Spectral-Spatial Graph Reasoning Network for Hyperspectral Image Classification

Di Wang, Student Member, IEEE, Bo Du, Senior Member, IEEE, Liangpei Zhang, Fellow, IEEE

Abstract—In this paper, we propose a spectral-spatial graph reasoning network (SSGRN) for hyperspectral image (HSI) classification. Concretely, this network contains two parts that separately named spatial graph reasoning subnetwork (SAGRN) and spectral graph reasoning subnetwork (SEGRN) to capture the spatial and spectral graph contexts, respectively. Different from the previous approaches implementing superpixel segmentation on the original image or attempting to obtain the category features under the guide of label image, we perform the superpixel segmentation on intermediate features of the network to adaptively produce the homogeneous regions to get the effective descriptors. Then, we adopt a similar idea in spectral part that reasonably aggregating the channels to generate spectral descriptors for spectral graph contexts capturing. All graph reasoning procedures in SAGRN and SEGRN are achieved through graph convolution. To guarantee the global perception ability of the proposed methods, all adjacent matrices in graph reasoning are obtained with the help of non-local self-attention mechanism. At last, by combining the extracted spatial and spectral graph contexts, we obtain the SSGRN to achieve a high accuracy classification. Extensive quantitative and qualitative experiments on three public HSI benchmarks demonstrate the competitiveness of the proposed methods compared with other state-of-the-art approaches.

Index Terms—Adaptively, non-local self-attention mechanism, spectral-spatial graph reasoning network (SSGRN), hyperspectral image (HSI) classification

I. INTRODUCTION

Relying on the excellent characteristics that the intrinsic properties of the targets can be identified by automatically extracting the effective features in an end-to-end manner. The deep learning technologies are being extensively employed in processing of the hyperspectral image (HSI). The HSI includes abundant spectral information which is carried by hundred of bands and vision representations presented by high spatial resolution, effectively serving the precision agriculture [1], environmental monitoring [2], anomaly detection [3] and so on. Among these fields, the land cover classification is always a foundamental and hot topic, where each pixel in the whole scene is needed to be assigned a certain and unique semantic category.

In early HSI classification community, the deep features are usually obtained with the networks of fully connected structure [4]–[6]. For example, Chen et al. [4] directly uses the stacked autoencoder (SAE) to extract the spectral and spatial features, respectively. They adopt the layer-wise stacked pre-training strategy to ensure the stability of network learning and progressively deepen the networks, while Li et al. [5] employs another deep neural network (DNN) called belief network based on the restricted Boltzmann machine to classify the HSI. However, these networks require sufficient computations since each neuron needs to connect with all the units in next layer, demanding to optimize a large amount of parameters, and inevitably generates the overfitting phenomenon. In addition, the input form is strictly restricted that only the 1-D data can be fed into the networks. Thus, in order to obtain the spatial information, Chen et al. has to flatten the image patches to fit the SAE, and the spatial structures are unavoidably destroyed, affecting the classification accuracy.

To tackle these problems, in recent years, many researchers begin to interpret the HSI for pixel-level classification using convolutional neural networks (CNNs), where the fixed size convolution kernels slide on the whole image and the corresponding parameters are constantly updated according to the global statistical characteristics. Thus, compared with DNNs receiving 1-D spectral vectors, the CNNs that process the clipped spatial patches around the target pixel have larger vision fields since they can perceive the neighborhood information of the target pixel and reduce the overfitting because there are fewer parameters needed to be learned. In the past years, regarding the abundant spectral channels in HSI, the mainstream deep learning-based HSI parsing methods are all on the foundation of classification networks, where the output is a group of category probability vector and the input is mainly spatial patches or spectral vectors of the target pixel. For example, Hu et al. [7] extracts the spectral features with a five layer 1-D CNN. In [8], the 2-D CNN is utilized to obtain spatial features, while the spectral-spatial joint features are acquired in [9]–[12]. In addition to being directly used as an extractor for single scale feature, the architectures of CNN can be modified to generate the enhanced features for further accuracy improving [13]–[23].

However, as the networks deepen, their abilities are soon saturated since the limited size of input patches in spite of possess an extraordinarily large theoretical receptive field. To this end, many fully convolutional network (FCN) based deep learning methods are recently developed, where the whole image can be directly input into the network and simultaneously classify all pixels. With the FCN, it’s naturally consider stacking the convolutional layers to increase the receptive field [24]–[27]. Specifically, in [24], the dilated convolution is used in spatial part to expand the vision field. However, the ordinary
and dilated convolution are all local operators, meaning it is still difficult to model the dependence between a specific pixel and other long-range positions, thus the networks can’t fully leverage the contexts.

In order to capture the non-local contexts, [28] firstly introduces the self-attention [29] mechanism into computer vision field. They compute the similarity between spatial points in a projected latent semantic space and aggregate these points by weighted addition. Then, [30] and [31] perform the scene parsing of natural image by directly regarding the pixel as the spatial points, where each pixel has a global view of the image. However, these networks inevitably have high computational complexity since each pixel need to compute the similarity with all the other pixels, requiring \(O(N^2)\) calculations, where \(N = HW\) is the pixel number, \(H\) and \(W\) are height and width. To shrink the computations, [32] only captures the contexts in same row or column, decreasing the complexity to \(O(N\sqrt{N})\) since usually \(H = W = \sqrt{N}\), while the complexity in [33] is \(O(NP)\), where \(P\) is the remaining pixel number after conducting pyramid pooling [34]. But the straightforward pyramid pooling neglects the intrinsic information in image, so [35] and [36] obtain the district features through combining the contexts in a same category under the guide of ground truth. However, they have to use the coarse segmentation mask when inferring without the truth label, bringing the unnecessary errors. Different from these approaches, the descriptors in our methods are the gathering centers obtained by separately aggregating the features in the homogeneous areas. Concretely, these areas are generated by superpixel segmentation that based on the pixel spectral-spatial similarities [37], [38]. Thus, the proposed methods not only require a lower complexity compared with the original non-local module, but produces the more effective descriptors since the homogeneous regions can be adaptively obtained whether in training or testing.

It should be noticed that the input and output of HSI classification task are similar with the semantic segmentation of natural image, meaning some patterns in natural image segmentation field may still play a role for HSI classification. However, different from the rich and complex natural image database which has certain channels, HSI scene usually involves only a single image with hundreds of bands that are determined with the sensor type. Thus, one of the key parts in HSI processing can be distinguished from the natural image operation is to explore how to better exploit the rich spectral information. In fact, the adjacent bands in HSI contain contextual information [14], [15]. Thus, we adopt a similar idea that reasonably aggregating the channels to generate spectral descriptors for better spectral contexts capturing.

The conventional CNN only aggregate the contexts in a regular region. However, the objects in images usually have an irregular distribution. Therefore, it is unsuitable to place these spatial descriptors in a form of regular grid. It is more appropriate to solve this problem from graph perspective where each descriptor is considered as a graph node. Then the graph contexts can be well extracted by conducting graph convolutional network (GCN) [39], which is first implemented in a transductive pattern where each pixel is regarded as a node and the whole image need to be fed into the network, requiring a large amount of computing resource. There are two solutions to address this problem, one is inductive learning, which is attempt to achieve this procedure using mini-batch training strategy [40], while another is with the representative descriptors because of the redundant information lying in adjacent pixels. For example, [41] generates the descriptors by directly segmenting the original image using superpixel and obtains the multiscale features through combining the node contexts in different neighborhoods, corresponding to different adjacent matrices, and performs the superpixel-level labeling. [42] further optimizes the adjacent matrices producing procedure with the generalized mahalanobis distance and constructs the projection matrix to implement the mapping of pixel to region and region to pixel. Different from above approaches, we generate the descriptors based on the intermediate features in network, thus our methods can adaptively produce the more flexible homogeneous areas for the more powerful descriptors obtaining. In addition, we conduct the graph reasoning based on the similarities between the descriptors in a global view to acquire the more effective graph contexts. This can be implemented with the help of self-attention mechanism. What’s more, besides the spatial graph reasoning, we also implement the graph convolution on the spectral descriptors to capture the relationships of different spectral channels. To our best knowledge, it is the first time to put the superpixel segmentation procedure into network for training in an end-to-end manner and implement graph reasoning between spectral bands for HSI classification.

We call our method as spectral-spatial graph reasoning network (SSGRN), which includes the spatial graph reasoning subnetwork (SAGRN) and spectral graph reasoning subnetwork (SEGRN), the main contributions of this paper can be summarized as follows:

1) We propose an end-to-end spectral-spatial graph reasoning network named SSGRN. Compared with existing spectral-spatial joint networks, our model can adaptively capture the inter-relationship of the representative descriptors and fully exploit the graph contexts.

2) To generate the effective descriptors, we design the spatial subnetwork SAGRN, where the homogeneous areas are adaptively generated by involving the superpixel segmentation procedure into network and training in an end-to-end manner.

3) A spectral subnetwork SEGRN is proposed to capture the contextual information lying in different bands using graph reasoning. As far as we know, it is the first time the relationships of spectral channels are explored from graph perspective for HSI classification.

4) Benefiting from the proposed graph reasoning modules. Our networks yield significant improvement in three HSI classification benchmarks compared with other GCN-based methods.

The remainder of this paper is organized as follows. Section II gives an introduction of related works. Section III describes the proposed networks. Experiments and related comprehensive analyses are presented in section IV. Finally, Section V concludes this paper.
II. RELATED WORK

A. Semantic Segmentation for Natural Image

Since the fully convolutional networks are used by FCN [43] for semantic segmentation into natural scene, the FCN-based methods keep making progress. UNet [44], SegNet [45], DeconvNet [46], RefineNet [47] and DFNet [48] adopt encoder-decoder architectures to carefully recover the details in upsampling procedure. To perceive the global information, GCN [49] uses the larger kernel, BiSeNet [50] adopts the global pooling, while an effective encoding layer is introduced in EncNet [51] on top of the network. There are also some networks extracting the features from the perspective of categories [35], [36]. In the aspect of multiscale feature generating, multiple atrous convolution filters are employed in DeeplabV3 [52] to construct the ASPP module to regularly capture the objects in different range of the target pixel. DenseASPP [53] further increases the receptive field with the dense connection between channels.

B. Graph Convolutional Network

According to the convolution theorem, for functions \( f, g \)
\[
F(f \ast g) = F(f) \cdot F(g)
\] (1)

Thus, the convolution operation can be rewritten as
\[
f \ast g = F^{-1}(F(f) \cdot F(g))
\] (2)

where \( F \) and \( F^{-1} \) are fourier transform (FT) and inverse fourier transform (IFT), meaning the graph convolution \( G \) of filter \( g \) and node \( x \) in graph \( G \) is the multiplication in transformed fourier domain
\[
G(g \ast x) = F^{-1}(F(g) \cdot F(x))
\] (3)

In fact, the basis of FT is the eigenvectors that make up the orthogonal spectral matrix (spectral matrix) which is produced by implementing the spectral decomposition (SD) \( D \) on the laplacian matrix \( L_G \) of graph \( G \), which is defined as
\[
L_G = D^G - A^G
\] (4)

where \( D^G \) and \( A^G \) are the degree matrix and adjacency matrix, \( D_{ii}^G = \sum_{j=1}^{K} A_{ij}^G \) is the nodes number. Actually, the symmetric normalized laplacian matrix \( \bar{L}^G \) is more commonly used in GCN
\[
\bar{L}^G = \frac{-1}{\sqrt{D^G}} L^G \frac{-1}{\sqrt{D^G}} = I_K - \frac{-1}{\sqrt{D^G}} A^G \frac{-1}{\sqrt{D^G}}
\] (5)

\( I_K \) is the identity matrix. Then the SD on \( \bar{L}^G \) is
\[
D(\bar{L}^G) = U \tilde{\Lambda} U^{-1} = U \tilde{\Lambda} U^T
\] (6)
since \( U = \{ u_1, u_2, \ldots, u_K \} \) is the orthogonal spectral matrix \((U U^T = I_K)\) and \( u_i, i = 1, \ldots, K \) are the FT basis. Thus, \( U^T \) and \( U \) can be regarded as the discrete form of FT and IFT. \( \tilde{\Lambda} \) is the diagonal matrix of eigenvalues. Therefore, the equation (5) can be presented as
\[
G(g \ast x) = U ((U^T g) \cdot (U^T x)) = U g_\theta(\tilde{\Lambda}) \cdot (U^T x)
\] (7)

where \( U^T g \) is the filter that need to be trained and can be considered as the function of \( \tilde{\Lambda} \) since the eigenvalues are corresponding to the frequencies of fourier domain.

Implementing equation (7) needs high computational complexity that mainly lying in computing the multiplication of \( U \) and eigenvalue decomposition. In practice, the \( g_\theta(\tilde{\Lambda}) \) is approximated by \( k \)th truncated expansion of chebyshev polynomials [54].

\[
g_\theta(\tilde{\Lambda}) \approx \sum_{i=0}^{k} \theta_i^x T_i(\tilde{\Lambda})
\] (8)

where \( \theta_i^x \) is the chebyshev coefficients. \( \tilde{\Lambda} = \frac{2}{\lambda_{max}} \tilde{\Lambda} - I_K \) and \( \lambda_{max} \) is the maximum value in \( \tilde{\Lambda} \). In addition, it’s obvious that
\[
G(g \ast x) \approx \sum_{i=0}^{k} \theta_i^x T_i(\bar{L}^G) x
\] (9)

where \( \bar{L}^G = \frac{2}{\lambda_{max}} \bar{L}^G - I_K \) since \((U \tilde{\Lambda} U^T)^k = U \tilde{\Lambda}^k U^T\). It can be seen that the graph convolution \( G(\cdot) \) depends on the \( k \)th-order neighborhoods since the scaled laplacian matrix \( \bar{L}^G \) is conducted \( k \) times, while the largest eigenvalue in \( \tilde{\Lambda} \) is 1.

The chebyshev polynomial \( T(\cdot) \) is defined in recursion form
\[
T_0(x) = 1,
T_1(x) = x,
T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x), \quad i > 1
\] (10)

To simplify the computation, limiting \( i = 1, \lambda_{max} = 2 \) and \( \theta = \theta_0^x = -\theta_1^x \) to approximate the graph convolution \( G(g \ast x) \) as the linear function of laplacian matrix \( \bar{L}^G \), when combining equation (5) then the graph convolution
\[
G(g \ast x) = \theta_0^x x + \theta_1^x \bar{L}^G x
\]
\[
= \theta_0^x x + \theta_1^x \left( \bar{L}^G - I_K \right) x
\]
\[
= \theta_0^x x - \theta_1^x \left( \frac{-1}{\sqrt{D^G}} A^G \frac{-1}{\sqrt{D^G}} \right) x
\] (11)
\[
= \theta \left( I_K + \frac{-1}{\sqrt{D^G}} A^G \frac{-1}{\sqrt{D^G}} \right) x
\]

However, the eigenvalues of \( I_K + \frac{-1}{\sqrt{D^G}} A^G \frac{-1}{\sqrt{D^G}} \) are in range [0,2], which may causes gradient explode or vanish when repeating this operation. Thus, [39] gives a renormalization trick that set \( \bar{A} = A^G + I_K \), \( \bar{D}_{ii} = \sum_{j=1}^{K} \bar{A}_{ij} \) and the GCN can be formulized as
\[
G^{(l+1)} = \left\{ \begin{array}{ll}
X, & l = -1 \\
\sigma (ZG^{(l)}W^{(l)}), & l > -1
\end{array} \right.
\] (12)

where \( X \) is the original input, \( Z = \frac{-1}{\sqrt{D}} \hat{A} \frac{1}{\sqrt{D}} \) and \( G^{(l)} \) are the output and trainable parameter matrix in \( l \)th layer, \( \sigma \) is the activation function and the ReLU is used in this paper.
III. PROPOSED METHODS

From the introduction of GCN, it can be seen that the critical parts for graph reasoning are attempt to obtain the $Z$ and $X$ in a more effective way, and this idea will be applied to all parts in the proposed methods.

A. Spatial Graph Reasoning Subnetwork

The hundreds of bands and high spatial resolution in HSI provide strong spectral-spatial relationships of the pixels, and these contextual information is easily utilized with superpixel segmentation, which can be implemented by adopting the SLIC algorithm [55] to generate a series compact superpixels. The original SLIC algorithm is difficult to directly placed into the network for end-to-end training since the existing undifferentiable min-max operations. [59] addresses this problem by transforming this operation to the differentiable weighted addition. With this technique, we successfully obtain the superpixels inside the network and generate the more effective descriptors. In addition, it can be noticed that the generated superpixels have brought the graph structure and the number of clustered districts is significantly less than the original pixels, benefitting to conduct the graph reasoning in a high efficiency.

In SAGRN, the obtained superpixels and descriptors are in fact the homogeneous regions and $D = \{d_1, d_2, \ldots, d_K\} \in \mathbb{R}^{K \times C}$, then the $X = \{x_1, x_2, \ldots, x_K\}$ can be easily obtained from $D$ using linear mapping. Concretely, each descriptor is computed through taking the mean value of the features in the corresponding area. This procedure is shown as follows

$$S = S(F)$$

$$d_i = \frac{\sum_{j=1}^{HW} I(S_j = i) \cdot F_j}{\sum_{j=1}^{HW} I(S_j = i)} \quad i = 1, \ldots, K$$

$$x_i = \xi (d_i)$$

where $F \in \mathbb{R}^{C \times H \times W}$ is the input feature, $S$ and $R$ represent the superpixel generating procedure and the corresponding segmentation result map. There are a total of $K$ descriptors and $i$ is the index, $C$, $H$ and $W$ are separately the number of channels, height and width of $F$. $I(S_j = i)$ is the binary indicator that judges whether the value of $j$th pixel in $S$ equal to $i$. $\xi$ represents the $1 \times 1$ convolutional layer, which is used to conduct the information integration of different channels.

To obtain $Z$, different from the conventional adjacent matrix that only considers the neighbor nodes, we treat it as a dense graph that each node possesses the relationship with all the other nodes to better capture the graph contexts. Specifically, the relationship of each pair of nodes is measured by computing their similarity in a mapped latent space

$$Z_{ij} = \frac{\exp(\phi(d_i)^T \psi(d_j))}{\sum_{k=1}^{K} \exp(\phi(d_i)^T \psi(d_k))} \quad i = 1, \ldots, K$$

where $\phi$ and $\psi$ are the mapping functions, which all can be implemented with the $1 \times 1$ convolution. Then, we normalize the similarity matrix using softmax function.

After obtaining the $Z \in \mathbb{R}^{K \times K}$ and $X \in \mathbb{R}^{K \times C}$, the spatial graph reasoning is achieved by directly adopting the GCN formula

$$G(Z, X) = \sigma (Z \cdot W \cdot X)$$

where $W \in \mathbb{R}^{C_1 \times C_2}$ is the trainable parameters and the ReLU is used as the activation function $\sigma$. It can be seen that each district-level feature is enhanced since it has the global view that capturing the contextual information lying in all the other nodes.

At last, these enhanced nodes need to be reprojected for recovering the shape of a pixel-level feature. In fact, the construction of pixel-level feature depends on the combination of reasonable descriptors since they possess specific connotations various from each other. In this paper, the node vectors after reasoning are considered as a group of bases which can span an effective feature space, where the information in any point can be inferred based on the linear aggregation of these vectors for the more complex semantic understanding. To this end, assume $G = G(Z, X) = \{g_1, g_2, \ldots, g_K\} \in \mathbb{R}^{K \times C_2}$, then the affinities $A \in \mathbb{R}^{K \times HW}$ between feature $F$ and node set $G$ are firstly measured in a new transformed space

$$A_{ij} = \frac{\exp(\rho(g_i)^T \eta(f_j))}{\sum_{h=1}^{HW} \exp(\rho(g_i)^T \eta(f_h))} \quad i = 1, \ldots, K$$

Then the target pixel-level feature $F_{sa\_main} \in \mathbb{R}^{HW \times C_3}$ is subsequently obtained by linearly combining these descriptors, where the affinity matrix $A$ is served as the corresponding weights. Thus, $F_{sa\_main} = A^T \zeta(G)$, and we subsequently reshape $F_{sa\_main}$ to $\mathbb{R}^{C_3 \times H \times W}$. Here, $\rho$, $\eta$ and $\zeta$ are implemented with $1 \times 1$ convolution and the subscript $main$ means the main branch to distinguish the later introduced auxiliary path in network. In above procedure, $C = C_1 = C_2 = C_3$ for convenience. After obtaining the $F_{sa\_main}$, through a $3 \times 3$ convolutional layer followed by a group normalization (GN) layer, a ReLU function, a $1 \times 1$ convolutional layer and a bilinear upsampling function, the probability matrix $P_{sa\_main} \in \mathbb{R}^{C_n \times H \times W}$ is acquired for computing loss, where $C_n$ is the number of categories. This procedure is symbolized as follows.

$$P_{sa\_main} = \delta_{sa\_main}(F_{sa\_main})$$

However, in the early stage of training, the disorganized high-level features may unfavorable to the homogeneous region generating, and the classification quality is suffered with the affected descriptors. To produce the more stable superpixels, we add an auxiliary branch to achieve a fast convergence and obtain the corresponding probability matrix $F_{sa\_aux}$ in a similar way as the main branch.

$$P_{sa\_aux} = \delta_{sa\_aux}(F)$$

In the proposed methods, the loss function is defined to $L(\cdot) = l(P(\cdot), Y)$, where $Y$ is the ground truth and $l$ is implemented with the cross entropy loss, thus the total loss of SAGRN is

$$L_{sa} = L_{sa\_main} + L_{sa\_aux}$$

The diagram of SAGRN is presented in Figure [1](a).
Backbone
Downsample
Grouping
Ground truth
Input image
Addition
(a)
(b)
Reproject
Reproject

Fig. 1. Architecture of the proposed SSGRN. Firstly, the feature $F$ is obtained after the received image through the backbone network. Then the $F$ is separately fed into two parts of (a) SAGRN and (b) SEGRN. In spatial part, a superpixel map $S$ is firstly generated based on $F$ and used to produce the descriptor set $D$. Then a graph convolution is conducted within $D$ and the generated $G$ is employed to reconstruct the pixel-level feature $F_{sa\text{-}main}$ under the guide of $F$ for classification. An auxiliary branch is imposed on $F$ to improve the quality of $S$. In spectral subnetwork, to obtain the pixel-level feature $F_{se}$, grouping, gradient convolution and reconstruction are sequentially implemented on the foundation of the downsampled $F$. In the end, a skip connection is built on $F$ to aggregate with $F_{sa\text{-}main}$ and $F_{se}$ for final classification.

B. Spectral Graph Reasoning Subnetwork

In SEGRN, we adopt the similar thinking as SAGRN. The homogeneous areas in SAGRN are regarded as the clustering of pixels. Thus, it is natural to consider aggregating the channels in a reasonable way. In order to obtain the spectral descriptors, inspired by [14], [15], we directly take the mean value of the aggregated features in channel direction, which are realized by grouping the feature maps of adjacent bands.

For the input feature $F' \in \mathbb{R}^{C \times H' \times W'}$ with $C$ bands: $F' = \{b_1, b_2, \ldots, b_C\}$, assume they are separately assigned to $M$ groups $F' = \{r_1, r_2, \ldots, r_M\}$, then the $i$th group

$$r_i = \left\{\frac{b_{C \cdot (i-1)+1}}{M}, \frac{b_{C \cdot (i-1)+2}}{M}, \ldots, \frac{b_{C \cdot i}}{M}\right\}$$

Thus the $i$ spectral descriptor can be obtained through

$$d_i = \frac{\sum_{j=1}^{C/M} b_{C \cdot (i-1)+j}}{C/M}$$

The remaining steps are similar with SAGRN. It should be noticed that the $F'$ is downsampled from $F$ with average pooling before conducting graph reasoning to save the computing resources. Therefore, $H' < H, W' < W$, and we group the downsampled $F'$ to generate the spectral descriptors. The pixel-level feature $F_{se} \in \mathbb{R}^{C \times H \times W}$ is reconstructed after reasoned descriptor linear combination and the bilinear interpolation is employed to guarantee the feature sizes are aligned for later aggregation. Through SEGRN, we successfully perform the graph reasoning in spectral direction since the contextual information lying in different channels is perceived and obtain an enhanced feature where each channel is improved by capturing the relationships with other bands.

At last, the loss of SEGRN is computed by

$$L_{se} = l(\delta_{se}(F_{se}), Y)$$

The diagram of SEGRN is depicted in Figure 1 (b).

C. Spectral-Spatial Graph Reasoning Network

The feature $F$ is obtained through the backbone network composing with three blocks. Each block includes a convolutional layer followed by a GN layer and a ReLU function. There is 2x downsampling after the first block to reduce the memory consumption. The whole network is trained from the scratch and no pre-trained parameters of the existing popular models need to be loaded.

After passing through the SAGRN and SEGRN, we obtain the corresponding enhanced features $F_{sa\text{-}main}$ and $F_{se}$. To preserve the original information, we adopt the residual skip connection, thus the spectral-spatial fused feature $F_{fused} \in \mathbb{R}^{C \times H \times W}$ is defined as

$$F_{fused} = F_{sa\text{-}main} + F_{se} + F$$
And the corresponding loss is also obtained

$$L_{\text{fused}} = l(\delta_{\text{fused}}(F_{\text{fused}}), Y)$$  \hspace{1cm} (24)

At last, the total loss of the proposed SSGRN is computed by

$$L_{\text{ss}} = L_{\text{sa}} + L_{\text{sc}} + L_{\text{fused}}$$  \hspace{1cm} (25)

The whole diagram of SSGRN is shown in Figure 1.

IV. EXPERIMENTS

In this section, we first introduce the datasets and implementation details, then we conduct a series of qualitative and quantitative comprehensive analyses. The comparison between the proposed methods with other state-of-the-art approaches will be presented in the end.

A. Dataset

1) Indian Pines: This scene was gathered at North-western Indiana by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor in 1992, consisting 200 bands with size of 145 × 145 pixels that in 20m spatial resolution after water absorption bands were removed and in the wavelength range of 0.4-2.5µm. 16 vegetation classes are involved in this scene, related information such as original image visualization, ground truth, configuration of training, validation and testing samples has been presented in Figure 2 and Table I.

2) Pavia University: This scene was obtained over Pavia University at Northern Italy by Reflective Optics System Imaging Spectrometer (ROSIS) in 2001, consisting 103 bands with size of 610 × 340 pixels that in 1.3m spatial resolution and in the wavelength range of 0.43-0.86µm. 9 categories are included in this data, which is shown in Figure 3 and Table II.

3) University of Houston: This scene was acquired over the University of Houston campus and its neighbor regions by ITRES-CASI 1500 sensor in 2012, containing 144 bands with size of 349 × 1905 pixels that in 2.5m spatial resolution and in the wavelength range of 0.4-1.0µm. 15 categories are included in this data, which is depicted in Figure 4 and Table III. This scene was also used in the 2013 IEEE GRSS Data Fusion Contest.

B. Implementation Details and Experimental Settings

We employ Pytorch to implement these methods. The base learning rate is set to 1e-3, which is adaptively ad-
We evaluate the different parts of the proposed methods on Indian Pines dataset, including the branches of spatial or spectral graph reasoning through the aggregating of the homogeneous regions or channels, are need to be configured manually. In view of this, we analyze the influence of the descriptor number on network accuracies, and the results are described in Figure 5. It can be apparently observed that the accuracies increase as the descriptor number growing. Concretely, the accuracies increase quickly in early stage, but the growth rate gradually slow down as the descriptor number further increasing, and the turning points are separately 4 and 32 in SAGRN and SEGRN, respectively. In addition, the accuracies keep it low when the spectral descriptors less than 16 because of the spectral information losing if a large amount bands are aggregated at once. In our experiments, in order to achieve high accuracies while reduce the computational complexity as much as possible, we set the descriptor number as 256 in both SAGRN and SEGRN.

C. Parameter Analysis

In the proposed methods, only the number of descriptors in SAGRN and SEGRN, which are exactly graph nodes number that separately determines the scale of spatial and spectral graph reasoning through the aggregating of the homogeneous regions or channels, are need to be configured manually. All experiments are implemented on the Intel Xeon Gold 5118 2.30GHz processor and a single NVIDIA Tesla V100 GPU.

D. Ablation Study

We evaluate the different parts of the proposed methods on Indian Pines dataset, including the branches of spatial or spectral graph reasoning (SAGR or SEGR) and the auxiliary branch in SAGRN. In addition, the influence of multiple parameters is also analyzed. The parameter details of the proposed methods are shown in Table IV, and the parameters are separately represents the “channel number”, “kernel size”, “stride”, “padding size” and “dilation rate” of the corresponding layer. In addition, the table also indicates whether the GN and ReLU are used after the convolutional layer. In addition, the table also indicates the “channel number”, “kernel size”, “stride”, “padding size” and “dilation rate” of the corresponding layer.

The classification consistency to penalize the model possessing category preference. All experiments are repeated with 10 times, and we record the mean value in the experiments, including overall accuracy (OA), average accuracy (AA) and Kappa coefficient (Kappa). The OA is the most popular evaluation criteria, which is calculated through dividing the number of correct pixels by the number of pixels that need to be classified. However, the OA is usually affected by the phenomenon of unbalance categories. To tackle this problem, the AA and Kappa are separately computed based on the confusion matrix, where the recall values of all categories are averaged to get AA, while the Kappa is used to measure the classification consistency to penalize the model possessing category preference. All experiments are implemented on the Intel Xeon Gold 5118 2.30GHz processor and a single NVIDIA Tesla V100 GPU.

C. Parameter Analysis

In the proposed methods, only the number of descriptors in SAGRN and SEGRN, which are exactly graph nodes number that separately determines the scale of spatial and spectral graph reasoning through the aggregating of the homogeneous regions or channels, are need to be configured manually. In view of this, we analyze the influence of the descriptor number on network accuracies, and the results are described in Figure 5. It can be apparently observed that the accuracies increase as the descriptor number growing. Concretely, the accuracies increase quickly in early stage, but the growth rate gradually slow down as the descriptor number further increasing, and the turning points are separately 4 and 32 in SAGRN and SEGRN, respectively. In addition, the accuracies keep it low when the spectral descriptors less than 16 because of the spectral information losing if a large amount bands are aggregated at once. In our experiments, in order to achieve high accuracies while reduce the computational complexity as much as possible, we set the descriptor number as 256 in both SAGRN and SEGRN.

D. Ablation Study

We evaluate the different parts of the proposed methods on Indian Pines dataset, including the branches of spatial or spectral graph reasoning (SAGR or SEGR) and the auxiliary branch in SAGRN. In addition, the influence of multiple parameters is also analyzed. The parameter details of the proposed methods are shown in Table IV, and the parameters are separately represents the “channel number”, “kernel size”, “stride”, “padding size” and “dilation rate” of the corresponding layer.
TABLE VI
MODEL COMPLEXITIES IN DIFFERENT COMBINATIONS OF THE PROPOSED METHODS ON INDIAN PINES DATASET

| Method           | Params(M) | FLOPs(G) | OA(%) | AA(%) | Kappa(%) |
|------------------|-----------|----------|-------|-------|----------|
| FCN              | 0.49      | 8.70     | 91.06 | 96.35 | 89.71    |
| FCN+RCCA($R = 2$) | 0.86      | 16.00    | 93.40 | 95.38 | 92.38    |
| FCN+SAGR         | 0.98      | 12.46    | 89.87 | 91.81 | 88.31    |
| FCN+SAGR+AB      | 1.28      | 15.54    | 94.04 | 95.66 | 93.11    |
| FCN+CAM          | 0.78      | 12.42    | 94.34 | 97.52 | 93.47    |
| FCN+SEGR         | 1.09      | 12.00    | 95.64 | 97.88 | 94.97    |

TABLE VII
ACCURACIES VERSUS DIFFERENT TRAINING SAMPLE PERCENTS OF THE PROPOSED METHODS ON INDIAN PINES DATASET

| Training sample (%) | 100 | 80  | 60  | 40  | 20  |
|---------------------|-----|-----|-----|-----|-----|
| OA                  | 94.04 | 90.76 | 88.60 | 75.69 | 52.01 |
| SAGRN OA            | 95.66 | 92.99 | 92.38 | 82.54 | 62.39 |
| Kappa               | 93.11 | 89.33 | 86.80 | 72.26 | 45.60 |
| AA                  | 95.64 | 94.07 | 92.95 | 88.45 | 82.17 |
| SEGRN AA            | 97.88 | 97.39 | 97.06 | 93.15 | 87.56 |
| Kappa               | 94.97 | 93.16 | 91.86 | 86.76 | 79.58 |
| OA                  | 97.87 | 96.70 | 95.53 | 92.52 | 86.07 |
| SSGRN OA            | 98.77 | 98.30 | 97.85 | 96.45 | 91.92 |
| Kappa               | 97.52 | 96.17 | 94.83 | 91.39 | 84.02 |

E. Model Complexity and Stability

In order to more comprehensively analyze the proposed methods, we assess the parameter number (Params) and computational complexity of SAGRN and SEGRN, while the latter is shown in the form of floating-point operations per second (FLOPs), and the results are shown in Table VI. Basing on the backbone network of FCN, we simultaneously compare the performance of the proposed SAGR and SEGR with other existing commonly employed non-local contexts capturing modules that utilize self-attention mechanism, including PAM [30], RCCA [32] and CAM [30], where the recurrent number $R$ of RCCA is set to 2 to keep consistent with the original literature. It can be seen that our methods perform better than other modules since they achieve competitive accuracies with fewer computations. Concretely, for the square Indian Pines dataset, the inner product operations in PAM and recurrent RCCA are separately implemented $N^2$ and $4N\sqrt{N}$ times, while the proposed SAGR only needs $K^2 + NK$ times, where $N$ is the pixel number of the input feature, $K$ is the number of spatial descriptor, in practice $K \ll N$. In SEGRN, the complexity of SEGR is less than CAM since the feature $F$ is downsampled before implementing graph reasoning. The “AB” in Table VI represents the auxiliary branch in SAGRN. Although the accuracies of FCN+PAM on Indian Pines dataset are close to FCN+SAGR+AB, it may not tackle some larger scenes such as the University of Houston dataset at one time due to the issue that out of memory, which unavoidably affects the classification. Actually, in our experiments, the
combination of FCN, SAGR and AB equals to SAGRN, while FCN+SEGR equals to SEGRN.

We also evaluate the stability of the proposed methods. Specifically, we separately adopt different percent of training samples of Table IV and the corresponding accuracies are shown in Table VII. It is obvious that the SEGRN is more robust than SAGRN since its OA higher than 80% even if only 20% of the training samples are available, while SAGRN drops almost half the accuracies. Nevertheless, by fusing the extracted spatial and spectral graph contexts, more than 90% of pixels are classified correctly in SSGRN when only 40% training samples are known, reflecting the stability of the proposed methods in the case of few samples.

F. Visualization

Besides the above quantitative analyses, to more intuitively understand our methods, we separately visualize the affinity matrix A of SAGRN and SEGRN, which indicates the similarities between the selected different descriptors with the vectors lying in different spatial positions or channels of the feature F, and the results are shown in Figure 6 where the responsibility intensity is represented with different colors or the height in direction of vertical axis. In Figure 6 it can be seen that different descriptors highlight different areas, demonstrating that they separately possess close relationships with the corresponding region. In other words, the connotations of these descriptors are certainly the meaning of the emphasized areas. These concepts serve as the basic components and can be organized by linearly aggregating the reasoned graph nodes in G to generate more complex semantic information for the understanding of other positions. While in Figure 7 different descriptors pay unequal attentions on each band. Thus, the responsibility curves have various shapes.

We also evaluate the distinguishability of the extracted features of the proposed SSGRN in a three-dimensional space by utilizing the t-SNE dimension reduction [57], where the distributions of reference sample at the corresponding input image, feature F generated by backbone network, probability matrix $P_{sa\_aux}$ of auxiliary branch, $F_{sa\_main}$ of SAGRN, $F_{se}$ of SEGRN and the fused feature $F_{fused}$ are separately shown in Figure 8. It can be observed that the samples on original image space are mixed-up and difficult to be identified, while after the encoding of backbone network, the obtained F began to be separable, and the auxiliary branch further strengthen the distinctions. Benefitting from the graph reasoning modules, the pixel representations on $F_{sa\_main}$, $F_{se}$ and $F_{fused}$ possess high separability with large inter-class distances, and the points of each category constitute a unique manifold, indicating that the latent patterns of corresponding categories have been perceived, demonstrating the effectiveness of the proposed SSGRN.
TABLE VIII
ACCURACIES OF DIFFERENT ALGORITHMS ON INDIAN PINES DATASET (%)

| Metric | 1-DCNN | 2-DCNN | SSFCN | GCN | Mini-GCN | MDGCN | CADGCN | SAGRN | SEGRN | SSGRN |
|--------|--------|--------|-------|-----|---------|-------|--------|-------|-------|-------|
| OA     | 73.19±3.61 | 82.94±1.84 | 86.74±0.82 | 70.22±1.51 | 77.60±2.94 | 93.76±0.60 | 89.36±0.82 | 94.04±1.83 | 93.64±0.32 | 97.87±0.55 |
| AA     | 89.52±3.74 | 80.55±2.06 | 84.81±0.93 | 66.27±1.53 | 74.44±3.09 | 92.79±0.77 | 87.71±0.93 | 91.13±1.88 | 94.97±0.60 | 97.52±0.63 |
| Kappa  | 65.42±8.54 | 72.93±13.41 | 68.54±12.21 | 56.78±18.87 | 52.34±24.56 | 40.82±33.04 | 30.56±40.12 | 19.78±47.89 | 12.34±55.67 | 6.95±63.41 |

Class 1  86.92±3.73  99.23±2.31  99.23±2.31  76.15±3.99  76.92±8.41  100.00±0.00  95.38±1.25  93.31±4.25  100.00±0.00  99.23±0.31
Class 2  86.18±8.94  80.15±3.36  85.13±1.29  79.02±4.29  75.58±9.11  93.14±3.37  75.33±4.23  92.70±6.19  95.86±1.90  98.24±2.49
Class 3  66.70±9.11  86.47±2.41  91.15±1.80  86.68±4.51  85.64±9.91  92.52±1.67  81.38±5.34  91.52±2.41  96.30±0.23  96.58±1.18
Class 4  82.63±9.60  99.42±0.72  94.82±2.22  85.62±4.40  73.72±3.96  82.05±2.30  90.58±3.58  99.17±0.74  99.27±0.47  100.00±0.00
Class 5  92.22±3.35  95.91±3.37  94.93±9.05  86.48±3.52  74.10±3.42  95.84±2.93  96.64±3.00  99.81±0.50  99.52±0.27  99.57±0.04
Class 6  93.39±4.28  95.13±3.81  91.87±6.72  92.94±1.32  91.13±3.14  96.00±5.03  95.18±3.68  99.86±0.23  98.67±0.38  97.77±0.19
Class 7  93.75±8.39  100.00±0.00  97.50±5.00  91.25±8.40  62.00±1.91  95.83±1.34  83.07±3.56  95.00±1.09  87.50±3.67  100.00±0.00
Class 8  98.39±1.04  99.52±0.65  99.10±1.00  93.54±2.82  97.86±1.08  99.75±0.46  98.76±0.79  99.13±1.16  99.95±0.16  100.00±0.00
Class 9  86.67±6.33  83.93±5.00  95.00±6.74  83.83±9.93  0.00±0.00  77.78±0.61  0.00±0.00  100.00±0.00  100.00±0.00  100.00±0.00
Class 10 97.39±6.53  85.29±2.07  91.30±1.29  74.46±4.82  74.27±5.62  91.74±1.67  83.53±5.31  91.28±3.67  98.85±1.62  92.66±2.14
Class 11 56.92±8.31  64.44±4.42  74.82±5.12  53.18±8.12  71.21±5.98  92.88±1.74  89.93±3.73  90.32±1.17  97.20±1.45  97.71±2.07
Class 12 78.22±9.89  93.96±2.22  90.18±2.97  66.93±4.48  93.96±1.53  94.37±0.70  89.11±2.95  88.84±4.73  96.75±1.26  99.39±0.18
Class 13 98.10±9.22  99.99±0.20  99.99±0.29  97.24±1.31  99.14±1.21  94.21±0.84  99.71±0.64  99.71±0.67  100.00±0.00  100.00±0.00
Class 14 78.75±7.61  84.95±7.92  91.66±1.84  87.16±2.61  73.33±0.44  98.40±5.06  94.20±1.19  99.57±1.19  99.39±1.56  99.23±0.23
Class 15 74.97±9.39  87.98±1.12  88.11±1.81  70.17±3.46  98.11±1.31  93.87±0.26  98.95±1.49  97.55±0.87  96.65±0.23  96.60±0.42
Class 16 96.67±4.44  98.89±2.22  97.22±4.48  91.67±5.69  90.00±8.12  82.51±3.50  100.00±0.00  98.89±0.70  100.00±0.00

G. Performance Comparison

We conduct the comparison between the proposed methods with other state-of-the-art approaches, including the CNN-based networks: 1-DCNN [9] for spectral classification with the spectral vector, 2-DCNN [9] for spatial classification using the spatial patches, the FCN-based algorithm SSFCN [24] for spectral-spatial joint classification and the approaches on the foundation of graph convolution, including GCN [39], Mini-GCN [40], MDGCN [41] and CADGCN [42].

The implementation of GCN in the pattern of transductive learning, where the whole image is directly fed into the model since each pixel is regarded as a graph node, and it unavoidably consumes too much computing resource. Mini-GCN simplifies the problem by using a shrunk adjacency matrix that is obtained through computing only with the nodes in current mini-batch, while the graph nodes in MDGCN are obtained with the segmented superpixel and the classification is performed in superpixel-level, too. The CADGCN also uses the superpixels to generate the graph nodes and performs the pixel-region-pixel transformation.

Table VIII lists the classification accuracies of each algorithm, where the mean value and standard deviation are...
Hyperspectral Image Classification Based on Spatial Contextual Information Perception

Abstract

Index Terms

I. INTRODUCTION

Simultaneously reported. It can be seen that the FCN-based SSFCN obviously perform better than the CNN-based 1-DCNN and 2-DCNN, showing the importance of spectral-spatial combination and large vision field in HSI classification. However, the accuracies of SSFCN are still limited because SSFCN only perceives the local information since its convolutions are all local operators. For the graph convolution-based methods, the redundant calculations of adjacent pixels in GCN affect the classification while Mini-GCN attempts to use the simplified adjacency matrix to accelerate this procedure with mini-batch training strategy. However, the pixels lacking of the representativeness of various contexts are directly set as the descriptors in GCN and Mini-GCN, degrading the classification. Some effective descriptors are obtained in MDGCN and CADGCN with the help of superpixel segmentation on the original image, thus they get higher accuracies than GCN and Mini-GCN. In addition, it needs to be noticed that the adjacency matrices in MDGCN and CADGCN are all calculated by only considering the relationships of neighbor nodes. Compared with the above methods, our SAGRN performs better since the more effective descriptors can be flexibly and adaptively generated through the homogeneous areas that are produced by implementing superpixel segmentation on the intermediate features of the network. Moreover, the SAGRN possesses the global view since the relationships between each node and all the other nodes are measured. By implementing the spectral reasoning, the SEGRN successfully captures the graph contexts lying in different channels. To our best knowledge, it is the first time that the relationships of spectral bands are perceived through graph convolution in HSI classification community. At last, combining the spatial and spectral graph contexts that are separately obtained by SAGRN and SEGRN, the proposed SSGRN performs the best and achieves the overall accuracy of 97.87%, 98.04% and 95.59% on Indian Pines, Pavia University and University of Houston dataset.

We select and predict the networks having relative higher accuracies shown in Table VIII-X including SSFCN, MDGCN, CADGCN and the proposed methods, and the results are depicted in Figure 9-11. It can be seen that the classification maps are consistent with the accuracies presented in Table VIII-X. We can obviously find that the proposed methods have fewer noises (compare with MDGCN) and misclassifications (compare with SSFCN and CADGCN), generating the discriminative classification maps where the objects possess continuous surfaces and well-maintained edges.

V. CONCLUSION

In this paper, we propose a network called SSGRN to classify the HSI. Considering that the irregular distribution of land objects and the various relationships of different spectral bands, the corresponding contextual information is more suitably extracted from the perspective of graph. Concretely, this network contains two subnetworks that separately extract the spatial and spectral graph contexts. In spatial subnetwork SAGRN, in order to generate the more effective descriptors for graph reasoning, different from the previous approaches implementing superpixel segmentation on the original image, we perform this procedure on the intermediate features inside network based on the pixel spectral-spatial similarities to more flexibly and adaptively produce the homogeneous regions and the descriptors are acquired through separately aggregating these regions. In addition, the SAGRN introduces less error...
only maintain high accuracies but reducing the computational resource consumption and still perform stable with less training samples, while the performance comparison evaluations demonstrate the competitiveness compared with other state-of-the-art methods.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Paolo Gamba and Prof. Saurabh for providing the Pavia University and University of Houston datasets. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

REFERENCES

[1] X. Zhang, Y. Sun, K. Zhang, L. Zhang, and S. Wang, “Crop classification based on feature band set construction and object-oriented approach using hyperspectral images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 9, no. 9, pp. 4117–4128, Sep. 2016.
[2] X. Yang and Y. Yu, “Estimating soil salinity under various moisture conditions: An experimental study,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 5, pp. 2525–2533, May 2017.
[3] S. Matteoli, M. Diani, and G. Corsini, “A tutorial overview of anomaly detection in hyperspectral images,” IEEE Aerosp. Electron. Syst. Mag., vol. 25, no. 7, pp. 5–28, 2010.
[4] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep Learning-Based Classification of Hyperspectral Data,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 2094–2107, June 2014.
[5] T. Li, J. Zhang, and Y. Zhang, “Classification of hyperspectral image based on deep belief networks,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct 2014, pp. 5132–5136.
[6] P. Zhou, J. Han, G. Cheng, and B. Zhang, “Learning compact and discriminative stacked autoencoder for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 7, pp. 4823–4833, 2019.
[7] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” J. Sensors, vol. 2015, 2015.
[8] W. Zhao and S. Du, “Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 8, pp. 4544–4554, Aug 2016.
[9] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 6232–6251, Oct 2016.
[10] Z. Zhong, J. Li, Z. Luo, and M. Chapman, “Spectral-spatial residual network for hyperspectral image classification: A 3-d deep learning framework,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 847–858, Feb 2018.
[11] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, “Simultaneous spectral-spatial feature selection and extraction for hyperspectral images,” IEEE Trans. Cybern., vol. 48, no. 1, pp. 16–28, 2018.
[12] Z. Feng, S. Yang, M. Wang, and L. Jiao, “Learning dual geometric low-rank structure for semisupervised hyperspectral image classification,” IEEE Trans. Cybern., vol. 51, no. 1, pp. 346–358, 2021.
[13] Y. Xu, B. Du, F. Zhang, and L. Zhang, “Hyperspectral image classification via a random patches network,” ISPRS-J. Photogramm. Remote Sens., vol. 142, pp. 344–357, 2018.
[14] Y. Xu, L. Zhang, B. Du, and F. Zhang, “Spectral-spatial unified networks for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 10, pp. 5893–5909, Oct 2018.
[15] D. Wang, B. Du, L. Zhang, and Y. Xu, “Adaptive spectral-spatial multiscale contextual feature extraction for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., pp. 1–17, 2020.
[16] H. Sun, X. Zheng, X. Lu, and S. Wu, “Spectral–spatial attention network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 5, pp. 3232–3245, 2020.
[17] L. Mou and X. X. Zhu, “Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 1, pp. 110–122, 2020.

Fig. 11. Classification map of different methods on University of Houston dataset. (a) SSFCN. (b) MDGCN. (c) CADGCN. (d) SAGRN. (e) SEGRN. (f) SSGRN.
WANG et al.: SSGRN FOR HSI CLASSIFICATION

[18] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, “Hyperspectral image classification with attention-aided cnns,” IEEE Trans. Geosci. Remote Sens., pp. 1–13, 2020.

[19] X. Tang, F. Meng, X. Zhang, Y. Cheung, J. Ma, F. Liu, and L. Jiao, “Hyperspectral image classification based on 3-d octave convolution with spatial-spectral attention network,” IEEE Trans. Geosci. Remote Sens., pp. 1–18, 2020.

[20] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, “Residual spectral-spatial attention network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., pp. 1–14, 2020.

[21] H. Lee and H. Kwon, “Going deeper with contextual cnn for hyperspectral image classification,” IEEE Trans. Image Process., vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[22] M. Zhang, W. Li, and Q. Du, “Diverse region-based cnns for hyperspectral image classification,” IEEE Trans. Image Process., vol. 27, no. 6, pp. 2623–2634, 2018.

[23] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, “Exploring hierarchical convolutional features for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 11, pp. 6712–6722, 2018.

[24] Y. Xu, B. Du, and L. Zhang, “Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification,” IEEE Trans. Big Data., pp. 1–1, 2019.

[25] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, “Fpga: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 8, pp. 5612–5626, 2020.

[26] D. Wang, B. Du, and L. Zhang, “Fully contextual network for hyperspectral scene parsing,” IEEE Trans. Geosci. Remote Sens., pp. 1–16, 2021.

[27] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, “Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnns,” IEEE Trans. Cybern., vol. 50, no. 1, pp. 100–111, 2020.

[28] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” CoRR, vol. abs/1711.07971, 2017.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern. Recognit. (CVPR), 2019, pp. 3141–3149.

[31] Y. Yuan and J. Wang, “Ocnet: Object context network for scene parsing,” CoRR, vol. abs/1809.00916, 2018.

[32] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ocnet: Criss-cross attention for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 603–612.

[33] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, “Asymmetric non-local neural networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 593–602.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern. Recognit. (CVPR), 2017, pp. 6230–6239.

[35] Y. Yuan, X. Chen, and J. Wang, “Object-Contextual Representations for Semantic Segmentation,” arXiv e-prints, p. arXiv:1909.11065, Sep. 2019.

[36] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding, “Acfnet: Attentional class feature network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 6797–6806.

[37] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, “Cascade superpixel regularized gabor feature fusion for hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 5, pp. 1638–1652, 2020.

[38] J. Jiang, J. Ma, and X. Liu, “Multilayer spectral-spatial graphs for label noisy robust hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., pp. 1–14, 2020.

[39] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” CoRR, vol. abs/1609.02907, 2016. [Online]. Available: http://arxiv.org/abs/1609.02907

[40] D. Hong, L. Gao, X. Zhang, A. Plaza, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., pp. 1–13, 2020.

[41] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, “Multiscale dynamic graph convolutional network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 5, pp. 3162–3177, 2020.

[42] S. Wan, C. Gong, P. Zhong, S. Pan, G. Li, and J. Yang, “Hyperspectral image classification with context-aware dynamic graph convolutional network,” IEEE Trans. Geosci. Remote Sens., pp. 1–16, 2020.