruction network, making the whole network is end-to-end trainable. The attention module of NBAN is also a lightweight block makes our framework can be plugged into many network architectures. We conduct extensive experiments on three real datasets and prove our method is significantly better than many compared BS methods on classification performance. NBAN makes sure the selected bands are representative for the whole band set, so our method has more advantages when the subset size is small. Specifically, the use of attention score matrix makes the process of the band selection more explanatory. Meanwhile, we also summarize the relationship between some band noise and the degree of the correlation between the spectral bands.

Besides, in the process of the experiment we summarize the effects of noise band and information entropy on band correlation and classification performance. Then we find NBAN has a powerful ability to avoid noise bands due to its nonocal attention module. However, there may be some information that we ignore in the attention score matrix. In the future work, we will pay more attention to improving the interpretability of the framework and reducing the complexity of the model. The above-mentioned will be our future works.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive suggestions and criticisms. They would also like to thank the proposer of the contrast algorithm, especially Prof. M. Gong and Dr. W. Zhang who provided the source codes of the MOBS method and OPBS.

References

[1] M. Wang, Y. Wan, Z. Ye, X. Gao, and X. Lai, "A band selection method for airborne hyperspectral image based on chaotic binary coded gravitational search algorithm," *Neurocomputing*, vol. 273, no. 17, pp. 57–67, 2017.

[2] C. M. Gevaert, J. Suomalainen, J. Tang, and L. Kooistra, "Generation of spectral-temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3140–3146, Jun. 2015.

[3] J. Pontius, R. P. Hanavan, R. A. Hallett, B. D. Cook, and L. A. Corp, "High spatial resolution spectral unmixing for mapping ash species across a complex urban environment," *Remote Sens. Environ.*, vol. 199, pp. 360–369, 2017.

[4] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 10901.

[5] X. Liu, Q. Hu, Y. Cai, and Z. Cai, "Extreme learning machine-based ensemble transfer learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3892–3902, 2020, doi: 10.1109/JSTARS.2020.3006879.

[6] G. Edelman, E. Gaston, T. van Leeuwen, P. Cullen, and M. C. G. Aalders, "Hyperspectral imaging for non-contact analysis of forensic traces," *Forensic Sci. Int.*, vol. 223, no. 1–3, pp. 28–39, Nov. 2012.

[7] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, Mar. 2019.

[8] X. Liu, R. Wang, Z. Cai, Y. Cai, and X. Yin, "Deep multigrained cascade forest for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8169–8183, Oct. 2019.

[9] F. Melgani and L. Bruzzone, "Deep multigrained cascade forest for hyperspectral image classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[10] S. Sawant and M. Prabukumar, "A survey of band selection techniques for hyperspectral image classification," *J. Spectral Imag.*, vol. 9, 2020, Art. no. a5.

[11] A. Agarwal, T. El-Ghazawi, H. El-Askary, and J. Le-Moigne, "Efficient hierarchical-PCA dimension reduction for hyperspectral imagery," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, 2007, pp. 353–356.

[12] J. Wang and Chein-I Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[13] W. Li, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[14] X. Jiang, X. Song, Y. Zhang, J. Jiang, J. Gao, and Z. Cai, "Laplacian regularized spatial-aware collaborative graph for discriminant analysis of hyperspectral imagery," *Remote Sens.*, vol. 11, 2018, Art. no. 29.

[15] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 544–557, Jan. 2016.

[16] W. Sun, L. Tian, Y. Xu, D. Zhang, and Q. Du, "Fast and robust self-representation method for hyperspectral band selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 5087–5098, Nov. 2017.

[17] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 544–557, Jan. 2016.

[18] M. Zhang, M. Gong, and Y. Chan, "Hyperspectral band selection based on multi-objective optimization with high information and low redundancy," *Appl. Soft Comput.*, vol. 70, pp. 604–621, 2018.

[19] P. Hu, X. Liu, Y. Cai, and Z. Cai, "Band selection of hyperspectral images using multiobjective optimization-based sparse self-representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 452–456, Mar. 2019.

[20] W. Sun, L. Zhang, B. Du, W. Li, and Y. Mark Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2784–2797, Jun. 2015.

[21] C. Chang, Q. Du, T.-L.Sun, and M. L. G.Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[22] K. Sun, X. Geng, and L. Ji, "A new sparsity-based band selection method for target detection of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 329–333, Feb. 2015.

[23] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 631–644, Feb. 2015.

[24] W. Zhang, X. Li, Y. Dou, and L. Zhao, "A geometry-based band selection approach for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4318–4333, Aug. 2018.

[25] Y. Cai, X. Liu, and Z. Cai, "BS-Nets: An end-to-end framework for band selection of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1969–1984, Mar. 2020.

[26] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

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

[28] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.*, vol. 2015, 2015, Art. no. 258619.

[29] Z. Dong, Y. Cai, Z. Cai, X. Liu, Z. Yang, and M. Zhuge, "Cooperative spectral-spatial attention dense network for hyperspectral image classification," IEEE Geosci. Remote Sens. Lett., to be published, doi: 10.1109/LGRS.2020.2989437.

[30] A. Vaswani *et al.*, "Attention is All You Need," *Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[31] F. Wang *et al.*, "Residual attention network for image classification," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 6450–6458, doi: 10.1109/CVPR.2017.683.

[32] M. Jaderberg, K. Simonyan, A. Zisserman, and k. Kavukcuoglu, "IEEE example: Spatial transformer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. M. Lee Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2015, pp. 2017–2025.

[33] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 1412–1421.

[34] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Int. Conf. Mach.*, 2015, pp. 2048–2057.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[36] W. Ning, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image in painting," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107448.

[37] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Comput. Vis. - (ECCV) - 15th Europ. Conf.*, vol. 11211, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.

[38] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[39] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, doi: 10.1109/CVPR.2018.00813.

[40] B. Zhang, D. Xiong, J. Su, H. Duan, and M. Zhang, "Bilingual autoencoders with global descriptors for modeling parallel sentences," in *Proc. 26th Int. Conf. Comput. Linguistics: Tech. Papers*, Dec. 2016, pp. 2548–2558.

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. M. Ghahramani C. Welling Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2014, pp. 3104–3112.

[42] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, "Deep subspace clustering networks," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. S. Luxburg, H. Bengio Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2017, pp. 24–33.

[43] Y. Cai, Z. Zhang, Z. Cai, X. Liu, X. Jiang, and Q. Yan, "Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3018135.

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

[45] W. Sun and Q. Du, "Graph-regularized fast and robust principal component analysis for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3185–3195, Jun. 2018.

[46] W. Zhang, X. Li, and L. Zhao, "A fast hyperspectral feature selection method based on band correlation analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1750–1754, Nov. 2018.

[47] X. Geng, K. Sun, L. Ji, and Y. Zhao, "A fast volume-gradient-based band selection method for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7111–7119, Nov. 2014.

[48] J. Yin, Y. Wang, and J. Hu, "A new dimensionality reduction algorithm for hyperspectral image using evolutionary strategy," *IEEE Trans. Ind. Inform.*, vol. 8, no. 4, pp. 935–943, Nov. 2012.

Tiancong Li is currently working toward the Ph.D. degree in geosciences information engineering with the School of Computer Science, University of Geosciences, Wuhan, China.

His current research interests include deep learning, image processes, and hyperspectral image classification.

Yaoming Cai received the B.Eng. degree in information security from the China University of Geosciences, Wuhan, China, in 2016, where he is currently working toward the Ph.D. degree in geosciences information engineering with the School of Computer Science.

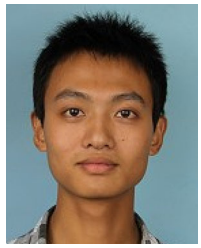
His research interests include machine learning, pattern recognition, evolutionary computation, and hyperspectral image classification.

Zhihua Cai received the B.Sc. degree in computing mathematics from Wuhan University, Wuhan, China, in 1986, the M.Sc. degree in computer software and theory from the Beijing University of Technology, Beijing, China, in 1992, and the Ph.D. degree in geodetection and information technology from the China University of Geosciences, Wuhan, China, in 2003.

He is currently a Faculty Member with the School of Computer Science, China University of Geosciences. He has authored or coauthored over 50 research papers in journals and international conferences, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: PART B-CYBERNETICS, *Applied Soft Computing*, *Information Sciences*, *Knowledge-Based Systems*, *Knowledge and Information Systems*, etc. His main research interests include data mining, machine learning, evolutionary computation, and their applications.

Xiaobo Liu received the M.S. degree in computer science and the Ph.D. degree in geosciences information engineering from the China University of Geosciences, Wuhan, China, in 2008 and 2012, respectively.

He is currently an Associate Professor with the School of Automation, China University of Geosciences. His research interests include machine learning, evolutionary computation, and hyperspectral remote sensing image processes.



Qiubo Hu received the B.S. degree in computer science in 2019 from the China University of Geosciences, Wuhan, China, where he is currently working toward the M.S. degree in computer science.

His current research interests include machine learning and data mining.