Subspace Clustering for Hyperspectral Images via Dictionary Learning with Adaptive Regularization

Shaoguang Huang, *Member, IEEE,* Hongyan Zhang, *Senior Member, IEEE,* and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*—Sparse subspace clustering (SSC) has emerged as an effective approach for the automatic analysis of hyperspectral images (HSI). Traditional SSC-based approaches employ the input HSI data as a dictionary of atoms, in terms of which all the data samples are linearly represented. This leads to highly redundant dictionaries of huge size and the computational complexity of the resulting optimization problems becomes prohibitive for large-scale data. In this paper, we propose a scalable subspace clustering method, which integrates the learning of a concise dictionary and robust subspace representation in a unified model. This reduces significantly the size of the involved optimization problems. We introduce a new adaptive spatial regularization for the representation coefficients, which incorporates spatial information of HSI and improves the robustness of the model to noise. We derive an effective solver based on alternating minimization and alternating direction method of multipliers (ADMM) to solve the resulting optimization problem. Experimental results on four representative hyperspectral images show the effectiveness of the proposed method and excellent clustering performance relative to the state-of-the-art.**

***Index Terms*—Hyperspectral images, clustering, subspace representation.**

I. Introduction

HYPERSPECTRAL images (HSIs) generated from airborne sensors or from satellites measure the objects on the Earth's surface with hundreds of spectral bands. With their rich spectral information, HSIs offer far better discrimination between different materials than conventional panchromatic and multispectral images, facilitating a wide range of applications including precision agriculture [1–3], environmental monitoring [4], defense and security [5], food safety [6] and mineralogy [7]. As a fundamental technique in these applications, clustering aims to group pixels into different clusters according to the inherent similarity of data points in an unsupervised way. In contrast to supervised classification, clustering requires no labelled training samples, which allows a wider application in practice.

In general, existing clustering methods can be roughly categorized into five groups: centroid-based clustering methods, density-based clustering methods, biological clustering methods, spectral-based clustering methods and deep learning based clustering methods. The centroid-based clustering methods such as k-means [8] and fuzzy c-means (FCM) [9] group data points by minimizing their distances to the iteratively updated cluster centroids. The density-based clustering methods, including [10, 11], identify the clusters of data by locating regions of high density that are separated from one another by regions of low density. Biological clustering methods obtain clustering results by mimicking biological systems [12]. Spectral-based clustering methods unveil cluster structure of data by making use of the spectrum (eigenvalues) of similarity matrix of the data to perform dimensionality reduction before clustering in k-means [13–16]. Deep learning based clustering methods [17–20] often consist of two steps: deep features extraction and clustering by applying the learned features in conventional clustering algorithms. The first step learns effectively non-linear and discriminative features of HSI by neural networks such as autoencoder, which leads to a better clustering performance. In pariticular, spectral-based clustering methods have been widely applied in various applications due to their excellent performance [21]. Graph construction in such clustering methods plays an essential role in the final clustering accuracy. The commonly used graph, built by k nearest neighbours (KNN) approach with Euclidean distance in the origianl data space, is sensitive to noise and often fails to capture the intrinsic data structure especially for the data points distributed near the intersection of two subspaces [22].

Sparse subspace clustering (SSC) method [23] builds a sparse graph by solving a sparse representation related problem in a self-representation model (i.e, the input data is employed as a dictionary), and achieves the state-of-the-art clustering performance. In general, SSC approach groups data points into different clusters in two steps: similarity matrix construction and spectral clustering. Basically, it models a high-dimensional data space as a union of low-dimensional subspaces, and estimates the similarity matrix with sparse coefficients of input data that are optimized in a subspace representation model. The key insight is that each data point in the subspace $\mathcal{S}_i$ can be represented by a linear combination of a few others from the same subspace $\mathcal{S}_i$. Thus, SSC starts from a self-representation model, i.e., $\mathbf{Y} = \mathbf{Y}\mathbf{C}$, and infers sparse coefficients $\mathbf{C}$ of the input data $\mathbf{Y}$ by solving a sparse coding problem ($\mathbf{C} \neq \mathbf{I}$). In particular, the non-zero entry C_{ij}

This work was supported in part by the Flanders AI Research Programme under Grant 174B09119, in part by the Bijzonder Onderzoeksfonds (BOF) under Grant BOF.24Y.2021.0049.01 and in part by the National Natural Science Foundation of China under Grant 61871298 and Grant 42071322. *(Corresponding author: Hongyan Zhang.)*

S. Huang and A. Pižurica are with the Department of Telecommunications and Information Processing, TELIN-GAIM, Ghent University, 9000 Ghent, Belgium (e-mail: Shaoguang.Huang@ugent.be; Aleksandra.Pizurica@ugent.be).

H. Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zhanghongyan@whu.edu.cn).

indicates explicitly that the data points $\mathbf{y}_i$ and $\mathbf{y}_j$ are belonging to the same subspace. This enables a direct and efficient construction of similarity matrix $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^T|$ where $|\mathbf{C}|$ takes the absolute values of $\mathbf{C}$, which is further applied in the spectral clustering framework to obtain clustering results.

However, SSC model solves separately the coefficients for each data point. In case of HSI, SSC groups pixels by using their spectral signatures alone, which means that the spatial information is not taken into account, resulting in sensitivity of performance to noise and within-class spectral variability [24, 25]. It has been widely proved that using spatial information together with spectral information can effectively improve the performance in various HSI processing tasks including supervised classification [26], denoising [27–29], change detection [30] and super-resolution [31]. Similarly, incorporating spatial information proves to be beneficial in HSI clustering as well [24, 32–37]. These methods take into account the spatial information either by directly introducing spatial regularization in the clustering models [24, 32, 33, 35–37] or by post-processing approaches such as Local Bilateral Filtering [34]. Due to the integration of spatial and spectral information, these methods undoubtedly achieve improved clustering accuracy. However, as they employ the whole input HSI data as the dictionary, which is typically huge and redundant in practice, the subspace representation is less efficient and less informative. Moreover, the resulting optimization problems are computationally expensive due to the high complexity in the order of $\mathcal{O}(q^3)$, where q is the total number of pixels in HSI. This prohibits their applications on large-scale HSI.

A clear way to mitigate this problem is to replace the original self-representation dictionary with a more compact, yet equally expressive dictionary. With a smaller dictionary, the number of sparse coefficients to be solved decreases as well, reducing thereby the overall computational complexity. However, the related research in the literature is rather limited. A recent sketched SSC model of [38], proposed for the clustering tasks in computer vision, lowers the computational complexity by using a sketched dictionary with a random projection technique. It was shown that applying such method in HSI clustering directly often yields poor performance [39]. To improve its clustering performance, a spatial regularization was introduced, achieving a lower computational complexity and improved clustering accuracy [39]. In [40], graph regularized sparse coding (GRSC) was introduced for general data representation, which takes into account of global similarities of data points by a graph Laplacian regularization in representation domain and yields improved performance. However, in case of HSI the important local spatial information is not considered in GRSC. Also its high computational complexity limits the applications on large-scale data. The authors in [41] presented a cascaded clustering model consisting of sparse dictionary learning and anchored subspace regression to calculate the sparse coefficients. The sparse dictionary is obtained by multiplying a fixed wavelet dictionary with a learned sparse matrix. The underlying fixed wavelet dictionary poses some limitations in terms of adapting to the actual data structure. Moreover, the spatial information is not exploited in the dictionary learning. Also, the cascaded approach to calculate coefficients solves multiple sparse representation related optimization problems, which results in a higher computational burden.

In this paper, we propose a novel dictionary learning based subspace clustering method for HSI with an adaptive joint total variation spatial regularization. The contributions of this paper can be summarized from three aspects. First, different from the traditional SSC-based clustering methods, which use the whole redundant input HSI data to construct the dictionary, we utilize a compact dictionary that is adaptively learned from the input data. The small dictionary reduces the number of sparse coefficients to be solved, which significantly lowers the overall computational complexity. Second, we take into account the spatial information by incorporating a novel adaptive joint total variation constraint in the subspace clustering model. The joint total variation (JTV) is formulated by adopting an $\ell_{1,2}$ norm penalty on the difference matrix of coefficients, which encodes effectively the dependencies of spatially neighbouring pixels in the low-dimensional subspace and promotes the coefficient vectors of most neighbouring pixels to be similar. The weights for the difference matrix in the JTV are updated iteratively in our optimization algorithm, which enables our model to treat pixels in homogeneous regions and edges differently. Third, we develop an efficient optimization algorithm for the resulting optimization problem using alternating minimization and alternating direction method of multipliers (ADMM). Extensive experiments are conducted and the results demonstrate the superior performance of the proposed method in terms of both quantitative and visual evaluations.

The rest of this paper is organized as follows. Section II briefly introduces the clustering of HSIs with the SSC model. Section III describes the proposed model and the resulting optimization problem. Section IV presents the experimental results on the real hyperspectral data sets and the comparisons with other methods. Section V concludes the paper.

II. HSI CLUSTERING WITH THE SSC MODEL

We denote by $\mathbf{Y} \in \mathbb{R}^{B \times MN}$ the flattened 2-D matrix from the original 3-D HSI data cube with a size of $M \times N \times B$, where M and N represent the height and the width of the HSI, respectively, and B denotes the number of bands. Each vector $\mathbf{y}_i \in \mathbb{R}^B$ represents the spectral signature of each pixel in HSI. We assume that there are t classes in the data. SSC partitions the high-dimensional data space into a union of lower dimensional subspaces. The pixels belonging to one class constitute a subspace. The key idea is that among infinitely many possibilities to represent a data point $\mathbf{y}_i$ in terms of other points, a sparse representation will select a few points that belong to the same subspace as $\mathbf{y}_i$. This is known as the subspace preserving property [23]. Thus, SSC starts from a self-representation model where the input data matrix $\mathbf{Y}$ is employed as a dictionary: $\mathbf{Y} = \mathbf{YC}$ and infers the coefficient matrix $\mathbf{C} \in \mathbb{R}^{MN \times MN}$ by solving the following optimization problems:

$$\begin{aligned} &\arg\min_{\mathbf{C}} \ \|\mathbf{C}\|_1 + \frac{\beta}{2}\|\mathbf{Y} - \mathbf{YC}\|_F^2 \\ &s.t. \ \ \mathrm{diag}(\mathbf{C}) = \mathbf{0}, \ \ \mathbf{1}^T\mathbf{C} = \mathbf{1}^T, \end{aligned} \tag{1}$$

where $\|\mathbf{C}\|_1 = \sum_i \sum_j |C_{ij}|$; $\mathbf{1}$ is an all-one vector; diag$(\mathbf{C})$ is a vector with its i-th element being C_{ii}; $\mathbf{0}$ is an all-zero vector and β is a parameter, which controls the balance between the data fidelity and the sparsity of the coefficient matrix. The first constraint is introduced to avoid the trivial solution of representing a sample by itself and the second constraint ensures that each data point is an affine combination of other data points when the data lie in a union of affine subspaces.

The model in (1) can be solved by the ADMM algorithm [42]. The coefficients matrix $\mathbf{C}$ yields directly the correlation structure among the pixels, i.e. a non-zero entry C_{ij} indicates that the samples $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same class. This leads to the construction of similarity matrix $\mathbf{W} \in \mathbb{R}^{MN \times MN}$ by $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^T|$. Clustering results are obtained by employing the similarity matrix $\mathbf{W}$ within the spectral clustering [43]. Specifically, the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{MN \times MN}$ is first formed by

$$\mathbf{L} := \mathbf{D}_w - \mathbf{W} \tag{2}$$

where $\mathbf{D}_w \in \mathbb{R}^{MN \times MN}$ is a diagonal matrix with $D_{w_{ii}} = \sum_j W_{ij}$ [44]. Afterwards, the t eigenvectors $\{\mathbf{v}_k\}_{k=1}^t$ of $\mathbf{L}$ corresponding to the t smallest eigenvalues of $\mathbf{L}$ are calculated via singular-value decomposition (SVD). Finally, the clustering result is obtained by applying the matrix $\mathbf{V} = [\mathbf{v_1}, ..., \mathbf{v_t}] \in \mathbb{R}^{MN \times t}$ to the k-means clustering method.

III. Dictionary learning based subspace clustering method with an adaptive joint total variation regularization

In this section, a novel subspace clustering model is proposed for HSI. The proposed method employs a dictionary learning strategy to model the underlying data subspaces. Moreover, a novel adaptively weighted joint total variation regularization is integrated into our model to improve the homogeneity of a clustering map. Finally, we develop an efficient optimization algorithm for the resulting model based on alternating minimization and ADMM.

A. Dictionary learning based subspace clustering

Most of the current SSC methods rely on a self-representation model, where the whole HSI data is employed as the dictionary to model the low-dimensional subspaces of data. HSIs usually have a rather small number of classes, and the spectral signatures within one class show very high similarity, which suggests that a HSI contains tremendous redundant information [45]. Thus using the whole HSI input data to model the data subspaces results in a poor representation ability and high computational complexity. Moreover, SSC-based methods represent a data point by a linear combination of a few other data points. When the data points are noisy, each data point is represented by other noisy data. Therefore, the obtained similarity matrix $\mathbf{W}$ is deteriorated, leading to a degraded clustering performance in the subsequent spectral clustering. It is thus of interest to construct efficiently compact dictionaries to model the underlying low-dimensional subspaces of a HSI. It has been demonstrated that learning a dictionary from data instead of using a predefined one can effectively improve the performance of data analysis [46–52]. This motivates us to learn a compact dictionary to model the low-dimensional subspaces of HSIs in our subspace clustering method.

The objective function with respect to the dictionary $\mathbf{D} \in \mathbb{R}^{B \times n}$ and sparse matrix $\mathbf{A} \in \mathbb{R}^{n \times MN}$ can be formulated as follows:

$$\underset{\mathbf{D},\mathbf{A}}{\arg\min}\ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1, \quad s.t.\ \ \mathbf{D} \geq 0, \tag{3}$$

where λ is a penalty parameter to control the sparsity of $\mathbf{A}$. The constraint on dictionary, $\mathbf{D} \geq 0$, requires that the atoms are nonnegative in agreement with the positive spectral intensities. Sparse coefficients matrix $\mathbf{A}$ indicates the contribution of atoms to the input data in the subspace representation.

The optimization problem in (3) is efficiently solved by alternating minimization switching between sparse coding and dictionary learning steps. As similar data points often yield similar representation coefficients and dissimilar data points often yield different coefficients, we can view matrix $\mathbf{A}$ as extracted features of input data $\mathbf{Y}$. After the sparse matrix $\mathbf{A}$ is obtained, we particularly employ it for the construction of similarity matrix. A sparse graph often yields better performance than a fully-connected graph as sparse graphs have much less spurious connections between the points belonging to different classes [53]. Hence, we construct the similarity matrix by using a KNN graph. For each $\boldsymbol{a}_i$ being the i-th column of $\mathbf{A}$, we find the first k nearest neighbours in Euclidean distance, denoted as $N_k(\boldsymbol{a}_i)$. Then the similarity matrix $\mathbf{W}$ is calculated as

$$W_{ij} = \begin{cases} w_{ij} & \boldsymbol{a}_i \in N_k(\boldsymbol{a}_j) \text{ or } \boldsymbol{a}_j \in N_k(\boldsymbol{a}_i) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where w_{ij} is obtained by using a Gaussian kernel function with parameter σ:

$$w_{ij} = e^{\frac{-\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2}{2\sigma^2}}. \tag{5}$$

Finally, the obtained sparse similarity matrix $\mathbf{W}$ is fed into the spectral clustering method to produce clustering results. We refer to our initialized method as the dictionary learning based subspace clustering method (DLSC).

B. Subspace clustering model with AJTV regularization

The optimization problem in (3) boils down to calculating the sparse coefficients vector of each data point separately and independently. This process is sensitive to noise and better results can be expected when incorporating prior knowledge about the spatial dependencies among the neighbouring pixels and their respective coefficient vectors.

Spatially adjacent pixels in a HSI are likely to belong to the same cluster characterized by a given type of spectral responses. Thus they will also be well represented by the same set of prototype responses (atoms) and in similar proportions.

In sparse coding formulation, this means that pixels within a local region are likely to be well represented as linear combinations of the same few atoms from a dictionary. Each row of $\mathbf{A}$, $\boldsymbol{a}^i \in \mathbb{R}^{1 \times MN}$, is the vector of responses of all the

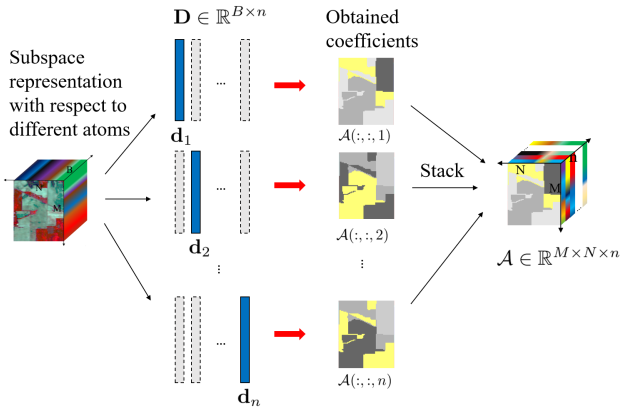

$\mathbf{D} \in \mathbb{R}^{B\times n}$
Obtained coefficients
Subspace representation with respect to different atoms
B
N
M
$\mathbf{d}_1$
$\mathbf{d}_2$
$\mathbf{d}_n$
$\mathcal{A}(:,:,1)$
$\mathcal{A}(:,:,2)$
$\mathcal{A}(:,:,n)$
Stack
N
M
n
$\mathcal{A} \in \mathbb{R}^{M\times N\times n}$

Fig. 1. A motivation of the introduction of JTV regularization. As similar data points have similar response to the atoms, coefficient matrices should be piece-wise smooth in spatial dimensions and have similar edges to the original HSI, which becomes apparent after reshaping them to a 3-D cube.

data points $\mathbf{Y}$ to an atom $\mathbf{d}_i$. For the purpose of visualization, let us reshape $\boldsymbol{a}^i$ into a 2-D matrix $\mathcal{A}(:,:,i) \in \mathbb{R}^{M\times N}$ as shown in Fig. 1. Its (m,n)-th entry shows the contribution of the atom $\mathbf{d}_i$ in representing the (m,n)-th pixel in the underlying HSI, i.e., $\mathcal{A}(:,:,i)$ is an activation map for $\mathbf{d}_i$. Observe that these activation maps are locally smooth with sharp transitions among neighbouring regions, we encode this property by introducing an adaptive joint total variation spatial regularization.

Let us first apply the anisotropic total variation (TV) norm [54] to each layer $\mathcal{A}(:,:,i)$:

$$\begin{aligned}\|\mathcal{A}(:,:,i)\|_{TV} = \sum_{m=1}^{M}\sum_{n=1}^{N} & |\mathcal{A}(m+1,n,i) - \mathcal{A}(m,n,i)| \\ & + |\mathcal{A}(m,n+1,i) - \mathcal{A}(m,n,i)| \quad (6)\end{aligned}$$

assuming periodic boundary conditions: $\mathcal{A}(M+1,n,i) = \mathcal{A}(1,n,i)$ $(1 \leq n \leq N)$ and $\mathcal{A}(m,N+1,i) = \mathcal{A}(m,1,i)$ $(1 \leq m \leq M)$. For the sake of more compact notation, let us rewrite the two terms in the expression above using finite difference operators that act on a reshaped image (in a raster scanning way) of size $MN \times 1$. In our notation,

$$\|\mathcal{A}(:,:,i)\|_{TV} = \|\boldsymbol{a}^{i^T}\|_{TV} = \|\mathbf{H}_x\boldsymbol{a}^{i^T}\|_1 + \|\mathbf{H}_y\boldsymbol{a}^{i^T}\|_1, \quad (7)$$

where $\mathbf{H}_x$ and $\mathbf{H}_y$ are the forward finite-difference operators in the horizontal and vertical directions, respectively, which correspond to the two terms from (6) when applied to the reshaped 1-D image data $\boldsymbol{a}^{i^T}$.

We apply the TV norm in (7) to each layer $\boldsymbol{a}^i$, and aggregate these to what we denote the TV norm of $\mathcal{A}$:

$$\begin{aligned}\|\mathcal{A}\|_{TV} &= \sum_{i=1}^{n}\|\mathcal{A}(:,:,i)\|_{TV} \\ &= \sum_{i=1}^{n}\|\boldsymbol{a}^{i^T}\|_{TV} \\ &= \sum_{i=1}^{n}\|\mathbf{H}_x\boldsymbol{a}^{i^T}\|_1 + \|\mathbf{H}_y\boldsymbol{a}^{i^T}\|_1 \\ &= \|\mathbf{H}_x\mathbf{A}^T\|_1 + \|\mathbf{H}_y\mathbf{A}^T\|_1. \quad (8)\end{aligned}$$

We model this way spatial dependencies in the activation maps $\boldsymbol{a}^i$ for each atom and aggregate them over all. What we still did not take into account is that the coefficients vectors $\boldsymbol{a}_i \in \mathbb{R}^n$ and $\boldsymbol{a}_j \in \mathbb{R}^n$ of two neighbouring pixels $\mathbf{y}_i$ and $\mathbf{y}_j$ should be similar as well. We require that the vector $\boldsymbol{a}_j$ comprising activations of all atoms $\mathbf{d}_k$, $k \in \{1,...,n\}$ at spatial location j is close (in the Euclidean sense) to $\boldsymbol{a}_i$ when i and j are spatially adjacent. We incorporate this constraint by replacing the ℓ_1 norm in (7) with the $\ell_{1,2}$ norm defined as $\|\mathbf{X}\|_{1,2} = \sum_i\sqrt{\sum_j X_{ij}^2}$. The resulting expression that we refer to as the joint TV (JTV) norm of $\mathcal{A}$ is:

$$\|\mathcal{A}\|_{JTV} = \|\mathbf{H}_x\mathbf{A}^T\|_{1,2} + \|\mathbf{H}_y\mathbf{A}^T\|_{1,2}. \quad (9)$$

We next explain how the JTV norm promotes neighbouring pixels to yield similar representations. Let $\{\mathbf{y}_j\}_{j\in\mathcal{N}_i}$ be the spatially adjacent pixels of $\mathbf{y}_i$, $\mathbf{y}_i = \mathbf{D}\boldsymbol{a}_i$ and $\{\mathbf{y}_j = \mathbf{D}\boldsymbol{a}_j\}_{j\in\mathcal{N}_i}$, where $\mathcal{N}_i$ is the index set comprising the spatial neighbours of $\mathbf{y}_i$ in horizontal and vertical directions. The spatial constraint (9) can be reformulated as

$$\begin{aligned}\|\mathcal{A}\|_{JTV} &= \sum_{i=1}^{MN}\sum_{j\in\mathcal{N}_i}\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 \\ &= \sum_{i=1}^{MN}\sum_{j\in\mathcal{N}_i} r_{ij} \\ &= \|\mathbf{R}\|_1, \quad (10)\end{aligned}$$

where $r_{ij} = \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2$ and is the (i,j)-th entry of $\mathbf{R}$. It is clear that the sparsity of $\mathbf{R}$ leads to most of r_{ij} to be zeros, which indicates that $\boldsymbol{a}_i$ and its spatial neighbours $\{\boldsymbol{a}_j\}_{j\in\mathcal{N}_i}$ are often the same. This facilitates the construction of a better similarity matrix in (4), improving thereby the performance in spectral clustering.

Note that the JTV norm in (9) treats all the pixels equally with the same weight 1. It would be more reasonable to promote the spatial continuity more in the homogeneous regions than in the edge areas. To allow this, we generalize the JTV term by introducing two diagonal matrices (for the horizontal and vertical directions), which contain weight coefficients for each pixel. We refer to the resulting expression as the adaptive joint total variation:

$$\|\mathcal{A}\|_{AJTV} = \|\mathbf{W}_x\mathbf{H}_x\mathbf{A}^T\|_{1,2} + \|\mathbf{W}_y\mathbf{H}_y\mathbf{A}^T\|_{1,2}, \quad (11)$$

$\mathbf{W}_x$ and $\mathbf{W}_y$ are two diagonal matrices where the diagonal elements are the weights corresponding to each pixel. We integrate this novel AJTV norm with our DLSC model in (3) and derive the final model, termed IDLSC, as follows.

$$\begin{aligned}\underset{\mathbf{D},\mathbf{A}}{\arg\min}\ & \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1 + \lambda_{tv}(\|\mathbf{W}_x\mathbf{H}_x\mathbf{A}^T\|_{1,2} \\ & + \|\mathbf{W}_y\mathbf{H}_y\mathbf{A}^T\|_{1,2}), \quad s.t.\ \ \mathbf{D} \geq 0, \quad (12)\end{aligned}$$

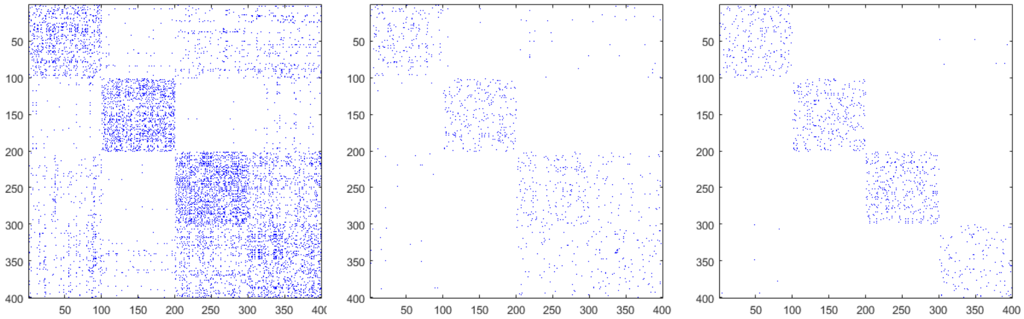


Fig. 2. Examples of similarity matrices in real data obtained by SSC (left), our DLSC in (3) (middle) and our final IDLSC clustering model in (15) (right). There are four clusters and 100 samples per cluster are randomly selected from *Indian Pines* image. The selected data is sequentially arranged by class.

where λ_{tv} is a penalty parameter for the AJTV regularization term. Here, we update $\mathbf{W}_x$ and $\mathbf{W}_y$ iteratively according to the gradient information of $\mathbf{A}$:

$$\mathbf{W}_{x_{ii}}^{(r+1)} = \frac{1}{1+ug_{x_i}^{(r)}} \tag{13}$$

$$\mathbf{W}_{y_{ii}}^{(r+1)} = \frac{1}{1+ug_{y_i}^{(r)}}, \qquad (i = 1, 2, ..., MN) \tag{14}$$

where r is the iteration number; $g_{x_i}^{(r)} = \|(\mathbf{H}_x \mathbf{A}^{(r)^T})^i\|_2$, $g_{y_i}^{(r)} = \|(\mathbf{H}_y \mathbf{A}^{(r)^T})^i\|_2$; u is a constant parameter with $u \geq 0$. Thereby, the value of weight is in the range of $[0, 1]$. When $r = 0$, we set $\mathbf{W}_x$ and $\mathbf{W}_y$ to identity matrices.

From (13) and (14) we can see that when the difference of sparse vectors between neighbouring pixels is small, i.e. the values of g_{x_i} and g_{y_i} are small, the corresponding weights $\mathbf{W}_{x_{ii}}$ and $\mathbf{W}_{y_{ii}}$ are relatively large, which increases the penalties on the i-th rows of $\mathbf{H}_x\mathbf{A}^T$ and $\mathbf{H}_y\mathbf{A}^T$ in next iteration. Compared with the coefficients matrix in model (3), it becomes more discriminative now in (12) due to the incorporation of the weighted JTV constraint. The benefit derived from such AJTV norm is consequently indicated in the structure of the resulting similarity matrix. In Fig. 2, we show the similarity matrices of SSC model, our DLSC model in (3) and our final IDLSC model in (12). There are 4 clusters and each has 100 pixels extracted from the common benchmark dataset *Indian Pines*. All the pixels are sequentially arranged according to their labels. In Fig. 2, we can see that the similarity matrix of the proposed IDLSC shows a more clear block-diagonal structure and much less incorrect connections compared to both SSC (left) and our original model (middle). It is known that the ideal similarity matrix should be block-diagonal, connecting only the data points from the same class [23]. The IDLSC model proves to perform better in this aspect than other methods.

Denote by $\mathbf{H} = [\mathbf{H}_x; \mathbf{H}_y]$ the combined TV operator and by $\mathbf{W}_h = \text{Diag}([\text{diag}(\mathbf{W}_x); \text{diag}(\mathbf{W}_y)])$ a new diagonal matrix. We obtain a simplified formulation of the IDLSC model from (12):

$$\arg\min_{\mathbf{D}\geq\mathbf{0},\mathbf{A}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1 + \lambda_{tv}\|\mathbf{W}_h\mathbf{H}\mathbf{A}^T\|_{1,2}, \tag{15}$$

where diag$(\cdot)$ represents a vector whose entries are the diagonal elements of a matrix and Diag$(\cdot)$ denotes a diagonal matrix with diagonal entries from a vector.

C. Optimization algorithm

In this section, we develop an efficient optimization method for our resulting model (15) using the alternating minimization. In general, it consists of two main updating steps: sparse coding and dictionary learning.

1) Sparse coding step: Firstly, we solve the following sparse coding problem with respect to $\mathbf{A}$ when dictionary $\mathbf{D}$ is fixed.

$$\arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1 + \lambda_{tv}\|\mathbf{W}_h\mathbf{H}\mathbf{A}^T\|_{1,2} \tag{16}$$

Directly solving this sparse coding problem is difficult and there is no closed-form solution in the literature. To solve the problem in (16), we introduce three auxiliary variables $\mathbf{B}, \mathbf{Z} \in \mathbb{R}^{n\times MN}$ and $\mathbf{V} \in \mathbb{R}^{2MN\times n}$ and reformulate model (16) equivalently to

$$\begin{aligned}&\arg\min_{\mathbf{A},\mathbf{B},\mathbf{Z},\mathbf{V}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2 + \lambda\|\mathbf{Z}\|_1 + \lambda_{tv}\|\mathbf{W}_h\mathbf{V}\|_{1,2}\\ &\qquad s.t. \quad \mathbf{A}=\mathbf{B}, \mathbf{A}=\mathbf{Z}, \mathbf{H}\mathbf{A}^T=\mathbf{V}\end{aligned} \tag{17}$$

Then we can solve the constrained problem (17) based on the ADMM algorithm [42]. The augmented Lagrangian function of (17) is derived as

$$\begin{aligned}&\frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2 + \lambda\|\mathbf{Z}\|_1 + \lambda_{tv}\|\mathbf{W}_h\mathbf{V}\|_{1,2} + \frac{\mu}{2}\|\mathbf{A}-\mathbf{B}+\frac{\mathbf{Y}_1}{\mu}\|_F^2\\ &+\frac{\mu}{2}\|\mathbf{A}-\mathbf{Z}+\frac{\mathbf{Y}_2}{\mu}\|_F^2 + \frac{\mu}{2}\|\mathbf{H}\mathbf{A}^T-\mathbf{V}+\frac{\mathbf{Y}_3}{\mu}\|_F^2,\end{aligned} \tag{18}$$

where $\mathbf{Y}_1, \mathbf{Y_2}$ and $\mathbf{Y}_3$ are multipliers and μ is a weighting parameter. ADMM calculates each of the variables $\{\mathbf{A}, \mathbf{B}, \mathbf{Z}, \mathbf{V}\}$ iteratively by solving one while fixing others, and the resulting sub-problems often can be solved easily.

The objective function with respect to $\mathbf{A}$ is given by

$$\begin{aligned}\mathbf{A}^{r+1} &= \arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{A}-\mathbf{B}^r+\frac{\mathbf{Y}_1^r}{\mu}\|_F^2\\ &+\frac{1}{2}\|\mathbf{A}-\mathbf{Z}^r+\frac{\mathbf{Y}_2^r}{\mu}\|_F^2 + \frac{1}{2}\|\mathbf{H}\mathbf{A}^T-\mathbf{V}^r+\frac{\mathbf{Y}_3^r}{\mu}\|_F^2\end{aligned} \tag{19}$$

By setting the first-order derivative to zero, we can obtain

$$\mathbf{A}(\mathbf{H}^T\mathbf{H}+2\mathbf{I}) = \mathbf{Z}^r + \mathbf{B}^r - \frac{\mathbf{Y}_1^r}{\mu} - \frac{\mathbf{Y}_2^r}{\mu} + (\mathbf{V}^{r^T} - \frac{\mathbf{Y}_3^{r^T}}{\mu})\mathbf{H}. \tag{20}$$

Matrix $\mathbf{A}$ can be efficiently calculated by using the fast Fourier transform (FFT) method:

$$\mathbf{A}^{r+1} = \mathcal{F}^{-1}\left[\frac{\mathbf{G}}{2+(\mathcal{F}(\mathbf{H}_x))^2+(\mathcal{F}(\mathbf{H}_y))^2}\right] \tag{21}$$

where $\mathbf{G} = \mathcal{F}(\mathbf{Z}^r + \mathbf{B}^r - \mathbf{Y}_1^r/\mu - \mathbf{Y}_2^r/\mu + (\mathbf{V}^{r^T} - \mathbf{Y}_3^{r^T}/\mu)\mathbf{H})$, and $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the FFT and the inverse FFT, respectively.

By fixing $\mathbf{A}, \mathbf{Z}$ and $\mathbf{V}$, we can obtain $\mathbf{B}$ by solving the following sub-problem:

$$\mathbf{B}^{r+1} = \arg\min_{\mathbf{B}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2 + \frac{\mu}{2}\|\mathbf{A}^{r+1} - \mathbf{B}^r + \frac{\mathbf{Y}_1^r}{\mu}\|_F^2. \tag{22}$$

The solution can be obtained by setting the first-order derivative to zero:

$$\mathbf{B}^{r+1} = (\mathbf{D}^T\mathbf{D} + \mu\mathbf{I})^{-1}(\mathbf{D}^T\mathbf{Y} + \mu\mathbf{A}^{r+1} + \mathbf{Y}_1^r). \tag{23}$$

Next, matrix $\mathbf{Z}$ is updated by solving following sub-problem:

$$\mathbf{Z}^{r+1} = \arg\min_{\mathbf{Z}} \lambda\|\mathbf{Z}\|_1 + \frac{\mu}{2}\|\mathbf{A}^{r+1} - \mathbf{Z} + \frac{\mathbf{Y}_2^r}{\mu}\|_F^2. \tag{24}$$

By introducing the following soft-thresholding operator:

$$\mathcal{R}_{\triangle}(x) = \begin{cases} sgn(x)(|x|-\triangle) & |x| \geq \triangle \\ 0 & otherwise \end{cases} \tag{25}$$

the problem in (24) can be solved by [55–58]

$$\mathbf{Z}^{r+1} = \mathcal{R}_{\frac{\lambda}{\mu}}(\mathbf{A}^{r+1} + \frac{\mathbf{Y}_2^r}{\mu}). \tag{26}$$

The objective function with respect to $\mathbf{V}$ is given by

$$\mathbf{V}^{r+1} = \arg\min_{\mathbf{V}} \lambda_{tv}\|\mathbf{W}_h\mathbf{V}\|_{1,2} + \frac{\mu}{2}\|\mathbf{H}\mathbf{A}^{(r+1)^T} - \mathbf{V} + \frac{\mathbf{Y}_3^r}{\mu}\|_F^2, \tag{27}$$

Denote by $\mathbf{v}_i$ the i-th row of $\mathbf{V}$ and $\mathbf{u}_i$ the i-th row of $\mathbf{H}\mathbf{A}^{(r+1)^T} + \frac{\mathbf{Y}_3^r}{\mu}$, the problem (27) can be solved in a row-wise manner as follows.

$$\mathbf{v}_i^{r+1} = \arg\min_{\mathbf{v}_i} w_{ii}\lambda_{tv}\|\mathbf{v}_i\|_2 + \frac{\mu}{2}\|\mathbf{u}_i - \mathbf{v}_i\|_2^2. \tag{28}$$

Then $\mathbf{v}_i^{r+1}$ can be updated by

$$\mathbf{v}_i^{r+1} = (1 - w_{ii}\lambda_{tv}/\mu/\|\mathbf{u}_i\|_2)_+\mathbf{u}_i, \tag{29}$$

where $(x)_+$ is an operator defined as $(x)_+ = max(x,0)$. Then we update the weighting matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ by (13) and (14), and the multipliers $\mathbf{Y}_1, \mathbf{Y}_2$ and $\mathbf{Y}_3$ by

$$\begin{aligned} \mathbf{Y}_1^{r+1} &= \mathbf{Y}_1^r + \mu(\mathbf{A}^{r+1} - \mathbf{B}^{r+1}) \\ \mathbf{Y}_2^{r+1} &= \mathbf{Y}_2^r + \mu(\mathbf{A}^{r+1} - \mathbf{Z}^{r+1}) \\ \mathbf{Y}_3^{r+1} &= \mathbf{Y}_3^r + \mu(\mathbf{H}\mathbf{A}^{(r+1)^T} - \mathbf{V}^{r+1}). \end{aligned} \tag{30}$$

These steps are updated iteratively until stop criterion is satisfied.

Algorithm 1 The proposed IDLSC method

```
Input: A HSI data $\mathbf{Y}$, $\lambda, \lambda_{tv}, t, n, k, u$;
Initialize $\mathbf{D}, \mathbf{B}, \mathbf{Z}, \mathbf{V}$ and $\mathbf{W}_h$;
while not converged do
    Sparse coding:
    while not converged do
        Update $\mathbf{A}$ by (21)
        Update $\mathbf{B}$ by (23)
        Update $\mathbf{Z}$ by (26)
        Update $\mathbf{V}$ by (29)
        Update $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ by (30)
        Update $\mathbf{W}_h$ by (13) and (14)
    end while
    Dictionary update:
    while not converged do
        Update $\mathbf{D}$ by (35)
        Update $\mathbf{S}$ by (36)
        Update $\mathbf{Y}_4$ by (34)
    end while
end while
Construct similarity matrix $\mathbf{W}$ by (4).
Apply $\mathbf{W}$ into spectral clustering.
Output: A clustering result of the HSI $\mathbf{Y}$.
```

2) Dictionary learning step: When $\mathbf{A}$ is fixed, the objective function respect to dictionary $\mathbf{D}$ is given by

$$\mathbf{D} = \arg\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2, \quad s.t. \;\; \mathbf{D} \geq 0. \tag{31}$$

We first introduce an auxiliary matrix $\mathbf{S}$ and let $\mathbf{D} = \mathbf{S}$, then we obtain an equivalent problem of (31):

$$\arg\min_{\mathbf{D},\mathbf{S}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2, \quad s.t. \;\; \mathbf{S} \geq 0, \mathbf{D} = \mathbf{S}, \tag{32}$$

The augmented Lagrangian function of (32) is derived as follows

$$\arg\min_{\mathbf{D},\mathbf{S}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 + \frac{\mu_1}{2}\|\mathbf{D}-\mathbf{S}+\frac{\mathbf{Y}_4}{\mu_1}\|_F^2 + \iota_+(\mathbf{S}), \tag{33}$$

where $\iota_+(S) = \sum_{i=1}^B \sum_{j=1}^n \iota_+(\mathbf{S}_{ij})$ is the indicator function and $\iota_+(s)$ is zero if s belongs to the nonnegative orthant and $+\infty$ otherwise.

We obtain the optimal solution of (33) based on ADMM algorithm by iteratively updating $\mathbf{D}, \mathbf{S}$ and $\mathbf{Y}_4$ as follows.

$$\begin{cases} \mathbf{D}^{r+1} & = \arg\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 + \frac{\mu_1}{2}\|\mathbf{D}-\mathbf{S}^r+\frac{\mathbf{Y}_4^r}{\mu_1}\|_F^2 \\ \mathbf{S}^{r+1} & = \arg\min_{\mathbf{S}} \frac{\mu_1}{2}\|\mathbf{D}^{r+1}-\mathbf{S}+\frac{\mathbf{Y}_4^r}{\mu_1}\|_F^2 + \iota_+(\mathbf{S}) \\ \mathbf{Y}_4^{r+1} & = \mathbf{Y}_4^r + \mu_1(\mathbf{D}^{r+1} - \mathbf{S}^{r+1}). \end{cases} \tag{34}$$

$\mathbf{D}^{r+1}$ can be obtained by setting the first-order derivative to zero:

$$\mathbf{D}^{r+1} = (\mathbf{Y}\mathbf{A}^T + \mu_1\mathbf{S}^r - \mathbf{Y}_4^r)(\mathbf{A}\mathbf{A}^T + \mu_1\mathbf{I})^{-1}. \tag{35}$$

Matrix $\mathbf{S}$ can be updated by

$$\mathbf{S}^{r+1} = (\mathbf{D}^{r+1} + \frac{\mathbf{Y}_4^r}{\mu_1})_+. \tag{36}$$

The two steps of sparse coding and dictionary learning are updated iteratively until $|\mathbf{D}^{r+1} - \mathbf{D}^{r}|_\infty < \epsilon$. The complete clustering method is summarized in Algorithm 1.

Remark 1: Computing the inverse in (23) and (35) can be expensive when the involved matrices are big. The matrix size of $\bar{\mathbf{D}} = \mathbf{D}^T\mathbf{D} + \mu\mathbf{I}$ in (23) and $\bar{\mathbf{A}} = \mathbf{A}\mathbf{A}^T + \mu_1\mathbf{I}$ in (35) is $n \times n$. Note that $\bar{\mathbf{D}}$ and $\bar{\mathbf{A}}$ are symmetric and positive definite. Therefore, one can efficiently solve $\bar{\mathbf{D}}\mathbf{B} = \mathbf{F}$ by "*B=linsolve($\bar{D}$,F,opts)*" in MATLAB. When n is small, we can also efficiently solve $\bar{\mathbf{D}}\mathbf{B} = \mathbf{F}$ by "*B=$\bar{D}\backslash F$*" in MATLAB. In our implementation, we use "*B=$\bar{D}\backslash F$*" to solve the problem $\bar{\mathbf{D}}\mathbf{B} = \mathbf{F}$ as n is often small as indicated in Fig. 18. Similarly, we solve $\mathbf{D}\bar{\mathbf{A}} = F$ by "*D=F/$\bar{A}$*".

Next, we analyse the computational complexity of the proposed optimization algorithm. In each iteration of sparse coding, the time complexity is $\mathcal{O}(MNn\log(MN))$ for updating $\mathbf{A}$, $\mathcal{O}(MMn^2)$ for updating $\mathbf{B}$, $\mathcal{O}(2MNn)$ for updating $\mathbf{V}$ and $\mathcal{O}(2MNn)$ for updating $\mathbf{W}$. $\mathbf{Z}$ is updated by the thresholding operator in (26) whose time complexity is negligible. Since $\log(MN) < n$ in most cases, the sparse coding has a time complexity of $\mathbf{O}(I_1MNn^2)$, where I_1 is the number of iterations in sparse coding. The update of $\mathbf{D}$ in (35) is $\mathcal{O}(MNBn + MNn^2 + n^3)$. The complexity to update $\mathbf{S}$ by (36) is neglectable. As $n \ll MN$ and n is smaller than B in most cases, the time complexity for updating $\mathbf{D}$ is $\mathcal{O}(MNBn)$. Thus, the time complexity for dictionary update step in Algorithm 1 is $\mathcal{O}(I_2MNBn)$, where I_2 is the number of iterations in (34). Finally, we obtain the overall time complexity, $\mathcal{O}(\max(I_1MNn^2, I_2MNBn))$, of Algorithm 1 in each iteration of the outer loop.

IV. Experiments

To evaluate the performance of the proposed method, we compare it with two classical clustering methods FCM [9] and k-means [8], the powerful density-based clustering method CFSFDP [11], the original SSC method [23], the state-of-the-art spatial-spectral clustering methods L2-SSC [35] and JSSC [24] and four recently proposed Sketch-SSC [38], SS-SDAR methods [41], Hx-NMF [59] and DS3C [18]. Hx-NMF integrates graph learning and subspace clustering in a unified non-negative matrix factorization (NMF) framework, which does not rely on external clustering algorithms. DS3C is an end-to-end deep subspace clustering method in which multi-scale auto-encoder is designed to extract spatial-spectral features at different scales and self-representation layers are integrated with multi-scale auto-encoder to learn subspace representation in the deep feature domain. Compared to the SSC, L2-SSC and JSSC, which employ the whole input data as dictionary, Sketch-SSC, SS-SDAR and our method IDLSC utilize compact dictionaries in the subspace clustering models. In addition, the results of our initial DLSC model are also reported.

We evaluate all the methods on three well-known HSIs, including Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) *Indian Pines* image, Hyperspectral Digital Imagery Collection Experiment (HYDICE) *Urban* image and National Center for Airborne Laser Mapping (NCALM) *University of Houston* image. The clustering performance is measured by the metrics: overall accuracy (OA), averaged precision rate (APR) and the running time. Also visual clustering results are reported. To calculate the OA and APR from confusion matrix, we first find the best match between the clustering results and ground truth by an optimal mapping function obtained by the Kuhn-Munkres algorithm [60]. Note that the label information is used only for the purpose of evaluating clustering performance and not used in the clustering methods. All clustering methods are unsupervised methods. Let $n_{i,j}$ be the number of pixels in class i that are labeled as class j and p_i the precision rate for the i-th class $p_i = n_{i,i}/\sum_j n_{j,i}$ [61]. Then, APR is given by $\sum p_i/t$. We set the number of clusters, t, to the number of classes in the ground truth. All the methods except SS-SDAR (partly written with C code) and DS3C were implemented in MATLAB on a computer with an Intel© core-i7 3930K CPU with 64 GB of RAM. The DS3C method was implemented in TensorFlow and was run in Google Colab with a Tesla P100 GPU with 25 GB of RAM.

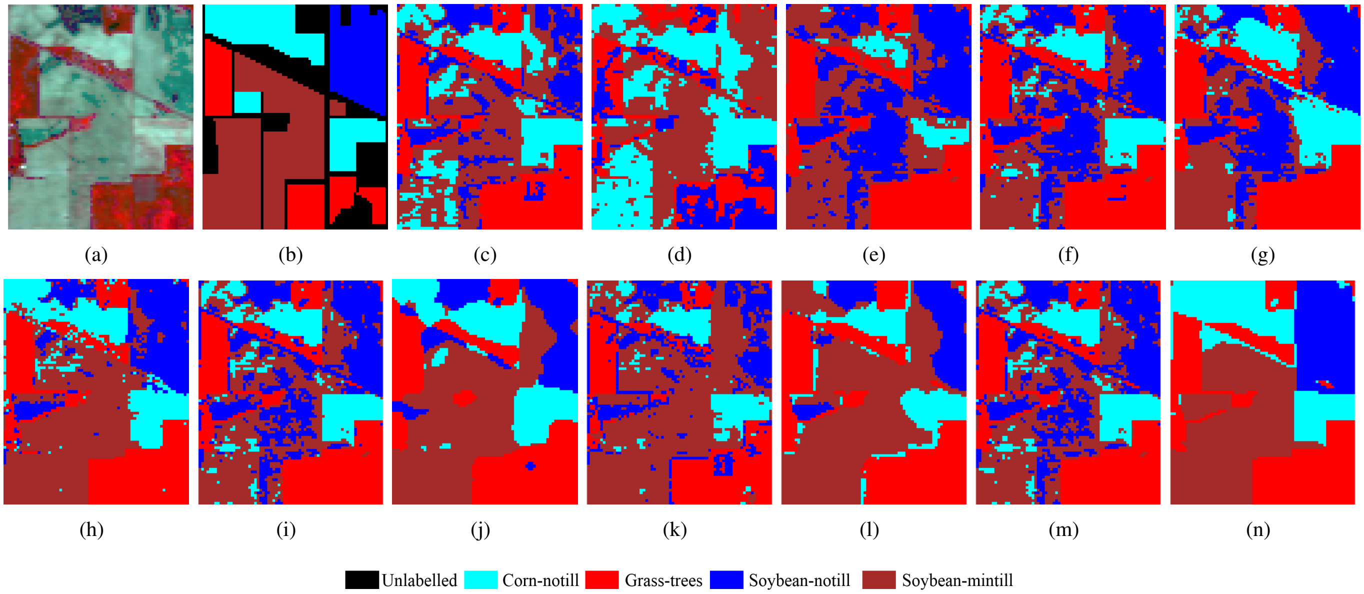
The parameters of compared methods are tuned to achieve their best performance in terms of OA. Specifically, we tune the parameter of SSC α_z in the range of $\{100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800\}$. In L2-SSC, the parameter α is set to 10^{-3} and β is tuned in the range of $\{100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800\}$. The number of super-pixels in JSSC is set in the range of $\{10, 20, 40, 80\}$ and λ is tuned in the range of $\{1, 10, 100, 1000\}$. In Sketch-SSC, we tune the parameter α in the range of $\{10, 100, 1000, 10000, 100000\}$. The parameter K in SS-SDAR is searched within the set $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ and λ_2 is set to 0.1 as suggested by [41]. In Hx-NMF, we vary parameters α and β in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, respectively. In DS3C, the architecture of auto-encoder and parameters α, β are the same as that in [18], and we set the number of epochs to 500.

A. AVIRIS Data Set: Indian Pines Image

This image was captured by the AVIRIS sensor over the Indian Pines region in North-western Indiana on June 12, 1992, with 20-m spatial resolution per pixel and 10-nm spectral resolution per band in the range between 100 nm and 2500 nm. *Indian Pines* image contains 220 bands and each band has a spatial size of 145×145. During the test, 20 spectral bands in 104-108, 150-163 and 200 are removed due to water absorption. This image consists of 16 classes. We select four classes from the image and the test image has a total number of 85×70 pixels [32]. The four classes in the test hyperspectral data are shown in Table I. The corresponding false-color composite image and ground truth are shown in Fig. 3 (a) and (b), respectively. It is known that supervised classification on this data is very difficult due to the spectral noise and the high similarity of data points belonging to different classes (e.g., class 1, 3 and 4), which can be seen in Fig. 4. Unsupervised clustering methods do not use any labelled training samples, so the accessed prior information is more limited than the supervised classification, which makes the clustering on this data even more challenging.

TABLE I
QUANTITATIVE EVALUATION OF DIFFERENT CLUSTERING METHODS ON PART OF *INDIAN PINES*

No.	Class name	FCM	k-means	CFSFDP	SSC	L2-SSC	JSSC	Sketch-SSC	SS-SDAR	Hx-NMF	DS3C[1]	DLSC	IDLSC
1	Corn-notill	62.39	69.85	28.46	60.00	61.09	74.03	62.19	69.25	50.65	48.36	57.61	**88.18**
2	Grass-trees	94.66	53.84	**100**	98.36	99.32	**100**	**100**	90.96	91.51	**100**	99.97	**100**
3	Soybean-notill	44.13	0	82.38	76.91	79.37	86.20	68.80	62.02	41.94	60.25	62.09	**100**
4	Soybean-mintill	63.83	57.59	50.73	50.68	54.89	87.79	58.87	91.37	88.10	90.90	68.91	**94.84**
	OA(%)	65.34	50.17	59.10	65.11	67.78	86.40	68.12	81.35	72.40	77.57	70.35	**95.04**
	APR(%)	65.8	44.61	73.01	72.40	74.63	86.51	72.64	82.04	77.19	79.97	74.05	**95.48**
	Time(in seconds)	6	**3**	9	543	624	270	**3**	29	78	874	30	164


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

(h) (i) (j) (k) (l) (m) (n)

■ Unlabelled ■ Corn-notill ■ Grass-trees ■ Soybean-notill ■ Soybean-mintill

Fig. 3. The part of *Indian Pines*. (a) False color image, (b) Ground truth, and Clustering maps of (c) FCM, (d) k-means, (e) CFSFDP, (f) SSC, (g) L2-SSC, (h) JSSC, (i) Sketch-SSC, (j) SS-SDAR, (k) Hx-NMF, (l) DS3C, (m) DLSC and (n) IDLSC.

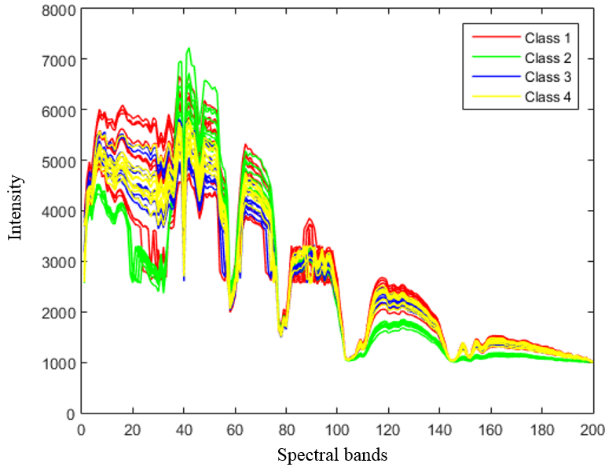


Fig. 4. Randomly selected spectral curves (ten per class) in *Indian Pines* image.

The clustering results of different clustering methods are reported in Table I with quantitative evaluations mentioned above. The clustering maps are shown in Fig. 3. The best result in Table I is marked in bold and the second best result is underlined. We set λ and λ_{tv} in our method as 5×10^{-6} and 5×10^{-2} for this data set, respectively. Other parameters are set as $n = 70, k = 30$ and $u = 0.5$ for the experiments in all the data sets. In general, Fig. 3 and Table I indicate that our IDLSC method outperforms other clustering methods in terms of overall accuracy. There are several observations to be made here.

Firstly, representation-based clustering methods SSC, L2-SSC, JSSC, Sketch-SSC, SS-SDAR, Hx-NMF, DLSC and IDLSC often achieve higher clustering accuracy than the classical clustering methods FCM, k-means and CFSFDP, which demonstrates the effectiveness of representation-based clustering for HSI. The superior performance mainly comes from the fact that the similarity matrix obtained in representation-based methods can capture well the nonlinear and low dimensional manifold structure of the input data, which is further exploited in the spectral clustering.

Secondly, the spatial-spectral clustering methods achieve higher accuracy than spectral-based methods. Specifically, L2-SSC, JSSC, SS-SDAR and IDLSC perform better than SSC in terms of clustering accuracy and especially our method IDLSC yields the best result with 29.83% improvement over SSC. The reason for the lower accuracy of SSC is that the sparse coding

[1] The whole data including unlabelled pixels is fed into DS3C model rather than only the labelled pixels as in [18], which leads to a different result from the published work [18].

in SSC is computed independently which ignores the spatial dependencies of pixels. In contrast, spatial-spectral clustering methods incorporate spatial information in the clustering models, resulting in improved clustering accuracy. The improved performance of IDLSC over DLSC also demonstrates the importance of the spatial information. It is observed that the deep subspace clustering model DS3C outperforms the spectral-based methods SSC, Sketch-SSC, Hx-NMF and DLSC. The improvement can be mainly attributed to the discriminative deep features extracted by encoder. Note that a query data point of DS3C is a data cube, including the central pixel and its neighbouring pixels in a square window, which means that DS3C also incorporates spatial information to a certain degree. However, the self-representation layer placed after encoder computes coefficients vectors independently, which neglects the spatial connection of data points, leading to inferior performance to other spatial-spectral methods JSSC, SS-SDAR and IDLSC. The clustering results in Fig. 3 show that the spatial-spectral methods often produce smoother clustering maps than spectral-based methods, indicating the superior ability to preserve the homogeneity of results. It is obvious that our result in Fig. 3 (n) exhibits the best consistency in local region and best agreement with the ground truth, which is clearly verified in Table I. For example, the accuracies for both class 2 and class 3 reach up to 100%, and the accuracies for class 1 and 4 are 88.18% and 94.84%, respectively, which are much higher than that in other state-of-the-art methods.

Thirdly, the representation-based methods usually take longer time than the classical methods FCM, k-means and CFSFDP, especially the traditional SSC-based methods SSC, L2-SSC and JSSC. Their long running time is due to the huge size of optimization problems, involving large self-representation dictionaries. Compared with SSC, L2-SSC and JSSC, other representation-based methods including Sketch-SSC, SS-SDAR, DLSC and IDLSC consume less running time because of the employed compact dictionaries. As the dictionary size is much smaller than the self-representation dictionary, the number of variables to be optimized is significantly reduced (up to hundreds times), decreasing thereby the overall complexity. This clearly shows the benefit of using compact dictionaries in subspace clustering. DS3C uses more running time than other methods, which is mainly caused by the employed two self-representation layers which contains huge amount of variables to be optimized.

Lastly, observe that our IDLSC shows a notable improvement over Sketch-SSC and SS-SDAR, both of which utilize compact dictionaries in the subspace clustering models. The clustering accuracy of Sketch-SSC on the *Indian Pines* data set was only about 68%, which is clearly insufficient for any practical use. This poor performance is mainly due to the noise and spectral variability. Similarly, Hx-NMF, which does not make use of spatial information, does not obtain satisfactory clustering performance here as well. The accuracy obtained by SS-SDAR although better (around 81%) is still unsatisfactory. This can be partly attributed to the weaker representation ability of a fixed wavelet dictionary. In contrast, our method yields an accuracy of above 95%. The superior performance confirms the advantages of using the adaptive dictionary and the novel AJTV spatial regularization.

B. HYDICE Data Set: Urban Image

The second data set that we use for evaluation is *HYDICE Urban* image, which was captured by the HYDICE sensor during a flight campaign over Copperas Cove, near Fort Hood, TX, USA. This data has a spatial size of 307×307 and contains 210 bands corresponding to the wavelengths ranging from 400 nm to 2500 nm. After removing the bands 1-4, 76, 87, 101-111, 136-153 and 198-210, which are seriously polluted by the atmosphere and water absorption, the remaining 162 bands are used in the experiments. For computational efficiency, a typical subset of data with a size $150\times160\times162$ is used as the test data, which includes seven land-cover objects as shown in Table II. The false-color image and ground truth are shown in Fig. 5 (a) and (b), respectively. Fig. 6 presents the randomly selected spectral curves in each class, where we can learn that this dataset has more complicated land-covers. In particular, class 3, 4 and 5 have large spectral variations within class and also share very high spectral similarity cross classes.

We report the clustering performance of all the methods except DS3C in Table II and Fig. 5, where quantitative and visual evaluations are presented, respectively. Due to out of memory, DS3C cannot be run in our available computing devices. In this data, we set $\lambda = 6\times10^{-3}$ and $\lambda_{tv} = 10^{-3}$ and other parameters the same to the first data set. Generally, from Table II and Fig. 5, we can learn the similar observations to that in the previous experiment. Our method IDLSC consistently achieve the best clustering performance in terms of OA and APR, which demonstrates the effectiveness and superiority of our approach. We observe that on some classes such as “parking lot”, “trees”, “sparse vegetation” and “concrete road” other methods yield better classification accuracy. However, the accuracies that our method yields on these classes are comparable to the corresponding best results. Classical methods k-means and CFSFDP obtain very poor results especially in the classes “tree” and “asphalt road”. In contrast, the remaining representation-based methods mostly yield much higher overall accuracies and cluster the classes “tree” and “asphalt road” very well, indicting the superior ability of representation-based methods to capture the complicated data structure. The accuracy of SSC is not acceptable as it only consider spectral information. The spatial-spectral clustering methods L2-SSC and JSSC incorporate spatial information to improve the accuracy to a certain degree. But their high computational complexities caused by the large self-representation dictionary dramatically increase the running time, which is less attractive in practice. Sketch-SSC method proposed for computer vision task uses a sketched dictionary by random projection to reduce the overall complexity, but the clustering accuracy in HSI is very poor due to the effect of noise and spectral variability. SS-SDAR obtains a comparatively higher accuracy of 83.01% with the sparse dictionary and post filtering technique, but the representation ability of dictionary is not optimal because of the fixed wavelets dictionary, which leaves room for a better performance. Compared with SS-SDAR, our method achieves a significant improvement in terms of OA, with the completely

TABLE II
QUANTITATIVE EVALUATION OF DIFFERENT CLUSTERING METHODS ON THE *HYDICE URBAN* ("OOM" MEANS "OUT OF MEMORY")

No.	Class name	FCM	k-means	CFSFDP	SSC	L2-SSC	JSSC	Sketch-SSC	SS-SDAR	Hx-NMF	DS3C	DLSC	IDLSC
1	Roof	88.22	90.20	94.74	96.24	96.48	89.48	97.72	87.64	88.95	OOM	**99.42**	99.41
2	Parking lot	93.52	96.63	98.98	93.82	95.75	93.33	88.15	**99.96**	95.49	—	67.88	92.75
3	Grass	48.49	43.56	41.71	73.55	75.17	72.62	70.46	83.64	63.46	—	68.45	**84.29**
4	Trees	1.24	37.68	9.01	**92.65**	90.27	92.13	90.41	74.02	**92.65**	—	83.85	84.43
5	Sparse vegetation	77.25	85.18	**99.26**	40.83	77.56	56.91	71.74	63.16	92.12	—	60.38	95.84
6	Asphalt road	69.03	2.71	0	71.26	46.58	72.13	48.93	71.37	13.99	—	90.62	**93.08**
7	Concrete road	74.39	70.18	52.55	53.10	**98.78**	97.89	46.83	96.90	94.46	—	93.24	90.22
	OA(%)	71.02	65.04	64.26	76.37	82.32	80.95	75.76	83.01	76.47	OOM	78.86	**92.37**
	APR(%)	65.64	59.53	67.28	77.05	83.20	81.18	75.22	82.95	78.54	—	79.32	**92.47**
	Time(in seconds)	39	**17**	158	31047	20111	10907	24	480	968	—	173	1100

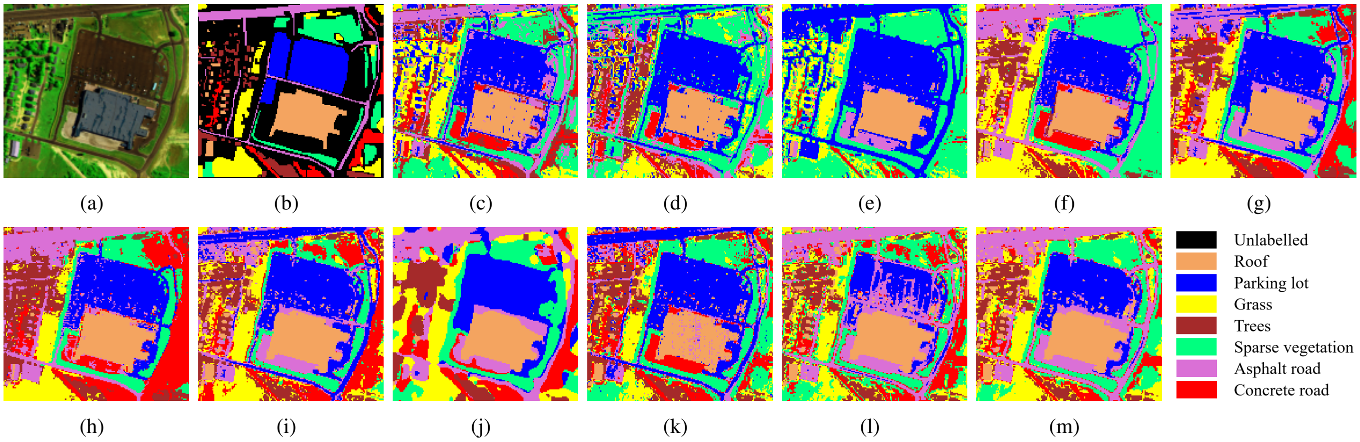


Fig. 5. *HYDICE Urban* image. (a) False color image, (b) Ground truth, and Clustering maps of (c) FCM, (d) k-means, (e) CFSFDP, (f) SSC, (g) L2-SSC, (h) JSSC, (i) Sketch-SSC, (j) SS-SDAR, (k) Hx-NMF, (l) DLSC and (m) IDLSC.

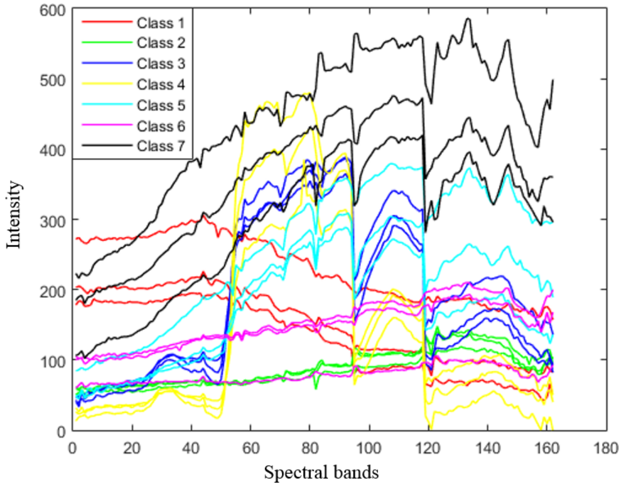


Fig. 6. Randomly selected spectral curves (three per class) in *HYDICE Urban* image.

learned dictionary and adaptive joint TV constraint, which collaboratively promote the discriminability and consistency in subspace representation. Hx-NMF yields an OA of 76.47%, which is comparable to SSC. Our method spends slightly more running time than Sketch-SSC and SS-SDAR (partly implemented by C code), but it is admissible in consideration of the improved accuracy. Notably, in comparison with the SSC, L2-SSC and JSSC, our method achieves a remarkable improvement both in accuracy and running time. Also our clustering map in Fig. 5 (m) presents the best visual result where the detailed and smooth regions are mostly consistent with the false-color image and ground truth.

C. NCALM Data Set: University of Houston

The third dataset we use for evaluation is *University of Houston* image, which was gathered by the NCALM sensor during a flight over the University of Houston campuse, Texas, USA, in June 2012. This dataset has been used in the 2013 IEEE GRSS Data Fusion Contest (DFC) [62]. The hyperspectral image has a spatial size of 349 × 1905, comprising 144 spectral bands from the 380 nm to 1050 nm. We are interested in a 130 × 130 × 144 subset of this image, captured over Robertson stadium on the Houston Campus and its surroundings. Fig. 7 (a) and (b) show the false-color image and corresponding ground truth. In total there are seven classes in the test dataset as shown in Table III. The spectral curves shown in Fig. 8 indicates the challenge of clustering on this data due to the large spectral variations within class and high similarity between several classes.

Table III and Fig. 7 present the clustering performance on this data with quantitative evaluations and visual clustering maps, respectively. The result of DS3C is not shown because of out of memory. We set $\lambda = 0.1$ and $\lambda_{tv} = 0.05$ and other parameters the same as that in previous experiments. The results in Table III show that our method continually outperforms other state-of-the-art clustering methods with the highest

TABLE III

QUANTITATIVE EVALUATION OF DIFFERENT CLUSTERING METHODS ON THE *NCALM UNIVERSITY OF HOUSTON* ("OOM" MEANS "OUT OF MEMORY")

No.	Class name	FCM	k-means	CFSFDP	SSC	L2-SSC	JSSC	Sketch-SSC	SS-SDAR	Hx-NMF	DS3C	DLSC	IDLSC
1	Concrete	80.40	80.77	46.50	40.54	46.50	46.50	48.66	46.87	87.11	OOM	53.28	**93.98**
2	Grass-1	99.88	99.54	**100**	99.88	99.30	**100**	**100**	**100**	98.61	—	90.85	99.91
3	Grass-2	0	0	0	36.20	58.06	70.43	0.07	48.57	70.07	—	76.52	**98.85**
4	Parking lot	63.31	62.91	99.70	97.60	99.10	98.55	95.41	**99.75**	89.72	—	95.16	98.20
5	Roof	88.46	92.31	53.08	95.38	**100**	**100**	**100**	95.38	99.23	—	**100**	**100**
6	Trees	52.30	69.98	84.26	**89.83**	83.05	82.57	1.4	62.95	24.70	—	0	0.68
7	Asphalt	82.52	85.03	0.13	51.32	30.19	68.68	29.66	59.62	1.13	—	64.78	**85.43**
	OA(%)	68.74	70.25	63.93	73.17	73.77	79.80	62.24	75.67	72.87	OOM	73.35	**89.35**
	APR(%)	55.85	58.54	60.60	79.59	73.21	82.58	58.05	77.31	66.31	—	71.79	85.75
	Time(in seconds)	21	**10**	66	11502	11859	3958	12	260	469	—	112	383

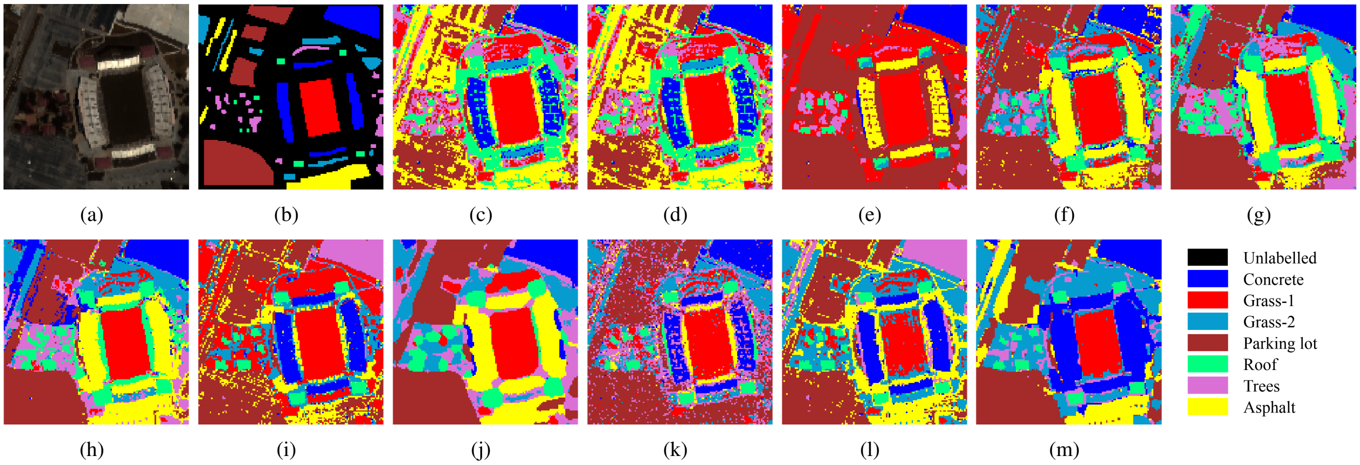


Fig. 7. *NCALM University of Houston* image. (a) False color image, (b) Ground truth, and Clustering maps of (c) FCM, (d) k-means, (e) CFSFDP, (f) SSC, (g) L2-SSC, (h) JSSC, (i) Sketch-SSC, (j) SS-SDAR, (k) Hx-NMF, (l) DLSC and (m) IDLSC.

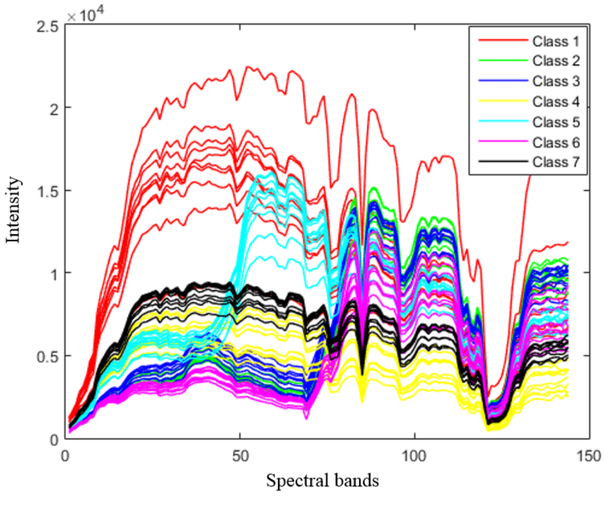
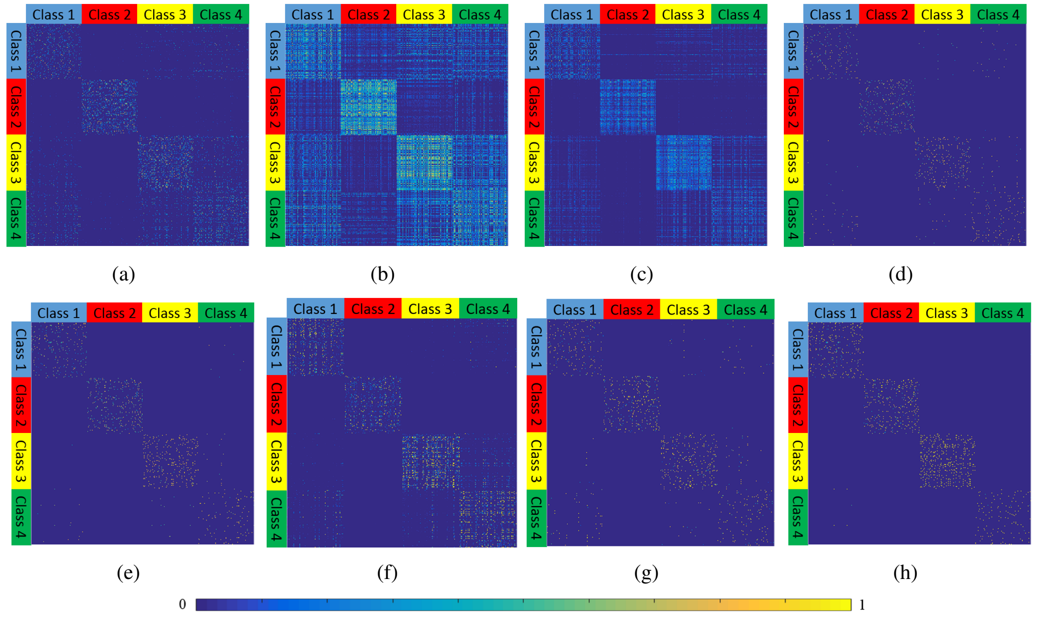


Fig. 8. Randomly selected spectral curves (ten per class) in *NCALM University of Houston* image.

overall accuracy. We also achieve the highest accuracies in the classes "concrete", "grass-2", "roof" and "asphalt", and comparable accuracies in the classes "grass-1" and "parking lot". The accuracy of "trees" is very low. The reason mainly attributes to the failure of IDLSC in discrimination between "grass-2" and "trees", which results in that most of "trees" are clustered into the group of "grass-2" as shown in Fig. 7 (m). A possible solution to alleviate this problem is to use a different set of parameters. But the risk is a degraded overall accuracy. It is noticed again in Table III that representation-based methods yield better performance than FCM, k-means and CFSFDP in terms of accuracy. We can also find that k-means algorithm is the most efficient clustering method, which takes the least running time in all the three datasets. Compared with the spectral-based SSC method, the spatial-spectral methods L2-SSC and JSSC obtain improved clustering accuracy. In comparison with SSC, L2-SSC and JSSC, Sketch-SSC, SS-SDAR, Hx-NMF, DLSC and IDLSC take much less running time due to the employed compact dictionary. Our IDLSC method yields a notable improvement over SSC, L2-SSC and JSSC in terms of accuracy and speed. Compared with SS-SDAR method, our method achieves a significant accuracy improvement with the comparable running time. The results in Fig. 7 reveal that the clustering map of our method in Fig. 7 (m) is more congruous with the ground-truth than others especially for the classes "concrete" and "asphalt". Overall, the results in this experiment verifies again the advantages of our approach.

D. Discussion and Analysis

1) The Analysis of Similarity Matrices: As most of the representation-based methods apply the same spectral clustering to obtain clustering results with their corresponding



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 9. Similarity matrices obtained by (a) SSC, (b) L2-SSC, (c) JSSC, (d) Sketch-SSC, (e) SS-SDAR, (f) DS3C, (g) DLSC and (h) IDLSC in the *Indian Pines* image.

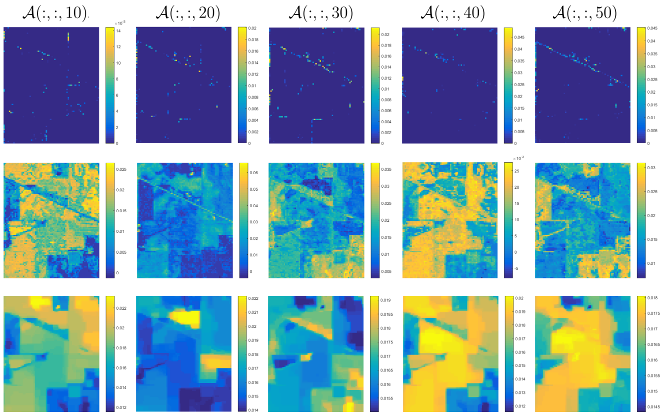


Fig. 10. Comparisons between the coefficients obtained by SSC (first row), DLSC (second row) and IDLSC (third row). Five layers of coefficients cube $\mathcal{A}$ are shown.

similarity matrices, it is very interesting to visualize and compare the resulting similarity matrices in different methods. For simplification, we assume all the data points are sequentially arranged by class. The similarity matrices obtained by different methods are shown in Fig. 9, which are corresponding to the randomly selected pixels (200 per class) in the *Indian Pines* image. Clearly, our similarity matrix in Fig. 9 (g) resembles more block-diagonal structure than others, which means it imposes less wrong connections between the data points belonging to different classes. In contrast, the similarity matrices of other methods have many incorrect links between different classes, e.g., classes (1,4) and (3,4). This observation is consistent with the results in Table I, where the accuracies on the classes 1, 3 and 4 are much lower than class 2 in most of the clustering methods.

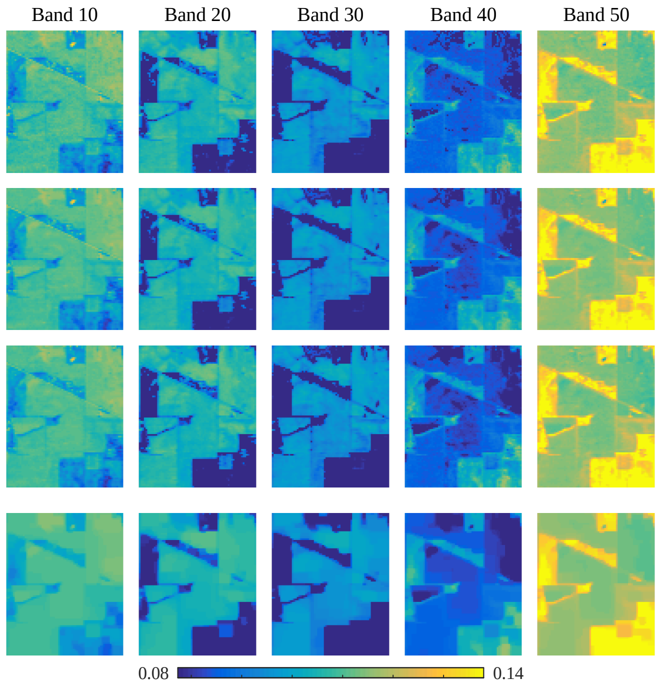
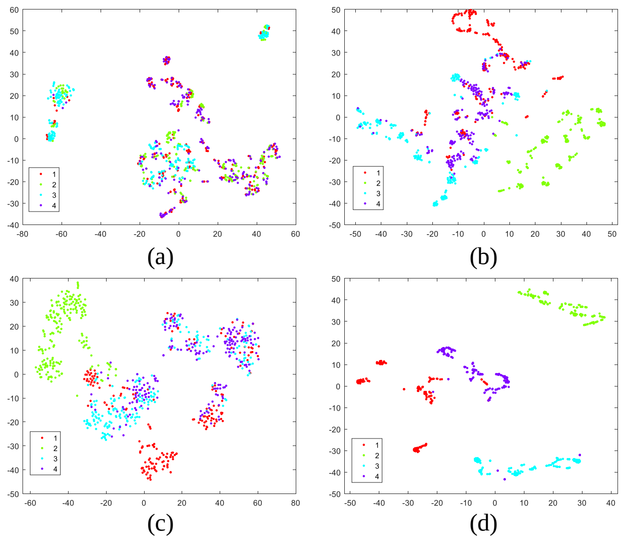


Fig. 11. Comparisons between the original HSI (first row) and the reconstructed data by using the coefficients from SSC (second row), DLSC (third row) and IDLSC (last row). Five bands are shown.

2) Visualization of the Learned Features: We show in Fig. 10 the coefficients of SSC, DLSC and IDLSC on the data set *Indian Pines*. The coefficient maps of SSC (first row) are much more sparse compared with DLSC (second row) and IDLSC (third row). This is because SSC uses the highly redundant self-representation dictionary while DLSC and IDLSC use compact dictionary that is learned from the input data. The



(a)

(b)

(c)

(d)

Fig. 12. Feature visualization by applying (a) raw data, (b) coefficients of SSC, (c) coefficients of DLSC and (d) coefficients of IDLSC in the dimension reduction algorithm t-SNE. The dimension of data is reduced to two.

atoms in the learned dictionary are often more discriminative than those of SSC. This results in denser coefficients in each $\mathcal{A}(:,:,i)$ for the methods DLSC and IDLSC. Compared with SSC, the coefficients of DLSC and IDLSC show much stronger spatial correlations in local regions as shown in Fig. 10. This is important for the subsequent construction of KNN graph in our method as the neighbours searched in the feature domain mostly belong to the same cluster. Due to the use of spatial regularization, the coefficient maps of IDLSC are much smoother than DLSC as shown in Fig. 10, which means that the coefficients of most neighbouring pixels are similar, reducing thereby the within-cluster variance of features.

We obtain reconstructed data by $\hat{\mathbf{Y}} = \mathbf{DA}$. The reconstructed data $\hat{\mathbf{Y}}$ is reshaped to 3-D hyperspectral data and five bands are shown in Fig. 11. Generally, all three methods show good approximations to the input data. SSC (second row) and DLSC (third row) yield closer data reconstructions to the raw data (first row) compared with IDLSC (last row). The bands of the reconstructed data in IDLSC are smoother than the raw data $\mathbf{Y}$ and the data approximations obtained by SSC and DLSC. This is reasonable because of the smoothness of the coefficients in IDLSC. However, the coefficients resulting in better data reconstructions do not necessary perform better in the clustering task.

We apply the learned coefficients in the dimension reduction algorithm: t-distributed stochastic neighbor embedding (t-SNE) [63] and show the results in Fig. 12, where the dimension of coefficients vector is reduced to two and the data points belonging to different clusters are annotated by different colors. The results show that compared to the raw data in Fig. 12 (a), the coefficients learned by SSC, DLSC and IDLSC show improved separability. IDLSC yields the best result where four clusters are almost perfectly separated. The significantly improved separability with IDLSC facilitates the subsequent construction of a more block-diagonal similarity matrix of KNN graph as most of the neighbours defined in Euclidean distance in the feature domain are from the same cluster. This improves thereby the accuracy in spectral clustering.

TABLE IV
DIFFERENT NOISE ADDED INTO INDIAN PINES FOR ROBUST ANALYSIS

Schemes	Noise	Noise level
GN1	Gaussian noise	SNR varying between 30 and 40 dB in each band
GN2	Gaussian noise	SNR varying between 20 and 30 dB in each band
GN3	Gaussian noise	SNR varying between 10 and 20 dB in each band
IN1	Impulse noise	20% corrupted pixels in bands 30-40
IN2	Impulse noise	40% corrupted pixels in bands 30-40
IN3	Impulse noise	60% corrupted pixels in bands 30-40
MN1	Mixed noise	GN3+IN1
MN2	Mixed noise	GN3+IN2
MN3	Mixed noise	GN3+IN3

3) The Influence of Noise: We show the influence of noise on the clustering accuracies of different methods on the data set *Indian Pines* in Fig. 13. Three kinds of noise are considered: Gaussian noise, impulse noise and a mixture of the former two. We add different levels of noise in the data as shown in Table IV. For instance, in scheme GN1 we add Gaussian noise such that signal-to-noise ratio (SNR) varies between 30 and 40 dB in each band, and in scheme IN1 we introduce impulse noise with 20% of corrupted pixels in bands 30-40. ”Raw” in Fig. 13 is the case using the raw HSI.

Generally, the accuracies of all the methods decrease after adding the noise. The mixed noise tends to deteriorate the performance most severely, and impulse noise has more influence on the clustering performance than Gaussian noise. It is observed that in all schemes our method IDLSC yields consistently the highest accuracies. With Gaussian noise, the performance of representation-based methods SSC, L2-SSC, Sketch-SSC, SS-SDAR, HX-NMF, DLSC and IDLSC is more stable than conventional FCM, k-means and CFSFDP, which yield significantly decreased accuracies when high-level of Gaussian noise is added, i.e., the scheme GN3. With impulse noise, spectral-based methods, such as FCM, k-means, CFSFDP, Sketch-SSC, Hx-NMF, DS3C and DLSC, yield poor clustering results compared with “Raw”. This is mainly caused by the significantly increased within-cluster variability from the added impulse noise. While due to the adopted spatial constraints in spatial-spectral clustering methods JSSC, SS-SDAR and IDLSC, they are able to yield more smoothing features in spatial dimensions and thereby reduce the within-cluster variance in feature domain, leading to significantly improved accuracy compared with the spectral-based methods. It is observed that L2-SSC, SS-SDAR and IDLSC yield comparable performance to “Raw” in MN1, but in MN2 and MN3 their accuracies are dropped sharply. The deep learning based model DS3C performs well on low levels of Gaussian noise such as GN1 and GN2 but worse in other schemes. The proposed IDLSC yields comparable accuracies to “Raw” in the schemes of GN1-GN3, IN1 and MN1, but performs worse on high-levels of impulse noise and mixed noise in IN2, IN3, MN2 and MN3.

4) The Analysis of Parameters: The parameters in our approach include two regularization parameters λ and λ_{tv}, the number of neighbours in KNN graph k, the parameter of

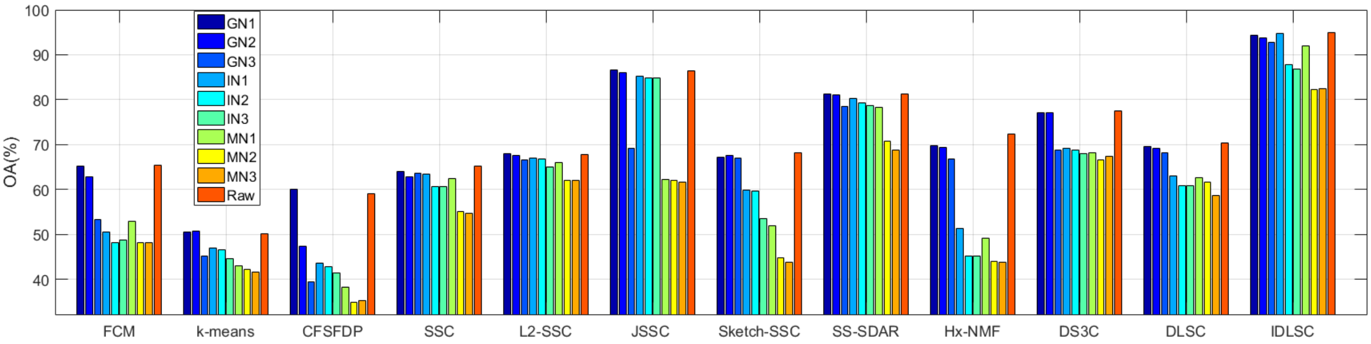


Fig. 13. The influence of noise on the clustering accuracy of different methods on part of *Indian Pines*.

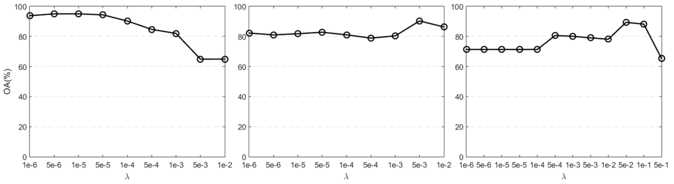


Fig. 14. The overall accuracies with respect to λ in three datasets: *Indian Pines* (left), *HYDICE Urban* (middle) and *University of Houston* (right).

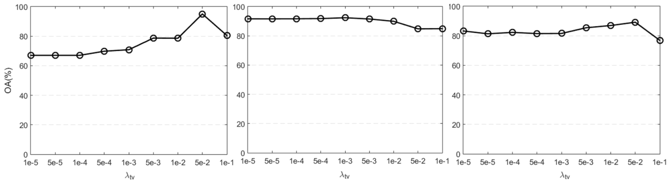


Fig. 15. The overall accuracies with respect to λ_{tv} in three datasets: *Indian Pines* (left), *HYDICE Urban* (middle) and *University of Houston* (right).

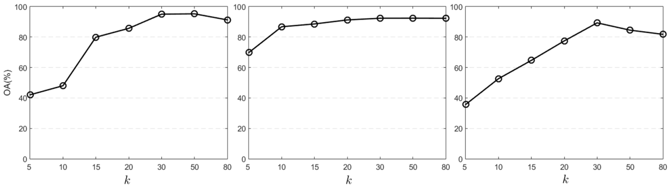


Fig. 16. The overall accuracies with respect to k in three datasets: *Indian Pines* (left), *HYDICE Urban* (middle) and *University of Houston* (right).

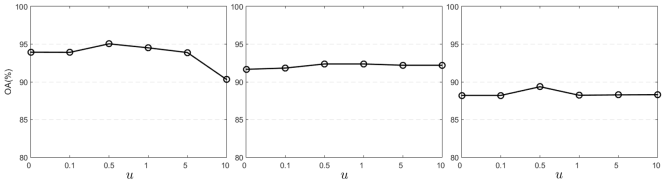


Fig. 17. The overall accuracies with respect to u in three datasets: *Indian Pines* (left), *HYDICE Urban* (middle) and *University of Houston* (right).

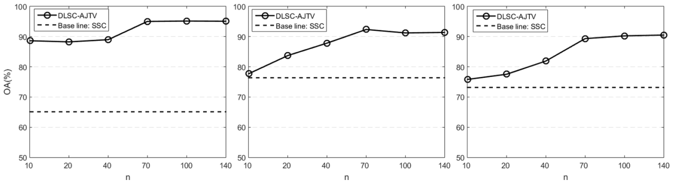


Fig. 18. The overall accuracies with respect to n in three datasets: *Indian Pines* (left), *HYDICE Urban* (middle) and *University of Houston* (right).

weights u and the dictionary size n. We conduct experiments to analyse the parameters on three data sets: *Indian Pines*, *HYDICE Urban* and *University of Houston*, and report the corresponding results in Figs. 14 – 18, showing the curves of overall accuracy with respect to the parameters.

The regularization parameters λ and λ_{tv} in (15) control the balance between three terms, i.e. data fidelity, sparsity of coefficients and consistency of coefficients. The experimental results in Fig. 14 and 15 show a stable performance of our method over a relatively wide ranges of these parameters. On *HYDICE Urban* and *University of Houston* the performance is very stable with respect to both parameters. In *Indian Pines*, the parameters with larger λ_{tv} and smaller λ often performs better, indicating the importance of incorporating spatial information in this data. The results in *HYDICE Urban* and *University of Houston* show that the parameter λ has slightly more effect on the clustering performance than λ_{tv}. It is also noticed that in most cases our method performs better than the original SSC method with higher accuracy.

We also present the clustering results with respect to other parameters including k, u and n in Figs. 16 – 18. The optimal settings of these three parameters are less dependent on the particular data sets, which implies that we can fix them for all the data sets, achieving satisfactory clustering performance. Fig. 16 indicates that the accuracies increase dramatically at the beginning when k is small, and then reach to the peak at $30 \leq k \leq 50$ before the following decrease at $k > 50$. Hence, we set $k = 30$ in our experiments for all the tested data sets, which is also the reason why the similarity matrix in our method is often sparser than that in SSC, JSSC and L2-SSC, as indicated in Fig. 2. Such sparse similarity matrix is particularly beneficial for big data because of the reduced memory requirement. The results in Fig. 17 show that $u = 0.5$ is a reasonable choice for all the three data sets. In Fig. 18, we can notice that a larger dictionary often yields better accuracy. The clustering accuracy gradually increases along with increasing n, and then saturates. Notably, when the dictionary size is very small (e.g., $n = 10$), our method still outperforms the original SSC method on the three data

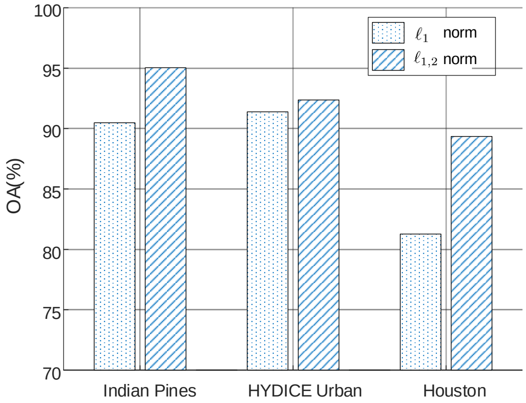


Fig. 19. Performance comparisons between the models with ℓ_1 norm based TV constraint and $\ell_{1,2}$ norm based TV constraint on three data sets.

sets, which demonstrates the great potential of using compact dictionary in subspace clustering. According to the parameter study, we suggest to set $\lambda = 10^{-3}, \lambda_{tv} = 5 \times 10^{-2}, n = 70, k = 30$ and $u = 0.5$, which yields reasonably good clustering performance in different data sets.

5) Ablation Study: To validate the benefit of $\ell_{1,2}$ norm based AJTV, we compare our method with its modified version where $\|\mathbf{WHA}^T\|_{1,2}$ in the proposed model (15) is replaced with the ℓ_1 norm based regularization $\|\mathbf{WHA}^T\|_1$. The results in Fig. 19 reveal that the model with $\ell_{1,2}$ norm based TV regularization yields consistently better clustering performance in terms of OA than the ℓ_1 norm regularized model on the three data sets. Especially on the data set *University of Houston*, the OA improvement is remarkable: more than 8%. We observe that even with ℓ_1 norm, our method outperforms the reference methods in Tables I – III on the three data sets.

6) Performance on Unbalanced Data Set: The experimental results on the data set *Indian Pines* with 16 classes are shown in Table V. The 16 classes in the *Indian Pines* are known to be unbalanced [61]. For instance, the class "oats" has only 20 samples while the class "soybean-mintill" has a total number of 2455 samples. The results in Table V show that clustering on this data set is difficult: most of the tested clustering methods yield the OA even below 40%. IDLSC yields the best clustering performance in terms of OA and APR. Compared with JSSC, which achieves the second best result, our method obtains a significant OA improvement of more than 12%. SS-SDAR obtains a comparable APR but shows a much worse OA and a lower running speed in comparison with IDLSC. Compared with Hx-NMF, our method shows better OA, APR and running time. Hx-NMF is less efficient on large-scale data sets where it needs to update huge similarity matrices. Compared to SSC, L2-SSC and JSSC, IDLSC shows a significant speed improvement due to the learned compact dictionary. Sketch-SSC and k-means are the fastest methods but their results are inferior to most others in terms of the accuracy.

7) Convergence Study: We show in Fig. 20 the evolution of objective function of the proposed model with respect to the number of iterations on the three data sets. The results reveal that the objective function monotonically decreases to a stable level on the three HSIs. Roughly, our proposed model converges within 20 iterations. In the data sets *Indian Pines* and *NCALM University of Houston*, the objective function drops sharply in the first several iterations and then saturates, demonstrating the fast convergence of our algorithm.

TABLE V
CLUSTERING RESULTS ON UNBALANCED DATA SET: *INDIAN PINES* WITH 16 CLASSES ("OOM" MEANS "OUT OF MEMORY")

Methods	OA(%)	APR(%)	Time (in seconds)
FCM	31.57	31.71	65
k-means	37.41	37.36	10
CFSFDP	35.99	14.96	131
SSC	34.80	34.87	16906
L2-SSC	42.10	39.20	20769
JSSC	50.90	44.38	18326
Sketch-SSC	36.93	20.04	**7**
SS-SDAR	43.12	51.53	516
Hx-NMF	41.02	26.14	734
DS3C	OOM	OOM	OOM
DLSC	40.68	39.61	103
IDLSC	**63.45**	**52.30**	306

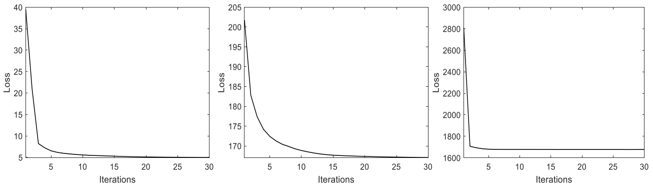


Fig. 20. The evolution of loss of the proposed model with respect to the number of iterations on three data sets: *Indian Pines* (left), *HYDICE Urban* (middle) and *NCALM University of Houston* (right).

V. CONCLUSION

In this paper, we propose a novel subspace clustering method for hyperspectral remote sensing images by using dictionary learning technique with an adaptive joint total variation regularization. In particular, we employ a compact dictionary learned from the data to model the low-dimensional subspaces of data, which enables a more efficient subspace clustering method. To capture the important local geometric data structure, a joint total variation with a $\ell_{1,2}$ norm is incorporated in our model. To discriminate the data points in different regions, the joint TV is adaptively weighted according to the coefficients matrix. Consequently, the consistency and discriminability of the coefficients get improved in the subspace representation, which increases the robustness of model to noise and spectral variability, facilitating simultaneously the construction of a desired similarity matrix. Furthermore, we develop an effective solver to obtain the solution for the resulting optimization problem based on alternating minimization and alternating direction method of multipliers. The experiments on real HSIs show the effectiveness of our algorithm and its superiority over the state-of-the-art.

VI. ACKNOWLEDGEMENT

The authors would like to thank the Hyperspectral Image Analysis Group and the NSF-Funded Center for Airborne

Laser Mapping, University of Houston, for providing the *University of Houston* image.

References

[1] B. Datt, T. R. McVicar, T. G. Van Niel, D. L. Jupp, and J. S. Pearlman, "Preprocessing eo-1 hyperion hyperspectral data to support the application of agricultural indexes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1246–1259, 2003.

[2] M. A. Lee, Y. Huang, H. Yao, S. J. Thomson, and L. M. Bruce, "Determining the effects of storage on cotton and soybean leaf samples for hyperspectral analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2562–2570, 2014.

[3] H. Zhang, J. Kang, X. Xu, and L. Zhang, "Accessing the temporal and spectral features in crop type mapping using multi-temporal sentinel-2 imagery: A case study of Yi'an county, Heilongjiang province, China," *Comput. Electron. Agric.*, vol. 176, p. 105618, 2020.

[4] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, 2014.

[5] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proc. IEEE*, vol. 97, no. 6, pp. 1031–1055, 2009.

[6] Y.-Z. Feng and D.-W. Sun, "Application of hyperspectral imaging in food safety inspection and control: a review," *Crit. Rev. Food Sci. Nutr.*, vol. 52, no. 11, pp. 1039–1058, 2012.

[7] N. Fox, A. Parbhakar-Fox, J. Moltzen, S. Feig, K. Goemann, and J. Huntington, "Applications of hyperspectral mineralogy for geoenvironmental characterisation," *Miner. Eng.*, vol. 107, pp. 63–77, 2017.

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

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, 2013.

[10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[11] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[12] Y. Zhong, L. Zhang, and W. Gong, "Unsupervised remote sensing image classification using an artificial immune network," *Int. J. Remote Sens.*, vol. 32, no. 19, pp. 5461–5483, 2011.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. AM. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[14] A. Goh and R. Vidal, "Segmenting motions of different types by unsupervised manifold clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[15] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 217–240, 2012.

[16] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Sparsity-based clustering for large hyperspectral remote sesning images," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[17] Y. Cai, Z. Zhang, Z. Cai, X. Liu, and X. Jiang, "Hypergraph-structured autoencoder for unsupervised and semisupervised classification of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, 2021.

[18] 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, 2021.

[19] J. Sun, W. Wang, X. Wei, L. Fang, X. Tang, Y. Xu, H. Yu, and W. Yao, "Deep clustering with intraclass distance constraint for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4135–4149, 2021.

[20] K. Li, Y. Qin, Q. Ling, Y. Wang, Z. Lin, and W. An, "Self-supervised deep subspace clustering for hyperspectral images with adaptive self-expressive coefficient matrix initialization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3215–3227, 2021.

[21] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, 2011.

[22] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2790–2797.

[23] ——, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[24] S. Huang, H. Zhang, and A. Pižurica, "Semisupervised sparse subspace clustering method with a joint sparsity constraint for hyperspectral remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 989–999, 2019.

[25] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Nonlocal means regularized sketched reweighted sparse and low-rank subspace clustering for large hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4164–4178, 2021.

[26] S. Jia, X. Deng, J. Zhu, M. Xu, J. Zhou, and X. Jia, "Collaborative representation-based multiscale superpixel fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7770–7784, 2019.

[27] Y. Yuan, X. Zheng, and X. Lu, "Spectral–spatial kernel regularized for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3815–3832, 2015.

[28] H. Zhang, L. Liu, W. He, and L. Zhang, "Hyperspectral image denoising with total variation regularization and nonlocal low-rank tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3071–3084, 2020.

[29] H. Zhang, J. Cai, W. He, H. Shen, and L. Zhang, "Double low-rank matrix decomposition for hyperspectral image denoising and destriping," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[30] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote sensing image spatiotemporal fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4273–4286, 2020.

[31] C. Yi, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral image super-resolution based on spatial and spectral correlation fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4165–4177, 2018.

[32] H. Zhang, H. Zhai, L. Zhang, and P. Li, "Spectral–spatial sparse subspace clustering for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3672–3684, 2016.

[33] H. Zhai, H. Zhang, X. Xu, L. Zhang, and P. Li, "Kernel sparse subspace clustering with a spatial max pooling operation for hyperspectral remote sensing data interpretation," *Remote Sens.*, vol. 9, no. 4, p. 335, 2017.

[34] J. Xu, N. Huang, and L. Xiao, "Spectral-spatial subspace clustering for hyperspectral images via modulated low-rank representation," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2017, pp. 3202–3205.

[35] H. Zhai, H. Zhang, L. Zhang, P. Li, and A. Plaza, "A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 43–47, 2017.

[36] S. Huang, H. Zhang, and A. Pižurica, "Joint sparsity based sparse subspace clustering for hyperspectral images," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 3878–3882.

[37] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Total variation regularized collaborative representation clustering with a locally adaptive dictionary for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 166–180, 2019.

[38] P. A. Traganitis and G. B. Giannakis, "Sketched subspace clustering," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1663–1675, 2018.

[39] S. Huang, H. Zhang, Q. Du, and A. Pizurica, "Sketch-based subspace clustering of hyperspectral images," *Remote Sens.*, vol. 12, no. 5, 2020.

[40] Y. Yankelevsky and M. Elad, "Theoretical guarantees for graph sparse coding," *Appl. Comput. Harmon. Anal.*, vol. 49, no. 2, pp. 698–725, 2020.

[41] N. Huang and L. Xiao, "Hyperspectral image clustering via sparse dictionary-based anchored regression," *IET Image Process.*, vol. 13, no. 2, pp. 261–269, 2018.

[42] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[43] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[44] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[45] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral image restoration using low-rank matrix recovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4729–4743, 2013.

[46] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2691–2698.

[47] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, 2012.

[48] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[49] W. Fu, S. Li, L. Fang, and J. A. Benediktsson, "Contextual online dictionary learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1336–1347, 2017.

[50] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, 2018.

[51] X. Han, J. Yu, J. Luo, and W. Sun, "Reconstruction from multispectral to hyperspectral image using spectral library-based dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1325–1335, 2018.

[52] Y. Yuan, D. Ma, and Q. Wang, "Hyperspectral anomaly detection via sparse dictionary learning method of capped norm," *IEEE Access*, vol. 7, pp. 16 132–16 144, 2019.

[53] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[54] L. Condat, "Discrete total variation: New definition and minimization," *SIAM J. Imaging Sci.*, vol. 10, no. 3, pp. 1258–1290, 2017.

[55] M. A. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.

[56] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[57] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Sim.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[58] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

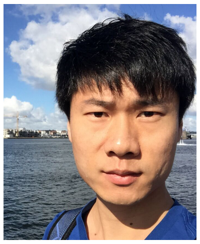
[59] Q. Wang, X. He, X. Jiang, and X. Li, "Robust bi-stochastic graph regularized matrix factorization for data clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[60] L. Lovász and M. D. Plummer, *Matching theory*. American Mathematical Soc., 2009, vol. 367.

[61] B. Xue, C. Yu, Y. Wang, M. Song, S. Li, L. Wang, H.-M. Chen, and C.-I. Chang, "A subpixel target detection approach to hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5093–5114, 2017.

[62] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, 2014.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.



Shaoguang Huang (Member, IEEE) received the M.S. degree in telecommunication and information system from Shandong University, Jinan, China, in 2015 and the Ph.D. degree in computer science engineering from Ghent University, Belgium, in 2019.

He is currently a Post-Doctoral Research Fellow with the Group for Artificial Intelligence and Sparse Modelling (GAIM), Ghent University, Belgium. His area of interests includes image processing, sparse representation, clustering, hyperspectral image analysis and machine learning.

Dr. Huang served as a Guest Editor for *Remote Sensing*. He is also a Reviewer of the international journals, including IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Biomedical Circuits and Systems, *Knowledge-Based Systems, Pattern Recognition, Remote Sensing*, *Computers & Geosciences*.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2005 and 2010, respectively.

He has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, since 2016. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China. He has authored/co-authored more than 100 research articles and eight patents. His research interests include image reconstruction for quality improvement, hyperspectral information processing and agricultural remote sensing.

Dr. Zhang scored first in the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest in 2019 and 2021 organized by the IEEE Image Analysis and Data Fusion Technical Committee. He has served as the Session Chair of the 2016 IEEE International Symposium on Geoscience and Remote Sensing (IGARSS) Conference and the 2015 IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) Conference. He also serves as an Associate Editor for *Photogrammetric Engineering* & *Remote Sensing* and *Computers* & *Geosciences*. He is also a Reviewer of more than 30 international academic journals, including IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Image Processing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters.

Aleksandra Pižurica (Senior Member, IEEE) received the Diploma in electrical engineering from the University of Novi Sad, Serbia, in 1994, the Master of Science degree in telecommunications from the University of Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Belgium, in 2002.

She is a Professor in statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov Random Field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Prof. Pižurica received the scientific prize "de Boelpaepe" for 2013–2014 awarded by the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis. The work of her team has been awarded twice the Best Paper Award of the IEEE Geoscience and Remote Sensing Society Data Fusion contest, in 2013 and 2014. She has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2012 – 2016), Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2016 – 2019) and an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2016 – 2020). She was a Lead Guest Editor for the *EURASIP Journal on Advances in Signal Processing* Special Issue "Advanced Statistical Tools for Enhanced Quality Digital Imaging with Realistic Capture Models" in 2013.