A Complementary Spectral-Spatial Method for Hyperspectral Image Classification

Lulu Shi, Chunchao Li, Teng Li and Yuanxi Peng

***Abstract*—In hyperspectral image classification, using spatial information as a supplement to spectral information is an effective way to improve classification accuracy. In this paper, a novel robust complementary method using spectral–spatial information is proposed to reduce the information loss in feature extraction, thus improving the classification effect. In short, two complementary feature extraction stages are used to get the probability maps for decision fusion. In the stage of pre-processing feature extraction, we propose an adaptive cubic total variational smoothing method (ACTVSP), which is first proposed and applied in the remote sensing research field, to obtain the first-stage probability map. At the same time, we utilize edge-preserving filtering in the post-processing stage and obtain the second probability map by the means of pixel-level classifier. Finally, the probability-like maps obtained in the above two stages are integrated by decision fusion rules. Experiments on ten public data sets show that the effectiveness of our proposed method and demonstrate the superior of distinguishing different land covers on the basis of very few training samples. Therefore, it can be applied to practical applications in different scenes.**

***Index Terms*—Hyperspectral classification, Feature extraction, Adaptive cubic total variation, Augmented Lagrange multiplier.**

I. Introduction

Hyperspectral images (HSIs) are made up of hyperspectral imager aboard the aerospace craft to capture three-dimensional images. Every pixel in the image contains abundant spectral information, space, and radiation, which have characteristic of spectrum close to continuous and mapping unity. These characteristics make HSIs widely used in many fields of actual target classification and recognition. Such as mineral exploration [1], land cover [2], [3], agricultural production [4], [5] and military target detection [6], [7]. At the same time, the dimension of hyperspectral image is often much larger than the number of available training samples, which brings many challenges to its classification and recognition applications. Hyperspectral images classification is one of the research hotspots in remote sensing.

The goal of classification of HSIs is to assign a category label to each pixel in the image according to the sample characteristics. Since different ground objects have different spectral curves, many methods using spectral features have been proposed for hyperspectral image classification. For example, support vector machine (SVM) [8], K-nearest Neighbor (KNN) [9], naive Bayes (NB) [10] and sparse representation classification [11] are all based on the spectral features of hyperspectral images. Among them, SVM using kernel transform technology [12] obtains a separate hyperplane through a small number of training samples, and then distinguishes different types of hyperspectral pixels, which avoids the phenomenon of Houghes and improves the classification effect a bit. Unfortually, the experimental results show that most machine learning classification algorithms based on spectral features are not ideal. The main reason lies in the properties of hyperspectral data. There are ”same thing different spectrum” and ”foreign matter same spectrum” problems that lead to highly nonlinear, eventually making some classification models based on statistical pattern recognition difficult to deal with the original hyperspectral images.

In order to address the above limitation, a number of researchers have introduced spatial information of hyperspectral images on the basis of spectral information. As in [13], [14], the spatial correlation between spatial pixels was used as a supplement to spectral information to enhanced and improved the classification performance of per-pixel classifier. The theoretical basis of this operation is that pixels belonging to a small area have a high probability of belonging to the same category, that is, adjacent pixels are associated to a large extent. Besides, many spectral–spatial combined classification methods [15]–[18] have been proposed in the past ten years, and the experimental results show that when using spectral-spatial information, the accuracy and robustness of classification results are greatly improved. According to the different processing ideas of spatial–spectral information combination, the existing methods can be mainly divided into two classification methods based on pre-processing and post-processing. The classification method based on pre-processing is to extract spatial features through certain structures and rules, then fuse them with spectral features, and finally feed them into the classifier. The spatial feature extraction method based on morphological contour is one of the most representative methods. For example, extended morphological profiles (EMPs) [19] have been proposed for constructing spectral–spatial features which are adaptive definitions of the neighborhood of pixels. In [20], invariant attribute profiles (IAPs) has been proposed to identify the same material from different scenes or locations in space. In [21], [22], a series of open and close operators with different scales are used to extract texture information from

This work was partially supported by the National Natural Science Foundation of China (No. 91948303-1, No. 61803375, No. 12002380, No. 62106278, No. 62101575, No. 61906210) and the National University of Defense Technology Foundation (No. ZK20-52). (Corresponding author: Teng Li, Yuanxi Peng.)

L. Shi, C. Li and Y. Peng are with the College of Computer Science, National University of Defense Technology, Changsha 410073, China, and also with the State Key Laboratory of High Performance Computing, Changsha 410073, China. e-mail: (sll@nudt.edu.cn; lcc@nudt.edu.cn; pyx@nudt.edu.cn)

T. Li is with the Beijing Institute for Advanced Study, National University of Defense Technology, Beijing 100020, China, and also with the College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China. e-mail:(liteng09@nudt.edu.cn)

images. Furthermore, spectral–spatial kernels were proposed to classify hyperspectral images in the kernel space. In [15], [23] and [24], researchers proposed the use of composite kernel, morphological kernel and graph kernel to improve SVM classifier [8] respectively.

Differently, the post-processing based classification method only uses spectral information to pre-classify HSIs by a pixel-by-pixel classifier at first, and then corrects the pre-classification results according to the spatial dependence between pixels to further improve the classification accuracy. For example, in [25], multiple logistic regression (MLR) is first used as a classifier to pre-classify HSIs to obtain a posterior probability, and then a Markov regularization term [26] describing prior probability is used for post-processing to obtain a new classification result map. In [27], [28], the post-processing is carried out by introducing the total variation regularization term to adaptively adjust the weight of pixels in spatial neighborhood. In [29], based on extended random walk (ERW), the correlation between adjacent pixels is used to optimize the classification probability.

Besides, many spectral–spatial information based deep methods have been proposed for hyperspectral image classification in recent years, such as convolutional neural networks (CNNs) [30]–[32]. In [33], compact and discriminative stacked autoencoder has been proposed to learn a low dimensional feature space. To address bottlenecks in complex scenarios that require fine-grained categorization, [34] provided a baseline solution by developing a general multimodal deep learning (MDL) framework. In [35], a classification method based on multi-view deep neural network is proposed to extract the spatial-spectral features of the scene with limited samples. However, compared with traditional machine learning methods, deep learning methods usually require more training samples and more parameter adjustment.

The above two methods are the main ideas of combination of spectral and spatial information. But there is a problem in both pre-processing and post-processing, that is, some information will be lost. The classification method based on pre-processing is to classify the image after feature extraction [36]. In the process of feature extraction, part of the texture information of the image will be lost, especially the target of small size is easy to be ignored. However, the classification method based on post-processing firstly classifies the image at the pixel level, which can avoid the problem of ignoring details of the method based on pre-processing. The accompanying problem is that it will be affected by the noise of the original image and lead to misclassification [37]. Even if the obtained probability map is optimized later, the probability distortion cannot be completely compensated. Therefore, the classification method based on post-processing also loses part of image information. From the above analysis, it can be known that since the information which two methods lost is different, it may be possible to combine them by some ways to reduce the loss.

To overcome the aforementioned issue, the final decision fusion based on the result of pre-processing and post-processing classification method was proposed. Recently, a fusion framework [38] based on pre-processing and post-processing classification method is proposed to improve classification effect. The main motivation of this integration is that the large scale objects can be extracted in the pre-processing stage, and the small scale objects can be modeled in the post-processing stage. Although the method uses the spatial correlation of pixels as the basis in the stage of pre-processing feature extraction, the differential processing of spectral information is neglected in SP method.

Considering the differential processing of spectral information, in this paper, an adaptive complementary spectral–spatial method is proposed. To be more specific, our method is mainly divided into two parts. In the first stage, an adaptive cubic total variational smoothing method (ACTVSP) is proposed, which is a pre-processing classification method. It should be pointed out that the adaptive parameters are the embodiment of different treatments for different bands. After the spectral and spatial smoothing of HSIs and feature extraction, an pixel-level classifier SVM [8] is used for classification. The second stage is based on the post-processing classification method. Firstly, the pixel direction classification map obtained by SVM classifier [8] is represented as multiple probability maps (the probability that the pixel belongs to a specific category), and edge-preserving filtering (EPF) [39] is carried out to classify each pixel according to the maximum probability. Finally, the probability-like maps obtained in two stages are integrated by decision fusion rules. The main motivation of the proposed method is that the rough structure profile of HSIs can be extract in the first stage, but details will be ignored and edges contour will be fuzzy. Meanwhile, class probability obtained after edge-preserving filtering optimization in the second stage can enhance the details and edge contour as a supplement to the first stage. To sum up, the main contributions of this paper are as follows:

- An adaptive hyperspectral image feature extraction method based on spectral–spatial joint feature is proposed, i.e., ACTVSP. It can smooth the spatial and spectral directions of the whole hyperspectral image while maintaining the geometric characteristics, reducing the impact of noise on classification.
- A robust complementary spectral–spatial method which absorb the pre-training thought of deep learning is proposed from a new perspective, which improves the performance of hyperspectral classification effectively. In a word, it is to make up for the information loss of feature extraction with the probability map obtained by post-processing.
- In order to prove the generalization ability and robustness of the algorithm, experiments are carried out on ten public hyperspectral data sets, including both classical and newer data sets. State-of-art classification effects were obtained under a set of fixed parameter values and a fixed number of training samples for all data sets.

The rest of this paper is organized as follows. Section II introduces the background of the proposed method. Section III introduces the proposed method. Section IV is experiment and analysis. And the conclusion is given in Section V.

II. Background

A. Vectorization Representation

A hyperspectral image (HSI) is a three-dimensional (3D) datacube. In this paper, it is expressed as $f(x, y, z)$, where (x, y) and z denotes the coordinate in spatial dimension and spectral dimension, respectively.

For the convenience of description, we use matrices and vectors to represent images and variables. For example, suppose that HSI has M rows, N columns, and T bands, we stack the input HSI $f(x, y, z)$ into a column vector of size $MNT \times 1$ according to the lexicographic order. For ease of reading, we use the bold letter $\mathbf{f}$ to represent the vectorized version of the HSI:

$$\mathbf{f} = \mathbf{vec}(f(x, y, z)) \tag{1}$$

where **vec** represents the vectorization operator.

B. HSI Modeling

Feature extraction of hyperspectral images faces many challenges. How to model hyperspectral images is an important problem to be considered first. Due to the observed hyperspectral images with a lot of noise [40], this paper adopts the following methods to model HSI:

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \eta \tag{2}$$

where $\mathbf{f} \in \mathbb{R}^{MNT\times 1}$ is a vector of size $M \times N \times T$ which denotes the unknown hyperspectral image, the matrix $\mathbf{H} \in \mathbb{R}^{MN\times MN\times T}$ is a three-dimensional convolution operation for linear transformations, and $\mathbf{g} \in \mathbb{R}^{MNT\times 1}$ is a vector denoting the observed hyperspectral image. $\eta \in \mathbb{R}^{MNT\times 1}$, is a vector denoting the noise. Our goal is to obtain HSI $\mathbf{f}$ from observed HSI $\mathbf{g}$ using the proposed model by smoothing the HSI.

C. Cubic Total Variation Operator

We define an 3-D total variation operator $\mathbf{D} = [\mathbf{D_x}^T, \mathbf{D_y}^T, \mathbf{D_z}^T]^T$, where $\mathbf{D_x}$, $\mathbf{D_y}$ and $\mathbf{D_z}$ are the first-order forward finite difference operators along the horizontal, vertical and spectral dimension, respectively. The specific formula is expressed as follows:

$$\mathbf{D_x}\mathbf{f} = \mathbf{vec}(f(x+1, y, z) - f(x, y, z)) \tag{3}$$

$$\mathbf{D_y}\mathbf{f} = \mathbf{vec}(f(x, y+1, z) - f(x, y, z)) \tag{4}$$

$$\mathbf{D_z}\mathbf{f} = \mathbf{vec}(f(x, y, z+1) - f(x, y, z)) \tag{5}$$

In order to control the smoothness of spatial and spectral dimensions more flexibly, this paper introduces the following three scaling factors $(\beta_x, \beta_y, \beta_z)$, multiplied by $\mathbf{D_x}$, $\mathbf{D_y}$ and $\mathbf{D_z}$ respectively, and obtains $\mathbf{D} = [\beta_x \mathbf{D_x}^T, \beta_y \mathbf{D_y}^T, \beta_z \mathbf{D_z}^T]^T$. Meanwhile, the total variation norm is defined as:

$$\|\mathbf{f}\|_{TV_2} = \sum_p \sqrt{\beta_x^2 [\mathbf{D}_x\mathbf{f}]_p^2 + \beta_y^2 [\mathbf{D}_y\mathbf{f}]_p^2 + \beta_z^2 [\mathbf{D}_z\mathbf{f}]_p^2} \tag{6}$$

where $[\mathbf{f}]_p$ denotes the pth component of the vector $\mathbf{f}$.

III. Proposed Method

The schematic diagram of this method is shown in Fig. 1, which mainly includes three key steps: dimension reduction, complementary feature extraction and probability fusion. The details are as follows:

A. Dimension Reduction

In order to reduce the execution time, we adopts the averaging fusion method to reduce the spectral dimension of HSI for the subsequent steps. In [41] proves that average dimension reduction is an effective tool in computational efficiency and protection of pixel reflection. Specifically, the original image $\mathbf{g}$ is firstly divided into B subsets according to spectral dimension. The reduction of the dimension of the data is then performed on each subset.

B. Complementary Spectral–Spatial Adaptive Method

Based on adaptive cubic total variation model [28] for feature extraction can smooth hyperspectral image well. However, it is easy to blur the edge profile. It is necessary to choose appropriate parameters to achieve a balance between maintaining the edge contour and smoothing the image. The classification effect obtained by using the ACTVSP method alone for feature extraction is largely dependent on the selection of parameters, and tends to blur details and edges. Therefore, the robustness is not strong. The solution of this paper is to use the edge-preserving filter [39] to supplement the probability, to compensate for the edge ambiguity in the feature extraction stage, so as to improve the classification accuracy. Besides, the fusion of the two methods will enhance the robustness of the overall method. Although there are parameters that need to be adjusted for edge-preserving filtering, for the two complementary stages, it is only necessary to select the parameters with good overall effect but no need for the best, and the ideal effect can be achieved after combination. Specifically, the method proposed in this paper is as follows:

1) Weight parameter

Assuming the dimension reduced HSI f is a $M \times N \times B$ matrix. It should be pointed out that, for the convenience of explanation, the f mentioned in the description of weight parameters is in matrix form, but the actual operation is to convert f into the corresponding vector $\mathbf{f}$ for processing. In [28], the author thinks that different pixels in the spatial dimension of HSI should be multiplied by different weights to ensure different smoothness of different pixels. But it ignored the fact that different bands of spectral dimension also need different degrees of processing. One of the important properties of HSIs compared to ordinary images is that HSIs have many different spectral bands. Therefore, we take this into account and modify the spatial dimension adaptive parameter proposed in [28] so that it can be applied to the adaptive processing of spectral dimension. The weight parameters set in this paper are different for each spectral band. We set the weight parameter matrix as $W \in \mathbb{R}^{M\times N\times B}$. For the same spectral band, the weight parameters of all pixels are the same. Setting W to the same size as HSI can facilitate the solution after W vectorization.

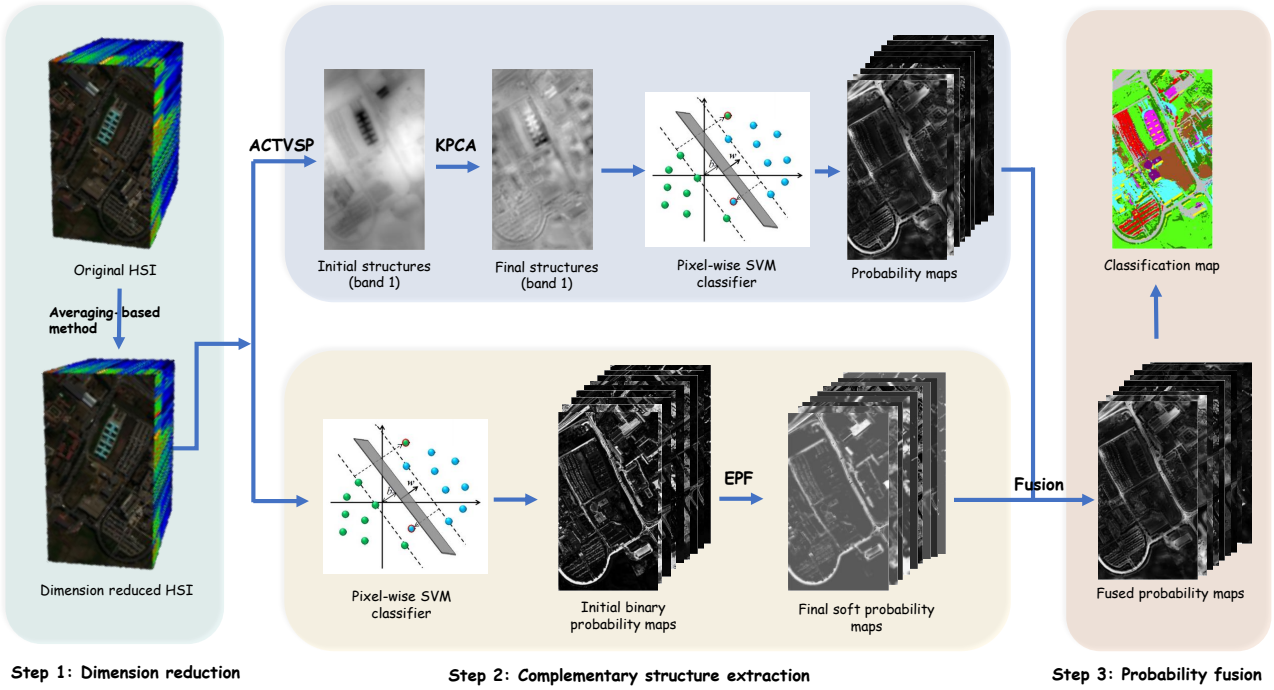


Fig. 1. The schematic diagram of the proposed method.

The weight parameters of all pixels in ith band are:

$$\tau_i = \frac{1}{1 + \alpha \sum_{MN} \|\mathbf{f}\|_{TV_{2,i}}} \tag{7}$$

where $\|\mathbf{f}\|_{TV_{2,i}}$ represents the cubic total variation norm of the element of vector $\mathbf{f}$ corresponding to the ith band, while $\sum_{MN} \|\mathbf{f}\|_{TV_{2,i}}$ represents the sum operation of the norm of all the elements of vector $\mathbf{f}$ corresponding to the ith band.

In Equation (7), to define the spectrally weighted parameter of each band, we use parameter α to control the contribution of each spectral band information to the adaptive parameter weight τ_i, the range of parameter τ is between [0, 1]. In addition, due to the small value of τ, the iterative solution process of the subsequent augmented Lagrange algorithm [42] is too fast, resulting in poor feature extraction effect. Therefore, a correction value is needed to balance the iteration rate. Through the experiment, the correction value of $\frac{NB}{M}$ is the best. When other values are taken, the subsequent classification accuracy is very low, so the correction value is not discussed in the parameter analysis part.

Therefore, the final weight parameters are as follows:

$$[W]_i = \frac{NB\tau_i}{M} \tag{8}$$

For the sake of distinction, $[W]_i$ does not refer to the value of the ith parameter, but to the value of the ith bands of the weight matrix W.

Consistent with the representation of $\mathbf{f}$, we use the bold letter $\mathbf{W}$ to represent the vectorized version of matrix W, i.e.,

$$\mathbf{W} = \mathbf{vec}(W(x, y, z)) \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{MNB\times 1}$. And W_i is the ith element of vector $\mathbf{W}$.

Such parameter definition is based on facts. First, the cubic total variation value of a band is positively correlated with the edge contour of the band. When the edge of a band is more obvious, the corresponding weight parameter is smaller, which can avoid serious distortion caused by excessive smoothing.

2) ACTVSP Model

In this work, an adaptive feature extraction smoothing model is proposed. The core optimization problem we solve is to minimize the following formula:

$$\arg\min_{\mathbf{f}} \quad \frac{1}{2}\|\mathbf{Hf} - \mathbf{g}\|_2^2 + \lambda ACTVSP(\mathbf{f}) \tag{10}$$

where λ is the regularization parameter, which controls the trade-off between the data fidelity and regularization item, and the regularization item is expressed as:

$$ACTVSP(\mathbf{f}) = \mathbf{W}^T \|\mathbf{f}\|_{TV_2} \tag{11}$$

For the convenience of description, refer to the treatment method in [42], we define $\|\mathbf{Df}\|_2 \overset{\text{def}}{=} \|\mathbf{f}\|_{TV_2}$. Even though they're not exactly equal, we did this to make the description of the algorithm clearer.

Thus, problem (10) can be written as:

$$\arg\min_{\mathbf{f}} \quad \frac{1}{2}\|\mathbf{Hf} - \mathbf{g}\|_2^2 + \lambda \mathbf{W}^T \|\mathbf{Df}\|_2 \tag{12}$$

To solve Problem (12), we introduce an intermediate variables $\mathbf{u}$ to replace $\mathbf{Df}$. In this way, problem (12) can be transformed into an equivalent problem:

$$\begin{aligned} &\arg\min_{\mathbf{f},\mathbf{u}} \quad \frac{1}{2}\|\mathbf{Hf} - \mathbf{g}\|_2^2 + \lambda \mathbf{W}^T \|\mathbf{u}\|_2 \\ &s.t. \quad \mathbf{u} = \mathbf{Df} \end{aligned} \tag{13}$$

Therefore, the augmented Lagrange function to solve the problem (13) is:

$$L(\mathbf{f},\mathbf{u},\mathbf{y}) = \frac{1}{2}\|\mathbf{Hf}-\mathbf{g}\|_2^2+\lambda\mathbf{W}^T\|\mathbf{u}\|_2-\mathbf{y}^T(\mathbf{u}-\mathbf{Df})+\frac{\rho}{2}\|\mathbf{u}-\mathbf{Df}\|_2^2 \tag{14}$$

where ρ is a fixed parameter representing the step size of gradient ascent, $\|\mathbf{u}-\mathbf{Df}\|_2^2$ is quadratic penalty term, and $\mathbf{y}$ is the Lagrange multiplier associated with the constraint $\mathbf{u}=\mathbf{Df}$.

Then, through continuous iteration, the saddle point of the Lagrange function $L(\mathbf{f},\mathbf{u},\mathbf{y})$ is finally found, that is, the solution of the original problem (10). Referring to [42] of video denoising solutions, we use the alternating direction method (ADM) [43], [44] to iteratively solve $\mathbf{f}$ and $\mathbf{u}$ subproblems iteratively, which results in the following updates:

Step 1 :Update $\mathbf{f}_{k+1}$

$$\mathbf{f}_{k+1} = \arg\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Hf}-\mathbf{g}\|_2^2-\mathbf{y}_k^T(\mathbf{u}_k-\mathbf{Df})+\frac{\rho}{2}\|\mathbf{u}_k-\mathbf{Df}\|_2^2 \tag{15}$$

By dropping the indices k, and take the partial with respect to $\mathbf{f}$, solution of the subproblem (15) is found by considering the normal equation:

$$\mathbf{H}^T\mathbf{Hf}-\mathbf{H}^Tg+\mathbf{D}^T\mathbf{y}-\rho\mathbf{D}^T\mathbf{u}+\rho\mathbf{D}^T\mathbf{Df}=0 \tag{16}$$

By arranging equation (16), we can get:

$$(\mathbf{H}^T\mathbf{H}+\rho\mathbf{D}^T\mathbf{D})\mathbf{f} = \mathbf{H}^T\mathbf{g}+\rho\mathbf{D}^T\mathbf{u}-\mathbf{D}^T\mathbf{y} \tag{17}$$

after the triple block-circulant convolution matrix $\mathbf{H}$ be diagonalized by the three-dimensional DFT matrixt, one solution of Equation (17) is:

$$\mathbf{f} = \mathscr{F}^{-1}\left[\frac{\mathscr{F}(\mathbf{H}^T\mathbf{g}+\rho\mathbf{D}^T\mathbf{u}-\mathbf{D}^T\mathbf{y})}{|\mathscr{F}[\mathbf{H}]|^2+\rho(|\mathscr{F}[\mathbf{D}_x]|^2+|\mathscr{F}[\mathbf{D}_y]|^2+|\mathscr{F}[\mathbf{D}_z]|^2)}\right] \tag{18}$$

where $\mathscr{F}$ denotes the three-dimensional Fourier Transform operator [45]–[47].

Step 2 :Update $\mathbf{u}_{k+1}$

$$u_{k+1} = \arg\min_{\mathbf{u}} \lambda\mathbf{W}^T\|\mathbf{u}\|_2-\mathbf{y}_k^T(\mathbf{u}-\mathbf{Df}_{k+1})+\frac{\rho}{2}\|\mathbf{u}-\mathbf{Df}_{k+1}\|_2^2 \tag{19}$$

The $\mathbf{u}$-subproblem can be solved using the shrinkage formula [48]. Letting $\mathbf{v}=\mathbf{Df}+\frac{\mathbf{y}}{\rho}$, $\mathbf{u}$ is given by the following formula:

$$\mathbf{u} = \max\{\ \|\mathbf{v}\|_2-\frac{\lambda\mathbf{W}^T}{\rho},0\}\ \cdot\frac{\lambda\mathbf{W}^T\mathbf{v}}{\|\mathbf{v}\|_2} \tag{20}$$

We can easily find that:

$$\mathbf{u}_x = \max\{\ \|\mathbf{v}\|_2-\frac{\lambda\mathbf{W}^T}{\rho},0\}\ \cdot\frac{\lambda\mathbf{W}^T\mathbf{v}_x}{\|\mathbf{v}\|_2} \tag{21}$$

where $\mathbf{v}_x=\beta_x\mathbf{D}_x\mathbf{f}+\frac{\mathbf{y}_x}{\rho}$, $\mathbf{v}_y$ and $\mathbf{v}_z$ can be derived in a similar way.

Step 3 :Update $\mathbf{y}_{k+1}$

$$\mathbf{y}_{k+1} = \mathbf{y}_k-\rho(\mathbf{u}_{k+1}-\mathbf{Df}_{k+1}) \tag{22}$$

Repeat these three steps until they converge at each pixel. From this, we can get the image $\mathbf{f}$ after feature extraction. To further enhance the separability of pixels between different species, the kernel principal component analysis (KPCA) [49] is applied on $\mathbf{f}$ obtained by the above steps, and K principal components are preserved for classification.

Then, the SVM classifier [8] is conducted to obtain the class probability $\mathbf{P}_1$.

3) Edge-preserving Feature Extraction

Through the first stage of image smoothing and feature extraction, the hyperspectral image obtained will improve the classification effect, but then face the problem that the image edge contour becomes blurred after texture smoothing. At the same time, because of too many parameters, it is difficult to debug the best parameter combination in practical application. Edge-preserving filtering [39] can make up for the above problems in the first stage and enhance the robustness of the overall algorithm while sharpening the edges of the contour to prevent excessive smoothing. To be more specific, the dimensionally reduced images are first fed into a spectral classifier SVM [8] to obtain multiple initial probability maps. For example, $\mathbf{p}=(\mathbf{p}_1,...,\mathbf{p}_n)$ represents one of the probability maps, where $\mathbf{p}_{i,n}\in[0,1]$ is the initial probability that pixel i belongs to the nth class.

Then edge-preserving filtering is used to optimize the obtained probability map and align the probability map with real object boundaries. The optimization probability is modeled as the weighted average of its neighborhood probability, that is:

$$\acute{\mathbf{p}}_{i,n} = \sum_j W_{i,j}(I)\mathbf{p}_{j,n} \tag{23}$$

where i and j represent the ith and jth pixel points respectively, and the filtering weight W is selected so that the filtering retains the edge of the specified guiding image I. The guided filteris used in this article, and the filtering weight $W_{i,j}(I)$ of the guided filter can be expressed as follows:

$$W_{i,j}(I) = \left(\frac{1}{|\omega|^2}\right)\sum_{k\in\omega_i,k\in\omega_j}\left(1+\frac{(I_i-m_k)(I_j-m_k)}{\sigma_k^2+\epsilon}\right) \tag{24}$$

where ω_i and ω_j are local windows around pixel i and j, respectively, m_k and σ_k are the mean and variance of I in ω_k, and $|\omega|$ is the number of pixels in ω_k. Gray-scale guidance image obtained after principal component analysis is used as guidance image I.

Finally, a classification based on maximum probability is used to obtain the final probability, according to the following Equation:

$$\acute{\mathbf{p}}_i = \arg\max_n \acute{\mathbf{p}}_{i,n} \tag{25}$$

The final $\acute{\mathbf{p}}$ is $\mathbf{P}_2$ of the second stage probability map we need.

4) Probability Fusion

The class probability obtained through the first stage may cause the edges to be blurred compared with the original image due to excessive smoothing of the image. In the second stage, the class probability is obtained by using edge-preserving filtering [39] algorithm, and the edge is relatively clear. In this paper, weighted decision fusion rule is used to merge the class probability obtained in the two stages, so as to obtain the final class label. In this way, the classification probability

of the same target is not only higher than that of other types of land cover, but also the edge will be very clear, which can more accurately divide the area covered by different ground objects. In terms of pixels, the final label for each pixel is determined based on maximum probability:

$$\mathbf{P} = \arg\max_i \{\ \mu \mathbf{P}_1^i + (1-\mu)\mathbf{P}_2^i\} \tag{26}$$

where μ is a free parameter, and $\mathbf{P}_1^i$ is the class probability map obtained by the classification of ACTVSP. $\mathbf{P}_2^i$ denotes the class probability map of classification of EPF. $\mathbf{P}$ is the final classification result.

The complete and detailed algorithm is given in Algorithm 1 for easy understanding.

Algorithm 1 Algorithm process:

Input observed HSI $\mathbf{g}$
step1.**ACTVSP**
Initialize $\mathbf{f_0} = \mathbf{g}, \mathbf{u_0} = \mathbf{Df_0}, \mathbf{y} = 0, k = 0$
while $\|\mathbf{f}_{k+1} - \mathbf{f}_k\| >$ tol do
  Calculate $\mathbf{W}$.
  Solve the f-subproblem using Equation (18).
  Solve the u-subproblem using Equation (20).
  Update the Lagrange multiplier y using Equation (22).
end
  Get probability map $\mathbf{P_1}$ via SVM.
step2.**EPF**
  Obtain multiple initial probability maps.
  Optimize and align the obtained probability map by Equation (23).
  Get probability map $\mathbf{P_2}$ using Equation (25).
step3.**Probability Fusion**
  Get final probability map $\mathbf{P}$ using Equation (26).

C. Parameter Optimization and Selection of SVM Model

In this paper, SVM is selected as the classifier for both pre-processing method ACTVSP and post-processing method EPF.

For the SVM with Gaussian kernel function, there are two main training parameters: C is the penalty coefficient, which controls over-fitting of the model, and gamma (γ) is a parameter that comes with the kernal function, which controls the degree of nonlinearity of the model.

For the determination of C and gamma parameters, we use the cross-validation method and the grid search algorithm as the parameter optimization method. The C and gamma parameters are combined according to step size 1 in a certain range of values (The certain value range is pointed out in Section IV-B-(2)). Then, under different (C, gamma) combinations, the training set samples are divided into n groups. One group is used as validation data samples, and the other $n-1$ groups are used as training data. Each set of data is taken in turn as a verification sample, so that under a combination of (C, Gamma) , n calculations are required. The n-times model accuracy was averaged as a score for the models in this group (C, Gamma).

In this way, we can get the score of the model under different (C, gamma) combinations. We choose the group with the highest score (C, gamma) as the final parameter value of SVM model. If there are different combinations have the same score, we first consider the parameter C, the smaller value is prior. Bacause under the condition that the accuracy of the model is guaranteed, the smaller the value of C is, the higher the fault-tolerance of the model is. Thus can avoid over-fitting.

The above parameter optimization process can make the parameters of SVM model change dynamically with the training set, and avoid the tedious manual adjustment process.

IV. Experiments

A. Data sets

To verify the validity of our method, we use a classical data set, i.e., University of Pavia, a data set used for hyperspectral change detection in the past, i.e., the China Data set, and a recent data set, i.e., Houston 2013.

1) University of Pavia: The University of Pavia image was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS-03) optical sensor with a spectral coverage ranging from 0.43 to 0.86 μm, which covers the campus at the University of Pavia, Pavia, Italy. This image contains 103 bands of size 610 $\times$ 340 pixels with a spatial resolution 1.3 m per pixel. Fig. 2 shows the false color composite image, training and test samples, ground truth. The number of training and test samples is shown in Table VI.

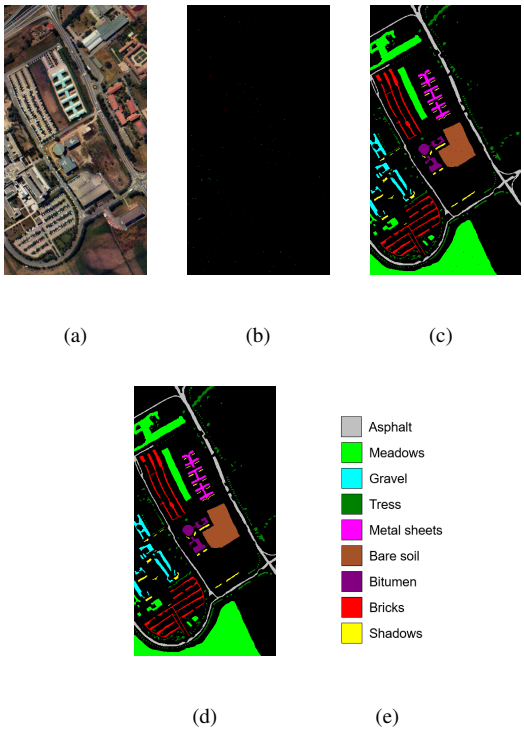


Fig. 2. University of Pavia data set. (a) False color composite image. (b) Training samples. (c) Testing samples. (d) Ground truth. (e) Class names.

2) The China Data set: The China data set belongs to a farmland near the city of Yuncheng Jiangsu province in China,

which was captured for hyperspectral change detection on May 3, 2006, and April 23, 2007 [50]. This scene is mainly a combination of soil, river, tree, building, road and agricultural field. For this data set, all changes related to the type of land cover and river. And the ground truth provided by the author was divided into four main categories. In addition, this data set belongs to Hyperion sensors. This image contains 154 bands of size 420 × 140 pixels. Fig. 3 shows the false color composite image, training and test samples, ground truth. The number of training and test samples is shown in Table VII.

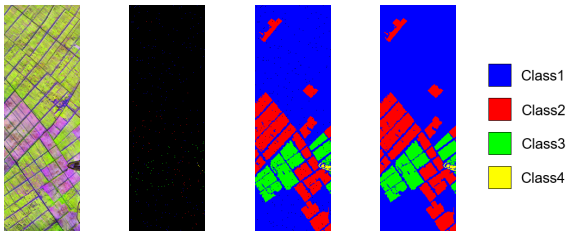


Fig. 3. The China data set. (a) False color composite image. (b) Training samples. (c) Testing samples. (d) Ground truth. (e) Class names.

3) Houston 2013: The Houston 2013 image was distributed by the 2013 GRSS Data Fusion Contest. This image was acquired by the compact airborne spectrographic imager over the University of Houston campus and the neighboring urban area on June 23, 2012. Furthermore, it contains 144 bands of size 349 × 1905 pixels with a spatial resolution of 2.5 m, and the spectral coverage ranges from 0.38 to 1.05 μm. Fig. 4 shows the false color composite image, training and test samples, ground truth. The number of training and test samples is shown in Table VIII.

B. Experimental Setup

1) Evaluation Indexes: In order to quantitatively evaluate the classification performance of all research methods, four popular and widely used quantitative indicators, i.e., individual class accuracy (CA), overall accuracy (OA), average accuracy (AA), and Kappa coefficient, were used to evaluate the classification performance of HSI. Where CA represents the percentage of correctly classified pixels for each class. OA represents the percentage of correctly classified pixels in the total number of pixels. AA measures the average of the percentage of pixels correctly identified for each land cover. The Kappa coefficient calculates the percentage of identified pixels corrected by the number of agreements that would be expected purely by chance. All experiments were randomly selected and repeated for 10 times to obtain the mean and standard deviation of CA, OA, AA and Kappa coefficients. The final result is obtained by taking the average value of the four evaluation indexes obtained from the above experiments.

2) Competitive approaches: In this paper, the following seven state-of-art classification methods are used for comparison, including four machine learning methods and three deep learning methods:

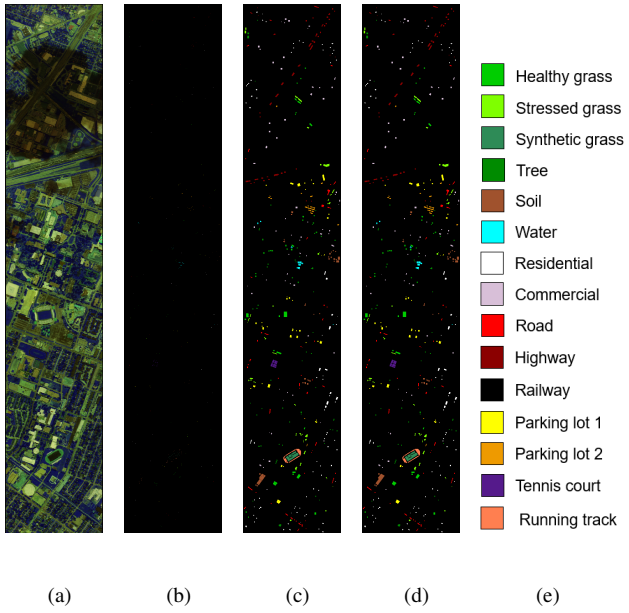


Fig. 4. Houston 2013 data set. (a) False color composite image. (b) Training samples. (c) Testing samples. (d) Ground truth. (e) Class names.

Support vector machine (**SVM**) [8] is performed using the LIBSVM library, where the Gaussian kernel is adopted. Among them, the optimal parameters selection of kernel adopts five-fold cross-validation. The penalty factor is from 10^{-2} to 10^{4}, and the kernel width varies between $2^{-4} \sim 2^{4}$.

Weighted markov random field (**WMRF**) [26] uses the total variational regularization model to smooth the a posteriori distribution to describe the spatial information in the hidden domain.

Random patches network (**RPNet**) [51] is fed into a spectral classifier using random patch as a convolution kernel by learning the deep convolution features of HSI.

Structural profile (**SP**) [38] proposes an adaptive texture smoothing method to construct the structural contour of hyperspectral images and a hyperspectral image classification framework based on dual spatial information fusion.

Random multigraphs ensemble learning (**RMGE**) [52] propose a novel graph-based semi-supervised learning classification model for hyperspectral images.

Spectralformer [53] rethink hyperspectral image classification from a sequential perspective with transformers, and propose a novel backbone network called SpectralFormer for hyperspectral image classification.

Weighted feature fusion (**WFCG**) [54] propose a weighted feature fusion framework called WFCG for HSI classification, by using the characteristics of superpixel-based GAT and pixel-based CNN, which proved to be complementary.

All comparison methods used the default parameter settings given in the corresponding literature for comparison.

C. Analysis of Parameters

In the proposed method, there are six parameters need to be fixed, i.e., the number of the dimension reduced data B, the

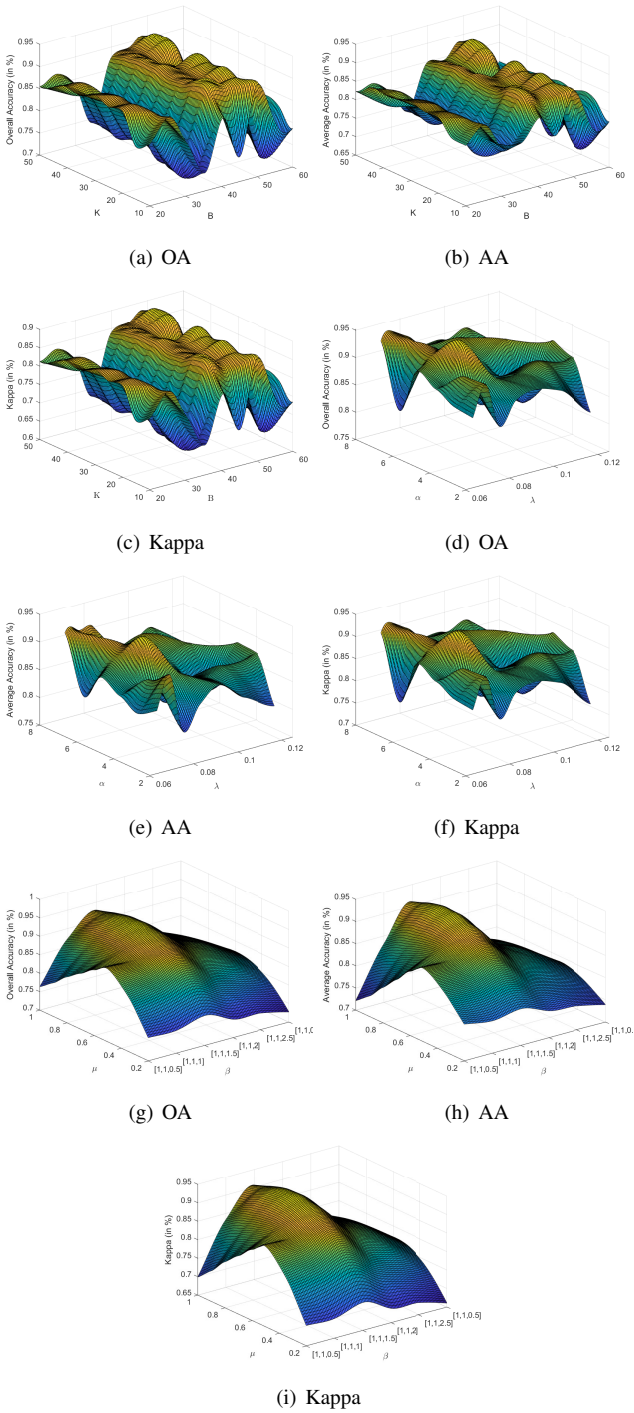
constants $(\beta_x, \beta_y, \beta_z)$, the number of kernel principle components K used for classification, the regularization parameter λ, the adaptive parameter contribution coefficient α, and the fusion weight μ. An experiment is performed on the University of Pavia data set to initialize the aforementioned parameters. When the parameters B and K are analyzed, $(\beta_x, \beta_y, \beta_z)$, λ, α and μ are set to be (1,1,1), 0.083, 6 and 0.7, respectively. Fig. 5(a) to 5(c) shows the influence of different values of B and K on OA, AA, Kappa and computing time. It can be concluded that, in general, when B and K are smaller, the classification accuracy of our method tends to decline and the calculation time is shorter. Moreover, when B and K are set as 50 and 45, respectively, OA and AA of the method reach the best classification accuracy, and Kappa coefficient is also large.

Besides, based on the most suitable values of B and K obtained from the above experiments (i.e., B and K are set as 50 and 45 respectively), Fig. 5(d) to 5(f) shows the influence of parameters λ and α. Obviously, when α is equal to 5 and λ is around 0.083, OA, AA and Kappa can achieve the best results. Fig. 5(g) to 5(i) shows the influence of parameters μ and β when other parameters are set to the most appropriate value obtained by experiments (B, K, λ and α are set as 50, 45, 0.083 and 5 respectively). It can be concluded that, when β is equal to [1,1,1] and μ is equal to 0.7, the three evaluation indexes achieved the best effect. Subsequent experiments based on optimal parameters verify that the proposed method can achieve good classification performance on all data sets.

D. Analysis of the Effects of Different Data Dimension

First of all, the purpose of dimensionality reduction for original data is to reduce run time and redundant information. Table I lists the classification effects of data with original data and different dimensionality reduction values on the University of Pavia data set by using the method proposed in this paper. The number of training samples is 10 pixels randomly selected from each class. As can be seen from Table I, for the data with appropriate values for dimensionality reduction, both the overall accuracy of classification and the running time are significantly improved. For example, on the University of Pavia data set, the best classification effect can be achieved by sending the data after dimensionality reduction to 50 bands into subsequent work. However, it does not mean that the higher the dimension reduction intensity is, the better the effect of improving the classification accuracy will be. Excessive dimension reduction will lead to the neglect of useful information and the increase of running time. The insufficient dimension reduction will reduce the running time, but it will cause a lot of redundant information to be retained. In both cases, the classification accuracy will decrease. In general, due to the large amount of useless information in hyperspectral images, dimensionality reduction by selecting appropriate values of original data at the beginning can improve the classification effect.

Fig. 6 compares the probability map obtained by using the original data and the data with a dimensionality reduction value of 50. It can be clearly seen that the classification probability map obtained by using the dimensionality reduction data is very similar, and even clearer in detail.



(a) OA (b) AA (c) Kappa (d) OA (e) AA (f) Kappa (g) OA (h) AA (i) Kappa

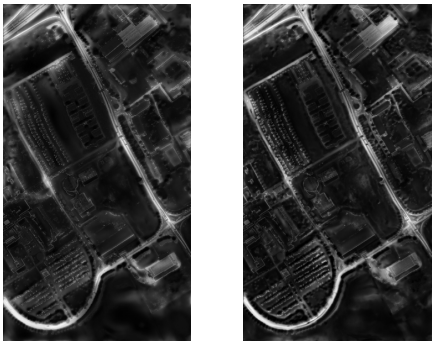
Fig. 5. The influence of the parameters, i.e, B, K, λ, α, μ, and $(\beta_x, \beta_y, \beta_z)$, on the classification performance of the proposed method.

E. Ablation Experiment

The two main components of the proposed method are feature extraction by ACTVSP model and edge-preserving filtering [39]. The first stage can be divided into ACTVSP feature extraction and dimension reduction using KPCA [49]. Experiments were conducted on the University of Pavia data set using the control variable method, where the number of training samples was 10 pixels randomly selected from each

TABLE I
CLASSIFICATION EFFECT OF DIFFERENT DATA DIMENSION.

Data type	Raw data	Dimension reduced data					
		20	30	40	50	60	70
OA (in %)	88.31	84.23	88.36	90.65	**95.07**	81.23	81.97
AA (in %)	84.14	80.72	83.91	85.66	**92.91**	81.38	77.11
Kappa (in %)	84.88	79.83	84.81	87.90	**93.52**	76.04	77.16
time (in s)	149.21	**40.58**	50.28	60.61	86.76	98.37	117.62



(a) (b)

Fig. 6. Probability maps obtained by the classification of the University of Pavia data set. (a) Using original image classification. (b) Using dimension reduced image.

class. First, we analyzed the influence of different dimension reduction method for classification effect. It should be noted that we used the same best combination of parameters for all the experiments which was acquired in Section IV-C. Table II shows the classification performance obtained by combining several different dimension reduction methods with other components. Specifically, we selected six widely used the method for comparison, i.e., principal component analysis (PCA) [55], linear discriminant analysis (LDA) [56], probabilistic PCA (ProbPCA) [57], factor analysis (FA) [58], classical multidimensional scaling (MDS) [59] and kernel PCA (KPCA) [49]. It can be seen from the Table II that KPCA method combined with other components can obtain the highest classification accuracy. The reason is that KPCA projects the data of the original space into the high-dimensional feature space through kernal mapping for processing.

Furthermore, the results in Table III demonstrate the validity of other different components. First of all, it can be seen from Table III that only using one of EPF, ACTVSP and KPCA for feature extraction, or even the fusion method of ACTVSP and EPF for classification results are unsatisfactory. In addition, it can be observed that classification result acquired by the combination of KPCA and EPF are very poor, which can prove that KPCA does not play a key role in the final classification results. After image smoothing and feature extraction by ACTVSP and dimension reduction by KPCA [49], the final classification effect is greatly improved. On this basis, the classification effect obtained by probability combination with EPF is the best in the experimental method. This can be explained. To illustrate visually, Fig. 7 shows the separability of different categories of pixels after combining them using different methods.

TABLE II
CLASSIFICATION EFFECT ON THE UNIVERSITY OF PAVIA DATA SET WITH DIFFERENT DIMENSION REDUCTION METHODS.

	PCA	LDA	ProbPCA	FA	MDS	KPCA
OA (in %)	90.75	86.13	88.89	63.45	92.57	**95.07**
AA (in %)	86.51	79.19	86.14	69.54	92.14	**92.91**
Kappa (in %)	87.86	81.76	85.73	55.01	90.03	**93.52**
time (in s)	70.51	70.47	317.45	114.11	**68.85**	72.22

As can be seen from Fig. 7, the separability of the image processed by ACTVSP is improved, because of pixels of the same kind are gathered together, and then KPCA processing [49] is performed to separate pixels of different kinds more widely, the separability of pixels is further improved, thus improving the classification efficiency.

To show more intuitively that our methods can improve separability between categories, we choose several widely used baseline methods [36] in the hyperspectral community for comparison, including Gabor [13], LPP [60], LDA [56], CGDA [61], JPlay [62] and SP [38]. The experiment was performed on the University of Pavia data set. For the sake of convenience, we set the number of spectral dimensions after the feature extraction to 45. Fig. 8 shows the scatter diagrams of the first three bands, which can represent class separation capability of features extracted by different methods. As we all know, the higher the degree of separation between classes, the better the classification result obtained by the classifier.

It can be observed from Fig. 8, compared with other single feature extraction methods, SP and our method can effectively reduce intra-class differences and increase inter-class differences. Because these two methods smooth the hyperspectral image by adding the variational regularization term, and retain the main features. However, our method further widens the inter-class gap and further reduces the intra-class gap compared with SP. This can be explained because our method adds smoothing to the spectral segment during the pre-processing stage, which removes as much redundant information as possible and makes the main features clearer.

F. Description of Robustness and Smoothing Effect

To prove the robustness and generalizability of our method, we conducted experiments on ten public data sets, including Indian Pines 1992 (220 bands version), the University of Pavia, Pavia Center Part1, Pavia Center Part2, Salinas, the Kennedy Space Center, the China data set, Indian Pines 2010, Botswana and Houston 2013 data set. All data sets were randomly selected from 20 training samples in each category and the rest as test samples. In addition, all data sets were experimented using the same set of fixed parameters (B, K, λ, α, μ and $(\beta_x, \beta_y, \beta_z)$ equal to 50, 45, 0.083, 5, 12 and [1,1,1] respectively). Among all data sets, some differ greatly in spatial size, for example, Indian Pines 1992 contains 145 $\times$ 145 pixels, while Houstone contains 349 $\times$ 1905 pixels. Some differ greatly in spectral dimensions, for example, Indian

TABLE III
CLASSIFICATION EFFECT OF DIFFERENT COMPONENTS.

	EPF	ACTVSP	KPCA	KPCA+EPF	ACTVSP+KPCA	ACTVSP+EPF	ACTVSP+KPCA+EPF
OA (in %)	66.35	65.45	46.18	70.88	92.86	74.03	**95.07**
AA (in %)	67.76	68.69	53.86	76.89	90.92	69.02	**92.91**
Kappa (in %)	58.62	57.77	36.32	62.75	90.60	67.88	**93.52**
time (in s)	**0.40**	72.64	5.51	0.41	70.69	73.53	72.22

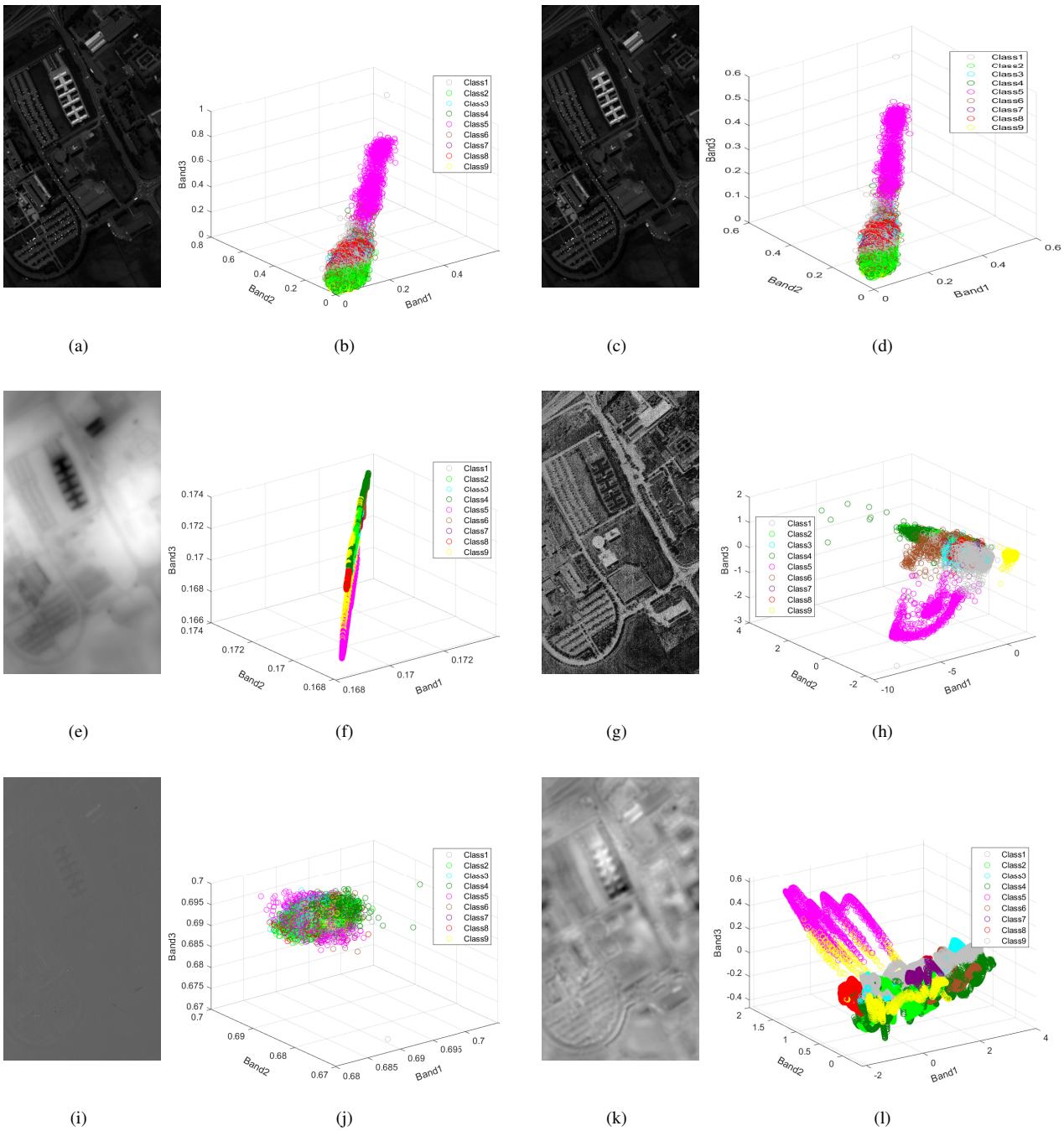


Fig. 7. Influence of different components on pixel separability of the University of Pavia data set. (a) Dimensionality reduced image. (b) Pixel separability of the first three bands about(a) (different symbols denotes pixels belonging to different classes). (c) EPF processed image. (d) Pixel separability after EPF. (e) Images extracted by ACTVSP. (f) Pixel separability after ACTVSP. (g) KPCA processed image. (h) Pixel separability after KPCA. (i) KPCA+EPF processed image. (j) Pixel separability after KPCA+EPF. (k) ACTVSP+KPCA processed image. (l) Pixel separability after ACTVSP+KPCA.

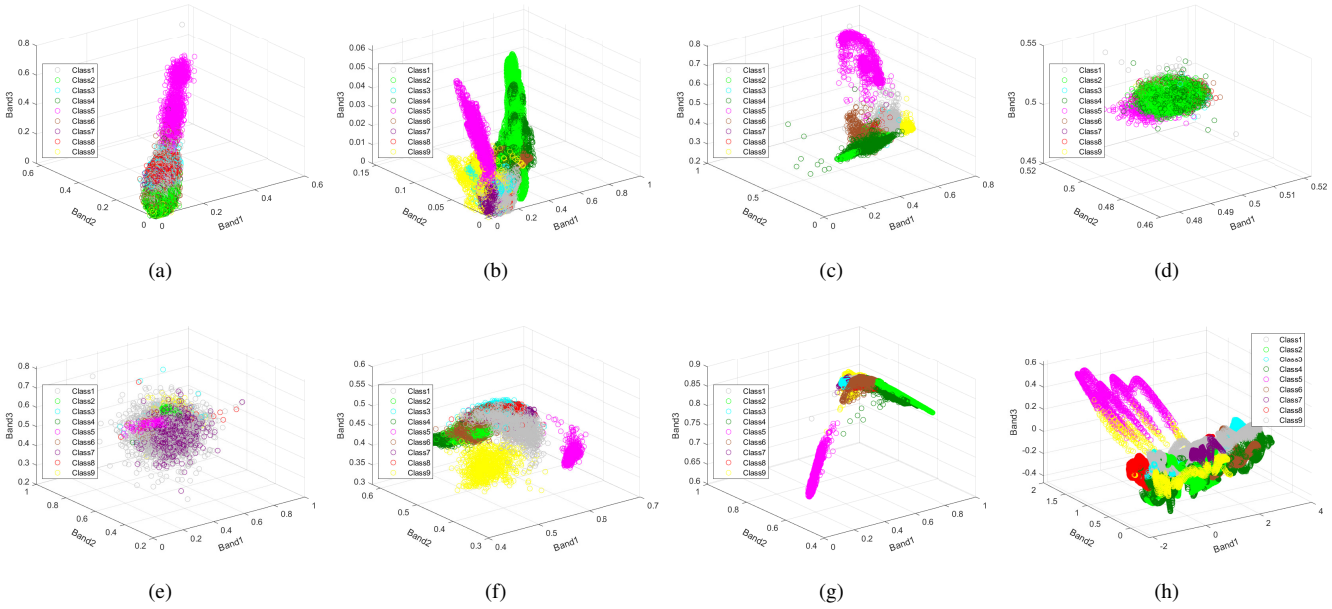


Fig. 8. The pixel separability of baseline methods. (a) The original HSI. (b) Gabor [13]. (c) LPP [60]. (d) LDA [56]. (e) CGDA [61]. (f) JPlay [62]. (g) SP [38]. (h) Our method.

TABLE IV
CLASSIFICATION PERFORMANCE ON TEN PUBLIC DATA SETS.

	Indian Pines 1992	Pavia University	Pavia Center Part1	Pavia Center Part2	Salinas	KSC	China	Indian Pines 2010	Botswana	Houston 2013
OA (in %)	91.08	96.29	97.46	97.46	97.92	99.07	85.46	96.30	99.19	85.84
AA (in %)	87.40	94.40	92.33	92.33	98.32	98.81	68.66	95.09	99.30	86.78
Kappa (in %)	89.82	95.12	96.42	96.42	97.68	98.96	70.88	95.46	99.12	84.70
time (in s)	13.57	88.64	279.08	284.88	37.07	122.26	14.63	137.58	116.32	225.09

Pines 1992 contains 220 bands, while the University of Pavia contains 103 bands. In addition, the proportion of different classes in almost all data sets also varies greatly. For example, the number of four categories in the China data set are 40407, 12229, 6019 and 145. Therefore, taking 20 random points as training samples for each class will bring great challenges to classification. Table IV shows the classification results. It can be observed that, despite many of the problems mentioned above, our approach achieves over 85% overall accuracies on all experimental data sets, and over 90% on eight data sets. Although there are six parameters listed in Section IV-C that need to be adjusted, but the purpose of adjustment is to achieve the best classification effect. In practice, we do not need to spend a lot of energy on parameter selection, but only need to select a set of fixed universally applicable parameters can achieve better results. To get the best results, we just need to make a few minor adjustments. This is why we say that the proposed method is of robustness.

Moreover, the objective evaluation of smoothing effect on band1, band2, band3 and band4 are shown in Table V. At the same time, for comparison, we also quantitatively evaluate the smoothing effect of SP method on the above four bands. It can be concluded that the proposed method does have a certain smoothing effect with regard to quantitative measurements aspect. Although our method uses an adaptive smoothing strategy for spectral dimensions, the SP method smoothes HSIs in spatial dimensions, but the quantitative evaluation shows that the smoothing effect is similar for HSIs. This further illustrates that rich spectral bands are a very important property of HSIs. In recent years, many methods of hyperspectral image classification focus on the processing of spatial information rather than making full use of spectral information. Our approach is to regress to the essence and to introduce spatial information on the basis of spectral information. It turns out that this can also obtain a good feature extraction effect, and then improve the classification effect. In addition, it can be seen from Fig. 9(d), 10(d) and 11(d) that the classification map obtained by SP has over-fitting phenomenon, which is caused by the excessive smoothing of HSIs in the pre-processing stage. Therefore, although our method and SP method are very close to each other in the evaluation index of denoising, SP method will cause more information loss.

TABLE V
OBJECTIVE EVALUATION OF SMOOTHING EFFECT ON DIFFERENT BANDS.

Method	Evaluation Index	Band1	Band2	Band3	Band4
Our method	MSE	1.831026	**0.080332**	**0.047911**	0.038629
	PSNR	45.50386	**59.08193**	**61.32643**	62.26171
SP	MSE	**1.695135**	0.221813	0.148387	**0.026773**
	PSNR	**45.83876**	54.67094	56.41684	**63.85387**

G. Classification Results

1) University of Pavia:

The first experiment was conducted on the University of Pavia data set. Among them, 10 pixels are randomly selected from each class for the training samples, accounting for 0.2% of ground truth. In all contrast methods, SVM [8], WMRF [26], SP [38] and RMGE [52] are traditional machine learning methods. RPNet [51], Spectralformer [53] and WFCG [54] are deep learning methods.

It can be clearly seen from Fig. 9 that many noise labels are misclassified in the visual ground object classification map of SVM [8]. The reason is that SVM does not process original HSI and only uses spectral information while ignoring spatial information, which leads to unsatisfactory overall accuracy of classification. The WMRF method [26] uses the spatial adaptive total variational regularization method to make the classification more smooth in space by smoothing HSI images, so as to improve the classification accuracy. However, there are many mislabeled pixels in this classification map, because when the number of training samples is limited, the classification map obtained by WMRF method will appear excessive smoothness. RPNet [51], SP [38] use the dual information of spatial and spectral to reduce the influence of noise on classification results. Compared with SVM , RPNet and SP have better classification effect. RMGE [52] also uses spatial and spectral information. Even though it retains the fine structure of the local region by modeling multiple anchor graphs, this semi-supervised approach of representing the whole by parts is not as good as methods which deal with each spectral segment. Besides, due to limited training samples, Spectralformer [53] have large areas of mislabeled pixels. Among the three deep methods, WFCG [54] performs best, because it is a feature fusion method that use the characteristics of superpixel-based GAT and pixel-based CNN. However, compared with our method, it needs more training samples, so the classification effect is slightly worse than our method. SP method has better robustness compared with other methods due to its fusion framework. However, due to the large number of parameters, unreasonable parameter setting will lead to over-fitting phenomenon, small samples will be ignored, and excessive image smoothing will easily lead to misclassification. For example, in Fig. 9(d), small targets within the main distribution areas of meadows, metal sheets and bricks are almost all smoothed out. In contrast, the proposed method combines spectral and spatial information, draws on the noise robustness characteristics of SP framework, and proposes a complementary method, which can extract objects while preserving object edges. Compared with other methods, the classification accuracy is higher when the ground object scene is complicated, the edge is not obvious and the distribution is scattered.

Quantitative results obtained by different research methods are shown in Table VI. As can be seen from the table, our method has the highest classification accuracy in OA, AA and Kappa coefficients. In addition, although only three classes of our method have the highest CA, six of the nine classes have CA of more than 95%, and seven of them have CA of more than 90%. This shows that compared with other comparison methods, the method presented in this paper is widely applicable to the classification of most different types of land types on the University of Pavia data set.

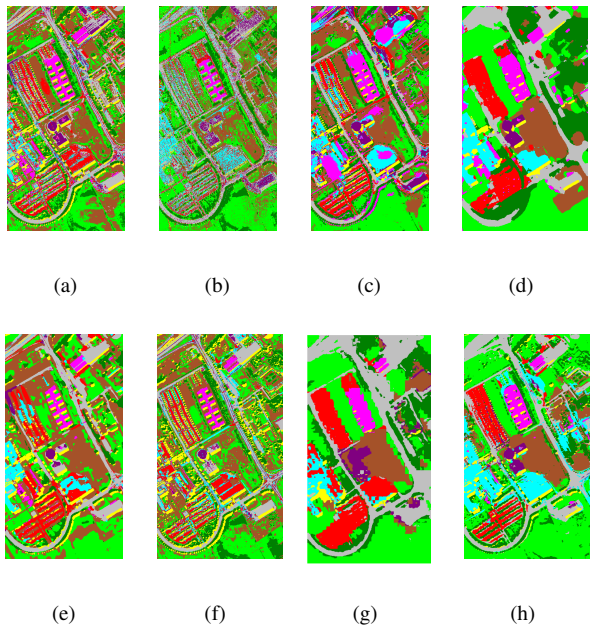


Fig. 9. Classification maps of all studied approaches on the University of Pavia data set. (a) SVM [8], (b) WMRF [26], (c) RPNet [51], (d) SP [38], (e) RMGE [52], (f) Spectralformer [53], (g) WFCG [54] and (h) Our method.

2) The China Data set:

The second experiment was conducted on the China Data set. Since the total number of samples of different categories is too large, we randomly select 200, 70, 50, 5 labeled points as training samples for four categories, accounting for 0.49%, 0.57%, 0.84% and 3.6% of ground truth respectively. In addition to class 4, the inter-class similarities of the other three classes are very strong. The classification maps of all methods are shown in Fig. 10.

Through observation, due to the small number of training samples and the small proportion in the image of class 4, the classification effect of all the methods on class 4 is the worst. Among them, the three deep methods (RPNet [51], Spectralformer [53] and WFCG [54]) for class 4 misclassification phenomenon is more obvious, WFCG method even completely misclassified class 4. One important reason is that deep methods generally require more training samples than traditional machine learning methods. Besides, it can be concluded that since SVM [8] only uses spectral information as the classification basis, the obtained classification map is also greatly affected by noise, and the misclassification phenomenon is serious. The same goes for WMRF [26]. Different from the excessive smoothness of the University of Pavia data set, WMRF has the problem of insufficient smoothness on the China data set, which leads to the phenomenon of misclassification of noise labels. Since class 1 is similar to class 2, all the methods except our method and RPNet did not classify class 2 in the upper left corner correctly, but

TABLE VI
CLASSIFICATION PERFORMANCE OF ALL STUDIED APPROACHES FOR THE UNIVERSITY OF PAVIA DATA SET, INCLUDING SVM [8], WMRF [26], RPNET [51], SP [38], RMGE [52], SPECTRALFORMER [53], WFCG [54] AND OUR METHOD.

Class name	Training set	Testing set	Classification accuracies of different approaches (in %)							
			SVM	WMRF	RPNet	SP	RMGE	Spectralformer	WFCG	Our method
Asphalt	10	6621	88.07	67.54	97.16	**97.27**	65.81	72.56	95.92	96.73
Meadows	10	18639	91.34	76.97	97.07	97.58	93.03	50.43	**99.10**	98.99
Gravel	10	2089	55.82	68.60	64.72	**99.14**	68.70	36.38	93.03	77.92
Tress	10	3054	63.85	90.93	94.04	76.24	28.82	94.07	**99.09**	76.68
Metal sheets	10	1335	90.77	99.33	63.04	79.13	**99.93**	99.10	89.77	97.23
Bare soil	10	5019	32.21	62.74	47.49	99.27	66.10	69.08	99.76	**99.78**
Bitumen	10	1320	48.44	68.48	57.94	**100.00**	73.46	46.52	86.49	91.32
Bricks	10	3672	63.15	55.07	82.42	82.99	44.98	77.26	71.49	**97.61**
Shadows	10	937	99.89	43.54	99.43	99.64	79.94	**100.00**	99.81	99.89
	OA		64.41	72.24	79.66	94.22	75.03	63.29	94.57	**95.07**
	AA		70.39	70.35	78.15	92.36	68.97	71.71	92.72	**92.91**
	Kappa		57.10	64.39	74.37	92.34	65.42	55.01	92.77	**93.52**

TABLE VII
CLASSIFICATION PERFORMANCE OF ALL STUDIED APPROACHES FOR THE CHINA DATA SET, INCLUDING SVM [8], WMRF [26], RPNET [51], SP [38], RMGE [52], SPECTRALFORMER [53], WFCG [54] AND OUR METHOD.

Class name	Training set	Testing set	Classification accuracies of different approaches (in %)							
			SVM	WMRF	RPNet	SP	RMGE	Spectralformer	WFCG	Our method
class1	200	40207	88.45	90.67	91.77	92.02	92.60	89.98	90.22	**93.24**
class2	70	12159	86.40	86.49	86.95	88.02	84.59	80.64	90.94	**93.40**
class3	50	5969	68.31	68.40	71.30	91.50	73.70	46.51	**94.60**	82.66
class4	5	140	13.33	57.14	51.61	56.04	**79.31**	46.51	0.00	61.74
	OA		85.35	84.37	88.78	91.01	88.96	83.51	90.65	**92.14**
	AA		69.54	75.68	75.41	81.90	82.55	67.14	68.94	**82.76**
	Kappa		67.55	67.51	76.04	80.42	76.53	64.66	78.96	**83.10**

misclassified class 2 in the upper left corner into class 1. However, because of the similarity between class 1 and class 3, and the interference of noise, RPNet, as a deep method, produced serious misclassification when the training samples are limited.

It can be seen from the classification maps that SP [38] and WFCG have good classification effects. An important reason is that these methods all make use of the dual information of spectral and spatial. Among them, the classification map of SP has the phenomenon of over-fitting, which is because the feature extraction process of SP is a denoising process in essence, leads to over-smoothing and insufficient extraction of edges and details. Compared to the University of Pavia data set, the China data set has a high degree of inter-class similarity, so it's a big challenge to distinguish between different classes. Meanwhile, due to the increase in the proportion of different classes, it is easy to ignore the relatively small area. Therefore, the extraction of edges and details is a problem that can not forget. From the classification map, we can see that our method has excellent performance in distinguishing similar classes and preserving edges and details.

Quantitative results obtained by different research methods are shown in Table VII. As can be seen from the table, our method also has the highest classification accuracy in OA, AA and Kappa coefficients. This also confirms the effectiveness of the method.

3) Houston 2013:

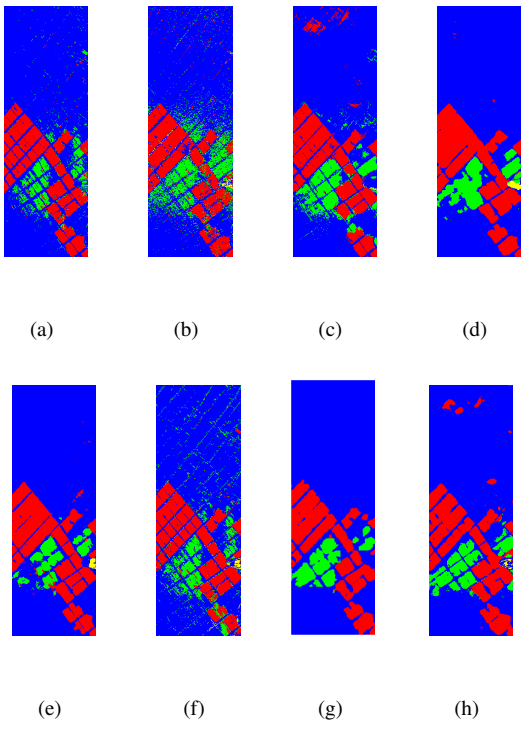
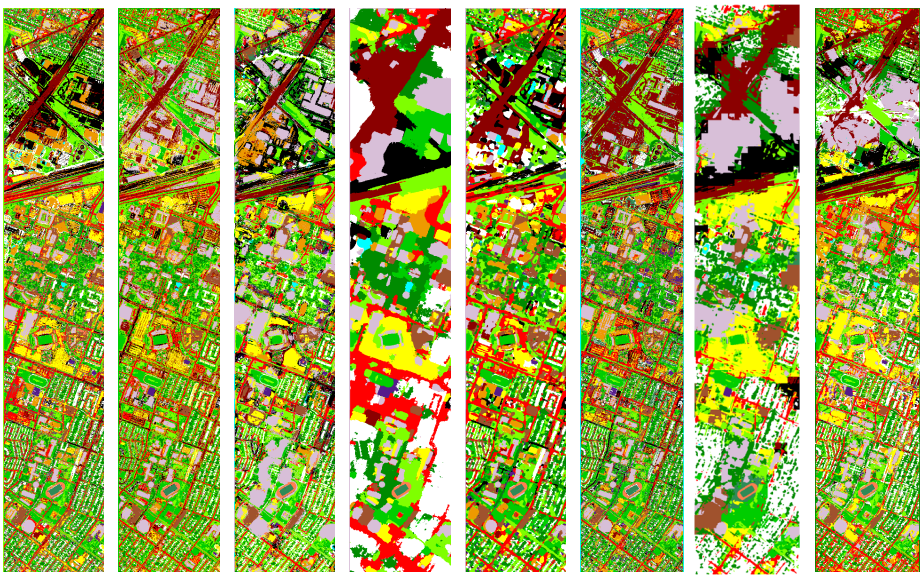


Fig. 10. Classification maps of all studied approaches on the China data set. (a) SVM [8], (b) WMRF [26], (c) RPNet [51], (d) SP [38], (e) RMGE [52], (f) Spectralformer [53], (g) WFCG [54] and (h) Our method.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 11. Classification maps of all studied approaches on the Houston 2013 data set. (a) SVM [8], (b) WMRF [26], (c) RPNet [51], (d) SP [38], (e) RMGE [52], (f) Spectralformer [53], (g) WFCG [54] and (h) Our method.

The third experiment was conducted on Houston 2013 data set, and 20 pixels were randomly selected from each class as training samples, accounting for 2% of ground truth. This data set is more challenging than the first two. The classification maps of all methods are shown in Fig. 11. Just like the previous two data sets, SVM [8] has a serious misclassification of noise labels. The WMRF [26] method with spatial constraint has little effect on improving noise phenomenon.

Three deep methods, i.e., RPNet [51], Spectralformer [53] and WFCG [54], still do not work well in the condition of limited training samples. Besides, the classification effect of RMGE [52] is even worse than that of SVM, which shows that the semi-supervised method may not be very suitable for challenging data sets. In addition, SP [38] and our method are more stable than the above methods, but the classification results of SP method appear over-fitting phenomenon, proving that our method performs better in details such as small-size targets.

Quantitative results obtained by different research methods are shown in Table VIII. It can be seen from the table that the accuracy of the individuals obtained by our method is above 85% in 14 out of 15 kinds of ground objects. Meanwhile, the classification effect of the four similar categories of healthy grass, Stressed grass, synthetic grass and tree is better. Three of the classes had an individual accuracy of around 98% and one class had an individual accuracy of around 90%. Our method also has the highest OA, AA and Kappa. The above analysis can prove that our method also has a good classification effect for challenging data sets.

H. Computing Cost

In this work, all experiments were carried out with Matlab R2020B on a laptop computer with Intel(R) Core(TM) I7-10875H CPU processor (2.30 GHz), 16GB memory and 64-bit operating system. The running times of all research methods on the three data sets used are shown in Table IX. It can be observed that, in this work, the three experimental data sets in this paper are the China data set, the University of Pavia and Houston 2013 in descending order according to spatial size and spectral dimension. It can also be seen from Table IX that the larger the data scale, the longer the running time, and the two are linearly correlated.

At the same time, the running time of the proposed method is reasonable. Taking the China data set as an example, the running time of our method is 25.10 seconds, which is in the middle of the eight methods compared. In addition, in all the comparison methods, SVM [8] has excellent performance in terms of time, mainly because this method does not need feature extraction or probability optimization of HSI original data like other methods, and directly carries out pixel-level classification of the original image body. This makes the method take less time than the other seven methods, which is understandable because SVM has at least one step less than the other methods.

V. Conclusion

In this paper, a complementary spectral–spatial method for hyperspectral image classification is proposed. It is mainly composed of two stages, i.e., texture smoothing pre-processing feature extraction method based on ACTVSP and post-processing feature extraction method based on edge preserving

TABLE VIII
CLASSIFICATION PERFORMANCE OF ALL STUDIED APPROACHES FOR THE HOUSTON 2013 DATA SET, INCLUDING SVM [8], WMRF [26], RPNET [51], SP [38], RMGE [52], SPECTRALFORMER [53], WFCG [54] AND OUR METHOD.

Class name	Training set	Testing set	Classification accuracies of different approaches (in %)							
			SVM	WMRF	RPNet	SP	RMGE	Spectralformer	WFCG	Our method
Healthy grass	20	1231	95.42	**98.70**	91.24	86.60	84.97	85.05	96.06	96.66
Stressed grass	20	1234	86.71	96.19	83.57	87.35	70.81	78.61	**97.15**	89.49
Synthetic grass	20	677	99.70	99.85	96.14	**100.00**	99.86	93.06	96.06	**100.00**
Tree	20	1224	**98.37**	92.81	97.16	83.00	63.91	91.75	73.91	97.83
Soil	20	1222	90.20	98.53	94.82	**100.00**	92.51	81.59	90.14	98.86
Water	20	305	94.10	96.39	**98.33**	85.88	98.15	84.59	91.69	92.40
Residential	20	1248	82.58	70.59	91.47	79.94	52.68	62.42	88.04	**91.90**
Commercial	20	1224	88.61	71.81	82.69	**99.18**	44.45	55.72	82.68	87.57
Road	20	1232	66.32	66.96	91.12	85.88	65.10	68.75	**96.24**	89.11
Highway	20	1207	81.10	85.75	88.11	82.88	**93.72**	67.19	81.48	87.43
	20	1215	66.62	65.02	78.45	**97.81**	90.45	64.36	92.31	93.20
Parking lot1	20	1213	65.68	56.97	84.15	**93.94**	66.59	24.57	62.39	86.29
Parking lot2	20	449	32.80	53.45	63.31	80.08	90.19	35.86	**100.00**	77.59
Tennis court	20	408	85.53	**100.00**	91.05	99.76	**100.00**	90.93	**100.00**	90.87
Running track	20	640	97.22	99.06	96.34	**100.00**	97.58	99.69	97.73	98.31
	OA		80.52	82.12	88.13	90.07	76.71	70.57	85.75	**91.96**
	AA		82.07	83.47	88.53	90.82	80.73	72.28	89.73	**91.83**
	Kappa		78.96	80.67	87.17	89.26	74.86	68.24	84.57	**91.31**

TABLE IX
THE COMPUTING TIME OF DIFFERENT APPROACHES.

Data sets	Computing time of all studied approaches (in seconds)							
	SVM	WMRF	RPNet	SP	RMGE	Spectralformer	WFCG	Our method
University of Pavia	**3.60**	519.00	7.36	37.12	161.39	343.55	48.41	72.22
The China data set	**6.11**	31.02	7.31	15.05	321.26	713.03	9.63	25.10
Houston 2013	**26.88**	68.46	124.65	158.44	72.24	244.05	85.32	187.14

filtering [39], and then the two probability maps are fused. Our method can smooth the image, effectively remove invalid information, and keep the edge contour information and details from being ignored while extracting features. At the same time, adopting KPCA [49] after ACTVSP can improve the separability between classes, so that the same kind of pixels gather together and different kinds of pixels stay away, which is conducive to improving the classification effect. In addition, our experimental results on three real hyperspectral data sets show that our method has superior performance on both complex and simple data sets. Besides, experiments on ten public data sets show that the generalization and robustness is strong. Moreover, it can also be concluded that when the number of training samples is limited, the classification accuracy of this method is superior to other recently proposed hyperspectral image classification methods, and the running time is also in the middle position among the comparison methods.

Although our method has strong robustness for data sets in different situations, But there are still some drawbacks:

- In some scenarios that require high classification effects, there are six parameters that need to be manually adjusted, which is a little troublesome in actual use and will affect the application to some extent.
- The running time can be further reduced.
- The proposed algorithm only considers the spectral adaptive weight without considering the spatial domain adaptive weight.

Based on the drawbacks listed above, the focus of our future research is as follows.

- How to reduce the parameters that need to be adjusted or how to automatically select the optimal parameters.
- The ways of reducing algorithm running time and improving execution efficiency by optimizing algorithm.
- How to combine the adjusted spectral adaptive weight with the spatial domain pixel adaptive weight.

ACKNOWLEDGMENT

The authors would like to thank Prof. Melba Crawford and Prof. Dr. Danfeng Hong for sharing the Indian Pines 2010 data set, the IEEE GRSS Image Analysis and Data Fusion Technical Committee for distributing the Houston 2013 data set and authors of all public data sets used in this paper.

REFERENCES

[1] I. C. C. Acosta, M. Khodadadzadeh, L. Tusa, P. Ghamisi, and R. Gloaguen, "A machine learning framework for drill-core mineral mapping using hyperspectral and high-resolution mineralogical data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4829–4842, 2019.

[2] N. Joshi, M. Baumann, A. Ehammer, R. Fensholt, K. Grogan, P. Hostert, M. R. Jepsen, T. Kuemmerle, P. Meyfroidt, E. T. Mitchard *et al.*, "A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring," *Remote Sensing*, vol. 8, no. 1, p. 70, 2016.

[3] R. Manandhar, I. O. Odeh, and T. Ancev, "Improving the accuracy of land use and land cover classification of landsat data using post-classification enhancement," *Remote Sensing*, vol. 1, no. 3, pp. 330–344, 2009.

[4] N. Patel, C. Patnaik, S. Dutta, A. Shekh, and A. Dave, "Study of crop growth parameters using airborne imaging spectrometer data," *International Journal of Remote Sensing*, vol. 22, no. 12, pp. 2401–2411, 2001.

[5] Q. Feng, J. Liu, and J. Gong, "Uav remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote sensing*, vol. 7, no. 1, pp. 1074–1094, 2015.

[6] Y. Yuan, Q. Wang, and G. Zhu, "Fast hyperspectral anomaly detection via high-order 2-d crossing filter," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 620–630, 2014.

[7] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4955–4965, 2013.

[8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[9] E. Blanzieri and F. Melgani, "Nearest neighbor classification of remote sensing images with the maximal margin principle," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1804–1811, 2008.

[10] B. Bigdeli, F. Samadzadegan, and P. Reinartz, "A multiple svm system for classification of hyperspectral remote sensing data," *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 4, pp. 763–776, 2013.

[11] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3973–3985, 2011.

[12] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, vol. 1. IEEE, 2003, pp. 288–290.

[13] T. C. Bau, S. Sarkar, and G. Healey, "Hyperspectral region classification using a three-dimensional gabor filterbank," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 9, pp. 3457–3464, 2010.

[14] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3681–3693, 2015.

[15] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE geoscience and remote sensing letters*, vol. 3, no. 1, pp. 93–97, 2006.

[16] W. Duan, S. Li, and L. Fang, "Superpixel-based composite kernel for hyperspectral image classification," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2015, pp. 1698–1701.

[17] J. Zabalza, J. Ren, J. Zheng, J. Han, H. Zhao, S. Li, and S. Marshall, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4418–4433, 2015.

[18] C. Li, X. Tang, L. Shi, Y. Peng, and Y. Tang, "A two-staged feature extraction method based on total variation for hyperspectral images," *Remote Sensing*, vol. 14, no. 2, p. 302, 2022.

[19] J. A. Benediktsson, M. Pesaresi, and K. Amason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1940–1949, 2003.

[20] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3791–3808, 2020.

[21] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A new approach to mixed pixel classification of hyperspectral imagery based on extended morphological profiles," *Pattern Recognition*, vol. 37, no. 6, pp. 1097–1116, 2004.

[22] A. Plaza, P. Martínez, J. Plaza, and R. Pérez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 466–479, 2005.

[23] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial–spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognition*, vol. 45, no. 1, pp. 381–392, 2012.

[24] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt, "Spatio-spectral remote sensing image classification with graph kernels," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 741–745, 2010.

[25] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 809–823, 2011.

[26] L. Sun, Z. Wu, J. Liu, L. Xiao, and Z. Wei, "Supervised spectral–spatial hyperspectral image classification with weighted markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1490–1503, 2014.

[27] J. Liu, Z. Wu, J. Li, A. Plaza, and Y. Yuan, "Probabilistic-kernel collaborative representation for spatial–spectral hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 2371–2384, 2015.

[28] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3660–3677, 2012.

[29] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 144–153, 2014.

[30] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6712–6722, 2018.

[31] V. Kumar, R. S. Singh, and Y. Dua, "Morphologically dilated convolutional neural network for hyperspectral image classification," *Signal Processing: Image Communication*, vol. 101, p. 116549, 2022.

[32] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2020.

[33] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4823–4833, 2019.

[34] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[35] A. Sellami and S. Tabbone, "Deep neural networks-based relevant latent representation learning for hyperspectral image classification," *Pattern Recognition*, vol. 121, p. 108224, 2022.

[36] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 60–88, 2020.

[37] H. Zhang, "Hyperspectral image denoising with cubic total variation model," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, vol. 7, pp. 95–98, 2012.

[38] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[39] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification with edge-preserving filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2666–2677, 2013.

[40] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sensing*, vol. 10, no. 3, p. 482, 2018.

[41] P. Duan, X. Kang, S. Li, P. Ghamisi, and J. A. Benediktsson, "Fusion of multiple edge-preserving operations for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 10 336–10 349, 2019.

[42] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen, "An augmented lagrangian method for total variation video restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3097–3111, 2011.

[43] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast alternating direction optimization methods," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2014.

[44] G. I. Marchuk, "Splitting and alternating direction methods," *Handbook of numerical analysis*, vol. 1, pp. 197–462, 1990.
[45] G. H. Golub and C. F. Van Loan, "Matrix computations, secondedition," 1989.
[46] M. K. Ng, *Iterative methods for Toeplitz systems*. Numerical Mathematics and Scie, 2004.
[47] E. Huggins, "Introduction to fourier optics," *The Physics Teacher*, vol. 45, no. 6, pp. 364–368, 2007.
[48] C. Li, *An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing*. Rice University, 2010.
[49] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.
[50] M. Hasanlou and S. T. Seydi, "Hyperspectral change detection: An experimental comparative study," *International journal of remote sensing*, vol. 39, no. 20, pp. 7029–7083, 2018.
[51] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS journal of photogrammetry and remote sensing*, vol. 142, pp. 344–357, 2018.
[52] Y. Miao, M. Chen, Y. Yuan, J. Chanussot, and Q. Wang, "Hyperspectral imagery classification via random multigraphs ensemble learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 641–653, 2021.
[53] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
[54] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, 2022.
[55] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[56] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*. Springer, 2013, pp. 237–280.
[57] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
[58] A. Lavanya and S. Sanjeevi, "An improved band selection technique for hyperspectral data using factor analysis," *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 2, pp. 199–211, 2013.
[59] J. Wang, "Classical multidimensional scaling," in *Geometric structure of high-dimensional data and dimensionality reduction*. Springer, 2012, pp. 115–129.
[60] X. He and P. Niyogi, "Locality preserving projections," *Advances in neural information processing systems*, vol. 16, 2003.
[61] N. H. Ly, Q. Du, and J. E. Fowler, "Collaborative graph-based discriminant analysis for hyperspectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2688–2696, 2014.
[62] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–484.

Lulu Shi received B.E. degree from the Changan University, Xi'an, China, in 2020. She is currently studying for a master's degree at the College of Computer, National University of Defense Technology, Changsha, China. Her main research interests are hyperspectral image analysis and machine learning.

Chunchao Li received the B.Sc degree from Huazhong University of Science and Technology, Wuhan, China, in 2018, and the M.S degree from the National University of Defense Technology, Changsha, China, in 2020. He is currently pursuing the Ph.D. degree with the State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, China. His research interests include image processing, parallel computing, and intelligent perception.

Teng Li received the B.S., M.S., and Ph.D. degrees from National University of Defense Technology in 2013, 2015, and 2020, respectively. His research interests include machine learning, hyperspectral image processing and representation learning.

Yuanxi Peng was born in 1966. He received the B.S. degree in computer science from Sichuan University, Chengdu, China, in 1988, and the M.S. and Ph.D. degrees in computer science from the National University of Defense and Technology (NUDT), Changsha, China, in 1998 and 2001, respectively. He was a Visiting Professor with the Department of Electronic and Computer Engineering, University of Toronto, Toronto, ON, Canada, from 2010 to 2011. He has been a Professor with the Computer School, NUDT, since 2011. His research interests are in the areas of hyperspectral image processing, high-performance computing, multi and many-core architectures, on-chip networks, cache coherence protocols, and architectural support for parallel programming.