Mixed Residual Convolutions with Vision Transformer in Hyperspectral Image Classification

Ying Cao
School of Communications and Electronics
Jiangxi Science and Technology Normal University
Nanchang, China
Candy_2000_0108@163.com

Yang Wang
School of Communications and Electronics
Jiangxi Science and Technology Normal University
Nanchang, China
yangwang0620@163.com

Zhijian Yin
School of Communications and Electronics
Jiangxi Science and Technology Normal University
Nanchang, China
zhijianyin@aliyun.com

Zhen Yang
School of Communications and Electronics
Jiangxi Science and Technology Normal University
Nanchang, China
yangzhenphd@aliyun.com

2022 IEEE 22nd International Conference on Communication Technology (ICCT) | 978-1-6654-7067-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCT56141.2022.10073347

Abstract—**In recent years, deep learning methods represented by convolutional neural networks (CNN) have gradually become a research focus in the field of hyperspectral image classification (HSI). Although the proposed CNN-based methods have the advantages of spatial feature extraction, they are difficult to handle the spectral feature. We propose a mixed residual convolutions with Vision Transformer model (MRViT), which uses ViT Transformer (ViT) to overcome this limitation. First, using principal component analysis (PCA) and a channel shift strategy to construct a module to process HSI data. Then, a mixed residual convolution is constructed to extract spectral-spatial features. Finally, the important feature information is enhanced by the ViT model. In this paper, four datasets are used for experimental analysis, which confirms that the performance of MRViT model is superior to other HSI classification models. The research content of this paper enriches related research in HSI classification, and also provides certain reference value for subsequent research.**

Index Terms—**convolutional neural network (CNN), hyperspectral image (HSI) classification, principal component analysis (PCA), channel shift strategy, mixed residual convolutions, Vision Transformer (ViT)**

I. Introduction

Hyperspectral remote sensing technology is a remote sensing information acquisition technology developed based on imaging spectroscopy. It has extensive and important applications in the fields of atmospheric detection, such as space remote sensing, earth resources survey, military reconnaissance, environmental monitoring, agriculture and ocean remote sensing. HSI classification is an important method for extracting information from hyperspectral images. This method uses principal component analysis (PCA) [1] in data processing, then divide each pixel in the image into different ground object categories by the corresponding criterion and the spectral-spatial information. So people can grasp the spatial distribution rule of ground objects and achieve more increasingly higher levels of application.

In recent years, many deep learning classification methods have appeared in HSI classification. Especially, many methods in HSI data processing and HSI classification have been completely changed by CNN-based models. CNN-based model has the characteristics of local connection and shared weight, which have the advantages of spatial feature extraction.

According to the input information of the model, the HSI classification methods based on CNN can be divided into three types: the CNN model based on spectral information, the model based on spatial information and the model based on spectral-spatial information. The model based on spectral feature information like 1D-CNN [2], which receives spectral feature vector as model input and classifies hyperspectral image pixels after model. Due to HSI also have rich spatial information, the model based on spatial feature information, such as 2D-CNN [2], which uses spatial information in hyperspectral image classification; The method based on spectral-spatial feature information, like 3D-CNN [3], which combines spectral-spatial information in a stereo as input and classifies the pixels after 3D-CNN. Based on the above three models, researchers proposed a mixed 2D-3D model [4] for the application of hyperspectral image classification.

Although these CNN-based methods have the advantages of spatial feature extraction, they are difficult to handle the spectral feature. In order to solve this issue, Transformer is introduced as feature extraction modules in CNN-based models. On the basis of these model, SpectralFormer [5], Multimodal Hyperspectral Unmixing [6], Convolutional Transformer Network [7], Spatial-Spectral Transformer [8], spectral–spatial connected attention mechanism [9] and other attention models have also been applied to hyperspectral image classification, and these model methods have achieved excellent performances in classification.

Originally, the Transformer was mainly used in the field of natural language processing (NLP) [10]. Last year, a new model named ViT [11] performed well in the field of im-

978-1-6654-7067-4/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on May 04,2023 at 07:16:39 UTC from IEEE Xplore. Restrictions apply.

age classification, which could effectively capture and utilize the space information between adjacent image patches. The Transformer could establish global dependencies between the sequences of input vectors. Nowadays, some researchers also try to apply the model into HSI classification, but most of their works are based on the improved method in feature extraction module.

We proposed MRViT model in this article. In order to enhance the feature extraction capability in mixed residual convolution module, we used the channel shift strategy after PCA and used ViT model to enhance characters representation after convolution module.

The main contributions of this article are summarized as follows.

- we proposed MRViT model, an end-to-end network model, which not only used a mixed residual convolutions, but also combines the ViT model. The mixed residual convolution module could effectively extract more spectral-spatial feature information, so that more distinguishing features can be extracted and used later. The ViT model could well capture the relationship between spectral sequences in HSI pixels, which can enable MRViT to acquire a excellent performance in experiment.
- In the design of the mixed convolution model, we combine the residual network idea. Thus, the phenomenon of decreasing accuracy is effectively alleviated and the overall accuracy in hyperspectral images classification is improved.
- Inspired by the Convolutional Block Attention Module [12] and MCNN [15], before the feature information enters the mixed residual convolution module, a channel shift strategy is adopted to emphasize some spectral bands, which can help the latter module to extract more spectral information.

II. Related Work

A. Principal Component Analysis

The function of PCA is to map data into low-dimensional subspace through linear transformation. To be specific, PCA seeks to maximize the internal information of data after dimensionality reduction, and judge the importance of the projected direction by measuring the data variance. Its mathematical process can be simply expressed as follows. W is defined as a matrix composed of column vectors containing all feature mapping vectors, which can retain information in the data well. Meanwhile, an optimized objective function can be obtained by algebraic linear transformation of the matrix, the function can be seen as follow:

$$\min_{w} \operatorname{tr}\left(\mathbf{W}^{\mathbf{T}}\mathbf{A}\mathbf{W}\right), \qquad \text{s.t. } \mathbf{W}^{\mathbf{T}}\mathbf{W} = \mathbf{I} \tag{1}$$

Where tr is the trace of the matrix, and A is the covariance matrix. Therefore, the output data of PCA can be represented by $\mathrm{Y} = \mathrm{W}^{\mathrm{T}}\mathrm{X}$, and the optimal matrix W is composed of the eigenvectors corresponding to the largest n eigenvalues in front of the data covariance matrix as column vectors, thus reducing the original dimension of X to the n dimension.

B. 3D and 2D convolutions

The mixed residual convolution module including 3D and 2D convolutions. The 3D and 2D convolution formula in the neural network, which is defined as follows:

$$\begin{aligned} f_{i,j}^{x,y,z} &= \Phi\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{i,j,m}^{p,q,r} f_{(i-1)m}^{(x+p)(y+q)(z+r)} + b_{i,j}\right) \\ f_{i,j}^{x,y} &= \Phi\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{i,j,m}^{p,q} f_{(i-1)m}^{(x+p)(y+q)} + b_{i,j}\right) \end{aligned} \tag{2}$$

In Eq.(2) , $f_{i,j}^{x,y,z}$ represents the output variable at the position of $(\mathrm{x},\mathrm{y},\mathrm{z})$ of the jth feature graph of ith layer, where $\Phi(\bullet)$ is the activation function and m is the feature cube related to the jth feature cube in the (i−1)th layer. P_i , Q_i and R_i represent the height, width and channel number of the 3D convolution kernel respectively. In this case, R_i stands for spectral dimension. $\mathcal{W}_{i,j}^{p,q,r}$ is the value of position weight parameters $(\mathrm{p},\mathrm{q},\mathrm{r})$ connected to the mth feature graph, and $\mathrm{b}_{i,j}$ is the deviation of the jth feature graph in the ith layer.

In brief, although 2D convolution can extract spatial feature information, it cannot obtain significant feature information in continuous spectral bands. While 3D convolution can simultaneously extract spatial and spectral feature information, it requires more computational costs. Therefore, neither 2D convolution nor 3D convolution alone to extract hyperspectral image features is not the best choice.

C. The idea of residual module

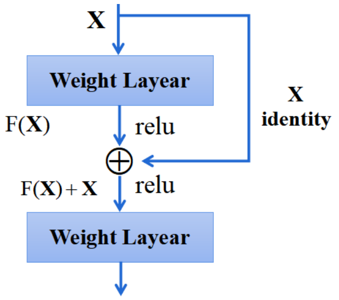

Fig. 1. The residual idea of MRViT. The residual idea is also known as a jump connection.

In the residual network module shown in Fig. 1, we can transform the problem of eigenvector processing into solving the residual mapping function F of the network, which is also known as a jump connection. In Eq.(3), $\mathrm{H}(\mathrm{X})$ represents the mapping value, which can also be understood as the observation value, and X is the feature map of the ResNet [13] output of the previous layer.

$$\mathrm{F}(\mathrm{x}) = \mathrm{H}(\mathrm{x}) - \mathrm{x} \tag{3}$$

Provided that the dimension of the jump join X is inconsistent with the dimension of the residual mapping result $\mathrm{F}(\mathrm{X})$,

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on May 04,2023 at 07:16:39 UTC from IEEE Xplore. Restrictions apply.

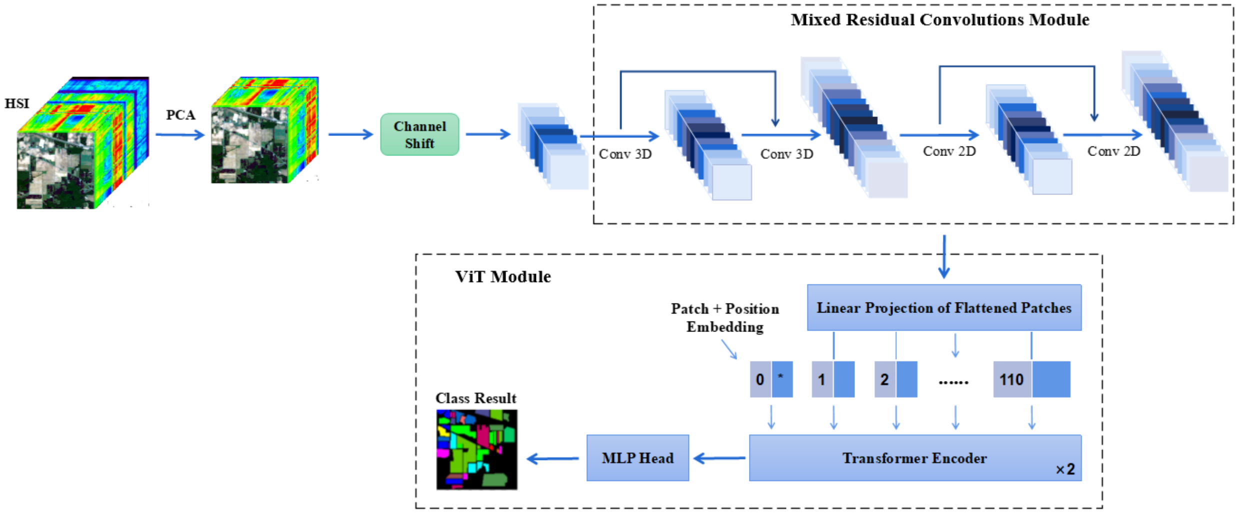

Fig. 2. Architecture of the proposed MRViT Model. The model including three modules: data processing module with PCA and channel shift, the mixed residual convolution module and the ViT model

it is important to note that they cannot be added directly. If we want them to have the same dimension, we should raise the dimension of X , then calculate.

III. Method

The Fig. 2 shows the overall framework of the MRViT model, the model is composed of the PCA operation, the channel shift strategy, the mixed convolution module and the ViT module.

A. PCA and Channel Shift

Before the feature information enters the mixed convolution module, some important features are extracted by PCA and channel shift operations. PCA reduces the dimension of spectral band and maintains the integrity of spatial information as much as possible.

As mentioned above, PCA measures the importance of each direction by comparing the variance of the data in the projection space. As we all know, the greater variance of the data, the more information it contains. Therefore, we can infer that the feature information of the spectral bands in the HSI data is also arranged in descending order after the PCA operation. The bands with more retained spectral information contribute more to the subsequent feature extraction process. In other words, we can determine the relatively more important spectral bands by sorting, which is the essential characteristic of PCA.

Based on the sorting characteristic of PCA and the edge effect in the convolution process, a channel shift operation uses in the data when the data after dimensionality reduction. The core idea of channel shift is to move the important spectral channel to the central position of the data. Conversely, the less important spectral channel can be placed at the edge of the data. The process of channel shift is shown in the Fig. 3 This strategy can increase the number of effective extraction of spectral features in convolution module and keep the important spectral channels in the central position of the effective convolution domain. After the operation, which can make the mixed residual convolution module extract abundant spectral-spatial feature information.

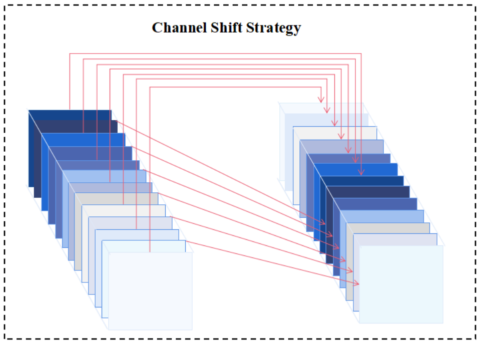

Fig. 3. The idea of channel shift strategy. The deeper color in the central position denotes the abundant information in the spectral channel.

B. Mixed Residual Convolution Module

The mixed residual convolution module of this paper uses two convolution methods, including 3D convolution and 2D convolution. Since the 2D-CNN model mainly considers the spatial correlation of each channel in a hyperspectral image, it can extract the local spatial feature information of each pixel. The 3D-CNN model also can utilize the correlation information between different channels, it can simultaneously extract the spatial-spectral information of HSI data to improve the capability of feature representation.

In this paper, we proposed a feature extraction module which combines 3D convolutions and 2D convolutions, and the residual idea is integrated into the model. In general, if the number of model network layers are increasing, the feature information extracted by the model will also be richer. However, with the deepening of network layer, the optimization

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on May 04,2023 at 07:16:39 UTC from IEEE Xplore. Restrictions apply.

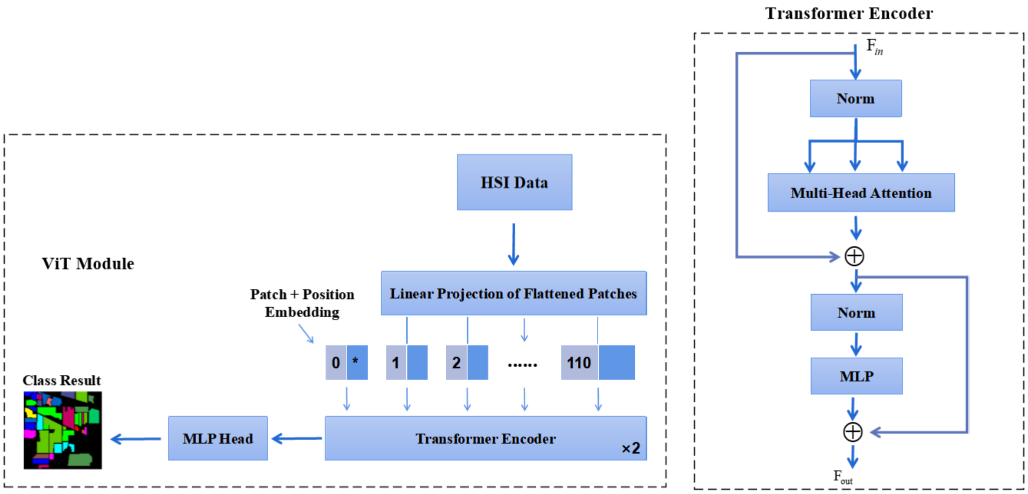

Fig. 4. Architecture of the ViT Model. The model is consists of three modules: Linear Projection of Flattened Patches module, Transformer Encoder module, MLP Head module.

effect of training model will become increasingly worse, even the accuracy of training data and test data will continue to decrease. The emergence of residual idea effectively alleviates the phenomenon of accuracy decline, and improvs the overall accuracy in hyperspectral images classification.

C. *Vision Transformer*

Transformer was originally proposed for NLP, while ViT is an attempt to apply Transformer to Computer Vision. The model framework of ViT can be seen in Fig. 4, which is consists of three modules: Linear Projection of Flattened Patches module, Transformer Encoder module and MLP Head module.

The working process of Vision Transformer can be summarized as follows. To begin with, a picture is divided into Patches with a given size. As the data input into Transformer module is vector sequence, it needs to be transformed by embedding module. Then, in front of the vector sequence, adding $[\mathbf{class}]$ vector and taking Position Embedding operation. When the vector sequence passing the Transformer Encoder module, only need to extract the corresponding results of $[\mathbf{class}]$ vector, and the MLP Head module can get the final class result.

Transformer Encoder is the most important module in ViT. Location embedding is used to mark the location information of each semantic marker before the data enters the module. Each vector sequence can be represented by $[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots, \mathbf{t}_n]$, and a classification vector $\mathbf{t}_0^{\text{class}}$should be added to connect the vector sequence before entering Encoder, which is used to perform the final classification task. Then, position coding is carried out for these vector sequences, and finally the vector sequences of semantic markers are expressed as follows:

$$\mathbf{T}_{\text{input}} = \left[\mathbf{t}_0^{\text{class}}, \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\right] + \text{PE} \tag{4}$$

The Transformer Encoder module is used to establish deep relationships between semantic markers. It contains MSA, MLP and two normalized layers, with jumping connections designed before the MSA block and MLP layer. MSA module can effectively capture the correlation between feature sequences, and it combines from different heads in order to learn more feature information. In the calculation of this module, involving three weight matrices W_Q, W_K and W_V, so the corresponding sequences of Q, K and V can be obtained after matrix transformation. In MSA module, each head part involves the calculation of three weight matrices, and the same formula is used to calculate the attention value of each head. Finally, the attention results of each head are connected to form the final attention value. This process can be expressed by the following equation:

$$\mathbf{A} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

$$\begin{gathered}\text{MSA}(\mathbf{A}, \mathbf{K}, \mathbf{V}) = \\ \text{Concat}\left(\mathbf{SA}_1, \mathbf{SA}_2, \mathbf{SA}_3, \ldots, \mathbf{SA}_n\right)\mathbf{W}\end{gathered} \tag{6}$$

Where d_k refers to the dimension of the vector sequence K, n is the total number of heads, W is the parameter matrix.

Finally, the multi-attention results are input into the MLP module for further processing. The MLP module consists of two fully connected layers, between which have an activation function (GELU).

IV. Experiment and Analysis

In order to verify the feasibility of the model, four HSI data sets are selected for the experiment, which are Indian Pines, Pavia University, Salinas and KSC. The proposed model is also compared with other hyperspectral classification models, which are SVM [14], 2D-CNN [2], 3D-CNN [3], MCNN [15], DPRN [16] and SSTFF [17].

During the Training, in order to ensure the reliability of the comparison effect, we set the patch size to 13, reduce the dimension of PU after PCA into 100, and the dimension of other datasets after PCA into 110. In addition, batch size , the

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on May 04,2023 at 07:16:39 UTC from IEEE Xplore. Restrictions apply.

learning rate and training epochs are set to 64, 0.001 and 300. In the process of training, the training ratio of these datasets are set to 0.05, 0.1, 0.005 and 0.2. In order to evaluate the performance of different classification methods, we adopted three well-known numerical indicators that is overall accuracy (OA), average accuracy (AA) and Kappa coefficient (Kappa) to evaluate the classification results. OA represents the ratio between the number of correctly classified samples to the total test samples, AA represents the average of accuracies in all classes, and Kappa is an available measure of agreement between the ground truth map and classification map. The experimental results of all models, which using four data sets, are shown in the following tables.

TABLE I
CLASSIFICATION ACCURACY (%) AND SCORE OF OA,AA,KAPPA FOR INDIAN PINES

	SVM	2D-CNN	3D-CNN	MCNN	DPRN	SSTFF	MCVT
OA	78.99±2.66	93.79±1.45	95.73±2.43	96.00±1.24	96.17±1.37	96.39±0.64	**98.22±0.47**
AA	73.27±1.24	94.69±2.53	**96.22±1.58**	82.93±5.51	92.00±1.42	92.68±0.51	93.39±0.73
Kappa	76.77±2.11	93.10±3.16	95.40±1.84	95.42±0.97	95.82±1.64	95.88±0.73	**97.97±0.62**

TABLE II
CLASSIFICATION ACCURACY (%) AND SCORE OF OA,AA,KAPPA FOR PAVIA UNIVERSITY

	SVM	2D-CNN	3D-CNN	MCNN	DPRN	SSTFF	MCVT
OA	82.17±1.69	93.50±2.61	89.36±1.48	95.28±1.48	95.02±1.21	99.42±0.21	**99.56±0.27**
AA	70.93±2.57	92.61±5.14	84.27±2.14	93.93±0.95	90.30±1.77	98.44±0.37	**99.05±0.30**
Kappa	79.41±1.74	91.57±2.44	84.69±3.42	94.36±1.22	93.47±1.48	**99.53±0.24**	99.42±0.37

TABLE III
CLASSIFICATION ACCURACY (%) AND SCORE OF OA,AA,KAPPA FOR SALINAS

	SVM	2D-CNN	3D-CNN	MCNN	DPRN	SSTFF	MCVT
OA	88.64±2.47	88.59±2.94	90.89±1.04	93.37±1.03	95.58±1.08	96.39±0.44	**96.73±0.45**
AA	91.21±1.84	89.57±3.16	92.00±3.26	93.91±1.21	97.32±0.94	**98.63±0.41**	98.09±0.39
Kappa	88.37±2.19	89.42±2.68	89.74±2.17	93.23±1.57	95.31±1.30	**97.23±1.05**	96.36±0.97

TABLE IV
CLASSIFICATION ACCURACY (%) AND SCORE OF OA,AA,KAPPA FOR KSC

	SVM	2D-CNN	3D-CNN	MCNN	DPRN	SSTFF	MCVT
OA	89.10±2.33	89.27±2.08	90.98±1.41	98.07±1.16	97.82±0.84	99.08±0.39	**99.66±0.19**
AA	87.11±1.97	86.66±1.37	87.93±0.36	97.41±1.03	96.50±1.17	98.50±0.65	**99.37±0.21**
Kappa	87.86±2.41	85.89±2.23	90.00±1.19	96.79±2.27	97.57±0.44	99.03±0.27	**99.62±0.17**

V. Conclusion

This article proposes a MRViT model for improving the performance of HSI classification. The model integrates a mixed residual convolution module and ViT structure organically. The mixed residual convolution module can make full use of spectral information and spatial information when the HSI data do the PCA and channel shift operations. Such operations make the analysis of data characteristics more sufficient. The ViT model can better capture the relationship between spectral sequences vectors. The experimental results demonstrate that our proposed network architecture can effectively enhance the accuracy of HSI classification.

In the future, based on the MRViT model, we will study an end-to-end transformer network to solve some shortcomings, such as feature enhancement and feature fusion, and extract abundant spatial-spectral features, thereby further improving the classification accuracy. In addition, we will make effort to find better strategies for more universal applications of our model.

Acknowledgment

This work was supported by the Jiangxi Provincial Natural Science Foundation (20202BABL202014), the National Natural Science Foundation of China (62061019), the key Project of Jiangxi Education Department (GJJ201107), and the Key Laboratory of System Control and Information Processing, Ministry of Education (Scip202106).

References

[1] Kang, X., Xiang, X., Li, S., Benediktsson, J. A. (2017). PCA-based edge-preserving features for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(12), 7140-7151.

[2] Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 54(10), 6232-6251.

[3] Hamida, A. B., Benoit, A., Lambert, P., Amar, C. B. (2018). 3-D deep learning approach for remote sensing image classification. IEEE Transactions on geoscience and remote sensing, 56(8), 4420-4434.

[4] Roy, S. K., Krishna, G., Dubey, S. R., Chaudhuri, B. B. (2019). HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 17(2), 277-281.

[5] Hong,D. , Han, Z. , Yao, J. , Gao, L. , Zhang, B. , Plaza, A. , et al. (2021). Spectralformer: rethinking hyperspectral image classification with transformers.

[6] Han, Z. , Hong, D. , Gao, L. , Yao, J. , Zhang, B. , Chanussot, J. . (2022). Multimodal hyperspectral unmixing: insights from attention networks. IEEE Transactions on Geoscience and Remote Sensing, 60.

[7] Zhao, Z., Hu, D., Wang, H., Yu, X. (2022). Convolutional Transformer Network for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

[8] He, X., Chen, Y., Lin, Z. (2021). Spatial-Spectral Transformer for Hyperspectral Image Classification. Remote Sensing, 13(3).

[9] Guo, W., Ye, H., Cao, F. (2022). Feature-Grouped Network With Spectral–Spatial Connected Attention for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-13.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[12] Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[13] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[14] Melgani, F., Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on geoscience and remote sensing, 42(8), 1778-1790.

[15] Zheng, J., Feng, Y., Bai, C., Zhang, J. (2020). Hyperspectral image classification using mixed convolutions and covariance pooling. IEEE Transactions on Geoscience and Remote Sensing, 59(1), 522-534.

[16] Paoletti, M. E., Haut, J. M., Fernandez-Beltran, R., Plaza, J., Plaza, A. J., Pla, F. (2018). Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 57(2), 740-754.

[17] Sun, L., Zhao, G., Zheng, Y., Wu, Z. (2022). Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on May 04,2023 at 07:16:39 UTC from IEEE Xplore. Restrictions apply.