2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) | 978-1-6654-7069-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/WHISPERS56178.2022.9955116

HYPERSPECTRAL IMAGE CLASSIFICATION BASED ON MULTI-LEVEL SPECTRAL-SPATIAL TRANSFORMER NETWORK

Hao Yang[1], *Haoyang Yu*[1], *Danfeng Hong*[2], *Zhen Xu*[1], *Yulei Wang*[1] *and Meiping Song*[1]

[1] Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, 116026, China

[2] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100094, China

ABSTRACT

Deep learning methods have been widely used in hyperspectral image classification (HSIC). In recent years, Convolutional Neural Network (CNN) has become a mainstream model of deep learning for HSIC. Although the CNN-based method has made great progress, it still faces a series of challenges such as insufficient use of long-distance information, limited receiving domain, and high computational overhead. In order to overcome these issues, this paper proposes a multi-level spectral-spatial transformer network (MSTNet) for HSIC through the image-based classification framework. The proposed network learns feature representation through a transformer encoder, and integrates multi-level features through a decoder to generate classification results. Finally, the experimental results on two real hyperspectral data sets verified the superiority of the method.

Index Terms— Hyperspectral image classification, Transformer, Self-attention, Convolutional Neural Network

1. INTRODUCTION

In recent years, hyperspectral images (HSI) have received widespread attention due to their rich spectral and spatial information [1,2]. Compared with ordinary images, HSI provide finer ground features. In the task of hyperspectral image classification (HSIC), deep learning methods have become a research hotspot due to their excellent performance [3,4].

The HSIC model based on deep convolutional neural network (CNN) is an excellent classification model [5]. Hong et al [6]. proposed a general multimodal deep learning (MDL) framework for the performance bottleneck problem of fine classification in complex scenes, which became an effective solution. Yu et al [7]. proposed a dual-channel convolutional network (DCCN) for global and multi-scale information utilization of HSI data. These CNN-based methods focus on local information and do not fully utilize global information. In recent years, some Transformer-based models have been used in image processing. Dosovitskiy et al [8]. proposed a pure transformer directly applied to image patch sequences for classification, called Vision transformer (ViT). Hong et al [9]. proposed a new class of backbone network SpectralFormer based on the sequential perspective of the Transformer, which can adapt to pixel and small-block inputs.

In this paper, we propose a Transformer-based model for HSIC called MSTNet to make full use of global information. The main contributions of this method are as follows: (1) CNN is replaced by Transformer for feature extraction, which makes full use of global information. (2) The image-based classification framework is used to replace the patch-based classification framework, which reduces the redundancy of information. (3) The proposed MSTNet is more lightweight and efficient.

2. PROPOSED APPROACH

In this section, we first introduce the image-based classification framework used, and then introduce our proposed MSTNet network.

2.1. Image-based Classification Framework for HSI

In the HSIC deep learning method, the traditional method usually uses the patch-base framework. The center of the training pixel and its neighbors are formed into a patch, and input into the model to predict the center of the label. In the process of testing, this method will gradually generate patches and input the model, and then, the predicted label is reshaped according to the original image size. Obviously, the patch-based method has the following shortcomings. First, the size of the patch limits the acceptance domain of the model. Secondly, the suitable size of the patch is affected by the spatial resolution of the image. When the size of the patch changes, the model will no longer be applicable, so it is difficult to design a general classification model. Finally, in the testing phase, the patch-based method has a lot of redundant information, requires a lot of computing resources, and is inefficient in time.

In this article, we introduce an image-based classification framework, which has a good performance on the problems

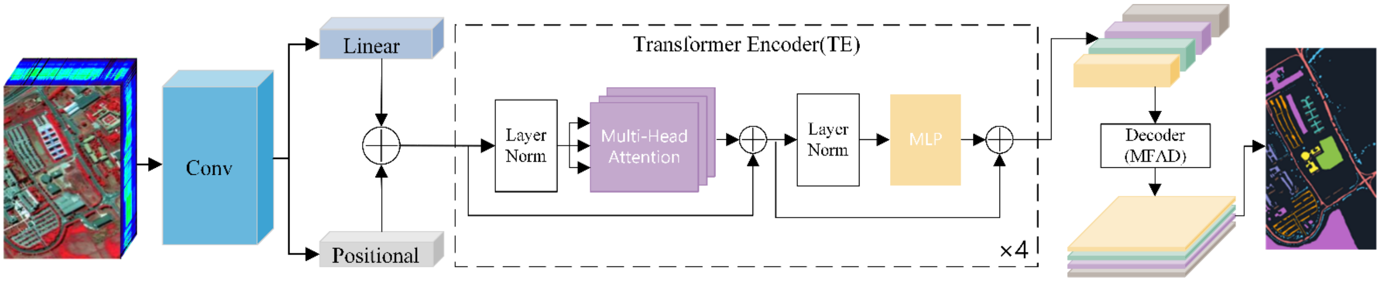

Fig. 1 An illustration of the proposed MSTNet.

in the above patch-based method and can be applied to any semantic segmentation model. In the training phase, the image containing the training sample is used as input, and the predicted labels of all corresponding pixels are output. Use the position of the marked sample as a mask to select the corresponding pixel. In the testing phase, we introduce images to the model and predict the labels corresponding to all pixels.

Compared with the traditional patch-based method, the image-based method has a faster reasoning process and saves computing resources.

2.2. Multi-level Spectral-Spatial Transformer Network

We propose a multi-level spectral-spatial transformer network (MSTNet), the network structure is shown in the figure 1. In MSTNet, the main modules are transformer encode (TE) and Multi-Level Feature Aggregation Decoder (MFAD). The brief workflow of the network is as follows: First, input HSI into MSTNet through the image-based classification framework. Then the CNN layer is used for processing to generate feature maps to reduce the spatial resolution. Secondly, in order to match the input size of the transformer, linear projection and position embedding are used to reconstruct the three-dimensional feature map into a two-dimensional sequence and input TE to perform MSA learning, and then input the learning result into the MLP layer. Finally, output the feature map from TE to MFAD to generate a multi-level fusion feature map.

a) Transformer Encoder (TE)

Figure 1 shows the structure of TE, which includes two and a multi-head self-attention (MSA). LN is usually carried out before the start of the training phase. The purpose is to standardize the data distribution to alleviate the disappearance of gradients and enhance the generalization ability of the network and the model description ability. MSA uses multiple connected independent self-Attention components, which are the key components in the transformer model. MLP consists of two fully connected layers and a nonlinear Gaussian Error Linear Unit (GELU).

b) Multi-Level Feature Aggregation Decoder (MFAD)

After obtaining the feature results in TE, in order to complete the pixel-level segmentation, we need to reshape the features of the encoder from the 2D embedding shape to the 3D feature map. In order to improve the presentation ability, multi-level feature aggregation is introduced in the decoder. As shown in the figure 1, in MFAD, the feature of the encoder is first converted from 2D to 3D, and then the convolutional layer is used to reduce the feature channel. Increase the spatial resolution of the feature map through two up-sampling layers. And, by introducing a top-down aggregation structure to improve the flow of information between levels. Finally, the multi-level features are connected by cascading channels to generate fusion features, and they are up sampled to the original spatial resolution.

3. EXPERIMENTAL RESULTS

3.1. Experimental Data and Setup

In this section, we will use the Indian Pines dataset collected by the AVIRIS sensor and the Pavia University dataset collected by the ROSIS spectral imager to evaluate the performance of MSTNet. At the same time, we will evaluate the performance of MSTNet, RBF-SVM [10], 3DCNN [11], SSRN [12] and UNet [13]. The computing environment consisted of the following specifications: i7-7820X CPU, 32 GB of RAM, and a GTX 2080TI 11GB GPU.

3.2. Experiments with the Pavia University dataset

In this experiment, we used the Pavia University dataset and the Indian Pines dataset to test and compare the performance of different classification methods. Figures 2 and 3 show the corresponding classification diagrams. Combining the results of Table 1 and Table 4, we can get the following conclusions:

1) Compared with RBF-SVM, 3DCNN, SSRN, these patch-based classification methods have higher classification accuracy, indicating that the combination of spectral information and spatial information is effective, and at the same time It also reflects that the deep learning method is

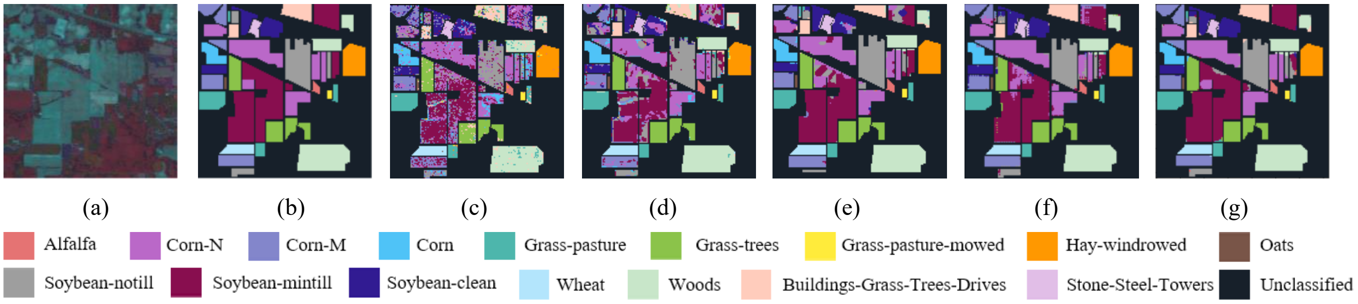

Fig. 2. Classification maps obtained by the different tested methods for the AVIRIS Indian Pines data set. In all cases, 697 labeled samples in total (50 per class) were used. The overall classification accuracies are given in the parentheses. (a) True-color composite image. (b) Reference map. (c) RBF-SVM (71.49%). (d) 3DCNN (79.24%). (e) SSRN (87.69%). (f) UNet (93.24%). (g) MSTNet (97.18%).

Table 1. Overall, Average, K Statistic, and Individual Class Accuracies for the Indian Pines Data with 50 Training Samples Per Class. The Highest Accuracies are Highlighted in Bold.

Class	RBF-SVM	3DCNN	SSRN	UNet	MSTNet
1	78.90%	88.24%	97.87%	80.70%	82.14%
2	61.14%	70.96%	90.32%	86.77%	**99.20%**
3	57.27%	72.73%	86.19%	87.99%	**98.20%**
4	57.84%	80.20%	93.05%	82.29%	95.18%
5	84.03%	91.94%	91.55%	94.62%	**98.34%**
6	92.26%	98.50%	99.12%	99.04%	**99.17%**
7	79.41%	90.32%	**100.00%**	**100.00%**	96.55%
8	97.99%	97.54%	97.42%	98.76%	99.79%
9	62.96%	83.33%	**100.00%**	86.95%	**100.00%**
10	66.14%	71.77%	82.77%	89.81%	89.98%
11	63.20%	71.86%	87.19%	95.91%	**99.14%**
12	68.79%	71.26%	84.59%	**94.12%**	93.69%
13	94.01%	95.79%	**100.00%**	97.61%	98.55%
14	89.65%	95.08%	98.38%	**98.81%**	97.81%
15	58.28%	72.39%	77.89%	94.37%	93.91%
16	93.33%	**96.84%**	92.08%	86.11%	93.00%
OA	71.49%	79.24%	87.69%	93.24%	**97.18%**
AA	75.33%	84.30%	92.40%	92.12%	**95.91%**
Kappa	68.00%	76.67%	86.11%	92.33%	**96.80%**

suitable for HSIC tasks.

2) Image-based classification methods such as UNet and MSTNet have higher classification accuracy than patch-based classification methods, reflecting the superiority of image-based classification methods.

3) The proposed MSTNet has the best performance, reflecting the effectiveness and potential of using Transformer to extract long-distance features.

3.3. Time efficiency comparison experiment analysis

Table 2 Overall, Average, K Statistic, and Individual Class Accuracies for the Pavia University Data with 50 Training Samples Per Class. The Highest Accuracies are Highlighted in Bold.

Class	RBF-SVM	3DCNN	SSRN	UNet	MSTNet
1	88.05%	89.65%	96.44%	**99.89%**	99.62%
2	90.34%	88.15%	92.68%	**99.70%**	99.23%
3	71.75%	88.59%	91.94%	95.79%	**99.43%**
4	92.11%	92.64%	96.33%	98.35%	98.05%
5	98.25%	99.56%	99.90%	99.78%	99.78%
6	73.05%	72.86%	96.41%	99.49%	**99.98%**
7	78.22%	79.38%	**100.00%**	99.85%	99.32%
8	80.11%	90.47%	**98.47%**	93.16%	98.12%
9	99.90%	98.88%	99.60%	98.96%	96.87%
OA	85.94%	85.99%	92.05%	98.80%	**99.18%**
AA	85.75%	88.91%	96.86%	98.33%	**98.93%**
Kappa	81.75%	81.93%	89.84%	98.41%	**98.91%**

Table 3. The training and testing times for patch-based and image-based classification.

	Indian Pines			University of Pavia		
	SSRN		MSTNet	SSRN		MSTNet
Patch size	7	11	-	7	11	-
Train Time of One Epoch (s)	2.48	4.53	0.066	1.03	1.59	0.28
Test Time of One Epoch (s)	4.94	8.69	0.052	40.35	60.75	0.25

In HSIC, the time efficiency of deep learning methods is one of the important properties. We selected the SSRN model in Patch-based and the image-based MSTNet model as a comparison, and set 7 and 11 patch sizes for SSRN. According to the results in Table 3, we have the following conclusions: 1) MSTNet time efficiency is better than SSRN. 2) When the patch increases, the time efficiency of SSRN will decrease.

The above results show that the image-based classification

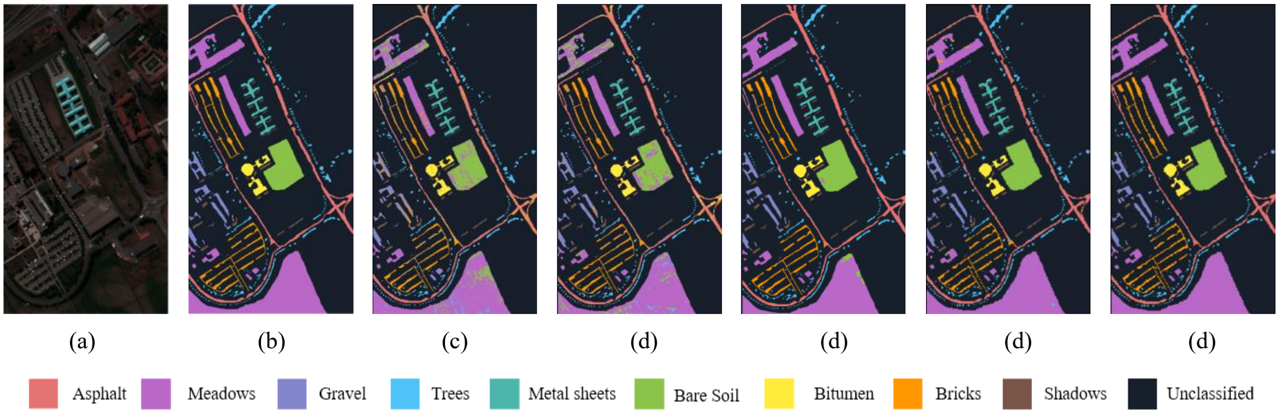

Fig. 3. Classification maps obtained by the different tested methods for the ROSIS University of Pavia data set. In all cases, 450 labeled samples in total (50 per class) were used. The overall classification accuracies are given in the parentheses. (a) True-color composite image. (b) Reference map. (c) RBF-SVM (85.94%). (d) 3DCNN (85.99%). (e) SSRN (92.05%). (f) UNet (98.80%). (g) MSTNet (99.18%).

method is more advantageous than the patch-based method in terms of time efficiency. With the patch-based method, when the patch increases, the redundant information will increase correspondingly, which makes the time cost more.

4. CONCLUSIONS

A Multi-level Spectral-Spatial Transformer Network for Hyperspectral Image Classification (MSTNet) based on an image classification framework is proposed. The main contribution of this method is to use the self-attention mechanism to collect long-distance spatial and spectral information to improve classification accuracy. In addition, incorporating an image-based classification framework improves the time efficiency of classification. Experiments on MSTNet on two real hyperspectral data sets show that the model has excellent performance in classification accuracy, time efficiency, and robustness.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China under Grant 42101350 and Grant 61801075, and the Fundamental Research Funds for Central Universities under Grant 3132022249.

REFERENCES

[1] H. Yu et al., “Neighborhood activity-driven representation for hyperspectral imagery classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4506-4517, Aug. 2020.

[2] B. Zhang, X. Sun, L. Gao and L. Yang, "Endmember Extraction of Hyperspectral Remote Sensing Images Based on the Ant Colony Optimization (ACO) Algorithm," IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 7, pp. 2635-2646, July 2011.

[3] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 7, pp. 5966-5978, July 2021.

[4] K. Zheng et al., "Coupled Convolutional Neural Network With Adaptive Response Function Learning for Unsupervised Hyperspectral Super Resolution," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 2487-2502, March 2021.

[5] V. Slavkovikj et al., Proceedings of the 23rd ACM International Conference on Multimedia, Association for Computing Machinery, Brisbane, Australia, 2015.

[6] D. Hong et al., "More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 5, pp. 4340-4354, May 2021.

[7] H. Yu, H. Zhang, Y. Liu, K. Zheng, Z. Xu and C. Xiao, “Dual-channel convolution network with image-based global learning framework for hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, Early Access.

[8] A.Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” International Conference on Learning Representations (ICLR), 2021.

[9] D. Hong et al., "SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers," IEEE Transactions on Geoscience and Remote Sensing, Early Access.

[10] H. Yu, L. Gao, J. Li, S. Li, B. Zhang and J.A. Benediktsson, “Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields,” Remote Sensing., vol. 11, no. 3, Apr. 2016.

[11] A. B. 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.

[12] Z. Zhong, J. Li, Z. Luo and M. Chapman, "Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 2, pp. 847-858, Feb. 2018.

[13] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Nov. 2015.