Spectral Pyramid Graph Attention Network for Hyperspectral Image Classification

Tinghuai Wang\textsuperscript{1}, Guangming Wang\textsuperscript{1}, Kuan Eeik Tan\textsuperscript{1}, and Donghui Tan\textsuperscript{2}

\textsuperscript{1}Huawei Helsinki Research Center, Finland
Email: tinghuai.wang@huawei.com

\textsuperscript{2}Huawei Technologies, China

Abstract—Convolutional neural networks (CNN) have made significant advances in hyperspectral image (HSI) classification. However, standard convolutional kernels neglect the intrinsic connections between data points, resulting in poor region delineation and small spurious predictions. Furthermore, HSIs have a unique continuum data distribution along the high dimensional spectrum domain - much remains to be addressed in characterizing the spectral contexts considering the prohibitively high dimensionality and improving reasoning capability in light of the limited amount of labeled data. This paper presents a novel architecture which explicitly addresses these two issues. Specifically, we design an architecture to encode the multiple spectral contextual information in the form of spectral pyramid of multiple embedding spaces. In each spectral embedding space, we propose graph attention mechanism to explicitly perform interpretable reasoning in the spatial domain based on the connection in spectral feature space. Experiments on three HSI datasets demonstrate that the proposed architecture can significantly improve the classification accuracy compared with the existing methods.

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

The rapid development of hyperspectral sensors enables the observation of hundreds of continuous bands throughout the electromagnetic spectrum with high spectral resolution. The rich spectral signatures of hyperspectral images (HSIs) facilitate the study of the chemical properties of scene materials remotely. Hyperspectral image classification has consequently been playing an increasingly important role in various fields, such as mining, agriculture, environmental monitoring, and land-cover mapping.

Various approaches have been proposed to address the hyperspectral image classification problem. Early approaches mainly adopted traditional machine learning methods, e.g., KNN [1], SVM [2], graphical model [3], extreme learning machine [4], dictionary learning [5], and among others trained on hand-crafted features from HSI data to fully exploit the spectral information. Nonetheless, accurately classifying different land-cover categories using only the spectral information regardless of the spatial constraint is difficult. Methods [6, 7] exploiting both spatial and spectral information have also been proposed to address this issue.

Early methods largely relied on hand-crafted features empirically designed for HSI data which have limited discriminative power. Inspired by the success of deep neural networks (DNN) in natural image classification, deep learning based methods have been proposed for hyperspectral image classification which have significantly boosted the performance thanks to the strong representation capability. The first DNN approach was proposed by Chen et al. [8], which utilized stacked autoencoders to learn high-level features. Recurrent Neural Network (RNN) based architecture was proposed by Mou et al. [9]. More powerful end-to-end Convolutional Neural Network (CNN) [10]–[17] based architectures have advanced the state-of-the-art recently. Lee et al. [18] explored local contextual interactions by jointly exploiting local spatio-spectral relationships of neighboring individual pixel vectors. Song et al. [19] introduced residual learning [20] to build very deep network for extracting more discriminative features for HSI classification.

Despite of the remarkable performance by the CNN-based methods, they suffer from several drawbacks. Specifically, standard convolutional kernels work in regular receptive fields for feature response, whose weights are fixed given their position within the small convolution window. Such position-determined weights lead to the isotropy of the kernel with respect to the feature attributes of neighboring locations. This limitation of the standard convolutional kernel neglects the intrinsic or extrinsic structural connections between data points, resulting in poor region delineation and small spurious predictions. To address this problem, graph convolutional network (GCN) based approach S\textsuperscript{2}GCN has been proposed by Qin et al. [21] which operated on a graph constructed on the local data points and is able to aggregate and transform feature information from the neighbors of every graph node given their relative spatial distance. However, there are several limitations of this approach. Firstly, the graph is constructed based on the rigid pixel lattice which inherently limited the scope of interaction between data points. Secondly, the relation between nodes only accounts for their spatial distance neglecting their correlation in the feature space. Furthermore, S\textsuperscript{2}GCN only conducts reasoning at a single feature space which is potentially limited to the capacity of the early feature extraction block of the network. Thirdly, S\textsuperscript{2}GCN is a graph-based semi-supervised approach which predicts unlabeled data with the presence of labeled data which potentially harms its...
accuracy on dataset without any available labels.

To explicitly address the above issues, we present a novel architecture to encode the multiple spectral contextual information in a hierarchical manner, forming multiple spectral embedding spaces. In multiple spectral embedding spaces, we propose graph attention mechanism to explicitly perform context based reasoning in the spatial domain based on the connection in spectral feature space. Our graphs are dynamically constructed based on the intrinsic structure of the data, i.e., nodes of a same neighborhood are assigned with different importance, enabling thorough interaction between comprising data points while increasing the model capacity. After the graph reasoning, an attention based aggregation reminiscent of multi-head attention [22] with focus on specific spectral channels given the contextual scopes is applied on the spectral feature pyramid rather than a simply averaging in GAT [23].

This is based on the observation that features from different contextual level carry different degrees of discriminative power for classification. Extracting the key discriminative features from each contextual level and enforcing this information during the aggregation can effectively improve the classification accuracy.

II. METHOD

We propose a novel end-to-end graph attention architecture for spectral feature learning and reasoning of HSIs, which consists of mainly two blocks, i.e., spectral feature pyramid learning and graph attention based reasoning, as illustrated in Fig. 1.

A. Spectral Pyramid Learning

We propose to utilize atrous convolution to probe the HSI signal along the spectral dimension at multiple sampling rates for capturing varying spectral contextual information, forming a spectral pyramid. It brings us mainly two advantages: (1) multiple spectral contextual information is captured without introducing large convolutional kernel or fully connected layer which significantly reduces the number of trainable model parameters and computational complexity (2) large effective fields-of-views is achieved without resorting to resampling features which in turn causes downsampled feature map.

As illustrated in Fig. 1 we adopt 3D convolutions of kernel size $3 \times 1 \times 1$ and dilation rates $\{(1, 1, 1), (12, 1, 1), (24, 1, 1), (36, 1, 1)\}$ respectively. With the spatial kernel size $1 \times 1$, the 3D convolution is equivalent to 1D convolution over the spectral vector at each pixel location. The output $y(s, i, j)$ of the atrous convolution at signal location $(s, i, j)$ with a filter $w[k]$ of length $K$ is defined as

$$y(s, i, j) = \sum_{k=1}^{K} x[s + r \cdot k, i, j]w[k]$$

where $r$ indicates the stride with which we sample the input signal, and $r = 1$ corresponds to the standard convolution.

A spectral pooling layer is also proposed to gather the global spectral context. Specifically, the spectral pooling layer consists of $[\text{AdaptiveAvgPool3d} + \text{Conv3d}[1 \times 1 \times 1] + \text{BatchNorm3d} + \text{ReLU}]$, inspired by [24] in the spatial domain.

As illustrated in Fig. 1 the atrous spectral pyramid comprises 5 streams, i.e., feature maps from 4 Conv3D with different dilation rates and spectral level features from spectral pooling. The spectral pyramid thus captures spectral contextual information at multiple scales while maintaining a very low computational complexity.

A spectral pooling layer is also proposed to gather the global spectral context. Specifically, the spectral pooling layer consists of $[\text{AdaptiveAvgPool3d} + \text{Conv3d}[1 \times 1 \times 1] + \text{BatchNorm3d} + \text{ReLU}]$, inspired by [24] in the spatial domain.

As illustrated in Fig. 1 the atrous spectral pyramid comprises 5 streams, i.e., feature maps from 4 Conv3D with different dilation rates and spectral level features from spectral pooling. The spectral pyramid thus captures spectral contextual information at multiple scales while maintaining a very low computational complexity.

Residual blocks [20] are applied on each feature stream to further aggregate local spatial and spectral contexts and transform feature embeddings. Specifically, two bottleneck modules with filter size $(64, 128)$ respectively and expansion factor of 4 are applied per feature stream. $h' = \{h'_i\} \in \mathbb{R}^{d'}$. 

Fig. 1: Illustration of the our spectral pyramid graph attention network (SPGAT).
between features. A pair of features, we compute the feature differences which
computes the attention weights based on the concatenation of
softmax operation. As opposed to GAT [23], which
is an attention mechanism to compute the attention
\[ \alpha_{i,j} = \frac{e^{\text{LeakyReLU}(W_{\theta} h_i^T W_{\phi} h_j)}}{\sum_{j=1}^{N} e^{\text{LeakyReLU}(W_{\theta} h_i^T W_{\phi} h_j)}}, \]
where \( W_{\theta} \), \( W_{\phi} \) and \( W_{\psi} \) are the trainable parameters of embedding functions \( \theta(\cdot) \), \( \phi(\cdot) \) and \( \psi(\cdot) \) respectively. To make the attention weights comparable across nodes, they are normal-
ized by softmax operation. As opposed to GAT [23], which computes the attention weights based on the concatenation of a pair of features, we compute the feature differences which is more efficient and explicit to characterize the correlation between features.

Finally, the normalized attention weights are used to perform a weighted combination all nodes to obtain a new set of features,

\[ h'_i = \text{LeakyReLU}(\sum_{j=1}^{N} \alpha_{i,j} W_{\xi} h_j), \]
where \( W_{\xi} \) indicates the weights of a 1-D convolutional layer \( \xi(\cdot) \) before applying a LeakyReLU nonlinearity.

After obtaining a new set of feature streams from the multiple graph attention layers, we propose an attention based aggregation with focus on specific spectral channels given the contextual scopes, rather than a simply averaging in GAT [23]. As illustrated in Fig. 3, spectral attention module takes
the feature maps from neighboring contextual levels as input and computes spectral level attention coefficients to guide the upper level feature to focus on certain spectral channels. This is based on the observation that features from different contextual level carry different degrees of discriminative power for classification. Extracting the key discriminative features from each contextual level and enforcing this information during the aggregation can effectively encode the contextual information at multiple scales.

III. EXPERIMENTAL RESULTS

In this section, we conduct extensive experiments on three publicly available hyperspectral image datasets, and present results using four metrics including per-class accuracy, overall accuracy (OA), average accuracy (AA), and Kappa coefficient. The network architecture of our proposed SPGAT is identical for all the datasets. Specifically, two graph attention layers per contextual level are used, with graphs constructed on 7×7 image patches. Learning rate of 0.001 and Adam optimizer are adopted. We follow [25] for selecting the training and testing sets. The number of training epochs is set to 500. All the reported accuracies are calculated based on the average of ten training sessions to obtain stable results.

A. Datasets

The University of Pavia dataset captured the University of Pavia with the ROSIS sensor in 2001. It consists of 610×340 pixels with a spatial resolution of 1.3 m×1.3 m and has 103 spectral channels in the wavelength range from 0.43 µm to 0.86 µm after removing noisy bands. This dataset includes 9 land-cover classes as listed in Table I. The false color image and ground-truth map are shown in Fig. 4.

The Indian Pines dataset was collected by Airborne Visible/Infrared Imaging Spectrometer sensor which consists of 145×145 pixels with a spatial resolution of 20 m×20 m and has 220 spectral channels covering the range from 0.4 µm to 2.5 µm. Fig. 5 exhibits the false color image and ground-truth map of the Indian Pines dataset. This dataset includes 16 land-cover classes as listed in Table II.

The Kennedy Space Center dataset which was taken by AVIRIS sensor over Florida with a spectral coverage ranging from 0.4 µm to 2.5 µm, contains 224 bands and 614 × 512 pixels with a spatial resolution of 18 m. This dataset comprises...
13 land-cover classes as listed in Table III and Fig. 6 exhibits the false color image and ground-truth map.

B. Classification Results

We quantitatively and qualitatively evaluate our proposed method and compare with various recent deep learning based methods to demonstrate its effectiveness.

1) The University of Pavia Dataset: Table I presents the quantitative results obtained by different methods on the University of Pavia dataset, where the highest value in each row is highlighted in bold. Our proposed SPGAT outperforms the competing methods on 7 out of 9 categories and exhibits the best overall results, i.e., OA, AA and Kappa. In general, all methods except SSLSTMs [25] and SPGAT fail to capture larger context, which consequentially produce inferior results. SSLSTMs applies LSTM to learn both spatial and spectral features, however it fails to characterize the large scale spectral contexts due to the limitation of LSTM models on high dimension feature. On the contrary, our SPGAT encodes the spectral contexts in a hierarchical manner and utilizes graph attention models to explicitly perform multi-scale inference in the spatial domain based on the connection in spectral feature space.

Fig. 4 shows a visual comparison of SPGAT and SSLSTMs on the University of Pavia dataset. We can observe that SSLSTMs suffers from mis-classifications even in large regions of the same category, due to the lack of inference in the spatial space as well as the failure to encode the most discriminative spectral features. We can see that our SPGAT is able to produce both accurate and coherent predictions.

2) The Indian Pines Dataset: Table II presents the quantitative results obtained by different methods on the Indian Pines dataset. In addition to tradition methods, we also compare with methods based on graph neural networks, i.e., GCN [26] and S^2GCN [21], to demonstrate the effectiveness of our method. Overall, our method surpasses all the compared methods, with a significant margin of 5.11% comparing to the best competing method S^2GCN. GCN and S^2GCN both construct their graphs based the pixel lattice, whereas S^2GCN encodes the spatial distance between adjacent nodes in graph which enables it to capture the local intrinsic structure in HSI data. However, SPGAT has the advantages over the compared GNN based methods as follows: (a) SPGAT constructs data-driven graph rather than a rigid spatial structure (b) SPGAT assigns adjacency matrix with both connectivities and strengths between nodes which fully characterize their correlation in spectral feature space (c) SPGAT features the multiple spectral contextual graph based inference which enables robustness and accuracy in generalizing to HSI data captured by various sensors.

3) The Kennedy Space Center Dataset: Table III presents the quantitative results obtained by different methods on the Kennedy Space Center dataset. Similar to the results on the University of Pavia dataset, SPGAT produces the best OA accuracy among all compared methods, and SSLSTMs achieves the second best. It is worth-noting that two other graph based approaches, i.e., GCN [26] and ELP-RGF [35] also show promising results compared with traditional methods, which demonstrates the advantage of conducting graph reasoning.

C. Ablation Study

In this section, we conduct ablation study on the University of Pavia dataset to investigate the effectiveness of various novel parts of our proposed architecture. To show the importance of generating multiple spectral contextual feature, we remove the branches where the dilation rate is larger than 1, keeping the branch with dilation rate of 1 (SPGAT-1). In order to demonstrate the effectiveness of attention graph layer, we compare with the baseline model replacing GAT with GCN [26] layers (SPGCN). Finally we show that the spectral attention block aggregates features preserving the discriminative information across multiple spectral contextual level compared with averaging (SPGAT-Avg). As summarized in Table IV, introducing multiple spectral contextual embedding spaces enables an OA gain of 3.67%, whilst replacing graph attention layer with GCN leads to OA drop of 2.55%. We also observe that the spectral attention block also boosts the performance by 0.31%.

IV. Conclusions

In this paper, we proposed a novel architecture for hyperspectral image classification. We demonstrated that using Atrous convolution to probe the HSI signal along the spectral dimension at multiple sampling rates could efficiently and effectively encode varying spectral contextual information. We further proposed graph attention based reasoning in each spectral embedding space which produced significantly boosted classification accuracy compared with existing methods.
| Class                        | R-PCA CNN | PPF-CNN | CD-CNN | SS-CNN | 3D-CNN | SSLSTMs | SPGAT |
|-----------------------------|-----------|---------|--------|--------|--------|---------|-------|
| Asphalt                    | 92.43     | 97.42   | 94.60  | 97.40  | 96.72  | 98.63   | 98.60 |
| Meadows                    | 94.84     | 95.76   | 96.00  | 99.40  | 96.31  | 98.74   | 96.11 |
| Gravel                     | 90.89     | 94.05   | 95.50  | 98.84  | 97.15  | 96.57   | 97.83 |
| Trees                      | 93.99     | 97.52   | 95.90  | 96.16  | 96.16  | 98.43   | 99.23 |
| Painted metal sheets       | 100.0     | 100.0   | 100.0  | 100.0  | 99.81  | 99.94   | 100.0 |
| Bare Soil                  | 92.86     | 99.13   | 94.10  | 98.70  | 94.87  | 99.43   | 100.0 |
| Bitumen                    | 93.89     | 98.96   | 96.19  | 100.0  | 97.44  | 99.31   | 100.0 |
| Self-Blocking Bricks       | 91.18     | 93.62   | 88.80  | 94.57  | 98.23  | 97.98   | 99.92 |
| Shadows                    | 99.33     | 99.60   | 99.5   | 99.87  | 98.04  | 99.39   | 100.0 |
| OA                         | 95.87     | 96.48   | 96.73  | 98.41  | 96.55  | 98.48   | 98.92 |
| AA                         | 94.38     | 97.03   | 95.77  | 98.22  | 97.19  | 98.51   | 99.07 |
| Kappa                      | -         | -       | -      | -      | 93.50  | 97.56   | 97.86 |

TABLE I: Per-class accuracy, OA, AA (%), and Kappa coefficient achieved by different methods on the University of Pavia dataset.

| Class                         | LapSVM | S^2SL | SSGEL | GCN | S^2GCN | LBMSELM | SPGAT |
|-------------------------------|--------|-------|-------|-----|--------|---------|-------|
| Alfalfa                       | 91.81  | 84.96 | 100.0 | 100.0| 100.0  | 98.03   | 96.22 |
| Corn-notill                   | 72.89  | 70.45 | 84.72 | 60.92| 92.22  | 89.97   | 99.27 |
| Corn-mintill                  | 63.45  | 65.21 | 84.70 | 45.99| 94.42  | 98.43   | 95.86 |
| Grass-pasture                 | 77.40  | 87.81 | 84.05 | 89.08| 98.03  | 96.11   | 99.95 |
| Grass-trees                   | 96.39  | 96.30 | 90.28 | 84.83| 98.07  | 99.58   | 98.78 |
| Grass-pasture-mowed           | 100.0  | 91.24 | 100.0 | 98.2 | 100.0  | 96.80   | 98.78 |
| Hay-windrowed                 | 97.25  | 98.51 | 100.0 | 97.53| 99.69  | 100.0   | 98.78 |
| Oats                          | 100.0  | 95.55 | 26.30 | 61.01| 100.0  | 96.80   | 98.78 |
| Soybean-notill                | 73.44  | 83.91 | 76.68 | 65.90| 97.41  | 74.49   | 98.15 |
| Soybean-mintill               | 61.67  | 69.23 | 86.34 | 51.57| 83.95  | 92.38   | 98.29 |
| Soybean-clean                 | 66.55  | 83.47 | 77.04 | 91.35| 94.28  | 98.79   | 99.95 |
| Wheat                         | 98.41  | 99.05 | 99.03 | 70.91| 100.0  | 96.82   | 98.78 |
| Woods                         | 89.05  | 93.81 | 99.79 | 66.87| 99.01  | 93.80   | 98.78 |
| Buildings-Grass-Trees-Drives  | 80.92  | 67.94 | 76.46 | 63.38| 97.19  | 94.49   | 98.78 |
| Stone-Steel-Towers            | 99.55  | 85.81 | 100.0 | 62.22| 94.85  | 94.94   | 98.78 |
| OA                            | 75.71  | 79.85 | 86.53 | 66.87| 91.64  | 87.47   | 96.75 |
| AA                            | 84.72  | 84.61 | 83.41 | 63.38| 94.92  | 89.37   | 97.42 |
| Kappa                         | 72.62  | 77.25 | 84.72 | 62.22| 90.41  | 85.61   | 96.30 |

TABLE II: Per-class accuracy, OA, AA (%), and Kappa coefficient achieved by different methods on the Indian Pines dataset.

| Class                        | LapSVM | SSLP-SVM | GCN | ELF-RGF | SSSSTMs | SPGAT |
|-------------------------------|--------|----------|-----|---------|---------|-------|
| Scrub                         | 87.17  | 87.19    | 86.91| 100.0   | 99.56   | 98.86 |
| Willow swamp                  | 95.63  | 77.38    | 83.29| 99.75   | 90.41   | 98.86 |
| Cabbage palm hammock         | 70.90  | 85.87    | 87.57| 93.06   | 99.72   | 98.86 |
| Cabbage palm/oak hammock     | 83.97  | 51.97    | 24.08| 75.39   | 99.56   | 98.86 |
| Slash pine                   | 79.08  | 41.13    | 63.36| 55.95   | 93.79   | 98.62 |
| Oak/broadleaf hammock        | 89.62  | 36.43    | 61.01| 95.64   | 95.15   | 92.53 |
| Hardwood swamp               | 96.34  | 72.06    | 91.20| 98.79   | 100.0   | 94.06 |
| Graminoid marsh              | 93.34  | 74.67    | 78.20| 99.10   | 98.40   | 98.65 |
| Spartina marsh               | 98.12  | 89.52    | 85.39| 97.21   | 99.57   | 99.89 |
| Cattail marsh                | 92.90  | 75.53    | 84.28| 84.79   | 100.0   | 99.89 |
| Salt marsh                   | 94.92  | 84.47    | 94.68| 99.95   | 99.47   | 99.95 |
| Mud flats                    | 94.22  | 68.16    | 82.14| 94.37   | 99.31   | 99.95 |
| Water                        | 99.08  | 99.15    | 98.99| 100.0   | 99.88   | 100.0 |
| OA                            | 91.25  | 75.52    | 83.60| 93.21   | 97.89   | 98.15 |
| AA                            | 90.41  | 72.72    | 78.60| 91.84   | 97.28   | 98.81 |
| Kappa                        | 90.25  | 72.88    | 81.70| 92.45   | 97.65   | 97.84 |

TABLE III: Per-class accuracy, OA, AA (%), and Kappa coefficient achieved by different methods on the Kennedy Space Center dataset.

|                   | SPGAT-1 | SPGAT-CN | SPGAT-Avg | SPGAT |
|-------------------|---------|----------|-----------|-------|
| OA                | 95.23   | 96.37    | 98.61     | 98.92 |

TABLE IV: OA achieved by different baselines on the University of Pavia dataset.
Fig. 4: The university of Pavia dataset: (a) False color image (b) Groundtruth map (c) Prediction of SSLSTMs [25] (d) Prediction of the proposed SPGAT

Fig. 5: Indian Pines dataset: (a) False color image (b) Groundtruth map (c) Prediction of LBMSELM [34] (d) Prediction of the proposed SPGAT

Fig. 6: Kennedy Space Center dataset: (a) False color image (b) Groundtruth map (c) Prediction of SSLSTMs [25] (d) Prediction of the proposed SPGAT

REFERENCES

[1] L. Ma, M. M. Crawford, and J. Tian, “Local manifold learning-based k-nearest-neighbor for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4099–4109, 2010.

[2] B.-C. Kuo, C.-S. Huang, C.-C. Hung, Y.-L. Liu, and I.-L. Chen, “Spatial information based support vector machine for hyperspectral image classification,” in *2010 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 832–835.

[3] L. Shi, L. Zhang, J. Yang, L. Zhang, and P. Li, “Supervised graph embedding for polarimetric sar image classification,” *IEEE Geoscience
S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, “Two-stream deep learning approach for remote sensing image classification,” in Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2019, pp. 3962–3965.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” arXiv preprint arXiv:1802.02611, 2018.

F. Zhou, R. Hang, Q. Liu, and X. Yuan, “Hyperspectral image classification using spectral-spatial lstms,” Neurocomputing, vol. 328, pp. 39–47, 2019.

T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proceedings of International Conference on Learning Representations, 2017.

K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, “Deep supervised learning for hyperspectral data classification through convolutional neural networks,” in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2015, pp. 4959–4962.

W. Li, G. Wu, F. Zhang, and Q. Du, “Hyperspectral image classification using deep pixel-pair features,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 844–853, 2016.

H. Lee and H. Kwon, “Going deeper with contextual cnn for hyperspectral image classification,” IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4843–4855, 2017.

S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, “Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 8, pp. 4520–4533, 2017.

L. Yang, S. Yang, P. Jin, and R. Zhang, “Semi-supervised hyperspectral image classification using spatio-spectral laplacian support vector machine,” IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 3, pp. 651–655, 2013.

I. Dópido, J. Li, P. R. Marpu, A. Plaza, J. M. B. Dias, and J. A. Benediktsson, “Semisupervised self-learning for hyperspectral image classification,” IEEE transactions on geoscience and remote sensing, vol. 51, no. 7, pp. 4032–4044, 2013.

J. Cao and B. Wang, “Embedding learning on spectral–spatial graph for semisupervised hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1805–1809, 2017.

F. Cao, Z. Yang, J. Ren, W. Chen, G. Han, and Y. Shen, “Local block multilayer sparse extreme learning machine for effective feature extraction and classification of hyperspectral images,” IEEE Trans. Geoscience and Remote Sensing, vol. 57, no. 8, pp. 5580–5594, 2019.

B. Cui, X. Xie, S. Hao, J. Cui, and Y. Lu, “Semi-supervised classification of hyperspectral images based on extended label propagation and rolling guidance filtering,” Remote Sensing, vol. 10, no. 4, p. 515, 2018.

L. Wang, S. Hao, Q. Wang, and Y. Wang, “Semi-supervised classification for hyperspectral imagery based on spatial-spectral label propagation,” ISPRS journal of photogrammetry and remote sensing, vol. 97, pp. 123–137, 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in Proceedings of International Conference on Learning Representations, 2018.