PASSNet: A Spatial–Spectral Feature Extraction Network With Patch Attention Module for Hyperspectral Image Classification

Renjie Ji, Kun Tan, *Senior Member, IEEE*, Xue Wang, Chen Pan, and Liang Xin

***Abstract*— Convolutional neural networks (CNNs) have achieved success in hyperspectral image (HSI) classification, but the performance is constrained by the limited reception field. In this regard, vision transformer (ViT) is introduced recently, which is of powerful capabilities in long-range feature extraction for HSI classification. However, transformers are computation intensive and poor for local feature extraction. The motivation for this study is to build a lightweight hybrid model, which ensembles the respective inductive bias from CNNs and global receptive field from transformers. In this work, we propose a concise and efficient framework—the spatial–spectral feature extraction network with patch attention module (PAM) (PASSNet), to simultaneously extract both local and global features. Specifically, we design an innovative plugin called PAM, which can be easily integrated into both CNNs and transformers blocks to extract spatial–spectral features from multiple spatial perspectives. Besides, a novel partial convolution (PConv) operation is introduced, with a reduced computational cost than vanilla convolution operation. Through coupling the local attention from the CNNs with the global receptive fields in the transformers, the proposed PASSNet exhibits a superior classification performance on three well-known datasets with a small training sample size.**

***Index Terms*— Hyperspectral image (HSI) classification, partial convolution (PConv), patch attention module (PAM), vision transformer (ViT).**

I. Introduction

In RECENT years, with the rapid development of remote sensing technology, hyperspectral sensors are considered a revolution, with their unprecedented spectral, spatial, and temporal resolutions [1], [2]. Currently, hyperspectral images (HSIs) have been widely applied in the fields of land-use monitoring [3], water quality parameter estimation [4], and so on. In order to make better use of HSI data in these areas, in particular, HSI classification has received long-standing and widespread interest [5].

Currently, with its powerful capabilities to extract local spatial and spectral features, convolutional neural networks (CNNs) have shown great performances in HSI classification. In the face of high-dimensional HSI data, a multiscale 3-D deep CNN (M3D-DCNN), proposed by He et al. [6], employs 3-D CNNs to extract spatial and spectral feature from HSI data. With the widespread use of group convolution and attention mechanisms, more lightweight CNN networks have been developed. For example, Cui et al. [2] designed a lightweight spectral–spatial attention network (LSSAN) for HSI classification, which maintained a high accuracy while significantly reducing the amount of computation. Gao et al. [7] proposed a multiscale residual network (MSRN) for HSI classification, with mixed depthwise separable convolution, which achieved better results with less computation effort. In summary, the CNNs have effectively improved HSI classification performance. However, due to the fixed convolutional kernel size design, it is challenging to capture the long-range dependencies of the complex HSI cube. Therefore, in order to achieve better classification performance, more CNN layers are often required, which contradicts the goal of lightweight modeling.

As is well known, in recent years, with the emergence of models, such as the vision transformer (ViT) model [8] in image processing and computer vision, transformer backbone networks have also been introduced in HSI classification [9], [10], [11]. Hong et al. [9] presented the spectralformer (SF) network, which can learn spectrally local sequence information from groupwise spectral embeddings. Mei et al. [10] designed the group-aware hierarchical transformer (GAHT) with grouped pixel embedding module, which obtains the feature information from the spatial–spectral context of HSI data. Recently, some researchers have attempted to use a CNN to first extract the shallow features before transformers. Sun et al. [11] introduced a new model named the spectral–spatial feature tokenization transformer method, which combines a transformer and a CNN, to extract both spectral–spatial features and high-level semantic features. In summary, it is feasible to apply transformer backbone networks for HSI classification, benefit from the larger global receptive field and better feature extraction capability for sequence data. However, since all the positions are required to be computed at each step, transformers incur a significant computational cost.

The motivation of this letter is to better tradeoff the network structure and the local-global feature extraction. Thus, we propose a lightweight hybrid model named PASSNet. PASSNet consists of two convolutional blocks and two transformer blocks, which can employ the strengths of CNNs in local feature extraction and the benefits of transformers in long-range context modeling. In detail, first, we designed a lightweight shallow feature information extraction module

Manuscript received 15 July 2023; revised 21 September 2023; accepted 28 September 2023. Date of publication 6 October 2023; date of current version 18 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42171335; in part by the Shanghai Municipal Science and Technology Major Project under Grant 22511102800; and in part by the National Civil Aerospace Project of China under Grant D040102. *(Corresponding author: Kun Tan.)*

Renjie Ji, Kun Tan, and Xue Wang are with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China (e-mail: ecnu.jirenjie@gmail.com; tankuncu@gmail.com; wx_ecnu@yeah.net).

Chen Pan and Liang Xin are with the Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China (e-mail: panpan_tj@126.com; lxin8764@163.com).

Digital Object Identifier 10.1109/LGRS.2023.3322422

based on partial convolution (PConv). Second, we developed a new attention module named the patch attention module (PAM), which can be embed in convolution or self-attention modules easily. In the convolutional blocks, PAM can help to better extract local spatial–spectral feature information from multiple spatial perspectives. Besides, a simplified version of the PAM is also integrated into the multihead self-attention (MHSA) layer in the transformer blocks, which contributes blend local–global spatial and spectral features, while reducing the computational cost. The main contributions of this letter can be summarized as follows.

1) We propose an effective attention mechanism module named PAM, designed specifically for the HSI classification task with small input patches. PAM can strengthen CNNs and transformers in extracting local and global spatial–spectral features.
2) We introduce the lightweight and effective PASSNet, consisting of two convolutional blocks and two transformer blocks, which enable better extraction of local and information from HSI cubes.
3) We conduct a comparison analysis on three well-known datasets, where the proposed PASSNet framework obtained superior results compared to other models.

II. Proposed Methodology

A. Partial Convolution

Recently, a novel module named PConv was proposed by Chen et al. [12], which can further reduce the number of training parameters and the computational cost. Fig. 1 illustrates the different structures of vanilla convolution, group convolution, and PConv. The PConv module first divides the input feature cube $X \in \mathbb{R}^{H\times W\times C}$ equally into n parts as $X_i \in \mathbb{R}^{H\times W\times (C/n)}$ along the channel dimension without overlapping. Then, a vanilla convolution module is applied to the first part, while the other parts remain unchanged and are then concatenated back with the convolved part. Thus, it can be seen that only the first part of the input feature cube in PConv needs to be computed through vanilla convolution

$$\widehat{X} = \mathcal{F}_{\text{PConv}}(X) = \text{concat}(\mathcal{F}_{\text{Conv}}(X_1), X_2, \ldots, X_i, \ldots, X_n) \tag{1}$$

where n is the number of divided groups, and $\mathcal{F}_{\text{Conv}}$ represents a vanilla convolution operation.

We designed a residual module with PConv to extract the shallow spatial–spectral features for HSIs, with a small number of model parameters and amount of computation. As shown in Fig. 2, each convolution module consists of a PConv ($n = 2$) accompanied by two pointwise convolutions (PWConvs). Furthermore, a residual structure, a batch normalization (BN) layer, and Gaussian error linear unit (GELU) activation are employed to keep the robustness of the model.

B. Patch Attention Module

The spatial–spectral attention mechanism has been widely applied in image classification. One of the most popular attention mechanisms is the squeeze and excitation (SE) module [13], but it ignores the spatial information, which is rich and critical in remote sensing images. Inspired by the SE module [13] and coordinate attention (CA) module [14], we propose the PAM, which attempts to extract the local spatial–spectral information from the broken ground surface in HSIs. Fig. 3 shows the detailed design of the proposed PAM module and its differences with SE and CA modules.

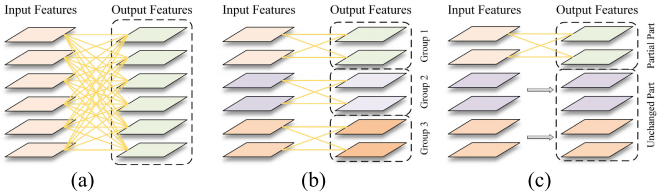


Fig. 1. Different structures. (a) Vanilla convolution. (b) Group convolution. (c) PConv.

Given the input $X \in \mathbb{R}^{H\times W\times C}$, first, two spatial extents of average pooling kernels, with the kernel size as $H \times 1$ and $1 \times W$, are applied to encode the input features by channel in the horizontal and vertical directions, respectively. Besides, due to the low spatial resolution of HSI, a surface object often composed of only a few pixels. Thus, a sliding and nonoverlapping average pooling kernel, with the fixed kernel size of 2×2, is used to extract the spatial features of the blocks. Thus, the spectral information at different positions is extracted. The output of the three average pooling operations at the cth channel can be formulated as follows:

$$\widehat{X_c^h} = \frac{1}{W}\sum_{i=1}^{W} X_c(h,i), \quad \widehat{X_c^w} = \frac{1}{H}\sum_{j=1}^{H} X_c(j,w)$$
$$\widehat{X_c^b} = \widehat{X_c^{i,j}} = \frac{1}{2\times 2}\sum_{m=0}^{1}\sum_{n=0}^{1} X_c^{2i+m,2j+n} \tag{2}$$

where $\widehat{X_c^h}$, $\widehat{X_c^w}$, and $\widehat{X_c^b}$ are the output of the horizontal, vertical, and block PAM modules, respectively. Differing from the SE module with global average pooling (GAP), the directional and block spatial structures are preserved in the PAM module.

In addition, we introduce the PWConv block to enable the extraction of local spatial–spectral features of HSIs. Specifically, transpose and flatten operations are first performed on the output of the PAM module, respectively, so that those vectors can be concatenated. Then, PWConv combined with BN and a nonlinear activation function $h_$swish are applied to concatenate the spatial features

$$X' = \text{concat}\left(\widehat{X_c^h}, \widehat{X_c^w}, \widehat{X_c^b}\right) \tag{3}$$
$$X' = h\_\text{swish}\left(\text{BN}\left(\mathcal{F}_{\text{PWConv}}\left(X'\right)\right)\right) \tag{4}$$

where $X' \in \mathbb{R}^{(H+W+HW/4)\times (C/r)}$ is the intermediate feature cube, and r is the reduction ratio to squeeze the number of channels, as in the SE block. After this, X' is split back into three separate tensors at their original size. Then, three individual PWConv and sigmod operations are applied to excite X'_h, X'_w, and X'_b, so that the number of channels is restored to match the input X

$$\begin{aligned} g^h &= f\left(\mathcal{F}_{\text{PWConv1}}\left(X'_h\right)\right) \\ g^w &= f\left(\mathcal{F}_{\text{PWConv2}}\left(X'_w\right)\right) \\ g^b &= f\left(\mathcal{F}_{\text{PWConv3}}\left(X'_b\right)\right) \end{aligned} \tag{5}$$

where f is the sigmoid function. Next, g^w is transposed back to its original shape, and nearest neighbor interpolation is applied to restore g^b to the original spatial size. Finally, the three outputs g^h, g^w, and g^b are considered as being attention weighted and are multiplied with the input features X

$$\widehat{X_{i,j}} = X_{i,j} \times g_i^h \times g_j^w \times g_{i,j}^b \tag{6}$$

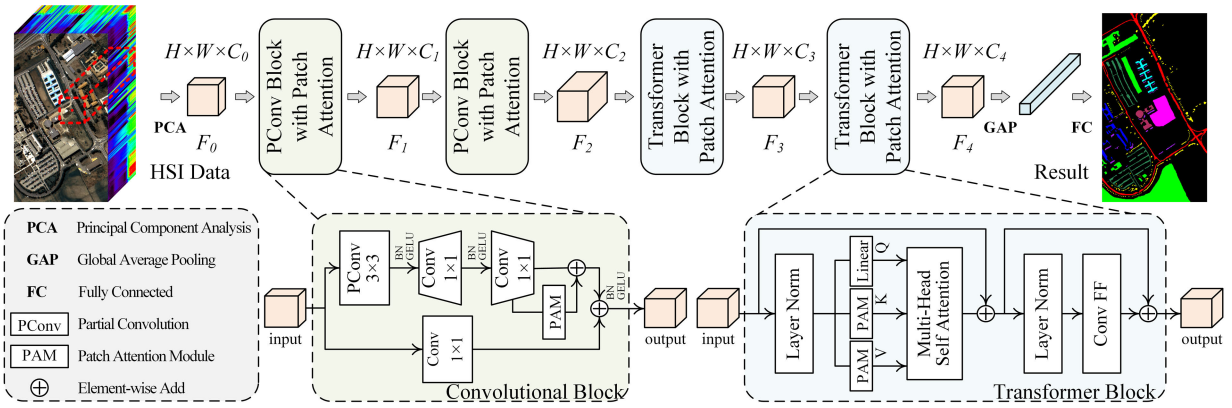
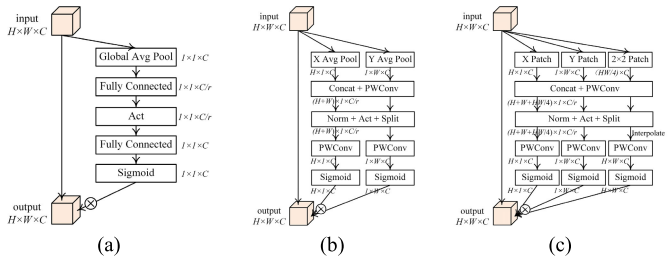
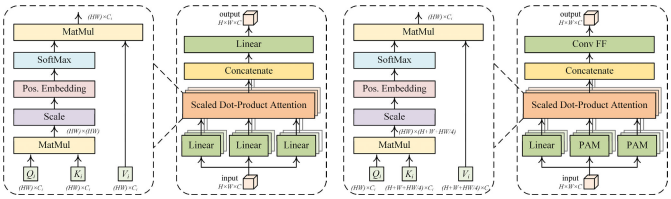


Fig. 2. Overall illustration of the proposed PASSNet framework for HSI classification.



(a) (b) (c)

Fig. 3. Different design. (a) SE module. (b) CA module. (c) Proposed PAM.



(a) (b)

Fig. 4. Different designs of MHSA attention block. (a) Original MHSA module. (b) PAM-enhanced MHSA module.

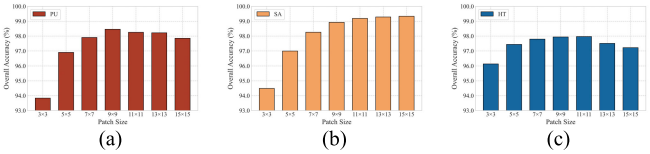
where $\widehat{X_{i,j}}$ is the final output of the PAM block. The proposed PAM is added to the PConv block to improve the extraction of local image features by the convolution module.

C. Transformer Block With Simplified PAM

Similar to most ViT modules, the transformer block in the proposed PASSNet consists of two parts: an MHSA layer and a feed forward (FF) layer. Since HSIs are very fragmented, the fusion of local and global features is very important for HSI classification, and we tried to migrate a simplified PAM block into the transformer block. The PAM module helps to enhance the local–global feature extraction capability of the MHSA layer, thus contributing to improved classification results with less computational cost.

As shown in Fig. 4(b), a simplified PAM operation is performed to extract the spatial features and reduce the feature size of the input key and value before the MHSA layer. After the concatenation of the three spatial features extracted by PAM, as shown in (3), only one sigmoid activation function is retained, and all the convolution operations in the PAM are now removed. Each pixel is treated as a token, and all the pixel tokens constitute the input sequence. Given the input $X \in \mathbb{R}^{(HW)\times C}$, the query Q', key K', and value V' can then be formulated as follows:

$$Q' = XW^{Q}, \quad K' = \mathrm{PAM}\left(XW^{K}\right), \quad V' = \mathrm{PAM}\left(XW^{V}\right) \tag{7}$$



(a) (b) (c)

Fig. 5. Sensitivity analysis for the PASSNet with different input patch sizes on the three datasets. (a) PU. (b) SA. (c) HT.

where Q' with only linear projection, and K' and V' are the PAM-enhanced key and value, with abundant local spatial features, respectively. Then, an MHSA layer with two heads is applied to extract global spatial and spectral feature.

For the FF layer, since HSIs feature is rich in spatial relationships, we reconstructed the FF layer with a convolution operation instead of linear projection, named as ConvFF. The output of the ConvFF layer can be computed as follows:

$$\begin{aligned} \hat{X} &= \mathrm{GELU}(\mathcal{F}_{\mathrm{PWConv}}(X)) \\ \hat{X} &= \mathrm{GELU}\left(\mathcal{F}_{\mathrm{DWConv}}\left(\hat{X}\right)\right) + X \\ \hat{X} &= \mathcal{F}_{\mathrm{PWConv}}\left(\hat{X}\right) \end{aligned} \tag{8}$$

where $\mathcal{F}_{\mathrm{PWConv}}$ is the PWConv with the expansion times as 4, and $\mathcal{F}_{\mathrm{DWConv}}$ is the depthwise convolutions (DWConvs) with the kernel size as 3.

Moreover, as shown in Fig. 2, layer normalization, residual structure, and absolute positional encoding are also introduced in the transformer blocks. These operations can improve the stability and learning efficiency of the model.

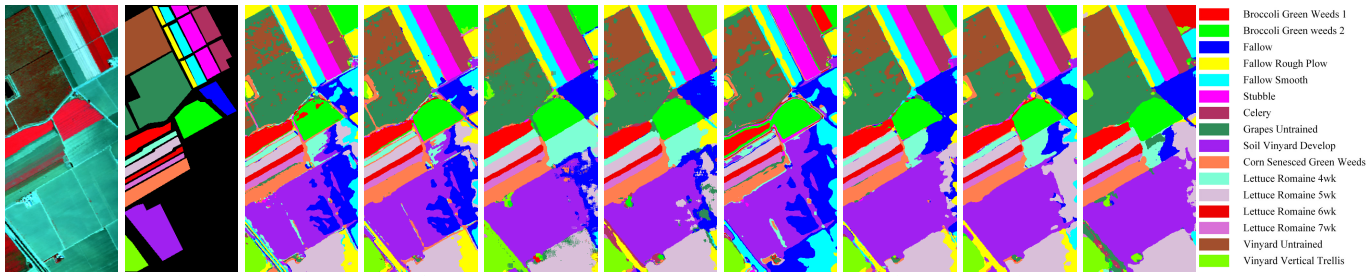
D. Overall Architecture

An overview of the proposed PASSNet is shown in Fig. 2. We aimed at designing a novel hybrid method to further improve the performance of HSI classification. The PASSNet consists of two convolutional blocks and two transformer blocks. The first two convolutional blocks with PAM are specifically designed to fully extract local features from HSI patch, and the latter two transformer blocks with PAM-enhanced MHSA are applied to extract and blend local–global spatial and spectral features.

The proposed PASSNet takes HSI patch data as the input, where each patch is labeled based on the label of its center pixel. First, the principal component analysis (PCA) method is applied to reduce the dimensionality of the HSI data to $C_0 = 30$. We then sequentially feed the feature cube through four blocks to obtain further feature maps. The four generated feature cubes keep the spatial size of $H \times W$ unchanged,

TABLE I
CLASSIFICATION RESULTS OF THE COMPARE CNN- AND TRANSFORMER-BASED METHODS ON THE THREE DATASETS

Datasets	Indicators	3-D CNN	M3D-DCNN	LSSAN	MSRN	SF	SPRLT	GAHT	PASSNet
PU	OA (%)	89.12±0.64	89.18±1.18	96.90±0.15	97.49±0.33	90.00±1.28	96.65±0.38	98.00±0.35	**98.48±0.31**
	AA (%)	84.28±2.06	83.76±1.64	94.80±0.85	95.63±0.62	83.74±1.42	93.61±1.24	96.66±0.71	**97.47±0.53**
	$\kappa\times100$	85.35±0.92	85.46±1.62	95.88±0.20	96.67±0.44	86.64±1.72	95.55±0.51	97.35±0.47	**97.99±0.41**
SA	OA (%)	90.45±0.81	92.16±0.85	96.78±0.51	96.66±0.45	89.68±0.90	95.85±0.38	95.92±0.79	**98.99±0.65**
	AA (%)	93.23±0.75	94.43±1.20	98.26±0.24	98.20±0.17	90.97±1.54	97.74±0.18	97.77±0.34	**99.40±0.24**
	$\kappa\times100$	89.36±0.90	91.28±0.95	96.41±0.57	96.28±0.51	88.51±1.01	95.38±0.43	95.46±0.88	**98.87±0.72**
HT	OA (%)	88.52±0.96	88.61±1.66	97.34±0.51	97.24±0.63	88.35±0.83	96.26±0.35	95.71±0.64	**98.05±0.34**
	AA (%)	88.36±1.14	88.27±1.74	97.21±0.46	97.34±0.57	86.95±0.97	96.01±0.46	95.85±0.58	**98.15±0.30**
	$\kappa\times100$	87.58±1.03	87.68±1.79	97.13±0.55	97.02±0.68	87.40±0.90	95.96±0.37	95.36±0.69	**97.89±0.37**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 6. Classification maps of the compared CNN- and transformer-based methods on the SA dataset. (a) False-color map. (b) Ground-truth map. (c) 3-D CNN. (d) M3D-DCNN. (e) LSSAN. (f) MSRN. (g) SF. (h) SPRLT. (i) GAHT. (j) PASSNet.

while the channel number changes, which are denoted at $C_1 = 64$, $C_2 = 128$, $C_3 = 64$, and $C_4 = 64$. Finally, a GAP layer and a fully connected module are applied to obtain the HSI classification results.

III. EXPERIMENTS AND ANALYSIS

A. HSI Dataset Description

To verify the validity of the proposed PASSNet framework, we conducted experiments on three well-known HSI datasets: the Pavia University (PU), Salinas (SA), and Houston 2013 (HT) datasets. We set a relatively small sample size for the training set. For the PU and SA datasets, 1% labeled samples were selected for the training. Due to the small number of samples tagged in the HT dataset, 5% labeled samples were selected. All the training sets were selected independently and randomly, and the other samples were applied for the testing.

B. Experimental Setups

The experiments were implemented in the PyTorch 1.13 environment on a desktop PC with an NVIDIA RTX 2070 GPU. For the proposed PASSNet, the learning rate was set to 0.001, and the AdamW optimizer was employed to update the training parameters. Furthermore, for all the models, we set the training batch size to 64 and the number of training epochs to 100. To ensure the optimal performance of the comparison models, all the other experimental setups remained the same as reported in the original articles, as much as possible.

The overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) are used to compare the classification results in a systematic manner. To minimize the experimental errors, we applied a random sampling strategy and repeated the process for five times. Finally, the experimental results are reported in the form of mean and standard deviation.

C. Influence of Patch Size

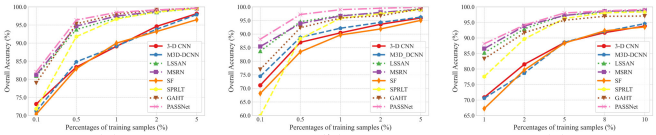
We conducted several experiments using different input patch sizes from the set of {3, 5, 7, 9, 11, 13, 15}. All the other training parameters were set as described in Sections III-A and III-B. Fig. 5 indicates that, for the PU and HT datasets, the OA shows an increasing trend with the patch size, until a certain point (9×9). For the SA dataset, the OA continues to improve within the range of the tested input patch sizes. To balance the classification accuracy and computational cost, we set the patch size to 9×9 for the subsequent experiments.

D. Comparison With Other Methods

In this section, we compare the proposed PASSNet with several commonly used deep learning methods. The CNN-based backbone models were 3-D CNN [15], M3D-DCNN [6], LSSAN [2], and MSRN [7], and the transformer-based backbone models were SF [9], the local transformer with spatial partition restore network (SPRLT) [16], and GAHT [10]. The proposed PASSNet is a hybrid architecture that combines the features of both CNN and transformer structures.

The classification result accuracy statistics for the comparison methods and the proposed PASSNet are listed in Table I. It can be found the PASSNet obtains the best performance on three well-known datasets. In detail, PASSNet outperforms the second-ranked method by approximately 0.48%/0.81% in the PU dataset, 2.21%/1.14% in the SA dataset, and 0.71%/0.81% in the HT dataset, in terms of OA/AA.

In addition, the classification results based on the same training samples are visualized to qualitatively compare these methods in terms of visual effects based on the SA dataset. As presented in Fig. 6, by adding the PAM to the convolutional and transformer blocks, PASSNet is able to extract more spatial–spectral features for HSI data. PASSNet is better at producing a clear and visually appealing visualization than



(a) (b) (c)

Fig. 7. Classification performance under different percentages of training samples on the three datasets. (a) PU. (b) SA. (c) HT.

TABLE II

TOTAL TRAINABLE PARAMETERS, FLOPS, AND PREDICTING TIMES FOR DIFFERENT COMPARISON METHODS ON THE SA DATASET

Indicators	3D-CNN	M3D-DCNN	LSSAN	MSRN	SF	SPRLT	GAHT	PASSNet
Params. (M)	0.39	0.60	0.04	0.14	0.36	0.84	0.97	0.24
FLOPs (M)	70.33	45.21	1.10	3.77	37.68	68.22	78.69	19.45
Pred. (s)	13.99	13.80	8.85	7.88	40.91	31.72	18.72	15.24

TABLE III

TOTAL TRAINABLE PARAMETERS AND OAS FOR DIFFERENT ABLATION STRUCTURES ON THE THREE DATASETS

Network composition	Computational cost (for the PU dataset)		OA (%)		
	Params. (M)	FLOPs (M)	PU	SA	HT
Baseline	0.2842	23.03	97.64	98.37	97.07
Baseline+PConv	0.2655	21.62	97.86	98.41	97.30
Baseline+PConv+PAM	0.2723	21.87	98.15	98.57	97.75
Baseline+PConv+PAM+PAM-MHSA	0.2438	19.45	**98.48**	**98.99**	**98.05**
Baseline+PConv+CA+PAM-MHSA	0.2420	19.28	98.27	98.78	97.71

the other methods, as it has a weaker sense of fragmentation, with fewer noise points and smoother, clearer boundaries.

E. Robustness Evaluation

To further evaluate the robustness of the proposed PASSNet, Fig. 7 illustrates the classification performance of all comparative methods across three datasets, considering varying percentages of training samples. It can be found with the increase of the number of training samples, PASSNet consistently achieves the best OA on three datasets, which shows its robustness. This is attributed to the integration of CNNs and transformer with PAM modules, which effectively harnesses the spatial and spectral characteristics of the limited training samples, resulting in an improved feature learning capacity.

F. Complexity Analysis

Table II shows the total trainable parameters, FLOPs, and predicting times of several methods on three datasets. Compared with other methods, it can be found that the proposed PASSNet has a medium number of parameters and FLOPs. The LSSAN method, as a lightweight CNN-based network with spatial and spectral attention mechanisms, boasts minimal parameters and FLOPs. The GAHT method owns the largest computational cost due to its incorporation of multiple transformer modules. In summary, the proposed PASSNet demonstrated superior classification performance and visual effect while maintain an acceptable computational burden.

G. Ablation Experiment

To verify the effectiveness of our proposed method, we conducted ablation experiments on various modules of PASSNet. The baseline network consists of the vanilla convolution and the original MHSA modules without any attention mechanism. Table III shows that the combination of PCA and PConv is positive for classification. The PAM module can also enhance the local–global feature extraction capability of convolution and MHSA layer. In summary, the proposed method is effective for HSI classification.

IV. CONCLUSION

In this letter, an effective and efficient spatial–spectral feature extraction framework called PASSNet has been proposed for HSI classification, which is a combination of convolutional and transformer blocks. First, we introduce a new convolution module PConv that can quickly extract the shallow semantic information for HSIs. In addition, we present a novel lightweight attention mechanism named PAM, which can be introduced into the CNNs and transformers blocks. Finally, based on a small training sample size, the PASSNet demonstrated an excellent classification ability on three well-known datasets.

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[2] Y. Cui, J. Xia, Z. Wang, S. Gao, and L. Wang, "Lightweight spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510114.

[3] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.

[4] C. Niu, K. Tan, X. Jia, and X. Wang, "Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery," *Environ. Pollut.*, vol. 286, Oct. 2021, Art. no. 117534.

[5] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: A review," *Sensors*, vol. 12, no. 4, pp. 4764–4792, Apr. 2012.

[6] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3904–3908.

[7] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021.

[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[9] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[10] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

[11] 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.

[12] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," 2023, *arXiv:2303.03667*.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[14] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[15] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben 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.

[16] Z. Xue, Q. Xu, and M. Zhang, "Local transformer with spatial partition restore for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4307–4325, 2022.