Hyperspectral Image Classification Based on Interactive Transformer and CNN With Multilevel Feature Fusion Network

Hao Yang, Haoyang Yu, *Member, IEEE*, Ke Zheng, Jiaochan Hu, Tingting Tao, *Student Member, IEEE*, and Qiang Zhang, *Member, IEEE*

***Abstract*— Due to the powerful feature information mining ability of deep learning, models such as convolutional neural network (CNN) and Transformer have gained a certain progress in hyperspectral image classification (HSIC). Characteristically, the CNN is good at extracting local information, but it has the limitation of insufficient receptive field. While the Transformer has the advantage of global representation, it ignores local details to some extent. Therefore, this letter proposes an interactive Transformer and CNN with a multilevel feature fusion network (ITCNet) for HSIC. Specifically, in the image-based framework, features with different perceptual fields and depths are extracted interactively by a multilayer Transformer and CNN, then fused through a multilevel feature fusion module for class prediction. Experimental results on two real datasets verify its efficiency, with improvements over other related methods.**

***Index Terms*— Convolutional neural network (CNN), hyperspectral remote sensing, image classification, image-based framework, transformer.**

I. Introduction

HYPERSPECTRAL image (HSI) can obtain spatial and spectral information about the observed target. They can better distinguish physical differences between surface materials through broad and dense spectral images relative to natural images [1]. The characteristics of HSI also benefit applications such as target detection and classification. HSI classification (HSIC) takes advantage of the spatial and spectral features to assign a category label to each pixel. In early research, HSIC mostly adopted manual feature extraction methods such as random forest [2], support vector machine (SVM) [3], and sparse representation [4]. With the massive increase in data and more complex application scenarios, the traditional methods reveal their limitations since their focus is mainly on extracting shallow features. In recent years, with the improvement of computer computing power, deep learning-based techniques have achieved excellent performance in various vision tasks and gradually become the mainstream of HSIC.

In deep learning, a convolutional neural network (CNN) is widely used due to its excellent feature extraction capabilities and transferability. In [5], 1-D-CNN was used to extract the spectral information of images and classify HSIs directly in the spectral domain. In [6], 2-D-CNN is used to extract spatial features while introducing residual learning to deepen the depth of the network. In [7], a 3-D-CNN is proposed to extract spatial and spectral features jointly. In [8], the residual network is improved to capture the spectral and spatial features in an end-to-end training approach. The CNN-based model has the unique advantage of feature representation between different channels and is good at extracting local features. However, its perceptual field is affected by the size of its convolution kernel and has a limited ability to extract and represent complex spatial and global features.

Transformer is a model introduced to computer vision in recent years from natural language processing. It maintains its excellent ability to model the dependencies between sequence elements. In [9], the proposed SpectralFormer learns local spectral features from the perspective of sequences. Although the Transformer-based model has advantages for global information extraction due to its attention mechanism and multilayer perceptron (MLP) structure, it also leads to difficulties in capturing local information [10]. In [11], a convolution transformer mixer (CTMixer) is proposed to combine the advantages of CNN and Transformer.

Both CNN-based and Transformer-based models mentioned above mainly use a patch-based framework, which divides pixels and their neighborhoods into patches and feeds the patches into the model one by one to predict the centers of the patches. The predicted labels are then aggregated and reduced to a predicted original image map. However, the operation of dividing patches also brings some drawbacks. For example, the size of a patch limits the model's understanding of the overall image, and neighboring patches have a large amount of overlap. This redundant computation grows exponentially as the patch grows, consuming many computational resources. It also limits the Transformer's ability for global feature extraction. In [12], a multilevel codec structure was

Manuscript received 27 April 2023; revised 25 July 2023; accepted 28 July 2023. Date of publication 7 August 2023; date of current version 21 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42101350, Grant 42201362, and Grant 42001280; in part by the China Postdoctoral Science Foundation under Grant 2022T150080 and Grant 2023T160073; and in part by the Fundamental Research Funds for Central Universities under Grant 3132023254 and Grant 3132023165. *(Corresponding author: Haoyang Yu.)*

Hao Yang, Haoyang Yu, and Qiang Zhang are with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: alain@dlmu.edu.cn; yuhy@dlmu.edu.cn; qzhang95@dlmu.edu.cn).

Ke Zheng is with the College of Geography and Environment, Liaocheng University, Liaocheng 252059, China (e-mail: zhengkevic@gmail.com).

Jiaochan Hu and Tingting Tao are with the College of Environmental Sciences and Engineering, Dalian Maritime University, Dalian 116026, China (e-mail: hujc@dlmu.edu.cn; ttt@dlmu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3303008

designed using the Transformer based on the image-based framework, and better performance was obtained, compared with the patch-based framework, the image-based framework has higher computational efficiency and more comprehensive information extraction capability. On one hand, it can process each pixel in parallel, avoiding the redundant computation caused by the large number of overlapping patches generated in the patch-based framework, thus reducing the computational cost. On the other hand, the image-based framework is not limited by the Patch size and can extract more comprehensive global information, thus improving the classification accuracy.

In general, CNN is good at extracting high-frequency information, which is better at the local level. At the same time, Transformer is good at extracting low-frequency information, which is better at the global level. Under the above conditions, the issue of reasonably combining the advantages of Transformer and CNN to use global and local HSI information synergistically is essential. Therefore, this letter proposes a parallel interaction structure between CNN and Transformer to extract HSI's local and global information in conjunction with the image-based framework. The method is named interactive transformer and CNN with multilevel feature fusion network (ITCNet), and its contributions are as follows.

1) In the image-based classification framework, the global features and local features of HSIs are extracted in parallel using the Transformer and CNN, which improves the limitation of traditional patches for the Transformer's long-range information extraction capability. It also enhances the efficiency of training and testing.
2) Integrate the sense field of image-based framework for complete HSI and the excellent extraction capability of multilevel feature fusion structure for multiscale features to achieve end-to-end prediction of complete HSI.

II. Proposed Classification Framework

A. Framework for Image-Based Classification for HSI

As shown in Fig. 1, an image-based framework is applied in this section. During the training process, a binary mask of the same size as the image will first be made, and the positions of a small number of selected pixel samples with labels will later be labeled. Thus, the position labeled mask for the training part will be obtained, while the rest of the labels will not be involved in the training of the model and hence are used as a test set. Whereas the HSI will be kept in full size for one time input to the model for training instead of dividing it again into separate patches or pixels. The optimization of the model parameters is achieved by calculating the cross entropy of the predicted labels and the true labels of the selected pixel points. In the testing phase, the model classifies the whole image at the pixel level and outputs the predicted labels of all pixels.

In the patch-based framework, the HSI is divided into fixed-size patches, which are fed into the model separately for training. In the inference process, the model will classify each patch independently to generate the prediction value of the center pixel of the patch, and then the prediction results will be arranged to generate the complete image classification. Compared to image-based, there are mainly the following differences.

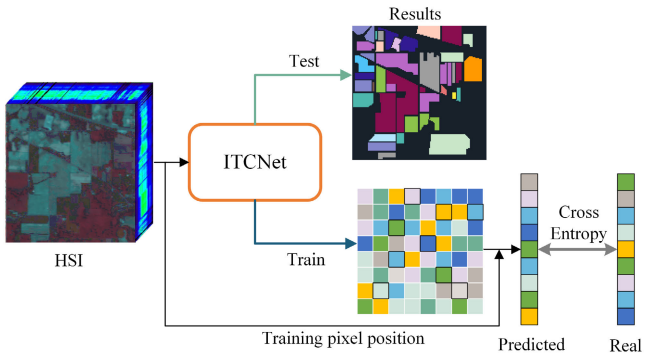

Fig. 1. Image-based framework for HSI.

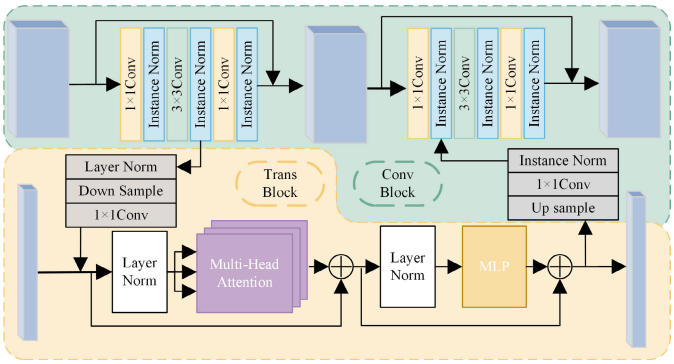

Fig. 2. GLIE module.

1) In patch-based, the patches involved in training are relatively independent of each other, and the parameters of the model only change with the current patch involved in training, not the whole image. Whereas in image-based, the model will update the parameters in the complete image.
2) In the inference stage, the patch-based prediction for the complete image is divided into two stages, first generating the prediction values of individual pixels and then arranging these prediction values to generate the prediction map. Whereas image-based prediction for an image is an end-to-end process that directly outputs a prediction map of the complete image.

In summary, the image-based method is an efficient and practical framework for HSIC, which can reduce computational costs while ensuring classification accuracy.

B. Interactive Concurrent Extraction With Global-Local Feature

CNN has the unique advantage of feature representation between different channels. It is good at extracting local features, but its perceptual field is affected by the size of the convolutional kernel and has limited ability to extract global features for complex spaces. In contrast, Transformer can model dependencies between sequence elements and thus has an advantage in representing global information.

To better utilize the rich spatial and spectral information in HSIs, a global–local interactive feature extraction (GLIE) module is designed in this part. As shown in Fig. 2, the module uses a two-branch parallel extraction strategy. The Transformer module includes two Layer Norm layers, a multihead attention (MSA) layer, and an MLP layer. The global representation

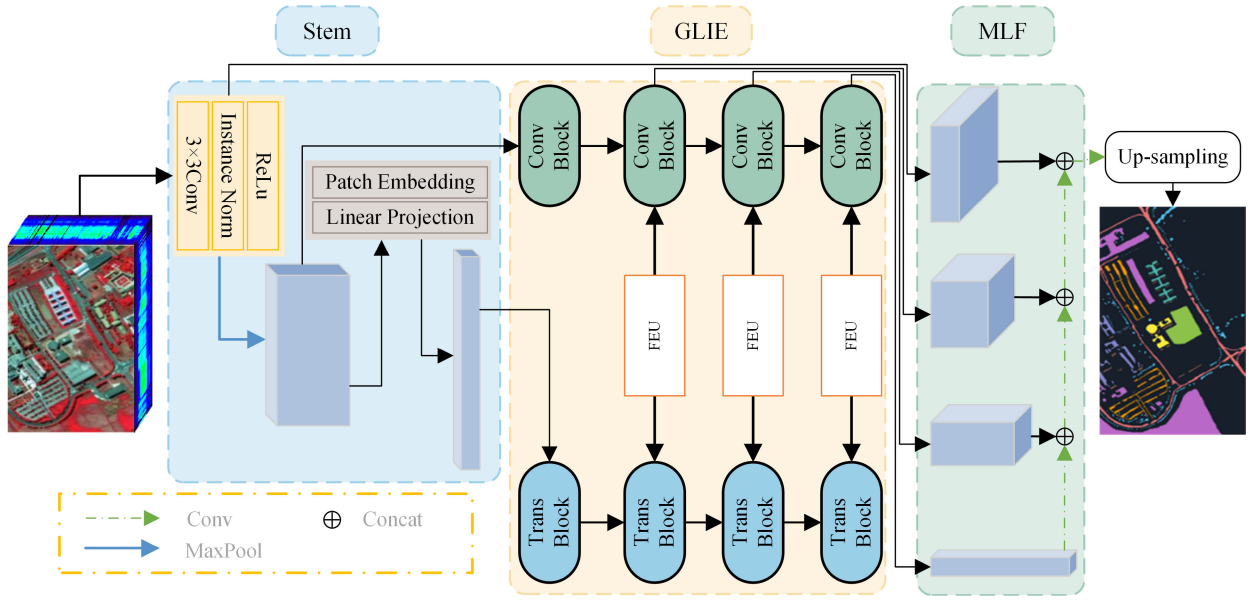

Fig. 3. Schematic of the ITCNet.

capability in the Transformer module is mainly derived from its designed MSA structure, and its process is represented as follows:

$$\text{MultiHead} = \text{Concat}(H_1, H_2, \ldots, H_N)\mathbf{W} \tag{1}$$

where $\mathbf{W}$ denotes the weight matrix learned by the model, Concat denotes concatenation over feature dimensions, H is the single-headed Self-attention, N represents the number of heads of self-attention, and MSA consists of multiple self-attention, where self-attention can be expressed as follows:

$$H = \text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V represent Query, Key, and Value, obtained by multiplying the input $\mathbf{X}$ by three different weight matrices $\mathbf{W}_Q, \mathbf{W}_k, \mathbf{W}_V$. d_k is the dimension of the Q and K. And the convolutional module consists of two 1×1 convolutional layers, one 3×3 convolutional layer, and three Instance Norm layers, which can be expressed as follows:

$$\mathbf{X}_n^{p+1} = N\left(\mathbf{W}_2\left(N\left(\mathbf{W}_1\left(N\left(\mathbf{W}_0\left(\mathbf{X}_n^p\right)\right)\right)\right)\right)\right) + \mathbf{X}_n^p \tag{3}$$

where $\mathbf{W}_0, \mathbf{W}_1$, and $\mathbf{W}_2$ represents the convolutional weight matrix, $\mathbf{X}_n^p$ is the pth layer and nth channel input features, $\mathbf{X}_n^{p+1}$ denotes the processed output features, and N is the Instance normalization.

The interaction of features between the Transformer and CNN is done through feature exchange unit (FEU), which uses 1×1 convolutional alignment channels, upsample and downsample operations to align spatial dimensions. Normalization is accomplished by Layer Normalization and InstanceNorm, the aim is to reduce the difference in scale of the data between the two branches. To extract features to multiple scales and depths, GLIE designs a multilayer composite structure and outputs parts of different sizes and numbers of channels.

C. Interactive Transformer and CNN With Multilevel Feature Fusion Network (ITCNet)

The overall network structure of the ITCNet is shown in Fig. 3, which consists of the Stem, GLIE, and the multilevel feature fusion module (MLF). The Stem part first varies the number of channels and integrates spectral information through convolution, InstanceNorm, and ReLu. MaxPooling then reduces the size of the original image to save computational resources. Features $\mathbf{X}$ and $\mathbf{X}'$ are then fed into the GLIE for feature extraction. Moreover, four features with different channels and resolutions are output to MLF at different network depths.

Afterward, MLF fuses the multiscale features and reflects them to the classification map of the original image dimensions to complete the pixel-level prediction. In the MLF fusion process, the most miniature size feature map is first aligned with the spatial size and the number of channels of the previous layer by an upsampling operation and a convolution layer, Upsampling is achieved by a bilinear interpolation method. The features are subsequently feature stitched with those of the last layer, followed by further feature fusion through two 3×3 convolutional layers and a ReLu layer. The resulting parts are then subjected to the same operation as the feature map of the previous layer. After fusion with the features of the uppermost layer, the resolution is upsampled to the original image size by a bilinear interpolation method to obtain predicted labels for all pixels.

III. Experimental Results and Analysis

A. Experimental Data and Setting

The first dataset is the Indian Pines scene, acquired by the Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS) sensor over northwest Indiana, United States. It includes 145×145 pixels and 200 spectral reflectance bands in the wavelength range 0.4–2.5 μm with a spatial resolution of 20 m. A total of 16 different feature classes are provided.

TABLE I

OVERALL, AVERAGE, K STATISTIC FOR THE INDIAN PINES DATASETS WITH 30 TRAINING SAMPLES PER CLASS. THE HIGHEST ACCURACIES ARE HIGHLIGHTED IN BOLD

Metrics	SVM-RBF	3DCNN	SSRN	DBDA	A^2S^2K-ResNet	CTMixer	Unet	MSTNet	ITCNET
OA	70.43%	72.56%	88.96%	87.50%	87.58%	89.20%	92.89%	93.03%	**96.05%**
AA	72.15%	75.35%	90.87%	89.84%	89.32%	91.03%	91.08%	88.82%	**95.81%**
Kappa	66.69%	68.95%	87.52%	85.86%	85.95%	87.81%	91.91%	92.08%	**95.50%**

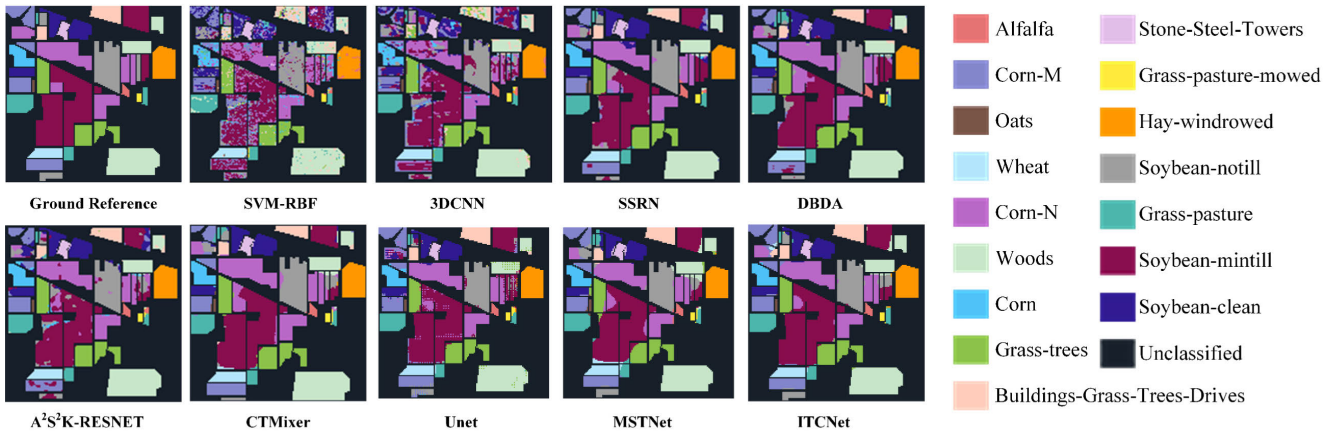

Fig. 4. Classification maps of different tested methods for the AVIRIS Indian Pines scene (30 samples per class).

TABLE II

OVERALL, AVERAGE, K STATISTIC FOR THE PAVIA UNIVERSITY DATASETS WITH 30 TRAINING SAMPLES PER CLASS. THE HIGHEST ACCURACIES ARE HIGHLIGHTED IN BOLD

Metrics	SVM-RBF	3DCNN	SSRN	DBDA	A^2S^2K-ResNet	CTMixer	Unet	MSTNet	ITCNET
OA	78.56%	80.65%	90.74%	92.26%	91.37%	89.77%	95.85%	95.80%	**97.43%**
AA	79.65%	83.05%	94.54%	95.34%	94.21%	94.15%	94.26%	93.23%	**95.78%**
Kappa	72.75%	75.00%	88.15%	90.06%	88.83%	86.95%	94.57%	94.48%	**96.61%**

TABLE III

TRAINING AND TESTING TIME FOR 100 EPOCH ON THE INDIAN PINES DATASETS

	Patch-based					Image-based		
	3DCNN	SSRN	DBDA	A^2S^2K-ResNet	CTMixer	Unet	MSTNet	ITCNET
Train Time(s)	20.61	83.35	310.02	72.67	19.53	2.77	2.49	2.59
Test Time(s)	1.22	4.62	8.72	4.12	2.37	0.02	0.02	0.04

The second dataset is the University of Pavia scene. This scene was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor during a flight campaign over Pavia, northern Italy, and includes 610 × 340 pixels and 103 spectral reflection bands in the wavelength range 0.43–0.86 μm with a spatial resolution of 1.3 m. A total of nine different feature classes are provided.

The comparative experimental model is (1) SVM with radial basis function (SVM-RBF) [3], (2) 3-DCNN [7], (3) spectral-spatial residual network (SSRN) [13], (4) A^2S^2K-ResNet [8], (5) double-branch dual attention (DBDA) [14], (6) CTMixer [11], (7) Unet [15], and (8) multilevel spectral–spatial transformer network (MSTNet) [12]. Here, (2)–(5) all use a patch-based framework, while (6) and (7) use an image-based framework. In the comparison experiments mentioned above, the hyperparameters were set up with the original paper. In ITCNet, the learning rate and epoch number are set to 0.0003 and 500, the network was trained using an Adam optimizer with parameters set as: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\beta_3 = 10^{-8}$. They are all implemented in a PyTorch-based environment using an Intel Core i7-11700 CPU, 32GB RAM, and an RTX 3060 12-GB GPU. This letter will construct the training set by 30 randomly selected pixels per class for each scene, and the other labeled samples will be used for testing evaluation.

The overall accuracy (OA), average accuracy (AA), and Kappa Statistic (κ) were used as evaluation metrics.

B. Results Analysis and Discussion

The classification results of different methods on the Indian Pine dataset are shown in Table I. The corresponding classification graphs of other methods on the Indian Pine dataset are shown in Fig. 4. From the above data, the following preliminary observations can be summarized.

1) Under the current sample conditions, the well-trained deep learning-based models are more suitable for complex scenarios than the SVM-RBF using shallow feature learning.

TABLE IV
OVERALL, AVERAGE, K STATISTIC FOR THE IP AND PU DATASETS WITH 30 TRAINING SAMPLES PER CLASS

Datasets	Methods	OA	AA	Kappa
Indian Pine	ITCNet-C	93.55%	93.58%	92.67%
	ITCNet-T	92.78%	83.84%	91.80%
	ITCNet	**96.05%**	**95.81%**	**95.50%**
Pavia U	ITCNet-C	94.06%	91.63%	92.26%
	ITCNet-T	94.85%	91.92%	93.28%
	ITCNet	**97.43%**	**95.78%**	**96.61%**

2) Compared to the patch-based framework models (3-DCNN, SSRN, DBDA, and A^2S^2K-ResNet), the image-based framework models (Unet, MSTNet, and ITCNet) improve the OA by more than 3.93%. This improvement is due to enhancing the receptive field brought by the image-based framework inputting the whole image. The image-based framework is more suitable than the patch-based framework when utilizing information at a distance.
3) In the image-based framework, MSTNet only uses the Transformer structure, and Unet only uses the CNN structure. There is no significant difference between MSTNet and Unet in OA, κ, while there is a slight difference in AA. This may be caused by the different focus of Transformer and CNN in extracting features, with MSTNet focusing more on long-range features and Unet relatively more on local features.
4) ITCNet with CNN and Transformer parallel structure for feature extraction were 3.02% higher in OA than MSTNet and 3.16% higher than Unet. This improvement mainly because ITCNet utilizes both global features extracted by Transformer and local features extracted by CNN when extracting features. It reflects that ITCNet has more advantages over the model using only Transformer versus the model using only CNN.
5) The training and testing time of the deep learning-based methods in this letter is shown in Table II. The image-based method is more time efficient compared to the patch-based method because the patch-based creates a large amount of redundant computation in the overlapping part between patches when dividing the patches.

Table III shows the classification maps and classification results of different methods on the Pavia University (PU) dataset, respectively. Compared with other methods, overall similar conclusions as on the Indian Pine dataset can be obtained. It is further verified that ITCNet has more global information, a larger perceptual field than Unet, and more local details than MSTNet. At the same time, the final multiscale feature fusion module designed by ITCNet also plays a corresponding role, which makes the extracted global and local information effectively utilized.

C. Ablation Experiments

To verify that the combination of Transformer and CNN modules in ITCNet is effective, experiments were conducted by removing the CNN branch in ITCNet (ITCNet-C) and the Transformer branch (ITCNet-T), respectively. Table IV shows that when the CNN branch was removed, OA decreased by 3.27% in IP and 2.57% in PU, respectively. When the Transformer branch was removed, OA decreased by 2.50% on the IP dataset and 3.37% on the PU dataset, respectively. This result shows the different focus of Transformer and CNN in extracting features. It also confirms that ITCNet forms an effective collaboration using the components extracted by Transformer and CNN.

IV. CONCLUSION

This letter proposes an ITCNet. Its main contributions include using the image-based framework to improve the global feature extraction capability of the transformer, increasing the perceptual field gap between the CNN and the transformer, and interacting the features at different scales for more comprehensive utilization. The classification experiments were conducted on two real hyperspectral datasets. The experimental results show that the proposed ITCNet yields better classification performance than other CNN-based and Transformer-based methods.

REFERENCES

[1] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.
[2] Z. Liu, B. Tang, X. He, Q. Qiu, and F. Liu, "Class-specific random forest with cross-correlation constraints for spectral–spatial hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 257–261, Feb. 2017.
[3] B.-C. Kuo, H.-H. Ho, C.-H. Li, C.-C. Hung, and J.-S. Taur, "A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 317–326, Jan. 2014.
[4] H. Yu, X. Shang, X. Zhang, L. Gao, M. Song, and J. Hu, "Hyperspectral image classification based on adjacent constraint representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 707–711, Apr. 2021.
[5] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Nov. 2015.
[6] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
[7] A. Ben Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.
[8] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.
[9] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
[10] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.
[11] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[12] 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.
[13] 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.
[14] 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.
[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.