A Triple-Path Spectral–Spatial Network With Interleave-Attention for Hyperspectral Image Classification

Ziqing Deng, Yang Wang, Bing Zhang, Linwei Li, Jihong Wang, Lifeng Bian, and Chen Yang

***Abstract*—To exploit hyperspectral image's (HSI) spectral–spatial information and reduce network complexity, a triple-path convolution neural network with an interleave-attention mechanism is constructed for high-precision classification. A hybrid branch is proposed to capture joint features, which are later utilized as complementary information for purely spectral and spatial features. Furthermore, the interleave-attention mechanism is elaborately designed to increase the interaction of data from spectral, spatial, and joint branches as they propagate through the network for feature integration. Meanwhile, two attention modules are adopted in the corresponding branch to optimize extracted features for better feature representation. We utilize several real HSI datasets to evaluate network performance, which demonstrates that the proposed triple-path network can obtain very satisfactory performance with fewer parameters and low computational complexity.**

***Index Terms*—Hyperspectral image (HSI) classification, interleave-attention, spectral and spatial attention block, spectral–spatial feature extraction, triple-path network (TP-Net).**

I. Introduction

HYPERSPECTRAL images (HSI) are obtained by a hyperspectral imaging spectrometer, which provides both detailed spatial structure and abundant spectral bands. HSI analysis has been adopted in many tasks, such as vegetation-cover classification, ground-changing object detection, natural resources segmentation, and hyperspectral unmixing [1]–[4]. Among those domains, the HSI's abundant spectral and spatial information is explored by researchers for object analysis. Also, because of HSI's low spatial resolution, it usually suffers from the material mixture effect, which influences the information analysis process. Therefore, how to extract the abundant HSI information effectively is essential to properly analyzing the characteristics of different objects and achieving object recognition [5]–[7].

Two types of methods are mainly used in the HSI classification tasks, i.e., traditional [8]–[10] and deep-learning methods [11]–[13]. At the beginning of classification studies, researchers usually directly utilized spectral information for HSI classification. Traditional methods include k-nearest neighbor [14], support vector machine (SVM) [15], and random forest [16], which directly use single-pixel sequence vectors as the input of the HSI classifier. To further improve the classification performance, researchers took spatial information into consideration. In 2017, Lu *et al.* [17] fused features extracted from three pixel levels into one composite feature, and SVM was adopted for classification. In 2019, Cao [18] and Dong [19] employed low-rank matrix factorization and proposed a feature learning method to obtain representative spectral–spatial features for high classification precision. It is clear that utilizing the HSI spectral and spatial features can effectively improve the classification performance, but traditional architectures heavily rely on hand-crafted features, which require professional prior knowledge.

The deep-learning-based method has also been introduced into HSI classification since it can automatically extract abstract features from input data and has achieved remarkable progress in computer vision fields [20]–[23]. So far, researchers have proposed various deep-learning-based methods [24] for HSI classification and achieved remarkable progress. In 2014, Chen *et al.* [25] utilized a deep-learning method to process hyperspectral data for the first time and adopted a stacked autoencoder to analyze the spectral information. In 2015, Hu *et al.* [26] constructed a multiple-layer convolutional neural network (CNN) that took each pixel's spectral bands as input to classify HSI. Spectral-based methods showed good performance, and subsequent studies have verified that spatial information also helps to improve performance.

To further explore the HSI spatial information, researchers combined spatial context with spectral information, which can considerably improve the classification performance. In 2017, Li *et al.* [27] proposed a 3-D CNN that directly utilized a 3-D convolution operation to extract HSI features for classification. To alleviate the declining-accuracy phenomenon in deep networks, in 2018, Zhong *et al.* [28] adopted a residual connection to 3-D CNN for deep abstract spectral–spectral feature

Manuscript received 10 May 2022; revised 17 June 2022 and 2 July 2022; accepted 15 July 2022. Date of publication 19 July 2022; date of current version 1 August 2022. This work was supported in part by the National Nature Science Foundation of China under Grant 62065003 and in part by Guizhou Provincial Science and Technology Projects under Grants Qiankehe-ZK[2022] Key-020 and General-105. *(Corresponding author: Chen Yang.)*

Ziqing Deng, Yang Wang, Bing Zhang, Linwei Li, Jihong Wang, and Chen Yang are with the Power Systems Engineering Research Center, Ministry of Education, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China (e-mail: dzq351089672@163.com; y.wang.gzu@foxmail.com; bzhang526@163.com; as_emiya5@163.com; wjihong20@163.com; eliot.c.yang@163.com).

Lifeng Bian is with the Frontier Institute of Chip and System, Fudan University, Shanghai 200433, China (e-mail: lfbian@fudan.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3192470

Codes are available on-line at: https://github.com/HyperSystemAndImageProc/A-Triple-Path-Spectral-Spatial-Network-with-Interleave-Attention-for-Hyperspectral-Image-Classificat.git

extraction. Except for the depth, the network width can also be an important factor in performance enhancement. Kang *et al.* [29] constructed a double-branch network that adopted residual and dense connections to extract spectral–spatial features and obtained satisfactory results. Chen *et al.* [30] proposed a light-weighted dual-path structure with 1-D and 2-D convolution for spectral–spatial feature utilization.

To further widen the network path for better performance, in 2019, Meng *et al.* [31] proposed a multipath network with residual connection to extract features in parallel and achieved high-accuracy classification results. To overcome the misclassifications of the irregular regions, in 2021, Xi *et al.* [32] proposed a multidirection network, which utilized rich spectral and spatial features through multidirection samples. To enhance the ability to measure different sample relations, Hong *et al.* [33] proposed a minibatch graph convolutional network (GCN) and combined it with CNN to work with irregular data. And Cai *et al.* [34] adopted a convolution layer after two-layer graphs to obtain the deep features' relationship. Furthermore, He *et al.* [35] utilized two GCNs to achieve feature extraction and label distribution learning, and the classification results demonstrated the GCN's potential for HSI processing. In 2022, Xi *et al.* [36] further involved different scales of neighborhoods in the GCN's adjacency matrices' construction. They also designed an innovative prototypical layer to enhance node features.

To better process sequential spectral signatures, Hong *et al.* [37] first introduced a transformer without any preprocessing operation into the HSI classification field, and they constructed the Spectral-Former that achieved significant improvement compared to other classic transformers. To further make full use of the transformer, Liu *et al.* [38] constructed a dual-path network (DP-Net) that utilized a transformer to process spatial and spectral information separately. And Hu *et al.* [39] further utilized contrastive learning to cooperate with a vision transformer to extract the HSI features in an unsupervised way and achieved satisfactory classification performance. From the above-mentioned methods, it is clear that spectral–spatial based methods can make full use of the HSI information, and CNN can be further combined with other networks to mine spectral–spatial information. Spectral–spatial based methods can exploit information from spectral and spatial domains, but different spatial areas and spectral bands usually contain information of varying importance. Thus, to focus on more informative features, an attention mechanism is introduced to the HSI classification fields.

Until now, various attention blocks have been proposed and adopted in the HSI classification field to emphasize target areas and suppress less useful regions. In 2018, Woo *et al.* [40] proposed a convolutional block attention module to achieve adaptive feature refinement from the channel and spatial dimensions. The module took advantage of two different spatial context features, which improved the representation powers of the network. To achieve feature refinement in the HSI classification field, in 2019, Ma *et al.* [41] adopted the submodule of CBAM as a spectral and spatial attention block in two branches to extract more discriminative features. To reduce the computational cost and increase the efficiency of the attention module, Wang *et al.* [42] designed an efficient channel attention block to learn effective attention weights by considering multiple channel interactions. And in 2020, Roy *et al.* [43] proposed an attention-based residual network (A^2S^2K), which adopted a feature recalibration mechanism to enhance extracted features. In 2021, Xue *et al.* [44] designed a second-order pooling operator to distinguish representative features, and the overall network achieved satisfactory classification results with the disjoint training samples [45]. Adopting disjoint training samples can greatly reduce the overlap between training and testing samples and provide a more objective and accurate evaluation. To achieve the full use of the attention mechanism, Hang *et al.* [46] constructed an attention-aided CNN, which utilized two attention-involved CNNs to achieve spectral and spatial feature extraction. Final results are obtained via an adaptively weighted summation method. It is clear that adopting the attention mechanism in the network can extract more discriminative spectral and spatial features for HSI classification.

In summary, lots of CNN-based methods have been constructed to extract HSI's spectral–spatial features and achieve high-accuracy classification results. But those methods usually have high computational complexity and massive network parameters. Recently, GCN and transformer have been introduced to the HSI classification field, which can further cooperate with CNN to model the relations between samples and process the sequential spectral data. Still, they are computationally expensive and have massive parameters [47]. Therefore, in this article, we aim to construct a triple-path network (TP-Net) that can achieve spectral–spatial feature utilization while having fewer parameters and lower computational complexity.

To take advantage of the joint information from the spectral and spatial domains, a hybrid information extraction branch is proposed. The extension is further combined with the spatial and spectral branches to construct a TP-Net for classification. The hybrid branch aims to explore the correlative information between spectral and spatial domains, which is later utilized as complementary information for purely spectral and spatial features. And by utilizing size-optimized convolution layers in different branches, the network parameters are further limited. In this article, the spectral and spatial features are extracted by corresponding channels and augmented by two modified attention mechanisms, including spectral and spatial attention modules. Except for the additive fusion of the three branches at the end of the network, the interleave-attention mechanism (Inter-Att) is elaborately designed to increase the interaction of data from different branches as they propagate through the network. Specifically, attention weight values obtained by one branch are further weighted for the other two branches for corresponding feature enhancement. Because the weight values are updated adaptively during the backpropagation process, one branch's updating weight values will affect other branches' features at the same time. By adopting the Inter-Att, the correlation between the branches can be enhanced, making the final fused features more representative. The main contributions can be summarized as follows.

1) A hybrid branch is constructed and works in parallel with the spectral and spatial branches to form a TP-Net for

HSI classification. The hybrid branch aims to capture the comprehensive spectral–spatial joint features and utilize them as complementary information of spectral and spatial features extracted by the other two branches.

2) An Inter-Att is proposed to increase interaction between different branches. By sharing each branch's obtained attention weight values with other branches, the features of each branch can be constrained by each other and make the final fused spectral–spatial features more representative. And the Inter-Att increases network performance, but it barely brings any increase in parameters.
3) Inspired by other researchers' work, two attention modules are introduced to extract features from informative spectral bands and spatial regions. In the spatial branch, a spatial attention block (SPA-Att) is developed and utilizes two different spatial context features to obtain the representative spatial weights. In the spectral branch, a channel attention block is constructed to obtain reliable channel weights by considering the local and cross-spectral band correlation.
4) Compared with several modern CNN methods, the proposed TP-Net can achieve higher classification accuracy with limited parameters.

The rest of this article is organized as follows. Section II introduces the prior knowledge that is related to the proposed network. Section III describes the proposed TP-Net's detailed information. Section IV gives a detailed description of datasets, related classification experiments, and analysis. Finally, Section V concludes this article.

II. Related Work

A. Cube-Based Methods for HSI Classification

Pixel-based methods [48], [49] mainly rely on abundant spectral information for HSI classification, whereas cube-based methods [50]–[52] take advantage of spectral and spatial information. Those methods have gained attention and allowed significant progress. We will introduce the detailed process in the remainder of this section.

In the early stages, researchers took advantage of a single pixel's abundant spectral information for classification. Mou *et al.* [48] utilized recurrent neural networks to analyze hyperspectral pixels as sequential data for classification. In [49], Li *et al.* utilized CNN to extract pixel-pair features and obtained the final label by voting strategy. Pixel-based methods utilize spectral information but lack spatial context information. In later stages, researchers used a pixel-centered 3-D data cube that naturally contains both spatial context and spectral information as the input of the classification network, which achieved remarkable progress. In [50], Zhang *et al.* proposed a region-based CNN that extracted spatial context features from the 3-D cube and further combined spectral information to gain more discriminative joint features. In [52], Wei *et al.* further combined multiscale strategies to exploit the 3-D cube's spatial and spectral information. In this article, the proposed TP-Net is also based on the 3-D data cube, but we further combined multipath, attention mechanism, and interleave-attention strategy to extract more discriminant features.

B. Convolution Operation for HSI Feature Extraction

The convolution operation is the key part of the CNN to extract the input data features automatically. By utilizing this operation, the feature maps can be obtained, and it can be described as follows:

$$x^{l+1} = w^{(l+1)}x^{l} + b^{(l+1)} \tag{1}$$

where x^{l} and x^{l+1} stand for the input and output of the $x+1$st convolutional layer, respectively, and w^{l+1} and b^{l+1} refer to the $l+1$st layer's weights and biases, respectively.

As for the HSI classification field, three types of convolution operations are utilized to extract features, which are 1-D, 2-D, and 3-D convolution operations. The 1-D convolution is commonly used for the extraction of spectral information, and it has been directly used in pixel-based methods for classification in the early stages. The 1-D convolution is more suitable to process sequence data, whereas the 2-D convolution is later introduced to handle HSI spatial information.

The 1-D and 2-D convolution operations can cooperate to achieve joint utilization of spectral and spatial information. For example, Yang *et al.* [53] constructed a two-channel network for HSI classification, which utilized 1-D CNN to achieve spectral information utilization in one channel and 2-D CNN for spatial features in the other. By fusing two channels' features, more representative spectral–spatial joint features can be obtained. Hybrid 1-D and 2-D convolution can achieve spectral and spatial information utilization, but the extracted features are relatively independent and lack correlation; 3-D convolution is further adopted in HSI feature extraction to directly extract the joint spectral–spatial features.

Usage of 3-D convolution can directly capture HSI's spectral and spatial features and simplify the feature fusion process, but its most prominent problem is the heavy network parameters. To reduce the network parameters, Zhong *et al.* in [28] adopted 3-D convolutions of size $(1 \times 1 \times X)$ for achieving spectral feature extraction. Furthermore, in [41], Ma *et al.* utilized two types of convolutions with specific sizes $(X \times X \times 1)$ and $(1 \times 1 \times X)$ to extract spatial and spectral features and further reduce the number of parameters. It is clear that by optimizing the configuration of 3-D convolution, the network parameters can be greatly reduced while maintaining high-precision performance.

In our case, we constructed a 3-D-convolution-based TP-Net to make full use of HSI's abundant spectral and spatial information. To alle///viate the problem of heavy network parameters, in the hybrid branch, we adopted a 3-D convolution of small kernel size to capture HSI's joint features. Meanwhile, in spatial and spectral branches, two types of convolutions with specific sizes $(X \times X \times 1)$ and $(1 \times 1 \times X)$ are adopted to achieve feature extraction. By utilizing size-optimized convolution layers, a TP-Net with limited parameters is constructed to achieve discriminative spectral–spatial feature extraction.

C. Attention Mechanism

Until now, attention mechanisms have been introduced in many tasks, i.e., text processing [54], changing object detection [55], visual question answering [56], and HSI classification.

By utilizing the attention mechanism, deep-learning-based networks can focus on information of the interested target and differentiate low-correlation information, which enhances network feature extraction ability.

When the attention mechanism is applied to the HSI classification field, it shows remarkable performance in stressing informative spectral bands and spatial regions [31], [33]. To select suitable parts from spectral bands, Dong *et al.* [57] proposed a band attention module (BAM) to emphasize important spectral bands, and classification results indicate that the BAM can effectively improve network performance. Except for the spectral domain, the attention mechanism is also utilized in the spatial domain to stress informative regions. Pande *et al.* [58] proposed an adaptive hybrid attention network constructed by two paths, one adopted 1-D convolution layers with the spectral attention block (SPE-Att) to obtain representative spectral features, and the other utilized 2-D convolution with the SPA-Att for discriminate spatial feature extraction. By fusing two paths' features, the network obtained more comprehensive HSI features. From the previous studies, it is obvious that the attention mechanism can be used as an effective strategy to extract more discriminating features and improve network performance. Based on the aforementioned attention blocks, we introduced two attention modules for spectral and spatial feature enhancement. For the spatial domain, we utilize two different spatial context features to obtain the representative spatial weights. As for the spectral domain, local and cross-spectral bands' correlation is considered and used for spectral weight acquisition. Furthermore, the Inter-Att is adopted to achieve better performance.

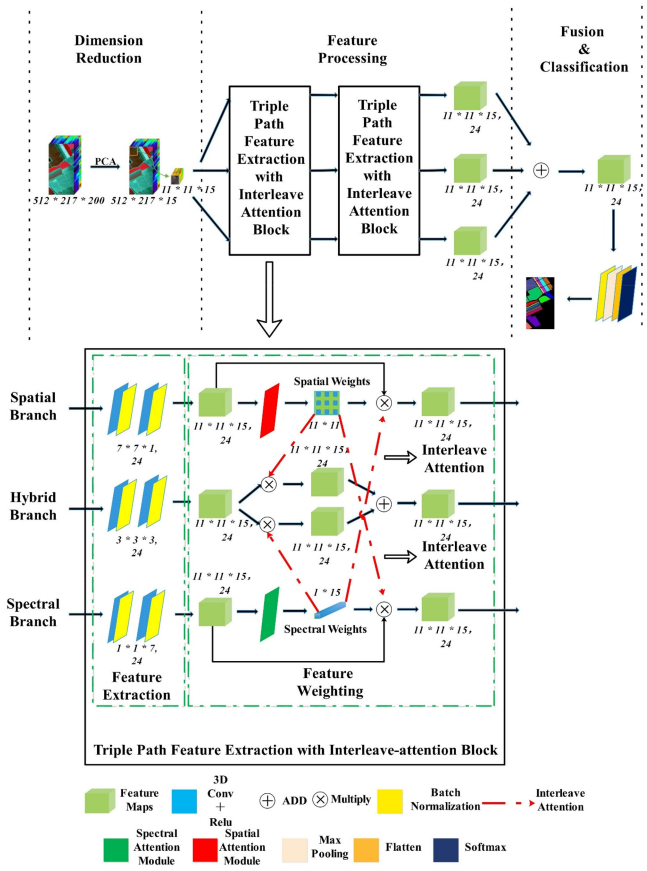


Fig. 1. Structure of the triple-path spectral–spatial network with interleave-attention.

III. Proposed Methods

In this section, we will first introduce the structure of the whole TP-Net and later give detailed information about its substructures, which are spectral and spatial attention blocks and the Inter-Att.

A. Triple-Path Spectral–Spatial Network With Interleave-Attention

The proposed network is mainly composed of two triple-path feature extractions with interleave-attention blocks. These blocks are constructed by three parallel branches, and they can be abstracted into the following three processes.

1) Feature extraction: Corresponding features are obtained from spatial, spectral, and spatial–spectral joint domains.

2) Feature weighting: We constructed spatial and spectral attention modules to emphasize informative bands and regions.

3) Inter-Att: By adopting this mechanism in different stages of the feature extraction process, the features of each branch can be constrained by each other and makes the final fused spectral–spatial features more representative.

Fig. 1 gives the whole TP-Net structure.

As shown in Fig. 1, we take the Salinas Valley (SV) dataset as an input example. The proposed network can be divided into the following three stages:

1) dimension reduction;

2) feature processing;

3) fusion and classification.

In the first stage, principal component analysis is introduced to reduce the spectral dimension to 15 and retain 99% of initial information [59], [60]. Then, the HSI data cube of size ($11 \times 11 \times 15$) is cropped and sent to the second stage for feature processing.

In the second stage, the input data are sent to the triple-path block for feature extraction and weighting. This block is mainly composed of three branches, i.e., hybrid, spectral, and spatial branches. In the hybrid branch, we used 3-D convolution layers of size ($3 \times 3 \times 3$, 24) to obtain comprehensive spectral–spatial joint features. As for the spectral and spatial branches, 3-D convolution layers of kernel sizes ($7 \times 7 \times 1$, 24) and ($1 \times 1 \times 7$, 24) are used to extract features from spatial and spectral domains separately. All the convolution layers' stride is (1, 1, 1), and we adopted the "same" padding strategy to keep the size of all the output feature maps unchanged. After two-times feature extraction, two attention modules are constructed and adopted in the spectral and spatial branches to achieve feature weighting. Meanwhile, the Inter-Att is elaborately designed to increase the interaction of data from different branches as they propagate through the network. To be specific, attention weights obtained by one branch are shared with the other two for feature readjustment, which enhances the correlation between the branches and makes the fused features more representative.

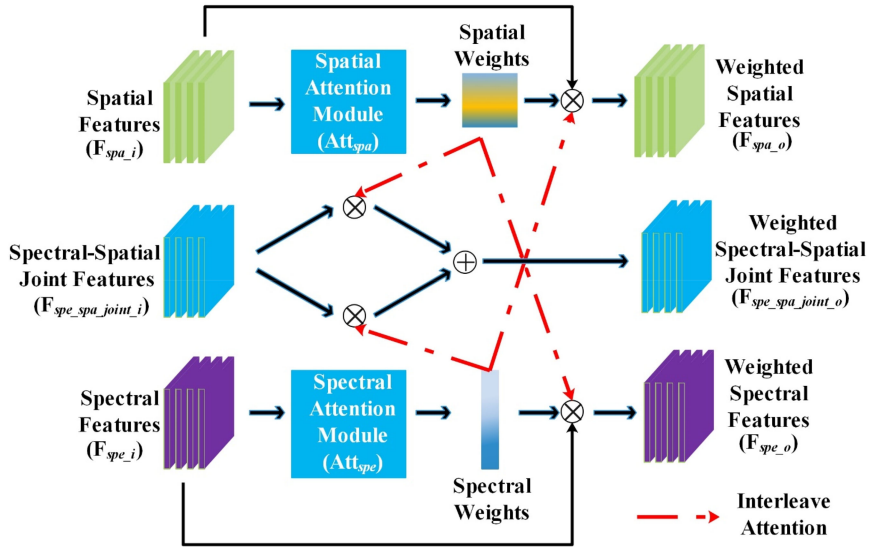


Fig. 2. Inter-Att adopted in our proposed network.

The above-mentioned steps constitute a triple-path feature extraction with an interleave-attention block, and two same blocks are stacked in the proposed network for deep abstract feature extraction. After obtaining three branches' weighted features, we fused those features for classification in the final stage.

In the final stage, we fused the three branches' well-extracted features to obtain representative spectral–spatial features of size ($11 \times 11 \times 15$, 24), and the fused features are sent into the Max-Pooling, Flatten layer, and Softmax classifier sequentially to obtain classification results.

B. Interleave-Attention Mechanism

To increase interaction between different branches for better feature fusion, the Inter-Att is proposed and adopted in different stages of the feature extraction process. To be specific, attention weights obtained by one branch are shared with the other two for feature readjustment, which enhances the correlation between the branches and makes the fused features more representative. Fig. 2 illustrates the proposed Inter-Att adopted in the network.

As shown in Fig. 2, first, two specially designed attention modules are adopted in spatial and spectral branches to obtain corresponding weight values from well-extracted features. Then, the obtained weights are shared with other branches for feature readjustment, which makes the features of each branch constrained by each other. Finally, three branches' features are enhanced from both spectral and spatial domains by feature weighting. The Inter-Att adopted in the triple branches can be described as follows:

$$F_{\text{spa\_}o} = F_{\text{spa\_}i} * \text{Att}_{\text{spa}}(F_{\text{spa\_}i}) * \text{Att}_{\text{spe}}(F_{\text{spe\_}i}) \quad (2)$$

$$F_{\text{spe\_spa\_}o} = F_{\text{spe\_spa\_}i} * [\text{Att}_{\text{spa}}(F_{\text{spa\_}i}) + \text{Att}_{\text{spe}}(F_{\text{spe\_}i})] \quad (3)$$

$$F_{\text{spe\_}o} = F_{\text{spe\_}i} * \text{Att}_{\text{spa}}(F_{\text{spa\_}i}) * \text{Att}_{\text{spe}}(F_{\text{spe\_}i}) \quad (4)$$

where F and Att stand for feature maps and attention block, respectively; subscripts spa, spe, and spe_spa stand for spatial, spectral, and spatial–spectral joint domains, respectively; subscripts i and o represent input and output, respectively.

C. Spectral and Spatial Attention Module

In the spectral branch, a dual-channel spectral attention module is introduced and adopted to emphasize informative spectral bands by evaluating local and cross-spectral bands' correlation. The module utilizes two parallel convolution channels with specified kernel sizes to obtain local and cross-spectral bands' weight values, which are further fused to obtain comprehensive spectral weights. By utilizing convolution's adaptive weight updating mechanism, the weight of each spectral band can be adaptively adjusted to make the feature extraction process more efficient. Fig. 3 illustrates the architecture of the spectral attention module.

First, the average pooling layer is utilized to process the input feature maps of size ($H \times W \times C$) to gain spectral vectors of size ($1 \times 1 \times C$), which ensures that the spatial information has no impact on the acquisition of spectral weight. Then, two parallel convolution channels are utilized to process the obtained vectors; one with a kernel size of ($1 \times 1 \times 1$) for local weight acquisition, and the other with a kernel size of ($1 \times 1 \times 3$) for cross weight acquisition. And through the Sigmoid activation function, the local and cross-spectral band weight vectors of size ($1 \times 1 \times C$) are obtained. Finally, two weight vectors are fused by addition operation and multiplied with the input feature maps to realize spectral feature weighting. The SPE-Att can be described as follows:

$$F_o = F_i * \{\text{Conv}_1[\text{AP}(F_i)] + \text{Conv}_3[GAP(F_i)]\} \quad (5)$$

where F_o and F_i are the output and input feature maps, respectively; Conv stands for the 3-D convolution operation, and the subscript number represents the kernel size; AP stands for the average pooling.

In the spatial branch, an adaptive-weight-adjustment-mechanism-based spatial attention module is introduced to

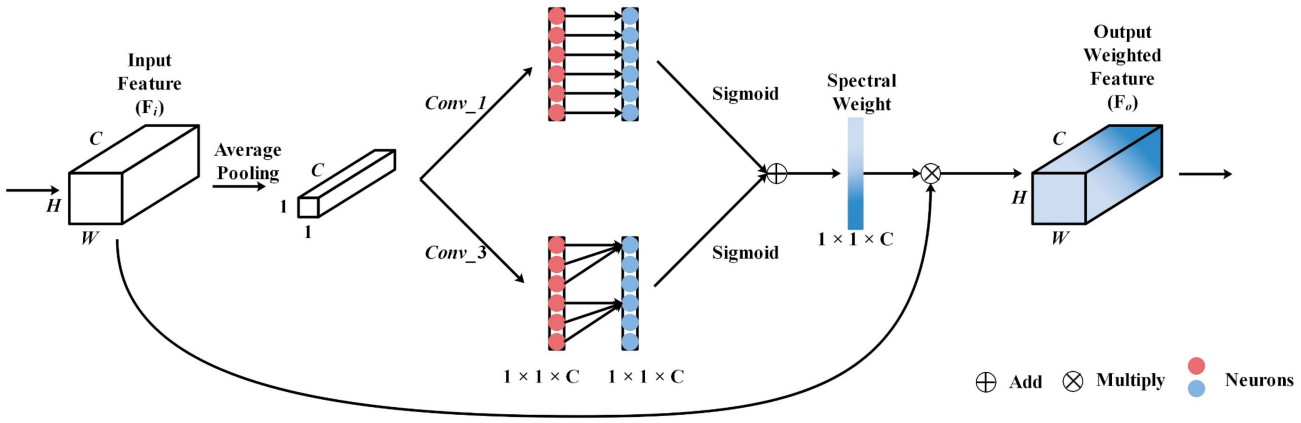


Fig. 3. Proposed spectral attention module.

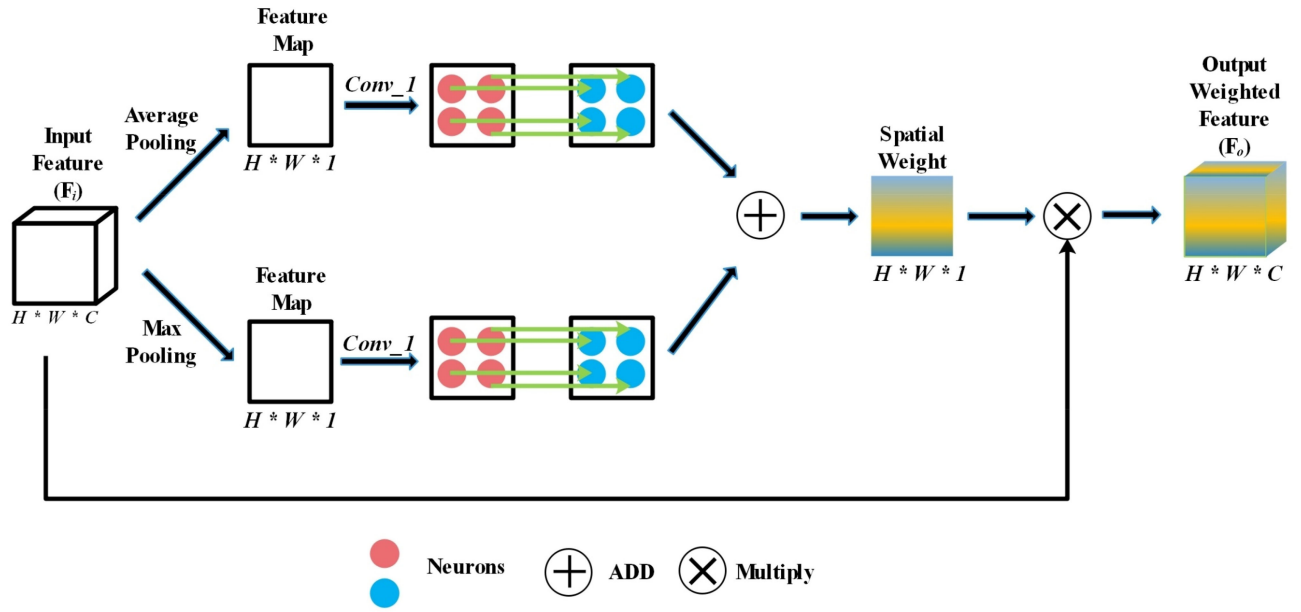


Fig. 4. Proposed spatial attention module.

stress meaningful spatial areas. The module utilizes convolution layers to process two different spatial context features from max-pooling and average-pooling layers in parallel. And by the fusing process, representative spatial weights can be obtained. Fig. 4 gives the detailed structure of the spatial attention module.

First, the input feature maps of size ($H \times W \times C$) are sent to two branches, in which two kinds of pooling layers are utilized to generate average and max feature maps of size ($H \times W \times 1$) and a convolution layer of size ($1 \times 1 \times 1$) is adopted to obtain spatial weight values of two different spatial context features. Then, the two spatial weights are fused by addition operation to obtain the comprehensive weights of size ($H \times W \times 1$). Finally, spatial weighted features can be obtained by multiplying the input with spatial weights. The spatial attention module is described as follows:

$$F_o = F_i * \{\text{Conv}_1[\text{AVE}(F_i)] + \text{Conv}_1[\text{MAX}(F_i)]\} \quad (6)$$

where F_o and F_i are the output and input feature maps, respectively; Conv stands for the 3-D convolution operation and the subscript number represents the kernel size; AVE and MAX stand for the average-pooling and max-pooling, respectively.

IV. Experiment Results

In this section, we utilized four widely used HSI datasets to verify network effectiveness. And the network configuration and classification results are also displayed. We also performed some ablation experiments to better explain the proposed network.

A. Datasets

In this article, four widely used HSI datasets are adopted in the experiment to evaluate the effectiveness of the proposed TP-Net, which are Indian Pines (IP), Pavia University (PU), SV, and Houston University (HU).

TABLE I
SAMPLE NUMBERS IN THE IP DATASET

Label	Class	Numbers
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings grass-trees-driverss	386
16	Stone-steel-towers	93
	Total	10 249

TABLE II
SAMPLE NUMBERS IN THE PU DATASET

Label	Class	Numbers
1	Asphalt	6631
2	Meadows	18 649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare soil	5029
7	Bitumen	1330
8	Self-blocking bricks	3682
9	Shadows	947
	Total	42 776

IP: The first dataset is obtained using an airborne spectrometer, which contains 16 classes of ground coverage, and after removing spectral bands that cover water absorption features, the remaining 200 bands with 145 × 145 spatial pixels are utilized for the experiment.

PU: The second dataset is gathered using a reflective optics imaging spectrometer, which contains 9 classes of ground coverage, and after removing 12 noisy bands, the remaining 103 bands with 610 × 340 spatial pixels are adopted in this article.

SV: The third dataset is also gathered using an airborne spectrometer, which contains 16 classes of ground coverage, and after removing 20 spectral bands that cover water absorption features, the remaining 204 bands with 512 × 217 spatial pixels are utilized.

HU: The fourth dataset is provided by the GRSS data fusion contest, which contains 15 classes of ground coverage. The data's spectral band number is 144, and the spatial resolution is 349 × 1905.

Tables I –IV give the detailed information of each dataset.

TABLE III
SAMPLE NUMBERS IN THE SV DATASET

Label	Class	Numbers
1	Brocoli-green-weeds-1	1969
2	Brocoli-green-weeds-2	3652
3	Fallow	1938
4	Fallow-rough-plow	1368
5	Fallow-smooth	2626
6	Stubble	3881
7	Celery	3509
8	Grapes-untrained	11 047
9	Soil-vinyard-develop	6079
10	Corn-senesced-green-weeds	3214
11	Lettuce-romaine-4wk	1048
12	Lettuce-romaine-5wk	1889
13	Lettuce-romaine-6wk	898
14	Lettuce-romaine-7wk	1050
15	Vinyard-untrained	7124
16	Vinyard-vertical-trellis	1771
	Total	53 063

TABLE IV
SAMPLE NUMBERS IN THE HU DATASET

Label	Class	Numbers
1	Healthy grass	1251
2	Stressed grass	1254
3	Synthetic grass	697
4	Trees	1244
5	Soil	1242
6	Water	325
7	Residential	1268
8	Commercial	1244
9	Road	1252
10	Highway	1227
11	Railway	1235
12	Parking Lot1	1233
13	Parking Lot2	469
14	Tennis court	428
15	Running track	660
	Total	15 029

B. Experimental Setting

In this article, five well-known CNN-based methods are selected for comparison, including the ResNet [61], the PyResNet [62], the SSRN [28], the HybridSN [63], A^2S^2K [43], and Attention-Based Second-Order Pooling Network (A-SPN) [44]. All networks are performed on a Tesla K40c 12-GB GPU, and the input spatial size is set to 11 × 11. For the proposed network, the batch size and learning rate are set to 16 and 0.0003, respectively. The RMSprop optimizer and categorical-cross-entropy loss function are selected for network training.

All methods are trained for 200 epochs, and we spilt each HSI dataset into three sets, i.e., training, validation, and testing set. To be specific, for the IP dataset, 20%, 10%, and 70% of samples are randomly selected for network training, validation, and testing,

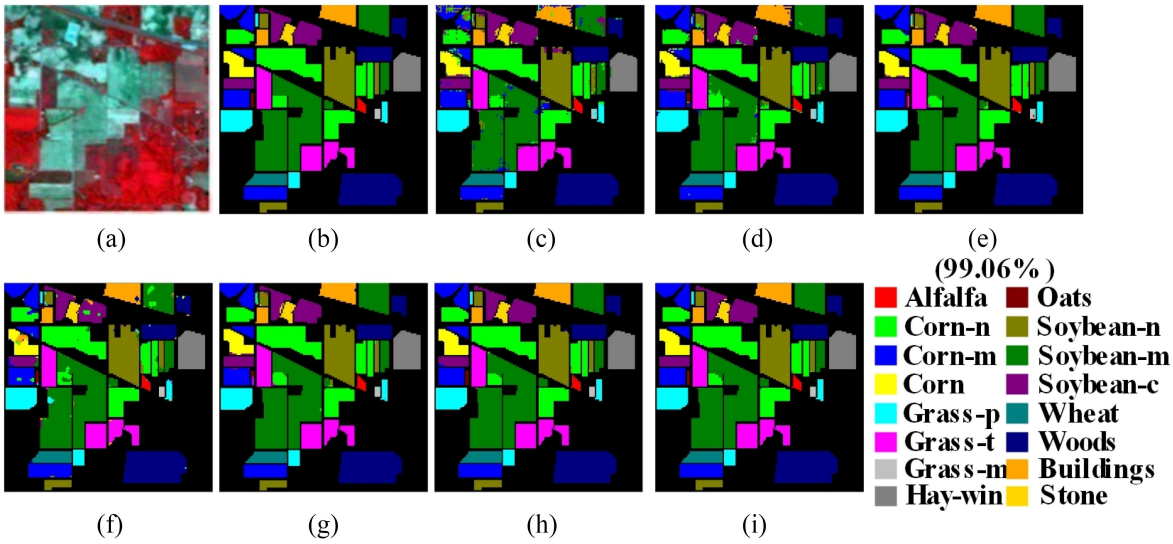


Fig. 5. Classification maps for the IP dataset. (a) and (b) False-color (FC) and ground truth (GT). (c)–(i) Classification maps of different networks. (a) FC. (b) GT. (c) ResNet (94.68%). (d) PyResNet (95.96%). (e) SSRN (99.06%). (f) HybridSN (98.47%). (g) A^2S^2K (99.26%). (h) A-SPN (99.52%). (i) Proposed (99.64%).

TABLE V
CLASSIFICATION RESULTS OF THE IP DATASET USING 20% TRAINING SAMPLES

Label	ResNet	PyResNet	SSRN	HybridSN	A^2S^2K	A-SPN	Proposed
1	100.00	96.42	93.54	100.00	96.77	99.72	100.00
2	90.87	91.89	99.76	99.21	98.52	99.50	99.82
3	83.85	93.38	99.18	98.35	99.17	99.66	100.00
4	100.00	95.45	96.45	99.24	99.25	100.00	100.00
5	97.53	98.25	98.62	99.15	99.31	99.76	100.00
6	99.54	99.54	97.79	99.26	98.00	100.00	100.00
7	100.00	100.00	90.90	96.35	100.00	100.00	100.00
8	100.00	99.64	100.00	99.21	100.00	100.00	100.00
9	85.71	100.00	100.00	100.00	100.00	98.75	100.00
10	87.68	96.08	97.97	98.42	98.46	99.58	99.74
11	98.26	99.57	99.39	98.36	99.79	99.72	99.29
12	79.21	93.27	96.48	98.54	94.93	98.75	99.58
13	100.00	98.27	100.00	99.01	100.00	99.81	100.00
14	96.47	95.88	99.48	99.26	99.74	100.00	100.00
15	98.94	99.00	97.82	98.09	100.00	100.00	100.00
16	81.69	93.33	98.24	99.48	89.23	99.86	96.00
OA (100%)	94.68	95.96	99.06	98.47	99.26	99.52	**99.64**
	± 3.053	± 1.182	± 0.323	± 0.037	± 0.412	± 0.012	± 0.036
AA (100%)	94.07	95.90	97.85	98.10	99.16	99.49	**99.64**
	± 3.658	± 1.953	± 1.682	± 0.505	± 0.399	± 0.008	± 0.026
K × 100	93.93	95.40	98.93	98.26	99.16	99.51	**99.60**
	± 3.500	± 1.341	± 0.369	± 0.049	± 0.471	± 0.031	± 0.044

respectively. For PU, SV, and HU datasets, the spilled ratios are 10%, 10%, and 80%. To quantify the experimental results, we adopted overall accuracy (OA), average accuracy (AA), and Kappa coefficient (K) in this article. For all networks, we repeat the whole training process five times to report the classification results with a form of mean accuracy ± the standard deviation.

C. Classification Results

1) Classification Results of the IP Dataset: All networks' classification maps and the corresponding accuracy indexes of the IP dataset are provided in Fig. 5 and Table V, and the best classification accuracies are in bold.

As shown in Table V, it is obvious that our TP-Net has gained the highest classification accuracies, with 99.64% OA, 99.64% AA, and 99.60% K, which verified the effectiveness of our network. And because the samples of the IP dataset are unbalanced and some categories have fewer samples, all methods obtained low classification accuracies in some categories. In terms of each category's classification accuracy, the proposed method can achieve at least an accuracy of more than 95%. Compared with ResNet and PyResNet, classification accuracies of classes 2, 3, 12, and 16, which are corn-notill, corn-mintill,

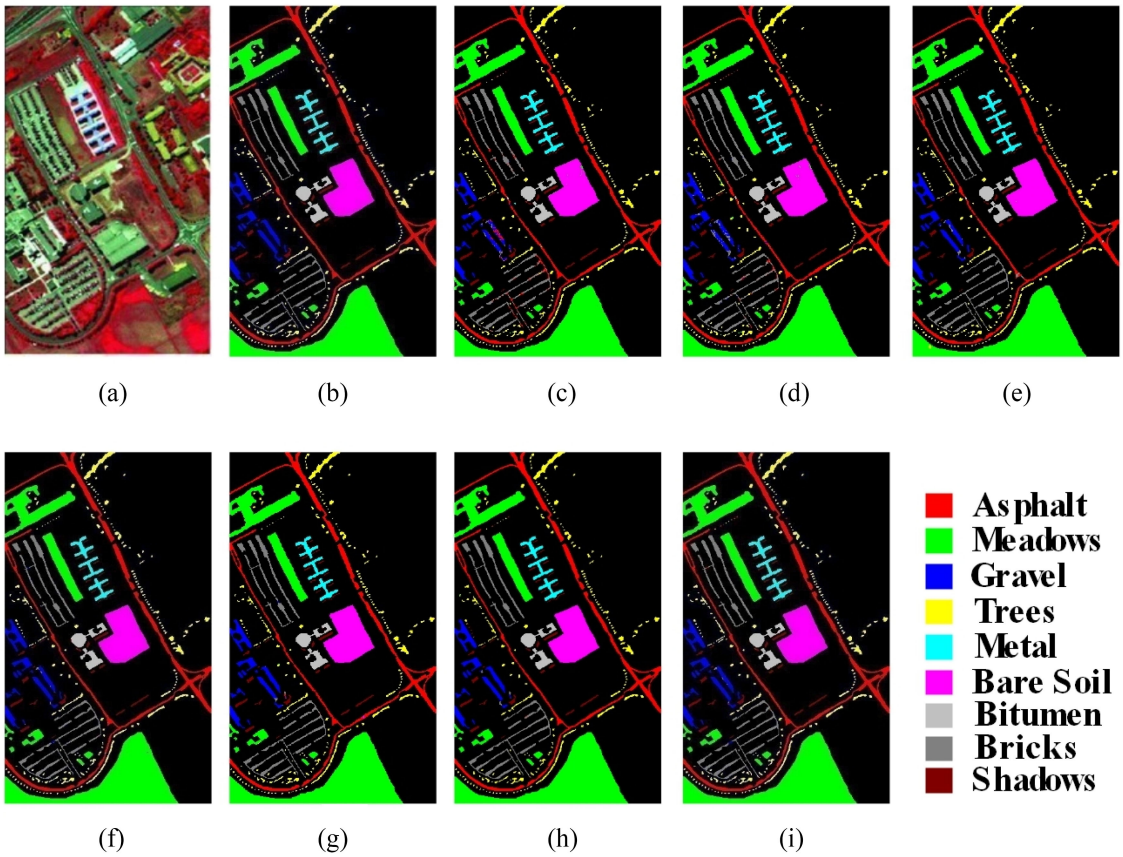


Fig. 6. Classification maps for the PU dataset. (a) and (b) False-color (FC) and ground truth (GT). (c)–(i) Classification maps of different networks. (a) FC. (b) GT. (c) ResNet (98.65%). (d) PyResNet (98.82%). (e) SSRN (99.52%). (f) HybridSN (99.64%). (g) A^2S^2K (99.51%). (h) A-SPN (99.70%). (i) Proposed (99.82%).

soybean-clean, and stone-steel-towers, respectively, have significantly improved by at least 5%. Compared to SSRN, the proposed network has improved 3% accuracies in 1, 4, 7, and 12 classes, which are alfalfa, corn, grass-pasture-mowed, and soybean-clean, respectively. When compared to HybridSN, the classification accuracy of the grass-pasture-mowed improved from 96.35% to 100%. As for A^2S^2K, classification accuracies of 1, 12, and 16, which are alfalfa, soybean-clean, and stone-steel-towers, respectively, have significantly improved by at least 3%. Compared to A-SPN, the proposed network has improved the accuracy of classes 1, 3, 5, 9, and 13, which are alfalfa, corn-mintill, grass-pasture, oats, and wheat, respectively, to 100%. As shown in Fig. 5, our TP-Net's classification map has less noise and each classification region's shape is smoother.

2) Classification Results of the PU Dataset: Fig. 6 shows each network's classification maps of the PU dataset; Table VI provides the corresponding results, and the best classification accuracies are in bold.

From Table VI, it is obvious that our TP-Net obtained the highest accuracies, with 99.82% OA, 99.63% AA, and 99.77% K. Compared to ResNet, the proposed network has significantly improved the classification accuracy of class 1, which is asphalt, from 92.37% to 100%. For PyResNet, the classification accuracy of class 8, which is self-blocking bricks, has improved from 92.21% to 99.22%. Compared to SSRN, the accuracies of classifying class 7, bitumen, and class 8, self-blocking bricks, are improved by at least 2%. As for HybridSN and A^2S^2K, the TP-Net has gained better accuracy in all classes. Compared to A-SPN, the proposed network has gained better OA and K, but lower AA. As can be seen from Fig. 6, all networks have obtained satisfactory classification maps with little noise since the samples of all classes are sufficient. Furthermore, it is obvious that the proposed TP-Net's classification map has no obvious misclassification area, and hot pixels can barely be seen.

3) Classification Results of the SV Dataset: All networks' classification maps and corresponding accuracy indexes of the SV dataset are provided in Fig. 7 and Table VII; the best classification accuracies are in bold.

As shown in Table VII, the proposed TP-Net obtained the highest classification accuracies, with 100% OA, AA, and K. It is clear that all networks have achieved excellent classification results since the samples of the SV dataset are balanced and sufficient. Compared to ResNet and PyResNet, the TP-Net has improved the classification accuracies of classes 4, 8, and 15, which are fallow-rough-plow, grapes-untrained, and vinyard-untrained, respectively, to 100%. As for SSRN, HybridSN, A^2S^2K, and A-SPN, the proposed network has gained accuracy improvement in all classes. As can be seen from Fig. 7, the

TABLE VI
CLASSIFICATION RESULTS OF THE PU DATASET USING 10% TRAINING SAMPLES

Label	ResNet	PyResNet	SSRN	HybridSN	A^2S^2K	A-SPN	Proposed
1	92.37	97.67	99.98	100.00	99.71	99.87	100.00
2	99.75	99.65	100.00	100.00	99.93	100.00	99.99
3	99.38	97.52	100.00	98.65	99.63	99.52	99.89
4	99.30	99.49	98.11	99.54	99.95	99.51	99.38
5	99.62	99.90	100.00	100.00	100.00	100.00	100.00
6	99.85	99.82	100.00	100.00	99.65	100.00	99.98
7	99.24	97.12	97.40	100.00	100.00	99.82	100.00
8	96.58	92.21	96.81	98.96	98.76	99.46	99.22
9	98.95	98.82	100.00	100.00	98.45	99.89	100.00
OA (100%)	98.65	98.82	99.52	99.64	99.51	99.70	**99.82**
	± 0.273	± 0.263	± 0.861	± 0.241	± 0.721	± 0.721	± 0.023
AA (100%)	98.31	98.36	99.38	99.40	99.24	**99.66**	99.63
	± 0.312	± 0.323	± 0.849	± 0.851	± 0.348	± 0.721	± 0.039
K × 100	98.22	98.44	99.36	99.53	99.35	99.67	**99.77**
	± 0.363	± 0.351	± 1.143	± 0.403	±0.954	± 0.721	± 0.032

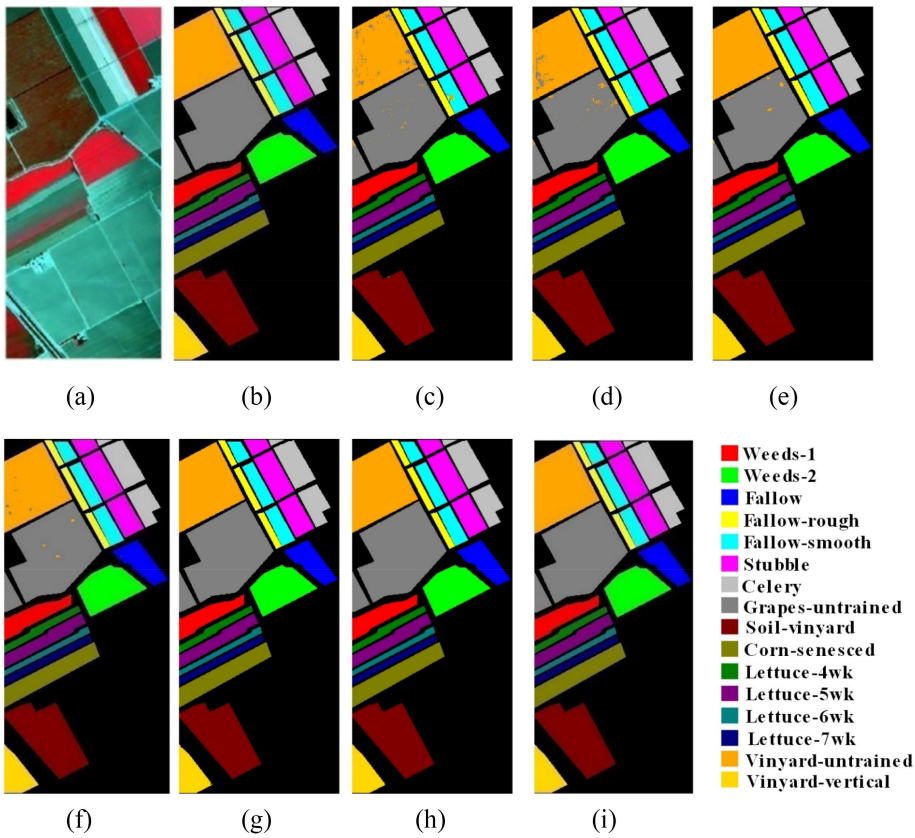


Fig. 7. Classification maps for the SV dataset. (a) and (b) False-color (FC) and ground truth (GT). (c)–(i) Classification maps of different networks. (a) FC. (b) GT. (c) ResNet (98.14%). (d) PyResNet (98.85%). (e) SSRN (99.93%). (f) HybridSN (99.80%). (g) A^2S^2K (99.95%). (h) A-SPN (99.96%). (i) Proposed (100.00%).

TP-Net's classification map owns no mislabeled pixels, and other methods contain more or less misclassified pixels.

4) Classification Results of the HU Dataset: All networks' classification maps and corresponding accuracy indexes of the HU dataset are given in Fig. 8 and Table VIII. The best classification accuracies are in bold.

As shown in Table VIII, it is clear that our TP-Net has achieved a significant accuracy improvement, with 99.72% OA, 99.76% AA, and 99.69% K. Compared to ResNet and PyResNet, classification accuracies of classes 1, 8, 10, 11, 12, and 13, which are healthy grass, commercial, highway, railway, parking lot1, and parking lot2, respectively, have improved by at

TABLE VII
CLASSIFICATION RESULTS OF THE SV DATASET USING 10% TRAINING SAMPLES

Label	ResNet	PyResNet	SSRN	HybridSN	A^2S^2K	A-SPN	Proposed
1	100.00	100.00	100.00	100.00	100.00	100.00	100.00
2	100.00	100.00	100.00	100.00	99.93	100.00	100.00
3	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	97.21	98.93	100.00	99.88	100.00	99.97	100.00
5	99.90	99.85	100.00	100.00	100.00	99.74	100.00
6	100.00	100.00	100.00	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	100.00	100.00
8	96.58	97.56	100.00	100.00	100.00	99.96	100.00
9	100.00	99.79	99.97	100.00	100.00	100.00	100.00
10	98.83	100.00	99.80	100.00	99.69	99.98	100.00
11	99.88	100.00	99.64	99.98	100.00	100.00	100.00
12	99.67	99.61	100.00	100.00	99.29	99.99	100.00
13	100.00	99.58	99.86	99.85	100.00	100.00	100.00
14	100.00	100.00	99.41	100.00	99.64	99.65	100.00
15	98.70	97.32	98.72	99.48	99.96	99.99	100.00
16	100.00	100.00	100.00	100.00	99.92	99.71	100.00
OA (100%)	98.14	98.85	99.93	99.80	99.95	99.96	**100.00**
	± 1.282	± 0.368	± 0.067	± 0.005	± 0.001	± 0.001	± 0.000
AA (100%)	99.08	99.27	99.93	99.88	99.92	99.93	**100.00**
	± 0.468	± 0.400	± 0.048	± 0.001	± 0.005	± 0.002	± 0.000
K × 100	97.93	98.72	99.92	99.78	99.95	99.95	**100.00**
	± 1.425	± 0.410	± 0.078	± 0.006	± 0.002	± 0.001	± 0.000

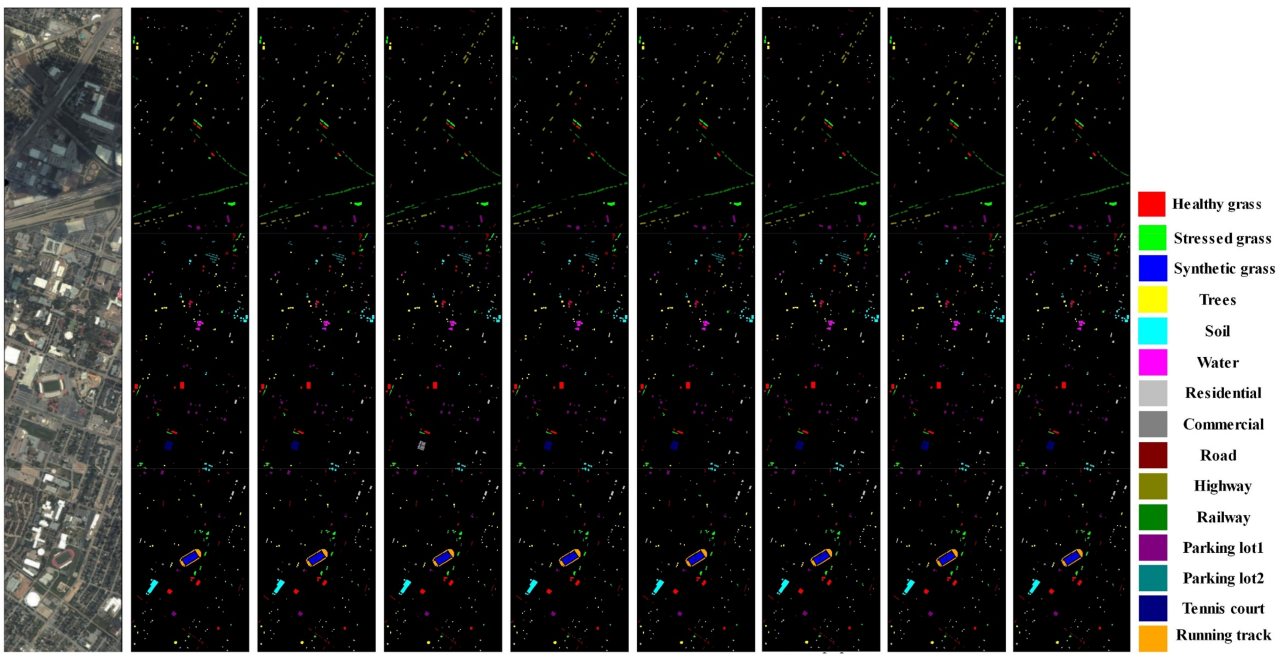


Fig. 8. Classification maps for the HU dataset. (a) and (b) False-color (FC) and ground truth (GT). (c)–(i) Classification maps of different networks. (a) FC. (b) GT. (c) ResNet (96.78%). (d) PyResNet (96.87%). (e) SSRN (98.98%). (f) HybridSN (98.55%). (g) A^2S^2K (99.18%). (h) A-SPN (99.25%). (i) Proposed (99.72%).

least 3%. Compared to SSRN and HybridSN, the proposed network has achieved an accuracy improvement in terms of each class. And compared to A^2S^2K and A-SPN, the proposed TP-Net has improved at least 1% in classes 9 and 13, which are road and parking lot2, respectively. The samples of the HU dataset are sparsely distributed in the spatial domain, which means the available context spatial information is limited. It demonstrates that the proposed network can achieve satisfactory classification results with less spatial context information.

TABLE VIII
CLASSIFICATION RESULTS OF THE HU DATASET USING 10% TRAINING SAMPLES

Label	ResNet	PyResNet	SSRN	HybridSN	A^2S^2K	A-SPN	Proposed
1	96.77	96.74	98.62	98.56	99.62	99.36	99.91
2	96.81	98.76	99.78	99.21	99.80	99.50	99.82
3	99.48	99.89	100.00	100.00	99.96	100.00	100.00
4	97.59	97.97	99.28	99.43	99.36	99.96	99.73
5	99.65	99.06	99.96	99.81	99.50	100.00	100.00
6	92.07	97.63	97.75	100.00	98.58	99.82	100.00
7	96.61	98.23	98.74	98.68	99.14	99.07	98.86
8	96.14	95.30	99.34	97.59	99.63	98.17	100.00
9	96.96	96.49	98.76	96.61	98.60	96.71	99.73
10	93.82	92.91	97.28	99.22	97.78	100.00	99.46
11	96.67	95.14	99.47	98.64	99.38	100.00	99.55
12	97.49	96.69	98.65	98.08	98.70	99.48	99.82
13	96.25	95.82	99.34	99.22	98.91	96.77	99.76
14	98.78	99.63	97.14	98.63	99.26	100.00	100.00
15	99.48	98.60	99.81	99.47	99.64	100.00	99.83
OA (100%)	96.78 ± 0.014	96.87 ± 0.012	98.98 ± 0.003	98.55 ± 0.020	99.18 ± 0.003	99.25 ± 0.041	**99.72** ± 0.001
AA (100%)	96.97 ± 0.015	97.26 ± 0.009	98.94 ± 0.004	98.62 ± 0.023	99.19 ± 0.004	99.25 ± 0.087	**99.76** ± 0.001
K × 100	96.52 ± 0.015	96.62 ± 0.013	98.90 ± 0.003	98.43 ± 0.025	99.11 ± 0.004	99.19 ± 0.060	**99.69** ± 0.002

D. Parameter Discussion

1) Investigation on Network Structure: To verify the effectiveness of the three proposed structures, including the following.

1) Hybrid-path network (HP-Net), which only uses a hybrid branch to achieve feature extraction.
2) DP-Net, which only contains two parallel paths to extract spectral and spatial features.
3) TP-Net, which directly utilized three paths to parallel extract deep abstract features.
4) TP-Net with SPA-Att, which added SPA-Att to enhance the spatial branch's features.
5) TP-Net with SPE-Att, which added SPE-Att to enhance the spectral branch's features.
6) TP-Net with spectral and spatial attention module (Att) was adopted to enhance features from spectral and spatial domains.
7) Based on TP-Net and Att, we further introduced the Inter-Att to make full use of the obtained spatial and spectral weights.

In this experiment, all hyperparameters remain the same as presented in Section IV-B. Table IX shows the detailed classification results using different structures.

As shown in Table IX, after adding various components, the accuracy of the network has been improved. From the DP-Net, HP-Net to TP-Net, the OA of four datasets has improved by 0.09%–0.56%, and the number of network parameters has increased about 99 072 and 47 592. After adding the SPA-Att and SPE-Att, the classification accuracies of four datasets have improved more or less, and the network parameter number has increased about only 3 K. After adding both spectral and spatial attention blocks, the AAs of each dataset show notable improvement. After adding the Inter-Att, it is clear that all classification accuracies have improved, whereas the number of network parameters increased by only 384. From the accuracy improvement, it is clear that the effectiveness of each structure is verified.

The above-mentioned experiment verified the effectiveness of the triple-path feature extraction with interleave-attention block, and the numbers of block are also essential to gain the most optimal performance. Table X gives the classification results of the proposed network under different triple-path blocks.

As can be seen from Table X, when the number of the triple-path feature extraction block is two, the proposed network achieves its best performance. However, when the block number continues to increase, the performance drops on four datasets and the parameter numbers increase significantly.

2) Investigation on Network Parameters:

a) Spatial Size of Input: Fig. 9 provides the classification accuracies of OA in terms of all input sizes. To a certain extent, a larger input spatial size contains more spatial context information, which can lead to more informative features. However, if the input size is too large, it may contain too much information from other classes, which may lead to the involvement of unnecessary features. And in this experiment, except the input parameter of spatial size is different, ranging from (7×7) to (15×15), the other hyperparameters remain the same as presented in Section IV-B.

As shown in Fig. 9, when the input size is (11×11), the proposed network has achieved its best performance. And it is clear that our network can obtain the best classification results.

TABLE IX
ACCURACY ANALYSIS FOR USING DIFFERENT STRUCTURES

Results	Datasets	HP-Net	DP-Net	TP-Net	TP-Net + SPA-Att	TP-Net + SPE-Att	TP-Net + Att	TP-Net + Att + Inter-Att
OA (100%)		98.58	98.62	99.14	99.35	99.39	99.53	**99.64**
AA (100%)	**IP**	98.34	96.90	98.96	99.24	99.17	99.50	**99.64**
K × 100		98.53	98.42	99.04	99.26	99.30	99.47	**99.60**
OA (100%)		99.31	99.44	99.53	99.58	99.57	99.60	**99.82**
AA (100%)	**PU**	99.20	99.11	99.34	99.41	99.47	99.53	**99.63**
K × 100		99.27	99.26	99.44	99.45	99.51	99.51	**99.77**
OA (100%)		99.54	99.48	99.76	99.80	99.83	99.96	**100.00**
AA (100%)	**SV**	99.56	99.53	99.68	99.74	99.75	99.94	**100.00**
K × 100		99.52	99.42	99.69	99.73	99.79	99.94	**100.00**
OA (100%)		99.28	99.23	99.41	99.50	99.58	99.65	**99.72**
AA (100%)	**HU**	99.21	99.29	99.38	99.47	99.53	99.66	**99.76**
K × 100		99.15	99.17	99.40	99.46	99.55	99.63	**99.69**
Parameters		53 464	104 944	152 536	155 128	157 071	160,024	160 408

TABLE X
ACCURACY ANALYSIS FOR DIFFERENT NUMBERS OF TRIPLE-PATH FEATURE EXTRACTION BLOCK

Results	Datasets	Numbers of triple-path feature extraction block			
		One	Two	Three	Four
OA (100%)		98.52	**99.64**	99.33	99.16
AA (100%)	**IP**	97.17	**99.64**	99.31	99.08
K × 100		98.31	**99.60**	99.22	99.09
OA (100%)		99.52	**99.82**	99.78	99.72
AA (100%)	**PU**	99.27	**99.63**	99.61	99.56
K × 100		99.50	**99.77**	99.71	99.70
OA (100%)		99.82	**100.00**	99.98	99.97
AA (100%)	**SV**	99.84	**100.00**	99.98	99.96
K × 100		99.81	**100.00**	99.98	99.97
OA (100%)		99.40	**99.72**	99.63	99.65
AA (100%)	**HU**	99.47	**99.76**	99.68	99.67
K × 100		99.36	**99.69**	99.60	99.63
Parameters		60 424	160 408	293 488	426 184

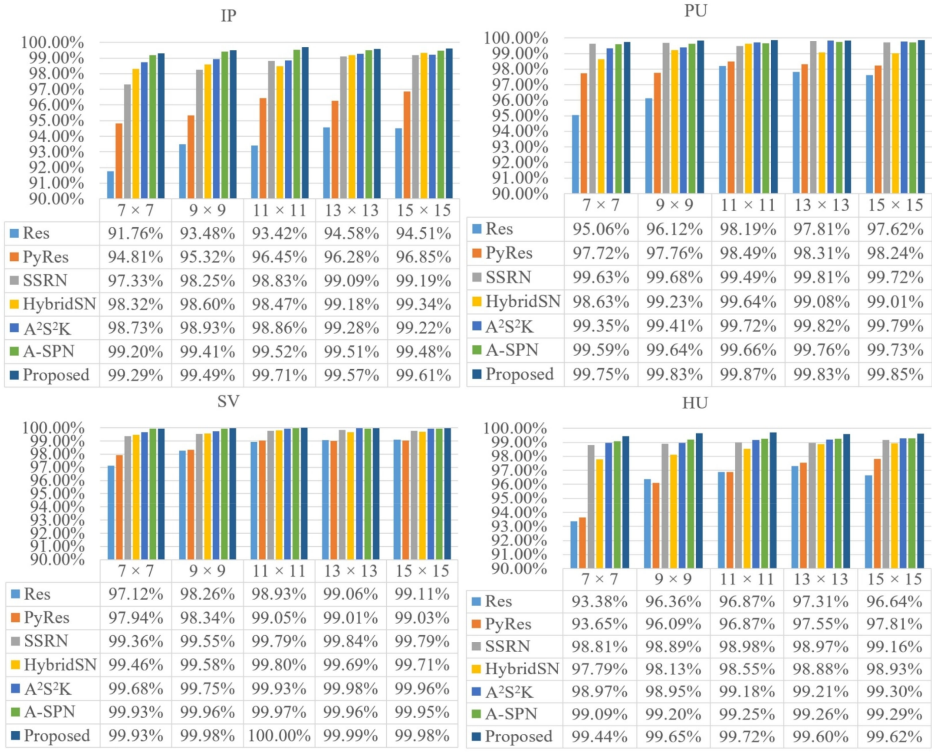
b) Training Proportions: Fig. 10 gives the detailed OA classification results of four datasets using different training proportions. Appropriate training proportion will be more beneficial to balancing training accuracy and training time. When the training proportion reaches a certain level, the performance improvement may be extremely limited. And in this experiment, except the parameter of training proportions is different, ranging from 1% to 25%, the other hyperparameters remain the same as presented in Section IV-B.

As we can see from Fig. 10, when the training proportions of IP, PU, SV, and HU are set to 20%, 10%, 10%, and 10%, respectively, the proposed network can obtain satisfactory classification results. In terms of all training proportions, our network can achieve better performance.

c) Principal Component Numbers: Table XI shows the classification accuracies of networks with different numbers of principal component. An appropriate principal component number will benefit the network training because the PCA operation can remove redundant spectral bands. However, if the number is set too low, some informative bands may be lost, which will lead to performance degradation.

As shown in Table XI, when the principal component number is 15, the network achieves its best performance. However, when the number exceeds 15, the performance drops to varying degrees.

3) Investigation on Network Parameter Numbers, FLOPs, and Running Time: All methods are executed on a Linux operating system with a Tesla K40c 12GB GPU and an Inter(R) Core i5-3470 CPU. The ResNet, PyResNet, and A^2S^2K are based on the PyTorch deep learning framework, and others are based on TensorFlow.



IP

	7 × 7	9 × 9	11 × 11	13 × 13	15 × 15
Res	91.76%	93.48%	93.42%	94.58%	94.51%
PyRes	94.81%	95.32%	96.45%	96.28%	96.85%
SSRN	97.33%	98.25%	98.83%	99.09%	99.19%
HybridSN	98.32%	98.60%	98.47%	99.18%	99.34%
A²S²K	98.73%	98.93%	98.86%	99.28%	99.22%
A-SPN	99.20%	99.41%	99.52%	99.51%	99.48%
Proposed	99.29%	99.49%	99.71%	99.57%	99.61%

PU

	7 × 7	9 × 9	11 × 11	13 × 13	15 × 15
Res	95.06%	96.12%	98.19%	97.81%	97.62%
PyRes	97.72%	97.76%	98.49%	98.31%	98.24%
SSRN	99.63%	99.68%	99.49%	99.81%	99.72%
HybridSN	98.63%	99.23%	99.64%	99.08%	99.01%
A²S²K	99.35%	99.41%	99.72%	99.82%	99.79%
A-SPN	99.59%	99.64%	99.66%	99.76%	99.73%
Proposed	99.75%	99.83%	99.87%	99.83%	99.85%

SV

	7 × 7	9 × 9	11 × 11	13 × 13	15 × 15
Res	97.12%	98.26%	98.93%	99.06%	99.11%
PyRes	97.94%	98.34%	99.05%	99.01%	99.03%
SSRN	99.36%	99.55%	99.79%	99.84%	99.79%
HybridSN	99.46%	99.58%	99.80%	99.69%	99.71%
A²S²K	99.68%	99.75%	99.93%	99.98%	99.96%
A-SPN	99.93%	99.96%	99.97%	99.96%	99.95%
Proposed	99.93%	99.98%	100.00%	99.99%	99.98%

HU

	7 × 7	9 × 9	11 × 11	13 × 13	15 × 15
Res	93.38%	96.36%	96.87%	97.31%	96.64%
PyRes	93.65%	96.09%	96.87%	97.55%	97.81%
SSRN	98.81%	98.89%	98.98%	98.97%	99.16%
HybridSN	97.79%	98.13%	98.55%	98.88%	98.93%
A²S²K	98.97%	98.95%	99.18%	99.21%	99.30%
A-SPN	99.09%	99.20%	99.25%	99.26%	99.29%
Proposed	99.44%	99.65%	99.72%	99.60%	99.62%

Fig. 9. Network accuracy under different input spatial sizes.

TABLE XI
ACCURACY ANALYSIS FOR DIFFERENT NUMBERS OF PRINCIPAL COMPONENT

Results	Datasets	Principal component numbers				
		5	10	15	20	25
OA (100%)	**IP**	99.02	99.42	**99.64**	99.56	99.44
AA (100%)		98.89	99.35	**99.64**	99.51	99.39
K × 100		99.01	99.38	**99.60**	99.57	99.42
OA (100%)	**PU**	99.41	99.79	**99.82**	99.78	99.67
AA (100%)		99.38	99.56	**99.63**	99.60	99.48
K × 100		99.40	99.72	**99.77**	99.71	99.56
OA (100%)	**SV**	99.78	99.93	**100.00**	99.98	99.97
AA (100%)		99.75	99.92	**100.00**	99.96	99.97
K × 100		99.78	99.92	**100.00**	99.98	99.97
OA (100%)	**HU**	99.21	99.53	**99.72**	99.65	99.68
AA (100%)		99.32	99.59	**99.76**	99.72	99.74
K × 100		99.19	99.49	**99.69**	99.63	99.66

Tables XII –XV give the network parameter number, computational complexity, and the running time, and it is clear that the proposed TP-Net owns the least parameter numbers in most cases, which are nearly half of compared networks. Many existing classification methods usually improve their network performance by deepening the network, which brings massive network parameters. Instead of deepening the network, we further widen it with three shallow branches to improve performance. And by utilizing size-optimized convolution layers in different branches, the network parameters are further limited. Meanwhile, the use of Inter-Att further improves the network performance but brings less than a 400-parameter increase. As for the network computational complexity, the proposed network is significantly reduced to only 0.3 million, which is the least among all networks. When comparing the training and testing time, the proposed TP-Net is faster than all compared networks except the HybridSN and A-SPN. It is worth noting that although their parameters are large, most of the parameters come from the full connection layer, which can be trained at a fast speed. Moreover, the proposed network involves massive convolution operations, which contain massive memory swaps, and this will significantly influence the training speed. From the above, it can

Fig. 10. Network accuracy under different training proportions.

TABLE XII
NETWORK PARAMETER NUMBERS, FLOPS, AND RUNNING TIME ON THE IP DATASET

FLOPs (M)	Dataset	Methods	Parameters (M)	Train/per epoch (s)	Test/(s)
84.1		ResNet	21. 91	65	165
85.1		PyResNet	22. 38	64	115
314.8	**IP**	SSRN	0.36	91	138
1.1		HybridSN	3. 93	1	2
340.2		A^2S^2K	0.37	104	47
2.5		A-SPN	0.64	4	2
0.33		Proposed	0.16	15	13

TABLE XIII
NETWORK PARAMETER NUMBERS, FLOPS, AND RUNNING TIME ON THE PU DATASET

FLOPs (M)	Dataset	Methods	Parameters (M)	Train/per epoch (s)	Test/(s)
43.3		ResNet	21. 60	135	913
43.8		PyResNet	22. 07	136	917
162.1	**PU**	SSRN	0.21	61	108
1.1		HybridSN	2. 14	1	4
175.2		A^2S^2K	0.22	78	147
0.38		A-SPN	0.10	2	4
0.32		Proposed	0.15	27	53

TABLE XIV
NETWORK PARAMETER NUMBERS, FLOPS, AND RUNNING TIME ON THE SV DATASET

FLOPs (M)	Dataset	Methods	Parameters (M)	Train/per epoch (s)	Test/(s)
85.7	SV	ResNet	21. 33	397	1150
86.8		PyResNet	21. 80	445	1159
321.1		SSRN	0.37	301	248
1.1		HybridSN	4. 00	3	5
347.1		A^2S^2K	0.37	359	269
2.6		A-SPN	0.66	3	6
0.33		Proposed	0.16	35	67

TABLE XV
NETWORK PARAMETER NUMBERS, FLOPS, AND RUNNING TIME ON THE HU DATASET

FLOPs (M)	Dataset	Methods	Parameters (M)	Train/per epoch (s)	Test/(s)
60.5	HU	ResNet	21.73	48	324
61.2		PyResNet	22.21	48	324
226.6		SSRN	0.27	27	20
1.1		HybridSN	2.88	2	4
244.9		A^2S^2K	0.28	33	52
1.2		A-SPN	0.31	1	2
0.32		Proposed	0.16	9	21

be concluded that the TP-Net can achieve the best classification results with the least parameters and computational complexity in a satisfactory running time.

V. CONCLUSION

In this article, a TP-Net with Inter-Att is proposed to extract representative spectral–spatial joint features for HSI classification. A hybrid branch is constructed to explore the correlative information between spectral and spatial domains, which is further integrated with purely spectral and spatial features for representative joint features. To increase the interaction of three branches, Inter-Att is elaborately designed and adopted in different stages of the feature extraction process. Through this mechanism, the features of each branch can be constrained by each other and make the final fused spectral–spatial features more discriminative. To further enhance the performance, spectral and spatial modules are introduced and adopted in corresponding branches to make the network more focused on informative regions and bands. Experimental results on four datasets demonstrate that compared to other networks, our TP-Net can achieve better performance with fewer parameters and lower computational complexity. In further work, we will try to introduce the Inter-Att to other HSI processing fields, such as superresolution, band selection, and so on, for full use of the attention mechanism. And we also encourage other researchers to verify the effectiveness of the proposed network with different indicators.

REFERENCES

[1] M. Masoud and S. Baharm, "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery," *Remote Sens.*, vol. 10, 2018, Art. no. 1119.

[2] W. Ma, Y. Xiong, and Y. Wu, "Change detection in remote sensing images based on image mapping and a deep capsule network," *Remote Sens.*, vol. 11, 2019, Art. no. 626.

[3] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial–spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[4] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[5] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[6] D. Wang, B. Du, L. Zhang, and Y. Xu, "Adaptive spectral–spatial multiscale contextual feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2461–2477, Mar. 2021.

[7] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[8] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and Wang L, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.

[9] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial–spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3534–3546, Jun. 2018.

[10] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[11] 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.

[12] E. Pan, "Spectral–spatial classification of hyperspectral image based on a joint attention network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 413–416.

[13] X. Kang, B. Zhuo, and P. Duan, "Dual-path network-based hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 447–451, Mar. 2019.

[14] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[15] 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.

[16] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[17] T. Lu, S. Li, L. Fang, X. Jia, and J. A. Benediktsson, "From subpixel to superpixel: A novel fusion framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4398–4411, Aug. 2017.

[18] X. Cao, Z. Xu, and D. Meng, "Spectral–spatial hyperspectral image classification via robust low-rank feature extraction and Markov random field," *Remote Sens.*, vol. 11, 2019, Art. no. 1565.

[19] C. Dong, M. Naghedolfeizi, D. Aberra, and X. Zeng, "Spectral–spatial discriminant feature learning for hyperspectral image classification," *Remote Sens.*, vol. 11, 2019, Art. no. 1552.

[20] K. Karthik, S. S. Kamath, and S. U. Kamath, "Automatic quality enhancement of medical diagnostic scans with deep neural image super-resolution models," in *Proc. IEEE 15th Int. Conf. Ind. Inf. Syst.*, 2020, pp. 162–167.

[21] H. Yanagisawa, T. Yamashita, and H. Watanabe, "A study on object detection method from manga images using CNN," in *Proc. Int. Workshop Adv. Image Technol.*, 2018, pp. 1–4.

[22] Y. Lu, Y. Shi, G. Jia, and J. Yang, "A new method for semantic consistency verification of aviation radiotelephony communication based on LSTM-RNN," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2016, pp. 422–426.

[23] A. H. Siregar and D. Chahyati, "Visual question answering for Monas tourism object using deep learning," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst.*, 2020, pp. 381–386.

[24] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[25] Y. Chen, Z. Lin, X. Zhao, and G. Wang, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[26] W. Hu, Y. Hhuang, and W. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, 2005.

[27] Y. Li, H. Zhang, and Shen Q, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, 2017, Art. no. 67.

[28] Z. Zhong, J. Li, Z. Luo, and Chapman M, "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.

[29] X. Kang, B. Zhuo, and P. Duan, "Dual-path network-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 16, no. 3, pp. 447–451, Mar. 2019.

[30] L. Chen, Z. Wei, and Y. Xu, "A lightweight spectral–spatial feature extraction and fusion network for hyperspectral image classification," *Remote Sens.*, vol. 12, 2020, Art. no. 1395.

[31] Z. Meng, L. Li, X. Tang, and Z. Feng, "Multipath residual network for spectral–spatial hyperspectral image classification," *Remote Sens.*, vol. 11, 2019, Art. no. 1896.

[32] B. Xi *et al.*, "Multi-direction networks with attentional spectral prior for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5500915, doi: 10.1109/TGRS.2020.3047682.

[33] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[34] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 19, 2022, Art. no. 8002005, doi: 10.1109/LGRS.2020.3026587.

[35] X. He, Y. Chen, and P. Ghamisi, "Dual graph convolutional network for hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502418, doi: 10.1109/TGRS.2021.3061088.

[36] B. Xi *et al.*, "Semisupervised cross-scale graph prototypical network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2022.3158280.

[37] D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[38] B. Liu *et al.*, "DSS-TRM: Deep spatial–spectral transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 103–114, 2022.

[39] X. Hu *et al.*, "Contrastive learning based on transformer for hyperspectral image classification," *Appl. Sci.*, vol. 11, no. 18, 2021, Art. no. 8670.

[40] S. Woo, J. Park, and J. Lee, "CBMA: Convolution block attention," 2018, *arXiv:1807.06521 [cs.CV].*

[41] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, 2019, Art. no. 1307.

[42] Q. Wang, B. Wu, and P. Zhu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Visi. Pattern Recognit.*, 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.

[43] 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.

[44] Z. Xue, M. Zhang, Y. Liu, and P. Du, "Attention-based second-order pooling network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9600–9615, Nov. 2021.

[45] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the sampling strategy for evaluation of spectral–spatial methods in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, Feb. 2017.

[46] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021, doi: 10.1109/TGRS.2020.3007921.

[47] M. Ahmad *et al.*, "Hyperspectral image classification-traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–998, 2022, doi: 10.1109/JSTARS.2021.3133021.

[48] L. Mou, P. Ghamisi, and X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[49] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[50] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[51] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[52] W. Wei, J. Zhang, and L. Zhang, "Deep cube-pair network for hyperspectral imagery classification," *Remote Sens.*, vol. 10, 2018, Art. no. 783.

[53] J. Yang, Y. Zhao, J. Chan, and C. Yi, "Hyperspectral image classification using two-channel deep convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Beijing, China, 2016, pp. 5079–5082.

[54] Z. Lin and M. Feng, "A structured self-attentive sentence embedding," in *Proc. Int. Conf. Learn. Representation*, 2017, pp. 1–15.

[55] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7281–7290.

[56] L. Peng, Y. Yang, Z. Wang, and Z. Huang, "MRA-Net: Improving VQA via multi-modal relation attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 318–329, Jan. 2022, doi: 10.1109/TPAMI.2020.3004830.

[57] H. Dong, L. Zhang, and Zhou B, "Band attention convolutional networks for hyperspectral image classification," 2019, *arXiv:1906.04379 [cs.CV].*

[58] S. Pande and B. Banerjee, "Adaptive hybrid attention network for hyperspectral image classification," *Pattern Recognit. Lett.*, vol. 144, pp. 6–12, 2021.

[59] X. Mei, E. Pan, Y. Ma, and X. Dai, "Spectral–spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, 2019, Art. no. 963.

[60] A. Paul, S. Bhoumik, and N. Chaki, "SSNET: An improved deep hybrid network for hyperspectral image classification," *Neural Comput. Appl.*, vol. 33, pp. 1575–1585, 2021.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[62] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, 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.

[63] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

Ziqing Deng received the B.S. degree in communication engineering from the Shenyang University of Technology, Shenyang, China, in 2020. He is currently working toward the M.S. degree in electronic and communication engineering at Guizhou University, Guiyang, China.

His research interests include computer vision, deep learning, and remote sensing image processing.

Yang Wang received the B.S. and M.S. degrees in electronic science and technology in 2018 and 2021, respectively, from Guizhou University, Guiyang, China, where he is currently working toward the Ph.D. degree in electronic science and technology.

His research interests include computer vision, machine learning, and image processing.

Bing Zhang received the B.S. degree in communication engineering in 2020 from Guizhou University, Guiyang, China, where he is currently working toward the M.S. degree in electronic and communication engineering.

His research interests include computer vision, deep learning, and medical image processing.

Linwei Li received the B.S. degree in electrical engineering and automation from Southwest Jiaotong University, Chengdu, China, in 2017. He is currently working toward the M.S. degree in electronic and communication engineering at Guizhou University, Guiyang, China.

His research interests include computer vision, deep learning, and remote sensing image processing.

Jihong Wang received the master's degree in microelectronics and solid electronics from Guizhou University, Guiyang, China, in 2005.

She is currently an Associate Professor with the College of Big Data and Information Engineering, Guizhou University. Her research interests include hyperspectral imaging systems, image processing, and embedded and optoelectronic systems.

Lifeng Bian received the B.S. degree in applied physics and microelectronics technology from the Hefei University of Technology, Hefei, China, in 1998, the M.S. degree in pattern recognition and intelligence system from the Institute of Intelligent Machines, Chinese Academy of Sciences (CAS), Beijing, China, in 2001, and the Ph.D. degree in condensed matter physics from the State Key Laboratory for Superlattices and Microstructures, Institute of Semiconductors, CAS, in 2004.

Until 2006, she was a Postdoctoral Researcher with the Paul-Drude-Institute for Solid-State Electronics, Berlin, Germany. Until 2021, she was a Professor with the Suzhou Institute of Nano-Tech and Nano-Bionics, CAS, Suzhou, China. She is currently a Researcher with the Institute of Chip and System Frontier Technology, Fudan University, Shanghai, China. Her research interests include hyperspectral imaging systems and semiconductor optical-electronic devices.

Chen Yang received the Ph.D. degree in microelectronics and solid electronics from the Institute of Semiconductor, Chinese Academy of Sciences, Beijing, China, in 2010.

From 2010 to 2012, he was a Research Engineer with Synopsys Co. Ltd. He was a Postdoctoral Researcher with the KTH Royal Institute of Technology, Stockholm, Sweden, from 2014 to 2015. He is currently a Professor with the College of Big Data and Information Engineering, Guizhou University, Guiyang, China. His research interests include remote sensing image processing, hyperspectral imaging systems, and machine learning.