Deep Spatial-Spectral Subspace Clustering for Hyperspectral Image

Jianjun Lei, *Senior Member, IEEE,* Xinyu Li, Bo Peng, Leyuan Fang, *Senior Member, IEEE,* Nam Ling, *Fellow, IEEE,* and Qingming Huang, *Fellow, IEEE*

***Abstract*—Hyperspectral image (HSI) clustering is a challenging task due to the complex characteristics in HSI data, such as spatial-spectral structure, high-dimension, and large spectral variability. In this paper, we propose a novel deep spatial-spectral subspace clustering network (DS^{3}C-Net), which explores spatial-spectral information via the multi-scale auto-encoder and collaborative constrain. Considering the structure correlations of HSI, the multi-scale auto-encoder is first designed to extract spatial-spectral features with different-scale pixel blocks which are selected as the inputs. Then, the collaborative constrained self-expressive layers are introduced between the encoder and decoder, to capture the self-expressive subspace structures. By designing a self-expressiveness similarity constraint, the proposed network is trained collaboratively, and the affinity matrices of the feature representation are learned in an end-to-end manner. Based on the affinity matrices, the spectral clustering algorithm is utilized to obtain the final HSI clustering result. Experimental results on three widely used hyperspectral image datasets demonstrate that the proposed method outperforms state-of-the-art methods.**

***Index Terms*—Hyperspectral image clustering, deep subspace clustering, multi-scale auto-encoder, self-expressiveness similarity constraint, deep learning.**

I. Introduction

HYPERSPECTRAL image (HSI) has attracted lots of attention and became a significant data source in a variety of remote sensing applications [1]–[4]. As one of the common and fundamental techniques, clustering for HSI has been widely used in environmental monitoring, geological mapping, mineral exploration, and vegetation survey [5]–[8]. Recently, many HSI classification methods based on deep learning have achieved promising performance [9]–[13]. However, deep networks that rely on supervised learning generally require a huge amount of annotated data. In practice, obtaining the labels of HSI pixels is a high time-consuming and labor-intensive progress. Therefore, it is significant to recognize the categories among HSI pixels in an unsupervised manner. To this end, HSI clustering method has attracted extensive attention. However, due to the high-dimension, large spectral variability, as well as no training labels, HSI clustering is still a challenging task in the field of remote sensing.

HSI clustering aims to divide pixels into different groups [14], in which similar pixels corresponding to one specific land-cover class are assigned into the same group. Traditional clustering methods designed for natural images [15]–[19] can be applied to HSI clustering. However, the performance is limited due to the complex characteristics of HSI data. Existing HSI clustering methods mainly include two categories, namely, spectral-only methods and spatial-spectral methods. The spectral-only methods [20], [21] generally cluster the HSI pixels by learning the feature representations in the spectral domain. Intuitively, a specific land-cover class is generally represented by an area with multiple pixels, thus the center pixel and its neighboring pixels are likely belonging to the same category. However, the methods only using spectral information ignore the spatial relationship between the neighboring pixels. Thus, spatial-spectral HSI clustering methods [22]–[24] were proposed to jointly utilize the spatial and spectral information, and achieved better performance than the spectral-only methods. In addition, considering the high-dimensional property of HSIs, some sparse subspace clustering (SSC) based methods [25]–[27] are proposed for HSI clustering. These methods learn a coefficient representation matrix and construct an affinity matrix to cluster the HSI data in the low-dimensional subspace. However, in practical applications, HSI data has large spectral variability and lies in a non-linear subspace. Although kernel-based spectral-spatial subspace clustering methods have been proposed to capture the non-linear characteristics of the HSI data [28], [29], these methods usually choose kernel functions based on experience, which cannot guarantee the feature space learned by the predefined kernel function is indeed suitable for clustering.

Recently, deep learning based clustering methods have been introduced to tackle the nonlinearity problem, and achieved promising performance in the task of natural image clustering [30]–[34]. Most existing deep clustering methods consist of two procedures, i.e., deep feature extracting and traditional clustering. As one of the typical unsupervised learning frameworks, auto-encoder has been used in clustering methods to learn the feature representations in an unsupervised manner. In these methods, the images are first encoded to obtain the feature representations by encoder, and then reconstructed images from the latent feature representations by decoder.

This work was supported in part by the National Natural Science Foundation of China under Grants 61722112, 61922029, 61620106009, and U1636214, and in part by the Natural Science Foundation of Tianjin under Grants 18ZXZNGX00110 and 18JCJQJC45800. (Corresponding author: Bo Peng.)

J. Lei, X. Li and B. Peng are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jjlei@tju.edu.cn; 601014361@qq.com; bpeng@tju.edu.cn).

L. Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: fangleyuan@gmail.com).

N. Ling is with the Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053, USA (e-mail: nling@scu.edu).

Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qmhuang@ucas.ac.cn).

Digital Object Identifier

Based on the extracted feature representations, traditional clustering algorithms or classification layers are applied for obtaining the final clustering result. Inspired by the success of deep natural image clustering methods, a Laplacian regularized deep subspace clustering method (LRDSC) [35] is proposed for HSI, where spatial-spectral features are extracted via a 3D auto-encoder network. Particularly, for HSI data, a specific land-cover class is generally represented by an area with multiple pixels, and thus how to make better use of the multi-scale spatial information and extract the discriminative spatial-spectral features are essential for the HSI clustering task.

In this paper, we focus on exploring the multi-scale spatial-spectral information from pixel blocks at different scales, and propose a deep spatial-spectral clustering network to learn the similarity relationship among the pixels for HSI clustering. Specifically, the multiple pixel blocks are selected for each central pixel to generate the multi-scale spatial-spectral feature, and the collaborative constrained self-expressive layers are introduced to capture the self-expressive subspace structures. The main contributions of this paper are summarized as follows.

1) To address the problems of deep subspace clustering for HSI, we propose a novel deep spatial-spectral subspace clustering network (DS^{3}C-Net) to generate discriminative spatial-spectral features.

2) A multi-scale auto-encoder is designed to explore spatial-spectral information with different-scale pixel blocks which are selected as the inputs.

3) The collaborative constrained self-expressive layers are designed to learn the subspace structures, and a self-expressiveness similarity constraint is introduced to collaboratively train the proposed network.

4) Experiments on three widely used HSI datasets demonstrate that the proposed method outperforms state-of-the-art methods.

The rest of the paper is organized as follows. Section II reviews the related works. Section III introduces the proposed network in detail. Section IV presents the experimental results and corresponding analyses. Finally, the conclusion is drawn in Section V.

II. Related Works

A. Traditional hyperspectral image clustering methods

Considering the spectral relationship in HSI, many spectral-only methods have been proposed for HSI clustering. For instance, Paoli *et al.* [14] proposed a clustering methodology to solve problems of class number estimation, feature extraction, and clustering simultaneously in an unsupervised way. Zhong *et al*. [20] proposed an artificial immune network for HSI clustering, in which the user-defined parameters are adaptively obtained and the inherent structure of HSI data is simulated by a biological model. In [21], an automatic fuzzy clustering method was proposed by transforming the clustering problem into a multi-objective problem, which better adapts to the complexity of remote sensing images. However, these spectral-only methods only focused on spectral information, while ignoring the spatial relationship between neighboring pixels.

The spatial-spectral clustering methods jointly utilize the spatial information and spectral information to get more discriminative features for HSIs. In [22], a spatial constraint based fuzzy C-means clustering method was proposed to exploit spatial contextual information. Lin *et al.* [23] proposed a dual-space transfer learning method to preserve local and global structures of HSI data. Murphy and Magioni [24] proposed a diffusion learning-based spatial-spectral clustering method (DLSS) to incorporate spatial and spectral information with a diffusion-inspired labeling. In [36], an extended diffusion learning-based clustering framework was proposed by incorporating spatial information into the underlying diffusion matrix. Besides, Zhang *et al.* [37] proposed an adaptive fuzzy local information C-means clustering algorithm, in which a fuzzy local similarity measure is used to incorporate local spatial and gray level relationships between the center pixel and its neighboring pixels. Morsier *et al.* [38] proposed a graph representation method that is discriminative of the cluster structure of the data. Zhai *et al.* [39] proposed a total variation regularized collaborative representation clustering algorithm to utilize the rich spatial-contextual information in HSI. The sparse subspace clustering (SSC) [25] method mapping data points into latent subspaces has also been applied into HSI task. Based on the typical SSC, Zhai *et al.* [40] and Sun *et al.* [41] proposed subspace clustering based hyperspectral image band selection methods, where subspace clustering model is used to capture the structure information from the learned representation. Tian *et al.* [28] proposed a kernel spatial-spectral based multi-view low-rank sparse subspace clustering method, where the Gaussian kernel is applied. Long *et al.* [29] proposed a Gaussian kernel dynamic similarity matrix based sparse subspace clustering method to calculate the similarity between pixel points by using the Gaussian kernel function. Xu *et al.* [42] proposed a spectral-spatial low-rank subspace clustering (SS-LRSC) algorithm to exploit the structure correlations by introducing a modulation strategy to the low rank representation matrix. Li. *et al.* [43] proposed a three-dimensional edge-preserving filtering based sparse subspace clustering method, where 3D edge-preserving filtering is introduced to extract the spectral-spatial information. Zhang *et al.* [26] proposed a spectral-spatial sparse subspace clustering (S^{4}C) algorithm, in which the spectral similarity of a local neighborhood is computed to exploit spatial information. In [27], a L2-norm regularized SSC (L2-SSC) method was proposed to exploit the spatial-spectral information, and achieved better clustering performance. However, SSC based methods generally utilize linear embedding functions to capture the subspace structure, but failing to capture the non-linear subspace structure of the complex high-dimensional HSIs with large spectral variability.

B. Deep clustering methods

Recently, some frameworks based on deep learning [30]–[34] have been proposed for clustering task. Deep embedded clustering (DEC) [30] is a pioneering work, which simultaneously learns feature representations and cluster assignments using deep networks. By utilizing a binary constrained pairwise-classification model, Chang *et al.* [32] proposed a

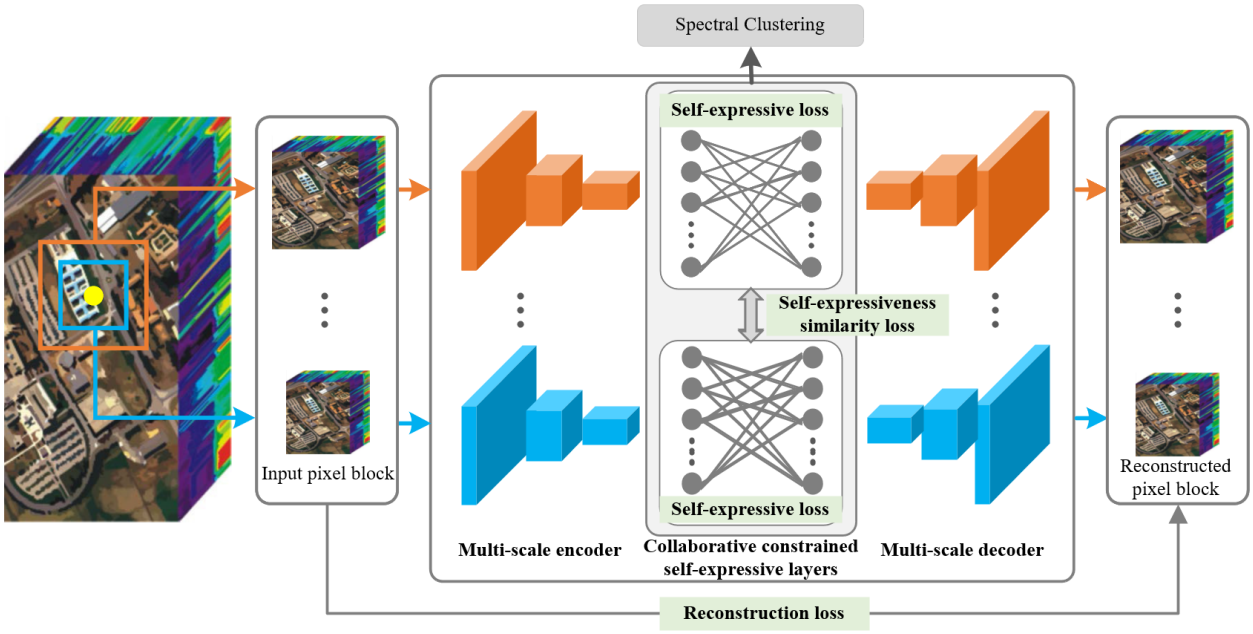

Fig. 1. Architecture of the proposed DS^{3}C-Net.

deep adaptive image clustering (DAC) method to recast the clustering problem into a binary pairwise-classification framework. Mukherjee *et al.* [33] and Chen *et al.* [34] proposed generative adversarial network (GAN) based natural image clustering methods, where GAN is used to simultaneously generate and cluster images by learning the discriminative features in the latent space. Based on the SSC, Ji *et al.* [31] proposed a deep subspace clustering network (DSC-Net), in which convolutional autoencoders are used to nonlinearly map the data into a latent space. Compared with the traditional clustering methods, these deep clustering methods can learn the nonlinear mappings from the data domain into the latent spaces, thus getting more robust clustering results. Inspired by the success of deep natural image clustering methods, Zeng *et al.* [35] proposed a Laplacian regularized deep subspace clustering (LRDSC) for HSI clustering, where spatial-spectral features are extracted via a 3D auto-encoder network. However, a specific land-cover class in HSI data is generally represented by an area with multiple pixels, making better use of the multi-scale spatial information is essential for the HSI clustering task. Therefore, different from the method that extracts the pixel features by a 3D structure, this paper focuses on exploring the multi-scale spatial-spectral information from pixel blocks at different spatial scales, and proposes a deep spatial-spectral clustering network to learn the similarity relationship among the pixels for HSI clustering.

III. Deep Spatial-Spectral Subspace Clustering Network

A. Architecture of DS^{3}C-Net

In this paper, a DS^{3}C-Net is proposed to effectively explore the spatial-spectral information for HSI. The architecture of the DS^{3}C-Net is shown in Fig. 1. Considering the structure correlations, different-scale pixel blocks around the center pixel are first selected as the inputs of the proposed network and the multi-scale auto-encoder is designed to extract spatial-spectral features in an unsupervised manner. Then, the collaborative constrained self-expressive layers are introduced between the encoder and decoder networks. Based on the spatial-spectral features extracted by encoders, self-expressive property of the HSI data is captured by optimizing the collaborative constrained self-expressive layers. By training the proposed network in a collaborative scheme, the final self-expressiveness coefficient matrix is obtained, and the spectral clustering is applied on the coefficient matrix for obtaining the HSI clustering result.

B. Multi-scale auto-encoder

HSI clustering aims to divide the pixels in HSI into different clusters. Generally, naive methods take the pixels as the inputs and cluster the pixels by extracting the features from the pixels directly. However, a specific land-cover class is generally represented by an area with multiple pixels, the center pixel and its neighboring pixels are likely belonging to the same category. Therefore, utilizing the spatial information around the center pixel will be beneficial for boosting the clustering performance. In addition, pixel blocks at different spatial scales emphasize the different characteristics of the input, i.e., larger pixel blocks provide richer spatial information and smaller spatial size pay more attention to the spatial correlation between the pixel and its adjacent pixels. In order to capture the spectral information and spatial correlation effectively, different-scale pixel blocks are used as the input of the DS^{3}C-Net to learn comprehensive spatial-spectral features.

In this paper, the original 3-D HSI pixel block $\mathbf{X}^{\text{ori}}_{Pb_{i,j}} \in \mathbb{R}^{w_i \times w_i \times d}$ is first converted to 2-D HSI pixel block $\mathbf{X}_{Pb_{i,j}} \in \mathbb{R}^{{w_i}^2 \times d}$, where $\mathbf{X}^{\text{ori}}_{Pb_{i,j}}$ and $\mathbf{X}_{Pb_{i,j}}$ denotes the original and

converted pixel block of the j-th center pixel at i-th scale, d denotes the number of spectral bands, and w_i denotes the spatial scale of the pixel block. Specifically, the center pixel of the pixel block is fixed as the starting point in 2-D data, and the other pixels in same block are arranged according to the distance from the center pixel. In this way, spatial and spectral information can be coped with simultaneously through 2-D convolution kernels without losing spectral information and spatial correlation. Let $\mathbf{X}_{Pb_i} = [\mathbf{X}_{Pb_{i,1}}, \mathbf{X}_{Pb_{i,2}}, ..., \mathbf{X}_{Pb_{i,m}}] \in \mathbb{R}^{m \times {w_i}^2 \times d}$ denotes pixel blocks of all pixels at i-th scale, $\mathbf{X}_{Pb_{i,j}}$ denotes the pixel block of the j-th pixel at i-th scale, and m denotes the total number of pixels. In the proposed DS^{3}C-Net, the $\mathbf{X}_{Pb_i}$ is mapped into the latent representations $\mathbf{Z}_{Pb_i} = [\mathbf{Z}_{Pb_{i,1}}, \mathbf{Z}_{Pb_{i,2}}, ..., \mathbf{Z}_{Pb_{i,m}}] \in \mathbb{R}^{m \times l_i}$ by the encoder network, $\mathbf{Z}_{Pb_{i,j}}$ denotes the latent representation of the j-th pixel at i-th scale, and l_i denotes the dimension of the latent representations at the i-th scale. The operation of encoder is formulated as:

$$\mathbf{Z}_{Pb_i} = f_{\Theta_{Ei}}(\mathbf{X}_{Pb_i}) \tag{1}$$

where $f_{\Theta_{Ei}}(\cdot)$ denotes the nonlinear map function of the encoder at the i-th scale, and Θ_{Ei} denotes the parameters of the encoder network. After that, the latent representation $\mathbf{Z}_{Pb_i}$ is embedded into the collaborative constrained self-expressive layers to explore the self-expressive property, and the self-expressive representation $\mathbf{C}_{pb_i}\mathbf{Z}_{Pb_i}$ is obtained, where $\mathbf{C}_{Pb_i} \in \mathbb{R}^{m \times m}$ corresponds to the parameters of self-expressive layer at i-th scale. By using the collaborative constrained self-expressive layers, the latent representations at different scales are interrelated to exploit the self-expressive property, which will be introduced in detail in Section C. In order to extract the features in an unsupervised manner, the multi-scale decoder is introduced to reconstruct the multi-scale input pixel blocks. Let $\hat{\mathbf{X}}_{Pb_i} \in \mathbb{R}^{m \times {w_i}^2 \times d}$ denotes the reconstructed pixel blocks of the i-th scale. The operation of decoder is formulated as:

$$\hat{\mathbf{X}}_{Pb_i} = \hat{f}_{\Theta_{Di}}(\mathbf{C}_{pb_i}\mathbf{Z}_{Pb_i}) \tag{2}$$

where $\hat{f}_{\Theta_{Di}}(\cdot)$ denotes the deconvolution operation of the decoder at the i-th scale, and Θ_{Di} denotes the corresponding parameters of the decoder network. Intuitively, the HSIs reconstructed with high quality indicate that the features can describe the input HSIs better. Thus, the HSIs reconstructed by the decoder should be as similar to the original HSIs as possible. To this end, the reconstruction error between the input and reconstruct pixel blocks is used as the global constraint for the network training. Therefore, the reconstruction loss L_{re} can be formulated as:

$$L_{\mathrm{re}} = \sum_{i=1}^{n} ||\mathbf{X}_{Pb_i} - \hat{\mathbf{X}}_{Pb_i}||_2^2 \tag{3}$$

where n denotes the number of the scales. By introducing the reconstruction loss function, the learned feature representations contain meaningful spatial-spectral information, which is beneficial for HSI clustering task.

C. Collaborative constrained self-expressive layers

In the proposed method, to capture the spectral information and spatial correlation effectively, different-scale pixel blocks of a pixel are selected as the input of the proposed network to learn comprehensive spatial-spectral features. Therefore, the pixel blocks at different scales share the same center pixel, and multiple scales are inputted into multi-stream encoders simultaneously. After the non-linear transformations with the encoders, the spatial-spectral features of different-scale pixel blocks are extracted in the network. To learn the self-expressive subspace structures of multi-scale pixel blocks, the collaborative constrained self-expressive layers are designed between the encoder and decoder networks. Here, the self-expressive property is related to that of local reconstruction which is first articulated in [25], which means that each data sample can be represented as a linear combination of other samples in the same subspace. Then, based on the self-expressive property, the clustering result is obtained by using the subspace clustering algorithm that has been demonstrated the mathematical and theoretical guarantees in [44], [45]. For the latent representation $\mathbf{Z}_{Pb_i}$ at i-th scale, the self-expressive property at i-th self-expressive layer is formulated as follows.

$$\mathbf{Z}_{Pb_i} = \mathbf{C}_{Pb_i}\mathbf{Z}_{Pb_i} + \mathbf{E}_{Pb_i} \quad \text{s.t. } \mathrm{diag}(\mathbf{C}_{Pb_i}) = \mathbf{0} \tag{4}$$

where $\mathbf{C}_{Pb_i} \in \mathbb{R}^{m \times m}$ denotes the self-expressiveness coefficient matrix, $\mathbf{E}_{Pb_i} \in \mathbb{R}^{m \times m}$ denotes the error matrix, and the constraint $\mathrm{diag}(\mathbf{C}_{Pb_i}) = \mathbf{0}$ is used to exclude trivial solutions. Based on the self-expressive property, the feature representation of each pixel is represented as a linear combination of the other pixels. In the proposed network, a fully-connected layer without bias and activation function is utilized as the self-expressive layer, in which $\mathbf{C}_{Pb_i}$ corresponds to the parameters of self-expressive layer at i-th scale. By introducing self-expressive layer at each stream, the network is trained to explore the self-expressive property of each scale pixel block. To obtain a more accurate and comprehensive coefficient matrix with the multiple self-expressiveness coefficient matrices in multiple streams, the collaborative constrained self-expressive layers are constrained by utilizing a self-expressive constraint and a self-expressiveness similarity constraint.

1) Self-expressive constraint: Self-expressive constraint L_{sc} is introduced to encourage the encoders to learn the feature representations suitable for subspace clustering. Self-expressive constraint works by minimizing the reconstruction error matrices between the latent representations acquired from the encoders $\mathbf{Z}_{Pb_i}$ and self-expressive representations generated by the self-expressive layers, which is formulated as:

$$\begin{aligned} L_{\mathrm{sc}} &= \sum_{i=1}^{n} ||\mathbf{Z}_{Pb_i} - \mathbf{C}_{Pb_i}\mathbf{Z}_{Pb_i}||_2^2 + \sum_{i=1}^{n} ||\mathbf{C}_{Pb_i}||_2 \\ \text{s.t.} \quad & \mathrm{diag}(\mathbf{C}_{\mathrm{Pb_i}}) = \mathbf{0} \end{aligned} \tag{5}$$

where $\mathbf{C}_{Pb_i}\mathbf{Z}_{Pb_i}$ denotes the output of self-expressive layers at the i-th scale, n denotes the number of the scales. The first term measures the difference between the la-

tent representations and self-expressive representations after self-expressiveness, which is used to encourage the self-expressiveness coefficient matrix $\mathbf{C}_{Pb_i}$ to capture the subspace structure in HSI. The second term denotes the regularization term of the weights of the self-expressive layers. It has been shown in [46] that the self-expressiveness coefficient matrix is constrained to satisfy the sparse property by minimizing certain norms of the coefficient matrix. Ideally, the sparse representation of a data point corresponds to the combination of a few points from the same subspace, and the nonzero elements in the self-expressiveness coefficient matrix correspond to the pixels from the same category. The regularization term is used to find the sparse representation of the given points. Since solving the sparse optimization program is in general NP-hard, the L2 norm is generally used as the relaxation [46]. By introducing the self-expressive constraint, the proposed network simultaneously learns the spatial-spectral feature and self-expressive property of the pixels in HSI.

2) Self-expressiveness similarity constraint: HSI clustering aims to divide each pixel into different clusters. In the proposed method, different-scale pixel blocks of a pixel are selected as the input of the proposed network to capture spectral information and spatial correlation effectively. The pixel blocks at different scales share the same center pixel, thus all the self-expressiveness coefficient matrices represent the self-expressive property of the same center pixels. Ideally, the self-expressiveness coefficient matrices of different self-expressive layers should be similar since the inputs of them share the same center pixel. To this end, self-expressiveness similarity constraint L_{ssc} is designed to ensure the similarity of the self-expressiveness coefficient matrices in the multiple self-expressive layers. By minimizing the difference between the weight matrices of the different self-expressive layers, the self-expressiveness similarity constraint is defined as:

$$L_{\mathrm{ssc}} = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} ||\mathbf{C}_{Pb_i} - \mathbf{C}_{Pb_j}||_2^2 \tag{6}$$

By jointly using the self-expressive constraint and self-expressiveness similarity constraint, the multiple self-expressiveness coefficient matrices should agree with each other. Then, the comprehensive self-expressiveness coefficient matrix is computed by $\mathbf{C}_{Pb} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_{Pb_i}$. Specifically, the self-expressiveness coefficient matrix $\mathbf{C}_{Pb}$ stands for how a pixel is represented by a linear combination of other pixels, and the nonzero elements in $\mathbf{C}_{Pb}$ denote the relevance of a pair of pixels in the corresponding row and column. By integrating multiple coefficient matrices of multi-scale spatial-spectral features, more comprehensive spatial-spectral clustering information is acquired from the collaborative constrained self-expressive layers. Based on the comprehensive self-expressiveness coefficient matrix, the similarity matrix can be computed as $|\mathbf{C}_{Pb}| + |\mathbf{C}_{Pb}|^{\mathrm{T}}$. The element of the similarity matrix in i-th row and j-th column denotes the similarity relationship between the i-th pixel and the j-th pixel. Finally, the spectral clustering algorithm is applied on the similarity matrix for obtaining the final clustering result.

D. Implementation details

To jointly learn the spatial-spectral features of the multi-scale pixel blocks and the comprehensive self-expressiveness coefficient matrices, the whole network is trained by using the combination of the reconstruction loss in Eq. (3), the self-expressive constraint in Eq. (5), and the self-expressiveness similarity constraint in Eq. (6). Thus, the overall loss function L_{total} is derived as:

$$\mathrm{L}_{\mathrm{total}} = L_{\mathrm{re}} + \alpha L_{\mathrm{sc}} + \beta L_{\mathrm{ssc}} \tag{7}$$

where α and β denote the hyper-parameters that balance the trade-off among different losses. The overall loss function L_{total} provide multi-constrains for the network training. Specifically, the reconstruction loss is utilized to encourage the encoders to learn meaningful spatial-spectral feature representations in an unsupervised manner. The self-expressive constraint aims to explore the self-expressive property of the pixels in HSI. The self-expressiveness similarity constraint is applied to ensure the similarity of the self-expressiveness coefficient matrices in multiple self-expressive layers and get more comprehensive self-expressive property. By adopting the loss function for collaborative training, the proposed network can learn more discriminative spatial-spectral features and comprehensive self-expressiveness coefficient matrices for clustering.

In the proposed deep spatial-spectral subspace clustering network, each encoder contains three convolutional layers with kernel size of 3×3 and stride size of 2×2. The channels of the three convolutional layers in the encoder network are 16, 32, and 64, respectively. For the collaborative constrained self-expressive layers, a fully-connected layer without bias and activation function is utilized in each stream, the size of which is equal to the number of pixels. Besides, the decoder networks, with the symmetrical structure to encoder networks, map the latent representations derived from self-expressive layers back to the original pixel blocks using deconvolutional layers. For the multi-scale auto-encoder, a two-stream auto-encoder using the pixel blocks with size of 5×5 and 7×7 as its input is adopted. During the training process, the learning rate is set to 1.0×10^{-3}. After the network is trained, the spectral clustering algorithm in [31] is applied based on the self-expressiveness coefficient matrix for the final clustering result.

IV. Experimental Results

A. Experiments settings

1) Datasets: To evaluate the effectiveness of the proposed method, three widely used datasets [1] , including the Indian Pines Dataset, the University of Pavia Dataset, and the Salinas Dataset are utilized in experiments. Indian Pines Dataset is acquired by the AVIRIS sensor. It is at a size of 145×145 and with 220 spectral bands in total. In the experiments, 20 water absorption and noisy bands are removed from the original dataset. University of Pavia Dataset is captured by

[1]http:///www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

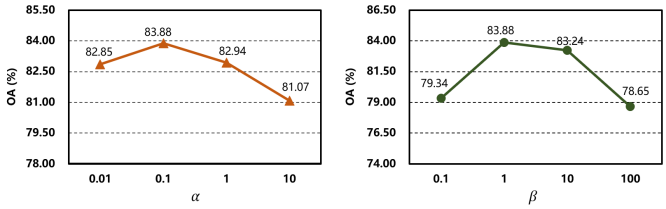

Fig. 2. Parameter analysis of the proposed method.

ROSIS-03 sensor, which is at a size of 610×340 and with 103 spectral bands in total. Captured by the AVIRIS sensor, Salinas Dataset is with the size of 512×217 and 224 spectral bands in total. Among them, 20 water absorption bands are discarded in the experiments. Considering the computational efficiency, the typical subsets of three datasets in the experiments are selected similar to [26]. Specifically, a subset with the size of 85×70 is selected on the Indian Pines Dataset, which includes corn-no-till, grass, soybeans-no-till, and soybeans-minimum-till. For the University of Pavia Dataset, a subset with the size of 200×100 is selected, which includes metal sheet, asphalt, meadows, trees, bare soil, bitumen, bricks, and shadows. For the Salinas Dataset, a subset with the size of 140×150 is selected, which includes vineyard untrained, grapes untrained, fallow smooth, fallow rough plow, stubble, and celery. To be noted, this paper focuses on the problem of unsupervised HSI clustering with no category label information of the pixels, and the whole dataset is used as the input to obtain the clustering result.

2) Metrics: Three metrics, i.e., overall accuracy (OA), average accuracy (AA), and kappa coefficient (KAPPA), are introduced for all the compared methods [47]. The details of the three metrics are given as follows.

OA denotes the proportion of correctly classified pixels in total pixels, which can be calculated as follows:

$$\mathrm{OA} = \sum_{i=1}^{P} \mathbf{T}_{ii} / H \tag{8}$$

where P denotes the number of classes, H denotes the number of pixels, and $\mathbf{T} \in R^{P \times P}$ denotes the confusion matrix of the clustering result. Specifically, the element $\mathbf{T}_{ij}$ represents the number of the pixels from the class j while clustered to class i.

AA represents the mean value of the ratio of correctly clustered pixels and total number of the corresponding class for each class, which is defined as follows:

$$\mathrm{AA} = \frac{\sum_{j=1}^{P} (\mathbf{T}_{jj} / \sum_{i=1}^{P} \mathbf{T}_{ij})}{P} \tag{9}$$

KAPPA describes the degree of agreement between ground truth and clustering labels by computing the measured accuracies.

$$\mathrm{KAPPA} = \frac{H \sum_{i=1}^{P} \mathbf{T}_{ii} - \sum_{i=1}^{P} (\sum_{j=1}^{P} \mathbf{T}_{ij} \sum_{j=1}^{P} \mathbf{T}_{ji})}{H^2 - \sum_{i=1}^{P} (\sum_{j=1}^{P} \mathbf{T}_{ij} \sum_{j=1}^{P} \mathbf{T}_{ji})} \tag{10}$$

All these metrics are capable of measuring the clustering performance, and the higher score indicates the better performance.

3) Comparison methods: To evaluate the proposed method, several clustering methods are considered for comparison, which include K-means [17], CFSFDP [15], SSC [25], S^4C [26], DLSS [24], and LRDSC [35]. Specifically, K-means, CFSFDP, and SSC are clustering methods designed for natural images, S^4C and DLSS are traditional HSI clustering methods, and LRDSC is a state-of-the-art deep clustering method for HSI. It is worth noting that the results of K-means, CFSFDP, SSC, and DLSS are obtained by running the public source code with original parameters; the result of LRDSC is obtained through our own implementation with the parameters used in [35]; and the result of S^4C is obtained from the authors. For a fair comparison and minimization of the influence caused by randomness, all the experiments are executed ten times and the average clustering performance are reported.

4) Parameter analysis: In the proposed method, there are mainly two parameters α and β in the overall loss function Eq. (7). Here, the fluctuations of OA value with the changes of parameters on the Indian Pines Dataset are shown in Fig. 2. In this experiment, β is first set as 1 to get the best α. As shown in Fig. 2, the highest OA value is obtained when α is set as 0.1. Then, α is fixed at 0.1 to find appropriate value of β. From Fig. 2, it can be seen that the best result is obtained when β is set as 1. Thus, the α and β are set as 0.1 and 1 in the experiments to get a satisfactory performance.

B. Comparison with state-of-the-art methods

1) Indian Pines Dataset: To validate the effectiveness of the proposed DS^3C-Net, comparison experiment is first conducted on the Indian Pines Dataset. The quantitative evaluation and visual clustering results are reported in Table I and Fig. 3. It can be seen that CFSFDP obtains an OA of only 59.48%, with a large number of misclustering and noise in the corresponding visual clustering result in Fig. 2(b). Although the other traditional clustering methods, such as S^4C and K-means, achieve better clustering performance, the overall accuracy and visual clustering results are still not satisfactory. By comparison, the deep clustering method LRDSC obtains relatively good clustering performance. More importantly, the proposed DS^3C-Net achieves the best performance. Comparing with LRDSC, the proposed DS^3C-Net obtains 9.78% OA increase, 10.62% AA increase, and 12.12% KAPPA increase. As can be seen from the visualization results, little noise exits in the center green area of the "Grass" class in Fig. 2(g), and less pixels of the "Soybeans-no-till" class are mis-clustered comparing with the other methods. It is mainly because that the proposed method not only extracts the spatial-spectral

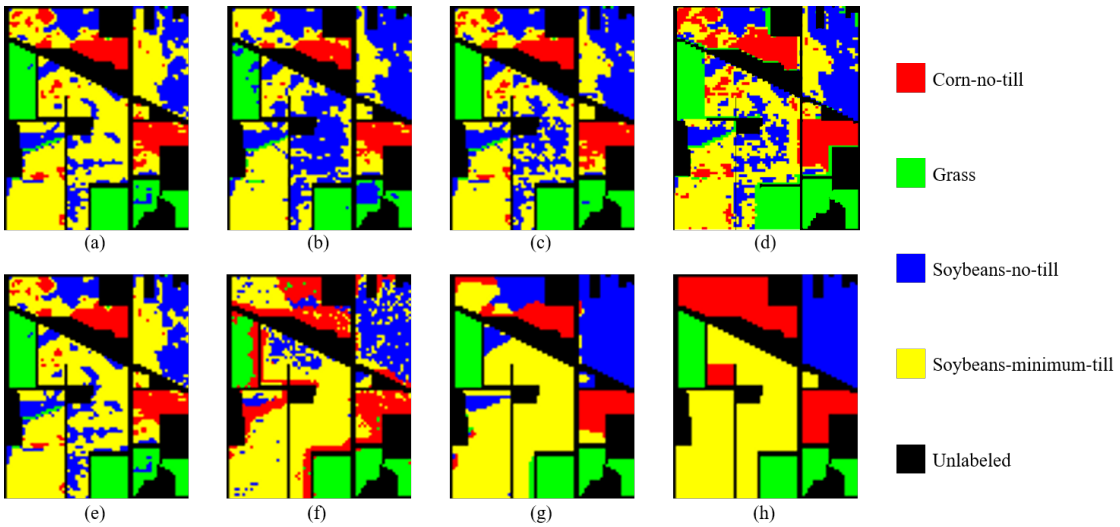

Fig. 3. Visual comparisons on the Indian Pines Dataset. (a) K-means. (b) CFSFDP. (c) SSC. (d) S^4C. (e) DLSS. (f) LRDSC. (g) DS^3C-Net (Ours). (h) Ground Truth.

TABLE I
CLUSTERING ACCURACY OF EACH CATEGORY AND THE THREE METRICS (OA, AA, AND KAPPA) (%) OF DIFFERENT CLUSTERING ALGORITHMS ON THE INDIAN PINES DATASET.

class	Method						
	K-means	CFSFDP	SSC	S^4C	DLSS	LRDSC	DS^3C-Net (Ours)
Corn-no-till	46.67	39.90	49.35	61.00	44.18	59.70	51.84
Grass	97.35	86.59	99.58	100.00	97.63	88.83	100.00
Soybeans-no-till	46.29	68.02	66.94	65.30	49.80	70.31	97.84
Soybeans-minimum-till	77.30	56.28	64.10	65.28	75.08	77.67	89.33
OA	68.25	59.48	67.01	70.08	67.36	74.10	**83.88**
AA	66.90	62.70	69.99	72.90	66.67	74.13	**84.75**
Kappa	59.22	51.34	59.88	58.25	58.33	67.77	**79.89**

features but also explores more comprehensive information via the multi-scale pixel blocks.

2) University of Pavia Dataset: The second experiment is conducted on the University of Pavia Dataset. The quantitative evaluation and visual clustering results are shown in Table II and Fig. 4. It can be seen that the experimental results are generally consistent with the first experiment. The clustering methods for traditional clustering methods perform worse than the deep clustering methods, since the discriminative spatial-spectral features can be explored by the deep network. Particularly, since different characteristics of different classes of spectrum, the relatively better clustering accuracy on different classes are achieved with different methods. It should be noted that the clustering performance of the proposed method for some class is zero. The reason might be that pixels of these classes show a high similarity with other classes. For example, the "bitumen" class is similar to the "metal sheet" class. As a result, all pixels of the "bitumen" class are mis-clustered to the "metal sheet" class. However, in terms of overall accuracy, the proposed method achieves generally better performance than the other methods. It can also be seen from Fig. 4 that, the smooth clustering results with least noise are also obtained by the proposed method. Specifically, the proposed method achieves a clustering map with least noise on the "bare soil", "bricks", and "asphalt" classes in Fig. 4 (g).

3) Salinas Dataset: The third experiment is conducted on the Salinas Dataset. The quantitative evaluation and visual clustering results are reported in Table III and Fig. 5. It can be seen that SSC achieves poor clustering accuracy with a low degree of overall accuracy. By comparison, the other methods such as LRDSC and S^4C achieve better clustering results, where most of the classes can be successfully distinguished. It is mainly because that the distribution of pixels on the Salinas Dataset is comparatively regular. Since the pixels belonging to the same category are distributed in the same concentrated area, most methods have obtained relatively better performance, thus our method improves relatively little. However, by comparison, the proposed method achieves the best clustering result with the OA of 86.98%. It can also be seen from the corresponding visual results, few errors occur at the edge of "Celery" and "Stubble" classes of the proposed method. It is mainly because that the spatial-spectral features extracted in the proposed method is more beneficial for distinguishing the pixels at the edge, which further verifies the effectiveness of the proposed method for HSI clustering.

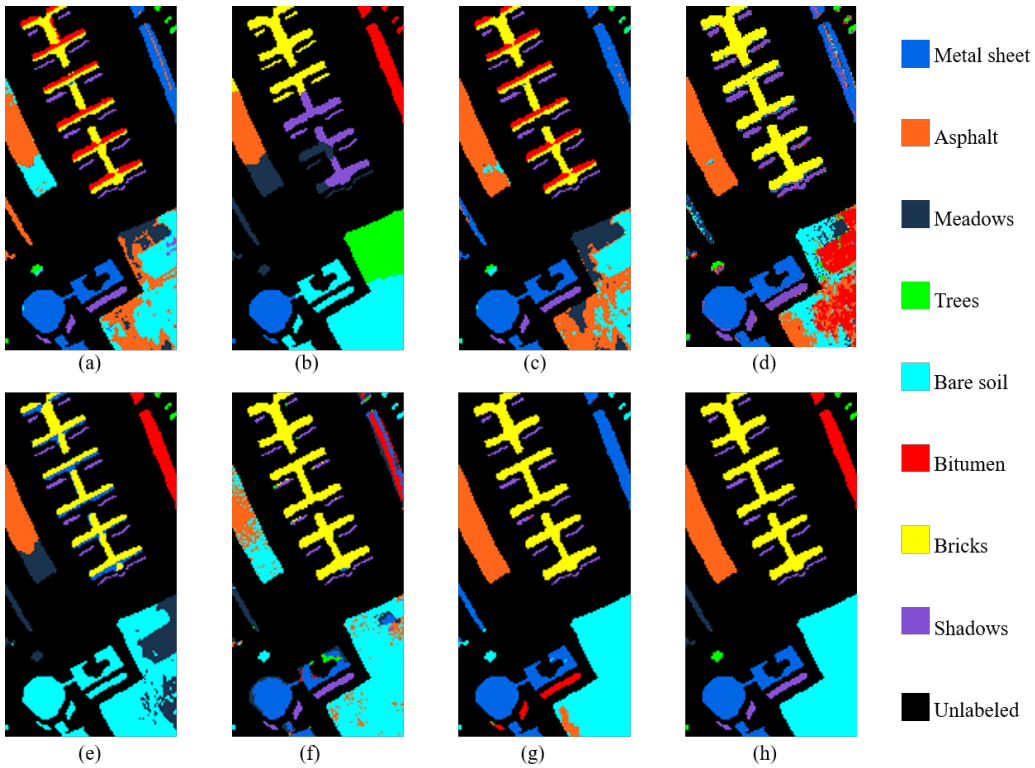

Fig. 4. Visual comparisons on the University of Pavia Dataset. (a) K-means. (b) CFSFDP. (c) SSC. (d) S^4C. (e) DLSS. (f) LRDSC. (g) DS^3C-Net (Ours). (h) Ground Truth.

TABLE II
CLUSTERING ACCURACY OF EACH CATEGORY AND THE THREE METRICS (OA, AA, AND KAPPA) (%) OF DIFFERENT CLUSTERING ALGORITHMS ON THE UNIVERSITY OF PAVIA DATASET.

class	Method						
	K-means	CFSFDP	SSC	S^4C	DLSS	LRDSC	DS^3C-Net (Ours)
Metal sheet	99.64	56.52	99.76	99.09	0.00	77.84	99.53
Asphalt	62.61	60.99	95.40	87.30	65.22	46.58	100.00
Meadows	0.00	100.00	2.80	60.64	99.07	87.85	0.00
Trees	83.82	0.00	48.53	98.61	45.59	0.00	0.00
Bare soil	50.92	58.47	32.64	31.93	70.23	89.42	96.10
Bitumen	0.00	100.00	0.00	0.00	100.00	48.91	0.24
Bricks	57.81	47.02	60.00	98.37	73.43	99.40	100.00
Shadows	100.00	15.24	100.00	99.09	59.56	93.63	58.73
OA	59.27	56.46	56.55	65.09	62.50	81.17	**86.87**
AA	56.85	54.78	54.89	**71.88**	64.14	67.95	56.82
Kappa	59.13	56.23	56.41	58.52	62.42	81.11	**86.85**

C. *Validity analysis of self-expressiveness similarity constraint*

In this paper, the self-expressiveness similarity constraint is designed to constrain the learning of the self-expressive layers in the proposed multi-scale network. In order to validate the effectiveness of the self-expressiveness similarity constraint, the experiments of the proposed method with and without self-expressiveness similarity constraint are conducted on the three datasets. The overall accuracy of the clustering results are shown in Table IV. It can be seen that the clustering performance of the method with self-expressiveness similarity constraint is better than that without self-expressiveness similarity constraint. Specifically, the proposed method with the self-expressiveness similarity constraint obtains the 0.92%, 0.74% and 0.41% overall accuracy increase on Indian Pines, University of Pavia and Salinas Dataset respectively. Since the pixel blocks at different scales describe the spatial-spectral information of the same center pixels, the self-representation coefficient matrix at each scale should be as similar as possible. By utilizing the self-expressiveness similarity constraint, the self-representation coefficient matrices of multiple scales are further constrained to ensure the consistency, so that the comprehensive coefficient matrix is obtained to boost the HSI clustering performance.

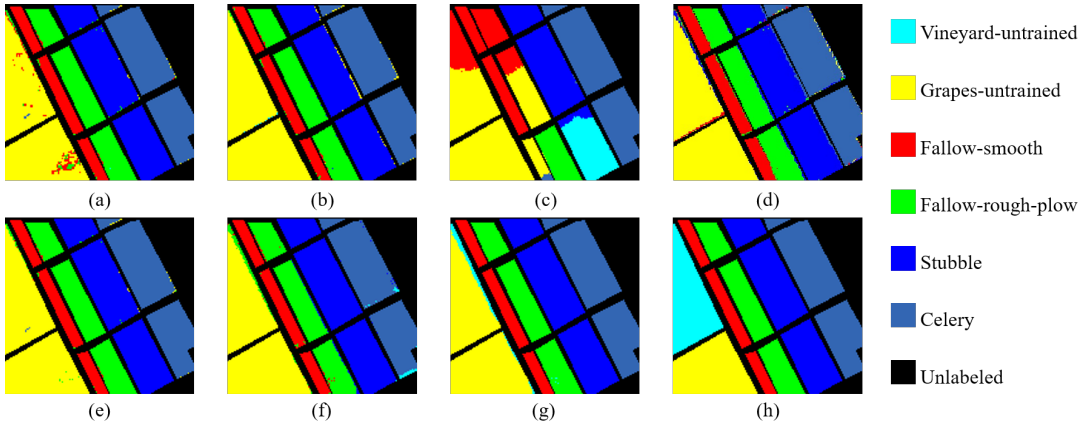

Fig. 5. Visual comparisons on the Salinas Dataset. (a) K-means. (b) CFSFDP. (c) SSC. (d) S^4C. (e) DLSS. (f) LRDSC. (g) DS^3C-Net (Ours). (h) Ground Truth.

TABLE III
CLUSTERING ACCURACY OF EACH CATEGORY AND THE THREE METRICS (OA, AA, AND KAPPA) (%) OF DIFFERENT CLUSTERING ALGORITHMS ON THE SALINAS DATASET.

class	Method						
	K-means	CFSFDP	SSC	S^4C	DLSS	LRDSC	DS^3C-Net (Ours)
Vineyard-untrained	0.00	0.46	0.00	0.00	0.00	0.00	8.79
Grapes-untrained	93.57	100.00	100.00	100.00	99.69	99.46	98.43
Fallow-smooth	99.10	98.29	74.61	99.26	99.35	99.19	100.00
Fallow-rough-plow	99.22	98.65	33.18	99.59	99.30	95.58	99.71
Stubble	99.47	98.66	65.71	99.77	99.70	99.97	100.00
Celery	99.52	99.26	100.00	99.84	99.46	98.04	100.00
OA	84.67	85.23	64.81	86.31	85.64	84.74	**86.98**
AA	81.81	82.55	62.25	83.07	82.92	82.04	**84.49**
Kappa	84.67	85.23	64.38	83.12	85.62	84.73	**86.96**

TABLE IV
EVALUATION OF THE SELF-EXPRESSIVENESS SIMILARITY CONSTRAINT (%).

Datasets	Indian Pines	University of Pavia	Salinas
without self-expressiveness similarity constraint	82.96	86.13	86.46
with self-expressiveness similarity constraint	**83.88**	**86.87**	**86.98**

D. Analysis of the network architecture

1) Analysis of the size of multi-stream pixel blocks: In this paper, the joint spatial-spectral features of HSI are learned by using the proposed multi-scale network, in which the pixel blocks at different scales are selected as the input of the network. Intuitively, the suitable pixel blocks which contains more useful information and less redundant information are conductive to explore more effective spatial information of the center pixel. In order to verify the influence of the sizes of the multi-stream pixel blocks, the experiments of the proposed network taking the pixel blocks of different sizes as inputs are conducted on the Indian Pines Dataset. The overall accuracy of the clustering results are shown in Table V, in which "1×1+3×3" denotes the center pixel and a neighborhood within the 3×3 block are selected as the inputs. Similarly, "3×3+5×5" denotes the neighborhoods within the 3×3 and 5×5 blocks are selected as the inputs, and so on. As shown in Table V, the first three lines with "1×1" center pixel as one of the inputs have the relatively poor performance, which verifies that only one pixel block cannot providing enough spatial information for the center pixel. Besides, it can be seen that the clustering performance increases with the size of the input blocks since the discriminative spatial-spectral features are extracted. Comparing the case of "1×1+3×3", the case of "3×3+5×5" obtains 2.78% overall accuracy increase, and the case of "5×5+7×7" obtains 5.6% overall accuracy increase. Although relatively better clustering performance can be obtained with the blocks of larger size, computation and memory for feature extraction and storage will increase sharply. Therefore, jointly considering the computing complexity and clustering performance, pixel blocks with the size of 5×5 and 7×7 are used in this paper.

2) Analysis of the number of multiple streams: In order to verify the influence of the number of multiple streams, the networks with different number of multiple streams are verified on the Indian Pines Dataset. For fair comparison between the single-stream and multi-stream network architecture, the self-expressiveness similarity constraint is not used in the

TABLE V
ANALYSIS OF THE SIZE OF MULTI-STREAM PIXEL BLOCKS.

Size of multi-stream pixel blocks	OA (%)
1×1 + 3×3	78.28
1×1 + 5×5	79.84
1×1 + 7×7	79.39
3×3 + 5×5	81.06
3×3 + 7×7	82.30
5×5 + 7×7	**83.88**

TABLE VI
ANALYSIS OF THE NUMBER OF MULTIPLE STREAMS.

Number of multiple streams	OA (%)
5×5	82.03
7×7	82.34
5×5+7×7	**82.96**
3×3+5×5+7×7	81.28

multi-stream network architecture. The comparison results are shown in Table VI, in which "5×5" and "7×7" indicates the single-stream network with the pixel block at the size of 5×5 and 7×7 respectively. As shown in the table, the clustering accuracy of the network with "7×7" is slightly better than that with "5×5". The two-stream network obtains the clustering performance of 82.96% overall accuracy, which outperforms the single-stream networks. It is mainly because that the spatial information with different characteristics are integrated via the multi-scale auto-encoder to obtain more comprehensive features. In addition, the clustering accuracy of the three-stream is lower than the single-stream and two-stream networks. With the increase of the number of streams, the multi-stream network becomes more and more complex. This brings great difficulties to the network learning, thus exerting negative impacts on the clustering performance. The experimental results show that choosing the pixel blocks with size of 5×5 and 7×7 as inputs is reasonable.

E. Analysis of computational complexity

In order to investigate the computational complexity of the proposed method, we compare the running time of different clustering methods on Indian Pines Dataset on a PC with a GPU of 48-Gb memory and a CPU @ 3.3GHz. The traditional clustering methods take relatively less running time to obtain the clustering results, i.e., K-means takes 0.55s, CFSFDP takes 1.10s, SSC takes 1.77s, and DLSS takes 6.98s. By comparison, the deep clustering method LRDSC takes 562s, which costs much more time than the traditional clustering methods. Comparing with LRDSC, the proposed method takes 545s, which achieves better clustering performance with the same level of computational complexity.

V. CONCLUSION

In this paper, we proposed a novel DS^{3}C-Net, which explores spatial-spectral information by designing a multi-scale auto-encoder network architecture. To extract the multi-scale spatial-spectral features, different-scale pixel blocks around the center pixel are selected as the inputs of the proposed network. Besides, the collaborative constrained self-expressive layers are introduced between the encoder and decoder to capture the self-expressive subspace structures. By designing a self-expressiveness similarity constraint, the proposed network is trained collaboratively so that the affinity matrixes of the feature representation are obtained. Finally, the spectral clustering algorithm is applied based on the affinity matrixes to obtain the HSI clustering result. Experimental results on three widely used datasets demonstrate the effectiveness of the proposed method.

In the future, we will focus on combining the proposed method with semi-supervised learning to exploit information from labled and unlabled data on HSI classification. Besides, further study may also include extending our method to self-supervised learning HSI classification.

REFERENCES

[1] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Trans. Circuits Syst. Video Technol.*, DOI: 10.1109/TCSVT.2020.2975566, pp. 1-14, 2020.

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

[3] X. Jin and Y. Gu, "Superpixel-based intrinsic image decomposition of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4285-4295, Aug. 2017.

[4] J. Lei, X. Luo, L. Fang, M. Wang, and Y. Gu, "Region-enhanced convolutional neural network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, p. 5693-5702, Aug. 2020.

[5] X. Jin, Y. Gu, and T. Liu, "Intrinsic image recovery from remote sensing hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 224-238, Jan. 2019.

[6] L. Sun, C. Ma, Y. Chen, Y. Zheng, H. J. Shim, Z. Wu, and B. Jeon, "Low rank component induced spatial-spectral kernel method for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, DOI: 10.1109/TCSVT.2019.2946723, pp. 1-14, 2019.

[7] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. and Remote Sens.*, vol. 57, no. 9, pp. 6916-6928, Sep. 2019.

[8] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral computational imaging via collaborative tucker3 tensor decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, DOI: 10.1109/TCSVT.2020.2975936, pp. 1-16, 2020.

[9] M. Han, R. Cong, X. Li, H. Fu, and J. Lei, "Joint spatial-spectral hyperspectral image classification based on convolutional neural network," *Pattern Recognit. Lett.*, vol. 30, pp. 38-45, 2018.

[10] W. Liu, X. Shen, B. Du, I. W. Tsang, W. Zhang, and X. Lin, "Hyperspectral imagery classification via stochastic HHSVMs," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 577-588, Feb. 2019.

[11] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690-6709, Sep. 2019.

[12] L. Fang, G. Liu, S. Li, P. Ghamisi, and J. A. Benediktsson, "Hyperspectral image classification with squeeze multibias network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1291-1301, Mar. 2019.

[13] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623-2634, Jun. 2018.

[14] A. Paoli, F. Melgani, and E. Pasolli, "Clustering of hyperspectral images based on multiobjective particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 12, pp. 4175-4188, Dec. 2009.

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

[16] B. Peng, J. Lei, H. Fu, C. Zhang, T.-S. Chua, and X. Li, "Unsupervised video action clustering via motion-scene interaction constraint," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 131-141, Jan. 2020.
[17] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100-108, Jan. 1979.
[18] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *New York, NY, USA: Springer-Verlag*, 2013.
[19] M. Maggioni and J. M. Murphy. "Learning by unsupervised nonlinear diffusion," *J. Mach. Learn. Research*, vol. 20, no. 160, pp. 1-56, Jan. 2019.
[20] 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, Aug. 2011.
[21] Y. Zhong, S. Zhang, and L. Zhang, "Automatic fuzzy clustering based on adaptive multi-objective differential evolution for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 5, pp. 2290-2301, Oct. 2013.
[22] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 4, pp. 1907-1916, Aug. 2004.
[23] J. Lin, C. He, Z. J. Wang, and S. Li, "Structure preserving transfer learning for unsupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1656-1660, Oct. 2017.
[24] J. M. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1829-1845, Mar. 2019.
[25] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765-2781, Nov. 2013.
[26] 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, Jun. 2016.
[27] 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, Jan. 2017.
[28] L. Tian, Q. Du, I. Kopriva, and N. Younan, "Kernel spatial-spectral based multi-view low-rank sparse sbuspace clustering for hyperspectral imagery," *Workshop Hyperspectral Image Signal Process.: Evolution Remote Sensi.*, Sep. 2018.
[29] Y. Long, X. Deng, G. Zhong, J. Fan, and F. Liu, "Gaussian kernel dynamic similarity matrix based sparse subspace clustering for hyperspectral images," *Int. Conf. Computational Intell. Security*, pp. 211-215, Dec. 2019.
[30] J. M. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, pp. 478-487, 2016.
[31] J. Pan, Z. Tong, H. Li, M. Salzmann, and I. Reid, "Deep subspace clustering networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, pp. 24-33, 2017.
[32] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 5880-5888, 2017.
[33] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "ClusterGAN: Latent space clustering in generative adversarial networks," in *Proc. 33rd AAAI Conf. Artificial Intell.*, pp. 4610-4617, 2019.
[34] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2172-2180, 2016.
[35] M. Zeng, Y. Cai, X. Liu, Z. Cai, and X. Li, "Spectral-spatial clustering of hyperspectral image based on Laplacian regularized deep subspace clustering," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2694-2697, 2019.
[36] J. M. Murphy and M. Maggioni, "Spectral-spatial diffusion geometry for hyperspectral image clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1243-1247, Jul. 2020.
[37] H. Zhang, Q. Wang, W. Shi, and H. Ming, "A novel adaptive fuzzy local information c-means clustering algorithm for remotely sensed imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5057-5068, Sep. 2017.
[38] F. Morsier, M. Borgeaud, V. Gass, J.-P. Thiran, and Devis Tuia, "Kernel low-rank and sparse graph for unsupervised and semi-supervised classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3410-3420, Jun. 2016.
[39] 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, Jan. 2019.
[40] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723-1740, Mar. 2018.
[41] W. Sun, J. Peng, G. Yang, and Q. Du, "Fast and latent low-rank subspace clustering for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3906-3915, Jun. 2020.
[42] J. Xu, N. Huang, and L. Xiao, "Spectral-spatial subspace clustering for hyperspectral images via modulated low-rank representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 3202-3205, Jul. 2017.
[43] A. Li, Q. A. Qin, Z. Shang, and Y. Tang, "Spectral-spatial sparse subspace clustering based on three-dimensional edge-preserving filtering for hyperspectral image," *Int. J. Pattern Recognit. Artificial Intell.*, vol. 33, no. 3, Mar. 2019.
[44] S. Mahdi and E. J. Candes, "A geometric analysis of subspace clustering with outliers," *Annals of Statistics*, vol. 40, no. 4, pp. 2195-2238, 2012.
[45] E. Elhamifar and R. Vidal, "Sparse subspace clustering," *IEEE Conf. Computer Vis. Pattern Recognit.*, 2009.
[46] P. Ji, M. Salzmann, and H. Li, "Efficient dense subspace clustering," *IEEE Winter Conf. Applications Computer Vis.*, pp. 461-468, 2014.
[47] S. Li, Q. Hao, G. Gao, and X. Kang, "The effect of ground truth on performance evaluation of hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7195-7206, Dec. 2018.

Jianjun Lei (M'11-SM'17) received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2007. He was a visiting researcher at the Department of Electrical Engineering, University of Washington, Seattle, WA, from August 2012 to August 2013. He is currently a Professor at Tianjin University, Tianjin, China. His research interests include 3D video processing, virtual reality, and artificial intelligence.

Xinyu Li received the B.S. degree at the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, in 2018. Currently, he is pursuing the M.S. degree at the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include computer vision, image processing and hyperspectral image clustering.

Bo Peng received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2020. She was a visiting research scholar at the School of Computing, University of Singapore, Singapore, from March 2019 to April 2020. She is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her research interests include computer vision, image processing, and vision understanding.

Leyuan Fang (S'10-M'14-SM'17) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015. From September 2011 to September 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council, where he was a Post-Doctoral Researcher with the Department of Biomedical Engineering, from August 2016 to September 2017. He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing. He is the associate editors of IEEE Trans. Image Processing, IEEE Trans. Geoscience and Remote Sensing, and Neurocomputing. Dr. Fang was a recipient of the 2nd-Grade National Award at the Nature and Science Progress of China in 2019.

Nam Ling (S'88–M'90–SM'99–F'08) received the B.Eng. degree from the National University of Singapore, Singapore, in 1981, and the M.S. and Ph.D. degrees from the University of Louisiana at Lafayette, Lafayette, LA, USA, in 1985 and 1989, respectively. From 2002 to 2010, he was an Associate Dean with the School of Engineering, Santa Clara University, Santa Clara, CA, USA. He was the Sanfilippo Family Chair Professor, and is currently the Wilmot J. Nicholson Family Chair Professor and the Chair with the Department of Computer Science and Engineering, Santa Clara University. He is/was also a Consulting Professor with the National University of Singapore, a Guest Professor with Tianjin University, Tianjin, China, a Guest Professor with Shanghai Jiao Tong University, Shanghai, China, a Cuiying Chair Professor with Lanzhou University, Lanzhou, China, a Chair Professor and Minjiang Scholar with Fuzhou University, Fuzhou, China, a Distinguished Professor with the Xi'an University of Posts and Telecommunications, Xi'an, China, a Guest Professor with the Zhongyuan University of Technology, Zhengzhou, China, and an Outstanding Overseas Scholar with the Shanghai University of Electric Power, Shanghai, China. He has authored or coauthored over 220 publications and seven adopted standard contributions. He has been granted nearly 20 U.S. patents so far. Dr. Ling is an IEEE Fellow due to his contributions to video coding algorithms and architectures. He is also an IET Fellow. He was named as an IEEE Distinguished Lecturer twice and was also an APSIPA Distinguished Lecturer. He was a recipient of the IEEE ICCE Best Paper Award (First Place) and the Umedia Best/Excellent Paper Award three times, six awards from Santa Clara University, four at the University level (Outstanding Achievement, Recent Achievement in Scholarship, President's Recognition, and Sustained Excellence in Scholarship), and two at the School/College level (Researcher of the Year and Teaching Excellence). He was a Keynote Speaker for IEEE APCCAS, VCVP (twice), JCPC, IEEE ICAST, IEEE ICIEA, IET FC Umedia, IEEE Umedia, IEEE ICCIT, and Workshop at XUPT (twice), and a Distinguished Speaker for IEEE ICIEA. He has served as a General Chair/Co-Chair for IEEE Hot Chips, VCVP (twice), IEEE ICME, Umedia (seven times), and IEEE SiPS. He was an Honorary Co-Chair for IEEE Umedia 2017. He has also served as a Technical Program Co-Chair for IEEE ISCAS, APSIPA ASC, IEEE APCCAS, IEEE SiPS (twice), DCV, and IEEE VCIP. He was a Technical Committee Chair for IEEE CASCOM TC and IEEE TCMM, and has served as a Guest Editor or an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I:REGULAR PAPERS, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE ACCESS, Springer JSPS, and Springer MSSP.

Qingming Huang (SM'08-F'18) is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia video analysis, image processing, computer vision and pattern recognition He has published more than 400 academic papers in prestigious international journals including IEEE Trans. Image Process., IEEE Trans. Multimedia, IEEE Trans. Circuits Syst. Video Technol., etc, and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, VLDB, etc. He is the associate editor of IEEE Trans. Circuits Syst. Video Technol., and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. Multimedia, IEEE Trans. Circuits Syst. Video Technol., IEEE Trans. Image Process., etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, PSIVT, etc.