# PSASL: Pixel-Level and Superpixel-Level Aware Subspace Learning for Hyperspectral Image Classification

Jie Mei, Yuebin Wang, Liqiang Zhang, Bing Zhang, Suhong Liu, Panpan Zhu, and Yingchao Ren

***Abstract*— The performance of hyperspectral image (HSI) classification relies on the pixel information obtained from hundreds of contiguous and narrow spectral bands. Existing approaches, however, are limited to exploit an appropriate latent subspace for data representation within the pixel-level or superpixel-level. To utilize spectral information and spatial correlation among pixels in HSI and avoid the "salt-and-pepper" problem generated in the pixel-based HSI classification, a novel pixel-level and superpixel-level aware subspace learning method called PSASL is developed. The PSASL constructs the subspace learning framework based on the reconstruction independent component analysis algorithm. The spectral–spatial graph regularization and label space regularization are developed as the pixel-level constraints. To avoid the "salt-and-pepper" problem generated in the pixel-based classification methods, superpixel-level constraints are introduced for integrating the data representations defined in the subspace and class probabilities of the pixels in the same superpixel. The subspace learning and the pixel-level regularization are combined with the superpixel-level regularization to form a unified objective function. The solution to the objective function is efficiently achieved by employing a customized iterative algorithm, and it converges very fast. A discriminative data representation and a universal multiclass classifier are learned simultaneously. We test the PSASL on three widely used HSI data sets. Experimental results demonstrate the superior performance of our method over many recently proposed methods in HSI classification.**

***Index Terms*— Feedback information, hyperspectral image (HSI) classification, pixel and superpixel, semisupervised learning, subspace learning.**

Manuscript received November 28, 2017; revised April 20, 2018 and September 19, 2018; accepted December 22, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0213600, in part by the National Natural Science Foundation of China under Grant 41371324 and Grant 41801241, and in part by the China Postdoctoral Science Foundation under Grant 2016M600953 and Grant 2018T110065. *(Jie Mei and Yuebin Wang contributed equally to this work.) (Corresponding author: Liqiang Zhang.)*

J. Mei, L. Zhang, S. Liu, and P. Zhu are with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: meijie@mail.bnu.edu.cn; zhanglq@bnu.edu.cn; liush@bnu.edu.cn; zlyxbmsl@163.com).

Y. Wang is with the School of Land Science and Technology, China University of Geosciences, Beijing 100083, China, and also with the School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China (e-mail: xxgcdxwyb@163.com).

B. Zhang and Y. Ren are with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China (e-mail: zb@radi.ac.cn; yingchaoren@gmail.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.2018.2890508

## I. Introduction

HYPERSPECTRAL image (HSI) classification has become a challenging research topic in the computer vision and remote sensing fields. Evaluation of the similarity of two pixels induced by the high dimensionality and the problem of limited training samples are two prominent challenges in HSI classification [1], [35]–[39]. In recent years, many methods have been proposed to classify HSIs [15], [33], [34], [40]–[46]. They are usually divided into two types of approaches, i.e., pixel-level-based and superpixel-level-based methods [28]–[30], [47]–[49].

Pixel-based HSI classification methods have appeared in [2], [13]–[16], [57], and [58]. In [2], an HSI classification algorithm based on discriminative conditional random field was developed. Ma *et al.* [13] combined the local manifold learning and $k$-nearest-neighbor (KNN) classifier for HSI classification. In this framework, locally linear embedding, local tangent space alignment, and Laplacian eigenmaps are investigated with these classifiers. A semisupervised graph-based method was proposed in [14]. It can well handle the special characteristics of the HSI, namely, high-input dimension of pixels, few labeled samples, and spatial variability of the spectral signature. Gao *et al.* [1] constructed a bilayer graph-based learning framework for HSI classification. A spectral–spatial feature learning method was proposed to obtain robust features of HSI in [16]. It combines the spectral feature learning and spatial feature learning into a hierarchical structure. He *et al.* [51] provided a comprehensive overview on the methods belonging to the category of spectral–spatial classification in a unified context. A method based on the generalization of concepts from mathematical morphology to multichannel imagery was introduced [52] to analyze the mixed pixel. The pixel-based classification approaches can well exploit discriminant spectral information of each pixel in the HSI. However, the classification accuracy is degraded with the redundant information among the pixels. Due to the noise and mixed spectral pixels in HSIs and lack of contextual information among pixels, the pixel-level classification methods are likely to generate noisy appearance in classification maps [17]. Dimensionality reduction and sparse representation are effective solutions to reduce redundant information among the pixels of HSIs. How to perform the dimensionality reduction of the high-dimensional features and

establish the neighboring relationship among the pixels from the high-dimensional features are the key toward a successful classification [1].

To a certain extent, dimensionality reduction is equal to subspace learning, that is, projects the original high-dimensional feature space to a low-dimensional subspace where the statistical properties like independent component analysis (ICA) [18] and principal component analysis (PCA) [19] can be well preserved. Based on a discriminative locally enhanced alignment technique, a dimensionality reduction method was proposed to maximize the distance between different classes and preserve the intrinsic geometric structure of the data by the use of labeled and unlabeled samples [20]. In [21], class separability, neighborhood structure preservation, and nearest feature line measurement were considered simultaneously to determine a transformation in the eigenspaces for dimensionality reduction. Sparse representation is also a powerful tool for extracting features from HSI [21]–[23], [59], [60]. A joint collaborative representation classification method with multitask learning was designed for HSI classification. By using the dictionaries of spectral, gradient, shape, and texture, spare features are extracted for HSI classification. A similar framework for dictionary training and feature extraction is also found in [22] and [23]. Different from [21] and [22], a multiscale adaptive sparse representation model is proposed in [23]. In the regions with different scales, the complementary information is incorporated for HSI classification. In [53], the task-driven dictionary learning (TDDL) algorithm was proposed for the supervised dictionary learning method. Sun *et al.* [53] proposed to enforce structured sparsity priors on the TDDL method in order to improve the performance of HSI classification.

For exploiting the contextual information among pixels, the HSI classification method needs to contain neighborhood covering or neighborhood importance with adaptive behavior across the HSI. A typical group of methods are superpixel/object/segmentation-based ones [54]–[56], where pixels within objects were collaboratively utilized for classification [51]. Among these works, superpixel-based classification is adopted to establish the neighboring relationship among the pixels. Superpixels are generated using the graph-based algorithms like normalized cuts [24] and entropy rate superpixel segmentation [25] or the gradient-descent-based algorithms like simple linear iterative clustering [26] and superpixels extracted via energy-driven sampling [27]. In [28], superpixels instead of pixels are applied to the graphical model to capture the contextual information and the spatial dependence among the pixels. Fang *et al.* [29] considered a superpixel in the HSI as a small spatial region whose size and shape can be adaptively adjusted for different spatial structures. In their approach, pixels within each superpixel were jointly represented by a set of common atoms from a dictionary via a joint sparse regularization. A superpixel-level sparse representation classification framework with multitask learning was developed in [30]. It exploited the class-level sparsity prior and the correlation of neighboring pixels to fuse multiple features. Based on the superpixels, a spectral–spatial adaptive sparse representation method was proposed for HSI compression by taking advantage of the spectral and spatial information of HSIs [31]. Sparse representation can transform spectral signatures of the pixels into sparse coefficients with very few nonzero entries.

The superpixel-based classification approach can well avoid the "salt-and-pepper" problem generated in the pixel-based classification methods. Yet, the classification accuracy is deteriorated if under-segmentation cannot be fully avoided in superpixel-based approaches [17]. Li *et al.* [17] presented a supervised HSI classification method by the probabilistic fusion of pixel-level and superpixel-level classifiers. This method generates superpixels from the first three principal components using the PCA, which is difficult to describe the spectral and spatial information of the pixels in each superpixel. Moreover, the way for generating superpixels does not consider the purity of each superpixel; thus, it decreases the classification accuracy. In our method, we take the spectral and spatial information and purity of the pixels in each of the superpixels into account.

HSIs exhibit strong dependencies across spatial and spectral neighbors, which have been proved to be useful for HSI classification [2]. In this paper, a pixel-level and superpixel-level aware subspace learning method (PSASL) is developed to effectively use spectral and spatial correlations among pixels in HSIs. Label space is utilized to extract pixel-based features by means of subspace learning. Then, these constraints including spectral–spatial information and label space are integrated into the subspace learning procedure. Based on the learned subspace of pixels, an adaptive mean shift-clustering algorithm is employed to generate superpixels, which provides feedback information to the subspace learning and cluster results in the pixel-level. The feedback information is also considered as the constraints of subspace learning. Thus, the subspace learning and the constraints defined in the pixel-level and superpixel-level form a unified objective function. The objective function is solved by means of a customized iterative algorithm. The overview of the PSASL for HSI classification is shown in Fig. 1. We test our method on three widely used HSI classification data sets. The experimental results show that our method outperforms other HSI classification methods.

The main contributions of this paper are summarized as follows.

1) The pixel-level regularization, superpixel-level regularization, and a single predictive linear classifier are explicitly integrated into a unified objective function for subspace learning. The proposed method is very effective and efficient for semisupervised HSI classification. It far outperforms the recently proposed methods in terms of classification accuracy.
2) The spectral–spatial graph regularization and label space regularization are developed as the pixel-level constraints for removing the redundant information in HSIs. Superpixel constraints are further utilized to provide feedback information to the subspace learning for avoiding the "salt-and-pepper" problem generated in the pixel-based classification methods.

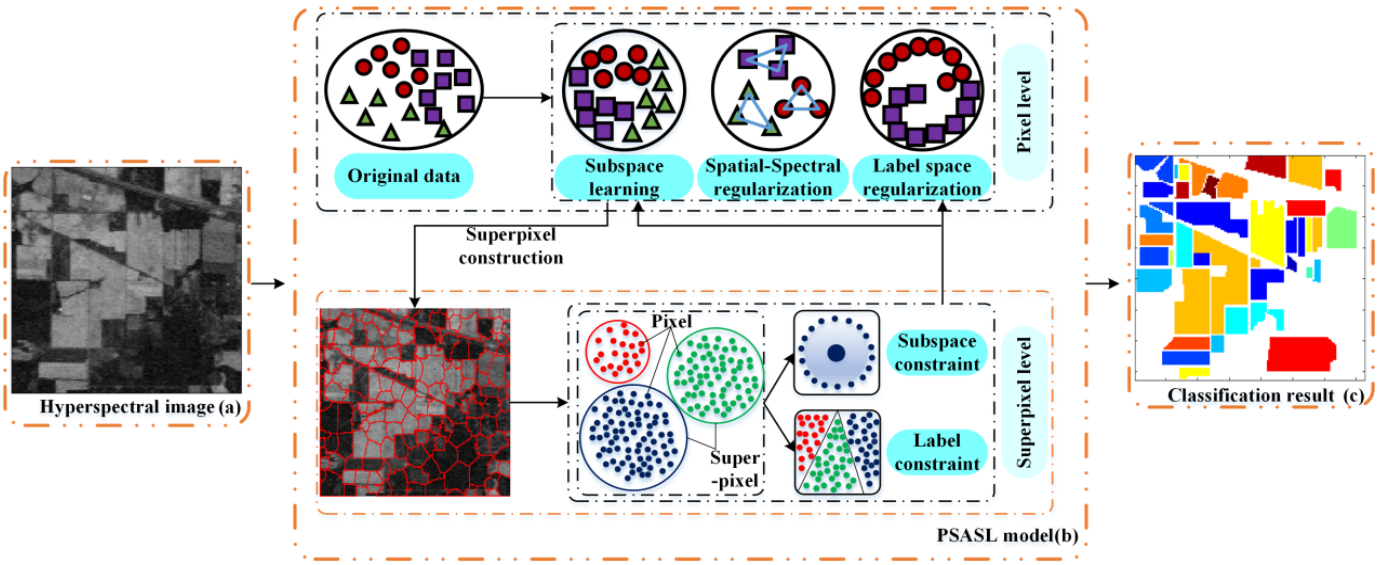



Fig. 1. Workflow of the PSASL for HSI classification. (a) Training HSI. (b) Learning process of the PSASL. (c) Classification results.

TABLE I
NOTATIONS AND DEFINITIONS

| Notation | Definition |
|---|---|
| X | Hyperspectral image pixels. |
| $\mathbf{Z}$ | The matrix of the original feature. $\mathbf{Z}\in\mathbb{R}^{d\times n}$. |
| $\mathbf{W}$ | The subspace learning matrix. $\mathbf{W}\in\mathbb{R}^{d\times d'}$. |
| $\mathbf{U}$ | The weight matrix. |
| $\mathbf{V}$ | The weight matrix. |
| $\mathbf{Y}$ | The label matrix. $\mathbf{Y}\in\mathbb{R}^{n\times c}$ |
| $\mathbf{F}$ | The predicted probability matrix. $\mathbf{F}\in\mathbb{R}^{n\times c}$ |
| $\mathbf{G}$ | The adaptive graph. |
| $\mathbf{H}$ | The projection matrix |
| $\mathbf{L}$ | The Laplacian matrix. |
| $\mathbf{I}$ | The identity matrix. |
| $\alpha,\beta,\gamma,\delta_s$ | They are used to balance the importance of the corresponding term. |
| $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ | They are used to balance the importance of the corresponding term. |
| $d$ | The original number of bands. |
| $\mathbf{b}$ | The bias. |
| $n$ | The number of image pixels. |
| $c$ | The number of classes. |
| $p_i$ | The location of image pixel $x_i$. |
| $\mathcal{N}_k(x_j)$ | The k-nearest neighbors of image pixel $x_j$. |
| $K$ | The number of chosen neighbors. |
| $\|\bullet\|_F$ | Frobenius norm. |
| $tr(\bullet)$ | The trace of the matrix. |

3) The solution to the objective function is efficiently achieved by employing a customized iterative algorithm, and it converges very fast.

For clarity, we illustrate important notations and definitions throughout this paper in Table I.

## II. PIXEL-LEVEL AND SUPERPIXEL-LEVEL AWARE SUBSPACE LEARNING

In this section, we first construct the model of subspace learning. Then, we integrate the constraints of spectral–spatial graph and label space in the pixel-level into subspace learning. Feedback information of the superpixel-level is provided for subspace learning in the pixel-level.

### A. Definition of Subspace Learning

The pixels $\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$ of an HSI is represented by the matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n] \in \mathbb{R}^{d\times n}$, where $d$ denotes the dimension of spectral vector (number of bands) and $n$ represents the number of different image pixels. To remove the redundant information, PSASL minimizes the following objective function based on the reconstruction ICA algorithms [3], [12]:

$$\Theta_1(\mathbf{W}) = \|\mathbf{W}\mathbf{W}^T\mathbf{Z} - \mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z}) \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d'}$ is the subspace learning matrix, which projects $\mathbf{Z}$ to a $d'$-dimensional feature space ($d' < d$). $\alpha$ is a tradeoff factor. $g$ is the nonlinear convex function which is defined as a smooth $l_1$ function, such as bnn[2], $\cosh(\mathbf{W}^T\mathbf{Z}) = (\exp(\mathbf{W}^T\mathbf{Z}) + \exp(-\mathbf{W}^T\mathbf{Z}))/2$. The function of cosh unites log function describing the penalty of the approximated orthonormal constraint.

By means of the first term in (1), discriminative features in the low-dimensional space can be obtained according to the error minimization between the reconstructed data and the original data.

### B. Pixel-Level Constraints

Subspace learning only considers the information of single pixels and ignores the relationships among the neighboring pixels. For example, if the pixels are close or have the similar data distributions, they have the higher probabilities belonging to the same class. To make the feature of a pixel more representative and discriminative in the subspace, the pixel-level constraint is considered in the subspace learning process, which mainly includes two components: spectral–spatial graph regularization and label space regularization.

*1) Spectral–Spatial Graph Regularization:* From (1), it is noted that the original redundant data can be reduced to

low-dimensional data. For two pixels $x_i$ and $x_j$ in an HSI, they are expected to satisfy the two following properties.

*Property I:* In terms of spectral information, if pixels $x_i$ and $x_j$ have similar data distributions, they have similar feature structures in a low-dimensional subspace.

*Property II:* In terms of spatial information, if the spatial distance between$x_i$ and $x_j$ is close, they have similar feature representations in a low-dimensional subspace.

Through the constraints of the spectral and spatial information in the HSI, the subspace learning has the ability to compress the highly correlated bands for making the feature representation more discriminative.

Inspired by Property I, an adaptive graph $\mathbf{G}_E$ is introduced to describe the relationships among the pixels during the subspace learning. As [4] and [5], the graph is constructed by selecting the nearest neighbors and defining the similarities among pixels. In $\mathbf{G}_E$, each vertex corresponds to one pixel $x_i$, and the nearest neighbors are selected according to the weight matrix $\mathbf{U}$. $\mathbf{U}$ is defined using the following function:

$$\mathbf{U}_{ij} = \begin{cases} \exp(-r_{ij}^2) & x_i \in \mathcal{N}_{k_1}(x_j) \text{ or } x_j \in \mathcal{N}_{k_1}(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$r_{ij} = \|\mathbf{W}^T\mathbf{z}_i - \mathbf{W}^T\mathbf{z}_j\|_2 = \sqrt{(\mathbf{z}_i - \mathbf{z}_j)^T(\mathbf{W}\mathbf{W}^T)(\mathbf{z}_i - \mathbf{z}_j)} \tag{3}$$

where $\mathcal{N}_{k_1}(x_j)$ denotes the $k_1$-nearest neighbors of the pixel $x_j$ according to the distance $r_{ij}$. From (2), it is observed that the similarities among pixels are computed according to the subspace learning matrix $\mathbf{W}$.

After $\mathbf{G}_E$ is constructed, we encode the learned pixel features which simultaneously preserve the local visual similarity among different pixels and satisfy the manifold assumption [6]. The corresponding objective function is defined as

$$\begin{aligned} \Theta_{21}(\mathbf{W}) &= \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{U}_{ij}\|(\mathbf{W}^T\mathbf{Z})_i - (\mathbf{W}^T\mathbf{Z})_j\|_2^2 \\ &= \text{tr}((\mathbf{W}^T\mathbf{Z})\mathbf{L}_E(\mathbf{W}^T\mathbf{Z})^T) \end{aligned} \tag{4}$$

where $\mathbf{L}_E = \mathbf{D}_E - \mathbf{U}$, $\mathbf{D}_E$ is a diagonal matrix of which the $(i, i)$th element equals to the sum of the $i$th row of $\mathbf{U}$.

Inspired by Property II, we introduced a graph $\mathbf{G}_A$ to express the spatial relationships among pixels. Like $\mathbf{G}_E$, $\mathbf{G}_A$ is also constructed by selecting the nearest neighbors and defining the similarities among pixels. Different from $\mathbf{G}_E$, the weight matrix $\mathbf{V}$ in $\mathbf{G}_A$ is defined using the following function:

$$\mathbf{V}_{ij} = \begin{cases} \exp\left(-\dfrac{\|P_i - P_j\|_2^2}{\sigma}\right) & x_i \in \mathcal{N}_{k_2}(x_j) \text{ or } x_j \in \mathcal{N}_{k_2}(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $P_i$ represents the coordinate of the pixel $x_i$. $\mathcal{N}_{k_2}(x_j)$ denotes the $k_2$-nearest neighbors of the pixel $x_j$ according to the spatial distance $\|P_i - P_j\|_2$. Through the constructed graph $\mathbf{G}_A$, the manifold approach is combined into the subspace learning by means of the following function:

$$\begin{aligned} \Theta_{22}(\mathbf{W}) &= \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{V}_{ij}\|(\mathbf{W}^T\mathbf{Z})_i - (\mathbf{W}^T\mathbf{Z})_j\|_2^2 \\ &= \text{tr}((\mathbf{W}^T\mathbf{Z})\mathbf{L}_A(\mathbf{W}^T\mathbf{Z})^T) \end{aligned} \tag{6}$$

where $\mathbf{L}_A = \mathbf{D}_A - \mathbf{V}$, $\mathbf{D}_A$ is a diagonal matrix of which the $(i, i)$th element equals to the sum of the $i$th row of $\mathbf{V}$.

Combining (4) and (6), the spectral–spatial graph regularization is formed as the following objective function:

$$\begin{aligned} \Theta_2(\mathbf{W}) &= \frac{1}{2}\sum_{i,j=1}^{n} (\mathbf{U}_{ij} + \beta\mathbf{V}_{ij})\|(\mathbf{W}^T\mathbf{Z})_i - (\mathbf{W}^T\mathbf{Z})_j\|_2^2 \\ &= \text{tr}((\mathbf{W}^T\mathbf{Z})(\mathbf{L}_E + \beta\mathbf{L}_A)(\mathbf{W}^T\mathbf{Z})^T). \end{aligned} \tag{7}$$

*2) Label Space Regularization:* For label space regularization, two pixels $x_i$ and $x_j$ in an HSI are expected to satisfy the two following properties.

*Property III:* If $x_i$ and $x_j$ have similar data distribution, they have the same labels in a low-dimensional subspace.

*Property IV:* If $x_i$ and $x_j$ have the same labels, their labels in a low-dimensional subspace are consistent with those in the original label space.

In light of the two properties III and IV, it is reasonable to assume that the intraclass pixels of the original high-dimensional feature space have more similar data structures in a low-dimensional subspace, and the interclass pixels have larger variations. For the semisupervised learning, the labels of some pixels in training data should be given in the HSIs.

Considering the above conditions, manifold smoothness with pixel labels is used to regularize the dimensionality reduction by jointly learning the relevance scores among different image pixels and subspace learning matrix $\mathbf{W}$. The objective function is as follows:

$$\begin{aligned} &\Theta_3(\mathbf{F}, \mathbf{W}) \\ &= \frac{1}{2}\sum_{i,j=1}^{n} (\mathbf{U}_{ij} + \beta\mathbf{V}_{ij})\|\mathbf{F}_i - \mathbf{F}_j\|_2^2 + \sum_{i=1}^{n}\|\mathbf{F}_i - \mathbf{Y}_i\|_2^2 \\ &= \text{tr}(\mathbf{F}^T(\mathbf{L}_E + \beta\ \mathbf{L}_A)\mathbf{F}) + \text{tr}((\mathbf{F} - \mathbf{Y})^T\mathbf{S}(\mathbf{F} - \mathbf{Y})) \end{aligned} \tag{8}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the diagonal matrix, in which the labeled pixel samples $\mathbf{S}_{ii} = 1$; otherwise, $\mathbf{S}_{ii} = 0$. $\mathbf{F}$ and $\mathbf{Y}$ indicate the predicted probability and the known labels of pixels. $\mathbf{Y} \subset \mathbb{R}^{n\times c}$ and $\mathbf{Y}_{ij}$ indicate the $j$th data of $\mathbf{Y}_i$. $\mathbf{Y}_{ij} = 1$ if $\mathbf{x}_i$ belongs to the $j$th class; otherwise, $\mathbf{Y}_{ij} = 0$. In the first term of (8), the classification scores and dimensionality reduction matrix can be optimized simultaneously. The second term introduces the actual label data integrating the first term to learn the classification results with local and global consistencies. Therefore, the neighboring pixels have the same labels and the pixel classification results are consistent with the given image pixel labels.

### C. Superpixel-Level Constraints

In order to avoid the "salt-and-pepper" problem generated in the pixel-based classification and offer the feedback information about the quality of subspace learning, superpixel-level constraints are introduced to further refine the HSI

classification results. Pixels $x_i$ and $x_j$ in the same superpixel with class purity are expected to satisfy the following property.

*Property V:* $x_i$ and $x_j$ have similar data representation and labeling information.

In light of property V, we should keep the superpixels with class purity from the HSI.

For an HSI, we generate $m$ superpixels $Q=\{q_1, q_2, \ldots, q_m\}$ using the adaptive mean shift-clustering algorithm [32]. Each pixel in a superpixel has similar spectral information, which can be represented by the cluster center of the pixels. Consider that the pixels contained in a superpixel have the same labels, the objective function shown in (9) refines the subspace learning matrix $\mathbf{W}$ from the superpixel-level, and the label matrix $\mathbf{F}$ defined in **the** pixel-level

$$\Theta_4(\mathbf{W},\mathbf{F})=\sum_{i=1}^{n}\exp\left(\frac{\left\|\mathbf{W}^T\mathbf{z}_i-\mu_t\right\|_2^2}{\delta_s}\right)+\gamma\ \|\mathbf{F}-\mathbf{F}_s\|_F^2 \quad (9)$$

where $\mu_t$ is the center of the superpixel that $\mathbf{z}_i$ belongs to, and $(F_s)_i$ is the statistical label information of the superpixel of $x_i$.

Equation (9) is utilized to provide the constraints of subspace learning and labeling information for the HSI classification, which can minimize the average "impurity" of the class distribution of the pixels in each superpixel in subspace learning. In this procedure, we name "optimal superpixel." The choice of the superpixel, thus, attempts to find a consistent overall segmentation, in which each segment contains pixels belonging to only one of the learned categories. This simple method allows us to consider full families of segmentation components, rather than a unique and predetermined segmentation. Once trained, the superpixel generation procedure is parameter-free and requires no thresholds.

At the beginning, the purity of the superpixels is low due to the insufficient learning of the subspace. As the number of the iterations increases, the feedback information from the superpixels to the subspace learning is more accurate. In the optimal stage, more accurate data representation is achieved in the procedure of subspace learning. As the superpixels with class purity are derived, the performance of HSI classification is, thus, enhanced.

### *D. Out-of-Sample Extension*

From (8), the classification results only reflect the probabilities of the vertices in $\mathbf{G}_E$ and $\mathbf{G}_A$ belonging to a certain class. Thus, we only obtain the classification results of the pixels that are vertices of the graphs. For the new image pixels called out-of-sample data, we need to add them into the graphs. Reconstruction of $\mathbf{G}_E$ and $\mathbf{G}_A$ is very time-consuming. In order to classify out-of-sample data efficiently, the linear regression is utilized to transform $\mathbf{F}$ into the linear classifier

$$\Theta_5(\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b})=\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}\|_F^2 \quad (10)$$

where $\mathbf{1}\in\mathbb{R}^{n\times 1}$ represent the elements that are equal to 1. We use the projection matrix $\mathbf{H}$ and bias $\mathbf{b}\in\mathbb{R}^{c\times 1}$ to classify the out-of-sample pixels.

Considering the above constraints, the final objective function that the subspace learning model integrates the pixel-level regularization and superpixel-level regularization is defined as

$$\min_{\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}}\ \Theta_1(\mathbf{W})+\lambda_1\Theta_2(\mathbf{W})+\lambda_2\Theta_3(\mathbf{F},\mathbf{W})+\lambda_3\Theta_4(\mathbf{W},\mathbf{F})+\lambda_4\Theta_5(\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}). \quad (11)$$

Then, the final objective function is formed as follows:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}}\ \Theta(\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b})\\ &=\min_{\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}}\|\mathbf{W}\mathbf{W}^T\mathbf{Z}-\mathbf{Z}\|_F^2+\alpha g(\mathbf{W}^T\mathbf{Z})\\ &\quad+\lambda_1\mathrm{tr}((\mathbf{W}^T\mathbf{Z})(\mathbf{L}_E+\beta\mathbf{L}_A)(\mathbf{W}^T\mathbf{Z})^T)\\ &\quad+\lambda_2(\mathrm{tr}(\mathbf{F}^T(\mathbf{L}_E+\beta\mathbf{L}_A)\mathbf{F})+\mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{S}(\mathbf{F}-\mathbf{Y})))\\ &\quad+\lambda_3\left(\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right)+\gamma\|\mathbf{F}-\mathbf{F}_s\|_F^2\right)\\ &\quad+\lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}\|_F^2 \quad (12)\end{aligned}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are constants to balance the corresponding terms.

## III. Optimization Algorithm

Given the nonlinear optimization in (12), solving the variables $\mathbf{W}$, $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{b}$ simultaneously is intractable by directly applying gradient descent or Newton's method due to the highly nonlinear nature of $\Theta$, which makes the gradient and the Hessian difficult to compute. In this paper, we adopt a customized iterative algorithm (Algorithm 1) to optimize the variables in the PSASL, which is in the spirit of the least-squares quantization [7]. In each iteration of the algorithm, the variables $\mathbf{W}$, $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{b}$ are optimized sequentially by taking the other three variables as constants. In this setting, $\Theta$ can be regarded as the linear function of $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{b}$, respectively, so that it can reach the minimum with respect to $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{b}$ without calculating the current gradient. Although the optimization of $\mathbf{W}$ is still nonlinear, the dimension and the difficulty of the optimization of $\mathbf{W}$ are significantly reduced compared with a gradient descent scheme that optimizes the four variables simultaneously. In the numerical tests, we find that the iterations always converge. As the unknown $\mathbf{H}$ and $\mathbf{b}$ are associated with the unknown $\mathbf{W}$ and $\mathbf{F}$, respectively, we only choose the starting points of $\mathbf{W}$ and $\mathbf{F}$ such that their values are close to the minimum. The initial guess of $\mathbf{W}$ is done by minimizing the first term in $\Theta_1$, while the initial values of $\mathbf{F}$ are set by the KNN algorithm. The convergence will stop once the number of iterations is larger than 30 or $|\Theta_t-\Theta_{t-1}|/\Theta_{t-1}<0.001$, where $\Theta_t$ is the value of the objective function in the $t$th iteration. With some algebra, the updating schedule is described in Sections III-A and III-B.

### *A. Optimization for H*

$\mathbf{H}$ is solved when $\mathbf{W}$, $\mathbf{F}$, and $\mathbf{b}$ are fixed. The optimization problem defined in (12) is written as follows:

$$\min_{\mathbf{H}}\Theta(\mathbf{H})=\min_{\mathbf{H}}\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}\|_F^2. \quad (13)$$

Equation (13) presents an unconstrained optimization problem. Assuming $\mathbf{A}=\mathbf{W}^T\mathbf{Z}$, we obtain the derivate of (13) with respect to $\mathbf{H}$, i.e.,

$$\frac{\partial\Theta(\mathbf{H})}{\partial\mathbf{H}}=2\mathbf{A}(\mathbf{A}^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}). \quad (14)$$

If $\mathbf{AA}^T$ is a singular square matrix, $\mathbf{B} = (\mathbf{AA}^T + \mu\mathbf{I})^{-1}$; otherwise, $\mathbf{B} = (\mathbf{AA}^T)^{-1}$, where $\mu$ is a small positive constant, and $\mathbf{I}$ is the identity matrix. By setting the derivative $\partial\Theta(\mathbf{H})/\partial\mathbf{H} = 0$, we obtain the following equation:

$$\mathbf{H} = \mathbf{B}(\mathbf{AF} - \mathbf{A1b}^T). \tag{15}$$

### *B. Optimization for b*

$\mathbf{b}$ is solved when $\mathbf{W}$, $\mathbf{F}$, and $\mathbf{H}$ are fixed. The optimization problem defined in (12) is rewritten as follows:

$$\min_{\mathbf{b}} \Theta(\mathbf{b}) = \min_{\mathbf{b}} \|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H} + \mathbf{1b}^T - \mathbf{F}\|_F^2. \tag{16}$$

Equation (16) presents an unconstrained optimization problem for optimizing $\mathbf{b}$. Based on some notations in the optimization for $\mathbf{H}$, we have the following:

$$\mathbf{b} = \frac{1}{n}(\mathbf{F}^T\mathbf{1} - \mathbf{H}^T\mathbf{A1}). \tag{17}$$

### *C. Optimization for F*

When the values of $\mathbf{W}$, $\mathbf{H}$, and$\mathbf{b}$ are fixed, (12) is written as

$$\begin{aligned}\min_{\mathbf{F}} \Theta(\mathbf{F}) = \min_{\mathbf{F}} \lambda_2(\mathrm{tr}(\mathbf{F}^T\mathbf{LF}) + \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{S}(\mathbf{F}-\mathbf{Y})))\\ + \lambda_3\gamma\|\mathbf{F}-\mathbf{F}_s\|_F^2 + \lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1b}^T-\mathbf{F}\|_F^2\end{aligned} \tag{18}$$

where $\mathbf{L} = \mathbf{L}_E + \beta\mathbf{L}_A$. By substituting the expression for $\mathbf{H}$ in (15) into (18), we obtain

$$\begin{aligned}\Theta(\mathbf{F}) = \lambda_2(\mathrm{tr}(\mathbf{F}^T\ \mathbf{LF}) + \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{S}(\mathbf{F}-\mathbf{Y})))\\ + \lambda_3\gamma\|\mathbf{F}-\mathbf{F}_s\|_F^2 + \lambda_4\|(\mathbf{A}^T\mathbf{BA}-\mathbf{I})(\mathbf{F}-\mathbf{1b}^T)\|_F^2.\end{aligned} \tag{19}$$

It is an unconstrained optimization problem. The derivative of (19) with respect to $\mathbf{F}$ is set to 0, then

$$\frac{\partial L(\mathbf{F})}{\partial\mathbf{F}} = 2\lambda_2(\mathbf{LF}+\mathbf{SF}-\mathbf{SY}) + 2\lambda_3\gamma(\mathbf{F}-\mathbf{F}_s) + 2\lambda_4\mathbf{CF} \tag{20}$$

where $\mathbf{C} = (\mathbf{A}^T\mathbf{BA}-\mathbf{I})^T(\mathbf{A}^T\mathbf{BA}-\mathbf{I})$. We have

$$\mathbf{F} = \left(\mathbf{L}+\mathbf{S}+\frac{\lambda_3\gamma}{\lambda_2}\mathbf{I}+\frac{\lambda_4}{\lambda_2}\mathbf{C}\right)^{-1}\left(\mathbf{SY}+\frac{\lambda_3\gamma}{\lambda_2}\mathbf{F}_s\right). \tag{21}$$

### *D. Optimization for W*

When the values of $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{b}$ are fixed, we rewrite (12) as follows:

$$\begin{aligned}&\min_{\mathbf{W}} \Theta(\mathbf{W})\\ &= \min_{\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}} \|\mathbf{WW}^T\mathbf{Z}-\mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z})\\ &\quad + \lambda_1\mathrm{tr}((\mathbf{W}^T\mathbf{Z})(\mathbf{L}_E+\beta\mathbf{L}_A)(\mathbf{W}^T\mathbf{Z})^T)\\ &\quad + \lambda_2\mathrm{tr}(\mathbf{F}^T(\mathbf{L}_E+\beta\ \mathbf{L}_A)\mathbf{F}) + \lambda_3\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right)\\ &\quad + \lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1b}^T-\mathbf{F}\|_F^2.\end{aligned} \tag{22}$$

The objective function defined in (22) is minimized through the unconstrained optimizer (e.g., L-BFGS or CG) to update $\mathbf{W}$. However, the update efficiency should not be overlooked. Aiming to enhance the performance of Algorithm 1, the auxiliary matrix $\mathbf{M}$ is introduced to separate (22). Then, (22) is transformed into a new style

$$\begin{aligned}\min_{\mathbf{W},\mathbf{M}} \Theta(\mathbf{W},\mathbf{M}) = \min_{\mathbf{W},\mathbf{M}} &\|\mathbf{WW}^T\mathbf{Z}-\mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z})\\ &+ \lambda_1\mathrm{tr}((\mathbf{M}^T\mathbf{Z})((\mathbf{L}_E)_\mathbf{M}+\beta\mathbf{L}_A)(\mathbf{M}^T\mathbf{Z})^T)\\ &+ \lambda_2\mathrm{tr}(\mathbf{F}^T((\mathbf{L}_E)_\mathbf{M}+\beta\mathbf{L}_A)\mathbf{F})\\ &+ \lambda_3\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right)\\ &+ \lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1b}^T-\mathbf{F}\|_F^2\\ \text{s.t. } \mathbf{W} = \mathbf{M}&\end{aligned} \tag{23}$$

where $(\mathbf{L}_E)_\mathbf{M}$ denotes the Laplacian matrix $\mathbf{L}_E$ constructed by the matrix $\mathbf{M}$.

Equation (23) is solved by means of the linearized alternating direction method with adaptive penalty (LADMAP [8]. The augmented Lagrangian function of (23) is

$$\begin{aligned}L(\mathbf{W},\mathbf{M},\mathbf{T},\nu) = &\|\mathbf{WW}^T\mathbf{Z}-\mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z})\\ &+ \lambda_1\mathrm{tr}((\mathbf{M}^T\mathbf{Z})((\mathbf{L}_E)_\mathbf{M}+\beta L_A)(\mathbf{M}^T\mathbf{Z})^T)\\ &+ \lambda_2\mathrm{tr}(\mathbf{F}^T((\mathbf{L}_E)_\mathbf{M}+\beta L_A)\mathbf{F})\\ &+ \lambda_3\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right)\\ &+ \lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1b}^T-\mathbf{F}\|_F^2\\ &+ \langle\mathbf{T},\mathbf{W}-\mathbf{M}\rangle + \frac{\nu}{2}\|\mathbf{W}-\mathbf{M}\|_F^2\end{aligned}$$

**Algorithm 1** PSASL

**Input:** Training set $\mathbf{Z}$, parameters $\alpha$, $\beta$, $\gamma$, $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$;

**Initialization**: $\mathbf{W}$; $\nu_0 = 0.1$, $\nu_{max} = 10^{10}$, $\rho = 1.1$; $t_1 = 0$, $t_2 = 0$

**While** stopping criterion is not met **do**

**1**: **While** stopping criterion is not met **do**

**1.1**: Update $\mathbf{W}$ by Eq. (25) with the unconstrained optimizer;

**1.2**: Update $\mathbf{M}$ by Eq. (28) with the unconstrained optimizer ;

**1.3**: Update Lagrange multipliers by $(\mathbf{T})_{new} = (\mathbf{T})_{old} + \nu(\mathbf{W}-\mathbf{M})$;

**1.4**: Update $\nu$ by $\nu = \min(\rho\nu, \max\nu)$;

**1.5**: Update $t_1 = t_1 + 1$;

**2**: Construct the graphs $\mathbf{G}_\mathrm{E}$ and $\mathbf{G}_\mathrm{A}$, and calculate $\mathbf{L}_\mathrm{E}$ and $\mathbf{L}_\mathrm{A}$;

**3**: Update $\mathbf{F}$ by $\mathbf{F} = \left(\mathbf{L}+\mathbf{S}+\frac{\lambda_3\gamma}{\lambda_2}I+\frac{\lambda_4}{\lambda_2}\mathbf{C}\right)^{-1}\left(\mathbf{SY}+\frac{\lambda_3\gamma}{\lambda_2}F_s\right)$;

**4**: Update $\mathbf{H}$ by $\mathbf{H} = \mathbf{B}\left(\mathbf{AF}-\mathbf{A1b}^T\right)$;

**5**: Update $\mathbf{b}$ by $\mathbf{b} = \frac{1}{n}\left(\mathbf{F}^T\mathbf{1}-\mathbf{H}^T\mathbf{A1}\right)$;

**6**: Update $t_2 = t_2 + 1$;

**7**: Obtain the optimal solution $\mathbf{W}$, $\mathbf{F}$, $\mathbf{H}$ and $\mathbf{b}$.

**Output:** The matrix $\mathbf{F}$, $\mathbf{H}$ and $\mathbf{b}$.

$$
\begin{aligned}
&= \|\mathbf{W}\mathbf{W}^T\mathbf{Z}-\mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z}) \\
&\quad + \lambda_1 \mathrm{tr}((\mathbf{M}^T\mathbf{Z})((\mathbf{L}_E)_{\mathbf{M}}+\beta L_A)(\mathbf{M}^T\mathbf{Z})^T) \\
&\quad + \lambda_2 \mathrm{tr}(\mathbf{F}^T((\mathbf{L}_E)_{\mathbf{M}}+\beta L_A)\mathbf{F}) \\
&\quad + \lambda_3 \sum_{i=1}^{n} \exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right) \\
&\quad + \lambda_4 \|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}\|_F^2 \\
&\quad + \frac{\nu}{2}\left\|\mathbf{W}-\mathbf{M}+\frac{\mathbf{T}}{\nu}\right\|_F^2 - \frac{1}{2\nu}\|\mathbf{T}\|_F^2 \qquad (24)
\end{aligned}
$$

where $\mathbf{T}$ is the Lagrangian multiplier, and $\nu > 0$ is a penalty parameter. Equation (24) describes an unconstrained problem, and it can be optimized with respect to $\mathbf{W}$ and $\mathbf{M}$ by fixing other variables. With some algebra, the updating schedule is described in the following section.

For updating $\mathbf{W}$, by considering other variants as the constants, (24) can be rewritten as

$$
\begin{aligned}
\min_{\mathbf{W}} L(\mathbf{W}) &= \min_{\mathbf{W}} \|\mathbf{W}\mathbf{W}^T\mathbf{Z}-\mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z}) \\
&\quad + \lambda_3 \sum_{i=1}^{n} \exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right) \\
&\quad + \lambda_4 \|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F}\|_F^2 \\
&\quad + \frac{\nu}{2}\left\|\mathbf{W}-\mathbf{M}+\frac{\mathbf{T}}{\nu}\right\|_F^2. \qquad (25)
\end{aligned}
$$

The derivative of (25) with respect to $\mathbf{W}$ is computed, and then, we have the following result:

$$
\begin{aligned}
\frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} &= 2(\ \mathbf{W}\mathbf{W}^T\mathbf{Z}\mathbf{Z}^T + \mathbf{Z}\mathbf{Z}^T\mathbf{W}\mathbf{W}^T - 2\mathbf{Z}\mathbf{Z}^T)\mathbf{W} \\
&\quad + \alpha\frac{\partial g(\mathbf{W}^T\mathbf{Z})}{\partial \mathbf{W}} + 2\frac{\lambda_3}{\delta_s}\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right) \\
&\quad \times \left(z_i\mathbf{z}_i^T W - \mathbf{z}_i\mu_t^T\right) \\
&\quad + 2\lambda_4\mathbf{Z}((\mathbf{W}^T\mathbf{Z})^T\ \mathbf{H}+\mathbf{1}\mathbf{b}^T-\mathbf{F})\mathbf{H}^T \\
&\quad + \nu\left(\mathbf{W}-\mathbf{M}+\frac{\mathbf{T}_1}{\nu}\right) \qquad (26)
\end{aligned}
$$

where

$$
\frac{\partial g(\mathbf{W}^T\mathbf{Z})}{\partial \mathbf{W}_{ij}} = \sum_{k=1}^{n}\tanh\left(\mathbf{W}_i^T\mathbf{z}_k\right)\mathbf{z}_{jk}. \qquad (27)
$$

Given a training data matrix $\mathbf{Z}$, we compute the function cost of (25) and the gradient using (26). Then, the objective function defined in (25) is minimized through the unconstrained optimizer (e.g., L-BFGS) to update $\mathbf{W}$.

For updating $\mathbf{M}$, by considering other variants as the constants, (25) can be rewritten as

$$
\begin{aligned}
\min_{\mathbf{M}} L(\mathbf{M}) &= \min_{\mathbf{M}} \lambda_1 \mathrm{tr}((\mathbf{M}^T\mathbf{Z})((\mathbf{L}_E)_{\mathbf{M}}+\beta L_A)(\mathbf{M}^T\mathbf{Z})^T) \\
&\quad + \lambda_2 \mathrm{tr}(\mathbf{F}^T((\mathbf{L}_E)_{\mathbf{M}}+\beta L_A)\mathbf{F}) \\
&\quad + \frac{\nu}{2}\|\mathbf{W}-\mathbf{M}+\frac{\mathbf{T}_1}{\nu}\|_F^2 \\
&= \min_{\mathbf{M}} \frac{\lambda_1}{2}\sum_{i,j=1}^{n}(\mathbf{U}_{ij}+\beta V_{ij})\|(\mathbf{M}^T\mathbf{Z})_i-(\mathbf{M}^T\mathbf{Z})_j\|_2^2 \\
&\quad + \frac{\lambda_2}{2}\sum_{i,j=1}^{n}(\mathbf{U}_{ij}+\beta V_{ij})\|\mathbf{F}_i-\mathbf{F}_j\|_2^2 \\
&\quad + \frac{\nu}{2}\left\|\mathbf{W}-\mathbf{M}+\frac{\mathbf{T}_1}{\nu}\right\|_F^2. \qquad (28)
\end{aligned}
$$

In (28), we use the chain rule to compute the derivative of $\mathbf{M}$. The objective function defined in (28) is minimized through the unconstrained optimizer to update $\mathbf{M}$.

The Lagrangian multiplier is updated as follows:

$$
(\mathbf{T})_{\text{new}} = (\mathbf{T})_{\text{old}} + \nu(\mathbf{W}-\mathbf{M}). \qquad (29)
$$

TABLE II
LAND COVER CLASSES WITH SAMPLES NUMBER FOR THE INDIAN PINES DATA SET

| Class | Land Cover Type | Semi (20 labeled samples) | | Out of Sample s |
|---|---|---|---|---|
| | | Labeled | Unlabeled | Testing |
| 1 | Corn-notill | 20 | 408 | 1000 |
| 2 | Corn-mintill | 20 | 229 | 581 |
| 3 | Grass-pasture | 20 | 125 | 338 |
| 4 | Grass-trees | 20 | 199 | 511 |
| 5 | Hay-windrowed | 20 | 123 | 335 |
| 6 | Soybean-notill | 20 | 272 | 680 |
| 7 | Soybean-mintill | 20 | 717 | 1719 |
| 8 | Soybean-clean | 20 | 158 | 415 |
| 9 | Woods | 20 | 360 | 886 |
| 10 | Buildings-Grass-Trees-Drives | 20 | 96 | 270 |

## IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed method PSASL for HSI classification. We first briefly describe the used HSI data. Afterward, we compare the classification results of the PSASL.

### A. Experimental Data Sets

Three HSI data sets are used to evaluate the performance of the proposed method.

The first data set is the Indian Pines data set which was gathered by AVIRIS sensor over the Indian Pines test site of North-Western Indiana in 1992. It consists of $145 \times 145$ pixels and 224 spectral reflectance bands in the wavelength range 0.4–2.5 $\mu$m with a spatial resolution of 20 m. The bands covering the region of water absorption (104–108, 150–163, and 220) are removed and hence, 200 out of the 224 bands are preserved. It contains 10 classes and 9620 labeled pixels. The detailed information about the data is listed in Table II.

The second data set is the Salinas data set, which was collected by the 224-band AVIRIS sensor over Salinas Valley, CA, USA. The image size is $512 \times 217$ pixels and is characterized by high spatial resolution (3.7-m pixels). As with Indian Pines data set, 20 water absorption bands (108–112, 154–167, and 224) out of 224 bands are discarded; thus, 204 bands are used in our experiment. The Salinas data set contains 16 classes and 54 129 labeled pixels, as shown in Table III.

TABLE III
LAND COVER CLASSES WITH SAMPLES NUMBER FOR THE SALINAS DATA SET

| Class | Land Cover Type | Semi (20 labeled samples) | | Out of Samples |
|---|---|---|---|---|
| | | Labeled | Unlabeled | Testing |
| 1 | Brocoli_green_weeds_1 | 20 | 181 | 1808 |
| 2 | Brocoli_green_weeds_2 | 20 | 353 | 3353 |
| 3 | Fallow | 20 | 178 | 1778 |
| 4 | Fallow_rough_plow | 20 | 119 | 1255 |
| 5 | Fallow_smooth | 20 | 248 | 2410 |
| 6 | Stubble | 20 | 376 | 3563 |
| 7 | Celery | 20 | 338 | 3221 |
| 8 | Grapes_untrained | 20 | 1107 | 10144 |
| 9 | Soil_vinyard_develop | 20 | 600 | 5583 |
| 10 | Corn_senesced_green_weeds | 20 | 308 | 2950 |
| 11 | Lettuce_romaine_4wk | 20 | 87 | 961 |
| 12 | Lettuce_romaine_5wk | 20 | 173 | 1734 |
| 13 | Lettuce_romaine_6wk | 20 | 72 | 824 |
| 14 | Lettuce_romaine_7wk | 20 | 87 | 963 |
| 15 | Vinyard_untrained | 20 | 707 | 6541 |
| 16 | Vinyard_vertical_trellis | 20 | 161 | 1626 |

TABLE IV
LAND COVER CLASSES WITH SAMPLES NUMBER FOR THE PAVIA UNIVERSITY DATA SET

| Class | Land Cover Type | Semi (20 labeled samples) | | Out of Samples |
|---|---|---|---|---|
| | | Labeled | Unlabeled | Testing |
| 1 | Asphalt | 20 | 643 | 5968 |
| 2 | Meadows | 20 | 1845 | 16784 |
| 3 | Gravel | 20 | 190 | 1889 |
| 4 | Trees | 20 | 286 | 2758 |
| 5 | Painted metal sheets | 20 | 115 | 1211 |
| 6 | Bare Soil | 20 | 483 | 4526 |
| 7 | Bitumen | 20 | 113 | 1197 |
| 8 | Self-Blocking Bricks | 20 | 348 | 3314 |
| 9 | Shadows | 20 | 75 | 852 |

The third data set is the University of Pavia data set (PaviaU), which was acquired by the ROSIS-03 sensor over an urban area, Northern Italy. The spatial size is $610 \times 340$ and the geometric resolution is 1.3 m. The 12 noisy bands are removed and 103 out of the 115 bands are used in our experiment. There are nine classes in PaviaU and 42 776 labeled pixels. The details are listed in Table IV.

In Tables II–IV, only 20 labeled samples are listed. Different numbers of labeled pixels are utilized to classify HSIs.

### B. Comparisons With Other Approaches

Since the PSASL is semisupervised, we mainly compare it with the five semisupervised approaches in terms of HSI classification accuracy, i.e., flexible manifold embedding (FME) [4], linear manifold regularization for large-scale semisupervised learning (LapRLS) [9], discriminating joint feature analysis for multimedia data understanding (SFSS) [10], learning a nonnegative sparse graph for linear regression (NNSG) [11], and learning a discriminative distance metric with label consistency (DDML-LC) for scene classification [50]. We also compare with two classical supervised methods SVM [61] and SVM-MRF [62].

In the FME, LapRLS, SFSS, and DDML-LC, the number of the nearest neighbors is selected from the set $\{3, 4, 5, 6, 7, 8, 9, 10\}$. The distances of the pixel features are calculated using $\mathbf{U}_{ij} = e^{-(\|\mathbf{Z}_i - \mathbf{Z}_j\|^2/\sigma)}$ in the FME and LapRLS, where $\sigma$ is the heat kernel, and it is selected from the set $\{10^{-9}, 10^{-8}, \ldots, 10^{8}, 10^{9}\}$. In the SFSS, $\mathbf{U}_{ij} = 1$ when $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are the nearest neighbors; otherwise, $\mathbf{U}_{ij} = 0$. In the DDML-LC, the optimized feature transformation matrix is used to compute the distances between pixels by means of the Mahalanobis distance. The defined graph learning functions can give the number of the nearest neighbors and compute the pixel feature distances in the NNSG. In the PSASL, the numbers of $k_1$ and $k_2$ are chosen from the set $\{3, 4, 5, 6, 7, 8, 9, 10\}$. $\sigma$ is also selected from the set $\{10^{-9}, 10^{-8}, \ldots, 10^{8}, 10^{9}\}$. In the FME, LapRLS, SFSS, NNSG, DDML-LC, and PSASL, the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are tuned from the set $\{10^{-9}, 10^{-8}, \ldots, 10^{8}, 10^{9}\}$, respectively.

In the Indian Pines data set, we randomly select 30% samples as the semisupervised classification data, and the rest are taken as the out-of-sample data, i.e., testing data. While there are too many pixels in the Salinas and PaviaU data sets, we randomly select 10% samples as the training data and the remaining pixels are the testing data. We further randomly select $s$ samples from the semisupervised classification data as the labeled data; $s$ is set to 3, 5, 10, and 20, respectively. The remaining data are unlabeled data (noted by Unlabel). For the two supervised methods SVM and SVM-MRF, to compare with other semisupervised methods, the same amount of labeled training data are randomly selected, such as 3, 5, 10, and 20 labeled samples, and the remaining data are used as testing data.

### C. Experimental Results

We report the HSI classification accuracy over randomly split unlabeled data set and unseen test data set, which are referred to as Unlabel and Test, respectively (see Tables V–VII and Figs. 2–7). To evaluate the classification results, three metrics of overall accuracy, average accuracy, and Kappa coefficient are used.

From the results listed in Tables V–VII and illustrated in Figs. 2–7, we derive the following observations.

1) Compared with the recently proposed methods, the PSASL achieves the best results for all Unlabel and Test samples. The classification accuracies on the three data sets obtained by the PSASL are higher than those obtained by other semisupervised methods. Compared with the supervised methods SVM and SVM-MRF, the PSASL also obtain better classification accuracy of testing data. It indicates that the PSASL is very effective on HSI classification.
2) With the number of the labeled pixels increasing as shown in Figs. 2–4, the classification accuracies for all the compared classifiers are increased. In addition, higher overall classification accuracies are obtained by the PSASL with varying numbers of training samples, demonstrating that the PSASL far outperforms the other compared methods. Since the PSASL embeds the spectral–spatial graph and label space constraint into the subspace learning, the constructed graph can well reflect

TABLE V

SEMISUPERVISED HSI CLASSIFICATION RESULTS (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE INDIAN PINES DATA SET

| Class | SVM | SVM-MRF | FME | | LapRLS | | SFSS | | NNSG | | DDML-LC | | PSASL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test |
| Corn-notill | 40.20 | 52.94 | 29.66 | 34.00 | 24.51 | 28.30 | 37.25 | 32.30 | 45.10 | 50.80 | 48.47 | 62.90 | 86.03 | 88.45 |
| Corn-mintill | 69.88 | 62.33 | 3.93 | 4.13 | 44.54 | 39.93 | 36.24 | 24.27 | 40.17 | 37.01 | 56.10 | 38.55 | 94.50 | 83.96 |
| Grass-pasture | 74.73 | 92.17 | 1.02 | 3.00 | 80.00 | 82.54 | 73.60 | 39.05 | 80.00 | 74.85 | 95.07 | 64.79 | 98.11 | 97.86 |
| Grass-trees | 80.28 | 89.37 | 90.45 | 87.67 | 85.43 | 80.82 | 80.40 | 85.32 | 92.96 | 91.39 | 99.54 | 91.59 | 97.51 | 99.53 |
| Hay-windrowed | 99.34 | 93.65 | 100 | 100 | 100 | 100 | 100 | 99.70 | 100 | 96.12 | 100 | 100 | 100 | 100 |
| Soybean-notill | 79.31 | 69.72 | 43.01 | 37.06 | 56.25 | 53.38 | 46.69 | 32.35 | 53.68 | 53.24 | 77.16 | 76.18 | 92.49 | 93.11 |
| Soybean-mintill | 44.07 | 59.15 | 77.96 | 76.14 | 50.21 | 48.78 | 52.44 | 56.69 | 51.88 | 33.93 | 69.07 | 60.83 | 77.52 | 65.03 |
| Soybean-clean | 54.28 | 73.43 | 20.89 | 22.17 | 41.14 | 36.87 | 31.65 | 3.13 | 62.03 | 60.00 | 73.14 | 65.78 | 95.21 | 95.49 |
| Woods | 89.08 | 79.31 | 99.17 | 99.77 | 91.39 | 90.51 | 69.44 | 77.06 | 70.83 | 79.44 | 100 | 98.98 | 91.89 | 93.55 |
| Buildings-Grass-Trees-Drives | 62.84 | 67.13 | 2.08 | 1.48 | 46.88 | 45.93 | 43.75 | 17.78 | 60.42 | 60.74 | 100 | 30.37 | 86.93 | 81.42 |
| OA | 63.49 | 68.53 | 55.86 | 54.75 | 57.57 | 56.75 | 54.15 | 49.06 | 60.03 | 56.82 | 76.37 | 69.35 | **92.13** | **89.85** |
| AA | 69.40 | 73.92 | 46.82 | 46.54 | 62.03 | 60.71 | 57.15 | 46.77 | 65.71 | 63.75 | 81.86 | 69.00 | **92.02** | **89.84** |
| Kappa | 0.59 | 0.64 | 0.47 | 0.46 | 0.51 | 0.51 | 0.47 | 0.41 | 0.54 | 0.51 | 0.73 | 0.65 | **0.91** | **0.89** |

*The best results are highlighted in bold.*

TABLE VI

SEMISUPERVISED HSI CLASSIFICATION RESULTS (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE SALINAS DATA SET

| Class | SVM | SVM-MRF | FME | | LapRLS | | SFSS | | NNSG | | DDML-LC | | PSASL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test |
| Brocoli_green_weeds_1 | 97.79 | 97.96 | 100 | 100 | 100 | 100 | 97.79 | 100 | 100 | 99.83 | 98.62 | 99.06 | 100 | 100 |
| Brocoli_green_weeds_2 | 97.98 | 98.62 | 97.17 | 71.49 | 68.84 | 67.94 | 89.80 | 77.63 | 70.54 | 89.35 | 99.34 | 98.60 | 100 | 99.72 |
| Fallow | 88.75 | 96.75 | 37.08 | 42.35 | 83.71 | 81.33 | 85.96 | 75.59 | 66.29 | 82.28 | 78.63 | 86.33 | 100 | 100 |
| Fallow_rough_plow | 98.98 | 99.46 | 100 | 99.76 | 99.16 | 99.84 | 100 | 99.92 | 99.16 | 99.12 | 98.73 | 97.29 | 94.12 | 97.48 |
| Fallow_smooth | 95.00 | 95.46 | 98.79 | 96.18 | 91.94 | 91.37 | 84.68 | 80.33 | 97.18 | 96.68 | 98.05 | 98.63 | 84.68 | 77.42 |
| Stubble | 98.63 | 98.70 | 98.94 | 99.83 | 100 | 99.86 | 98.40 | 99.80 | 100 | 99.72 | 99.86 | 100 | 97.07 | 95.21 |
| Celery | 99.21 | 99.11 | 98.82 | 99.94 | 99.41 | 99.60 | 99.70 | 99.53 | 99.70 | 99.75 | 99.94 | 99.81 | 100 | 99.70 |
| Grapes_untrained | 71.02 | 72.39 | 63.05 | 79.26 | 58.27 | 57.92 | 57.36 | 56.16 | 65.94 | 68.06 | 76.63 | 75.48 | 89.16 | 84.82 |
| Soil_vinyard_develop | 96.47 | 96.62 | 96.50 | 99.77 | 97.83 | 98.33 | 94.17 | 92.94 | 99.67 | 99.70 | 99.87 | 98.71 | 100 | 100 |
| Corn_senesced_green_weeds | 79.01 | 92.70 | 46.10 | 6.71 | 86.69 | 83.15 | 81.49 | 77.22 | 87.99 | 88.64 | 83.66 | 77.53 | 91.88 | 90.58 |
| Lettuce_romaine_4wk | 98.19 | 98.86 | 98.85 | 94.59 | 100 | 98.23 | 97.70 | 96.98 | 97.70 | 96.46 | 95.53 | 96.46 | 98.85 | 100 |
| Lettuce_romaine_5wk | 99.58 | 99.89 | 100 | 82.18 | 69.94 | 71.63 | 83.24 | 68.05 | 86.71 | 84.60 | 85.81 | 96.31 | 91.91 | 95.38 |
| Lettuce_romaine_6wk | 97.10 | 96.94 | 95.83 | 72.69 | 100 | 100 | 98.61 | 100 | 100 | 91.75 | 96.36 | 91.99 | 94.44 | 100 |
| Lettuce_romaine_7wk | 93.24 | 95.15 | 89.66 | 82.45 | 73.56 | 69.26 | 90.80 | 43.41 | 82.76 | 88.27 | 78.40 | 77.47 | 93.10 | 97.70 |
| Vinyard_untrained | 57.73 | 69.42 | 62.23 | 46.25 | 60.54 | 59.52 | 50.78 | 44.64 | 75.25 | 53.08 | 67.71 | 54.96 | 95.19 | 91.80 |
| Vinyard_vertical_trellis | 85.67 | 98.54 | 87.58 | 81.30 | 96.89 | 96.56 | 98.14 | 96.86 | 95.03 | 95.82 | 98.34 | 98.22 | 100 | 100 |
| OA | 84.81 | 88.24 | 79.80 | 76.34 | 79.65 | 79.47 | 79.10 | 75.38 | 84.06 | 83.59 | 87.64 | 85.71 | **94.90** | **93.21** |
| AA | 90.90 | 94.16 | 85.66 | 78.42 | 86.67 | 85.91 | 88.04 | 81.82 | 88.99 | 89.57 | 90.97 | 90.43 | **95.65** | **95.61** |
| Kappa | 0.83 | 0.87 | 0.77 | 0.74 | 0.77 | 0.77 | 0.77 | 0.73 | 0.82 | 0.82 | 0.86 | 0.84 | **0.94** | **0.92** |

*The best results are highlighted in bold.*

the relationships between image pixels in HSI. Then, the PSASL achieves much better HSI classification results.

3) From Figs. 5–7, the PSASL has more compact HSI classification results on the three data sets. It also validates that the superpixel generation procedure can provide useful feedback information for the subspace learning in the pixel-level.

### *D. Independent Analysis of the Regularization Terms*

To verify the contribution of each term in the objective function (12), we compare the independent regularization (IR) and the joint regularizations for HSI classification. As follows, the joint objective function can be divided into the three learning models.

1) *IR1:* IR1 is the subspace learning, which removes the redundant information by using the reconstruction ICA algorithm

$$\min_{\mathbf{W}} \|\mathbf{W}\mathbf{W}^T\mathbf{Z} - \mathbf{Z}\|_F^2 + \alpha g(\mathbf{W}^T\mathbf{Z}). \tag{30}$$

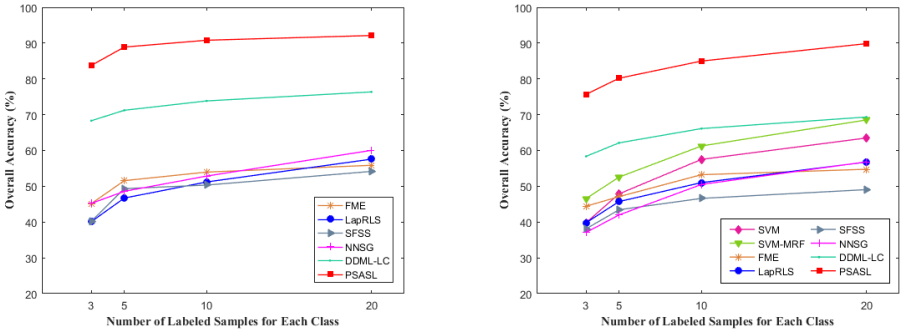
2) *IR2:* To make the subspace learning compress the highly correlated bands for making the feature representation more discriminative, IR2 adds the constraints of the spectral and spatial information in the HSI to the

TABLE VII

SEMISUPERVISED HSI CLASSIFICATION RESULTS (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE PAVIAU DATA SET

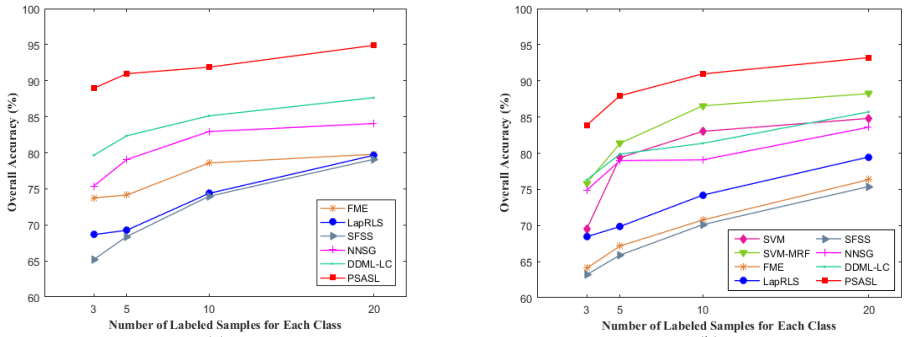
| Class | SVM | SVM-MRF | FME | | LapRLS | | SFSS | | NNSG | | DDML-LC | | PSASL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test | Unlabel | Test |
| Asphalt | 84.13 | 83.50 | 14.93 | 10.74 | 20.22 | 18.83 | 28.77 | 9.70 | 27.84 | 26.12 | 63.45 | 65.32 | 90.03 | 77.59 |
| Meadows | 69.54 | 70.46 | 79.46 | 76.69 | 51.49 | 49.67 | 50.41 | 49.27 | 72.30 | 50.13 | 76.10 | 56.37 | 68.99 | 64.32 |
| Gravel | 69.60 | 65.98 | 46.32 | 31.13 | 71.05 | 74.11 | 77.37 | 83.80 | 63.16 | 61.94 | 84.74 | 82.63 | 80.52 | 86.13 |
| Trees | 96.94 | 93.43 | 97.90 | 99.42 | 100 | 99.24 | 98.95 | 99.06 | 98.95 | 99.09 | 98.25 | 94.76 | 95.11 | 92.26 |
| Painted metal sheets | 91.32 | 94.44 | 100 | 100 | 100 | 100 | 99.13 | 100 | 100 | 100 | 100 | 100 | 100 | 99.19 |
| Bare Soil | 35.94 | 56.72 | 4.97 | 4.42 | 18.63 | 17.76 | 35.40 | 20.90 | 43.48 | 43.35 | 34.78 | 75.57 | 72.93 | 57.27 |
| Bitumen | 67.48 | 68.59 | 25.66 | 14.20 | 78.76 | 75.44 | 87.61 | 91.14 | 63.72 | 55.97 | 96.46 | 89.38 | 92.21 | 87.43 |
| Self-Blocking Bricks | 53.44 | 73.32 | 79.89 | 85.61 | 44.54 | 38.77 | 36.78 | 16.26 | 28.74 | 29.87 | 76.44 | 56.03 | 85.89 | 68.09 |
| Shadows | 80.26 | 84.39 | 97.33 | 99.77 | 100.00 | 99.65 | 98.67 | 99.88 | 90.67 | 90.61 | 100 | 97.33 | 99.01 | 98.96 |
| OA | 69.28 | 73.51 | 59.76 | 57.43 | 49.41 | 48.44 | 52.00 | 46.24 | 60.54 | 50.60 | 72.89 | 66.79 | **87.06** | **81.13** |
| AA | 72.07 | 76.76 | 60.72 | 58.00 | 64.97 | 63.72 | 68.12 | 63.34 | 65.43 | 61.90 | 81.14 | 79.71 | **87.19** | **81.25** |
| Kappa | 0.61 | 0.66 | 0.46 | 0.45 | 0.40 | 0.39 | 0.42 | 0.37 | 0.50 | 0.41 | 0.65 | 0.59 | **0.86** | **0.80** |

*The best results are highlighted in bold.*




(a) (b)

Fig. 2. Classification results for the Indian Pines data set with varying numbers of labeled samples. (a) Classification results of the unlabeled data by FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively. (b) Classification results of the testing data by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively.




(a) (b)

Fig. 3. Classification results of the Salinas data set with varying numbers of labeled samples. (a) Classification results of the unlabeled data by FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively. (b) Classification results of the testing data by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively.

objective function based on IR1

$$\min_{\mathbf{W}} \|\mathbf{W}\mathbf{W}^T\mathbf{Z}-\mathbf{Z}\|_F^2+\alpha g(\mathbf{W}^T\mathbf{Z}) +\lambda_1 \mathrm{tr}((\mathbf{W}^T\mathbf{Z})(\mathbf{L}_E+\beta\mathbf{L}_A)(\mathbf{W}^T\mathbf{Z})^T). \quad (31)$$

3) *IR3:* Based on IR2, IR3 further adds superpixel-level constraints to avoid the "salt-and-pepper" problem generated in the pixel-based classification and offer the feedback information about the quality of subspace learning

$$\min_{\mathbf{W}} \|\mathbf{W}\mathbf{W}^T\mathbf{Z}-\mathbf{Z}\|_F^2+\alpha g(\mathbf{W}^T\mathbf{Z}) + \lambda_1 \mathrm{tr}((\mathbf{W}^T\mathbf{Z})(\mathbf{L}_E+\beta\mathbf{L}_A)(\mathbf{W}^T\mathbf{Z})^T) + \lambda_3\left(\sum_{i=1}^{n}\exp\left(\frac{\|\mathbf{W}^T\mathbf{z}_i-\mu_t\|_2^2}{\delta_s}\right)+\gamma\|\mathbf{F}-\mathbf{F}_s\|_F^2\right). \quad (32)$$

For IR1, IR2, and IR3, the following classification function is learned independently from the feature learning for

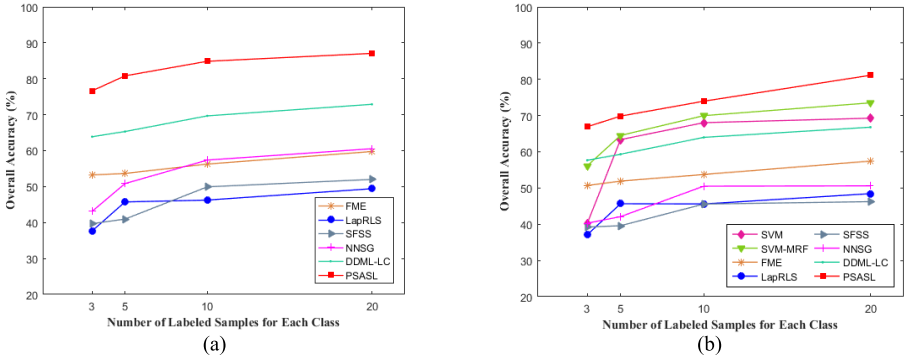



Fig. 4. Classification results of the PaviaU data set with varying numbers of labeled samples. (a) Classification results of the unlabeled data by FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively. (b) Classification results of the testing data by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and the PSASL, respectively.

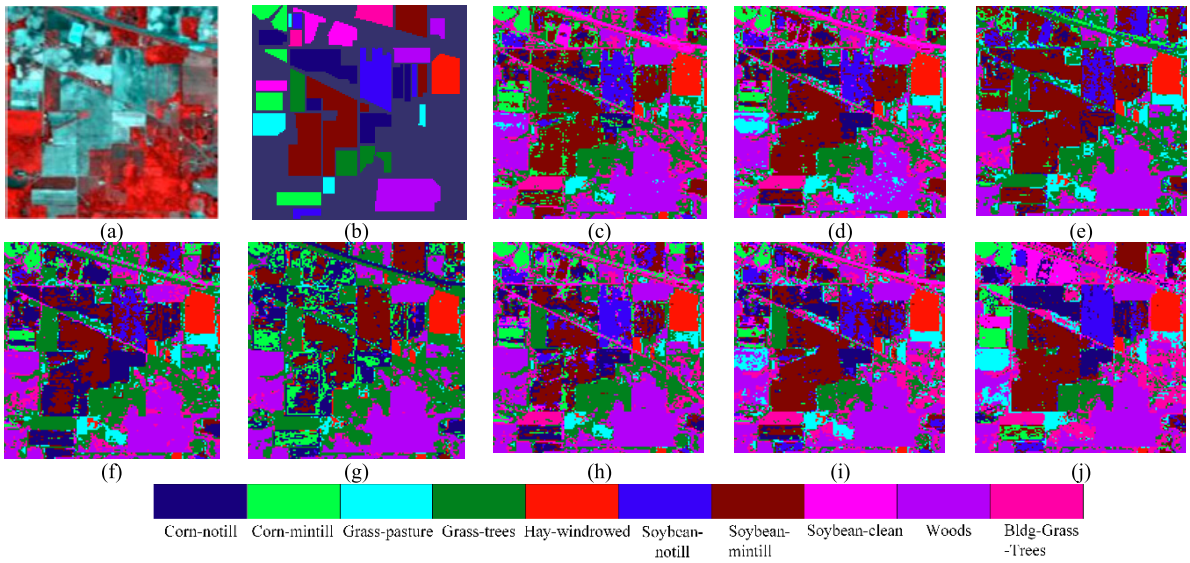



Fig. 5. Classification maps of the Indian Pine data set. (a) False-color image. (b) Ground truth. (c)–(j) Classification maps obtained by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and PSASL, respectively.

semisupervised HSI classification:

$$\min_{\mathbf{W},\mathbf{F},\mathbf{H},\mathbf{b}} (\mathrm{tr}(\mathbf{F}^T(\mathbf{L}_E + \beta\ \mathbf{L}_A)\mathbf{F}) + \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{S}(\mathbf{F}-\mathbf{Y}))) + \lambda_4\|(\mathbf{W}^T\mathbf{Z})^T\mathbf{H} + \mathbf{1}\mathbf{b}^T - \mathbf{F}\|_F^2. \quad (33)$$

We randomly select 30% samples of the Indian Pines data set, 10% samples of the Salinas and PaviaU data sets as the training data, and set $s = 20$ for the three data sets. The classification results of unlabeled data and testing data are reported in Tables VIII–X.

From Tables VIII–X, we have the following observations.

1) Since the spectral–spatial graph regularization make the feature representation more discriminative by compressing the highly correlated bands, the classification results of IR2 are better than IR1.
2) IR3 outperforms IR2 because the superpixel-level constraint in IR3 can avoid the "salt-and-pepper" problem generated in the pixel-based classification.
3) Compared with IR1, IR2, and IR3, PSASL obtains the best classification results due to the joint learning framework.

TABLE VIII
SEMISUPERVISED HSI CLASSIFICATION ACCURACY (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE INDIAN PINES DATA SET

| Method | Unlabeled Data | Test Data |
|---|---|---|
| IR1 | 85.75 | 76.32 |
| IR2 | 86.45 | 79.01 |
| IR3 | 88.79 | 83.07 |
| PSASL | **92.13** | **89.85** |

*The best results are highlighted in bold.*

### *E. Parameters Analysis*

In the PSASL, seven parameters $\alpha$, $\beta$, $\gamma$, $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ need to be tuned in each data set. We set $\alpha = 0.1$, $\beta = 0.001$, and $\gamma = 0.0001$ in the experiments and mainly discuss the influences of the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ on the HSI classification results.

The parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are related to the terms of spectral–spatial graph regularization, label space regularization, superpixel-level constraint, and out-of-sample extension, respectively. As shown in Fig. 8, we have to tune

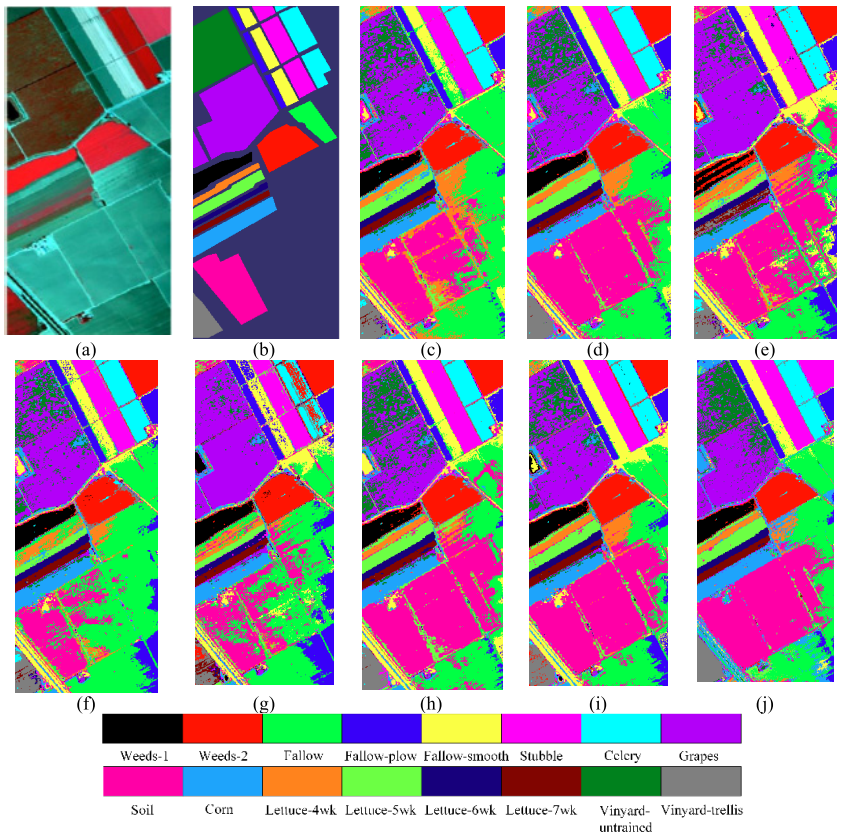



Fig. 6. Classification maps of the Salinas data set. (a) False-color image. (b) Ground truth. (c)–(j) Classification maps obtained by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and PSASL, respectively.

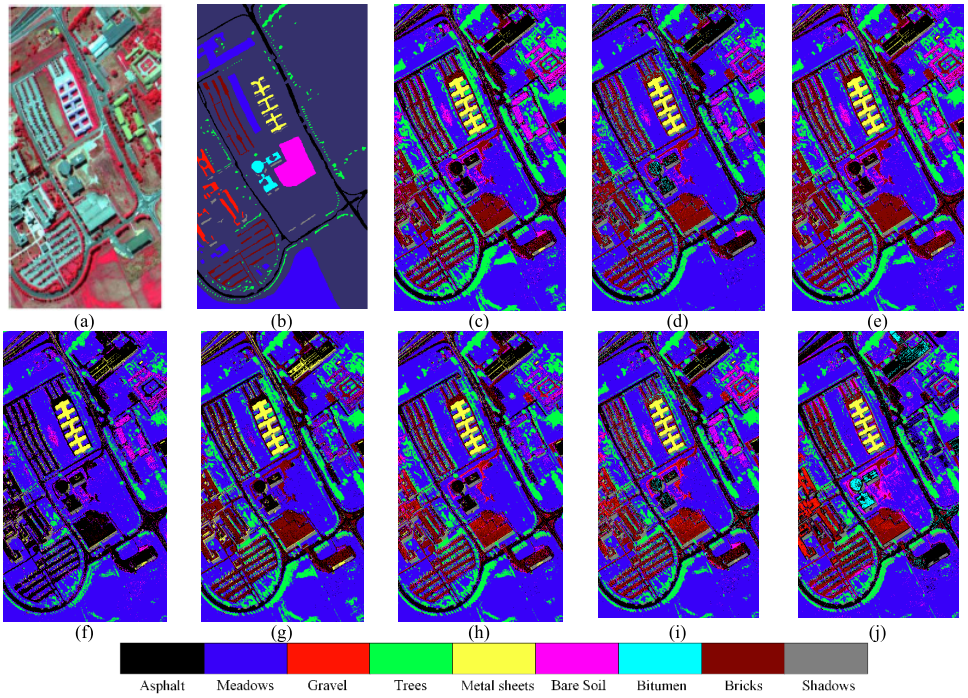



Fig. 7. Classification maps of the PaviaU data set. (a) False-color image. (b) Ground truth. (c)–(j) Classification maps obtained by SVM, SVM-MRF, FME, LapRLS, SFSS, NNSG, DDML-LC, and PSASL, respectively.

the parameters when we classify each data set. The parameters that make the HSI classification results the best are different for each data set since each data set has its own distinctive features. We find that the values of $\lambda_1$, $\lambda_2$, and $\lambda_3$ are larger than $\lambda_4$ on the three data sets, which proves that the spectral–spatial graph regularization, label space regularization, and superpixel-level constraint play more important roles in subspace learning than the out-of-sample extension. From Fig. 8,

TABLE IX

SEMISUPERVISED HSI CLASSIFICATION ACCURACY (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE SALINAS DATA SET

| Method | Unlabeled Data | Test Data |
|---|---|---|
| IR1 | 88.27 | 84.45 |
| IR2 | 90.03 | 85.52 |
| IR3 | 91.72 | 88.23 |
| PSASL | **94.90** | **93.21** |

*The best results are highlighted in bold.*

TABLE X

SEMISUPERVISED HSI CLASSIFICATION ACCURACY (%) BY SELECTING 20 LABELED SAMPLES FOR EACH CLASS ON THE PAVIAU DATA SET

| Method | Unlabeled Data | Test Data |
|---|---|---|
| IR1 | 77.12 | 68.69 |
| IR2 | 79.78 | 71.31 |
| IR3 | 82.23 | 75.10 |
| PSASL | **87.06** | **81.13** |

*The best results are highlighted in bold.*

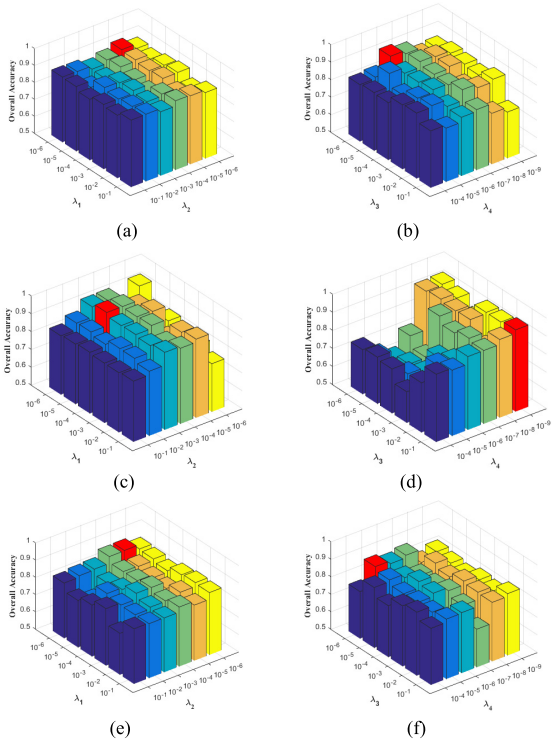



Fig. 8. Influences of different values of the parameters on the semisupervised classification results. (a) and (b) Influences of different values of $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ on the classification results for Indian Pines data set. (c) and (d) Influences of different values of $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ on the classification results for Salinas data set. (e) and (f) Influences of different values of $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ on the classification results for PaviaU data set.

it is noted that $\lambda_1$, $\lambda_2$, and $\lambda_3$ are nearly the same, which indicates that the constraints of the pixel-level and superpixel-level are equally important in the subspace learning.

According to the magnitude among different parts of the objective function, we obtain the initial values of the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$. The best values of these parameters should refer to the classification performance of the training data. In general, the best classification results for the three

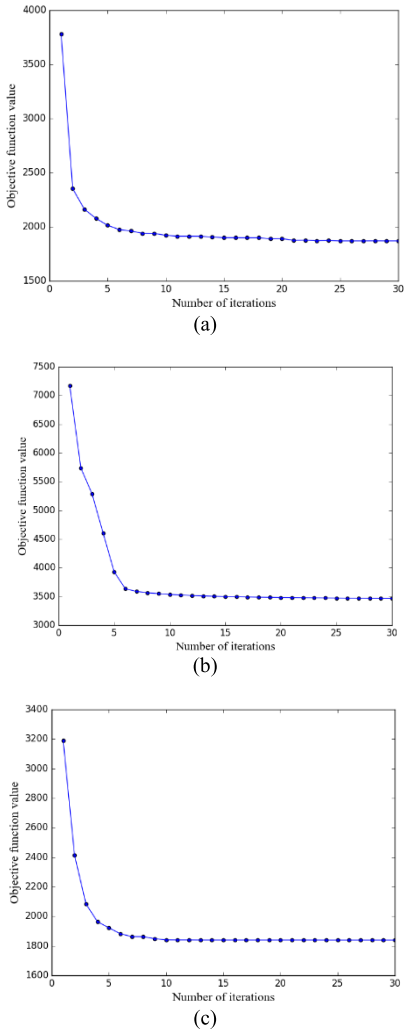



Fig. 9. Convergence processes of different data sets. (a) Indian Pines data set. (b) Salinas data set. (c) PaviaU data set.

data sets are achieved when the four parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are close to $10^{-6}$, $10^{-5}$, $10^{-6}$, and $10^{-5}$, respectively.

### *F. Algorithmic Convergence*

Solving the variables **W**, **F**, **H,** and **b** in (12) simultaneously is very difficult due to the highly nonlinear nature of (12). Inspired by the least-squares quantization, we adopt a customized iterative algorithm to optimize the variables. The objective function can converge to a local optimum by using Algorithm 1. Four variables **W**, **F**, **H,** and **b** need to be optimized in (12). In each iteration, with the help of the LADMAP, the process for optimizing **W** makes the objective function achieves a local minimum as other variables are fixed. The functions of optimizing **F**, **H,** and **b**are convex, and thus, they are convergent. With the optimized variables, the objective function can converge to a local optimum.

The convergence processes under different data sets are shown in Fig. 9. It is noted that (12) can converge to a local optimum (or even a global minimum) and converge very fast. Equation (12) usually reaches the convergence within about five iterations for each HSI data set. Therefore, the proposed solution in Algorithm 1 is very effective.

## V. Conclusion

In this paper, the PSASL, which is an approach for PSASL, is proposed for HSI classification. The main contribution of the PSASL lies in explicitly integrating the pixel-level regularization, superpixel-level regularization, and the single predictive linear classifier into the objective function for subspace learning. In order to avoid the "salt-and-pepper" problem generated in the pixel-based classification methods, superpixels are adopted to add the constraint of subspace learning and refine the HSI classification results. By means of keeping the purity of the distribution of category pixels in subspace learning and clustering results, the generated superpixels provide feedback information to the subspace learning and cluster results in the pixel-level. The solution to the objective function is efficiently achieved by employing a customized iterative algorithm, and it converges very fast.

Experimental results on three data sets show the effectiveness of the PSASL. The classification accuracies obtained by the PSASL are higher than those obtained by many recently proposed methods.

In future work, we will combine the PSASL with deep learning structure to automatically learn more representative features of the pixels for further enhancing performance of HSI classification.

## References


[1] Y. Gao, R. Ji, P. Cui, Q. Dai, and G. Hua, "Hyperspectral image classification through bilayer graph-based learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2769–2778, Jul. 2014.

[2] P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1890–1907, Jul. 2010.

[3] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.

[4] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[5] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 321–328.

[6] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.

[7] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[8] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 612–620.

[9] V. Sindhwani, P. Niyogi, M. Belkin, and S. Keerthi, "Linear manifold regularization for large scale semi-supervised learning," in *Proc. ICML Workshop Learn. Partially Classified Training Data*, 2005, pp. 1–4.

[10] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[11] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760–2771, Sep. 2015.

[12] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," *Natural Image Statist.*, vol. 39, pp. 151–175, Sep. 2009.

[13] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based $k$-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[14] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[15] 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.

[16] Y. Zhou and Y. Wei, "Learning hierarchical spectral–spatial features for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1667–1678, Jul. 2016.

[17] S. Li, T. Lu, L. Fang, X. Jia, and J. A. Benediktsson, "Probabilistic fusion of pixel-level and superpixel-level hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7416–7430, Dec. 2016.

[18] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten, "Hyperspectral image classification with independent component discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4865–4876, Dec. 2011.

[19] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[20] 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.

[21] J. Li, H. Zhang, L. Zhang, X. Huang, and L. Zhang, "Joint collaborative representation with multitask learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5923–5936, Sep. 2014.

[22] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[23] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.

[24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[25] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 2097–2104.

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[27] M. Van den Bergh, X. Boix, G. Roig, and L. Van Gool, "SEEDS: Superpixels extracted via energy-driven sampling," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 298–314, 2015.

[28] G. Zhang, X. Jia, and J. Hu, "Superpixel-based graphical model for remote sensing image mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5861–5871, Nov. 2015.

[29] L. Y. Fang, S. T. Li, X. D. Kang, and J. A. Benediktsson, "Spectral-spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4186–4201, Aug. 2015.

[30] J. Li, H. Zhang, and L. Zhang, "Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5338–5351, Oct. 2015.

[31] W. Fu, S. Li, L. Fang, and J. A. Benediktsson, "Adaptive spectral–spatial compression of hyperspectral image with sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 671–682, Feb. 2017.

[32] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[33] X. Huang and L. Zhang, "An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4173–4185, Dec. 2008.

[34] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018, doi: 10.1109/TCYB.2016.2605044.2017.

[35] G. Bilgin, S. Erturk, and T. Yildirim, "Unsupervised classification of hyperspectral-image data using fuzzy approaches that spatially exploit membership relations," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 673–677, Oct. 2008.

[36] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. application to hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1403–1413, Jun. 2010.

[37] L. P. Zhang, L. Zhang, D. Tao, and X. Huang, "A multifeature tensor for remote-sensing target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 374–378, Feb. 2011.

[38] H. Du, H. Qi, X. Wang, R. Ramanath, and W. E. Snyder, "Band selection using independent component analysis for hyperspectral image processing," in *Proc. 32nd Appl. Imag. Pattern Recognit. Workshop*, Oct. 2003, pp. 93–98.

[39] B. C. Kuo, C. H. Li, and J. M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.

[40] P. Zhong and R. Wang, "Modeling and classifying hyperspectral imagery by CRFs with sparse higher order potentials," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 688–705, Feb. 2011.

[41] B. B. Damodaran, R. R. Nidamanuri, and Y. Tarabalka, "Dynamic ensemble selection approach for hyperspectral image classification with joint spectral and spatial information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2405–2417, Jun. 2015.

[42] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 144–153, Jan. 2015.

[43] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.

[44] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans.Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.

[45] C. Persello, A. Boularias, M. Dalponte, T. Gobakken, E. Næsset, and B. Schölkopf, "Cost-sensitive active learning with lookahead: Optimizing field surveys for remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6652–6664, Oct. 2014.

[46] S. Sun, P. Zhong, H. Xiao, and R. Wang, "Active learning with Gaussian process classifier for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1746–1760, Apr. 2015.

[47] Y. Zhong and L. Zhang, "An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 894–909, Mar. 2012.

[48] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[49] 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.

[50] Y. Wang *et al.*, "Learning a discriminative distance metric with label consistency for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4427–4440, Aug. 2017, doi: 10.1109/TGRS.2017.2692280.2017.

[51] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[52] 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 Recognit.*, vol. 37, no. 6, pp. 1097–1116, 2004.

[53] X. Sun, N. M. Nasrabadi, and T. D. Tran, "Task-driven dictionary learning for hyperspectral image classification with structured sparsity constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4457–4471, Aug. 2015.

[54] J. Peng, Y. Zhou, and C. L. P. Chen, "Region-kernel-based support vector machines for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4810–4824, Sep. 2015.

[55] R. Roscher and B. Waske, "Shapelet-based sparse representation for landcover classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1623–1634, Mar. 2016.

[56] T. Lu, S. Li, L. Fang, L. Bruzzone, and J. A. Benediktsson, "Set-to-set distance-based spectral–spatial classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7122–7134, Dec. 2016.

[57] S. Mei, M. He, Y. Zhang, Z. Wang, and D. Feng, "Improving spatial–spectral endmember extraction in the presence of anomalous ground objects," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4210–4222, Nov. 2011.

[58] S. Mei, M. He, D. Feng, and Z. Wang, "Spectral-spatial endmember extraction for hyperspectral image," *IEEE Trans. Geosci. Remote Sens*, vol. 48, no. 9, pp. 3434–3445, Sep. 2010.

[59] B. Du, Y. Zhang, L. Zhang, and D. Tao, "Beyond the sparsity-based target detector: A hybrid sparsity and statistics based detector for hyperspectral images," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5345–5357, Nov. 2016.

[60] B. Du, M. Zhang, L. Zhang, R. Hu, and D. Tao, "PLTD: Patch-based low-rank tensor decomposition for hyperspectral images," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 67–79, Jan. 2017.

[61] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[62] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using Gaussian mixture models and Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 153–157, Jan. 2014.

**Jie Mei** is currently pursuing the master's degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include remote sensing image processing, and image-based and LiDAR-based segmentation and reconstruction.

**Yuebin Wang** received the Ph.D. degree from the School of Geography, Beijing Normal University, Beijing, China, in 2016.

He was a Post-Doctoral Researcher with the School of Mathematical Sciences, Beijing Normal University. He is currently an Assistant Professor with the School of Land Science and Technology, China University of Geosciences, Beijing. His research interests include remote sensing imagery processing and 3-D urban modeling.

**Liqiang Zhang** received the Ph.D. degree in geoinformatics from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2004.

He is currently a Professor with the Faculty of Geographical Science, Beijing Normal University, Beijing. His research interests include remote sensing image processing, 3-D urban reconstruction, and spatial object recognition.

**Bing Zhang** is currently a Full Professor and the Deputy Director of the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS), Beijing, China, where he has been leading lots of key scientific projects in the area of hyperspectral remote sensing for more than 20 years. He has authored more than 300 publications, including more than 190 journal papers. He has edited six books/contributed book chapters on hyperspectral image processing and subsequent applications. He has developed five software systems in the image processing and applications. His research interests include the development of mathematical and physical models and image processing software for the analysis of hyperspectral remote sensing data in many different areas.

Dr. Zhang was a recipient of the National Science Foundation for Distinguished Young Scholars of China in 2013 and the 2016 Outstanding Science and Technology Achievement Prize of CAS for his special achievements in hyperspectral remote sensing. He is currently serving as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is also the Guest Editor of the series of special issues of IEEE JSTARS, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the PROCEEDINGS OF IEEE, and *Pattern Recognition Letters*. He has been serving as a Technical Committee Member of the IEEE Workshop on Hyperspectral Image and Signal Processing Evolution in Remote Sensing since 2011, and as the President of the Hyperspectral Remote Sensing Committee of China National Committee of International Society for Digital Earth since 2012.

**Suhong Liu** received the B.S. degree in computer science from Southwest Jiaotong University, Chengdu, China, in 1988, the M.S. degree in geophysical well-logging from Jianghan Petroleum University, Jingzhou, China, in 1991, and the Ph.D. degree in cartography and remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1999.

She is currently a Professor with the Faculty of Geographical Science, Beijing Normal University, Beijing. Her research interests include spatiotemporal analysis of remotely sensed data and retrieval of land biophysical parameters from satellite data.

**Panpan Zhu** is currently pursuing the Ph.D. degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China.

Her research interests include remote sensing image processing, and image-based classification and retrieval.

**Yingchao Ren** is currently an Associate Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing imagery processing and geographical information science spatial analysis.