Multidimensional Information Expansion and Processing Network for Hyperspectral Image Classification

Zhen Yang, *Member, IEEE*, Ying Cao, Tao Zhang, *Member, IEEE*, Weiwei Guo, *Member, IEEE*, and Zenghui Zhang, *Senior Member, IEEE*

***Abstract*— In recent years, deep learning (DL) has been extensively used in the hyperspectral image (HSI) classification. The representative method is the convolutional neural network (CNN). However, due to the limitations of its inherent network backbone, CNNs still easily fail to mine some important information about HSIs, such as the sequence attributes of spectral signatures. To deal with this problem and make full use of the spectral–spatial information of HSIs, we propose a novel network named multidimensional information expansion and processing network (MIEPN) for HSI classification, which is mainly composed of one information expansion module (IEM), one feature information expansion and extraction module (FEEM), and one vision transformer (ViT) module. Briefly speaking, IEM expands and fuses HSI information in a 3-D space, yet FEPM pays more attention to digging deeper information. After these, the extracted information is input into the ViT module for HSI classification. Experiments carried out on several typical datasets demonstrate that the proposed network MIEPN can provide competitive results compared to the other state-of-the-art CNN-based methods.**

***Index Terms*— Feature expansion, feature extraction, hyperspectral image (HSI) classification, multidimensional information expansion and processing network (MIEPN), vision transformer (ViT).**

I. Introduction

HYPERSPECTRAL images (HSIs), collected in different and continuous spectral bands, belong to the category of 3-D data. Compared to traditional RGB images, HSIs contain more abundant information, enabling models better to perform some tasks, such as object detection, change detection, and geological exploration. Within early works, machine-learning-based methods, like support vector machines (SVMs) [1], are mainstream and usually applied to HSI classification. However, these methods cannot effectively obtain the deep representation of features.

Different from traditional machine learning, deep learning (DL) proposed in recent years holds a more powerful capability in feature extraction and automatic end-to-end training. Particularly, because of their better performance in image classification target detection, and so on, convolutional neural networks (CNNs) are popular in the DL family. Also, various CNN-based methods have been developed for HSI classification so far. For example, Chen et al. [2] developed a network model that stacked several 1D-CNNs, which took advantage of the spectral information for HSI classification. Unlike 1D-CNN, 2D-CNN is proposed to extract the spatial information for HSI classification. For instance, Paoletti et al. [3] developed a network named DPRN for HSI classification, wherein the residual block was utilized as well. Sun et al. [4] proposed a network FCSN to explore spectral–spatial features for HSI classification. Unfortunately, neither 1D-CNN nor 2D-CNN can simultaneously use the spectral and spatial information for HSI classification.

Actually, the joint use of spatial and spectral information has already been demonstrated to be more beneficial for HSI classification than their use independently. For example, Zhong et al. [5] proposed the spectral–spatial residual network (SSRN) to improve the accuracy of HSI classification. In recent works, some scholars have tried to fuse 2-D convolution and 3-D convolution together for HSI classification. Zheng et al. [6] proposed a mixed CNN with the covariance pooling, named MCNN-CP. Yang et al. [7] constructed a new cross-mixing residual network (CMR-CNN) for HSI classification, which adopted the 3-D residual structure and 2-D residual structure together to extract the features of HSIs. Recently, the attention mechanism has also been introduced into networks to optimize the discrimination of extracted features to improve the learning performance of CNNs. For instance, Li et al. [8] developed a double-branch dual-attention mechanism network (DBDA) to extract plenty of spectral–spatial features for HSI classification. Zheng et al. [9] proposed a rotation-invariant attention network (RIAN) to alleviate the problem that the

Manuscript received 3 July 2023; revised 28 August 2023; accepted 13 September 2023. Date of publication 18 September 2023; date of current version 5 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62261026, Grant 62201343, Grant 62071333, Grant 62271311, and Grant 62003251; in part by the General Project of Jiangxi Natural Science Foundation under Grant 20232ACB212006; in part by the Key Laboratory of System Control and Information Processing, Ministry of Education under Grant Scip202106; in part by the Shanghai Key Laboratory of Navigation and Location Based Services under Grant SKLNLBS2023001; in part by the Double First-Class Construction Foundation under Grant WH220503031; and in part by the European Space Agency (ESA)-Ministry of Science and Technology of China (MOST) Dragon 5 Program under Grant 58190. *(Corresponding author: Tao Zhang.)*

Zhen Yang is with the School of Communication and Electronics, Jiangxi Science and Technology Normal University, Nanchang 330000, China, and also with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China.

Ying Cao is with the School of Communication and Electronics, Jiangxi Science and Technology Normal University, Nanchang 330000, China.

Tao Zhang and Zenghui Zhang are with the Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhangtao8902@sina.cn).

Weiwei Guo is with the Center for Digital Innovation, Tongji University, Shanghai 200092, China.

Digital Object Identifier 10.1109/LGRS.2023.3316796

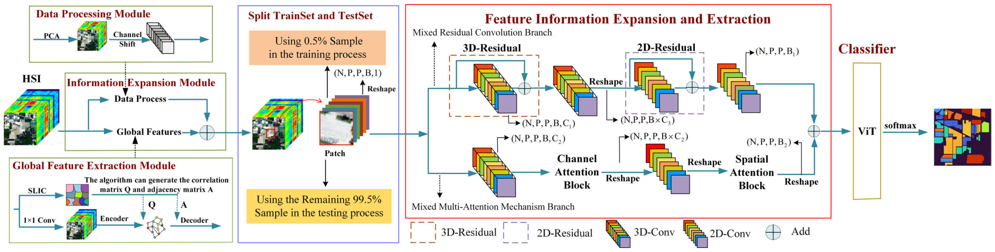
Data Processing Module
PCA
Channel Shift
HSI
Information Expansion Module
Data Process
Global Features
Global Feature Extraction Module
The algorithm can generate the correlation matrix Q and adjacency matrix A
SLIC
Q
A
1×1 Conv
Encoder
Decoder
Split TrainSet and TestSet
Using 0.5% Sample in the training process
(N, P, P, B, 1)
Reshape
Patch
Using the Remaining 99.5% Sample in the testing process
Feature Information Expansion and Extraction
Mixed Residual Convolution Branch
3D-Residual
2D-Residual
Reshape
(N, P, P, B, C₁)
(N,P,P,B×C₃)
(N, P, P, B₁)
(N, P, P, B, C₂)
(N, P, P, B×C₂)
Channel Attention Block
Reshape
Reshape
Spatial Attention Block
Reshape
(N, P, P, B₂)
Mixed Multi-Attention Mechanism Branch
Classifier
ViT
softmax
3D-Residual
2D-Residual
3D-Conv
2D-Conv
Add

Fig. 1. Architecture of the proposed network MIEPN, including IEM, FEEM, and ViT module.

Algorithm 1 Global Feature Extraction Method

```
Input: HSI $x \in R^{H \times W \times B}$. H, W, and B respectively represent the length of HSI, the width of HSI, the spectral number of HSI.
Output: Feature Vector
Initialize: the segmentation scale $\lambda$
1. Utilize the SLIC to partition the HSI into $P = (H \times W)/\lambda$ superpixels.
2. Construct the mapping association matrix Q between pixels and superpixels by $Q_{i,j} = \begin{cases} 1, & \text{if } \hat{S}_i \in G_i \\ 0, & \text{otherwise} \end{cases}$ $\hat{S} = \text{Flatten}(S)$, $G_i$ is the ith superpixel, $\hat{S}_i$ is the ith pixels in $G_i$.
3. Calculate the centroid of each superpixel by $F = \text{Encode}(S, Q) = \hat{Q}\,\text{Flatten}(S)$ to get the undirected graph nodes.
4. Construct the adjacency matrix A in superpixels by $A_{ij} = \begin{cases} 1, & \text{if } G_i \text{ and } G_j \text{ are adjancent} \\ 0, & \text{otherwise} \end{cases}$
5. Seed the undirected graph and the adjacency matrix A into GCN to get the features.
6. Assign the features of nodes to pixels by $S^* = \text{Decode}(F; Q) = \text{Reshape}(QF)$.
Note that the step 6 can fuse the global feature into pixel-level.
```

convolution is sensitive to the spatial rotation of inputs. Unfortunately, the attention mechanism cannot directly capture the spectral–spatial features on superpixel levels. Comparatively, the graph convolutional network (GCN), working on the superpixel-based nodes of HSIs, can solve this problem [10]. Liu et al. [10] proposed a heterogeneous deep network called CEGCN, in which CNN and GCN branches generated complementary spectral–spatial features at pixel and superpixel levels, respectively. Besides GCN, another technology called vision transformer (ViT) [11] is applied to HSI classification as well. For example, Sun et al. [12] constructed a spectral–spatial feature tokenization transformer (SSFTT) network for HSI classification.

Inspired by the above methods, in this letter, we propose a new network named multidimensional information expansion and processing network (MIEPN) to realize HSI classification, which consists of one information expansion module (IEM), one feature information expansion and extraction module (FEEM), and one ViT module, as shown in Fig. 1.

II. Methodology

In this section, the details of the proposed network MIEPN will be introduced and discussed carefully.

A. Information Expansion Module

As the HSI data contain redundant information, the CNN network is unable to extract more discriminative information for HSI classification. To deal with this problem, we hereafter design an IEM to reduce the redundant information and expand some discriminative information for HSIs. It is an HSI preprocessing method, composed of two submodules: the data-processing module and the global feature extraction module.

1) Data-Processing Module: It retains and emphasizes the important spectral information of HSI data. Since the original data has a lot of redundant information, in this module, PCA is adopted to retain the most important spectral information for HSI classification. Due to the sorting characteristic of PCA, the edge effect is prone to occur in the convolution process of feature extraction. It has proven that the channel shift strategy can make the CNN model more effective by using the processed data for HSI classification in [6], so we also adopt this strategy in this module.

2) Global Feature Extraction Module: Due to the large number of pixels in HSIs, many works have shown that considering each pixel as a node of a graph leads to a huge computational overhead and limits the applicability of the method. To solve this problem, we use the superpixels instead of pixels as nodes. To supplement the information in HSIs more comprehensively, we establish a module with GCN [10] to extract the global feature information of the superpixel level. The implementation details of the global feature extraction module are shown in Algorithm 1. The global feature extraction module extracts global information on HSIs as a feature expansion method, which processes HSIs in parallel with the data-processing module. Thus, it can be inferred that IEM integrating the spectral–spatial information at pixel- and superpixel-level simultaneously must be more conducive for HSI classification, whose effectiveness will be verified later.

B. Feature Information Expansion and Extraction Module

Generally speaking, 3-D convolution is used to extract spectral information, while 2-D convolution is used to extract spatial information. Moreover, the CNN-based methods, which only process HSIs with simple convolution operations, easily cause the ignorance of information in spectral and spatial space. For boosting performance, the FEEM module is designed to extract more discriminative information at the pixel level. The FEEM module consists of two branches. One is the mixed residual convolution branch, and the other is the mixed multiattention mechanism branch.

1) Mixed Residual Convolution Branch: Traditional CNN-based models, like 3D-CNN and 2D-CNN, could ignore some important features of HSIs when they are individually used in HSI classification. In general, if the number of network layers increases, the feature information extracted by models will also be richer. However, with the deepening of network layers, the optimization effect of the network could be worse, even if the accuracy keeps dropping. The emergence of residual structure effectively alleviates this phenomenon. Inspired by MCNN-CP and DPRN, we propose a novel branch that combines one 3-D residual structure and one 2-D residual structure together for HSI classification.

The change process of the patch data is presented in the red box of Fig. 1. In detail, N is the number of training samples, P represents the size of data, B is the number of spectral bands, the 3-D convolution kernels are, respectively, set to (32 × 3 × 3 × 3), (64 × 3 × 3 × 3), and (32 × 3 × 3 × 3), where (32 × 3 × 3 × 3) represents that the value of C_1 in the feature data is 32 and the convolution kernel (3 × 3 × 3) sequentially acts on the input feature maps to perform dot product with their weights and deviations. Similarly, the value of C_2 is 32 and C_3 is 64. The feature maps obtained by the 3-D convolution structure cannot be directly used as the input of the 2-D residual structure due to their different dimensions, so we should reshape the size of the path data obtained by the 3-D convolution. Here, we set the value of the 2-D convolution kernels into (64 × 3 × 3), (110 × 3 × 3), and (64 × 3 × 3), respectively. It also represents that the value of B_1 is 110 and the B_2 is 64 in the feature data. So, the dimension of the feature that fits into the ViT module is 174.

The 3-D and 2-D convolution formulas in the network are

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

where $f_{i,j}^{x,y,z}$ represents the output variable at the position of (x, y, z) of the jth feature graph of the ith layer, where $\Phi(\cdot)$ 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 3-D convolution kernel, respectively. In this case, R_i stands for spectral dimension. $w_{i,j}^{p,q,r}$ is the value of position weight parameters (p, q, r) connected to the mth feature graph, and $b_{i,j}$ is the deviation of the jth feature graph in the ith layer.

When the 3-D residual structure acts on the input data, the corresponding output could be denoted as

$$F_{i+1} = F_i + \varphi(F_i, W_i)$$
$$\varphi(F_i, W_i) = \Phi(W_{i-1}\cdot \mathrm{BN}(F_{i-1})) + b_{i-1} \quad (2)$$

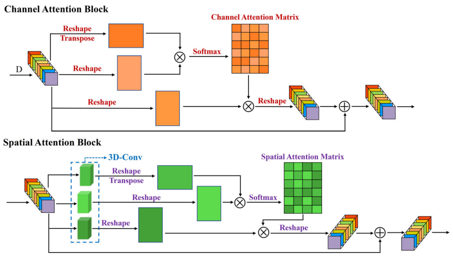

Fig. 2. Structure of channel attention and spatial attention mechanism.

where F_{i+1} is the output with the jth layer, $\varphi(\cdot)$ is the 3-D residual part, and W_{i-1} is determined by the convolution kernel of residual module. $\mathrm{BN}(\cdot)$ is the Batch normalization.

Similarly, the 2-D residual structure is expressed as

$$T_{i+1} = T_i + \sigma(F_i, W_i)$$
$$\sigma(F_i, W_i) = \Phi(W_{i-1}\cdot \mathrm{BN}(T_{i-1})) + b_{i-1} \quad (3)$$

where T_i represents the input in the ith layer of 2-D residual module $\sigma(\cdot)$, and W_{i-1} is the weight determined by the convolution kernel of 2-D residual module.

2) Mixed Multiattention Mechanism Branch: In general, 3-D convolution is used to extract spectral information, while 2-D convolution is used to extract spatial information. However, one shortcoming of CNNs is that all the spatial pixels and spectral bands own equivalent weights in the spatial and spectral domains. Obviously, different spectral bands and spatial pixels make different contributions to feature extraction. It has proven that the attention mechanism is a powerful technique to deal with this problem in [8], where the attention mechanism is named DBDA. Inspired by DBDA, we developed another branch in FEEM, wherein one 3-D convolution, one 2-D convolution, one channel attention, and one spatial attention are utilized to capture plenty of spectral and spatial features. The details of attention mechanisms are illustrated in Fig. 2. Without loss of generality, hereinafter, we depict the advantage of FEEM against the traditional feature extraction structure of CNN-based methods. In contrast to the previous CNN-based methods, the 3-D convolutions are utilized in these two branches, as a feature expansion method to expand the feature information of patch data, so that the data after convolutions can retain more feature information. It also combines a mixed multiattention mechanism to help the module obtain stronger spectral–spatial features. Therefore, compared to the traditional feature extraction module, the proposed module FEEM enables the model to get higher classification accuracy.

C. ViT Module

Traditional CNN-based models still have some problems. First, it is difficult for them to take into account global information in HSIs. Second, it is difficult for a patch-based classification framework to achieve a fast calculation process in HSI classification. Finally, it is difficult for CNNs to build a lightweight and efficient classification network [13]. Inspired by ViT, we focus on the transformer to overcome the above problems. In detail, we introduce the ViT model as the classifier of MIEPN, which only adopts two transformer encodes to capture the relationship between spectral sequences of patch data. The module framework is shown in Fig. 3.

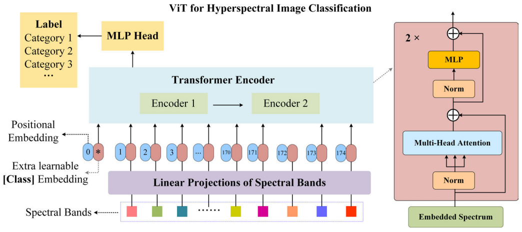

Fig. 3. ViT module of MIEPN, including two transformer encoders.

TABLE I

IMPACT OF DIFFERENT NUMBERS IN PCA FOR THE OA VALUES OF HSI CLASSIFICATION RESULTS

Datasets \ PCA	120	110	100	90	80
IP	70.45	**71.47**	70.79	59.39	67.18
PU	-	-	91.44	**92.23**	92.04
SA	96.02	**96.73**	96.18	95.84	95.96
KSC	**77.33**	68.99	63.81	65.77	68.38

TABLE II

IMPACT OF CONV-LAYERS FOR THE OA VALUES OF HSI CLASSIFICATION RESULTS

Conv-Layers \ Datasets		IP	PU	SA	KSC
3D-Conv	1	42.86	77.43	93.55	62.41
	2	59.77	84.43	94.91	63.04
	3	**62.71**	**86.99**	**95.51**	**65.08**
	4	54.57	84.47	93.75	63.01
2D-Conv	1	71.30	87.64	95.58	65.86
	2	71.8	88.52	95.85	68.91
	3	**74.14**	**94.59**	**96.34**	**70.73**
	4	66.13	84.86	94.58	67.35

III. EXPERIMENTS

A. Datasets and Experimental Setup

To test the feasibility of the model, four datasets are selected for the experiment in this letter, which are Indian Pines (IP), Pavia University (PU), Salinas (SA), and KSC. For the principal component value of PCA, we select it from [120; 110; 100; 90; 80]. The training ratio of these datasets is set to 0.005. Table I shows the OA values obtained under different principal component values. From it, we can see that IP and SA have the best results when the principal component value is 110. The principal component value of KSC is 120, which can get the best results. When the value is 90, PU obtains the best result. So, we set the value of PCA to [110; 90; 110; 120] for these four datasets in this letter.

To seek the best structure of FEEM, we have also done some experiments on the convolution layer. First, for designing the Conv-layer, the number is selected from [1; 2; 3; 4]. The final output channel of 3-D-conv is set to 64 while 2-D-Conv is set to 110 in this letter. The ratio of training is set to 0.005 for all of these four datasets. Table II shows the OA values obtained by different 3-D–2-D layer numbers. It is found that the OA values in these datasets are the best when the number of 3-D layers is 3. Similarly, we can also see that 3 is the best value for the number of 2-D layers.

In addition to reducing the dimension of these datasets, we also set some values of other parameters. During the training, the patch size, the learning rate, and training epochs are set to 13, 0.001, and 200, respectively. Note that the number of categories of each dataset is chosen as the batch size due to its smaller training ratio. We select 0.5% of each class as the training set of these datasets. To prove the effectiveness of our network, nine methods are also selected for comparison, including SVM [1], 2D-CNN [14], 3D-CNN [15], MCNN-CP [6], DPRN [3], DBDA [8], SSFTT [12], CEGCN [10], and CMR-CNN [7]. To quantitatively evaluate the performance of different methods, we adopt the indicators of overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa).

TABLE III

CLASSIFICATION RESULTS BY DIFFERENT METHODS ON THE FOUR DATASETS (USING 0.5% TRAINING SAMPLES)

Methods \ Datasets	IP			PU			SA			KSC		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
SVM [1]	43.59	24.11	32.94	76.00	71.15	74.00	86.68	88.9	85.12	47.25	36.65	39.73
2D-CNN [14]	55.33	30.24	47.61	85.24	80.02	79.94	92.42	93.54	91.54	53.17	41.17	49.98
3D-CNN [15]	48.16	27.42	36.67	76.17	76.11	75.75	92.64	95.03	91.78	57.28	46.89	53.02
MCNN-CP [6]	66.79	52.51	61.98	90.75	83.82	87.68	94.53	95.52	93.91	49.67	39.87	43.30
DPRN [3]	70.95	68.89	64.64	90.95	89.92	90.37	92.75	94.52	91.92	64.35	52.78	55.31
DBDA [8]	67.73	54.49	63.38	76.50	68.64	69.11	92.7	93.27	91.87	64.96	54.89	60.76
SSFTT [12]	71.96	71.30	67.93	94.26	**92.20**	92.20	94.57	91.88	93.96	59.90	51.58	54.53
CEGCN [10]	65.21	48.86	58.96	93.92	89.22	91.84	98.33	98.77	98.15	79.78	73.24	77.41
CMR-CNN [7]	70.49	70.21	65.33	94.34	88.75	92.46	97.43	97.14	97.71	62.27	53.33	57.53
MIEPN	**72.46**	**71.59**	**68.02**	**94.59**	90.93	**93.72**	**99.98**	**99.97**	**99.98**	**86.54**	**79.83**	**85.03**

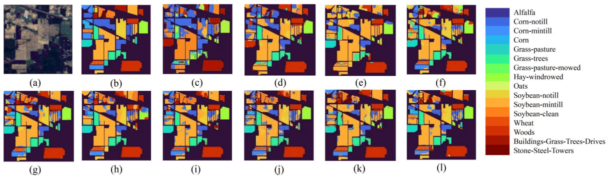

Fig. 4. Classification results of the IP dataset. (a) False color image. (b) Ground truth. (c) SVM. (d) 2D-CNN. (e) 3D-CNN. (f) MCNN-CP. (g) DPRN. (h) DBDA. (i) SSFTT. (j) CEGCN. (k) CMR-CNN. (l) MIEPN.

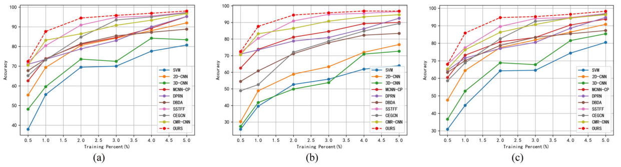

Fig. 5. Classification results of different methods with different training prevents on IP. (a) OA. (b) AA. (c) Kappa.

B. Results and Analyses

Table III displays the results of different methods on these four datasets. Due to space constraints, we here only analyze the IP results, which visual classification maps are also correspondingly shown in Fig. 4. Apparently, from Table III, MIEPN outperforms the other counterparts. Specifically, compared to MCNN-CP, MIEPN improves by 5.67%, 19.08%, and 6.04% on these metrics, respectively. It proves that the structure of residual and ViT are indeed beneficial for HSI classification. Similarly, from Fig. 4, it is easy to see that DPRN has the minimum areas of prediction error than MCNN-CP, which also means the structure of residual convolution can capture deeper features. SSFTT outperforms MCNN-CP in Table III, demonstrating that ViT is useful. In addition, MIEPN achieves 1.79%, 1.38%, and 2.69% improvements over CMR-CNN in terms of these three metrics, showing that ViT and attention mechanisms are really valuable for feature extraction. By comparing MIEPN and SSFTT, we can further demonstrate the effectiveness of the attention mechanism. Meanwhile, Table III reveals that CEGCN outperforms

TABLE IV

Ablation Test: The OA, AA, and Kappa Values of Different Combinations on IP (%) (Using 0.5% Training Samples). Note: The Data-Processing Module Was Added to the Network by Default When the Method Not Use the IEM Module

ViT	Mixed Residual Convolution Branch	IEM	Mixed Multi-Attention Mechanism Branch	OA	AA	Kappa
✓	×	×	×	73.18	53.25	68.69
✓	✓	×	×	74.26	56.46	70.34
✓	×	✓	×	73.58	54.84	69.62
✓	✓	✓	×	74.85	57.50	70.63
✓	×	✓	✓	74.64	55.84	70.62
×	✓	✓	✓	54.43	41.70	47.52
✓	✓	✓	✓	75.76	58.25	72.15

2D-CNN in terms of OA, and MIEPN outperforms DBDA. These results verify that the global information is appropriate for HSI classification. Thus, our method is valuable for a few sample classification issues via using global information and FEEM to extract spectral–spatial features, and ViT to classify. Moreover, Fig. 5 displays the accuracies of different methods on IP with different training ratios. It is obvious that MIEPN always achieves better results regardless of whether the training ratio is high or low. In particular, when the training ratio is 0.005, MIEPN shows a better classification performance in terms of OA, AA, and Kappa, which indirectly reflects that our method can extract more discriminant information with few training samples.

IV. Ablation Study

In this section, we perform some ablation experiments to better find the reason behind MIEPN. In Table IV, compared to the first row, the OA, AA, and Kappa values in the second row are increased by 1.08%, 3.21%, and 1.67%, respectively. It demonstrates that the proposed mixed residual convolution branch is effective in HSI classification. The OA, AA, and Kappa values in the third row are, respectively, 0.4%, 1.59%, and 0.95% higher than those in the first row. Also, it directly proves that the IEM module is valuable in the network. From the fourth row, it is easy to see that the residual structure and the global feature extraction module indeed help the network capture more discriminative information, since the OA, AA, and Kappa values are, respectively, 74.85%, 57.50%, and 70.63%, greater than the third row. The effectiveness of the multiattention mechanism for HSI classification can be easily demonstrated by comparing the third and fifth rows. Seeing the sixth row, one can find the module of ViT is important. Once we neglect it, the OA, AA, and Kappa degrade dramatically. A similar conclusion can be attained by comparing the sixth and seventh rows. The latter has better results, with OA increased by 21.33%, AA increased by 16.55%, and Kappa increased by 24.63%. So, using the ViT network as a classifier in our structure is more suitable for HSI classification. In summary, each module adopted in our method MIEPN is reasonable and effective.

V. Conclusion

In this work, we propose a novel network named MIEPN for HSI classification, which is composed of the IEM, FEEM, and ViT modules. Experiments show that: 1) the classification accuracy can be significantly improved when the global feature extraction module and the feature expansion operation of 3-D-Conv are simultaneously used and 2) residual structures and multiattention mechanism enable the network to extract more discriminative information. Nevertheless, future work is still needed on how to make the model lightweight so that it can be more flexible in HSI classification.

References

[1] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[2] 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.

[3] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, “Deep pyramidal residual networks for spectral–spatial hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[4] H. Sun, X. Zheng, and X. Lu, “A supervised segmentation network for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.

[5] 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.

[6] J. Zheng, Y. Feng, C. Bai, and J. Zhang, “Hyperspectral image classification using mixed convolutions and covariance pooling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, Jan. 2021.

[7] Z. Yang, Z. Xi, T. Zhang, W. Guo, Z. Zhang, and H.-C. Li, “CMR-CNN: Cross-mixing residual network for hyperspectral image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8974–8989, 2022.

[8] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, “Classification of hyperspectral image based on double-branch dual-attention mechanism network,” *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020.

[9] X. Zheng, H. Sun, X. Lu, and W. Xie, “Rotation-invariant attention network for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.

[10] Q. Liu, L. Xiao, J. Yang, and Z. Wei, “CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[11] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[12] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, “Spectral–spatial feature tokenization transformer for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

[13] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, “MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532513.

[14] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, “Hyperspectral image classification with deep learning models,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[15] Y. Li, H. Zhang, and Q. Shen, “Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network,” *Remote Sens.*, vol. 9, no. 1, p. 67, Jan. 2017.