Region-Aware Hierarchical Latent Feature Representation Learning-Guided Clustering for Hyperspectral Band Selection

Jun Wang, Chang Tang, *Member, IEEE*, Xinwang Liu, *Senior Member, IEEE*, Wei Zhang, *Member, IEEE*, Wanqing Li, *Senior Member, IEEE*, Xinzhong Zhu, *Member, IEEE*, Lizhe Wang, *Fellow, IEEE*, and Albert Y. Zomaya, *Fellow, IEEE*

Abstract—Hyperspectral band selection aims to identify an optimal subset of bands for hyperspectral images (HSIs). For most existing clustering-based band selection methods, they directly stretch each band into a single feature vector and employ the pixelwise features to address band redundancy. In this way, they do not take full consideration of the spatial information and deal with the importance of different regions in HSIs, which leads to a nonoptimal selection. To address these issues, a region-aware hierarchical latent feature representation learning-guided clustering (HLFC) method is proposed. Specifically, in order to fully preserve the spatial information of HSIs, the superpixel segmentation algorithm is adopted to segment HSIs into multiple regions first. For each segmented region, the similarity graph is constructed to reflect the bands-wise similarity, and its corresponding Laplacian matrix is generated for learning low-dimensional latent features in a hierarchical way. All latent features are then fused to form a unified feature representation of HSIs. Finally, k-means clustering is utilized on the unified feature representation matrix to generate multiple clusters from which the band with maximum information entropy is selected to form the final subset of bands. Extensive experimental results demonstrate that the proposed clustering method can achieve superior performance than the state-of-the-art representative methods on the band selection. The demo code of this work is publicly available at https://github.com/WangJun2023/HLFC.

Index Terms—Clustering, feature fusion, hierarchical latent feature learning, hyperspectral band selection.

Jun Wang is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: wang_jun@cug.edu.cn).

Chang Tang and Lizhe Wang are with the School of Computer Science and the Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences, Wuhan 430074, China (e-mail: tangchang@cug.edu.cn; lizhe.wang@gmail.com).

Xinwang Liu is with the School of Computer, National University of Defense Technology, Changsha 410073, China (e-mail: xinwangliu@nudt.edu.cn).

Wei Zhang is with the Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputing Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China (e-mail: wzhang@qlu.edu.cn).

Wanqing Li is with the School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2500, Australia (e-mail: wanqing@uow.edu.au).

Xinzhong Zhu is with the College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321017, China (e-mail: zxz@zjnu.edu.cn).

Albert Y. Zomaya is with the School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia (e-mail: albert.zomaya@sydney.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3191121.

Digital Object Identifier 10.1109/TCYB.2022.3191121

I. INTRODUCTION

IN RECENT years, hyperspectral images (HSIs) have been widely employed in many applications, for example, medical imaging processing [1], land cover classification [2], mineral exploration [3], etc. However, there is a large number of spectral bands available for HSIs and adjacent bands often produce relatively similar and redundant information [4] in the HSIs, selection of an optimal subset of bands is needed to reduce the redundant information.

Information redundancy in HSIs can be addressed via feature extraction [5]–[8] and feature selection [9]–[12], where the latter is also known as band selection. For the former approach, hyperspectral data are transformed into a lower-dimensional feature space according to a certain specific mapping, and then new features are constructed by combining some representative bands. Typical methods include principal component analysis (PCA) [13], LDA [14], and so on. Although these methods had achieved satisfying performance in data dimension reduction, due to the new features are generated from a linear combination of the original spectral bands, this will raise explicit spectral distortion and make the physical meaning of features difficult to be interpreted [15], [16]. Thus, this manner limits the applications of physical spectral measures. On the contrary, band selection only discards some redundant bands with preserving the original hyperspectral information, which is a benefit for subsequent quantitative analysis. Therefore, we mainly focus on the band selection.

Various band selection methods have been studied, which can be classified into supervised [17], semisupervised [18], and unsupervised [19] ones. Supervised and semisupervised band selection methods can achieve satisfying results in many cases

due to the advantages of labeled samples. However, it is labor-intensive and time-consuming to annotate HSIs. Therefore, unsupervised band selection methods are more practically feasible and often preferred.

Unsupervised band selection methods can be generally grouped into clustering-based [9] or ranking-based [20], and searching-based ones [21]. Clustering-based methods partition all bands into multiple clusters, and then the bands are selected to form the selected band subset in each cluster [22], [23]. Ranking-based methods calculate the weight of each band according to certain predefined metrics, and then the top-ranked bands are selected as the feature bands. Searching-based methods mainly select a subset of bands that produce optimal performance under the evaluation of a certain metric.

Despite the recent advance in unsupervised band selection, there are several limitations to existing methods. First, most existing methods tend to regard each band as a single feature, for example, [24]. In such a manner, the spatial structure information of HSIs is ignored. Second, hyperspectral data are often contaminated by noise during acquisition [25] and a mechanism to reduce the impact of the noise on selecting bands is needed. Third, regions corresponding to various land covers are often with different importance and such prior information has been neglected in previous methods, such as [26].

To address the above limitations, a band selection method via region-aware hierarchical latent feature representation learning-guided clustering (HLFC) is proposed. In order to fully preserve the spatial information of HSIs, HLFC segments the first principal component of HSIs into multiple regions via superpixel segmentation. For each segmented region, its corresponding Laplacian matrix is constructed to capture the spatial structure, and low-dimensional latent features are generated from the Laplacian matrix via a hierarchical strategy. The latent features are then fused to construct the complementary and common features in different regions. Finally, all bands are partitioned into multiple clusters based on the unified feature representation, and the band with maximum information entropy in each cluster is selected as the feature band.

The main contributions of this article include as follows.

1) A novel region-aware feature representation learning-guided clustering method for hyperspectral band selection. This method exploits spatial and spectral properties of HSIs and learns low-dimensional discriminative latent features of HSIs via a hierarchical strategy.

2) An effective feature fusion method to construct a global feature representation of HSIs. Different from the previous methods that keep the unified representation matrix fixed during fusion, our method couples the learning and fusion of features into a joint framework so that the two processes can reinforce each other.

3) An optimization algorithm is designed to solve the newly formulated HLFC problem.

4) Extensive experiments verify that HLFC can achieve considerable improvement over other representative band selection methods.

The remainder of this article is organized as follows. In Section II, we review some representative band selection methods, for example, clustering-based ones, ranking-based ones, and searching-based ones. Section III introduces the proposed method and an optimization solution to the proposed model is presented in Section IV. Extensive experiments and comparisons to the state-of-the-art methods are reported in Section V. Section VI concludes this article with remarks.

II. RELATED WORK

This section briefly reviews typical band selection methods, including clustering-based, ranking-based, and searching-based methods in relation to the proposed method.

A. Clustering-Based Methods

The clustering-based methods partition all original bands into multiple clusters and the representative band is selected to form the optimal feature band subset. There are two essential steps in these methods, one is how to group the bands into clusters, and another is how to select representative bands from multiple clusters. Yang *et al.* [27] adopted *k*-means clustering to partition all bands into multiple classes. Different from traditional clustering-based methods that select feature bands individually from each class, they select feature bands by traversing all classes simultaneously. However, because *k*-means clustering is very sensitive to the selection of initial clustering centers, different initialization methods are usually needed to determine a better solution. In addition, only when the number of samples is small, this method can achieve good performance. In order to solve the sensitivity of the *k*-means clustering algorithm to initial conditions, [28] proposed a sample-based AP clustering algorithm. Although it can obtain stable clustering results, the time complexity is relatively high. In addition, inspired by the graph theory in clustering, many graph-based methods have been proposed. Li *et al.* [29] constructed the affinity matrix to describe the similarity between pairwise bands, and then spectral clustering was utilized to generate multiple subgraphs from which the representative bands were selected. To preserve the local structure of HSIs, Sun and Du [30] proposed to construct the regularized Laplacian graph of HSIs, which was utilized to partition all hyperspectral bands into multiple classes. In addition, to improve the performance of feature selection on high-dimensional data, Song *et al.* [31] proposed a new hybrid evolutionary feature selection algorithm. Furthermore, other clustering-based evolutionary feature selection algorithms have also been developed, such as [32]–[34]. With the development of deep learning, some deep clustering methods have been proposed, for example, [35]–[37].

Overall, although these methods can achieve satisfying performance, the issues with them mainly lie in they regard each band as a single and high-dimensional feature vector and consider all pixels or all regions in the HSIs as equal importance. This does not accommodate well the cases where some regions in the HSIs are more important than others.
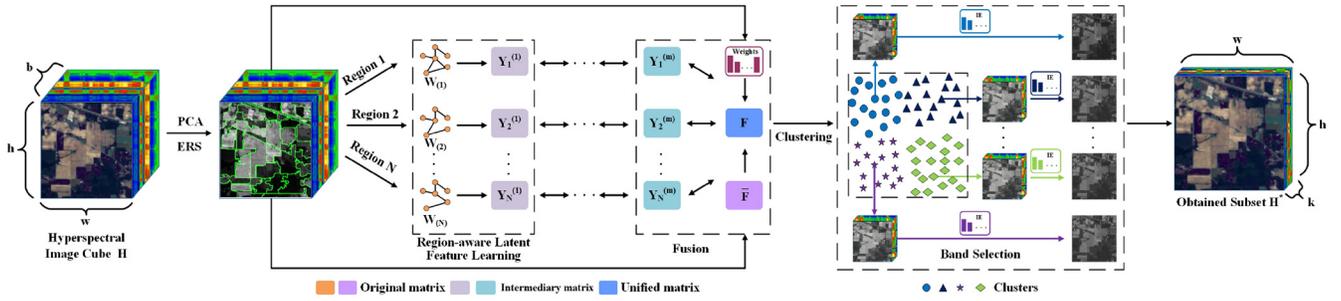
Fig. 1. Overview of the proposed HLFC. First, PCA and ERS are employed to segment an HSI cube $\mathbf{H} \in \mathbb{R}^{w \times h \times b}$ into multiple regions from which the similarity graph $\{\mathbf{W}_{(i)}\}_{i=1}^{N}$ is constructed. Second, latent feature matrices $\{\mathbf{Y}_{(i)}^{1}\}_{i=1}^{N}$, one for each region, are computed from the Laplacian matrices, referred to as region-aware latent feature learning. The dimensionality of the latent feature matrices is further reduced in a hierarchical manner to $\{\mathbf{Y}_{(i)}^{m}\}_{i=1}^{N}$. Third, a unified latent feature matrix $\mathbf{F} \in \mathbb{R}^{b \times d}$ is obtained by fusing $\{\mathbf{Y}_{(i)}^{m}\}_{i=1}^{N}$. Finally, the k-means algorithm is utilized to partition all bands into multiple clusters. The band with maximum information entropy in each cluster is selected to form the subset $\mathbf{H}^{*}$ of the bands.

B. Ranking-Based Methods

The ranking-based band selection methods quantify the weight of all bands via a predefined metric, and then the top-ranked bands are selected as the feature bands. Thus, the most important problem of this kind of method is how to determine the optimal metric. Most current methods mainly consider the information or correlation of bands, such as signal-to-noise ratio [38], information entropy [39], mutual information [40], and so on. In [41], information entropy was used to measure the importance of bands. Chang *et al.* [42] adopted PCA to prioritize the variance of all bands. However, neither of them took the correlation among bands into consideration. To address this, the Kullback–Leibler divergence was employed in [15] to remove the redundant bands. To be specific, the K–L divergence between the Gaussian distribution of all bands is calculated, and the difference between bands was evaluated through divergence. Finally, all bands are sorted according to the divergence and the band with maximum information (i.e., maximum divergence) is selected to form the optimal band subset. In addition, mutual information [43] and signal-to-noise [44] were also used to measure the importance of bands. Li *et al.* [45] proposed the constrained multiple-band selection method to select the feature bands, which intends to minimize the correlation of all selected feature bands.

In a word, ranking-based methods either select the most informative bands or the least correlated bands to form the feature band subset without considering explicitly discriminativeness. The proposed HLFC addresses this issue.

C. Searching-Based Methods

This kind of method regards band selection as a combinatorial optimization problem, and the optimal solution is obtained via some search algorithms. Based on the type of objective function and searching strategy, the searching-based methods can be roughly divided into evolutionary-based [46]–[48] and greedy-based methods [49]–[51]. The former mainly apply some evolutionary algorithms to obtain the optimal feature subset. For example, based on the variable-size clustering, He *et al.* [52] proposed a multitask artificial bee colony band selection method to simultaneously obtain different sizes of optimal band subset. To be specific, they modeled the band selection problem as a multitask optimization problem and

then designed a variable-size band clustering method as the objective function. Finally, multiple different sizes of band subsets are searched via a multi-micro-group bee colony algorithm. To address the local optimal and high computational burden of some existing PSO-based feature selection algorithms, Chen *et al.* [53] decomposed the high-dimensional feature selection task into several low-dimensional feature selection tasks from which the optimal feature subset is obtained by knowledge transfer. The greedy-based methods mainly adopt some greedy algorithms to obtain the optimal band subset. For instance, Geng *et al.* [50] proposed a volume-gradient-based band selection method to capture the subtle relationship between the volume of a subsimplex and the volume gradient of a simplex for HSIs. Based on the assumption that the optimal band subset should reconstruct the whole band with minimum error, they regarded the distance between each band and the hyperplane constructed by the rest bands as the prediction error and then tried to find the bands with the maximum volume of the parallelotope.

Although the most existing searching-based methods are very effective for band selection, they have high computational complexity as a result of searching all feasible solutions of band subset.

III. PROPOSED METHOD

A. Notations

Notations used throughout this article are briefly explained in this section. Bold uppercase letters and bold lowercase letters represent matrices and vectors, respectively, and scalars are denoted in a nonbold italic font. In addition, for a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, its trace and transpose are denoted by $\text{Tr}(\mathbf{M})$ and $\mathbf{M}^{\top}$, respectively. Its ijth entry and Frobenius norm are denoted as M_{ij} and $||\mathbf{M}||_F = (\sum_i \sum_j M_{ij}^2)^{[1/2]}$, respectively. If $\mathbf{M}$ is a 3-way tensor, then $\mathbf{M}_{(n)}$ is used to represent the nth frontal slices, that is, $\mathbf{M}(:,:,n)$. In addition, $\mathbf{M}^{(i)}$ is treated as the intermediate tensor of the ith level. $\mathbf{I}_n$ represents an identity matrix, and m_i denotes the ith element of vector $\mathbf{m}$.

B. Overview

Fig. 1 shows the block diagram of the proposed HLFC. First, PCA and entropy ratio segmentation (ERS) are employed

to segment an HSI cube $\mathbf{H} \in \mathbb{R}^{w \times h \times b}$ into N regions from which the similarity graphs $\{\mathbf{W}_{(i)}\}_{i=1}^{N}$ are constructed. Second, latent feature matrices $\{\mathbf{Y}_{(i)}^{1}\}_{i=1}^{N}$, one for each region, are computed from the Laplacian matrices, referred to as region-aware latent feature learning. The features are refined in a hierarchical manner to $\{\mathbf{Y}_{(i)}^{m}\}_{i=1}^{N}$ with reduced dimensionality. Third, a unified latent feature matrix $\mathbf{F} \in \mathbb{R}^{b \times d}$ is obtained by fusing $\{\mathbf{Y}_{(i)}^{m}\}_{i=1}^{N}$. Finally, the k-means algorithm is employed to group all bands into multiple clusters. And the band with maximum information entropy in each cluster is selected to form the subset $\mathbf{H}^{*}$ of the bands.

C. Region-Aware Hierarchical Latent Feature Learning

For an HSI cube, how to fully exploit its abundant spectral and spatial information to improve the performance on band selection is a critical factor. Although some existing representative band selection methods have achieved satisfying performance by only using spectral information, we think the utilization of spatial structure can further improve their performance. It is well known that the first principal component contains most of the information of original HSIs, which motivates us to use it for capturing the spatial structure. Considering that pixels in a region of the same land cover would have similar spectral properties [54], that is, similar pixel values in the HSIs, and different regions are of different importance. We use the entropy rate segmentation (ERS) [55], [56] to get a relatively accurate segmentation result that reflects the spatial structure of different objects from the first principal component, and then the segmented maps are extended to all bands and yields

$$\mathbf{H} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N\} \quad \text{s.t.} \quad \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \quad (\forall i, j, i \neq j) \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{w \times h \times b}$ denotes original HSIs; and N and $\mathcal{R}_i$ are the number of segmented regions and the ith region of HSIs, respectively.

In order to deal with redundant information and noise interference, discriminative latent features are learned from each region. Specifically, similarity matrices of segments in each band are constructed via the k-nearest neighbor graph. Suppose the similarity of two bands is relatively high in the original feature space, then the similarity property of them should remain the same as the previous in the new feature space. Thus, Laplacian matrices are generated from the graphs to learn the latent features, that is

$$\mathbf{L}_{(i)} = \mathbf{I} - \mathbf{D}_{(i)}^{-\frac{1}{2}} \mathbf{W}_{(i)} \mathbf{D}_{(i)}^{-\frac{1}{2}} \quad (2)$$

where $\mathbf{L}_{(i)}$ and $\mathbf{W}_{(i)}$ are the Laplacian matrix and the similarity matrix across different bands of the ith region, respectively. $\mathbf{D}_{(i)}$ is the diagonal matrix with each element $\mathbf{D}_{jj} = \sum_{i=1}^{b} \mathbf{W}_{ij}$. In the proposed method, k-nearest neighbor graph and Euclidean distance are adopted to construct similarity matrix $\mathbf{W}_{(i)}$. With respect to the Laplacian matrix of each segmented region, the latent features can be distilled via spectral embedding. Mathematically, the problem is formulated as follows:

$$\min_{\mathbf{Y}_{(i)}} \sum_{i=1}^{N} \text{Tr}\left(\mathbf{Y}_{(i)}^{\top} \mathbf{L}_{(i)} \mathbf{Y}_{(i)}\right) \quad \text{s.t.} \quad \mathbf{Y}_{(i)}^{\top} \mathbf{Y}_{(i)} = \mathbf{I}_d \quad (3)$$
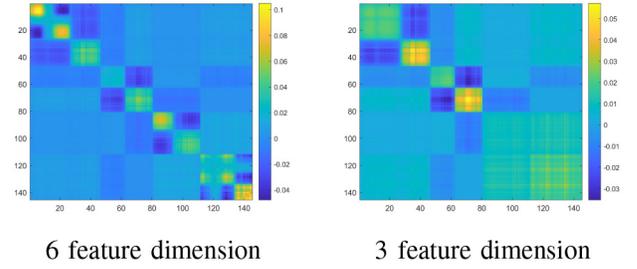
Fig. 2. Band-correlation matrix corresponding to different latent feature dimension for the Botswana dataset.

where $\mathbf{Y}_{(i)} \in \mathbb{R}^{b \times d}$ represents the latent feature matrix with d dimension for the ith region. Equation (3) can be easily solved by performing eigenvalue decomposition on $\mathbf{L}_{(i)}$. To be specific, the solution $\mathbf{Y}_{(i)}$ is formed by the eigenvectors corresponding to the first d smallest eigenvalues of $\mathbf{L}_{(i)}$.

Since the determination of optimal latent feature dimension is still an open problem, we can only set it empirically. When the latent feature dimension is fixed, the final hyperspectral band selection via clustering would highly depend on the quality of latent representation $\mathbf{Y}_{(i)}$. Accordingly, how to obtain a better latent representation with a fixed dimension becomes a critical problem. For a Laplacian matrix, the key cluster information is only embedded in its a few smallest eigenvectors, and different dimensions of eigenvectors would influence the final clustering performance due to the impact of noise. To observe the diverse information between different dimensions intuitively, the band-correlation matrix corresponding to different $\mathbf{Y}_{(i)}$ on the Botswana dataset is visualized in Fig. 2. As can be seen in Fig. 2, when the latent feature dimension is reduced from 6 to 3, the advantageous information in the top left and bottom right of the figure is lost, which indicates that the drop in feature dimension would significantly change the correlation among bands. Thus, in order to distill and preserve this kind of dimension-specific information, a hierarchical strategy is adopted to improve the representation capability of the generated latent features with fixed dimensions. Specifically, to obtain latent features with d dimension, intermediary latent feature matrices $\{\mathbf{Y}_{(i)}^{(1)}\}_{i=1}^{N} \in \mathbb{R}^{b \times d_1}$ are constructed via (3), where $b > d_1 > d$. Then, the second intermediary latent feature matrices $\{\mathbf{Y}_{(i)}^{(2)}\}_{i=1}^{N} \in \mathbb{R}^{b \times d_2}$ are derived from the first intermediary latent matrices, where $b > d_1 > d_2 > d$. Repeating the above process m times, the final latent feature dimension is gradually reduced to d, where $b > d_1 > d_2 > \cdots > d_{m-1} > d$. Without loss of generality, the problem can be expressed as

$$\max_{\mathbf{Y}_{(i)}^{(t)}} \quad \text{Tr}\left(\mathbf{Y}_{(i)}^{(t-1)} \mathbf{Y}_{(i)}^{(t-1)\top} \mathbf{Y}_{(i)}^{(t)} \mathbf{Y}_{(i)}^{(t)\top}\right)$$
$$\text{s.t.} \quad \mathbf{Y}_{(i)}^{(t)\top} \mathbf{Y}_{(i)}^{(t)} = \mathbf{I}_{d_t}, \mathbf{Y}_{(i)}^{(t)} \in \mathbb{R}^{b \times d_t}. \quad (4)$$

In such a manner, the latent features with predefined dimension are well distilled from each region.

D. Multiregional Latent Feature Fusion

For an HSI, the spatial information often exists in the pixels. According to the ground truth of each hyperspectral dataset, we can find that different regions present diverse

characteristics, which indicate that the spatial information of each region is different. That is to say, the pixels of diverse regions are different. In addition, we can also observe that the land covers are continuous in spatial distribution and the adjacent pixels show similar spectral features. Thus, one rational assumption is that pixels from one homogeneous region are more likely to be the same class, and each region carries different information and contributes differently to the selection of bands. In order to simultaneously capture the supplementary information and common information of various graphs, the most popular method is constructing a unified feature matrix by fusing the features of all graphs in clustering, such as [23] and [57]–[61]. Inspired by the above methods, the unified matrix for HSIs is constructed as follows:

$$
\max_{\mathbf{F}, \beta_i} \sum_{i=1}^{N} \beta_i \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(m)} \mathbf{Y}_{(i)}^{(m)\top} \mathbf{F}\mathbf{F}^\top\right)
$$

$$
\text{s.t. } \mathbf{F}^\top \mathbf{F} = \mathbf{I}_d, \sum_{i=1}^{N} \beta_i^2 = 1, \beta_i \geq 0 \tag{5}
$$

where β_i is the weight corresponding to the ith region, and $\mathbf{F} \in \mathbb{R}^{b \times d}$ presents the unified feature matrix with d dimension.

E. Prior Information Constraint

Since each region corresponds to the same type of land cover though the region would have different spectral properties in different bands, the underlying structure should be the same. This prior information is adopted to regularize the construction of the unified latent feature matrix, that is

$$
\max_{\mathbf{F}} \sum_{i=1}^{N} \lambda \mathrm{Tr}\left(\overline{\mathbf{F}}\ \overline{\mathbf{F}}^\top \mathbf{F}\mathbf{F}^\top\right) \quad \text{s.t. } \mathbf{F}^\top \mathbf{F} = \mathbf{I}_d \tag{6}
$$

where $\overline{\mathbf{F}} \in \mathbb{R}^{b \times d}$ denotes the average latent feature matrix of HSIs, and λ is a parameter. Furthermore, the solution for $\overline{\mathbf{F}}$ is similar to $\mathbf{Y}_{(i)}$ in (3), that is, via the spectral embedding for the average Laplacian matrix of the whole HSIs. Mathematically, the problem is modeled as follows:

$$
\min_{\overline{\mathbf{F}}} \sum_{i=1}^{N} \mathrm{Tr}\left(\overline{\mathbf{F}}^\top \mathbf{L}_{(i)}\overline{\mathbf{F}}\right) \quad s.t.\ \overline{\mathbf{F}}^\top \overline{\mathbf{F}} = \mathbf{I}_d. \tag{7}
$$

F. Overall Objective Function

Combining the items expressed in (3)–(6), the overall objective function is

$$
\max_{\mathbf{Y}, \mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\beta}} \sum_{i=1}^{N} -\gamma_i^{(1)} \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(1)\top} \mathbf{L}_{(i)}\mathbf{Y}_{(i)}^{(1)}\right)
$$

$$
+ \sum_{t=2}^{m} \sum_{i=1}^{N} \gamma_i^{(t)} \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(t-1)} \mathbf{Y}_{(i)}^{(t-1)\top} \mathbf{Y}_{(i)}^{(t)} \mathbf{Y}_{(i)}^{(t)\top}\right)
$$

$$
+ \sum_{i=1}^{N} \beta_i \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(m)} \mathbf{Y}_{(i)}^{(m)\top} \mathbf{F}\mathbf{F}^\top\right) + \beta_{N+1} \mathrm{Tr}\left(\overline{\mathbf{F}}\ \overline{\mathbf{F}}^\top \mathbf{F}\mathbf{F}^\top\right)
$$

$$
\text{s.t. } \mathbf{F}^\top \mathbf{F} = \mathbf{I}_d, \mathbf{Y}_{(i)}^\top \mathbf{Y}_{(i)} = \mathbf{I}, \sum_{i=1}^{N} \gamma_i^{(t)2} = \sum_{i=1}^{N+1} \beta_i^2 = 1
$$

$$
\gamma_i \geq 0, \beta_i \geq 0 \tag{8}
$$

where γ^t is the weight corresponding to each region in the tth layer. Obviously, the parameter λ can be automatically determined like β_i. To simply the expression, let β_{N+1} be λ in (8). For (8), it is worth noting that the intermediary latent feature matrices and the unified feature matrix reinforce each other. With the latent feature dimension decreasing in steps, the intermediary matrices are utilized to update the unified feature matrix $\mathbf{F}$, and then $\mathbf{F}$ will go back to boost the latent features distilling process. Thus, $\mathbf{F}$ can be tuned automatically for optimal hyperspectral band selection.

With the unified feature matrix $\mathbf{F}$ of HSIs, the k-means algorithm is utilized to generate multiple clusters. The band with the maximum information entropy from each cluster is selected as the feature band.

IV. OPTIMIZATION ALGORITHM

As seen from (8), the proposed model contains four variables $\mathbf{Y}$, $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$. Since it is difficult to directly solve each variable at once, we consider to solve it by an iteration algorithm. Instead of using existing iteratively optimization algorithms to get the optimal solutions, such as ALM [62] and its variant ADMM [63], we design a simpler optimization algorithm without introducing any parameters to derive the solution with respect to each variable.

A. Optimizing $\mathbf{Y}$

1) Optimizing $\{\mathbf{Y}_{(i)}^{(1)}\}_{i=1}^N$: When $\{\mathbf{Y}_{(i)}^{(t)}\}_{t=2}^m$, $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are fixed, (8) can be simplified to

$$
\max_{\left\{\mathbf{Y}_{(i)}^{(1)}\right\}_{i=1}^N} \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(1)\top}\left(-\gamma_i^{(1)} \mathbf{L}_{(i)} + \gamma_i^{(2)} \mathbf{Y}_{(i)}^{(2)} \mathbf{Y}_{(i)}^{(2)\top}\right)\mathbf{Y}_{(i)}^{(1)}\right)
$$

$$
\text{s.t. } \mathbf{Y}_{(i)}^{(1)\top} \mathbf{Y}_{(i)}^{(1)} = \mathbf{I}_{d_1}. \tag{9}
$$

2) Optimizing $\{\mathbf{Y}_{(i)}^{(t)}\}_{t=2}^{m-1}$: When $\{\mathbf{Y}_{(i)}^{(1)}\}_{i=1}^N$, $\{\mathbf{Y}_{(i)}^{(m)}\}_{i=1}^N$, $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are fixed, (8) can be reformulated as

$$
\max_{\mathbf{Y}_{(i)}^{(t)}} \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(t)\top}\left(\gamma_i^{(t)} \mathbf{Y}_{(i)}^{(t-1)} \mathbf{Y}_{(i)}^{(t-1)\top}\right)\mathbf{Y}_{(i)}^{(t)}\right)
$$

$$
+ \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(t)\top}\left(\gamma_i^{(t+1)} \mathbf{Y}_{(i)}^{(t+1)} \mathbf{Y}_{(i)}^{(t+1)\top}\right)\mathbf{Y}_{(i)}^{(t)}\right)
$$

$$
\text{s.t. } \mathbf{Y}_{(i)}^{(t)\top} \mathbf{Y}_{(i)}^{(t)} = \mathbf{I}_{d_t}, t \in \{2, \ldots, (m-1)\}, i \in \{1, \ldots, N\}. \tag{10}
$$

3) Optimizing $\{\mathbf{Y}_{(i)}^{(m)}\}_{i=1}^N$: When $\{\mathbf{Y}_{(i)}^{(t)}\}_{t=1}^{m-1}$, $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are fixed, (8) can be rewritten as

$$
\max_{\left\{\mathbf{Y}_{(i)}^{(m)}\right\}_{i=1}^N} \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(m)\top}\left(\gamma_i^{(m)} \mathbf{Y}_{(i)}^{(m-1)} \mathbf{Y}_{(i)}^{(m-1)\top} + \beta_i \mathbf{F}\mathbf{F}^\top\right)\mathbf{Y}_{(i)}^{(m)}\right)
$$

$$
\text{s.t. } \mathbf{Y}_{(i)}^{(m)\top} \mathbf{Y}_{(i)}^{(m)} = \mathbf{I}_{d_m}. \tag{11}
$$

B. Optimizing $\mathbf{F}$

Given $\mathbf{Y}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$, (8) can be simplified to

$$
\max_{\mathbf{F}} \sum_{i=1}^{N} \beta_i \mathrm{Tr}\left(\mathbf{Y}_{(i)}^{(m)} \mathbf{Y}_{(i)}^{(m)\top} \mathbf{F}\mathbf{F}^\top\right) + \beta_{N+1} \mathrm{Tr}\left(\overline{\mathbf{F}}\ \overline{\mathbf{F}}^\top \mathbf{F}\mathbf{F}^\top\right)
$$

$$
\text{s.t. } \mathbf{F}^\top \mathbf{F} = \mathbf{I}_d. \tag{12}
$$

The above objective functions, from (9) to (12), are all can be generalized into

$$\max_{\mathbf{U}} \operatorname{Tr}\left(\mathbf{U}^{\top} \mathbf{V} \mathbf{U}\right) \quad \text{s.t. } \mathbf{U}^{\top} \mathbf{U} = \mathbf{I}_d. \tag{13}$$

Equation (13) can be solved by performing eigenvalue decomposition on $\mathbf{V}$. Specifically, the optimal solution $\mathbf{U}$ is generated by the eigenvectors of $\mathbf{V}$ corresponding to the first d largest eigenvalues.

C. Optimizing $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

Given $\mathbf{Y}$, $\mathbf{F}$, and $\boldsymbol{\beta}$, the optimization problem of (8) is equivalent to

$$\max_{\boldsymbol{\gamma}} \ \sum_{i=1}^{N} \gamma_i^{(t)} \sigma_i^{(t)}$$

$$\text{s.t. when } t = 1, \sigma_i^{(1)} = -\operatorname{Tr}\left(\mathbf{Y}_{(i)}^{(1)\top} \mathbf{L}_{(i)} \mathbf{Y}_{(i)}^{(1)}\right)$$

$$\text{when } 2 \le t \le m, \sigma_i^{(t)} = \operatorname{Tr}\left(\mathbf{Y}_{(i)}^{(t-1)} \mathbf{Y}_{(i)}^{(t-1)\top} \mathbf{Y}_{(i)}^{(t)} \mathbf{Y}_{(i)}^{(t)\top}\right)$$

$$1 \le t \le m, \sum_{i=1}^{N} \gamma_i^{(t)2} = 1, \gamma_i \ge 0. \tag{14}$$

Considering that for $\forall t$, according to the Cauchy inequality, we can derive $\sum_{i=1}^{N} \gamma_i \sigma_i \le \sqrt{\sum_{i=1}^{N} \gamma_i^2 \sum_{i=1}^{N} \sigma_i^2} = \sqrt{\sum_{i=1}^{N} \sigma_i^2}$. When the equal sign holds, $(\gamma_1/\sigma_1) = (\gamma_2/\sigma_2) = \cdots = (\gamma_N/\sigma_N) = k$, we can derive that $\sum_{i=1}^{N} \gamma_i^2 = k^2 \sum_{i=1}^{N} \sigma_i^2 = 1$ and $k = \sqrt{(1/\sum_{i=1}^{N} \sigma_i^2)}$. So $\gamma_i = ([\sigma_i]/[\sqrt{\sum_{i=1}^{N} \sigma_i^2}])$ is the optimal solution of (14).

Furthermore, the solution for $\boldsymbol{\beta}$ is similar to $\boldsymbol{\gamma}$.

D. Convergence Analysis

In order to simplify the expression, (8) can be denoted as

$$\max_{\mathbf{Y}, \mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\beta}} \Theta\left(\mathbf{Y}_{(i)}^{(t)}, \mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\beta}\right). \tag{15}$$

For each iteration of the above optimization, we can derive that

$$\Theta\left(\left\{\mathbf{Y}_{(i)}^{(t)}\right\}^{s}, \mathbf{F}, \{\boldsymbol{\gamma}\}^{s}, \{\boldsymbol{\beta}\}^{s}\right)$$
$$\le \Theta\left(\left\{\mathbf{Y}_{(i)}^{(t)}\right\}^{s+1}, \mathbf{F}, \{\boldsymbol{\gamma}\}^{s+1}, \{\boldsymbol{\beta}\}^{s+1}\right) \tag{16}$$

where superscript s denotes the optimization at the sth iteration. By optimizing one variable and fixing others, the optimal solution can be obtained, thus (16) holds. Based on (16), the following similar inequality also holds for each iteration:

$$\Theta\left(\left\{\mathbf{Y}_{(i)}^{(t)}\right\}^{s}, \{\mathbf{F}\}^{s}, \{\boldsymbol{\gamma}\}^{s}, \{\boldsymbol{\beta}\}^{s}\right)$$
$$\le \Theta\left(\left\{\mathbf{Y}_{(i)}^{(t)}\right\}^{s+1}, \{\mathbf{F}\}^{s+1}, \{\boldsymbol{\gamma}\}^{s+1}, \{\boldsymbol{\beta}\}^{s+1}\right). \tag{17}$$

Consequently, (8) monotonically increases at each iteration. Furthermore, (8) is upper bounded. Therefore, the alternative iterative algorithm can be guaranteed to converge.

Algorithm 1 Hyperspectral Band Selection via HLFC

Input: Hyperspectral image cube $\mathbf{H} \in \mathbb{R}^{w \times h \times b}$, the number of selected bands k, the latent feature dimension d.
Output: The selected feature band subset $\mathbf{H}^{*}$.
1: Segment HSIs via PCA and superpixel segmentation.
2: Calculate the Laplacian matrix $\mathbf{L}_{(i)}$ of each segmented region via Eq. (2).
3: Initialize $\mathbf{Y}$ according to Eq. (3).
4: Initialize $\{\gamma_i\}_{i=1}^{N} = \frac{1}{\sqrt{N}}$, $\{\beta_i\}_{i=1}^{N+1} = \frac{1}{\sqrt{N+1}}$ and $t = 1$.
5: Calculate the information entropy for each band.
6: **while true**
7:   Update $\mathbf{F}$ by solving Eq. (12).
8:   Update $\mathbf{Y}$ by solving Eq. (9)-Eq. (11).
9:   Update $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ by solving Eq. (14).
10:   $t = t + 1$.
11: **end while** if $\frac{(obj^t - obj^{t-1})}{obj^{t-1}} \le 10^{-6}$ or $t \ge 100$.
12: Employ k-means algorithm on the unified latent feature matrix $\mathbf{F}$ to generate all clusters.
13: Select the band with maximum information entropy as the feature band from each cluster.
14: **Return** the selected feature band subset $\mathbf{H}^{*}$.

E. Time Complexity Analysis

The time complexity of HLFC mainly lies in the superpixel segmentation, solutions to $\mathbf{Y}$, $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$. For HSIs, the computation complexity of ERS is $\mathcal{O}(P\log P)$, where P denotes the total number of pixels in each band. For updating $\mathbf{Y}$, from (9) to (11), it costs $\mathcal{O}(b^3)$ in each iteration, where b represents the number of bands. For updating $\mathbf{F}$, that is, (12), it costs $\mathcal{O}(b^3)$ by adopting SVD on matrices of size $\mathbb{R}^{b \times b}$ in each iteration. For updating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, it costs $\mathcal{O}(Nm)$ and $\mathcal{O}(N)$, respectively. N represents the number of segmented regions, and m is the number of layers. Finally, the k-means costs $\mathcal{O}(bk^2)$, where k is the number of selected bands. Overall, the total time complexity of HLFC is $\mathcal{O}(P\log P + 2b^3 + bk^2 + N(m+1))$.

In summary, the details for solving HLFC are listed in Algorithm 1.

V. EXPERIMENTS

Extensive experiments are conducted on four public hyperspectral datasets to verify the effectiveness of HLFC. Results used different classifiers and different metrics are reported and compared to the state-of-the-art algorithms.

A. Experimental Setup

1) Datasets: Four public HSIs datasets are used for experiments, including Indian Pines Scene, Kennedy Space Center, Botswana, and Salinas.

Indian Pines: This dataset was captured by AVIRIS sensor over the Northwestern Indian in 1992. It contains 220 spectral bands, and each band is composed of 145×145 pixels. In addition, there are 16 different classes of land covers for each band, and the spatial resolution of this dataset is 20-m pixels. Due to the water absorption, we select 200 bands and discard

TABLE I
SUMMARY OF FOUR PUBLIC HYPERSPECTRAL DATASETS

Dataset	Pixels	Bands	Classes	Spatial resolution
Indian Pines	145×145	200	16	20-m pixels
KSC	512×614	176	13	18-m pixels
Botswana	1476×256	145	14	30-m pixels
Salinas	512×217	204	16	3.7-m pixels

20 spectral bands with the index of 104–108, 150–163, and 200 in the experiment.

KSC: This dataset was acquired by AVIRIS sensor over the Kennedy Space Center, Florida, in 1996. It contains 224 spectral bands, and each band consists of 512×614 pixels with 13 different classes of land covers. The spatial resolution of this dataset is 18-m pixels. In the experiment, we discard 48 bands due to water absorption.

Botswana: This dataset was obtained by NASA EO-1 satellite sensors over Okavango Delta, Botswana, between 2001 and 2004. It contains 242 spectral bands with the size of 1476×256 pixels, and there are 14 different classes of land covers. The spatial resolution of this dataset is 30-m pixels. In the experiment, we select 145 spectral bands by discarding 97 bands for the same reason as for other datasets.

Salinas: This dataset was collected by AVIRIS sensor in California. It consists of 224 spectral bands with the size of 512×217 pixels, and its spatial resolution of it is 3.7-m pixels. Each band contains 16 different classes of land covers. We select 204 spectral bands for the experiment by removing 20 bands due to water absorption.

In a word, the detailed information of these datasets is summarized in Table I.

2) Compared Methods: To verify the effectiveness of HLFC, the proposed method is compared to the state-of-the-art algorithms as briefly described as follows.

Uniform band selection (UBS) [42] partitions the whole HSIs uniformly, and the index of each segmented point is selected.

E-FDPC [64] groups the bands into clusters and then selects the feature bands by maximizing their intercluster distance and local density.

TOF [65] employs dynamic programming to obtain an optimal segmentation combination of HSIs and selects the band as the feature band according to [64] in each subcube.

OPBS [66] selects a band with maximum variance first, and then the band maximized the orthogonal projection to subspace is selected.

ASPS_MN [24] adopts a search algorithm to generate the optimal subspace. In each subspace, the band with maximum information entropy or minimum noise value is selected.

ONR [4] searches the optimal band combination by minimizing the reconstructing error of HSIs data.

FNGBS [67] adopts a coarse to fine method to segment HSIs into multiple cube, and the band is selected via the product of local density and information entropy.

3) Classification Settings: Classification experiments are conducted to verify the effectiveness of the proposed method

and all compared algorithms on the selected subset of bands. Three classical classifiers, for example, support vector machine (SVM), random forest (RF), and spatial–spectral kernel sparse representation classifier (KSRC) [68], are adopted to examine the classification performance. Since these classifiers are all supervised classification, for Indian_Pines, KSC, Botswana, and Salinas, we randomly choose 5%, 1% 1%, and 1% of the whole dataset as training samples, respectively. To the best of our knowledge, the optimal number of selected bands is still unknown. To make use of the advantages of competitors on hyperspectral band selection, as well as facilitate experimental comparisons, we remain the same number of selected bands as in their paper [69], that is, it is set to range from 5 to 50 every 5 intervals. Furthermore, three metrics are employed to estimate the precision of the classified pixels on three classifiers, including overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa). To reduce the randomness of the final results, we conduct each experiment ten times and the averages and standard deviations of these metrics are reported. The latent feature dimension d is set to 5 for all datasets. The hierarchical level m is set to two and each level's corresponding intermediary dimensions are set to $\{3d, 4d, 5d, \ldots, 10d\}$ and $\{2d, 3d, 4d, \ldots, 10d\}$, respectively. In order to determine the optimal intermediary feature dimension, grid search is employed in the experiments. Experiments are conducted on Windows10 with an Intel Core processor, 24-GB RAM, and MATLAB 2020a.

Segmentation of HSIs plays an important role in band selection. As far as we know, the determination of the optimal number of segments N is still an open problem, and it is determined experimentally. In [56], the textural information of the first principal component of HSIs is exploited to determine the number of segmented regions, and the objective function is constructed as follows:

$$N = T \times \frac{N_z}{P} \qquad (18)$$

where N_z denotes the number of nonzero values at the edge of the first principal component of HSIs, P denotes the total number of pixels of each band, and T is a fixed number. Followed this work, we can adaptively set the number of superpixels for all datasets by setting $T = 750$.

B. Results and Analysis

Table II lists OA, AA, and Kappa of the proposed method and the compared methods on the four public hyperspectral datasets with three different classifiers. In the experiments, the number of feature bands is set to 15, and the results include the average values of these indicators and their standard deviations over multiple runs of each method. The best results are highlighted in bold. As seen from Table II, the proposed method outperforms other competitors in most cases.

Since the optimal number of feature bands is difficult to determine in practice. Thus, we show the OA curves for each dataset with the number of feature bands from 5 to 50 with an interval of 5. According to Figs. 3–6, the following can be observed.

TABLE II
RESULTS OF ALL COMPETITORS ON FOUR PUBLIC HYPERSPECTRAL DATASETS (%)

Dataset	Classifier	Metric	UBS	E-FDPC	TOF	OPBS	ASPS_MN	ONR	FNGBS	HLFC
Indian Pines	SVM	OA	69.13±0.50	60.53±1.69	71.80±0.60	65.61±0.95	71.96±0.80	71.44±0.66	70.71±0.49	**72.63±0.66**
		AA	59.95±1.84	46.49±1.13	63.43±1.79	57.58±2.73	64.96±0.97	64.15±1.04	63.28±0.84	**66.95±0.81**
		Kappa	66.73±0.51	57.38±1.86	69.52±0.63	63.17±1.02	69.73±0.86	69.13±0.68	68.33±0.52	**70.43±0.68**
	RF	OA	53.23±1.88	60.56±0.44	53.42±2.29	62.73±0.85	53.34±2.31	53.82±1.40	54.95±1.27	**66.33±1.14**
		AA	35.92±1.78	45.26±1.16	35.63±2.39	50.85±0.97	35.30±3.22	36.67±0.90	37.19±1.31	**52.72±1.20**
		Kappa	50.28±1.98	57.88±0.44	50.55±2.20	60.15±0.84	50.53±2.40	50.94±1.38	52.11±1.24	**63.84±1.15**
	KSRC	OA	94.65±0.85	93.40±0.58	94.90±0.64	94.73±0.85	94.03±1.34	93.50±0.19	94.83±0.63	**95.50±1.25**
		AA	97.02±0.01	96.57±0.22	97.36±0.43	97.06±0.52	96.86±0.54	96.91±0.29	97.34±0.37	**97.55±0.57**
		Kappa	93.98±0.92	92.60±0.63	94.26±0.71	94.07±0.94	93.29±1.46	92.72±0.22	94.17±0.69	**94.92±1.38**
KSC	SVM	OA	59.02±2.64	52.94±2.86	59.85±1.53	43.08±2.37	59.69±2.04	56.52±3.07	61.13±2.33	**61.79±2.46**
		AA	39.99±2.80	33.21±2.77	40.74±2.12	25.10±1.88	41.23±2.09	36.80±2.96	42.14±2.84	**42.89±3.11**
		Kappa	56.00±2.75	49.71±2.95	56.86±1.62	39.57±2.41	56.74±2.14	53.39±3.16	58.20±2.44	**58.86±2.79**
	RF	OA	68.24±1.71	63.33±2.46	68.19±1.54	50.42±1.84	66.48±2.97	65.74±1.32	67.55±2.39	**68.54±3.24**
		AA	55.56±2.47	49.54±2.53	55.74±2.71	36.02±2.40	52.94±3.40	53.43±2.39	54.46±2.53	**56.07±3.32**
		Kappa	65.84±1.78	60.73±2.47	65.78±1.62	47.55±1.82	63.96±3.06	63.25±1.36	65.10±2.49	**66.16±3.44**
	KSRC	OA	83.09±3.14	80.40±3.97	83.35±2.93	55.93±3.96	82.33±3.38	84.75±3.15	85.48±2.11	**87.68±3.42**
		AA	78.64±3.52	74.69±4.91	80.56±1.25	47.95±2.03	77.45±3.24	79.89±4.44	80.65±2.79	**83.25±4.69**
		Kappa	81.59±3.34	78.70±4.21	81.91±3.08	53.29±3.78	80.78±3.55	83.32±3.40	84.11±2.26	**86.49±3.70**
Botswana	SVM	OA	65.89±4.24	65.11±2.41	66.53±4.96	68.89±1.88	68.62±4.42	70.71±3.21	67.29±3.09	**71.49±3.36**
		AA	62.24±4.04	61.28±2.49	63.46±4.73	64.90±2.17	65.39±4.89	67.39±3.25	63.24±3.16	**68.40±3.83**
		Kappa	64.43±4.30	63.62±2.45	65.13±4.99	67.47±1.91	67.20±4.51	69.32±3.26	65.86±3.14	**70.13±3.45**
	RF	OA	61.45±3.35	58.67±3.45	59.54±2.58	56.26±3.42	58.19±4.78	61.13±4.27	60.50±3.80	**61.62±2.69**
		AA	57.29±4.11	55.13±3.55	55.57±3.30	51.75±3.31	55.13±5.07	56.37±3.96	56.87±3.74	**57.46±2.40**
		Kappa	59.96±3.39	57.17±3.46	58.03±2.61	54.72±3.45	56.66±4.83	59.62±4.31	59.00±3.88	**60.12±2.72**
	KSRC	OA	86.92±3.76	89.44±2.98	91.98±1.59	85.71±3.38	91.70±2.67	88.42±3.86	90.67±2.58	**92.20±1.12**
		AA	87.14±3.13	89.04±3.71	91.59±1.65	85.87±3.46	91.83±2.23	88.41±3.50	90.40±2.85	**91.61±2.38**
		Kappa	86.09±3.94	88.72±3.15	91.41±1.69	84.82±3.54	91.12±2.82	87.67±4.05	90.03±2.74	**91.64±1.19**
Salinas	SVM	OA	86.82±0.42	88.71±0.35	88.74±0.54	88.24±0.58	87.38±0.70	88.96±0.48	88.51±0.41	**89.01±0.31**
		AA	91.79±0.35	92.12±0.68	92.18±0.90	91.60±1.03	91.01±1.27	92.48±0.45	92.01±0.53	**92.71±0.48**
		Kappa	86.82±0.45	87.75±0.37	87.76±0.58	87.24±0.63	86.32±0.75	88.01±0.52	87.52±0.44	**88.06±0.34**
	RF	OA	84.86±0.36	85.18±0.56	85.22±0.50	84.46±0.51	85.12±0.52	85.31±0.68	85.02±0.74	**85.67±0.52**
		AA	89.09±0.43	89.77±0.89	89.49±0.76	88.16±0.76	89.46±0.69	89.41±0.71	89.27±1.12	**90.44±0.62**
		Kappa	83.67±0.37	84.03±0.60	84.07±0.53	83.25±0.53	83.95±0.55	84.15±0.72	83.85±0.78	**84.54±0.57**
	KSRC	OA	97.33±0.39	97.97±0.35	98.08±0.28	95.95±0.76	97.00±0.88	97.82±0.41	97.44±0.45	**98.21±0.63**
		AA	98.83±0.25	99.04±0.11	99.10±0.10	98.35±0.26	98.61±0.36	99.05±0.19	98.83±0.00	**99.19±0.19**
		Kappa	97.05±0.43	97.76±0.38	97.87±0.31	95.55±0.82	96.69±0.96	97.59±0.45	97.17±0.49	**98.02±0.69**
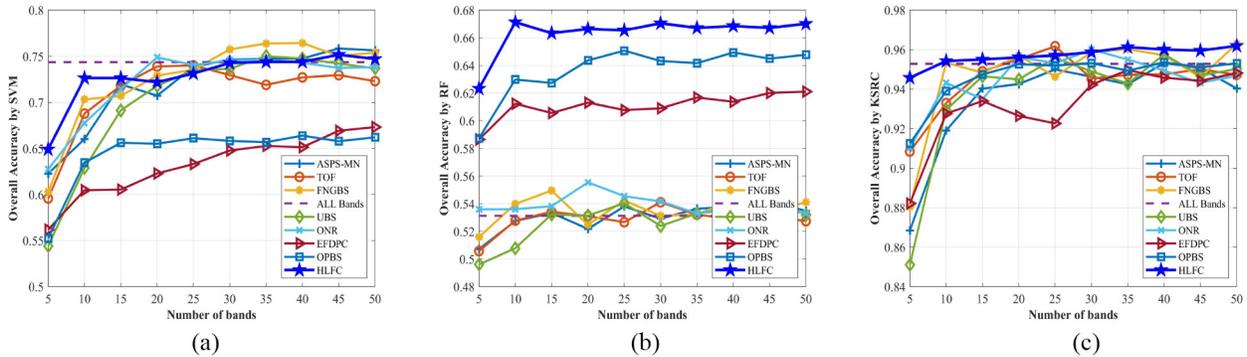
Fig. 3. OA of all competitors on the Indian Pines dataset. (a) Indian Pines-SVM. (b) Indian Pines-RF. (c) Indian Pines-KSRC.

1) The classification results of our proposed method outperform that of the compared methods, especially when the number of selected features band is small. For instance, taking Indian Pines dataset as an example, the proposed method achieves over 2% and 3% improvement compared with the second performer FNGBS and OPBS in
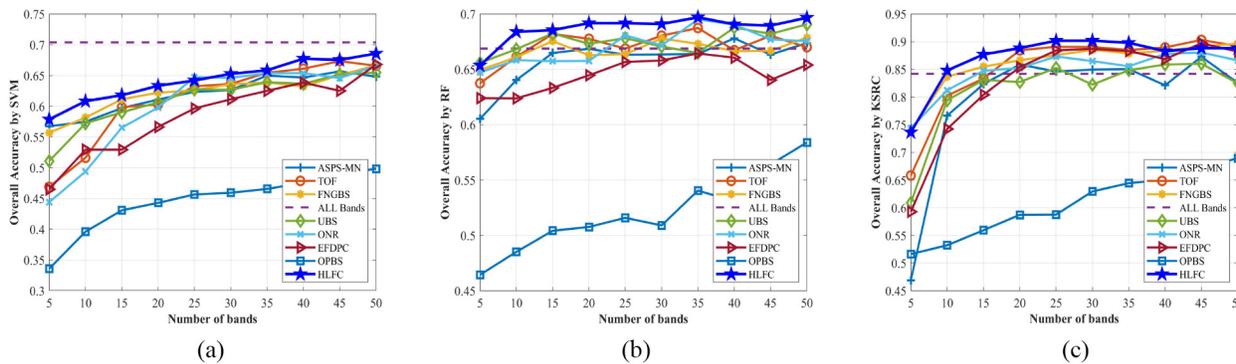
Fig. 4. OA of all competitors on the KSC dataset. (a) KSC-SVM. (b) KSC-RF. (c) KSC-KSRC.
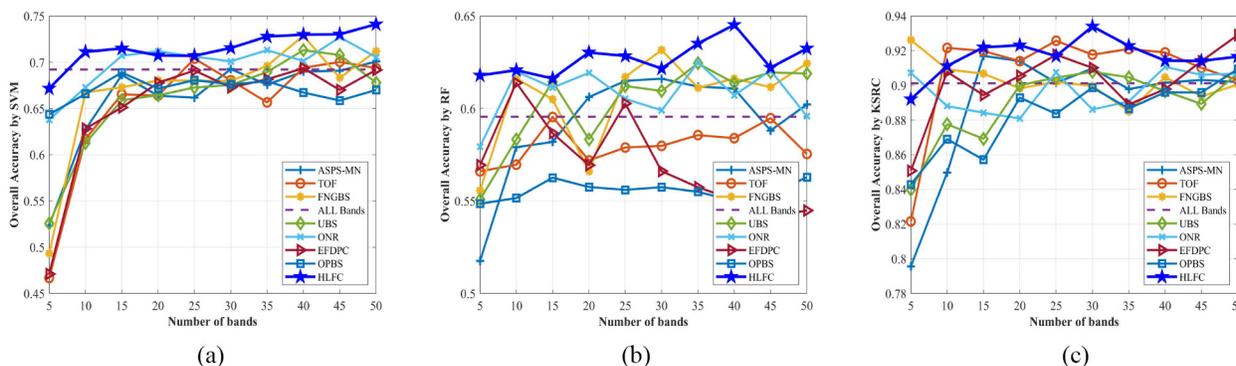
Fig. 5. OA of all competitors on the Botswana dataset. (a) Botswana-SVM. (b) Botswana-RF. (c) Botswana-KSRC.
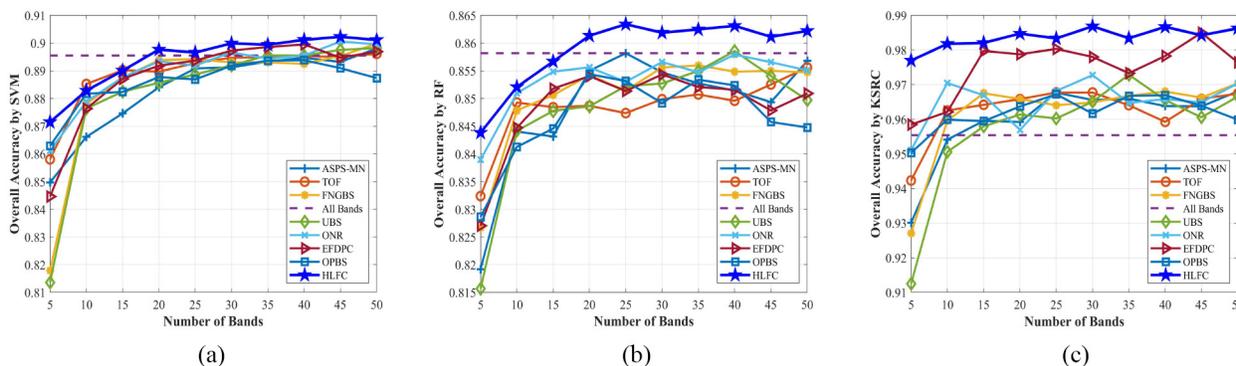
Fig. 6. OA of all competitors on the Salinas dataset. (a) Salinas-SVM. (b) Salinas-RF. (c) Salinas-KSRC.

SVM and RF classifiers, respectively. This demonstrates the effectiveness of our method. The primary reason can be attributed to the learning of latent feature representations and the utilization of spatial information of HSIs.

2) When the number of selected feature bands is small, the proposed method can achieve superior performance by using discriminative latent features, while most competitors using the original pixelwise features do not perform satisfyingly. In other words, these methods often lead to the selection of an increased number of feature bands in order to improve the final accuracy. For example, in these OA curves, the proposed method consistently outperforms other algorithms on four datasets with different classifiers when the number of feature bands is 10. As to

the other number of selected feature bands, the proposed method still achieves superior performance. When five feature bands are selected, compared with the second performer ONR, the proposed method yields more than 3% improvement on the Botswana dataset using SVM and RF classifiers, respectively, as seen in Fig. 5. In a word, the results have shown that the proposed method is a promising algorithm for hyperspectral band selection.

3) The proposed method has gained obvious improvement over other methods based on spectral property alone, such as ASPS_MN, FNGBS, and ONR. As shown in OA curves on the KSC dataset, the proposed method keeps performing better than comparable competitors when the number of selected bands increases. The compared methods do not pay sufficient attention to the spatial
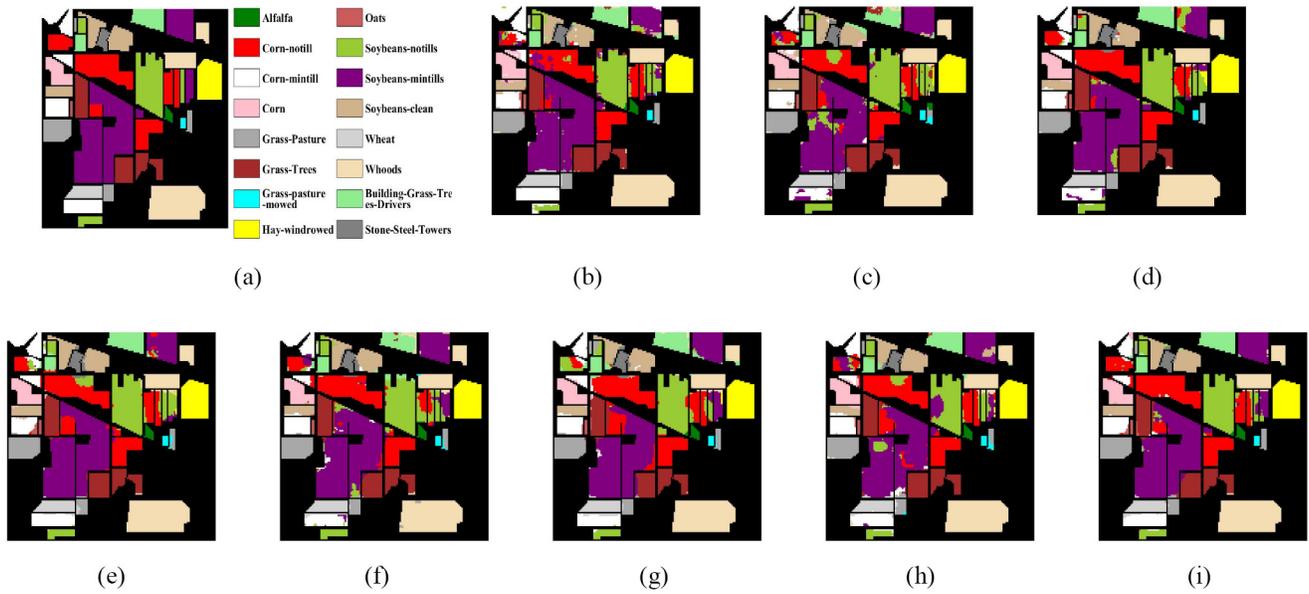
Fig. 7. Visual classification maps of all baselines on the Indian Pines dataset. (a) Ground truth. (b) UBS. (c) E-FDPC. (d) TOF. (e) OPBS. (f) ASPS_MN. (g) ONR. (h) FNGBS. (i) HLFC.
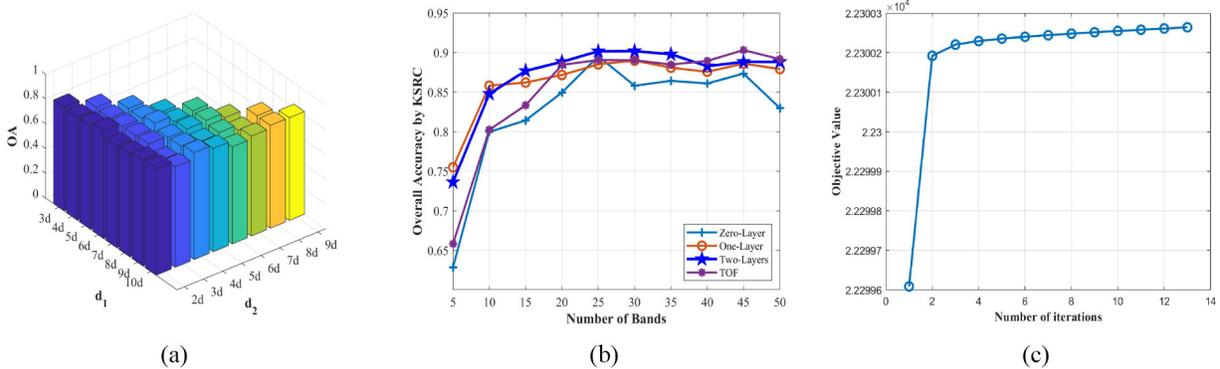
Fig. 8. (a) OA on the KSC dataset for different intermediary dimension. (b) OA on the proposed method with different number of layers and TOF. (c) Objective function value of HLFC on the KSC dataset.

information, resulting in unsatisfying performance while a fewer number of feature bands are selected. On the contrary, the proposed method can select the most discriminative bands and yield significant performance as a result of exploiting the spatial information of HSIs. As seen, compared with the second performer FNGBS, when the number of selected bands is 10, the proposed method improves OA by 2%, 2%, and 1% on the KSC dataset using SVM, RF, and KSRC classifiers, respectively, as seen in Fig. 4. In terms of AA and Kappa, the proposed method also can perform better than competitors. For example, as seen in Table II, the AA and Kappa value of the proposed method are 83.25% and 86.49%, respectively, on the KSC dataset with the KSRC classifier, while the second performer FNGBS only achieves 80.65% and 84.11%, respectively. Thus, these results have verified the advantages of the proposed method.

Fig. 7 shows the classification maps of all methods which have clearly shown their classification performance on the Indian Pines dataset.

C. Parameter and Convergence Study

As seen from (8), the proposed method only has one parameter m, that is, the size of layers. It should be noted that the intermediary feature dimension $\{d_i\}_{i=1}^m$ is fixed. That is, its size only depends on the latent feature dimension d, and its maximum value, and the minimum value is $10d$, $2d$, respectively. $d = 5$ works well for the four datasets in our experiments. In order to further study the influence of intermediary matrix dimension, the performance of HLFC with different intermediary latent feature dimensions on the KSC dataset is presented as shown in Fig. 8(a). To the best of our knowledge, the optimal intermediate feature dimension for each dataset is difficult to determine, so grid search is adopted in the experiments. From Fig. 8(a), we can find that the intermediary dimension has a strong impact on the final classification accuracy, and OA will improve when larger scale d_1 and d_2 are assigned. Furthermore, we compare the performance of the model with 0, 1, and 2 layers. Results are shown in Fig. 8(b). As seen, two layers outperform others on the KSC dataset with KSRC classification. This indicates that the hierarchical
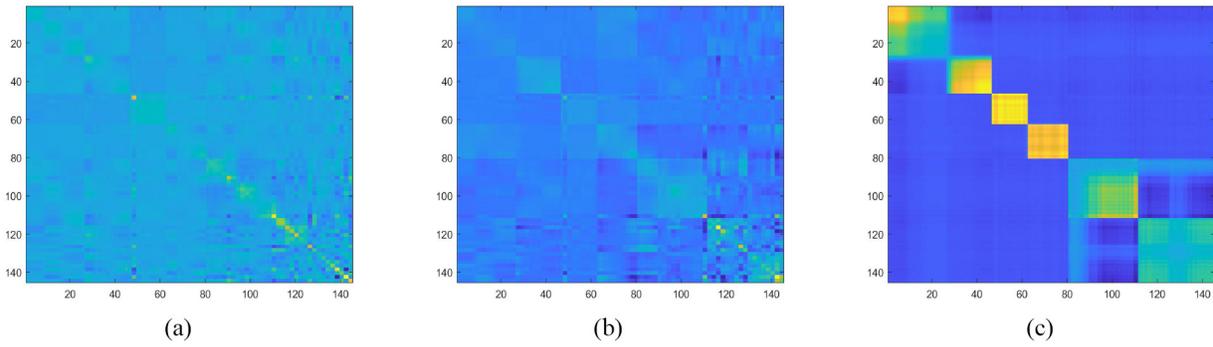
Fig. 9. Visualization of a typical unified latent feature matrix with (a) zero layer, (b) one layer, and (c) two layers on the Botswana dataset.

TABLE III
OA BY SELECTING DIFFERENT NUMBERS OF SUPERPIXELS ON FOUR PUBLIC HYPERSPECTRAL DATASETS WITH USING THE KSRC CLASSIFIER (%)

Datasets	T	N	5	10	15	20	25	30	35	40	45	50
	500	40	65.44	71.22	72.72	72.87	72.91	73.09	73.80	73.83	74.33	75.12
Indian_Pines	750	60	64.90	72.63	72.63	72.18	73.13	74.26	74.39	74.39	75.16	74.70
	1000	80	65.07	72.38	72.95	72.83	74.10	73.78	73.41	73.99	73.88	74.39
	500	45	58.44	60.90	62.32	63.27	64.20	65.18	67.21	66.45	68.09	67.76
KSC	750	58	57.86	60.82	61.79	63.32	64.19	65.26	65.85	67.74	67.53	68.56
	1000	90	57.79	61.79	61.46	63.26	63.73	64.94	66.41	66.74	67.32	67.97
	500	45	68.55	70.36	70.08	71.96	71.43	72.79	72.43	73.10	72.66	73.17
Botswana	750	58	67.16	71.12	71.49	70.73	70.70	71.54	72.78	72.99	73.02	74.10
	1000	90	67.12	71.61	71.33	73.20	72.77	72.74	73.85	71.84	74.38	75.02
	500	25	97.70	98.12	98.04	98.55	98.49	98.80	98.28	98.62	98.42	98.60
Salinas	750	38	97.70	98.18	98.21	98.48	98.34	98.69	98.34	98.67	98.43	98.63
	1000	50	97.70	97.96	98.24	98.52	98.49	98.75	98.33	98.65	98.51	98.65

TABLE IV
OA BY SELECTING DIFFERENT LAYERS ON THE INDIAN PINES DATASET WITH USING THE SVM CLASSIFIER (%)

Datasets	Layers	5	10	15	20	25	30	35	40	45	50
	2	64.90	72.63	72.63	72.18	73.13	74.26	74.39	74.39	75.16	74.70
Indian Pines	3	66.25	72.23	72.32	72.50	73.02	73.77	74.07	74.26	74.57	75.17
	4	65.38	72.66	73.34	72.51	72.44	73.22	73.64	74.01	73.81	75.02
	5	65.25	72.79	72.67	72.49	73.04	73.85	74.76	74.57	74.12	75.38

strategy adopted in the proposed method contributes to the final performance.

In Section IV, we have theoretically proved that the proposed optimization algorithm is convergent. Now, the experiments are conducted to verify its convergence property. In Fig. 8(c), the proposed method can quickly converge with several iterations.

As aforementioned, the optimal number of segmented regions is difficult to be determined, and it is usually set experimental. In order to study the impact of T on the classification accuracy, we set it to 500, 750, and 1000, respectively, and utilize the KSRC classifier to verify the effectiveness of HLFC on four public hyperspectral datasets, as seen in Table III. It can be observed from Table III that the final results are not sensitive to T, and the satisfactory performance can be obtained when $T = 750$.

D. Validation on the Unified Matrix

In Fig. 9, a typical unified latent feature matrix is visualized on the Botswana dataset. As seen, a much clearer clustering structure is shown in Fig. 9(c), hence, the hierarchical strategy is effective to generate the unified latent feature matrix.

E. Study of Different Layers

To the best of our knowledge, the determination of the optimal intermediate dimension is still an open problem, we can only obtain it by adopting a grid searching strategy. As seen from Table IV, we can find that when bigger layers are conducted, more improvements can be obtained. However, considering the model stability and time complexity, the two layers are recommended in the proposed method.

VI. CONCLUSION

In this article, a novel band selection method is proposed, namely, HLFC. HLFC adopts PCA and superpixel segmentation to segment HSIs into multiple regions to capture both spectral and spatial information. Then, latent features are generated to improve feature discrimination and reduce the computational complexity via a hierarchical strategy from each segmented region. These features are fused for band clustering and selection. Extensive experimental results demonstrate that HLFC is effective and outperforms other state-of-the-art methods.

For the HSIs, the two nonadjacent regions may contain strong correlations to some extent. In this article, the proposed method only exploits the local structure information of the whole HSIs while ignoring the nonlocal structure information. Thus, we will focus on how to employ the hybrid hypergraph to integrate the local and nonlocal spatial structure information of HSIs so as to guide the band selection in the future.

ACKNOWLEDGMENT

REFERENCES

[1] L. Ma *et al.*, "Deep learning based classification for head and neck cancer detection with hyperspectral imaging in an animal model," in *Proc. Med. Imag. Biomed. Appl. Mol. Struct. Funct. Imag.*, vol. 10137, 2017, Art. no. 101372G.

[2] B. Ayhan and C. Kwan, "Application of deep belief network to land cover classification using hyperspectral images," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 269–276.

[3] E. Bedini, "The use of hyperspectral remote sensing for mineral exploration: A review," *J. Hyperspectral Remote Sens.*, vol. 7, no. 4, pp. 189–211, 2017.

[4] Q. Wang, F. Zhang, and X. Li, "Hyperspectral band selection via optimal neighborhood reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8465–8476, Dec. 2020.

[5] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.

[6] 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.

[7] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1368–1379, Jul. 2001.

[8] C. Tang *et al.*, "Feature selective projection with low-rank embedding and dual Laplacian regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1747–1760, Sep. 2020.

[9] J. Wang *et al.*, "Hyperspectral band selection via region-aware latent features fusion based clustering," *Inf. Fusion*, vol. 79, pp. 162–173, Mar. 2022.

[10] C. Tang, X. Liu, E. Zhu, L. Wang, and A. Zomaya, "Hyperspectral band selection via spatial–spectral weighted region-wise multiple graph fusion-based spectral clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 19–27.

[11] J. Feng *et al.*, "Convolutional neural network based on bandwise-independent convolution and hard thresholding for hyperspectral band selection," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4414–4428, Sep. 2021.

[12] C. Tang *et al.*, "Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 1, 2021, doi: 10.1109/TKDE.2020.3048678.

[13] A. Datta, S. Ghosh, and A. Ghosh, "Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis," *Int. J. Remote Sens.*, vol. 38, no. 3, pp. 850–873, 2017.

[14] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial–spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[15] A. MartÍnez-UsÓMartinez-Uso, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[16] K. Sun, X. Geng, and L. Ji, "A new sparsity-based band selection method for target detection of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 329–333, Feb. 2014.

[17] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, Jan. 2011.

[18] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.

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

[20] G. Zhu, Y. Huang, S. Li, J. Tang, and D. Liang, "Hyperspectral band selection via rank minimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2320–2324, Dec. 2017.

[21] H. Su, Y. Cai, and Q. Du, "Firefly-algorithm-inspired framework with band selection and extreme learning machine for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 309–320, Jan. 2017.

[22] C. Tang *et al.*, "Learning a joint affinity graph for multiview subspace clustering," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1724–1736, Jul. 2019.

[23] C. Tang, Z. Li, J. Wang, X. Liu, W. Zhang, and E. Zhu, "Unified one-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, early access, May 5, 2022, doi: 10.1109/TKDE.2022.3172687.

[24] Q. Wang, Q. Li, and X. Li, "Hyperspectral band selection via adaptive subspace partition strategy," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4940–4950, Dec. 2019.

[25] Y. Wan, A. Ma, W. He, and Y. Zhong, "Accurate multi-objective low-rank and sparse model for hyperspectral image denoising method," *IEEE Trans. Evol. Comput.*, early access, May 10, 2021, doi: 10.1109/TEVC.2021.3078478.

[26] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7726–7738, Sep. 2021.

[27] R. Yang, L. Su, X. Zhao, H. Wan, and J. Sun, "Representative band selection for hyperspectral image classification," *J. Visual Commun. Image Represent.*, vol. 48, pp. 396–403, Oct. 2017.

[28] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Comput. Vis.*, vol. 3, no. 4, pp. 213–222, 2009.

[29] S. Li, J. Qiu, X. Yang, H. Liu, D. Wan, and Y. Zhu, "A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search," *Eng. Appl. Artif. Intell.*, vol. 27, pp. 241–250, Jan. 2014.

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

[31] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Z. Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data," *IEEE Tran. Cybern.*, early access, Mar. 17, 2021, doi: 10.1109/TCYB.2021.3061152.

[32] P. Wang, B. Xue, J. Liang, and M. Zhang, "Multiobjective differential evolution for feature selection in classification," *IEEE Trans. Cybern.*, early access, Dec. 7, 2021, doi: 10.1109/TCYB.2021.3128540.

[33] J. Li *et al.*, "Two-dimensional unsupervised feature selection via sparse feature filter," *IEEE Trans. Cybern.*, early access, Apr. 11, 2022, doi: 10.1109/TCYB.2022.3162908.

[34] C. Liu, C.-T. Zheng, S. Wu, Z. Yu, and H.-S. Wong, "Multitask feature selection by graph-clustered feature sharing," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 74–86, Jan. 2020.

[35] M. Zeng, Y. Cai, Z. Cai, X. Liu, P. Hu, and J. Ku, "Unsupervised hyperspectral image band selection based on deep subspace clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1889–1893, Dec. 2019.

[36] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial–spectral subspace clustering for hyperspectral image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2686–2697, Jul. 2021.

[37] Y. Cai, M. Zeng, Z. Cai, X. Liu, and Z. Zhang, "Graph regularized residual subspace clustering network for hyperspectral image clustering," *Inf. Sci.*, vol. 578, pp. 85–101, Nov. 2021.

[38] W. Jing, W. Lijiao, C. Jiantao, and L. Xiaorun, "Band selection based on signal-to-noise ratio estimation and hyperspectral anomaly detection," *Remote Sens. Technol. Appl.*, vol. 30, no. 2, pp. 292–297, 2015.

[39] P. Gao, J. Wang, H. Zhang, and Z. Li, "Boltzmann entropy-based unsupervised band selection for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 462–466, Mar. 2019.

[40] J. Feng, L. C. Jiao, X. Zhang, and T. Sun, "Hyperspectral band selection based on trivariate mutual information and clonal selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4092–4105, Jul. 2014.

[41] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogrammetric Eng. Remote Sens.*, vol. 70, no. 7, pp. 793–802, 2004.

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

[43] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 522–526, Oct. 2006.

[44] C. Yarra, S. Nagesh, O. D. Deshmukh, and P. K. Ghosh, "Noise robust speech rate estimation using signal-to-noise ratio dependent sub-band selection and peak detection strategy," *J. Acoust. Soc. America*, vol. 146, no. 3, pp. 1615–1628, 2019.

[45] H.-C. Li, C.-I. Chang, L. Wang, and Y. Li, "Constrained multiple band selection for hyperspectral imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2016, pp. 6149–6152.

[46] F. Xie, F. Li, C. Lei, J. Yang, and Y. Zhang, "Unsupervised band selection based on artificial bee colony algorithm for hyperspectral image classification," *Appl. Soft Comput.*, vol. 75, pp. 428–440, Feb. 2019.

[47] Y. Wan, A. Ma, Y. Zhong, X. Hu, and L. Zhang, "Multiobjective hyperspectral feature selection based on discrete sine cosine algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3601–3618, May 2020.

[48] Z. Yong, H. Chun-lin, S. Xian-fang, and S. Xiao-yan, "A multi-strategy integrated multi-objective artificial bee colony for unsupervised band selection of hyperspectral images," *Swarm Evol. Comput.*, vol. 60, Feb. 2021, Art. no. 100806.

[49] F. Zhang, Q. Wang, and X. Li, "Hyperspectral image band selection via global optimal clustering," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2017, pp. 1–4.

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

[51] H. Su, Q. Du, G. Chen, and P. Du, "Optimized hyperspectral band selection using particle swarm optimization," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2659–2670, Jun. 2014.

[52] C. He, Y. Zhang, D. Gong, X. Song, and X. Sun, "A multi-task bee colony band selection algorithm with variable-size clustering for hyperspectral images," *IEEE Trans. Evol. Comput.*, early access, Mar. 14, 2022, doi: 10.1109/TEVC.2022.3159253.

[53] K. Chen, B. Xue, M. Zhang, and F. Zhou, "Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimisation," *IEEE Trans. Evol. Comput.*, vol. 26, no. 3, pp. 446–460, Jun. 2022.

[54] H. Zhai, H. Zhang, L. Zhang, P. Li, and X. Xu, "Spectral-spatial clustering of hyperspectral remote sensing image with sparse subspace clustering model," in *Proc. 7th Workshop Hyperspectral Image Signal Process. Evol. Remote Sens. (WHISPERS)*, 2015, pp. 1–4.

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

[56] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2019.

[57] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.

[58] T. Zhou, C. Zhang, X. Peng, H. Bhaskar, and J. Yang, "Dual shared-specific multiview subspace clustering," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3517–3530, Aug. 2020.

[59] Z. Kang *et al.*, "Multi-graph fusion for multi-view spectral clustering," *Knowl. Based Syst.*, vol. 189, Feb. 2020, Art. no. 105102.

[60] C. Tang *et al.*, "CGD: Multi-view clustering via cross-view graph diffusion," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5924–5931.

[61] X. Li, H. Zhang, R. Wang, and F. Nie, "Multi-view clustering: A scalable and parameter-free bipartite graph fusion method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 330–344, Jan. 2022.

[62] E. G. Birgin and J. M. Martinez, "Improving ultimate convergence of an augmented lagrangian method," *Optim. Methods Softw.*, vol. 23, no. 2, pp. 177–195, 2008.

[63] C. Lu, J. Feng, S. Yan, and Z. Lin, "A unified alternating direction method of multipliers by majorization minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 527–541, Mar. 2018.

[64] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 88–102, Jan. 2016.

[65] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018.

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

[67] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5028–5039, Jun. 2021.

[68] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial–spectral kernel sparse representation for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2462–2471, Dec. 2013.

[69] B. Xu, X. Li, W. Hou, Y. Wang, and Y. Wei, "A similarity-based ranking method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9585–9599, Nov. 2021.

Jun Wang received the B.E. degree in information management and information system from Hubei University of Technology, Wuhan, China, in 2019. He is currently pursuing the master's degree with the School of Computer Science, China University of Geosciences, Wuhan, China.

His research interests are hyperspectral image processing and machine learning.

Chang Tang (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2016.

He was with the AMRL Laboratory, University of Wollongong, Wollongong, NSW, Australia, from September 2014 to September 2015. He is currently a Full Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. He has published 50+ peer-reviewed papers, including those in highly regarded journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, ICCV, CVPR, IJCAI, AAAI, and ACM MM. His current research interests include machine learning and computer vision.

Dr. Tang serves as an Associate Editor for *BioMed Research International* and *BMC Bioinformatics* and the Young Editor for *CAAI Transactions on Intelligence Technology* and *Computer Engineering*. He regularly serves on the Technical Program Committees or as the Area Chair of some top conferences, such as NIPS, ICML, CVPR, ICCV, ECCV, IJCAI, ICME, and AAAI.

Xinwang Liu (Senior Member, IEEE) received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2013.

He is currently a Full Professor with the School of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. He has published 60+ peer-reviewed papers, including highly regarded journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, NeurIPS, ICCV, CVPR, AAAI, and IJCAI. More information can be found at xinwangliu.github.io.

Xinzhong Zhu (Member, IEEE) received the Ph.D. degree from Xidian University, Xi'an, China, in 2018.

He is a Professor with the College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, China. His research interests include machine learning, computer vision, manufacturing informatization, robotics and system integration, and intelligent manufacturing. He is currently focusing on kernel learning and feature selection, multiview clustering algorithms, real-time object detection (e.g., pedestrian detection, vehicle detection, and general object detection) and deep learning, and their applications.

Prof. Zhu is a member of ACM.
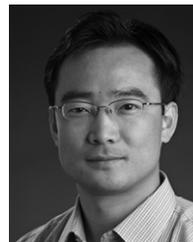
Lizhe Wang (Fellow, IEEE) received the B.E. and M.E. degrees from Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the Dr.Eng. degree (*magna cum laude*) from the University of Karlsruhe, Karlsruhe, Germany, in 2008.

He is currently a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences, Beijing, and also a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing, eScience, and remote sensing image processing.

Dr. Wang serves as an Associate Editor for the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, and the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING. He is a Fellow of IET and the British Computer Society.

Wei Zhang (Member, IEEE) received the B.E. degree from Zhejiang University, Hangzhou, China, in 2004, the M.S. degree from Liaoning University, Shenyang, China, in 2008, and the Ph.D. degree from the Shandong University of Science and Technology, Qingdao, China, in 2018.

He is currently an Associate Professor with the Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. His research interests include future generation network architectures, edge computing, and edge intelligence.

Albert Y. Zomaya (Fellow, IEEE) received the B.S. degree in electrical engineering from Kuwait University, Kuwait City, Kuwait, in 1987, and the Ph.D. degree in control engineering from Sheffield University, Sheffield, U.K., in 1990.

He is currently the Chair Professor of High Performance Computing and Networking with the School of Computer Science, University of Sydney, Sydney, NSW, Australia, where he is also the Director of the Centre for Distributed and High Performance Computing which was established in late 2009. He was an Australian Research Council Professorial Fellow from 2010 to 2014 and held the CISCO Systems Chair Professor of Internetworking from 2002 to 2007 and also was the Head of the School of Computer Science, University of Sydney from 2006 to 2007. Prior to his current appointment, he was a Full Professor with the Electrical and Electronic Engineering Department, The University of Western Australia, Perth, WA, Australia, where he also led the Parallel Computing Research Laboratory from 1990 to 2002. He served as an Associate, Deputy, and Acting Head of the Electrical and Electronic Engineering Department, The University of Western Australia, and held numerous visiting positions and has extensive industry involvement. He published more than 550 scientific papers and articles and is the author, coauthor, or editor of more than 20 books.

Prof. Zomaya served as the Editor-in-Chief for the IEEE TRANSACTIONS ON COMPUTERS from 2011 to 2014. He currently serves as a Founding Editor-in-Chief for the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING, the Founding Co-Editor-in-Chief for the *IET Cyber-Physical Systems*, and an Associate Editor-in-Chief (Special Issues) for *Journal of Parallel and Distributed Computing*. He also serves as an Associate Editor for 22 leading journals, such as the *ACM Computing Surveys*, *ACM Transactions on Internet Technology*, and IEEE TRANSACTIONS ON CLOUD COMPUTING. He is the Founding Editor of several book series, such as the *Wiley Book Series on Parallel and Distributed Computing*, *Scalable Computing and Communications* (Springer), and the *IET Book Series on Big Data*.

Wanqing Li (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from The University of Western Australia, Perth, WA, Australia, in 1997.

He was an Associate Professor with Zhejiang University from 1991 to 1992, a Senior Researcher and later a Principal Researcher with Motorola Research Lab, Sydney, NSW, Australia, from 1998 to 2003, and a Visiting Researcher with Microsoft Research, Redmond, WA, USA, in 2008, 2010, and 2013. He is currently an Associate Professor and the Director of Advanced Multimedia Research Lab, University of Wollongong, Wollongong, NSW, Australia. His research areas include machine learning, 3-D computer vision, 3-D multimedia signal processing, and medical image analysis.

Dr. Li serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and *Journal of Visual Communication and Image Representation*.