# Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach

Wenzhi Zhao and Shihong Du

***Abstract*—In this paper, we propose a spectral–spatial feature based classification (SSFC) framework that jointly uses dimension reduction and deep learning techniques for spectral and spatial feature extraction, respectively. In this framework, a balanced local discriminant embedding algorithm is proposed for spectral feature extraction from high-dimensional hyperspectral data sets. In the meantime, convolutional neural network is utilized to automatically find spatial-related features at high levels. Then, the fusion feature is extracted by stacking spectral and spatial features together. Finally, the multiple-feature-based classifier is trained for image classification. Experimental results on well-known hyperspectral data sets show that the proposed SSFC method outperforms other commonly used methods for hyperspectral image classification.**

***Index Terms*—Balanced local discriminant embedding (BLDE), convolutional neural network (CNN), deep learning (DL), dimension reduction (DR), feature extraction.**

## I. Introduction

NOWADAYS, more accurate and timely satellite imagery with high resolution in both spectral and spatial domains can be easily acquired due to sensor technology improvement. Newly developed hyperspectral sensors can collect hundreds of spectral bands as well as high spatial resolution at the same time [1]. Hyperspectral images (HSIs) are widely used in urban mapping, environmental management, crop analysis, and mineral detection. These applications often require the identification of the class of each pixel with a small number of training samples. However, it is hard to get higher interpretation accuracies when dealing with such increased spectral/spatial resolution imagery.

Two factors, in particular, have been recognized as an influence to the classification results obtained with hyperspectral imagery. For one, the high dimension [2] of the spectral information (i.e., hundreds of correlated spectral bands) produces the Hughes phenomenon which can significantly reduce classification accuracies. Different objects may share similar spectral properties (e.g., similar construction materials for both parking lots and roofs in the city area) [3] which make it even impossible to classify HSIs by using spectral information. On the other hand, classification accuracies are suffering a lot from greatly improved spatial resolution. Specifically, rich information provided by high-resolution images may increase the intraclass variation and decrease the interclass variation in both spectral and spatial domains [4] and lead to lower interpretation accuracies. It is critical to introduce the structural and contextual information in the spatial domain for image classification. Based on these two aspects of issues, it is commonly recognized that both effective dimension reduction (DR) methods and robust spatial feature extraction methods should be proposed [5].

Instead of using the full spectral bands for data processing, DR seeks low-dimensional representation for HSI interpretation [6], [7]. Therefore, DR could effectively find the class-specific subspace and also better interpret performance [8], [9]. Representative spectral-analysis-based DR algorithms can be classified into the unsupervised and supervised/semisupervised algorithms [10]. Unsupervised DR algorithms reveal the low-dimensional data structure without using any label information, e.g., principal component analysis (PCA), locally linear embedding [11], and neighborhood-preserving embedding [12]. However, in the field of HSI classification, instead of containing global mutual information, discriminative projections for each class should be explored [13]. Supervised DR methods attempt to learn metrics that keep data points within the same classes close while separating data points from different classes by using labeled samples [14], [15], e.g., linear discriminant analysis (LDA), nonparametric weighted feature extraction (NWFE) [16], local Fisher discriminant analysis (LFDA) [17], and local discriminant embedding (LDE) [18]. LDA explores the best projection to maximize the interclass distance while minimizing the intraclass distance. NWFE introduces nonparametric scatter matrices with training samples around the decision boundary. LFDA extends the LDA by assigning greater weights to closer connecting samples. LDE seeks the best projection to maximize the interclass scatter matrices by keeping away neighboring data points of different classes in a graph embedding framework [19]. However, in HSIs, especially for high-spatial-resolution ones, spectral information shows great variation for intraclass (e.g., roofs with shadows) and the similarity or confusion between different classes (e.g., roads and roofs are similar in spectral domain). To consider the intraclass variation as well as the similarity between the different classes in the spectral domain, we then proposed the balanced LDE (BLDE) algorithm. It introduces a balance objective to consider

Manuscript received October 12, 2015; revised January 25, 2016; accepted March 10, 2016. Date of publication April 8, 2016; date of current version June 1, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 41471315 and the Weng Hongwu Scientific Research Foundation of Peking University, China under Grant WHW201505. *(Corresponding author: Shihong Du.)*

The authors are with the Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China (e-mail: kdxiaozhi@gmail.com; dshgis@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2016.2543748

the intraclass criterion as well as the between-class ones in a graph embedding framework, to increase separability between classes. Although BLDE can effectively project spectral information into a low-dimension representation, it is impossible to distinguish different objects that share the same spectral properties. Therefore, spatial features are commonly incorporated with spectral ones for HSI classification.

Spatial features have been proven to be useful in improving the representation of HSI data and increasing interpretation accuracies. In recent years, many studies have been reported in spatial feature extraction. Extracting spatial features in remote sensing images commonly requires predefined spatial filter parameters which are determined according to the interpreter's performance and knowledge of the problem: gray-level co-occurrence matrix, wavelet texture, geometric image features [20], Gabor texture features, extended morphological features (EMPs), and attribute profiles [21]. These spatial features are aim-specific ones, which means that only one specific type of objects can be detected by each parameter configuration. However, the spatial property shows great variety at low levels (e.g., shape, texture, etc.), which makes it impossible to describe all types of objects by setting empirical parameters. Recently, deep learning (DL) [22], [23] has been described as one of the state-of-the-art machine learning techniques, and it shows great potential in the field of remote sensing classification. Instead of depending on hand-engineered spatial features, DL can generate high-level spatial features automatically and shows great robustness and effectiveness in image classification. In the field of remote sensing, Chen *et al.* [24] investigated the stacked autoencoder (SAE) based method to classify HSIs by input spectral/spatial information directly into the DL framework. Although SAE can also extract deep features from hierarchical architecture, the training samples (image patches) should be flattened to one-dimension to satisfy the input requirement of the SAE. Therefore, the flattened training samples neglected the spatial information that the original images may contain. Chen *et al.* [25] proposed a convolutional neural network (CNN) based 2-D structure extraction method to detect vehicles from high-spatial-resolution imagery. There are two flaws in this work. First, objects in remote sensing images tend to lie in different scales (i.e., roofs and roads with different sizes), and fixed detection windows obviously are not enough to discover such objects with great variation in shapes. Second, it overlooked the spectral information which is crucial for the remote sensing imagery process.

To overcome these drawbacks, we combined the CNN-based spatial features and the BLDE-based spectral features into image interpretation, and we propose a spectral–spatial feature based classification (SSFC) method in this study. In SSFC, the BLDE method is applied to reduce the high-dimension spectral information and extract spectral features in low dimensionality. The spatial features are obtained by training CNN on the few first principal component (PC) bands of the original data set. Feature fusion technique [26] is applied to generate spectral–spatial features. Finally, classifiers are trained based on the combined spectral–spatial features, as presented in Fig. 1. The characteristics of the proposed SSFC method are listed as follows. 1) A CNN-based spatial feature extraction method was proposed for the HSI classification. The extracted DL spatial features are more robust and effective than the traditional handcrafted ones. 2) The proposed BLDE can balance the locality preserving the scatter matrix and the between-class scatter matrix to obtain better discriminate projections. 3) The strategy of combining the spectral features with CNN-based spatial information can well reveal the intrinsic properties that the original data contain.

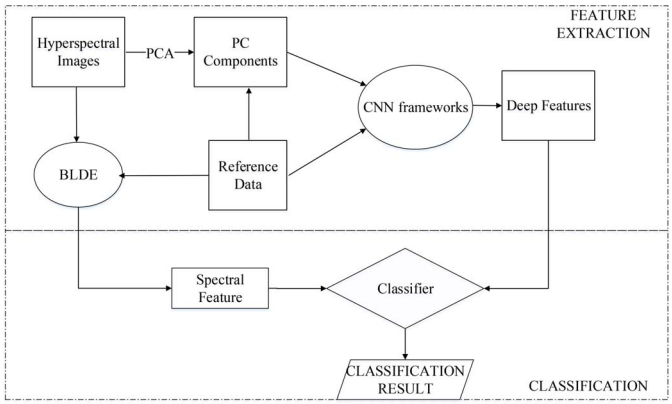



Fig. 1. Flowchart of the proposed SSFC-based classification.

The remainder of this paper is organized as follows. In Section II, the proposed SSFC is described in detail. The experimental results and analysis are presented in Section III, followed by the conclusion in Section IV.

## II. Proposed Approach

The structure of the proposed SSFC-based HSI classification algorithm is shown in Fig. 1. It can be divided into two main components. In the first step, the spectral and spatial features are extracted, respectively. For the spectral ones, DR methods are commonly recommended to reduce the spectral dimensionality; specifically, the BLDE algorithm is chosen to find a low-dimension representation for HSI images in this study. The CNN framework is applied to extract spatial-related deep features at high levels automatically. Then, the proposed spectral–spatial features can be obtained by stacking BLDE-based spectral features with CNN-based spatial features. Finally, the stacked features are inputted into an LR classifier, and classification results can be achieved.

### A. Spectral-Domain BLDE

To extract the spectral-domain features, we propose the BLDE method. It inherits the merits of the traditional LDE that estimates a linear mapping that simultaneously maximizes the local margin between different classes and keeps the samples from the within-class stay close. Also, it can overcome the singularity in the case of limited training samples, as shown in Fig. 2.

Assume that there are $M$ labeled samples $\{\mathbf{x}_i\}_{i=1}^{M} \subset \mathcal{R}^D$, with labels $\{l_i\}_{i=1}^{M}$. Two graphs are built to discover both geometrical and discriminant structures of the data manifold: the within-class graph $G_w$ and the between-class graph $G_b$. $N_w(\mathbf{x}_i)$ represents the neighbors sharing the same label with $\mathbf{x}_i$, while $N_b(\mathbf{x}_i)$ contains the neighbors that belong to different labels. $N_w(\mathbf{x}_i)$ and $N_b(\mathbf{x}_i)$ can be easily determined by nearest neighbor graphs: one nearest neighbor graph for same-class

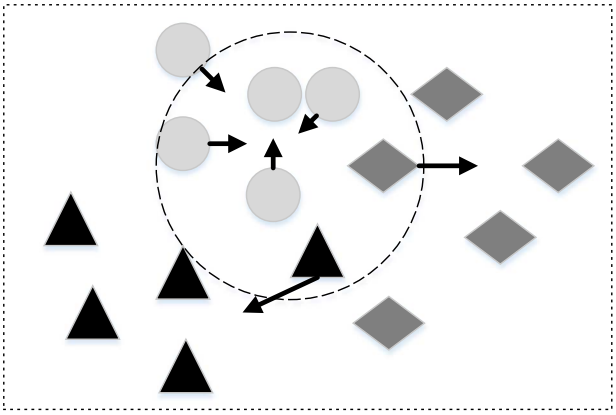


Fig. 2. Illustration of BLDE. The local margin between-classes are maximized, and the distances in the within-class graph are minimized.

samples (parameterized by $k_1$) and one nearest neighbor graph for between-class samples (parameterized by $k_2$). The parameters $k_1$ and $k_2$ can be chosen with empirical values. To construct $G_w$ based on the neighbor graph $N_w$, the affinity matrices $\mathbf{W}_w$ can be defined

$$\mathbf{W}_w = \begin{cases} \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)}{t} & \mathbf{x}_j \in N_w(\mathbf{x}_i) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Similarly, the affinity matrices $\mathbf{W}_b$ of $G_b$ can also be determined by the following equation:

$$\mathbf{W}_b = \begin{cases} 1 & \mathbf{x}_j \in N_b(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $t$ is the heat kernel parameter. The BLDE method computes a linear transform $V$ that simultaneously maximizes the local margins between different-class samples (3) and minimizes the distance between same-class samples (4) in spectral space. Mathematically, this corresponds to

$$\max_{\mathbf{V}} \frac{1}{2} \sum_{i,j} \left\|\mathbf{V}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|^2 W_{b,ij} \tag{3}$$

$$\min_{\mathbf{V}} \frac{1}{2} \sum_{i,j} \left\|\mathbf{V}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|^2 W_{w,ij}. \tag{4}$$

To gain more insight into (3) and (4), the optimization problem can now be formulated in trace form. For the same-class samples, the criteria can be represented as

$$\min J_s = \operatorname{tr}\left\{\mathbf{V}^T X(\mathbf{D}_w - \mathbf{W}_w)X^T\mathbf{V}\right\} \tag{5}$$

and for different-class criteria, the optimization objective can be

$$\max J_d = \operatorname{tr}\left\{\mathbf{V}^T X(\mathbf{D}_b - \mathbf{W}_b)X^T\mathbf{V}\right\} \tag{6}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of the matrix and $\mathbf{X}$ represents the data matrix, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$. $\mathbf{D}_w(\mathbf{D}_b)$ is the diagonal weight matrix, and the entries of $\mathbf{D}_w(\mathbf{D}_b)$ are column sums of $\mathbf{W}_w(\mathbf{W}_b)$. Commonly, the unified criterion is chosen to maximize the Fisher's ratio objective function [18] as

$$\max J_{\text{total}} = \frac{J_d}{J_s} = \frac{\operatorname{tr}\left\{\mathbf{V}^T X(\mathbf{D}_b - \mathbf{W}_b)X^T\mathbf{V}\right\}}{\operatorname{tr}\left\{\mathbf{V}^T X(\mathbf{D}_w - \mathbf{W}_w)X^T\mathbf{V}\right\}} \tag{7}$$

where $\operatorname{tr}(\mathbf{S})$ denotes the trace of the matrix $\mathbf{S}$. $\mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b$ and $\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w$ denote the Laplacian matrix associated with the graphs $G_b$ and $G_w$, respectively. Different from conventional LDE optimization objective, BLDE can simultaneously consider the intraclass scatter matrix and the between-class scatter matrix by introducing the following objective:

$$\max J_{\text{total}} = \frac{J_d - J_s}{J_d + J_s} = \frac{\operatorname{tr}\left\{\mathbf{V}^T(\widetilde{S}_b - \widetilde{S}_w)\mathbf{V}\right\}}{\operatorname{tr}\left\{\mathbf{V}^T(\widetilde{S}_b + \widetilde{S}_w)\mathbf{V}\right\}} \tag{8}$$

where the matrix $\widetilde{S}_b = X\mathbf{L}_b X^T$ denotes the locality-preserving between-class scatter matrix and the matrix $\widetilde{S}_w = X\mathbf{L}_w X^T$ denotes the locality-preserving within-class scatter matrix. However, for many applications in remote sensing image processing, the number of reference samples in the training set $N$ is much smaller than the number of features (hundreds of spectral bands). This is known as the singularity problem. To overcome the complication of singular matrices, we transform the original data onto a PCA subspace so that the resulting matrices $\widetilde{S}_b$ and $\widetilde{S}_w$ in BLDE which are less likely to suffer from the singularity problem.

Finally, the optimal BLDE projection $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v}_\ell]$ can be determined by finding the generalized eigenvectors that correspond to $\ell$ largest eigenvalues. The embedding of spectral features $\mathbf{z}_i$ is obtained by $\mathbf{z}_i = \mathbf{V}^T\mathbf{x}_i$.

### B. Spatial-Domain CNN Framework

*1) Brief Review of CNN:* A CNN [27] is usually composed of alternate convolutional and max-pooling layers (denoted as C layers and P layers) to extract hierarchical features from the original inputs (receptive field), subsequently with several fully connected layers (denoted by FC layers) followed to do classification, as shown in Fig. 3.

Considering a CNN with $L$ layers, we denote the output state of the $l$th layer as $\mathbf{H}^l$, where $l \in \{1, \ldots, L\}$, additionally using $\mathbf{H}^0$ to denote the input data. There are two parts of trainable parameters in each layer, i.e., the weight matrix $\mathbf{W}_l$ that connects the $l$th layer and its previous layer with state $\mathbf{H}^{l-1}$, and the bias vector $\mathbf{b}_l$.

As shown in Fig. 3, the input data are usually connected to a C layer. For a C layer, a 2-D convolution operation is performed first with convolutional kernels $\mathbf{W}_l$. Then, the bias term $\mathbf{b}_l$ is added to the resultant feature maps, in which a pointwise nonlinear activation operation $g(\cdot)$ is typically performed subsequently. Finally, a max-pooling layer is usually followed to select the dominant features over nonoverlapping square windows per feature map. The whole process can be formulated as

$$\mathbf{H}^l = \text{pool}\left(g(\mathbf{H}^{l-1} * \mathbf{W}_l + \mathbf{b}_l)\right) \tag{9}$$

where $*$ denotes the convolution operation and pool denotes the max-pooling operation.

Several C layers and P layers can be stacked one by one to form the hierarchical feature extraction architecture. Then, the resultant features are further combined into 1-D feature vectors by the FC layer. An FC layer first processes its inputs with nonlinear transformation by weight $\mathbf{W}_l$ and bias $\mathbf{b}_l$, and then, the pointwise nonlinear activation is followed:

$$\mathbf{H}^l = g(\mathbf{H}^{l-1} * \mathbf{W}_l + \mathbf{b}_l). \tag{10}$$

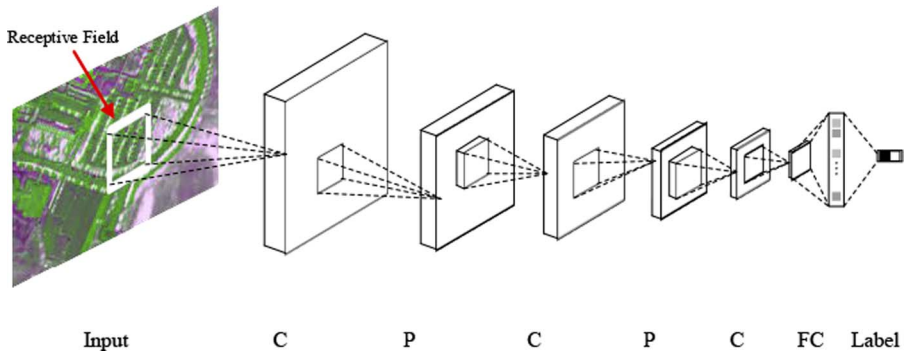



Fig. 3. Conventional CNN framework interspersed with convolutional layers and max-pooling layers. The fully connected layer followed to do classification.

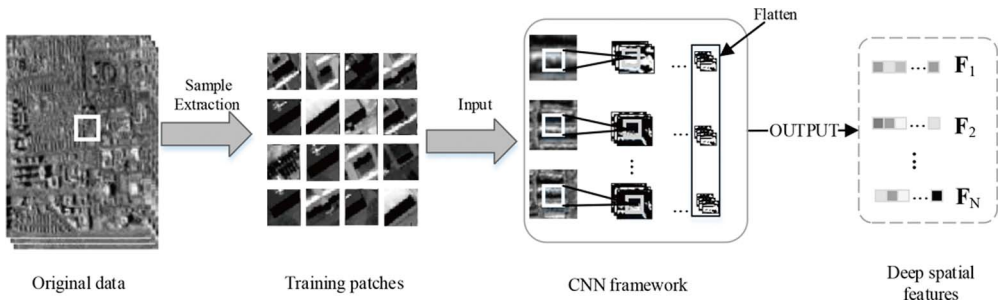



Fig. 4. Feature extraction with a CNN framework. Training samples are extracted from the original data. A CNN framework is trained with training samples. The features at the last layer of the CNN framework are flattened to form the feature vectors.

Several nonlinear activation functions have been proposed. Here, we choose the sigmoid activation function for its high capability and efficiency

$$g(x) = \left(1 + \exp^{(-x)}\right)^{-1}. \tag{11}$$

The last classification layer is usually a softmax layer, with the amount of neurons equaling the number of classes to be classified. We use a logistic regression layer with one neuron to do binary classification, which is similar to an FC layer. Therefore, the activation value represents the probability of the input belonging to the positive class.

The weights $\{\mathbf{W}_1, \ldots, \mathbf{W}_L\}$ and the biases $\{\mathbf{b}_1, \ldots, \mathbf{b}_L\}$ compose the model parameters, which are iteratively and jointly optimized through maximization of the classification accuracy over the training set.

*2) Spatial Feature Extraction With CNN:* Inspired by the conventional CNN, we developed a CNN-based architecture (as shown in Fig. 4) for remote sensing imagery spatial feature extraction. For the input layer, we extract image regions of fixed size centered on the ground-truth pixels to form the training samples. Suppose that there are $M$ training samples (patches) $S_i$, $i \in (1, \ldots, M)$ that are randomly chosen from the original image, and $\mathbf{t}_i$ represents the corresponding label of patch $S_i$. Training a CNN $f(\mathbf{W}, \mathbf{b}|S)$ with $L$ layers is equal to learning the filters $\mathbf{W}$ and the bias parameter $\mathbf{b}$ by optimizing the squared-error loss function of each layer. First, we initialize the parameters of $\mathbf{W}$ and $\mathbf{b}$. Then, based on the initialized CNN, the predicted labels of the last layer are $\mathbf{y}$

$$\mathbf{y}_i = \mathbf{W}^L H^{L-1} + \mathbf{b}^L, \quad i \in \{1, 2, \ldots, M\} \tag{12}$$

here, $\mathbf{y}_i$ is the predicted label based on $L$th layer CNN in response to $i$th input training sample. Based on the predicted labels, the squared-error loss function $\mathcal{L}$ can be written as

$$\min \mathcal{L} = \frac{1}{2} \sum_{i=1}^{M} \|\mathbf{t}_i - \mathbf{y}_i\|_2^2. \tag{13}$$

To minimize the loss function, the backward propagation algorithm is widely used to optimize the parameters $\mathbf{W}$ and $\mathbf{b}$. Specifically, it propagates the predict error $\mathcal{L}$ from the last layer to the first layers and modifies the parameter according to the propagated error at each layer. Commonly, stochastic gradient descent (SGD) algorithm is applied to achieve this goal. In SGD, the derivative of parameters $\mathbf{W}$ and $\mathbf{b}$ can be described as $\Delta k^l = (\partial \mathcal{L}/\partial \mathbf{W}^l)$ and $\Delta b^l = (\partial \mathcal{L}/\partial \mathbf{b}^l)$. Based on the gradient of parameter, the loss function can be optimized.

Once the loss function convergence is achieved, the optimal parameters of $\mathbf{W}$ and $\mathbf{b}$ of each layer can be determined. For an unlabeled image sample $S_{\text{unknown}}$, deep spatial feature $\mathbf{o}$ can be extracted by using the pretrained CNN framework

$$\mathbf{o} = f(\mathbf{W}, \mathbf{b}|S_{\text{unknown}}). \tag{14}$$

In SSFC, BLDE-based spectral features are stacked with spatial features extracted from CNN to form the spectral–spatial features $[\mathbf{z}, \mathbf{o}]$ for image classification.

### C. Computational Complexity

The computational complexities of the SSFC algorithm are analyzed according to the number of training samples. Assume that there are $N$ training samples and the dimensionality of the training data set is $R^d$. There are three parts in SSFC that

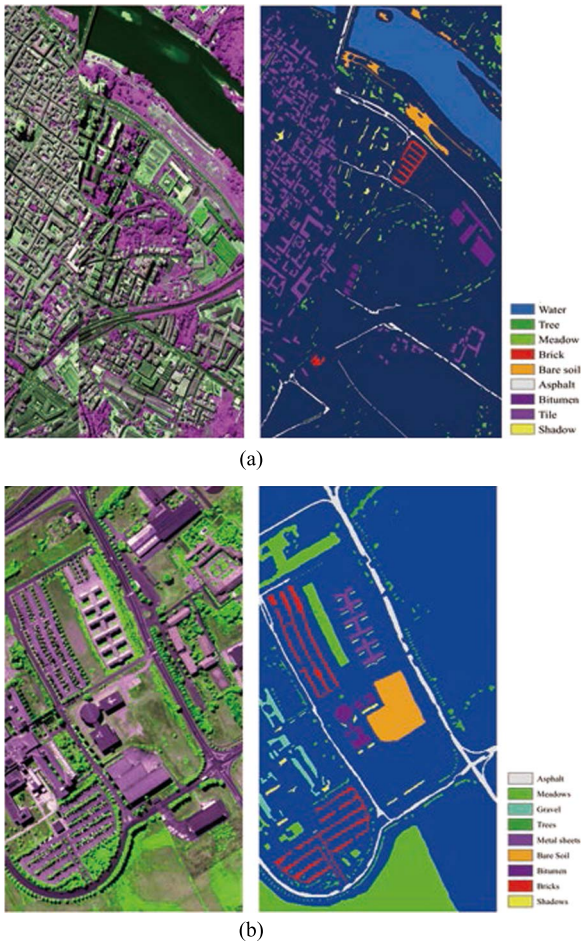



Fig. 5. False color images and ground-truth maps for data sets. (a) Pavia Center. (b) University of Pavia.

should be analyzed. First, the BLDE reduces the dimensionality of original spectral information by calculating the within-class scatter $J_s$ and between-class scatter $J_d$. The computational complexity for the BLDE method is $O(d^2N)$. Then, the CNN algorithm is applied to extract spatial features, which has the computational complexity of $O(d^2N^3)$ for single calculation. The entire complexity of the CNN framework is $O(d^2N^3) \times T$, where $T$ is the iteration number. Finally, the logistic regression method is used to classify the input sample features. The computational complexity of this step is $O(d^2N^2)$. In general, it costs more time to implement the SSFC method to classify HSI images. However, the proposed method should be more accurate and effective than the traditional spectral–spatial feature fusion method in terms of feature extraction.

## III. Experimental Results

### A. Description of Data Sets

Here, we examine SSFC with two popular HSI data sets and present experimental results showing the merits of the proposed approach. These data sets are selected for the following reasons: 1) both data sets are very high spatial resolution images, which brings great difficulty in classification, and 2) they correspond to different scenarios. One is peri-urban, and the other is a dense urban area (Fig. 5). The high dimensionality in spectral domain and complex spatial structures presented in the data sets are representative. Furthermore, all data sets cover a wide range of real cases. This allows us to assess the effectiveness of the proposed method and the validity of the derived conclusions. The overall accuracy (OA), average accuracy (AA), and kappa coefficient (kappa) are used to report the performance of the SSFC on these data sets.

TABLE I
Number of All Reference Data and Training and Test Samples for the Pavia Center and University of Pavia Data Sets

| Pavia Center | | University of Pavia | |
|---|---|---|---|
| Class | Reference data | Class | Reference data |
| Water | 66795 | Asphalt | 7189 |
| Tree | 8418 | Meadow | 19189 |
| Meadow | 3914 | Gravel | 2491 |
| Brick | 3493 | Tree | 3588 |
| Bare soil | 7404 | Metal sheet | 1610 |
| Asphalt | 10064 | Bare soil | 5561 |
| Bitumen | 8095 | Bitumen | 1705 |
| Tile | 44086 | Brick | 4196 |
| Shadow | 3339 | Shadow | 1178 |

The first test site is the hyperspectral digital imagery of Pavia center. The ROSIS sensor provides 115 bands with a spectral range from 0.43 to 0.86 $\mu$m. All noise-affected bands were deleted, with 103 useful channels left. The original size of this data set is 1096 $\times$ 1096, with the spatial resolution of 1.3 m per pixel. A 381-pixel-wide black band was removed. Therefore, a subset of this data set with the size of 1096 $\times$ 715 is applied in this experiment. There are nine classes considered: water, tree, meadow, brick, soil, asphalt, bitumen, tile, and shadow.

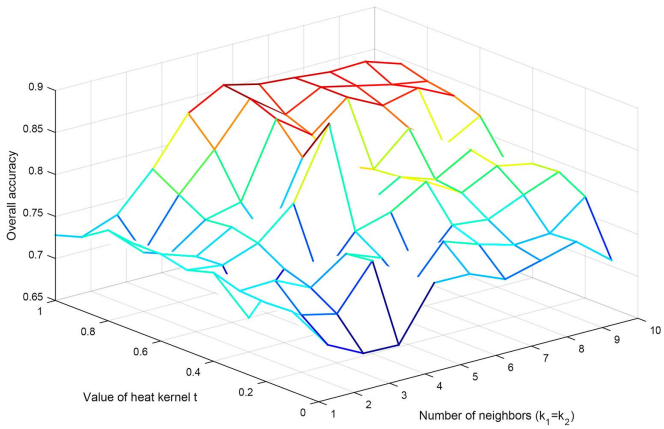
The second is the university scene of Pavia city. There are 103 spectral bands in this hyperspectral data set, with a spatial resolution of 1.3 m. The size of the university scene is 610 $\times$ 340 in pixels. Also, there are nine classes in this data set: asphalt, bitumen, gravel, metal sheet, bricks, shadow, meadow, soil, and tree.

Both data sets and reference maps are shown in Fig. 4. The training and test sets for each data set are listed in Table I, respectively. They are randomly chosen according to the reference data. Pixels from the training set are excluded from the test set in each case and vice versa.

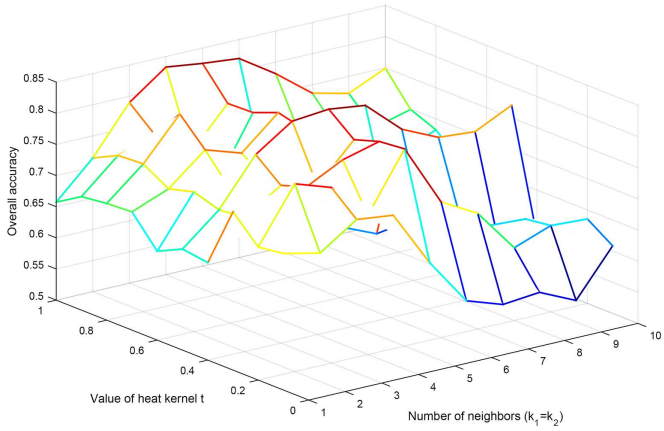
### B. Investigation of BLDE

In BLDE, we evaluated the parameter effect of intraclass and between-class neighbors $k_1, k_2$ and the heat kernel $t$ by setting a series of experiments. We investigated the suitable $k_1$, $k_2$, and $t$ by assigning $k_1 = k_2 \in [1, 2, \ldots, 10]$ and $t \in [0.1, 0.2, \ldots, 1]$. The experimental results are reported in Fig. 6. For the Pavia center data set, $k_1 = k_2 = 5$ and $t = 0.5$ reaches the highest point in Fig. 6. With the same configuration of $k_1$, $k_2$, and $t$, the classification accuracy in the University of Pavia data set is 82.64%, which is similar to the highest accuracy (83.21%). The conclusion coincides with the previous works in [28]. Therefore, for simplification, we set the parameter of BLDE to $k_1 = k_2 = 5$ and $t = 0.5$.

In SSFC, BLDE is used to extract spectral information by reducing the dimensionality in the spectral domain. To investigate the effectiveness of the proposed method, we randomly chose




(a)




(b)

Fig. 6. Evaluation of the parameter effect of the BLDE on different data sets. For all data sets, ten training samples are selected from each class, and the remaining samples are regarded as test samples. (a) Pavia Center. (b) University of Pavia.

$N = 10$, 30, and 50 samples from each class to form the training data set. The remaining samples are set as the testing data set. For each class, 300 samples are chosen as unlabeled test samples. The linear SVM classifier is used in the experiment, and its parameter is selected by fivefold cross-validation. Each experiment is repeated ten times with randomly chosen training samples.

The proposed BLDE is compared with other commonly used DR methods, including PCA, LDA, LDE, RLDE [28], LFDA [17], and NWFE as our comparison methods. We set the reduced dimensionality that varies from 1 to 15. The performances of these methods are reported in Fig. 7. The results show that BLDE outperforms other DR methods in terms of consistent classification results over a wide range of extracted features.

Specifically, the LFDA extended LDA by assigning greater weights to closer connecting examples, which means that "near" pairs are more important than "far" ones. However, in HSIs, especially for high-spatial-resolution ones, spectral information becomes unreliable with the increased spatial resolution, i.e., it commonly shows more spectral variation for intraclass (e.g., roofs with shadows) and spectral similarity or confusion between-class (e.g., roads and roofs are similar in spectral domain). To consider the intraclass variation as well as the similarity of between-class, all equal weights are suggested to be given when constructing a graph. Thus, LDE considers that all pairs are equally important and then maximizes the differences between distinct classes, which is more effective than LDA and LFDA for the HSI process. However, many classes are inseparable in the spectral domain; thus, instead of maximizing the discriminate objective function, keeping a balance between inseparable classes should be a better choice. Therefore, we proposed the BLDE as it considers the intraclass variation as well as the similarity between different classes.

To highlight the merit of BLDE, we compared the one-dimension reduced results with similar DR methods (i.e., LFDA and LDE). Generally, we compared them with two groups of data sets: one is asphalt and shadows selected from the Pavia center image, and the other is meadows and metal sheets selected from the Pavia University image. The projection results are reported in Fig. 8. Since LFDA focuses on "near" pairs, it cannot capture the true intraclass scatter well, leading to the undesirable projection. LDE obtains a projection that maximizes the intraclass variance, but the shape of the class boundary cannot be considered well, leading to an overlap of the two classes. Overall, BLDE shows better performance, especially for inseparable scenarios.

### C. CNN Configuration

As one of the state-of-the-art algorithms in the field of computer vision, CNN can effectively extract spatial-related high-level features. Different from the manual setting parameters of traditional spatial feature extraction methods, the CNN-based method can automatically learn the spatial-related parameters layer by layer. However, the configuration of CNN can greatly affect the classification accuracies in terms of spatial feature extraction as we reported in previous works [29]–[31]. To investigate the performance of CNN-based methods, we tested it mainly on the following two aspects.

*1) Feature Number:* In the framework of CNN, the number of features can determine the dimensionality of the extracted spatial features. To measure how the feature number can affect the classification accuracies, a series of experiments is conducted. The feature number varies from 10 to 30, and the whole CNN framework is constructed by a two-layer structure. Similar to BLDE, we have randomly chosen $N = 10$, 30, and 50 samples from each class to form the training data set. Then, for each class, 300 samples are chosen as unlabeled test samples. The overall accuracy is used to measure the performance of the CNN-based classification algorithm. The classification results are reported in Fig. 9.

As we can see from the results, the classification error rate shows the different values according to the increase of feature numbers for both Pavia center and University data sets. Compared to the Pavia center scene, the University data set shows remarkable stability over the different feature numbers. From Fig. 9, we can see that the overall classification error rate reaches its lowest point at 20 of the Pavia center data set. As for the University scene, the number of features will not significantly impact the classification accuracies. Therefore, we chose 20 output features for both the Pavia center and University data sets.

*2) Depth Effect:* The depth of CNN plays an important role in the classification accuracies since it controls the spatial

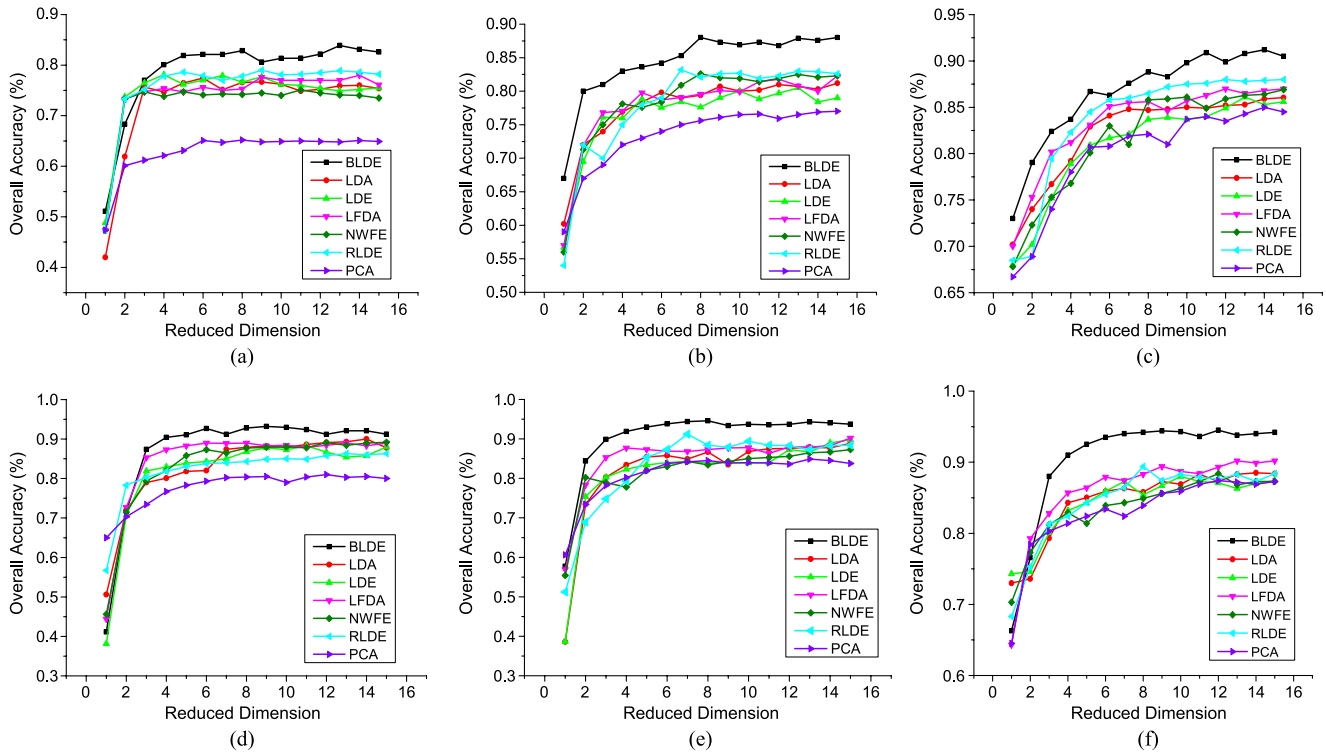



Fig. 7. Comparison of different DR methods; from left to right, 10, 30, and 50 training samples are selected, respectively. (a)–(c) Classification results on Pavia Center. (d)–(f) Classification results on University of Pavia.

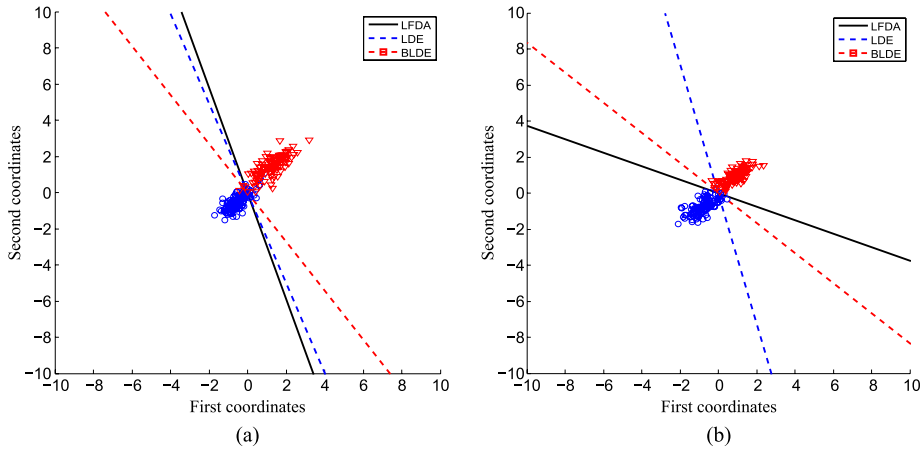



Fig. 8. Projections learned by BLDE, LDE, and LFDA in 2-D space, and two representative classes are presented for each data set. (a) Classes asphalt and shadows in the Pavia Center data set. (b) Classes meadows and metal sheets in the Pavia University data set.

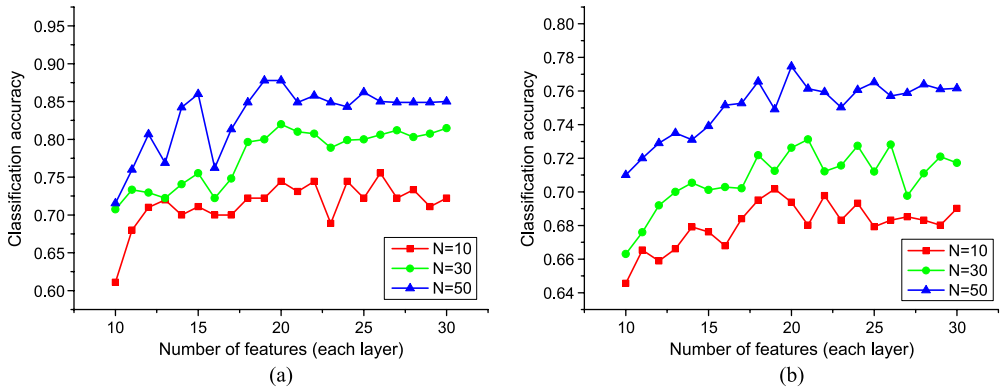



Fig. 9. Classification accuracies of CNN with $N = 10, 30,$ and 50 feature numbers for each layer. (a) Pavia Center. (b) Pavia University.

TABLE II
DEPTH EFFECT ON CNN-BASED CLASSIFICATION

| | Pavia Center | | | University of Pavia | | |
|---|---|---|---|---|---|---|
| | OA | AA | Kappa | OA | AA | Kappa |
| 1-layer | 93.15 | 90.41 | 91.11 | 79.19 | 84.35 | 73.24 |
| 2-layer | 97.67 | 95.62 | 96.73 | 86.11 | 91.98 | 82.35 |
| 3-layer | 98.65 | 97.33 | 98.12 | 87.97 | 89.43 | 84.40 |
| 4-layer | 98.95 | 97.73 | **98.52** | 94.10 | 93.45 | 92.78 |
| 5-layer | **99.68** | **98.07** | 98.15 | **95.98** | **95.38** | **95.24** |

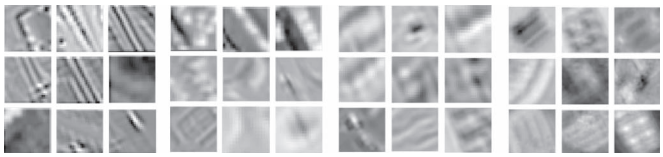



Fig. 10. Evolution of a randomly chosen subset of learned features through the training process. Nine feature maps are listed for each layer.

feature quality in terms of level of abstraction. To measure the effectiveness of the different depth configurations, a series of experiments is conducted on the Pavia center and University data sets. We have randomly chosen $N = 50$ samples from each class to form the training data set. Then, for each class, 300 samples are chosen as unlabeled test samples. Moreover, we set the number of features to 20 of each layer. To testify the effectiveness of the depth parameter, we set four different depths of CNNs from 1 to 5, and the feature number is 20. The experimental results are shown in Table II.

From Table II, we can see that the performance of the deeper CNN configuration is much better than the shallow ones. The reason for this phenomenon is that the spatial feature of high levels is much more abstract and effective for image representation. To better understand the abstract image features from CNN, a subset of the learned features during the training process is displayed in Fig. 10. At the first layer, we can see that features at this level tend to be edges, lines, and corners. Therefore, the features at this depth are low-level ones. Compared to shallow layers, the features at deeper layers seem to be more complex and abstract. After several subsampling and convolution steps, the feature maps at the fifth layer become class-specific and with semantic meanings. Therefore, in this study, we empirically set the size of the size of training set to $\mathcal{P} = 32$, and with a five-layer CNN, the feature number was set to 20 in this study.

### D. Comparison With Other Parametric Classification Methods

For evaluating the performance of the proposed SSFC, the classification results of SSFC are compared to other commonly used DR methods and spectral–spatial feature extraction methods, including PCA, LDA, LDE, RLDE, LFDA, NWFE, EMPs [32], and spectral+EMP (SEMP) [33]. Specifically, the EMPs are built using the area (related to the size of the regions) and standard deviation (which measures the homogeneity of the pixels enclosed by the regions) attributes. Following the work in [34], the threshold values of area are chosen in the range of {50, 500} and standard deviation ranging from 2.5% to 20%.

Here, LR and SVM are the default classifier for all of the comparison methods. The regularization terms of the LR classifier and linear SVM classifier are determined by fivefold cross-validation with the arrangement of $[2^{-2}, 2^{-1}, \ldots, 2^{11}, 2^{12}]$. The comparison results on the two HSI data sets are shown in Figs. 11 and 12, respectively. The results in Tables III and IV are the highest overall accuracies (in percent) among the first 30 features. Each number in the brackets corresponds to the optimal number of extracted spectral features (EMP also regarded as spectral features).

In SSFC, the first three PCs were selected as input images. The "Raw" method means that the original HSI images were directly used for classification without any feature extraction step. In the PCA-based classification algorithm, the original image was reduced into few PCs, which can be used for classification. The PCA classification method reduces the dimensionality in the spectral domain without any labels. However, it increases the variation in spatial domain (i.e., texture or shape variation). Therefore, the classification accuracies are not better than the "Raw" method, especially for the tile and road classes. Compared to unsupervised spectral DR methods, the local-preserve-based method shows better performance in terms of classification accuracies for both the Pavia center and University scene. The BLDE-based method balances both intraclass and interclass criteria by using a balancing parameter which outperforms the LDA and LDE in terms of overall accuracies. However, the classification results in the spectral domain are commonly no better than the spatial-based methods. The cause of this phenomenon is that the increased spatial resolution amplifies the within-class variation while decreasing the interclass separability. Thus, the spatial feature which overcomes spectral uncertainty is needed extremely, especially for such high-resolution images.

Different from spectral-domain DR methods, spatial-based methods try to find better spatial representation which can be more effective for HSI image classification, especially for large homogeneous areas. The EMP spatial feature extraction method was proposed to achieve this goal. However, spatial-based classification methods always result in poor performance without the combination of spectral information. Therefore, spectral–spatial-based methods that incorporate features from the spectral domain as well as spatial ones are widely used for HSI interpretation. In SEMP, spectral information and morphological features are stacked together for image classification. Therefore, spatial features as well as spectral ones can be simultaneously considered. However, such spatial-related features are hand-engineered ones and are time-consuming. The CNN-based method can automatically learn spatial-related features at high levels, which are more effective and robust than the predefined pixel-level features. In SSFC, both BLDE and CNN are introduced for HSI classification. In spatial domain, CNN extracts features automatically without any predefined parameters. Moreover, BLDE generates spectral features by balancing the intraclass scatter and between-class scatter.

The classification accuracies over different classification methods are reported in Tables III and IV. Generally, spectral-based methods show better performance in terms of water and shadows. EMP-based methods achieve the best results on asphalt and tree, but they fail to interpret classes with complex

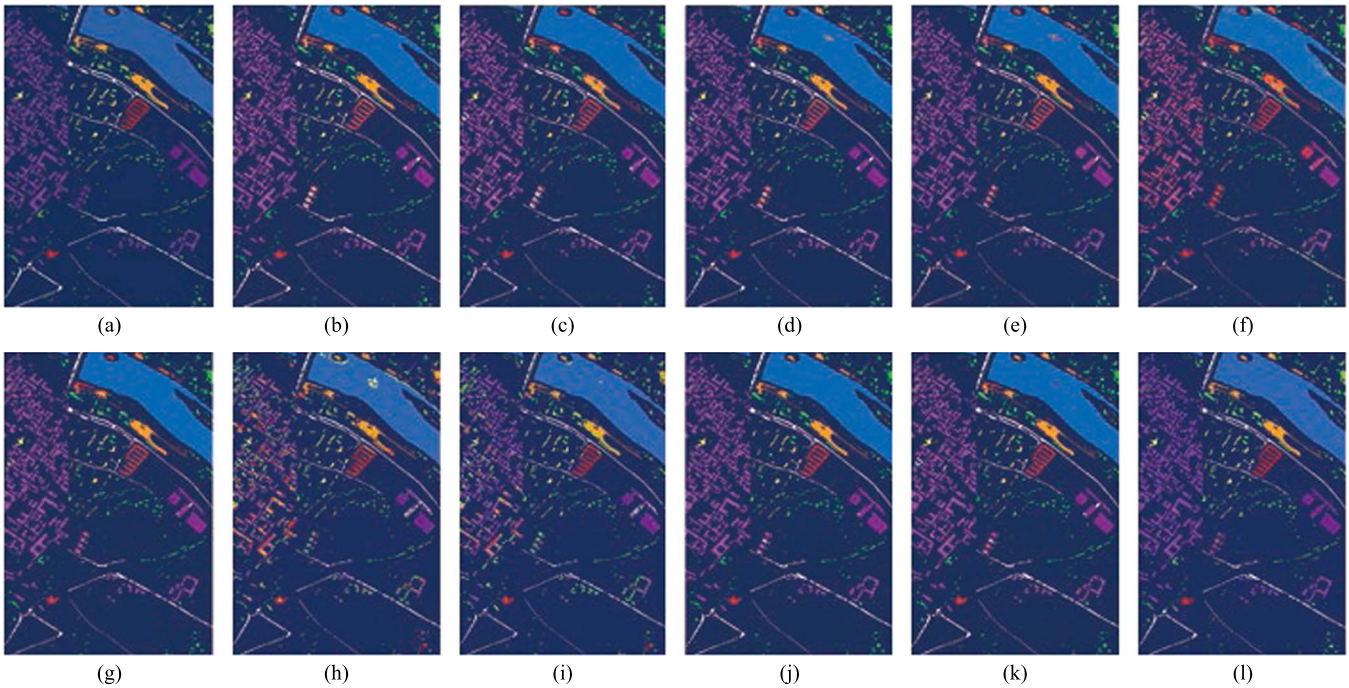



Fig. 11. Best classification results of the Pavia center scene. (a) Reference map of Pavia Center. (b)–(l) classification results by using RAW, PCA, LDA, LFDA, NWFE, LDE, RLDE, BLDE, EMP, SEMP, and SSFC.

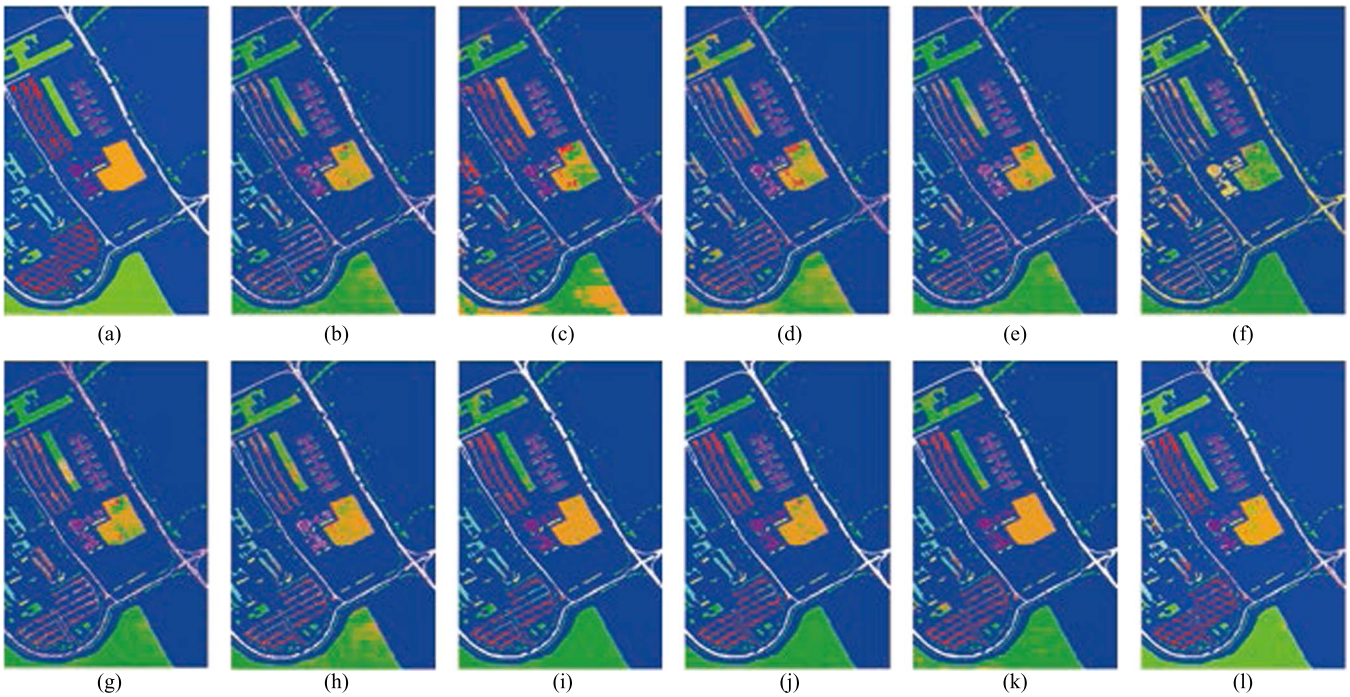



Fig. 12. Best classification results of University of Pavia. (a) Reference map of Pavia Center. (b)–(l) classification results by using RAW, PCA, LDA, LFDA, NWFE, LDE, RLDE, BLDE, EMP, SEMP, and SSFC.

shapes (e.g., bare soil and meadow). The classification results of the different classification methods are shown in Fig. 11. As we can see, graph-based DR methods produce "pepper" noise on classification results. However, the spatial-based methods can overcome the noise effect, but they fail to incorporate spectral information and produce false predictions. In SSFC-based classification results, "pepper" noise and spectral uncertainty are both greatly reduced. With the combination of DR and CNN methods, the best classification results are obtained, especially for the peri-urban data set.

TABLE III
HIGHEST OVERALL ACCURACIES (IN PERCENT) ON THE PAVIA CENTER DATA SET

| N | Method | Raw | PCA | LDA | LFDA | LSDA | NWFE | LDE | BLDE | EMP | SEMP | SSFC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | LR | 92.87 ± 0.94 | 93.32(30) ± 0.49 | 92.73(9) ±0.89 | 91.86(18) ±1.69 | 82.86(27) ±6.59 | 91.16(30) ±1.32 | 92.03(6) ±0.92 | 94.13(16) ±0.94 | 92.45(21) ±0.95 | 94.06(31) ±0.68 | 95.42(14) ±1.41 |
| | SVM | 92.43 ±1.18 | 93.21 (27) ±1.80 | 91.69 (11) ±2.42 | 90.36 (21) ±3.3 | 84.39 (29) ±6.21 | 91.32 (30) ±2.21 | 92.34 (6) ±1.34 | 93.89 (20) ±1.03 | 92.77 (27) ±0.77 | 93.76(29) ±1.07 | 95.77 (12) ±2.72 |
| 20 | LR | 94.13 ± 0.96 | 93.97 (30) ± 0.26 | 94.44 (10) ±0.69 | 94.84 (22) ±0.80 | 90.57 (29) ±1.42 | 913.24 (30) ±1.06 | 95.76 (6) ±0.82 | 96.51 (16) ±0.57 | 95.67 (25) ±1.17 | 97.82 (27) ±0.65 | 97.62 (14) ±0.86 |
| | SVM | 94.21 ±1.01 | 94.41 (29) ±0.73 | 93.76 ±1.00 | 91.57 (19) ±1.42 | 90.57 (30) ±3.79 | 94.03 (29) ±1.72 | 96.15 (6) ±1.91 | 95.30 (17) ±1.12 | 95.92 (27) ±2.38 | 97.21 (26) ±0.92 | 96.76 (13) ±1.89 |
| 30 | LR | 95.46 ±1.23 | 94.19 (30) ±0.68 | 95.64 (13) ±0.38 | 95.86 (20) ±0.36 | 92.54 (29) ±0.89 | 94.27 (30) ±0.95 | 96.04 (6) ±1.03 | 97.48 (19) ±0.49 | 96.61 (29) ±1.18 | 98.03 (26) ±0.45 | 98.97 (7) ±1.16 |
| | SVM | 94.78 ±1.46 | 94.02 (30) ±1.08 | 95.25 (12) ±0.73 | 93.70 (18) ±0.88 | 93.70 (30) ±2.60 | 94.18 (27) ±2.13 | 95.72 (8) ±1.01 | 97.28 (17) ±1.06 | 96.03 (30) ±0.79 | 97.87 (25) ±1.08 | 99.11 (10) ±1.29 |
| 40 | LR | 95.98 ±0.62 | 95.00 (28) ±0.93 | 96.30 (14) ±0.23 | 96.28 (21) ±0.27 | 93.92 (28) ±0.65 | 95.91 (26) ±0.67 | 96.51 (7) ±0.91 | 97.93 (13) ±0.57 | 97.55 (27) ±0.83 | 98.82 (26) ±0.76 | 99.15 (11) ±0.70 |
| | SVM | 95.70 ±0.81 | 94.24 (29) ±0.69 | 95.89 (13) ±0.46 | 95.23 (21) ±0.59 | 95.23 (30) ±1.03 | 96.32 (29) ±1.28 | 96.66 (8) ±1.24 | 97.12 (19) ±1.08 | 98.06 (26) ±2.19 | 99.01 (26) ±0.92 | 98.69 (9) ±1.29 |
| 50 | LR | 95.70 ±0.80 | 95.71 (30) ±1.03 | 96.50 (13) ±0.21 | 96.55 (21) ±0.23 | 95.68 (29) ±0.42 | 96.47 (28) ±0.19 | 97.01 (7) ±0.58 | 98.92 (16) ±0.73 | 97.83 (28) ±1.26 | 98.61 (26) ±1.05 | 99.87 (9) ±0.51 |
| | SVM | 95.99 ±0.53 | 95.75 (30) ±0.49 | 96.39 (13) ±0.35 | 95.94 (21) ±0.40 | 95.43 (29) ±1.05 | 96.54 (28) ±0.75 | 96.69 (7) ±0.79 | 97.95 (18) ±0.62 | 97.75 (29) ±2.07 | 98.25 (27) ±0.81 | 99.29 (11) ±0.92 |

TABLE IV
HIGHEST OVERALL ACCURACIES (IN PERCENT) ON THE UNIVERSITY OF PAVIA DATA SET

| N | Method | Raw | PCA | LDA | LFDA | LSDA | NWFE | LDE | BLDE | EMP | SEMP | SSFC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | LR | 76.52 ±4.35 | 77.72 (21) ±3.98 | 80.86 (7) ±3.12 | 79.02 (8) ±2.53 | 78.64 (11) ±3.52 | 77.38 (26) ±2.75 | 80.85 (6) ±2.56 | 81.99 (4) ±2.67 | 76.82 (15) ±3.05 | 78.17 (25) ±1.98 | 88.97 (9) ±2.32 |
| | SVM | 76.36 ±5.62 | 77.35 (16) ±4.36 | 79.34 (8) ±3.75 | 78.93 (15) ±3.69 | 79.82 (27) ±4.82 | 79.89 (27) ±4.26 | 80.06 (8) ±3.74 | 80.67 (6) ±4.21 | 77.11 (21) ±4.72 | 77.96 (25) ±3.77 | 88.42 (15) ±3.97 |
| 20 | LR | 85.29 ± 0.88 | 85.86 (19) ± 0.67 | 85.84 (6) ±0.98 | 85.66 (9) ±1.17 | 86.24 (8) ±2.03 | 86.98 (19) ±1.76 | 85.67 (6) ±1.31 | 89.24 (6) ±2.06 | 87.65 (14) ±1.68 | 88.86 (19) ±1.53 | 91.87 (8) ±1.42 |
| | SVM | 85.59 ±3.25 | 86.13 (28) ±3.34 | 85.06 (8) ±1.85 | 86.98 (21) ±2.51 | 85.45 (24) ±3.21 | 86.92 (30) ±3.87 | 86.41 (9) ±2.45 | 88.79 (9) ±3.07 | 88.37 (16) ±2.89 | 89.13 (21) ±3.29 | 91.23 (13) ±2.26 |
| 30 | LR | 89.92 ±1.21 | 90.01 (16) ± 0.64 | 88.23 (7) ±0.76 | 88.69 (10) ±0.82 | 90.02 (10) ±0.73 | 89.26 (20) ±0.59 | 90.11 (7) ±0.61 | 92.79 (6) ±0.49 | 91.86 (15) ±0.81 | 93.02 (14) ±0.73 | 94.24 (6) ±1.23 |
| | SVM | 89.85 ±1.90 | 89.79 (29) ±1.89 | 87.95 (6) ±1.3 | 90.19 (25) ±1.34 | 89.67 (19) ±1.52 | 90.13 (25) ±1.22 | 89.44 (7) ±1.82 | 92.65 (11) ±1.33 | 91.42 (23) ±2.02 | 93.26 (20) ±1.94 | 93.87 (12) ±2.41 |
| 40 | LR | 88.65 ±1.03 | 90.03 (20) ±0.43 | 89.03 (7) ±0.52 | 89.97 (9) ±0.54 | 90.32 (9) ±0.61 | 90.67 (21) ±0.72 | 91.03 (8) ±0.47 | 91.82 (10) ±0.59 | 92.71 (15) ±0.56 | 92.98 (18) ±0.76 | 94.23 (7) ±0.74 |
| | SVM | 90.48 ±1.21 | 90.51 (30) ±1.11 | 88.41 (8) ±1.66 | 90.04 (23) ±1.50 | 89.59 (18) ±1.24 | 91.14 (30) ±1.35 | 90.67 (10) ±1.26 | 92.31 (13) ±1.06 | 92.63 (16) ±1.82 | 93.44 (22) ±1.48 | 94.16 (9) ±1.37 |
| 50 | LR | 89.67 ±0.76 | 91.03 (22) ±0.61 | 90.62 (8) ±0.67 | 92.06 (12) ±0.53 | 91.67 (10) ±0.39 | 91.71 (22) ±0.41 | 90.37 (9) ±0.45 | 94.03 (10) ±0.63 | 93.02 (14) ±0.37 | 94.81 (16) ±0.42 | 96.98 (7) ±0.53 |
| | SVM | 91.34 ±1.13 | 90.83 (27) ±1.12 | 90.37 (8) ±0.95 | 91.22 (30) ±1.03 | 90.67 (21) ±0.87 | 92.03 (26) ±1.21 | 91.84 (11) ±0.77 | 93.27 (11) ±0.84 | 93.75 (20) ±0.91 | 95.10 (29) ±0.81 | 96.03 (8) ±0.73 |

## IV. CONCLUSION

In this paper, we have proposed a novel SSFC method for the classification of HSI data. There are two parts of this proposed method. First, BLDE is used to extract low-dimensionality spectral features from the original HSI data sets. Second, high-level spatial features are extracted by a CNN framework. Finally, both spectral and spatial features are stacked together, and an LR classifier is trained for classification. The proposed method has the following characteristics: 1) the balancing strategy overcomes the singularity, embodies the data diversity, and improves the classification accuracies; 2) the CNN-based spatial feature extraction technique overcomes the unpleasantness in spatial filtering parameter selection; and 3) combining the spectral and spatial discriminant features leads to excellent classification accuracies. The proposed method is compared to other well-known classification methods. The classification results on well-known data sets have shown that the proposed method has better performance in terms of classification accuracies.

However, the proposed method can still be revised in some aspects. For instance, how to set the size of training samples $\mathcal{P}$, which is also a key aspect in DL-based methods, has not been considered in the current framework yet. Also, it is difficult to interpret the significance of CNN features for possible improvements in the future. Therefore, our future work will focus on how to select the optimal observation scale in order to obtain the best classification accuracies for SSFC.

## ACKNOWLEDGMENT

The authors would like to thank Prof. Gamba from the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society for providing the ROSIS data set, Prof. H. Chen for sharing the LDE source code, Prof. C. Lin for providing the LIBSVM toolbox, and the handling editor and three anonymous reviewers for their detailed and constructive comments and suggestions, which greatly helped in improving the quality of this paper.

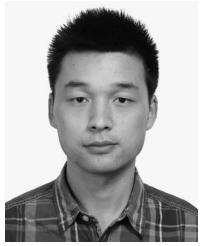
## REFERENCES

[1] C.-I. Chang, *Hyperspectral Data Exploitation: Theory and Applications*. New York, NY, USA: Wiley, 2007.

[2] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc. AMS Math Challenges Lecture*, 2000, pp. 1–32.

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

[4] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.
[5] A. Plaza, P. Martinez, R. Pérez, and J. Plaza, "Spatial/spectral endmember extraction by multidimensional morphological operations," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 9, pp. 2025–2041, Sep. 2002.
[6] C. Lee and D. A. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 4, pp. 792–800, Jul. 1993.
[7] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.
[8] M. Hasanlou and F. Samadzadegan, "Comparative study of intrinsic dimensionality estimation and dimension reduction techniques on hyperspectral images using K-NN classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1046–1050, Nov. 2012.
[9] H. Luo, L. Yang, H. Yuan, and Y. Y. Tang, "Dimension reduction with randomized anisotropic transform for hyperspectral image classification," in *Proc. IEEE Int. Conf. CYBCONF*, 2013, pp. 156–161.
[10] 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.
[11] Y. Fang *et al.*, "Dimensionality reduction of hyperspectral images based on robust spatial information using locally linear embedding," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1712–1716, Oct. 2014.
[12] J. Gui *et al.*, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognit.*, vol. 45, no. 8, pp. 2884–2893, Aug. 2012.
[13] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 55–66, Jan. 2014.
[14] S. Chen and D. Zhang, "Semisupervised dimensionality reduction with pairwise constraints for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 369–373, Mar. 2011.
[15] Q. Shi, L. Zhang, and B. Du, "Semisupervised discriminative locally enhanced alignment for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4800–4815, Sep. 2013.
[16] B.-C. Kuo and D. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.
[17] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, 2007.
[18] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE CVPR*, 2005, vol. 2, pp. 846–853.
[19] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[20] P. Dare and I. Dowman, "An improved model for automatic feature-based registration of SAR and spot images," *ISPRS J. Photogramm. Remote Sens.*, vol. 56, no. 1, pp. 13–28, Jun. 2001.
[21] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.
[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Adv. Neural Inf. Process. Syst.*, vol. 19, p. 153, 2007.
[24] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[25] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
[26] X. Kang, S. Li, and J. A. Benediktsson, "Feature extraction of hyperspectral images with image fusion and recursive filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3742–3752, Jun. 2014.
[27] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
[28] Y. Zhou, J. Peng, and C. Chen, "Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 1082–1095, Feb. 2015.
[29] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3368–3379, 2015.
[30] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015.
[31] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271616000137.
[32] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, Mar. 2005.
[33] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.
[34] B. Song *et al.*, "Remotely sensed image classification using sparse representations of morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5122–5136, Aug. 2014.



**Wenzhi Zhao** was born in Shandong, China, in 1990. He is currently working toward the Ph.D. degree in the Institution of Remote Sensing and Geographic Information System, Peking University, Beijing, China.

His research interests include hyperspectral data analysis, high-resolution image processing, deep learning techniques, and computational intelligence in remote sensing images.

**Shihong Du** received the B.S. and M.S. degrees in cartography and geographic information system from Wuhan University, Hubei, China, in 1998 and 2001, respectively, and the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2004.

He is currently an Associate Professor with Peking University, Beijing. His research interests include qualitative knowledge representation, reasoning and its applications, and semantic understanding of spatial data including GIS and remote sensing data.