Concern With Center-Pixel Labeling: Center-Specific Perception Transformer Network for Hyperspectral Image Classification

Chunyan Yu, *Senior Member, IEEE*, Yuanchen Zhu, Yulei Wang, *Member, IEEE*, Enyu Zhao, Qiang Zhang, *Member, IEEE*, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Self-attention-based approaches that leverage global context information for hyperspectral image (HSI) classification have gained increasing prominence. Nevertheless, due to the assignment of equivalent attention weight to all the tokens (pixels or patches), the existing self-attention mechanism inadvertently prioritizes the nonlabel-specified information over the instinct label-specified information, which generates attention shifts and redundancy in HSI classification. To alleviate the mentioned barrier, we propose the center-specific perception transformer (CP-Transformer) network, which is the first attempt to perform class-guided attention and filter interference factors for HSI classification feature representation. Specifically, the central-pixel focus attention (CFA) module is presented to compute the label-related attention between the center and other pixels. In this manner, CFA reduces computational complexity and closely aligns with the center-pixel labeling strategy. Besides, the spectral saliency focus attention (SSFA) module is developed to capture the spectral correlation by focusing salient bands to provide a beneficial supplement for spatial features. Moreover, the hierarchical integration network (HIN) constructs the inference network to integrate and rectify spatial-spectral features for HSI classification. The experiment results on four popular HSI datasets demonstrate that the proposed method achieves robust performance compared to other state-of-the-art methods. Our code will be released at https://github.com/Chirsycy/CP-Transformer**

***Index Terms*— Central focus attention, hierarchical integration network (HIN), hyperspectral image (HSI) classification, spectral-saliency focus attention.**

I. Introduction

HYPERSPECTRAL images (HSIs) [1], [2], [3], [4] capture a wide range of spectral bands and provide detailed information about the electromagnetic spectrum for each pixel. HSI classification is an essential task that aims to assign a unique semantic label to each pixel of the HSI by identifying the object characteristics. Nowadays, HSI classification has been extensively applied in various fields including land cover classification [5] and urban investigation [6].

Deep learning (DL) methods that automatically extract spectral-spatial features have attracted significant attention in HSI interpretation [7], [8], [9]. Early works such as auto encoders [10], [11], deep belief networks [12], [13] are applied to build categorical signatures, while have a limitation on learning practical features and achieve unsatisfactory performance. With the technological advancements in DL, the constitutional neural networks (CNNs) are empowered to perceive local spatial-spectral features and have been extensively employed in handling HSI classification tasks. Roy et al. [14] employed 3-D-2-D CNN networks by different morphology convolution kernels to extract semantic features for HSI classification. The cross-mixing residual network [15] and deep pyramidal residual network [16] adopted residual connectivity to expand the receptive field of features to fuse multilayer features. The adaptive spectral-spatial multiscale model [17] was composed of a spectral sub-network and a spatial sub-network, which extracts multiscale contextual information to promote feature representation. The online spectral information compensation network [18] transferred spectral information to a spatial feature extraction network during the multifeature fusion. Despite the successful works, CNN-based HSI classification models encounter challenges in capturing global pixel correlation and effectively extracting context features for classification.

To improve the feature representation and generalization of the irregular HSI data, graph neural networks (GNNs) have engaged in spatial-spectral feature mining for the recognition of HSI. GNNs excel at flexibly capturing relationships and connectivity by aggregating features from neighboring nodes. With the mentioned advantage, the GNN-based approaches effectively capture complex relationships between nodes, which incorporate contextual information to enhance the distinction of the extracted features for HSI classification. Dong et al. [19] built a weighted feature fusion of CNN and graph attention network to explore the spatial-spectral information of the hyperspectral data. Yu et al. [20] proposed an edge-inferring GNN with a task-guided self-diagnosis mechanism for few-shot HSI classification. Yu et al. [21] established a graph-polarized fusion network to integrate the super pixel-level and pixel-level spectral-spatial features from

Received 11 February 2025; revised 7 May 2025; accepted 19 May 2025. Date of publication 27 May 2025; date of current version 29 May 2025. This work was supported by the National Natural Science Foundation of China under Grant 62471079. *(Corresponding author: Yulei Wang.)*

Chunyan Yu, Yuanchen Zhu, Yulei Wang, Enyu Zhao, and Qiang Zhang are with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: yucy@dlmu.edu.cn; zhuyuanchen@dlmu.edu.cn; wangyulei@dlmu.edu.cn; zhaoenyu@dlmu.edu.cn; qzhang95@dlmu.edu.cn).

Xiaoqiang Lu is with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China.

Digital Object Identifier 10.1109/TGRS.2025.3573233

the two branches. While the mentioned solutions demonstrate satisfactory performance in HSI classification, the selection of appropriate nodes and edges in the graph-based model is critical, and poor choices negatively impact performance on HSI classification.

Recently, the attention mechanism that is inspired by the mechanism of human cognition has been extensively exploited to focus on the most relevant features or regions for classification. A few works [22], [23], [24], [25] have explored attention-based frameworks to improve the global discriminant modeling capability for HSI classification. By leveraging the self-attention mechanism, transformer networks [26], [27], [28] have become typically standard attention frameworks for HSI classification tasks, which capture the global correlation to extract more separable features, and address the limitation in CNNs to a certain extent. Typically, a new spectral spatial transformer (SST) [29] was proposed to establish relationships between the adjacent spectrum with a dense-layer transformer network. Spatial attention transformer (SAT) net [30] and spectral-spatial feature tokenization transformer (SSFTT) model [31] employed spectral and spatial attention to capture essential spatial-spectral information. Besides, some researchers presented diverse techniques [32], [33], [34], including establishing spectral-spatial fusion networks [36], [37], [38], [39], and adopting specialized sampling strategy steps [40], [41] to enhance the effectiveness of the transformer-based models. Moreover, researchers further integrate transformer architecture and CNN to generate global–local joint features for optimized feature assignment, and the typical work included convolution transformer mixer (CTMixer) [42], group-aware hierarchical transformer (GAHT) [43], and the model toward multilevel features and decision boundaries (ToMF-B) [44]. Although the transformer-based models perform outstandingly in capturing global information for feature extraction, they inevitably require substantial computational power and memory for HSI classification.

Motivation: As well known, the existing computation pattern of the self-attention mechanism is more suitable for natural image interpretation than HSI image recognition. For the natural image classification with the classic Vision Transformer Encoder (ViT) [45], token sequences are generated from one single image that is annotated with a unique class label. In this way, self-attention with equivalent attention weight measures the context information related to one specific category. Unlike the natural image classification, the samples of HSI classification are acquired by extracting sub-cubes from the HSI with a window size of 15 × 15 as illustrated in Fig. 1(a), and the label is aligned to the center pixel, which we refer to as center-pixel labeling in this article. Notably, the pixels in the sample patch easily encompass multiple categories with the labeling style. For instance, in the case of the Indiana Pines dataset, the category distribution of class-related pixels is depicted in Fig. 1(b). As reported, the majority of samples only contain fewer than 50% class-label pixels, especially the samples of class 2 merely contain fewer than 20% class-label pixels. Stated differently, most pixels in one HSI sample are non-label-related. Given this observation, unlike the natural image classification, the traditional

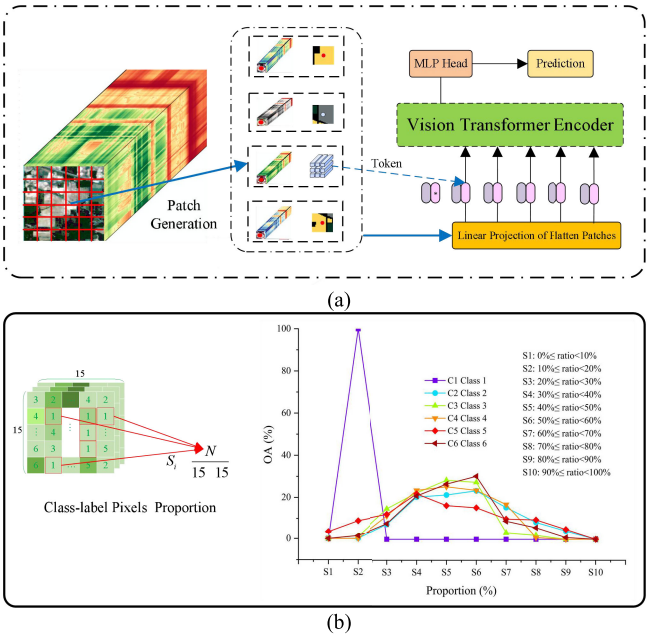

Fig. 1. Class-label-related classification issues on the Indian Pines dataset. (a) Classification procedure based on VIT network. (b) Class-label pixels proportion in samples with center-pixel labeling strategy.

self-attention mechanism in HSI classification as shown in Fig. 1(a) that allocates equivalent computation weight for all tokens results in attention shifts and redundancy, which tends to prioritize the background and pixels from other categories rather than the pixels related to the labeled class. Therefore, it is crucial to develop a self-attention mechanism that is specifically designed for the center-pixel labeling strategy for the HSI classification task.

In this article, to tackle the mentioned issues, we propose the center-specific perception transformer (CP-Transformer) network, which strengthens the class-related features and reduces the impact of nonlabel-related pixels. In the core component of the backbone, the central-pixel focus attention (CFA) mechanism is built to assess the category correlation between the center pixel and neighbor pixels, which are aligned with the center-labeling strategy of HSI. The spectral saliency focus attention (SSFA) module performs attention calculations to guide the model emphasizing the salient spectral. Besides, the hierarchical integration network (HIN) is built to preserve the category-oriented feature and filter confused semantic features.

The main contributions are summarized as follows.

1) To handle the attention shifts and redundancy caused by traditional self-attention computation mechanisms, we build a CP-Transformer network to explore attention information for HSI classification. Innovatively, the proposed model is specially designed to match the popular center-label strategy. Besides, the proposed model seamlessly integrates spatial attention, spectral attention, and feature fusion modules for effectively refining the spatial-spectral features.
2) To dynamically emphasize the spatial weight of class-related attention, the CFA module is presented as a center-specified computation pattern, which guides

the spatial attention toward class-oriented features and suppresses the expression of interference factors. Meanwhile, the CFA module merely measures the focus attention related to the labeling pixel, which decreases the computational complexity of attention from $O(n^2)$ to $O(n)$. In this way, CFA overcomes the low efficiency of self-attention computation in encoding complex contextual information of HSI classification.

3) To refine category features from the spectrum domain, we propose the SSFA module to emphasize the significant spectral information. Concretely, we compute spectral attention to concentrate on salient bands by regulating the prioritization in the spectral domain. Besides, cascading the SSFA module with the CFA module implements spatial-spectral aggregation and highlights class-related features for HSI classification further.
4) To reduce the computation consumption of the general decoder in transformer architecture, we develop the HIN as the substitute for the traditional feed forward network. HIN constructs a simple yet effective hierarchical convolution module to integrate spatial-spectral features, which eliminates the positional encoding and preserves class-related features.

II. Proposed Method

Question Definition: With the popular center-pixel strategy, we build the patch set denoted as $S = \{s_1, s_2, \ldots, s_n\} \in \mathbb{R}^{H\times W\times B}$, and the label set is denoted as $Y = \{y_1, y_2, \ldots, y_K\}$ where $H \times W$ and B represents the spatial size and band number, and K is the number of the classes, respectively. The objective of the HSI classification task is to assign each label from Y to the corresponding HSI patch within S.

Brief Overview: The flowchart of the proposed CP-Transformer is depicted in Fig. 2. As described, the model is fed with the input of samples from the patch set. In the backbone, we establish a four-stage architecture for discriminative feature representation. Each stage is composed of CFA, SSFA, and HIN modules, which comprehensively mine the spectral-spatial fusion information, maintain the spatial size of the feature map, and gradually decrease the channel number. Lastly, a global average pooling (GAP) layer is treated as a classifier to generate the prediction for HSI classification.

A. Central-Pixel Focus Attention Module

In our model, the initial pixel embedding (IPE) module consists of one transposed convolution layer (ConvTrans2d) with a kernel size of 1×1, a BatchNorm layer (BN), a Relu layer (Relu), and a maximum pooling layer (Maxpool). Fed with the input $s \in S$, the output map s^* is yielded to feed into the CFA block.

The CFA module is the core of each stage and aims to improve the spatial attention representation. With the CFA, the CP-Transformer enhances relevant long-range category-aware features and suppresses the irrelevant category information for HSI classification. Fig. 3 shows the structure of the CFA module. Initially, a 2-D LayerNorm operation is applied to normalize the input. Moreover, a transposed convolution operation with a 1×1 kernel is adopted to generate the pixel-oriented tensors k^{cfa} and v^{cfa}.

Next, to implement the center-specific perception attention, we design the computation pattern in terms of the sample label. In the central-labeling sample strategy, the central pixel certainly holds a dominant position in the HSI patch. Similarly, the center vector of feature maps in our network is also regarded as a computing core. As shown in (1), we regard the center vector of k^{cfa} as the query vector q^{cfa} and the neighbor vector $k^{\text{cfa}}_{(i,j)}$ as the context vector, where (i, j) denotes the ith and jth indices of the height and width, respectively, $h \times w$ is the spatial size of the input map

$$k^{\text{cfa}} := \begin{pmatrix} k^{\text{cfa}}_{11} & \cdots & k^{\text{cfa}}_{1w} \\ \vdots & q^{\text{cfa}} & \vdots \\ k^{\text{cfa}}_{h1} & \cdots & k^{\text{cfa}}_{hw} \end{pmatrix}. \tag{1}$$

To prevent numerical overflow issues, the L_2 norm is utilized to normalize each vector $k^{\text{cfa}}_{(i,j)}$, and the normalization process is denoted as follows:

$$k^{\text{cfa}}_{(i,j)} = \frac{k^{\text{cfa}}_{(i,j)}}{\sqrt{k^{\text{cfa}}_{(i,j)} \times (k^{\text{cfa}}_{(i,j)})^{\text{T}}}}. \tag{2}$$

Accordingly, the similarity map d^{cfa} between q^{cfa} and k^{cfa} is obtained by vector multiplication

$$\begin{aligned} d^{\text{cfa}} &= k^{\text{cfa}}_{ij} \times q^{\text{cfa}} \\ &:= \begin{pmatrix} k^{\text{cfa}}_{11} \times q^{\text{cfa}} & \cdots & k^{\text{cfa}}_{1w} \times q^{\text{cfa}} \\ \vdots & q^{\text{cfa}} \times q^{\text{cfa}} & \vdots \\ k^{\text{cfa}}_{h1} \times q^{\text{cfa}} & \cdots & k^{\text{cfa}}_{hw} \times q^{\text{cfa}} \end{pmatrix}. \end{aligned} \tag{3}$$

Unlike the Softmax function, the Softsign function eliminates exponential calculations and normalization. Since pixel weights are directly computed by the network, we employ the simple yet effective Softsign function as an activation function to smooth and normalize the similarity matrix d^{cfa}. The Softsign function with input λ is defined as follows:

$$\text{Softsign}(\lambda) = \frac{\lambda}{1 + |\lambda|}. \tag{4}$$

Afterward, the elementwise product operation is conducted between d^{cfa} and v^{cfa} to obtain the redirected attention x^{f_1} with the following equation:

$$x^{f_1} = \text{Softsign}(d^{\text{cfa}}) \odot v^{\text{cfa}}. \tag{5}$$

Further, we also employ the local convolution network as the local projection to enhance the diversity of spatial features, which is implemented by a transposed convolution operation with a 3×3 kernel and a subsequent BatchNorm layer. The specific equation is defined as follows:

$$x^{f_2} = \text{BN}\big(\text{ConvTrans2d}\big(x^*\big)\big). \tag{6}$$

Finally, the local information is combined with the attention map to prevent vanishing gradient and network degradation while enhancing the feature representation. The output x^{cfa} is generated with the following formula:

$$x^{\text{cfa}} = \text{drop}\big(x^{f_1} + x^{f_2}\big) \tag{7}$$

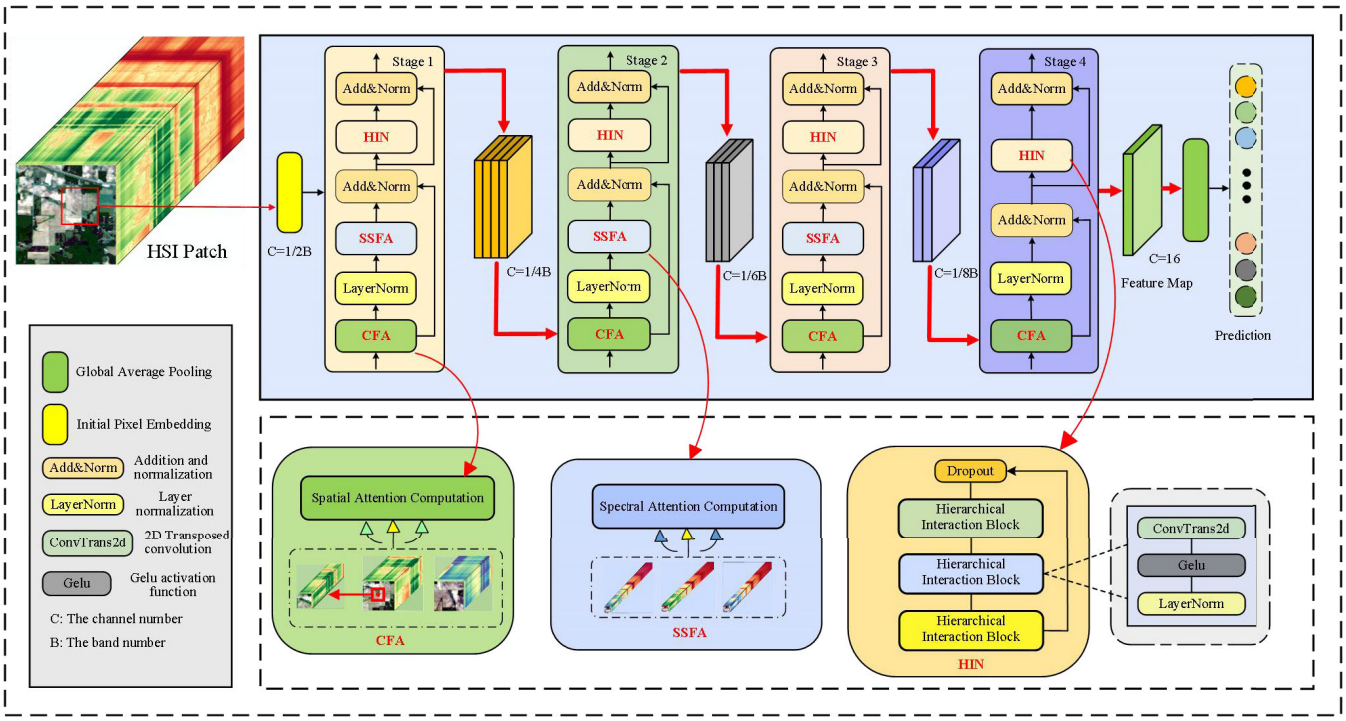

Fig. 2. Overall architecture of the CP-Transformer.

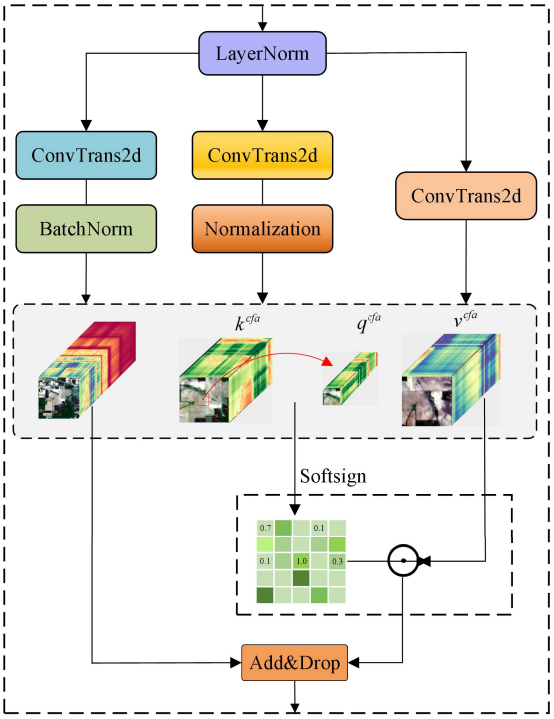

Fig. 3. Diagram of the overall structure of the CFA attention mechanism.

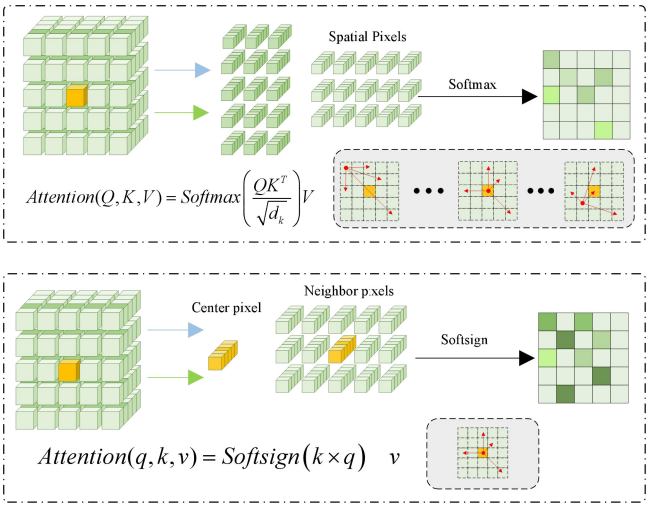

Fig. 4. Structural diagram of two types of attention computing. (a) Self-attention mechanism. (b) CFA mechanism. Notice: In this figure, q/Q, k/K, and v/V represents the query, key, and value of the attention calculation, respectively.

where the symbol drop means the dropout operation.

Comparison and Analysis: In this section, we compared the CFA with the traditional self-attention mechanism. As shown in Fig. 4(a), the calculation of traditional self-attention in one HSI sample is conducted between each pair of pixel tokens (or region-pixel tokens). Therefore, mixed-category tokens occupy a significant portion of attention, which leads the attention shifted to other classes. Meanwhile, the Softmax function requires scanning all pixels to select the most expressive pixels, which amplifies the interference region and suppresses the expression of the label-related regions with shifted attention. In contrast, the CFA module intends to mitigate category confusion and align with the center label style in the HSI classification. As illustrated in Fig. 4(b), CFA merely measures the category correlation between the central pixel and the neighbor vectors to explore the relatively spatial weight. Thus, the computational complexity is $O(n)$ in the CFA module rather than $O(n^2)$ in the traditional self-attention mechanism. Besides, unlike the Softmax function, the Softsign function

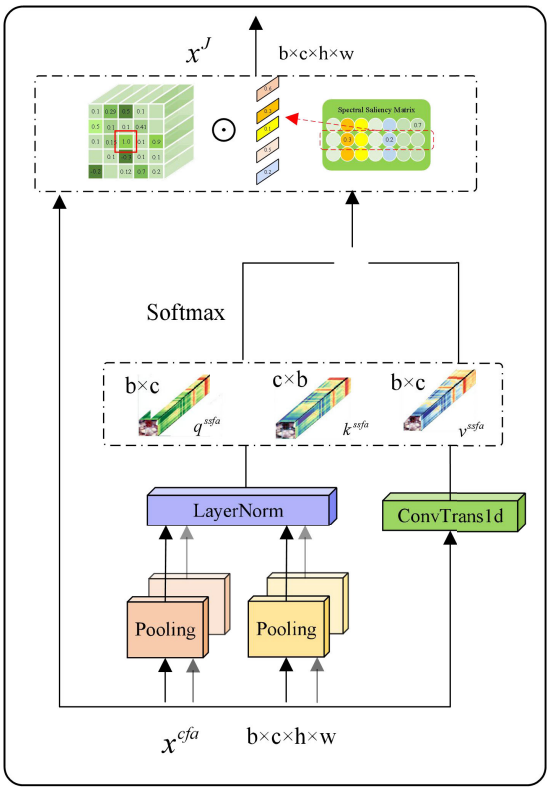

Fig. 5. Diagram of the overall structure of the SSFA attention mechanism.

increases the weight of all homogeneous class-label pixels and reduces the weight of the interference regions.

B. Spectral Saliency Focus Attention Module

As the beneficial supplementary for the CFA module, the SSFA is built to extract the salient spectral information for feature assembling. The structure of the SSFA is shown in Fig. 5. First, we employ the average pooling operation to obtain the average spectral vectors of the input tensor denoted as x'.

Next, we apply 1-D LayerNorm operation (LayerNorm*) to normalize the corresponding weight, obtain the projection matrix q^{ssfa} and k^{ssfa}. In addition, considering the interaction of different channels, we employ 1-D transpose convolution (ConvTrans1d) to project vectors x' to vectors v^{ssfa}.

To incorporate dependencies and enhance the spectral feature representation, the similarity matrix r^{s} is encoded as

$$r^{\text{s}} = \frac{q^{\text{ssfa}} \otimes (k^{\text{ssfa}})^{\text{T}}}{\text{Diag}(k^{\text{ssfa}} \otimes (k^{\text{ssfa}})^{\text{T}})} := \begin{pmatrix} \frac{r^{\text{s}}_{11}}{d^{\text{s}}_1} & \cdots & \frac{r^{\text{s}}_{1b}}{d^{\text{s}}_b} \\ \vdots & \ddots & \vdots \\ \frac{r^{\text{s}}_{b1}}{d^{\text{s}}_1} & \cdots & \frac{r^{\text{s}}_{bb}}{d^{\text{s}}_b} \end{pmatrix} \tag{8}$$

where $\otimes$ represents the matrix multiplication and Diag$(\cdot)$ means extraction of diagonal elements as the scaling factor vector $d^S_i, i \in \{1, \ldots, b\}$, and b denotes the batch number.

Accordingly, the recalibrated spectral weight matrix w^{s} is obtained by the following formula:

$$w^{\text{s}} = \text{Softmax}(r^{\text{s}})v^{\text{ssfa}}. \tag{9}$$

Lastly, joint feature maps x^{J} are obtained to amplify the spectral information with the following equation:

$$x^{\text{J}} = x^{\text{cfa}} + x^{\text{cfa}} \odot w^{\text{s}}. \tag{10}$$

C. Hierarchical Integration Network

In each stage of the CP-Transformer, we construct the HIN network to refine the spectral-spatial joint features for the inference part. The structure of HIN is illustrated in Fig. 2 and consists of three Hierarchical Integration Blocks (HIBs), which include the transposed convolution, LayerNorm, and Gelu activation function. First, we apply 2-D LayerNorm to obtain the initial feature map x^g_{in}, and then each output of the Pth layer x^g_P is achieved by (11). Notably, group-transposed convolution is employed in HIB to increase the diversity of learning features

$$x^g_p = \text{LayerNorm}(\text{Gelu}(\text{ConvTrans}(x^g_{P-1}))). \tag{11}$$

Afterward, the final output $\tilde{x}^g_P$ is obtained by feature connection and dropout operation, and the specific formula is defined as follows:

$$\tilde{x}^g_P = \text{drop}\left(x^g_3 + x^g_{\text{in}}\right). \tag{12}$$

D. Classification Procedure With the CP-Transformer

In the proposed model, the original training samples for HSI classification are fed into the IPE module first. Next, the obtained maps are further mined in the proposed four-stage model. In each stage, the spatial attention is calculated with the presented center-perception calculation mechanism in the CFA module. Subsequently, the SSFA module is responsible for the spectral attention refinement. The HIN module implements information aggregation further for the discriminative representation.

Comprehensively, in different stages, we try to maintain the spatial size of the feature map and gradually decrease the channel number for the compression mapping of spectral information. Lastly, a GAP layer is treated as a classifier to generate the prediction for HSI classification. In the training stage, the cross-entropy loss is employed as the final optimization function in the multiple-size training phase. The definition of the loss function is defined in the following formula:

$$\mathcal{J} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^k \log(\mathcal{H}(s_i)) \tag{13}$$

where N represents the number of specific samples, K denotes the total number of categories, x_i is the ith sample, y_i^k means the label of s_i, $\mathcal{H}(\cdot)$ denotes the prediction of the proposed model with the parameters of ψ.

The algorithm of the training of the CP-Transformer is outlined as in Algorithm 1.

E. Benefit and Limitation

Benefit: The proposed CP-Transformer offers two key advantages. 1) Enhanced self-attention relevance and sparsity. By concentrating specifically on the center-pixel-related attention computation, which matches the pixel style of the

labeling sample and improves the relevance and sparsity of attention computation. In this way, the model prioritizes the most informative pixels for accurate classification. 2) Reduced complexity. CFA reduces computational complexity by excluding the computation involving interference pixels, which make the method more computationally efficient and feasible compared to the traditional self-attention approach.

Algorithm 1 CP-Transformer

Input: $S = \{s_1, s_2, \ldots, s_n\}$, initial params of ψ
Output: ψ
Initialize ψ with random Gaussian values
$\{Q_i\} \leftarrow$ Get training set S_i from $S = \{s_1, s_2, \ldots, s_n\}$
For 1 *to Epoch* **do**
  For 1 *to Batch* **do**
  Randomly generate a mini-batch sample
  Feed samples and get the prediction
  $\mathcal{J} \leftarrow$ Calculate the cross-entropy prediction loss
  $\psi = \leftarrow \nabla_\psi(\mathcal{J})$ Update ψ via Adam optimizer
  End
End

Limitation: Since the spectral variability of different scenes disrupts the effective capture of attention information related to the center-labeled class, the performance of our method may be hindered in HSI cross-classification scenarios.

III. Experiments

In this section, the details of four experimental datasets and experimental settings are first introduced. Subsequently, we compare the results achieved by our method and other state-of-the-art methods. Furthermore, we discuss and analyze the experimental results to gain a comprehensive performance evaluation of our method.

A. Datasets

Four benchmark HSI datasets are selected for experiments, which are collected by various sensors in different regions and have found extensive utilization in numerous studies.

1) *Indian Pines:* The Indian Pines dataset involves 200 spectral bands after removing water absorption bands and 145 × 145 pixels with a spatial resolution of 20 m and 16 distinct classes.
2) *Houston:* The Houston dataset covers the University of Houston campus with a resolution of 349 × 1905, 144 spectral bands, and 15 classes of ground objects.
3) *KSC:* The kennedy space center (KSC) dataset contains 176 spectral bands after removing water noise, with 13 categories of ground objects and a spatial resolution of 512 × 614.
4) *Botswana:* The Botswana dataset covers 145 spectral bands to distinguish 14 types of ground objects and contains 256 × 1476 pixels.

B. Experiment Setting

The experiments are conducted on the standard hardware platform, which consists of an RTX 8000 GPU and an AMD Ryzen Threadripper 3990X CPU. Notably, the detailed

TABLE I
Distribution of Samples and the Background Color of Land Cover Classes on Indian Pines

NO	color	class	Test
1		Alfalfa	46
2		Corn-notill	1428
3		Corn-mintill	830
4		Corn	237
5		Grass-pasture	483
6		Grass-tress	730
7		Grass-pasture-mowed	28
8		Hay-windrowed	478
9		Oats	20
10		Soybean-notill	972
11		Soybean-mintill	2455
12		Soybean-clean	593
13		Wheat	205
14		Woods	1265
15		Buildings	386
16		Stone	93
Total			10249

class information of each dataset is uniformly reported in Tables I and II. The size of the sample is set to 15 × 15. For all four datasets, we randomly selected 20 samples per class as the training set. Additionally, 5% of the samples per class are allocated as the validation set, while all the samples are employed for the test set. Adam is employed as the optimizer with an initial learning rate of $1e{-}4$ and a weight decay set of $1e{-}6$. The dropout rate is set to 0.3, and the batch size is fixed at 200. To assess the effectiveness of our method, nine models are exploited for comparison. In which, the convolution-related models include residual spectral-spatial attention network (RSSAN) [?]db@bib:52 and pResNet [16], and the transformer-related network includes hierarchical transformer [47], spectral–spatial masked transformer (SSMTR) [48], spectral–spatial transformer network (SSTN) [49], robust vision transformer (RVT) [51], SSFTT [31], GAHT [43], and CTMixer [42]. All the methods are implemented by their original configurations and the best model parameters specified in their respective papers.

In the following experiments, we adopt the overall accuracy (OA), average accuracy (AA), and Kappa as objective criteria to evaluate the performance of all the compared methods. Each execution of all the approaches has been repeated ten times, and the OA calculation is reported in the form of mean ± standard deviation. Besides, four criteria are employed including, model parameters (Param), floating-point operations (FLOPs), training time (Train), and test time (Test) to measure the efficiency, complexity, and computational cost of the compared models.

C. Comparison With State-of-the-Art Methods

1) Classification Performance: The quantitative results of compared methods are reported in Tables III–VI. As recorded, the proposed method demonstrates superior performance in terms of OA, AA, and the Kappa coefficient. RSSAN and pResnet only obtain less than 80% OA on the Indian Pines dataset, which indicates that convolution-based methods have dataset, which indicates that convolution-based

TABLE II

DISTRIBUTION OF SAMPLES AND THE BACKGROUND COLOR OF LAND COVER CLASSES ON HOUSTON, BOTSWANA, AND KSC

		Houston		Botswana		KSC	
No	Color	class	Numbers	class	Numbers	class	Numbers
1		Healthy grass	1251	Water	270	Scrub	728
2		Stressed grass	1254	Hippo grass	101	Willow swamp	220
3		Synthetic grass	697	Floodplain grasses 1	251	CP hammock	232
4		Trees	1244	Floodplain grass 2	215	Slash pine	228
5		Soil	1242	Reeds	269	Oak/Broadleaf	146
6		Water	325	Riparian	269	Hardwood	207
7		Residential	1268	Firescar	259	Swamp	96
8		Commercial	1244	Island interior	203	Graminoid marsh	393
9		Road	1252	Acacia woodlands	314	Spartina marsh	496
10		Highway	1227	Acacia shrublands	248	Cattail marsh	365
11		Railway	1235	Acacia grasslands	305	Salt marsh	378
12		Parking Lot 1	1233	Short mopane	181	Mud flats	454
13		Parking Lot 2	469	Mixed mopane	268	Water	836
14		Tennis Court	428	Exposed soils	95	/	/
15		Running Track	660	/	/	/	/
Total			15029		3248		5752

TABLE III

CLASSIFICATION PERFORMANCE OF THE APPROACHES ON INDIAN PINES

Class	RSSAN	pResNet	HiT	SSMTR	SSTN	RVT	GAHT	SSFTT	CTMixer	Ours
1	97.21±1.40	96.25±3.50	**100.0±0.0**	**100.0±0.0**	96.31±4.35	99.57±0.91	**100.0±0.00**	99.78±0.69	**100.0±0.0**	**100.0±0.0**
2	40.41±15.1	65.23±4.23	46.22±5.04	58.46±4.27	61.63±5.26	42.41±6.93	65.04±4.65	80.80±4.22	90.80±1.32	**98.64±1.85**
3	60.21±9.20	70.12±7.08	63.01±10.4	67.37±7.23	33.65±15.1	48.52±3.09	73.45±5.29	83.84±4.67	94.57±2.20	**99.11±0.73**
4	76.42±6.20	98.71±2.31	90.04±5.64	92.45±2.53	100.0±0.0	94.39±2.03	97.72±2.40	98.06±1.57	99.87±0.28	**100.0±0.0**
5	88.37±1.40	84.22±5.63	85.75±9.24	78.88±5.78	73.23±5.34	84.54±5.60	89.69±1.96	92.92±2.72	95.92±1.26	**98.51±1.4**
6	87.22±3.22	93.96±2.27	92.96±2.12	96.62±1.07	94.37±2.4	88.35±2.96	96.67±1.00	97.55±1.45	98.36±0.43	**100.0±0.00**
7	97.01±0.41	96.24±1.30	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**
8	98.33±0.40	97.43±0.61	99.75±0.46	97.7±2.43	99.98±0.07	95.34±1.93	99.62±0.52	99.77±0.36	99.98±0.07	**100.0±0.00**
9	98.37±1.07	98.81±0.53	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.0**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**
10	62.21±7.20	70.32±3.37	59.44±4.51	73.53±2.13	72.9±7.81	62.33±5.24	73.52±2.71	75.7±3.34	85.43±3.24	**95.99±1.72**
11	71.72±3.62	73.65±2.97	69.04±2.49	70.02±3.27	72.32±5.64	61.36±1.65	74.26±3.15	81.96±5.16	91.17±1.71	**99.28±0.48**
12	50.67±8.50	62.15±3.21	53.09±10.3	66.31±7.55	45.73±13.6	61.55±9.11	82.67±5.12	82.09±5.16	88.31±2.01	**98.82±0.63**
13	92.51±1.66	97.22±1.81	98.63±1.35	96.63±1.65	100.0±0.0	99.36±0.4	98.1±3.14	99.12±1.39	99.76±0.26	**100.0±0.0**
14	69.28±5.71	80.38±1.70	85.91±3.39	88.07±4.11	92.31±3.74	83.16±1.64	94.4±1.2	97.43±1.09	97.94±0.54	**100.0±0.0**
15	92.41±3.35	96.67±0.71	93.73±3.44	89.35±3.95	89.07±7.45	79.33±8.00	95.26±1.79	94.66±2.46	96.87±1.05	**99.84±0.33**
16	93.23±2.27	94.52±1.58	99.57±0.59	98.71±1.95	100.0±0.0	99.89±0.34	99.68±0.52	99.24±0.52	99.24±0.52	**100.0±0.0**
OA	65.07±7.27	78.16±3.27	72.11±1.36	76.46±1.87	73.9±1.89	67.98±1.17	81.59±0.88	87.25±0.84	93.41±0.37	**99.04±0.29**
AA	74.94±5.67	85.16±2.21	83.57±1.41	85.88±1.55	83.22±0.87	81.26±1.32	90.00±0.54	92.68±0.57	96.14±0.22	**99.39±0.13**
Kappa	61.32±8.02	77.32±3.92	68.58±1.51	73.46±2.08	70.65±2.05	63.87±1.39	79.17±0.97	85.52±0.93	92.52±0.42	**98.91±0.33**

TABLE IV

CLASSIFICATION PERFORMANCE OF THE APPROACHES ON KSC

Class	RSSAN	pResNet	HiT	SSMTR	SSTN	RvT	GAHT	SSFTT	CTMixer	Ours
1	97.89±1.03	96.50±1.01	98.94±0.47	91.55±4.41	98.34±2.56	98.54±0.71	99.63±0.33	97.64±4.33	99.90±0.04	**100.0±0.00**
2	72.62±3.97	91.96±2.57	75.06±2.61	77.74±5.43	40.95±35.3	96.63±2.21	99.63±0.49	52.88±11.1	**100.0±0.00**	99.55±1.07
3	65.26±10.1	74.34±6.21	95.35±2.80	86.76±4.05	73.48±27.1	91.10±2.90	98.05±2.39	33.56±35.9	**100.0±0.00**	**100.0±0.00**
4	59.59±4.24	82.67±6.47	50.79±13.1	75.60±6.90	6.91±12.23	79.88±10.33	95.79±3.20	34.21±5.46	82.02±9.02	**95.08±6.64**
5	88.26±3.45	93.76±2.61	84.97±5.01	91.74±2.07	68.76±31.2	95.28±5.12	99.81±0.42	67.14±13.6	94.91±4.13	**100.0±0.00**
6	74.23±6.21	87.29±6.42	95.37±3.05	86.16±2.96	65.33±33.1	95.28±2.67	96.99±2.74	23.76±15.8	96.90±1.17	**99.87±0.41**
7	97.65±0.06	92.10±0.21	99.62±0.92	91.90±3.40	97.72±3.71	99.91±0.31	**100.0±0.00**	32.38±37.2	99.62±0.66	**100.0±0.00**
8	83.21±2.55	91.26±2.01	92.48±1.83	81.83±3.04	37.03±32.1	98.14±1.24	96.82±2.2	70.21±26.0	96.54±0.57	**100.0±0.00**
9	87.43±2.24	95.51±0.13	99.12±1.02	91.00±1.95	82.83±13.8	99.98±0.06	98.65±1.57	86.33±13.2	99.37±2.01	**100.0±0.00**
10	46.64±7.53	92.46±3.96	84.16±4.29	80.25±5.13	89.90±9.19	99.43±0.58	99.55±0.72	87.28±7.61	99.78±0.22	**100.0±0.00**
11	98.12±2.55	98.21±0.15	99.07±1.13	95.37±2.94	95.87±4.25	99.64±0.44	99.45±1.16	99.21±0.95	**100.0±0.0**	**100.0±0.00**
12	70.03±4.64	83.41±1.18	74.77±7.46	81.33±5.81	96.01±5.01	97.76±1.14	98.96±0.58	80.0±11.13	99.96±0.13	**100.0±0.00**
13	96.26±0.43	98.74±0.11	91.89±4.13	96.43±3.39	99.33±2.12	98.55±0.47	**100.0±0.00**	99.63±0.67	**100.0±0.00**	**100.0±0.00**
OA	80.66±4.13	92.32±1.35	89.51±2.34	88.13±1.45	80.12±3.42	97.16±0.47	98.92±0.43	78.41±3.01	98.46±0.42	**99.74±0.33**
AA	78.71±4.27	90.91±2.34	87.82±2.46	86.74±1.31	73.26±4.29	96.16±0.71	98.72±0.63	66.48±6.09	97.62±0.66	**99.58±0.52**
Kappa	78.09±3.96	91.07±1.51	88.32±2.59	86.81±1.61	77.85±3.82	96.84±0.52	98.79±0.48	75.69±3.42	98.28±0.46	**99.71±0.36**

methods have difficulty in separating the mixed categories features and generate poor performance. Transformer-based methods employ the ability to extract accurate category features to obtain better performance than the CNN methods. Notably, transformer-based methods yield unsatisfactory classification accuracy for class 2 in the Indian Pines dataset.

This is attributed to the fact that samples belonging to class 2 comprise less than 20% of class-labeled pixels, as depicted in Fig. 1(b). Consequently, the presence of nonlabel-class pixels disrupts the intended focus areas of the self-attention module, which leads to a shift in attention toward non-label features. RvT as the standard transformer network in the computer vision field obtains 67.98% OA on the Indian Pines dataset, which proves that the traditional self-attention module may not exactly match the style of HSI classification.

While SSMTR has attempted to capture advanced semantic information by utilizing reconstructed image pixels, the extracted features with limited samples may not be sufficient for the category representation and ultimately lead to reduced classification accuracy. SSTN, SSFTT, HiT, and GAHT achieve relatively acceptable performance by embedding different attention modules. However, attention shifts and redundancy in the self-attention mechanism hinder these models from focusing on the category information and shift the attention to interference pixels, which limit further performance improvement. CTMixer achieves a relatively competitive performance on four datasets, which benefit from the parallel-backbone framework. The result of CTMixer demonstrates that the convolution branch captures pixel-level local information to correct attention shifts and redundancy in the self-attention mechanism. Certainly, the lack of a spectral refining module prevents further performance improvement.

In comparison, the CP-Transformer achieves the best performance nearly at 99% OA on four datasets, which is attributed to two main aspects. First, we apply the CFA mechanism to extract accurate class-label features while avoiding the phenomenon of attention shifts and redundancy. Second, the SSFA module refines the spectral information to alleviate the spectral redundancy, leading the improved discrimination ability and excellent classification performance.

2) Visualization: The visual classification maps are shown in Figs. 6–9. As observed, the convolution-based methods usually learn the pixel-level local information, while the transformer-based methods tend to construct global relationships. However, due to the lack of spatial relationship distribution, convolution models including RSSAN and pResnet misclassify the boundary pixels and thus have difficulty maintaining the edge continuity of ground objects. Especially in Fig. 8(c) and (d), the result maps of Commercial and Trees are interfered with by other categories of pixels, which causes large misclassifications. The transformer-based models have encountered difficulties in identifying some discrete and irregular ground objects such as Corn-notill and Corn-mintill in Fig. 6(e)–(l), Commercial and Road in Fig. 8(e)–(l), and Hardwood in Fig. 9(e)–(l). The phenomenon shows that the self-attention mechanism in the loss of the reconstruction task.

Furthermore, SSTN, SSFTT, and CTMixer expect to employ the spectral-spatial joint feature to enhance the detail and edges of ground objects. However, SSTN, SSFTT, and CTMixer are influenced by attention shifts, and the correlation signal of the class-label pixel is attenuated or shifted to the other categories, which hampers the preservation of intricate edge details. Consequently, compared with other methods,

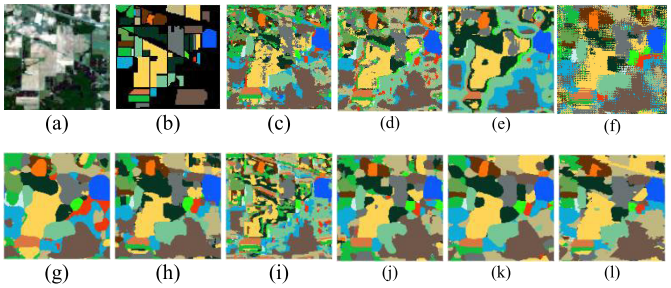

Fig. 6. Classification maps obtained by different classification methods for Indian Pines. (a) Three-band false color composite. (b) Ground truth. (c) RSSAN. (d) pResNet. (e) HIT. (f) SSMTR. (g) SSTN. (h) GAHT. (i) RVT. (j) SSFTT. (k) CTMixer. (l) CP-Transformer.

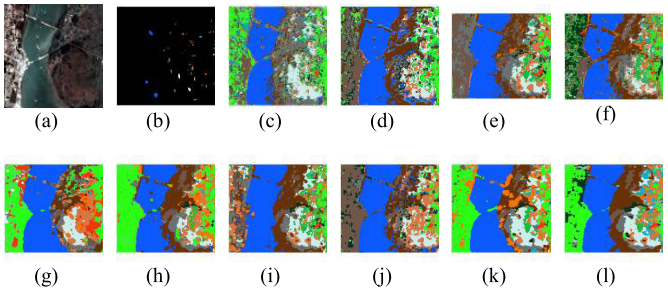

Fig. 7. Classification maps obtained by different classification methods for KSC. (a) Three-band false color composite. (b) Ground truth. (c) RSSAN. (d) pResNet. (e) HIT. (f) SSMTR. (g) SSTN. (h) GAHT. (i) RVT. (j) SSFTT. (k) CTMixer. (l) CP-Transformer.

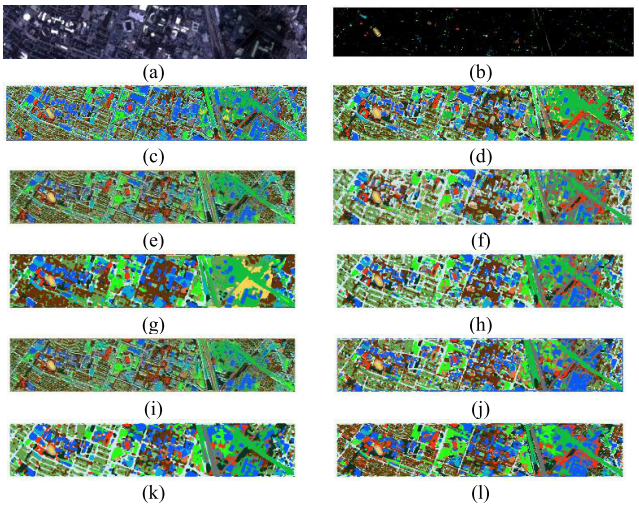

Fig. 8. Classification maps obtained by different classification methods for Houston. (a) Three-band false color composite. (b) Ground truth. (c) RSSAN. (d) pResNet. (e) HIT. (f) SSMTR. (g) SSTN. (h) GAHT. (i) RVT. (j) SSFTT. (k) CTMixer. (l) CP-Transformer.

the CP-Transformer accurately classifies multicategory ground objects including some discrete and mixed pixels, generates more continuous clutter boundaries, and further enhances the overall performance transformer model does not remove the interference factors and establish a clear and continuous classification boundary. Similarly, SSMTR attempts to reconstruct category relationship distribution, but the classification map generated by SSMTR is blurry and fragmented due to the object detail loss of the reconstruction task. Furthermore, SSTN, SSFTT, and CTMixer expect to employ the spectral-spatial joint feature to enhance the detail and edges

TABLE V
CLASSIFICATION PERFORMANCE OF THE APPROACHES ON BOTSWANA

Class	RSSAN	pResNet	HiT	SSMTR	SSTN	RvT	GAHT	SSFTT	CTMixer	Ours
1	98.52±0.41	99.30±0.10	99.33±0.94	79.7±8.79	95.04±7.34	99.11±0.94	95.52±1.55	94.34±3.45	98.82±1.69	**100.0±0.0**
2	91.09±3.35	97.41±0.60	99.31±1.55	97.43±2.34	90.69±15.86	100.0±0.0	99.9±0.31	99.60±0.83	**100.0±0.0**	**100.0±0.0**
3	95.70±1.15	96.13±1.31	96.06±3.29	81.08±13.35	89.6±13.47	90.24±4.24	99±1.08	97.09±1.49	**99.6±0.99**	99.52±0.92
4	79.14±5.13	94.23±1.27	98.04±2.58	95.35±5.96	90.14±20.67	99.39±0.58	100.0±0.0	99.35±1.37	**100.0±0.00**	99.86±0.23
5	85.16±5.19	92.19±1.57	85.32±4.47	81.38±5.00	37.51±28.73	90.01±3.60	92.68±4.16	90.30±3.92	92.98±2.14	**97.10±2.84**
6	90.09±3.91	90.19±3.21	97.43±1.35	81.90±7.25	44.61±37.58	91.01±4.75	97.06±2.95	90.19±4.25	98.96±1.41	**99.26±1.61**
7	97.20±1.01	97.21±2.51	**100.0±0.00**	96.95±4.22	97.92±5.16	88.38±5.31	87.76±5.05	**99.92±0.16**	99.19±0.82	99.54±1.07
8	90.61±1.90	99.31±0.37	99.31±1.38	87.29±8.02	82.12±30.16	97.83±2.97	98.28±4.28	99.31±1.25	**100.0±0.00**	**100.0±0.00**
9	88.13±2.67	98.41±0.31	95.61±1.27	91.53±10.37	97.87±5.55	95.35±4.83	99.90±0.22	99.01±2.48	**100.0±0.00**	99.97±0.10
10	81.23±4.01	98.13±0.82	98.83±1.25	95.00±6.83	91.01±21.2	96.21±6.35	99.48±1.66	99.88±0.38	**100.0±0.00**	99.96±0.13
11	90.16±1.62	99.02±0.37	92.36±5.07	95.90±4.75	87.05±12.95	94.36±5.80	98.79±2.01	98.20±2.78	99.93±0.21	**100.0±0.00**
12	91.16±1.39	97.11±0.61	94.81±4.08	91.94±6.69	98.01±4.68	95.14±4.70	97.90±2.59	95.97±5.40	99.95±0.17	**100.0±0.00**
13	91.30±2.97	98.88±0.23	94.18±4.05	95.00±4.84	87.95±28.69	93.81±3.80	**100.0±0.0**	99.74±0.61	**100.0±0.00**	**100.0±0.00**
14	97.37±0.41	99.16±0.44	93.26±3.37	83.16±0.01	81.37±4.66	82.74±1.33	83.16±0	86.42±5.62	83.16±0.02	**87.69±5.16**
OA	86.64±3.37	97.37±0.93	95.82±0.84	89.44±3.59	83.03±4.41	93.94±1.43	96.84±0.84	96.67±0.62	98.64±0.19	**99.25±0.46**
AA	87.95±2.91	97.52±0.61	95.99±0.88	89.54±3.36	83.64±4.29	93.83±1.35	96.39±0.79	96.38±0.70	98.04±0.17	**98.78±0.59**
Kappa	86.91±3.71	96.73±1.12	95.47±0.91	88.57±3.89	81.65±4.76	93.44±1.55	96.58±0.91	96.39±0.67	98.52±0.21	**99.19±0.50**

TABLE VI
CLASSIFICATION PERFORMANCE OF THE APPROACHES ON THE HOUSTON

Class	RSSAN	pResNet	Hit	SSMTR	SSTN	RvT	GAHT	SSFTT	CTMixer	Ours
1	92.51±1.42	93.84±1.33	81.84±1.61	77.13±3.02	91.54±6.37	82.10±5.95	84.84±2.17	92.86±2.17	89.33±1.47	**99.92±0.10**
2	81.26±4.53	97.13±0.31	91.50±3.30	97.02±1.45	84.88±5.12	91.76±2.70	97.94±2.27	96.45±2.01	97.41±0.11	**99.49±1.14**
3	95.41±4.11	96.56±0.27	97.39±0.63	98.82±0.65	95.62±1.32	98.14±0.76	98.8±0.75	98.52±1.01	98.51±0.39	**100.0±0.00**
4	87.22±3.63	93.89±1.78	84.74±5.31	84.92±3.51	90.00±2.28	83.43±8.21	93.03±1.76	93.33±1.3	95.85±2.12	**98.34±2.34**
5	98.11±0.21	97.76±0.41	97.63±1.03	92.58±1.91	**100.0±0.00**	96.32±2.47	98.05±2.39	96.28±2.47	99.78±0.22	**100.0±0.00**
6	85.23±4.69	92.92±1.43	92.55±6.13	95.76±1.65	82.77±5.46	98.89±1.62	98.18±0.47	91.20±5.01	98.40±0.39	**99.94±0.14**
7	89.59±1.51	85.65±3.36	60.87±5.52	83.30±2.77	51.55±28.21	81.36±3.55	90.24±2.13	87.54±3.68	91.45±0.93	**97.32±1.15**
8	65.06±4.13	61.56±7.62	60.53±1.57	73.76±2.01	59.65±12.71	69.61±4.54	80.34±2.97	84.86±1.79	87.33±2.21	**95.82±0.7**
9	78.53±5.19	83.91±5.86	68.93±6.23	83.12±1.93	54.86±10.91	72.33±4.49	86.16±3.96	86.13±2.09	87.89±1.23	**97.19±0.46**
10	69.76±2.20	77.85±9.34	76.33±5.59	90.73±5.01	70.11±14.32	76.69±6.94	92.31±4.09	96.81±1.72	97.97±2.67	**99.45±1.24**
11	71.46±6.10	85.15±5.03	59.37±12.29	90.23±1.96	62.55±15.32	77.60±5.22	95.66±1.49	95.34±2.31	97.80±0.95	**98.82±2.28**
12	79.48±7.21	86.51±3.19	72.10±2.90	81.90±4.25	47.15±8.84	67.43±8.56	85.3±1.77	90.57±2.56	91.87±3.25	**99.94±0.04**
13	69.94±12.1	95.76±1.62	55.74±5.40	91.64±1.07	91.45±5.51	80.6±3.47	97.25±2.17	97.87±3.48	**98.38±2.29**	96.63±0.96
14	89.25±2.43	98.86±0.34	96.87±0.78	99.93±0.22	99.91±0.23	99.25±1.03	**100.0±0.0**	94.51±4.53	**100.0±0.00**	99.91±0.13
15	96.73±1.34	99.04±0.31	98.97±1.28	98.97±1.66	99.96±0.14	99.91±0.16	97.05±1.46	99.83±0.43	**100.0±0.00**	**100.0±0.00**
OA	79.37±4.46	89.37±2.35	77.81±2.82	87.50±0.95	75.32±3.68	82.60±0.99	91.73±0.56	92.88±0.52	94.59±0.44	**98.75±0.31**
AA	80.16±3.15	90.34±1.52	79.69±2.76	89.32±0.81	78.8±3.05	85.03±0.91	93.01±0.51	93.47±0.5	95.46±0.40	**98.85±0.26**
Kappa	78.29±5.01	88.03±2.93	76.02±3.04	86.50±1.03	73.4±3.95	81.19±1.07	91.06±0.61	92.31±0.57	94.15±0.47	**98.65±0.33**

of ground objects. However, SSTN, SSFTT, and CTMixer are influenced by attention shifts, and the correlation signal of the class-label pixel is attenuated or shifted to the other categories, which hampers the preservation of intricate edge details. Consequently, compared with other methods, the CP-Transformer accurately classifies multicategory ground objects including some discrete and mixed pixels, generates more continuous clutter boundaries, and further enhances the overall performance.

3) Efficiency Comparison: We compare the computational efficiency of all the methods, and the results are shown in Table VII. Briefly speaking, the proposed CP-Transformer strikes a balance between satisfactory classification accuracy and computational cost. Specifically, CNN-based models demonstrate superior efficiency by utilizing shared CNN kernels for feature extraction, and thus the RSSAN model has a relatively balanced running cost. Compared with the traditional RVT and HiT models, our model is more lightweight owing to the HIN network, which employs a convolution network to build the main framework for HSI classification. Since residual operations bring low computational efficiency, pResnet exhibits a training parameter size that is four times larger than our model.

Besides, typical lightweight transformer-based frameworks such as SSTN and SSFTT have fewer Param and FLOPS, yet achieve unsatisfactory OA that is approximately lower 20% than the proposed method. The CFA attention mechanism aims to refine self-attention for HSI classification, which adapts to the sample labeling scheme and further reduces inference time compared with traditional transformer networks such as GAHT and CTMixer. Indeed, our method achieves excellent performance with a longer training time as a trade-off. The reason is attributed to two aspects. First, the PyTorch framework lacks the optimized calculation pattern for the CFA module. Second, the sequential connection between the CFA and the SSFA module increases training time.

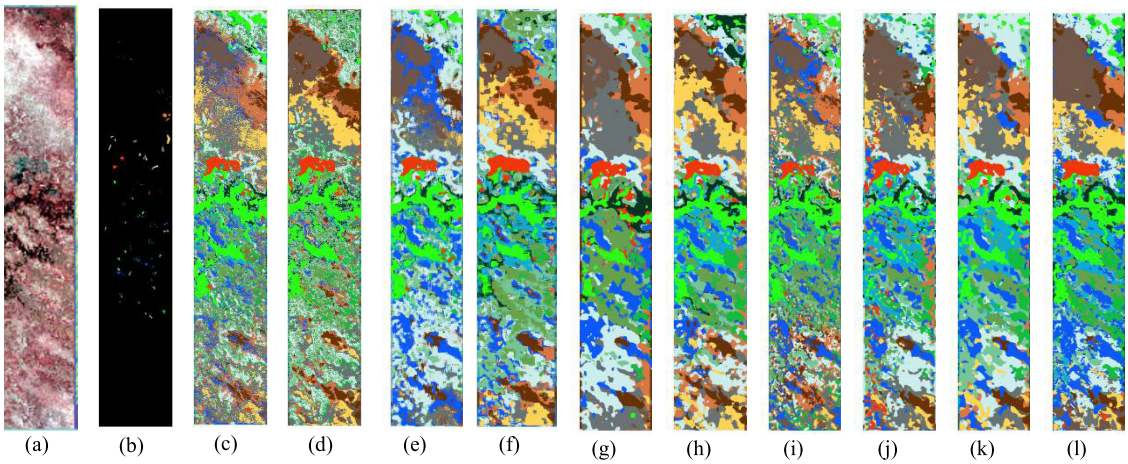

Fig. 9. Classification maps obtained by different classification methods for Botswana. (a) Three-band false color composite. (b) Ground truth. (c) RSSAN. (d) pResNet. (e) HIT. (f) SSMTR. (g) SSTN. (h) GAHT. (i) RVT. (j) SSFTT. (k) CTMixer. (l) CP-Transformer.

TABLE VII
COMPARISON OF THE COMPUTATION COST TRAINING TIME AND TEST TIME

Dataset	Model	RSSAN	pResNet	HiT	SSMTR	SSTN	RvT	GAHT	SSFTT	CTMixer	Ours
Indian	Param(M)	0.19	1.12	51.23	1.50	0.02	8.94	0.97	0.93	0.61	0.33
	FLOPS(M)	0.32	0.92	18.68	1.21	0.08	3.38	3.50	1.90	2.21	1.20
	Train(S)	210.46	476.25	2085.05	505.03	308.05	251.28	556.96	219.29	289.78	615.94
	Test(S)	4.21	5.41	9.09	4.27	4.14	6.17	5.47	3.97	4.79	4.63
Houston	Param(M)	0.19	1.15	46.13	1.48	0.04	8.87	0.97	0.67	0.65	0.33
	FLOPS(M)	0.34	0.96	14.18	1.13	0.15	3.35	3.50	1.36	2.34	1.17
	Train(S)	237.42	510.53	1848.	215.78	288.26	293.23	353.56	205.77	266.35	552.47
	Test(S)	5.73	6.81	12.39	6.11	5.86	8.67	8.11	5.50	6.74	6.85

D. Architectural Analysis of the CP-Transformer

1) Distribution of SSFA Module: In this part, we conducted experiments to investigate the effect on the SSFA module, the different allocation strategies obtain different classification results. The results are listed in Table VIII. As can be observed, without the SSFA module, the model only selects category features from the spatial domain and achieves about 95% OA on two datasets. Incorporating one single SSFA module, the model achieves an accuracy improvement of approximately 2.5%, which demonstrates that the SSFA module extracts beneficial spectral features to enhance discriminative category features for HSI classification. Moreover, the SSFA module exhibits a more significant performance on the Indian Pines dataset, since the Indian Pines dataset has lower spatial resolution yet richer spectral features. When the spatial information is insufficient for accurate classification, the SSFA module compensates by extracting sufficient spectral features, which enables better expressions of the spatial-spectral joint information.

2) Number of Stages: To better explore the effect of network depth, the base model (CP-Transformer without the SSFA module) has conducted several experiments on the Indian Pines dataset. As depicted in Fig. 10, the average classification accuracy exceeds about 94% OA for models with depths of 2, 3, and 4 stages, and the highest accuracy reaches 96%, which indicates that the appropriate depth model balances the parameters scale and the operational efficiency and brings relatively stable classification results. Besides, the OA will decline in five five-block model due to the excessive learnable parameters. Considering the overall situation, we set the four-block networks as the default framework.

TABLE VIII
CLASSIFICATION PERFORMANCE OF DIFFERENT LAYERS OF SSFA

1-layer	2-layer	3-layer	4-layer	Indian	Houston
×	×	×	×	94.72±0.71	96.72±1.06
✓	×	×	×	97.26±1.21	97.77±0.51
×	✓	×	×	97.93±0.63	98.19±0.40
×	×	✓	×	97.52±1.40	97.67±1.11
×	×	×	✓	94.72±2.86	97.59±0.64
✓	✓	×	×	98.61±0.32	97.68±0.19
×	✓	×	✓	97.63±0.92	98.19±0.15
×	✓	✓	×	98.62±0.40	97.53±0.67
×	×	✓	✓	97.95±0.86	97.18±0.89
✓	✓	✓	×	99.04±0.29	98.75±0.31
✓	✓	✓	✓	98.55±0.22	97.97±0.42

E. Generalization and Ablation Studies

1) Impact of Sample Size: To better display the effect of spatial generalization ability about all methods for different spatial sizes, we conducted a series of experiments on Indian Pines and Houston datasets, and the results are

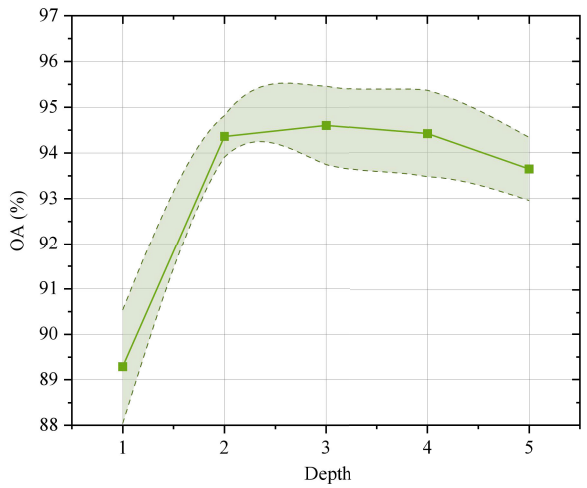

Fig. 10. OA results (%) of different depths on the Indian Pines dataset.

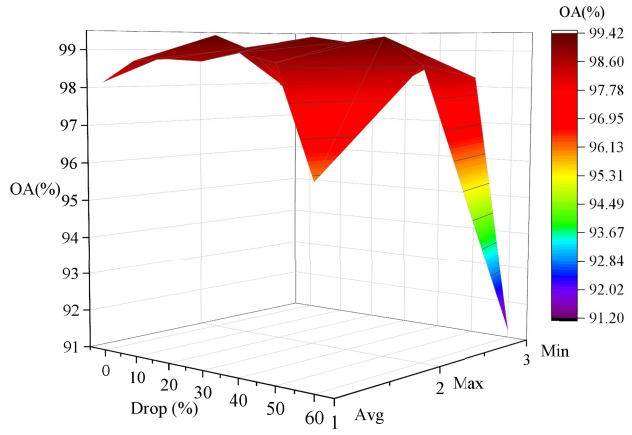

Fig. 13. Performance of different dropout rates in the CP-Transformer.

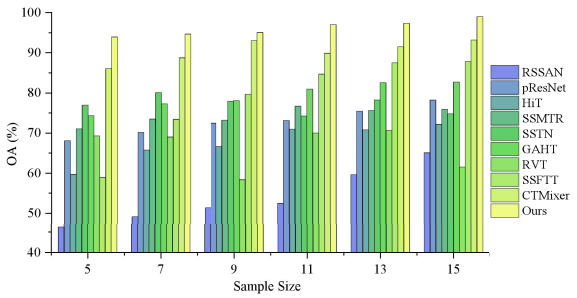

Fig. 11. Classification performance on Indian Pines with different sizes of training samples between the compared methods.

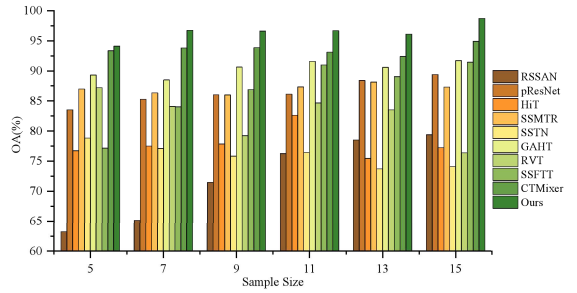

Fig. 12. Classification performance on Houston with different sizes of training samples between the compared methods.

shown in Figs. 11 and 12. As reported, the CP-Transformer consistently achieves superior classification accuracy of over 90% OA across different window sizes. The results prove that the CP-Transformer effectively utilizes the CFA mechanism to handle multiple-category distribution, which extracts contextual information from spatially homogeneous pixels and suppresses the expression of nonclass-label features. In particular, other methods struggle to obtain enough spatial information and generate classification accuracy of less than 60% on a 5×5 sample size. However, the CP-Transformer prioritizes the spectral information through the SSFA module, which effectively compensates for the lack of spatial information to obtain better classification accuracy.

2) Impact of Drop Proportion: The AA obtained with different drop rates is illustrated in Fig. 13. It can be observed that the model achieves optimal performance with a 20% dropout proportion. As the dropout proportion increases to 60%, the classification accuracy diminishes, which indicates an excessive discarding of feature data. Considering the varying training complexities of different datasets, we choose an intermediate dropout value of 0.3 as the default setting.

3) Impact of Training Sample Number: The results of different methods with different training sample numbers are illustrated in Fig. 14(a) and (b). In general, transformer-based methods require a large number of training samples to obtain better classification results. When coping with fewer trainable samples, these methods may have difficulty converging and yield poor performance, such as HiT, GAHT, and RVT only achieving less than 75% OA on five group experiments of the Indian Pines dataset. In comparison, our method is relativity not sensitive to the scale of training samples and obtains the best classification accuracy on each group experiment of two datasets. Besides, by effectively capturing spatial-spectral information, the CP-Transformer is more robust and requires fewer training samples to achieve excellent classification performance than the compared approaches. Especially, when the sample numbers are less than 1% (four samples per class) on Indian Pines and Houston datasets, the CP-Transformer still achieves a classification accuracy of approximately 90%.

4) Ablation Study: To verify the effect of the key components of the CP-Transformer network, four infrastructures have been constructed by SSFA, CFA, and HIN modules in different Ways. Table IX displays the performance of each model in terms of feature extraction and de-category mixing. As observed, the CPT-A (IPE + HIN) obtains over 90% OA on four datasets, 98% OA on the Botswana dataset especially. It illustrates that the HIN network can serve as an excellent hierarchical learning network. Besides, the CPT-B (IPE + CFA + HIN) and the CPT-C (IPE + SSFA + HIN) networks separately incorporate the CFA or SSFA modules to extract spatial or spectral features. By incorporating the SSFA module, OA in the CPT-C model improves by approximately 6% on the Indian Pines dataset, which indicates that the SSFA

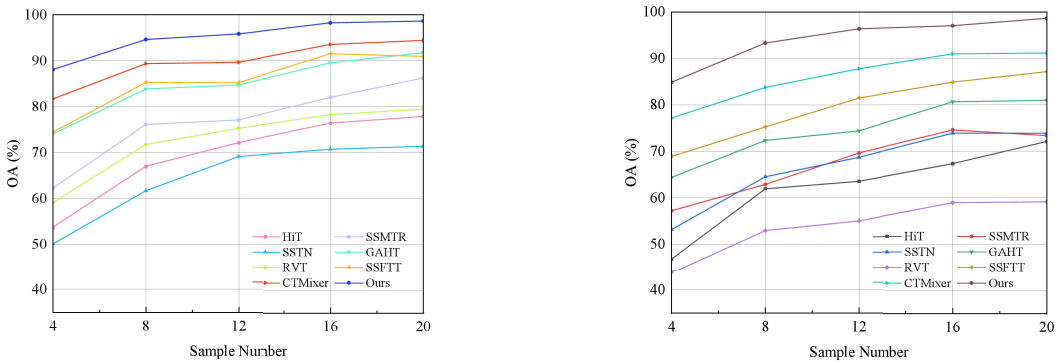

Fig. 14. Classification performance with different numbers of training samples of each category between the compared methods. (a) Houston. (b) Indian Pines.

TABLE IX
CLASSIFICATION PERFORMANCE OF DIFFERENT ARCHITECTURE

	Indian	Houston	Botswana	KSC
CPT-A	90.35±0.26	91.74±0.40	97.93±0.34	98.16±0.53
CPT-B	94.61±0.39	95.27±0.48	99.41±0.44	99.52±0.16
CPT-C	96.87±1.66	93.39±0.51	99.19±0.50	98.26±0.62
CPT-D	99.06±0.36	98.68±0.33	99.21±0.35	99.72±0.63

TABLE X
COMPARISON OF THE DIFFERENT MODEL STRUCTURE

Dataset	\	CP-T	CP-TS	CP-T*
Indian Pines	OA	**99.04±0.29**	92.67±0.69	83.19±87.62
	AA	**99.39±0.13**	95.7±0.33	92.09±92.93
	Kappa	**98.91±0.33**	91.57±0.67	81.14±85.95
Houston	OA	**98.75±0.31**	95.23±0.56	80.6±5.79
	AA	**98.85±0.26**	96.34±0.83	82.87±4.85
	Kappa	**98.65±0.33**	94.81±0.56	79.03±6.26

module extracts salient spectral features and reduces redundant information.

Meanwhile, the CPT-B with the CFA module achieves 4% OA improvement by enhancing the information description of class-label pixels and suppressing the representation of nonclass-label pixels. This phenomenon indicates that both local spatial features and the long-range contextual category relationships from the spatial domain are important for HSI classification. Finally, the CPT-D model (IPE + CFA + SSFA + HIN) integrating the CFA and SSFA modules effectively captures joint category information from both the spatial and spectral domains, which enables the joint transmission of class-label-related features to the HIN network and facilitates the extraction of more discriminative features. Compared to the CPT-A models, the CPT-D network achieves a remarkable increase of nearly 8% OA on the Indian Pines dataset and 7% OA in the Houston dataset.

5) Impact of the HIN: In this experiment, we compared three approaches to evaluate the impact of the HIN block on the Indian Pines and Houton datasets. Specifically, the implementations include CP-T (CFA + SSFA + HIN), CP-TS (CFA + SSFA + ResBlock), and CP-T* [CFA + SSFA + feed forward (FF)]. As shown in Table X, we can observe that the HIN block is more adaptive to our framework compared to the traditional FF block. The reason lies in the CFA module requires relative spatial positions of feature elements to compute attention. While in the FF block, the positional encoding features could be lost during the matrix transformation process. Rather than relying on additional positional encoding, we build the HIN that maintains the relative positional relationships of elements while keeping the feature map dimensions constant. Overall, in comparison to FF networks, the HIN computation approach is simpler and more appropriate for the attention calculation mechanism in the CFA of our model. Additionally, we conducted an experiment that employed a single-layer ResNet as a feature learning network for comparative experiments. The results indicate that the convolution-based block is more compatible with our model than the FF block, and improves performance by nearly 10%, which further validates the effectiveness of the HIN.

IV. CONCLUSION

In this article, we presented the category-guided transformer framework named CP-Transformer for HSI classification, which aims to alleviate the problem of attention issues caused by traditional self-attention mechanisms. As the essential component of our model, CFA suppresses the expression of interfering pixels and enhances the semantic priority of real category pixels. Significantly, CFA effectively reduces the computational cost from squared complexity to linear complexity by adjusting attention computation architecture. Furthermore, to extract the salient spectral information, the SSFA focuses on refining the spectral features to enhance category features. To integrate the spectral and spatial information, we employ a spatial-spectral fusion network denoted as the HIN module to combine the refined spectral and spatial features jointly for favorable information. The superiority of our approach has been extensively demonstrated through extensive experiments on multiple datasets. Since the spectral variability of different scenes disrupts the effective capture of attention information related to the center-labeled class, the

performance of our method may be hindered in HSI cross-classification scenarios. In the future, we plan to merge causal knowledge to improve the presented attention mechanism in collaborative learning frameworks, which is beneficial in the complex scenarios of HSI classification.

REFERENCES

[1] 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, doi: 10.1109/TIP.2018.2809606.

[2] C. Yu, Y. Zhu, M. Song, Y. Wang, and Q. Zhang, "Unseen feature extraction: Spatial mapping expansion with spectral compression network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5521915, doi: 10.1109/TGRS.2024.3420137.

[3] Y. Wang, X. Chen, E. Zhao, C. Zhao, M. Song, and C. Yu, "An unsupervised momentum contrastive learning based transformer network for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9053–9068, 2024.

[4] Y. Zhong, X. Wang, S. Wang, and L. Zhang, "Advances in spaceborne hyperspectral remote sensing in China," *Geo-Spatial Inf. Sci.*, vol. 24, no. 1, pp. 95–120, Jan. 2021.

[5] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 278–294, Jan. 2022, doi: 10.1080/10095020.2021.2017237.

[6] B. Li et al., "Integrating urban morphology and Land Surface Temperature characteristics for urban functional area classification," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 337–352, Jan. 2022.

[7] Q. Zhang, Y. Dong, Q. Yuan, M. Song, and H. Yu, "Combined deep priors with low-rank tensor factorization for hyperspectral image restoration," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[8] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.

[9] Y. Wang, X. Chen, E. Zhao, and M. Song, "Self-supervised spectral-level contrastive learning for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510515, doi: 10.1109/TGRS.2023.3270324.

[10] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "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, doi: 10.1109/JSTARS.2014.2329330.

[11] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015, doi: 10.1109/LGRS.2015.2482520.

[12] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015, doi: 10.1109/JSTARS.2015.2388577.

[13] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017, doi: 10.1109/TGRS.2017.2675902.

[14] 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, doi: 10.1109/LGRS.2019.2918719.

[15] Z. Yang, Z. Xi, T. Zhang, W. Guo, Z. Zhang, and H.-C. Li, "CMR-CNN: Cross-mixing residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8974–8989, 2022, doi: 10.1109/JSTARS.2022.3213865.

[16] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019, doi: 10.1109/TGRS.2018.2860125.

[17] 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, doi: 10.1109/TGRS.2020.2999957.

[18] J. Yang, B. Du, Y. Xu, and L. Zhang, "Can spectral information work while extracting spatial distribution—An online spectral information compensation network for HSI classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2360–2373, 2023, doi: 10.1109/TIP.2023.3244414.

[19] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[20] C. Yu, J. Huang, M. Song, Y. Wang, and C.-I. Chang, "Edge-inferring graph neural network with dynamic task-guided self-diagnosis for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535613, doi: 10.1109/TGRS.2022.3196311.

[21] Q. Yu, W. Wei, Z. Pan, J. He, S. Wang, and D. Hong, "GPF-Net: Graph-polarized fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5519622.

[22] J. Kang, Y. Zhang, X. Liu, and Z. Cheng, "Hyperspectral image classification using spectral–spatial double-branch attention mechanism," *Remote Sens.*, vol. 16, no. 1, p. 193, Jan. 2024.

[23] A. Jha, S. Bose, and B. Banerjee, "GAF-net: Improving the performance of remote sensing image fusion using novel global self and cross attention learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6343–6352, doi: 10.1109/WACV56688.2023.00629.

[24] T. Arshad, J. Zhang, S. C. Anyembe, and A. Mehmood, "Spectral spatial neighborhood attention transformer for hyperspectral image classification," *Can. J. Remote Sens.*, vol. 50, no. 1, May 2024, Art. no. 2347631, doi: 10.1080/07038992.2024.2347631.

[25] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501916, doi: 10.1109/TGRS.2021.3058549.

[26] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532513, doi: 10.1109/TGRS.2022.3186400.

[27] B. Zhang, Y. Chen, Z. Li, S. Xiong, and X. Lu, "SANet: A self-attention network for agricultural hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5501315, doi: 10.1109/TGRS.2023.3341473.

[28] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2536–2551, 2023, doi: 10.1109/TIP.2023.3270104.

[29] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535317, doi: 10.1109/TGRS.2022.3196661.

[30] Q. Hong et al., "SATNet: A spatial attention based network for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 22, p. 5902, Nov. 2022, doi: 10.3390/rs14225902.

[31] 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, doi: 10.1109/TGRS.2022.3144158.

[32] C. Zhao et al., "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, 2023, doi: 10.1109/TIP.2023.3287738.

[33] J. Yang, B. Du, and L. Zhang, "Overcoming the barrier of incompleteness: A hyperspectral image classification full model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14467–14481, Oct. 2024, doi: 10.1109/TNNLS.2023.3279377.

[34] Z. Li, Z. Xue, Q. Xu, L. Zhang, T. Zhu, and M. Zhang, "SPFormer: Self-pooling transformer for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502019, doi: 10.1109/TGRS.2023.3345923.

[35] D. Liao, C. Shi, and L. Wang, "A spectral–spatial fusion transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515216, doi: 10.1109/TGRS.2023.3286950.

[36] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[37] H. Yu, Z. Ling, K. Zheng, L. Gao, J. Li, and J. Chanussot, "Unsupervised hyperspectral and multispectral image fusion with deep spectral–spatial collaborative constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534114, doi: 10.1109/TGRS.2024.3472226.

[38] Z. Qiu, J. Xu, J. Peng, and W. Sun, "Cross-channel dynamic spatial–spectral fusion transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528112, doi: 10.1109/TGRS.2023.3324730.

[39] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When multi-granularity meets spatial–spectral attention: A hybrid transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401118, doi: 10.1109/TGRS.2023.3242978.
[40] X. Shang, S. Han, and M. Song, "Iterative spatial–spectral training sample augmentation for effective hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3131373.
[41] 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, doi: 10.1109/TGRS.2016.2616489.
[42] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3208935.
[43] 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, doi: 10.1109/TGRS.2022.3207933.
[44] C. Zhao, B. Qin, S. Feng, W. Zhu, L. Zhang, and J. Ren, "An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5546216, doi: 10.1109/TGRS.2022.3230378.
[45] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Oct. 2021, pp. 1–22.
[46] 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, doi: 10.1109/TGRS.2020.2994057.
[47] L. Huang, Y. Chen, and X. He, "Spectral–spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508718, doi: 10.1109/TGRS.2023.3264235.
[48] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715, doi: 10.1109/TGRS.2021.3115699.
[49] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "HiT: Hierarchical transformer with momentum contrast for video-text retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11895–11905, doi: 10.1109/ICCV48922.2021.01170.
[50] X. Mao et al., "Towards robust vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12032–12041, doi: 10.1109/CVPR52688.2022.01173.
[51] 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, doi: 10.1109/TGRS.2020.2994057.

Chunyan Yu (Senior Member, IEEE) received the Ph.D. degree in environmental engineering from Dalian Maritime University, Dalian, China, in 2012.

She is currently an Associate Professor with the Information Science and Technology College, Dalian Maritime University. Her research interests include image segmentation, hyperspectral image classification, and pattern recognition.

Yuanchen Zhu received the bachelor's degree from Qingdao Technology University, Qingdao, China, in 2022. He is currently pursuing the master's degree in computer science and technology with Dalian Maritime University, Dalian, China.

His research interests include hyperspectral image processing and deep learning.

Yulei Wang (Member, IEEE) received the B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively.

She is currently an Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her research interests include hyperspectral image processing and vital signs signal processing.

Dr. Wang was awarded by the China Scholarship Council in 2011 as a joint Ph.D. Student to study with the Remote Sensing Signal and Image Processing Laboratory, University of Maryland at Baltimore, Baltimore, MD, USA, for two years.

Enyu Zhao received the Ph.D. degree from the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. His research interests include quantitative remote sensing and hyperspectral image processing.

Qiang Zhang (Member, IEEE) received the B.E. degree in surveying and mapping engineering and the M.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017, 2019, and 2022, respectively.

He is currently an Xinghai Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. He has authored more than 20 journal articles in the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Earth System Science Data*, and *ISPRS Journal of Photogrammetry and Remote Sensing*. His research interests include remote sensing information processing, computer vision, and machine learning. More details could be found at https://qzhang95.github.io

Xiaoqiang Lu (Senior Member, IEEE) is currently a Full Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China. His research interests include intelligent optical sensing, pattern recognition, machine learning, and hyperspectral image analysis.