Hyperspectral Image Transformer Classification Networks

Xiaofei Yang, Weijia Cao, Yao Lu, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— Hyperspectral image (HSI) classification is an important task in earth observation missions. Convolution neural networks (CNNs) with the powerful ability of feature extraction have shown prominence in HSI classification tasks. However, existing CNN-based approaches cannot sufficiently mine the sequence attributes of spectral features, hindering the further performance promotion of HSI classification. This article presents a hyperspectral image transformer (HiT) classification network by embedding convolution operations into the transformer structure to capture the subtle spectral discrepancies and convey the local spatial context information. HiT consists of two key modules, i.e., spectral-adaptive 3-D convolution projection module and convolution permutator (ConV-Permutator) to retrieve the subtle spatial–spectral discrepancies. The spectral-adaptive 3-D convolution projection module produces the local spatial–spectral information from HSIs using two spectral-adaptive 3-D convolution layers instead of the linear projection layer. In addition, the Conv-Permutator module utilizes the depthwise convolution operations to separately encode the spatial–spectral representations along the height, width, and spectral dimensions, respectively. Extensive experiments on four benchmark HSI datasets, including Indian Pines, Pavia University, Houston2013, and Xiongan (XA) datasets, show the superiority of the proposed HiT over existing transformers and the state-of-the-art CNN-based methods. Our codes of this work are available at https://github.com/xiachangxue/DeepHyperX for the sake of reproducibility.**

***Index Terms*— 3-D convolution projection, convolution neural network (CNN), hyperspectral image (HSI) classification, transformers.**

I. Introduction

HYPERSPECTRAL images (HSIs) collecting hundreds of wavelength bands in spectral dimension at each pixel offer much more abundant spatial and spectral information for land cover recognition at a fine-grained level. This provides great potential in various high-precision earth observation missions, such as land cover identification [1], [2], urban change detection [3], [4], and environment monitoring [5], [6].

HSI classification extracts the spatial–spectral information to identify the pixels using various methods. The classification process mainly includes three steps: 1) image preprocessing (e.g., denoising [7], cloud removal [8], and missing data recovery [9]); 2) dimensionality reduction [10], [11]; and 3) feature extraction [12]–[14]. Among them, feature extraction is the key step to obtain high-precision classification results. Many hand-crafted-based methods of feature extraction for HSI classification have been developed in the past decade. For example, a support vector machine (SVM) [15] and K-nearest neighbor (KNN) [16] classifiers are widely used in HSI classification. Duan *et al.* [17] proposed a semisupervised method, called geodesic-based sparse manifold hypergraph, to extract HSI features by using nonlinear geodesic sparse hypergraphs. Luo *et al.* [11] proposed a multistructure unified discriminative embedding (MUDE) method for better representing the low-dimensional features, and achieved satisfactory classification results. However, these methods could not fit and represent a large number of complex data, resulting in the unsatisfactory performance.

Since deep learning architectures have great success in natural image recognition [18]–[22], many researchers apply them to HSI classification and propose various deep-learning-based HSI classification methods to extract more abundant features by inserting various models. For example, Boulch *et al.* [23] proposed an autoencoder (AE)-based network for HSI classification by using four AE layers followed by a Max-pooling layer. Yang *et al.* [24] presented convolution neural network (CNN)-based methods (e.g., 2-D-CNN and 3-D-CNN), stacking convolution operations (2-D or 3-D) followed by the BatchNorm (BN) [25] and rectified linear unit (ReLU) [26] activation function. It was noted that this article used a convolution layer with step size 2 to reduce dimension rather than the pooling layer. They have achieved satisfactory classification results, which attributes to the powerful ability to extract the local spatial context information. Other methods are recurrent neural network (RNN)-based HSI classification methods, which could handle the sequentiality data analysis. For example, Mou *et al.* [27] applied the RNNs to HSI classification and proposed an RNN classification network to mine the relationships of spectral bands. However, all these deep-learning-based HSI classification methods have some drawbacks.

1) *For CNNs:* All the CNN-based methods are susceptible to the lack of the ability to capture the subtle spectral

Manuscript received March 10, 2022; revised April 19, 2022; accepted April 27, 2022. Date of publication May 2, 2022; date of current version May 31, 2022. This work was supported by the University of Macau under Grant MYRG2018-00136-FST. *(Corresponding authors: Weijia Cao; Yicong Zhou.)*

Xiaofei Yang and Yicong Zhou are with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: xiaofei.hitsz@gmail.com; yicongzhou@um.edu.mo).

Weijia Cao is with the Department of Computer and Information Science, University of Macau, Macau, China, also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Yangtze Three Gorges Technology and Economy Development Company Ltd., Beijing 101100, China (e-mail: caowj@aircas.ac.cn).

Yao Lu is with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518057, China, and also with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: luyao2021@hit.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2022.3171551, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2022.3171551

discrepancies from the neighboring spectral bands, even though they have the powerful ability in capturing the local context information from HSIs. Second, CNN-based methods are overly concerned with spatial sequential information, resulting in misrepresenting the spectral sequential information in the extracted features and making it more difficult to mine and represent intrinsic and potentially spectral information.

2) *For RNNs:* The RNN-based methods suffer from severe gradient vanishing and hardly learn the long-term dependences due to their extreme dependence on the order of spectral bands, leading to a performance bottleneck in practical HSI classification.

Very recently, the transformer network [28], a novel deep learning mechanism, is proposed to solve natural image classification tasks from a sequence data perspective. Unlike CNNs and RNNs, transformer networks are more effective while analyzing the sequential data, mainly because of the self-attention techniques. This provides a new and effective approach for HSI classification. It is well known that the self-attention technique is the key module in transformers and can capture global information by encoding the positions. However, these transformer networks are proposed for natural image classification tasks. Some researchers applied the transformer to HSI classification. For example, He *et al.* [29] directly employed the transformer network for HSI classification and proposed the HSI-bidirectional encoder representations from transformer (BERT) by using bidirectional encoder representation from the transformer. He *et al.* [30] proposed a spatial–spectral transformer network for HSI classification. However, these two transformers directly used linear projection and do not consider the local spatial context information. In addition, the existing transformers have some problems, which restrict the further improvement of the performance for the HSI classification task. The problems are summarized as follows.

1) Although they perform well in solving the issue of long-term dependence of spectrum characteristics, they fail to capture the local spatial–spectral fusion information.
2) According to [31] and [32], due to the crucial 2-D convolution operations, transformer networks can capture the local spatial context information. However, 2-D convolution operations cannot satisfactorily retrieve the local spectral information.
3) Existing transformer networks encode spatial information sequentially using the flattening operation and linear projection, leading to the loss of the local spatial–spectral information and position information.

Based on the above-mentioned analysis, this article aims to propose a novel transformer-based classification network, called hyperspectral image transformer (HiT), to achieve the high-performance HSI classification by integrating the local feature and the global feature. Specifically, HiT consists of two key modules, namely, the spectral-adaptive 3-D convolution projection (SACP) module and the Conv-Permutator module. Different from the existing transformer-based HSI classification methods (such as [29] and [30]), our proposed HiT method is a vision transformer (ViT)-based method that could extract the local spatial–spectral fusion representations from the input HSIs using the SACP module and encode the representations using the Conv-Permutator module. More specifically, the SACP module is proposed to extract the local spatial information and adaptive capture the long-term spectral information. The Conv-Permutator module could separately encode the representations along the height, width, and spectral dimensions, respectively. In general, the main contributions of this article can be summarized as follows.

1) We propose a novel transformer-based HSI classification method called HiT. To the best of our knowledge, this article is the first to apply ViTs with convolution operation in the HSI classification.
2) We propose an SACP module to adaptively extract the spectral information, and further capture the spectral–spatial fusion information. To the best of our knowledge, this article is the first to employ the 3-D convolution operation to project the input HSIs.
3) We propose a new module, named Conv-Permutator, to capture much more spectral–spatial information by encoding the input representations along the height, width, and spectral dimensions, respectively.
4) The experimental results based on four benchmark datasets demonstrate that the proposed HiT outperforms the state-of-the-art transformers and CNN-based methods.

We organize the remaining of this article as follows. Section II discusses the related work about deep-learning-based HSI classification and transformer networks. Section III gives a brief illustration of the proposed HiT. Section IV illustrates the four benchmark HSI datasets, experimental settings, and experimental results and the corresponding analyses. Section V finally draws conclusions and a brief outlook of the future work.

II. Related Work

A. Deep-Learning-Based Methods for HSI Classification

Due to the great success of deep learning architectures (e.g., AEs, CNNs, and RNNs) on natural image recognition, many deep-learning-based methods are proposed for HSI classification [33]–[37]. For example, Chen *et al.* [38] and Boulch *et al.* [23] applied AEs into HSI classification, and proposed AE networks to extract deep features from HSIs, respectively. Mou *et al.* [27] and Hang *et al.* [39] utilized the RNNs to model the sequentiality data for HSI classification.

Since CNNs have the powerful ability in extracting the local spatial context information, many CNN-based approaches have been established for HSI classification. For example, Sharma *et al.* [40] proposed a 2-D-CNN-based approach for HSI classification by stacking 2-D convolution layers to capture the local spatial context information. Ran *et al.* [41] utilized two CNNs to extract spectral information and spatial context information, respectively. Some proposed modules were also inserted into the backbone networks to improve the classification performance. For example, Lorenzo *et al.* [42] proposed a new CNN-based method to classify the HSIs by applying the attention mechanism to

select the bands. Liu *et al.* [43] designed a semisupervised CNN-based method for HSI classification. Owing to the powerful ability of 3-D-CNNs in extracting the spatial–spectral fusion information, Yang *et al.* [24] and Chen *et al.* [44] proposed 3-D-CNN-based approaches for HSI classification. Lee and Kwon [45] designed a deep fully 3-D-CNN to capture the spatial–spectral information from HSIs. Li *et al.* [46] designed a new 3-D-CNN method without any pooling layers to produce a higher classification accuracy.

However, these deep learning approaches have their drawbacks in HSI classification. Although RNNs can model the sequentiality to represent the spectral bands, they may fail in processing the long-term dependences, because they are extremely dependent on ordered spectral bands. Although CNNs and their variants have achieved promising classification results, their inherent network and excessive concern with the local spatial information may not capture much more useful spectral sequentiality information. This will hinder them from generating a higher classification accuracy in HSI classification tasks.

B. ViTs for Image Classification

Recently, researchers rethink the natural image classification tasks from a sequence perspective using transformers networks and have proposed many transformer-based approaches for image classification. For example, Dosovitskiy *et al.* [47] first applied the transformer to the image classification task and proposed the ViT network. In ViT, the input images should be cropped into nine blocks, and then, the means of positional encoding are used to obtain the globally sequential information. However, the attention maps tend to be similar with the increase in the ViT's depth, which is called attention collapse. To overcome this issue, Zhou *et al.* [48] exchanged the attention weight of each head to recalculate the attention values using a learnable matrix, proposing the DeepViT for image classification. Nevertheless, these methods fail to capture the local spatial contextual information. Yuan *et al.* [49] downsampled the image sequence by unfolding to overlap the input image data in each token and proposed a new tokens-to-tokens (T2T) vision transformer for the image classification task.

Some researchers attempted to improve the performance of transformer networks by embedding the convolution techniques [50]–[57]. For example, Graham *et al.* [32] proposed a ViT in ConvNet's Clothing, called LeViT, by utilizing convolution projection instead of patchwise projection and adding extra nonlinearity in attention. Heo *et al.* [31] revisited the ViT and established a robust vision transformer (RvT) by combing some sturdy components, such as convolution layers, pooling layers, and more attention heads. All these ViTs, however, are established for the natural image classification tasks. Although there are some transformer-based HSI classification methods [29], [30], they fail in capturing the local spectral discrepancies in HSI classification tasks. In this article, we propose a novel ViT classification network named HiT to handle the HSI classification problems from a sequential perspective. HiT consists of two simple and effective modules, namely, SACP module used for capturing the local spatial–spectral fusion information and long-term spectral information, and Conv-Permutator with depthwise convolutions and pointwise convolutions utilized for separately encoding the representations along the height, width, and spectral dimensions, respectively.

III. Methodology

In this section, we will introduce the proposed HiT for HSI classification, including the details of SACP module and ConV-Permutator.

A. Overview of the Proposed HiT

In this article, we aim at designing a novel transformer-based method, i.e., HiT, for performing the high-precision classification of HSIs. HiT consists of two key modules for addressing the challenges in HSI classification, namely, SACP module that extracts the local spatial information and the long-term spectral information using the spectral-adaptive 3-D convolution layers and Conv-Permutator that separately encodes the representations along the height, width, and spectral dimensions using the depthwise and pointwise convolution layers. Thus, the proposed HiT can enhance the capturing capacity of the local spatial–spectral information and reduce the local information loss with deepening the networks. Fig. 1 illustrates an overview of the proposed HiT in the HSI classification task. We report the detailed definition of the proposed HiT in Table I. The first and second columns are the name of each block and its definition in terms of the layers. The third and fourth columns denote the input size and the output size of each block. It is worth noting that L and G denote the local branch and the global branch.

B. SACP Module

Our SACP is built with two spectral-adaptive 3-D convolution layers, shown in Fig. 2, which consists of two branches: local spatial branch L and global spectral branch G. The local spatial branch aims to learn the spatial location-sensitive importance map, and the global spectral branch adaptively aggregates the spectral information in a convolution manner. Suppose that $X \in R^{C\times S\times H\times W}$ is the input image (H and W are 15 in this article), where C represents the number of channels, and S, H, and W are its spectral–spatial dimensions. Thus, the SACP layer can be formulated as follows:

$$Y = G(X) \bigotimes \Big(L(X) \bigodot X \Big) \tag{1}$$

where $\bigotimes$ is the convolution operation and $\bigodot$ denotes the elementwise multiplication. It is worth noting that the above-mentioned two branches focus on different aspects of spectral–spatial information.

1) *For the Local Branch:* It attempts to capture the short-term spectral–spatial information and to attend the important features by using 3-D convolution.
2) *For the Global Branch:* It aims to incorporate long-range spectral information to conduct adaptive spectral aggregation by using 3-D convolution layers.

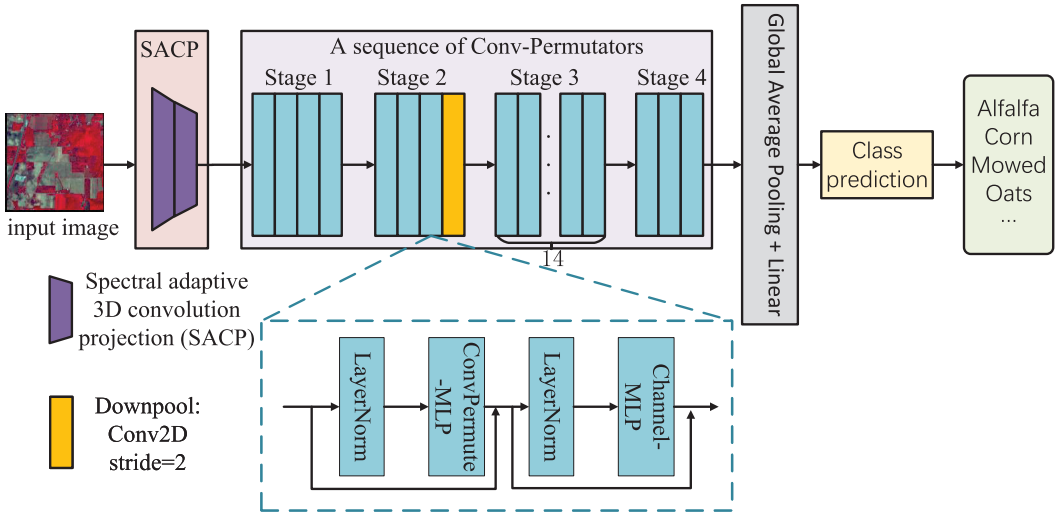


Fig. 1. Overall architecture of the proposed HiT. The SACP is spectral-adaptive 3D convolution projection module, which is used to extract the spatial–spectral representation from the input image, and then, this extracted representation is fed into a sequence of Conv-Permutators (e.g., four stages in this article) for feature encoding along the height, width, and spectral dimensions, respectively. Finally, a global average pooling layer and a fully connected layer are used for the class prediction. We note that the Conv3D and downpool layers are used to reduce the dimensions.

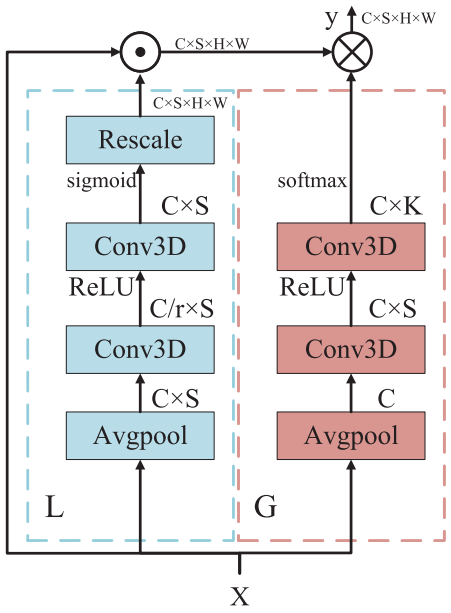


Fig. 2. Illustration of the proposed SACP layer. The SACP is composed of two branches: a local branch L and a global branch G. We note that $\otimes$ is the convolution operation and $\odot$ denotes the elementwise multiplication.

1) Local Branch: It is spatial location-sensitive and aims to leverage short-term spectral dynamics to perform local spatial–spectral feature extraction operations.

As shown in Fig. 2, the local branch is built by a sequence of 3-D convolution layers with ReLU [26]. Since the goal of the local branch is to capture short-term spatial–spectral information, we first average pooling the input image with the "AdaptiveAvgPool3d" operation, but not in the spectral dimension, and then set the first 3-D convolution kernel size K as (3, 1, 1) to learn an importance map solely based on a local spectral window. Specifically, the first Conv3D is followed by BN [14] and ReLU. Then, the second Conv3D with a Sigmoid activation yields the importance weights $W \in R^{C\times S}$ which are sensitive to spectral location. Finally, the spectral excitation is formulated as follows:

$$Z = F_{\text{rescale}}(W) \odot X = L(X) \odot X \tag{2}$$

$$W = L(X) = \text{Sigmoid}(\text{Conv3D}(\delta(\text{Conv3D}(X)))) \tag{3}$$

where $\odot$ and δ denote the elementwise multiplication and the activation function ReLU, respectively. In addition, $Z \in R^{C\times S\times H\times W}$. In order to match the size of X, $F_{\text{rescale}}(W)$ rescale the W to $\breve{W} \in R^{C\times S\times H\times W}$ by replicating in spatial dimension. Then, we can get the output of the local branch.

2) Global Branch: It incorporates global spectral information and learns to produce spectral-adaptive convolution kernel for dynamic aggregation.

As shown in Fig. 2, the global branch is built with two Conv3D layers, whose size is $1 \times 1 \times 1$. It is similar to the squeeze-and-excitation (SE) block, except that the convolution layer is the 3-D convolution layer. It is worth noting that the generated kernel is channelwise, which means that the global branch only models the spectral relations without considering the channel correlation. More formally, for the cth channel, the adaptive kernel is learned as follows:

$$\Theta_c = G(X)_c = \text{softmax}(F(W_2, \delta(F(W_1), \Phi(X)_c))) \tag{4}$$

where $\Theta_c \in R^K$ is generated adaptive kernel (aggregation weights) for the cth channel, K is the adaptive kernel size, and δ denotes the activation function ReLU. $F(W_1)$ and $F(W_2)$ denote the Conv3D layers, and $\Phi(\cdot)$ is the adaptive average pool operation. Since the learned adaptive kernel has the global receptive field, it could aggregate global spectral context. The learned aggregation weights $\Theta = \Theta_1, \Theta_2, \ldots, \Theta_c$ will be employed to perform spectral–spatial adaptive convolution.

3) Spectral-Adaptive Aggregation: In this step, we output the final features by combining the local branch and the global

TABLE I
ARCHITECTURAL DETAILS OF THE PROPOSED HiT

Block	Definition	Input size	Output size
Input	-	$1\times200\times15\times15$	$4\times100\times8\times8$
SACP-1	$\left\{\begin{array}{c} L:\begin{bmatrix} AvgPool \\ 3\times1\times1 Conv3D \\ 1\times1\times1 Conv3D \\ Sigmoid \end{bmatrix} \\ G:\begin{bmatrix} AvgPool \\ 1\times1\times1 Conv3D \\ 1\times1\times1 Conv3D \\ Softmax \end{bmatrix} \\ stride=(2,2,2) \end{array}\right\}$	$4\times100\times8\times8$	$4\times100\times8\times8$
SACP-2	$\left\{\begin{array}{c} L:\begin{bmatrix} AvgPool \\ 3\times1\times1 Conv3D \\ 1\times1\times1 Conv3D \\ Sigmoid \end{bmatrix} \\ G:\begin{bmatrix} AvgPool \\ 1\times1\times1 Conv3D \\ 1\times1\times1 Conv3D \\ Softmax \end{bmatrix} \\ stride=(2,1,1) \end{array}\right\}$	$4\times100\times8\times8$	$8\times50\times8\times8$
Conv-Permutator-1	$\left\{\begin{array}{c} LayerNorm \\ \begin{bmatrix} 1\times1 Conv2D \\ 1\times1 Conv2D \\ 1\times1 Conv2D \end{bmatrix} \\ LayerNorm \\ LayerNorm \end{array}\right\}\times4$	$400\times8\times8$	$400\times8\times8$
Conv-Permutator-2	$\left\{\begin{array}{c} LayerNorm \\ \begin{bmatrix} 1\times1 Conv2D \\ 1\times1 Conv2D \\ 1\times1 Conv2D \end{bmatrix} \\ LayerNorm \\ LayerNorm \end{array}\right\}\times3$	$400\times8\times8$	$400\times8\times8$
Downpooling	1×1 Conv2D stride=2	$400\times8\times8$	$400\times4\times4$
Conv-Permutator-3	$\left\{\begin{array}{c} LayerNorm \\ \begin{bmatrix} 1\times1 Conv2D \\ 1\times1 Conv2D \\ 1\times1 Conv2D \end{bmatrix} \\ LayerNorm \\ LayerNorm \end{array}\right\}\times14$	$400\times4\times4$	$512\times4\times4$
Conv-Permutator-4	$\left\{\begin{array}{c} LayerNorm \\ \begin{bmatrix} 1\times1 Conv2D \\ 1\times1 Conv2D \\ 1\times1 Conv2D \end{bmatrix} \\ LayerNorm \\ LayerNorm \end{array}\right\}\times3$	$512\times4\times4$	$512\times4\times4$
FC	LayerNorm	$512\times1\times1$	$16\times1\times1$
Softmax			

branch, which can be formulated as follows:

$$Y = G(X) \bigotimes Z \tag{5}$$

where Y is the final output features ($Y \in R^{C\times S\times H\times W}$).

In summary, our SACP presents an adaptive module and focuses on capturing different structures (i.e., short-term spatial–spectral and long-term spectral structures). It is worth noting that we finally reshape the output Y to $\check{Y} \in R^{(C\times S)\times H\times W}$, where the third dimension is the spectral dimension.

C. Conv-Permutator

According to Fig. 1, our Conv-Permutator module consists of two key components, i.e., ConvPermute and channel-master limited partnership (MLP) to encode the local spatial information and spectral information, respectively. The channel-MLP adopts a similar structure in ViP, consisting of two fully connected layers with a Gaussian error linear unit (GELU) [58] intermediate activation function. Sharing the similar processing of spatial encoding in ViP, the spatial–spectral information is processed along the height, width, and spectral dimensions, respectively. Different from ViP, we extract the spatial–spectral representations by utilizing the depthwise and pointwise convolution layers rather than linear projection. Given an input embedding D tokens $T \in R^{H\times W\times D}$, the outputs of Conv-Permutator can be formulated as

$$Y = \text{ConvPermute}(\text{LN}(T)) + T \tag{6}$$

$$Z = \text{channel-MLP}(\text{LN}(Y)) + Y \tag{7}$$

where LN denotes the LayerNorm, and Y and Z are the outputs of ConvPermute and channel-MLP, respectively. The ConV-Permutator in ConV-Permutator module will be introduced in detail as follows. The LayerNorm LN

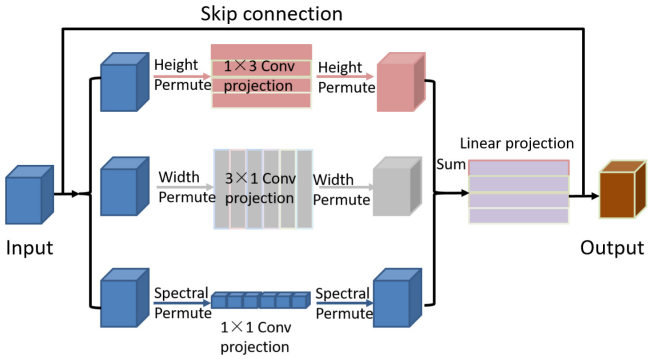


Fig. 3. Illustration of the proposed Conv-Permute module. The proposed Conv-Permute module first encodes the spatial–spectral features along the height, width, and spectral dimensions, respectively. The extracted features are then aggregated utilizing elementwise addition and a fully connected layer.

can be formulated as

$$y = \frac{x - E(x)}{\mathrm{Var}(x) + \epsilon} \tag{8}$$

where E is the mean, Var denotes the standard deviation, and ϵ is a very small constant, such as $1e-7$. x and y are the input feature maps and output feature maps, respectively.

ConvPermute: Fig. 3 illustrates the proposed ConvPermute. Since the features obtained from the previous projection module have three dimensions—H and W in the spatial domain, and S in the spectral domain—we propose a special architecture, named ConvPermute, that could divide the input features into three branches to encode the input representations along the height, width, and spectral dimensions, respectively.

Suppose that the hidden dimension S is 256, and the input representations are with size 8×8. To encode the spatial–spectral information, we conduct the height channel and width channel operations along the spatial dimensions, and spectral channel operation along the spectral dimension. The spectral channel operation captures the local spectral information from the input X using a simple pointwise convolution layer with weights $W_S \in R^{1\times1}$, generating the local spectral features X_S. Given the input $X \in R^{H\times W\times D}$, the spatial information encoder divides X into two branches: height-channel encoder and width-channel encoder. The height channel operation extracts the height local spatial information using a depthwise convolution layer with weights W_H, producing the local spatial information in height dimension X_H. The width channel operation utilizes a depthwise convolution layer with weights W_W to capture the width local spatial information, producing outputs of the local spatial information in height dimension X_W. We then simply fuse the output features obtained from three branches using the elementwise addition. To further improve the fused features, we recalibrate the importance of the three different branches by using a new fully connected layer. It can be calculated as

$$X_H = F_H(X) \tag{9}$$
$$X_W = F_W(X) \tag{10}$$
$$X_S = F_S(X) \tag{11}$$
$$\hat{X} = F(X_H + X_W + X_S) \tag{12}$$

where F_H and F_W are the depthwise convolution layers, and F_S denotes the pointwise convolution operation. $F(\cdot)$ denotes a fully connected layer with weights $W_P \in R^{C\times C}$.

Finally, we employ a skip connection [20] to avoid the vanishing gradient problem.

1) Depthwise and Pointwise Convolutions: Our proposed ConvPermute is built with depthwise convolution and pointwise convolution. While the standard convolution operation [as shown in Fig. 4(a)] simultaneously presents the channelwise and spatialwise computation, depthwise convolution is a spatial feature learning operation that applies a single convolution filter for each input channel and the pointwise convolution is a channel feature learning operation that combines the channels. Supposing an input feature map F_M of size $D_H \times D_W \times D_S$ and a standard convolution filter K of size $h \times w \times m \times n$, then the output feature map O_O can be acquired

$$O_O = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \tag{13}$$

where $i \in h$ and $j \in w$.

Different from the standard convolution, the output feature map $\acute{O}_O$ obtained with the depthwise convolution ($\acute{K}$ is the filter) [as shown in Fig. 4(b)] can be calculated

$$\acute{O}_O = \sum_{i,j} \acute{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m}. \tag{14}$$

When applying the pointwise convolution [as shown in Fig. 4(c), and its size 1×1] on the input feature map F_M, the output feature map $\dot{O}_O$ can be formulated

$$\dot{O}_O = \sum_{m} \dot{K}_{m,n} \cdot F_{k-1,l-1,m}. \tag{15}$$

Since the spectral information is embedded in the channel dimension, the pointwise convolution layer is used for capturing the spectral information.

In summary, the depthwise and pointwise convolutions are used for capturing the spatial correlations and the spectral correlations, respectively. It is worth noting that the sizes of depthwise convolution layers F_H and F_W are set to 3×1 and 1×3 in this article.

IV. Experiment

In this section, we first depict four benchmark HSI datasets, including the Indian Pines dataset, Pavia University dataset, Houston2013 dataset, and Xiongan (XA) dataset. Second, we will introduce the experimental setup, including the evaluation metrics used for this article, state-of-the-art backbone methods (such as AE-based methods, RNN-based methods, CNN-based methods, and transformer-based methods), and the implementation details. Finally, we conduct extensive experiments and ablation studies to evaluate the performance of the proposed HiT.

A. Datasets Description

1) Indian Pines Dataset: This hyperspectral dataset recording remote sensing images over North-Western Indiana, USA, was obtained in 1992 using the Airborne Visible Imaging

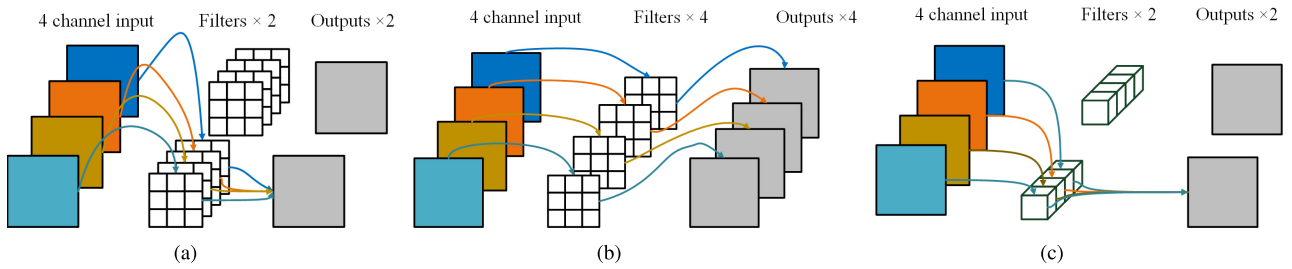


Fig. 4. Different convolution operations, including standard, depthwise, and pointwise convolution operations. It is worth noting that we show four input channels, two outputs for standard and pointwise convolution layers, and four outputs for depthwise convolution layers. The kernel size is 3 × 3. (a) Standard convolution. (b) Depthwise convolution. (c) Pointwise convolution.

Spectrometer (AVIRIS) sensor. The HSI consists of 145 × 145 pixels in the spatial dimension, and 220 bands in the spectral dimension. There are 200 spectral bands retained after removing 20 noisy bands. There are 16 categories in this dataset, including alfalfa, corn, and woods. We use 10% training samples and the other 90% samples are testing samples.

2) Pavia University Dataset: This dataset was collected using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor at Pavia University, Pavia, Italy. The Pavia University image consists of 610 × 340 in the spatial dimension, and 103 bands in the spectral dimension. This dataset contains nine land cover classes, including asphalt, gravel, and trees. Only 10% samples are set as the training samples, and 90% samples are testing samples.

3) Houston2013 Dataset: This hyperspectral dataset was captured by the innovative teaching methods for tomorrows' renewable specialists (ITRES) CASI-1500 sensor over the University of Houston and its surroundings in Texas, USA. It has been widely used for assessing the performance of the land cover classification [59]. There are 349 × 1905 pixels and 144 bands in the spatial and spectral dimensions of the Houston2013 images, respectively. It is noted that the Houston2013 dataset used in this article is a cloud-free version provided by the Geo-Science and Remote Sensing Society (GRSS) data fusion competition. There are 15 categories, including highway, road, and trees. The samples are divided into 10% training samples and 90% testing samples.

4) Xiongan Dataset: This is a new hyperspectral dataset, which was collected using the visible and near-infrared imaging spectrometer over the XA County and its neighboring Baiyangdian Lake areas in China, and provided by the Chinese Academy of Sciences in October 2017 [60]. The sensor can capture 250 spectral bands varying from 400 to 1000 nm, and the image size consists of 3750 × 1580 pixels. These data contain 20 land cover classes, including willow, rice, and corn. Only 1% samples are set for training and the other 99% samples are used for testing.

B. Experimental Setup

1) Evaluation Metrics: We evaluate the classification performance of all methods by using two widely used metrics, i.e., overall accuracy (OA) and kappa coefficient (κ).

2) Comparison With State-of-the-Art Backbone Methods: To conduct the following comparison experiments, we select several representative baselines and state-of-the-art backbone methods, including AEs (i.e., Boulch *et al.* [23]), RNNs (i.e., Mou *et al.* [27]), CNN-based methods (i.e., 2-D-CNN [24], R-2D-CNN [24], 3-D-CNN [24], and He *et al.* [37]), and transformer-based methods (i.e., ViT [47], Deep ViT [48], LeViT [32], and RvT [31]). These comparison methods are designed as follows.

1) For Boulch *et al.* [23], an AE-based method, the encoder layers are equal to the spectral bands of the input image. Each layer consists of one 1-D convolution operation followed by 1-D-Maxpool, BN [25], and ReLU [26] activation function.
2) For Mou *et al.* [27], an RNN-based method, there is one recurrent layer with the gated recurrent unit. Each layer has 64 neuron units.
3) The 2-D-CNN [24] comprises three 2-D convolution blocks and a softmax layer. Each 2-D convolution block consists of a 2-D convolution layer, a BN layer, a ReLU activate function, and an avg-pooling layer.
4) For R-2D-CNN [24], there are two CNNs with three 2-D convolution blocks. Each CNN has equal parameters to the 2-D-CNN.
5) The 3-D-CNN [24] consists of three 3-D convolution blocks and a softmax layer. Each 3-D convolution block comprises a 3-D convolution layer, a BN layer, a ReLU activation function, and a 3-D convolution layer with step size 2.
6) For He *et al.* [37], another 3-D-CNN-based method, different from the 3-D-CNN, includes a total of ten 3-D convolution layers and a softmax layer.
7) For ViT, a classical transformer-based method, following the ViT architecture [47], it contains a linear-projection component and transformer encoders.
8) The Deep ViT [48] is a ViT-based method by adding a linear layer following the self-attention modules.
9) The T2T [49] method uses a soft-split operation instead of a hard-split operation to improve the tokenization in ViT. As a result, the T2T can capture the finer local structure. This method consists of a patch embedding block without class token and a transformer block with the class token.
10) LeViT [32], a transformer-based method, includes four convolution embedding layers instead of linear projection. Similar to ViT, we also follow the LeViT archi-

TABLE II
CLASSIFICATION RESULTS OF THE INDIAN PINES DATASET

Class No.	AEs	RNNs	CNNs				Transformers					
	Boulch [23]	Mou [27]	2D-CNN [24]	3D-CNN [24]	He [37]	R-2D-CNN [24]	ViT [47]	Deep ViT [48]	T2T [49]	LeViT [32]	RvT [31]	HiT (Ours)
1	0.00	69.88	96.30	52.63	74.63	76.47	48.48	57.63	24.49	68.42	67.44	94.25
2	35.67	65.56	88.70	75.07	66.69	90.16	66.69	72.80	68.37	68.15	74.99	92.68
3	29.96	56.63	78.84	57.02	59.53	80.09	51.60	60.47	54.43	59.75	74.51	78.55
4	0.00	53.40	91.32	41.81	47.06	74.64	56.66	67.27	70.68	72.73	85.30	86.73
5	12.93	86.65	87.52	83.88	75.21	87.39	47.81	55.14	46.34	44.07	59.03	85.53
6	54.99	94.24	98.85	97.45	92.64	99.47	77.00	91.85	91.55	85.80	86.10	98.32
7	0.00	47.06	74.42	42.42	42.42	89.36	40.00	88.46	26.67	14.63	55.56	92.00
8	75.02	95.30	94.39	90.09	92.78	94.03	87.32	93.20	88.74	89.33	92.91	94.63
9	0.00	26.09	97.14	36.36	75.86	100	28.57	80.00	85.71	44.44	22.22	64.86
10	52.18	54.90	78.22	75.92	72.78	89.74	59.29	73.55	75.68	68.13	78.27	89.48
11	64.04	72.33	93.43	82.57	81.22	89.17	69.54	80.90	75.31	72.09	81.09	94.40
12	4.47	66.72	89.76	72.12	49.04	91.00	52.36	73.34	65.81	52.53	67.25	89.32
13	5.15	90.45	100	94.15	98.65	99.19	91.27	92.78	93.78	79.37	82.64	99.46
14	82.72	91.90	98.22	94.81	95.18	98.08	88.57	91.12	90.05	89.82	91.74	97.23
15	29.82	59.28	68.85	53.52	50.31	69.95	37.76	60.99	50.40	43.97	51.27	68.71
16	0.00	85.39	90.48	66.67	66.14	93.83	44.26	83.54	70.34	78.01	98.18	91.67
OA (%)	51.26	73.98	86.29	76.57	73.57	86.20	64.89	75.07	71.17	68.78	75.89	**87.54**
κ (%)	42.52	70.08	84.48	73.33	69.84	84.47	60.11	71.65	67.29	64.47	72.72	**85.93**

TABLE III
CLASSIFICATION RESULTS OF THE XA DATASET

Class No.	AEs	RNNs	CNNs				Transformers					
	Boulch [23]	Mou [27]	2D-CNN [24]	3D-CNN [24]	He [37]	R-2D-CNN [24]	ViT [47]	Deep ViT [48]	T2T [49]	LeViT [32]	RvT [31]	HiT (Ours)
1	64.10	68.68	96.17	82.45	84.78	96.30	78.56	89.21	34.24	87.75	79.30	96.92
2	45.27	80.53	98.68	89.27	86.31	98.80	81.38	92.16	46.03	90.26	85.22	99.12
3	0.00	72.55	98.56	72.28	85.47	99.00	87.07	92.80	52.30	90.71	86.85	98.69
4	94.75	97.89	99.26	97.44	98.53	99.20	98.24	98.81	92.78	98.41	98.62	99.33
5	71.06	75.34	98.20	88.97	86.05	80.24	83.96	91.96	40.68	91.47	86.55	98.22
6	9.69	82.64	98.86	88.10	87.79	98.50	80.27	91.32	38.93	89.98	76.99	98.98
7	0.00	87.95	99.23	96.47	97.49	98.20	91.73	97.85	7.01	96.92	74.31	99.46
8	92.23	95.65	99.20	96.67	97.31	99.20	97.04	98.22	89.34	98.42	96.32	99.42
9	92.67	95.50	99.61	97.20	95.47	98.70	97.95	98.66	95.45	98.24	98.54	99.68
10	93.81	95.88	99.17	97.80	98.16	99.80	98.55	99.08	84.04	98.82	98.68	99.84
11	0.00	8.45	93.23	8.54	19.75	87.80	21.06	67.82	0.20	43.87	31.97	91.62
12	0.00	58.65	95.45	75.28	74.50	95.50	69.98	83.27	52.50	83.26	76.97	94.40
13	75.12	84.59	97.08	89.17	87.61	97.50	86.33	92.18	70.14	91.91	90.13	97.81
14	0.00	31.90	84.64	53.89	63.67	77.00	55.92	80.24	5.17	72.80	68.23	90.57
15	0.00	67.22	93.34	74.77	77.26	92.50	79.52	88.48	9.09	88.11	78.74	94.71
16	0.00	38.11	81.88	44.84	40.35	7.00	47.68	67.74	23.71	62.93	56.05	82.80
17	0.00	2.40	38.26	8.04	0.13	29.00	7.60	36.71	0.00	25.33	12.58	59.86
18	71.29	76.49	94.95	81.78	85.36	96.90	82.66	90.32	56.34	90.23	83.76	97.19
19	0.00	52.34	97.28	75.81	67.62	97.30	73.70	86.84	45.92	87.05	83.19	97.15
20	0.00	88.48	95.89	90.90	89.77	96.90	90.32	94.10	86.32	92.95	92.62	96.35
OA (%)	71.14	82.16	97.11	88.23	88.07	97.38	86.17	92.50	62.00	91.96	88.03	**97.73**
κ (%)	65.57	79.36	96.64	86.35	86.10	96.90	83.92	91.30	55.25	90.68	86.12	**97.37**

tecture containing four convolution embedding layers, three stages comprised four multihead attention layers, and four MLP layers, respectively.

11) The RvT [31] consists of position-aware attention scaling and patchwise augmentation technologies. In detail, we follow the RvT architecture to design the network in the HSI classification task.
12) For the proposed HiT, we adopt two SACP layers to build an SACP module to embed the input images. Then, the 3-D token representations are separately processed along the height, width, and spectral dimensions, respectively. In detail, two depthwise convolution layers are used to process the height and width information and a pointwise convolution layer for processing spectral information. Finally, a global average pooling layer and linear layer are attached to predict the category.

3) Implementation Details: The proposed HiT and the compared methods were implemented on the PyTorch platform using a desktop PC with an Intel Core 7 Duo CPU (at 3.40 GHz), 64 GB of RAM, and one GTX R3090 GPU (24 GB of ROM). We adopt the Adam optimizer with batch size 100 and the learning rate 1e-3. It is noted that the transformer-based methods could not perform excellent results if the initialization of the learning rate is not 1e-3. We set the epochs on these four benchmark datasets to 100.

C. Results and Analysis

We conduct extensive experiments on four benchmark HSI datasets in terms of two well-known metrics, i.e., OA and κ. In detail, the results of these HSI datasets [e.g., Indian Pines, Pavia University (PaviaU), Houston2013, and XA datasets] are listed in Tables II–V, respectively.

According to the results, the AE-based methods (e.g., Boulch *et al.* [23]) achieve the worst classification results on all four benchmark datasets. Specifically, Boulch *et al.* [23] only achieves 51.26% and 42.52% in terms of OA and κ on the Indian Pines dataset. Owing to the powerful learning ability of the spectral sequential dimension, the RNN-based methods (e.g., Mou *et al.* [27]) produce better results than Boulch *et al.*'s [23] method. For example, 73.98% versus 51.26% in terms of OA on the Indian Pines dataset. Since CNNs have the powerful ability to capture the local spatial context information, the CNN-based methods,

TABLE IV
CLASSIFICATION RESULTS OF THE HOUSTON2013 DATASET

Class No.	AEs	RNNs	CNNs				Transformers					
	Boulch [23]	Mou [27]	2D-CNN [24]	3D-CNN [24]	He [37]	R-2D-CNN [24]	ViT [47]	Deep ViT [48]	T2T [49]	LeViT [32]	RvT [31]	HiT (Ours)
1	93.74	94.47	92.96	96.84	92.11	98.04	91.89	86.70	95.50	97.60	90.30	98.07
2	92.35	95.06	91.12	95.14	97.39	98.23	92.91	85.90	96.10	97.64	90.40	98.93
3	98.06	100	99.76	97.89	98.31	100	91.43	94.80	98.70	98.09	99.90	100
4	76.46	95.26	95.01	95.98	92.81	98.62	91.75	94.40	96.70	97.79	96.00	97.82
5	93.48	98.14	98.46	93.95	95.39	98.32	89.45	94.80	96.10	96.32	97.90	97.87
6	0.00	96.65	92.56	74.81	67.87	91.27	63.72	87.10	93.80	94.53	83.30	91.83
7	52.85	60.51	93.70	87.78	90.17	95.03	87.56	93.30	93.00	94.81	88.20	96.13
8	67.59	65.71	76.65	78.43	80.55	94.41	74.10	63.70	89.10	93.03	73.60	94.82
9	61.88	71.44	91.10	83.90	83.70	92.55	78.15	86.80	90.00	91.56	87.10	93.58
10	71.91	29.95	82.40	86.86	84.41	94.70	71.72	80.10	92.80	91.36	73.10	96.55
11	62.47	46.55	93.74	87.40	82.37	95.58	69.29	71.30	93.10	93.31	78.90	96.11
12	35.34	61.60	81.82	80.87	77.06	94.78	61.57	72.10	95.50	88.27	81.70	97.09
13	0.00	49.60	95.66	85.48	85.71	90.85	33.15	56.40	96.10	82.37	83.90	91.39
14	0.00	94.95	97.81	95.33	93.21	95.19	82.40	89.90	98.70	97.26	95.70	99.74
15	97.85	98.25	97.38	97.31	93.69	99.25	78.44	92.50	96.70	94.06	97.90	99.17
OA (%)	71.73	76.22	90.52	89.01	87.70	95.63	79.38	83.11	96.10	93.73	86.95	**96.35**
κ (%)	69.29	74.27	89.76	88.13	86.70	95.28	77.72	81.80	93.80	93.23	85.90	**96.06**

TABLE V
CLASSIFICATION RESULTS OF THE PAVIA UNIVERSITY DATASET

Class No.	AEs	RNNs	CNNs				Transformers					
	Boulch [23]	Mou [27]	2D-CNN [24]	3D-CNN [24]	He [37]	R-2D-CNN [24]	ViT [47]	Deep ViT [48]	T2T [49]	LeViT [32]	RvT [31]	HiT (Ours)
1	91.43	50.50	96.49	94.30	94.90	94.80	90.20	94.80	94.00	93.62	94.62	96.19
2	96.18	94.50	92.71	92.70	91.90	92.71	89.99	92.05	92.00	91.42	92.05	92.79
3	11.13	49.70	88.38	87.00	86.80	92.50	71.59	87.17	89.00	84.27	88.32	93.21
4	90.68	94.55	97.24	96.90	97.00	97.60	96.06	96.36	96.00	96.98	97.50	97.33
5	85.94	99.92	99.85	99.60	99.90	99.96	99.42	99.93	99.00	99.93	99.96	99.96
6	89.22	83.89	100.00	99.80	96.20	99.95	90.32	97.88	98.50	95.97	97.41	99.91
7	33.15	43.41	99.13	92.80	93.30	93.30	81.23	94.76	95.00	93.01	96.93	98.22
8	73.58	70.99	97.06	94.60	95.80	98.53	90.02	96.79	97.00	95.30	96.31	99.15
9	99.47	99.30	98.99	98.90	99.50	99.95	98.01	98.04	98.00	99.79	99.84	99.77
OA (%)	88.04	81.14	91.63	90.73	90.23	91.54	86.41	90.54	90.90	89.68	90.65	**92.00**
κ (%)	83.92	75.30	89.30	88.10	87.50	89.19	82.60	87.90	88.00	86.83	88.05	**89.77**

such as 2-D-CNN [24], R-2D-CNN [24], 3-D-CNN [24], and He *et al.* [37], observably perform excellent classification results. This also demonstrates the value and practicality of convolution operations in HSI classification. Without any convolution layers, transformer-based methods, e.g., ViT [47] and Deep ViT [48], are capable of capturing subtle spectral information and producing a competitive classification performance compared to the state-of-the-art CNN-based methods. Capturing the local spatial context information by using 2-D convolution operations, the LeViT [32] and RvT [31] outperform ViT [47] and Deep ViT [48] and achieve better classification results than CNN-based methods, such as the classification performances of LeViT [32] on Houston2013 and XA datasets.

On the other hand, the existing transformer-based methods have limitations in acquiring the local spectral discrepancies, leading to a performance bottleneck. We argue that the main factor leading to the improvement of our proposed HiT is the ways of image embedding and encoding spatial–spectral information. Different from current existing transformer-based methods, we adaptively embed the input images using an SACP module, which consists of two spectral-adaptive 3-D convolution layers to extract the spatial location-sensitive importance map and adaptive capture the long-term spectral information. We further separately encode the embedded representations along the height, width, and spectral dimensions, resulting in position-sensitive features. According to the results, the proposed HiT undoubtedly outperforms other transformer-based methods and state-of-the-art CNN-based methods. For example, the performance of HiT on Houston2013 dataset is 96.35%, which is better than R-2D-CNN [24] (96.35% versus 95.63%). Compared with some transformer-based methods, such as T2T [49] (96.10%), LeViT [32] (93.73%), Deep ViT [48] (83.11%), and RvT [31] (86.95%), our results are also better.

We utilize the visualization results for qualitative evaluation of the performance of all methods. Figs. 5–8 report the achieved classification result maps of Indian Pines, PaviaU, Houston2013, and Xiaongan datasets, respectively. These qualitative visualization results provide a rough finding that all methods achieve relatively smooth classification maps. Moreover, owing to the powerful local context information extraction ability of CNNs, their classification maps usually have less noise. On the other hand, the classification maps of the 3-D-CNN-based methods (e.g., 3-D-CNN and He) often have many pixels with incorrect recolonization. This is because a large number of training samples are urgently needed in training 3-D-CNNs. For transformer-based methods, they have the capability of extracting sequential representations from HSIs, producing classification maps approached to those of CNN-based methods. The proposed HiT performs satisfactory

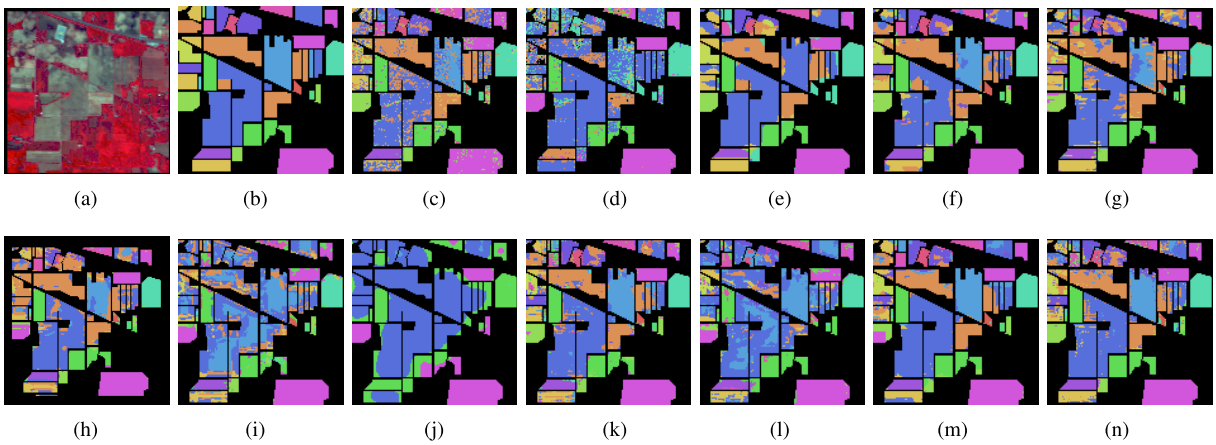


Fig. 5. Classification maps obtained by different methods on the Indian Pines dataset (with 10% training samples). (a) Input image. (b) Ground truth. (c) Mou. (d) Boulch. (e) 2-D-CNN. (f) R-2D-CNN. (g) 3-D-CNN. (h) He. (i) ViT. (j) Deep ViT. (k) T2T. (l) LeViT. (m) RvT. (n) HiT (Ours).

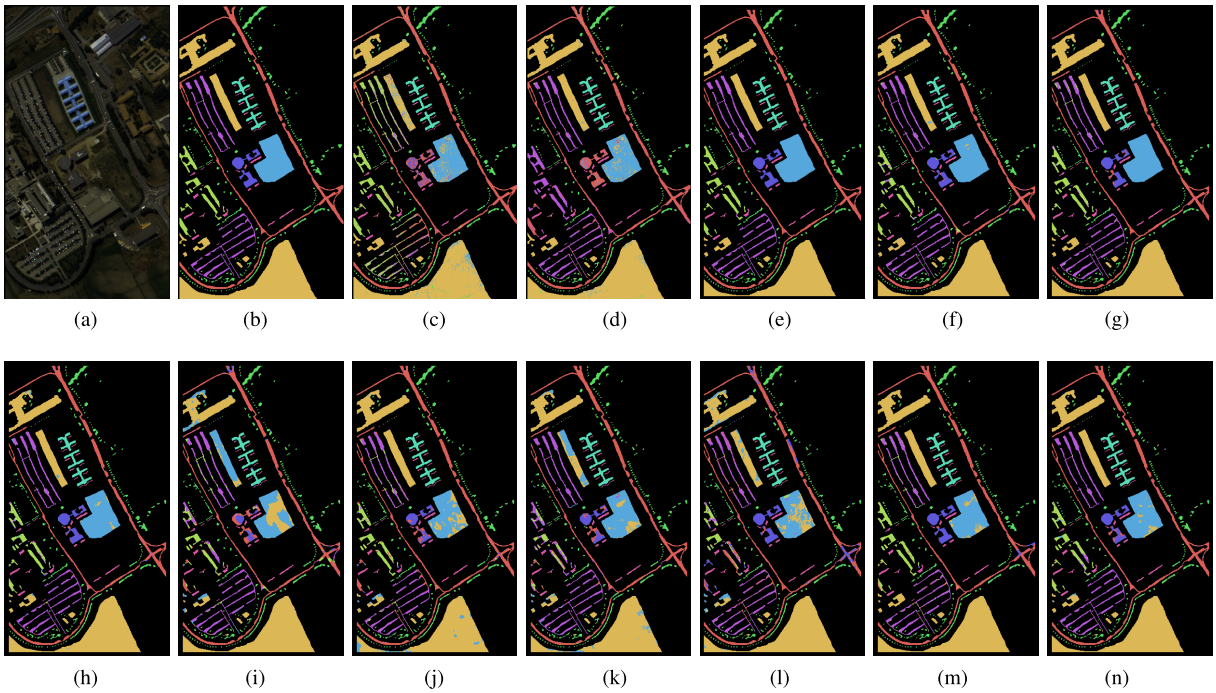


Fig. 6. Classification maps obtained by different methods on the Pavia University dataset (with 10% training samples). (a) Input image. (b) Ground truth. (c) Mou. (d) Boulch. (e) 2-D-CNN. (f) R-2D-CNN. (g) 3-D-CNN. (h) He. (i) ViT. (j) Deep ViT. (k) T2T. (l) LeViT. (m) RvT. (n) HiT (Ours).

classification results, attributing to utilizing an SACP module for enhancing the local spatial–spectral information and a Conv-Permutator for separately encoding spatial–spectral representations along the height, width, and spectral dimensions, respectively. It is worth noting that Figs. 7 and 8 only visualize the results of some areas from Houston2013 and XA datasets, and the specific visualization results are shown in the supplementary materials.

Finally, we evaluate the efficiency of various methods and present the results in Table VI. It is worth noting that the size of the input image is $1 \times 200 \times 15 \times 15$, and "bimg" denotes 100 images per batch. The term "param" denotes the parameters, and the term "Tp" is the shorthand for Throughput. We can observe that the proposed HiT performs admirably, but its efficiency in terms of speed is not optimal. This is also the limitation of the proposed HiT, which has a large number of parameters and does not perform well. Compared with the CNN-based methods, most of the transformer-based methods require lots of time to learn a satisfactory feature representation (such as 88.01 s versus 479.91 s in training time). This is mainly because most of the transformer-based methods need many repeated self-attention modules to learn abundant feature representation, leading to much larger sizes of transformer-based methods than those of CNN-based methods. When compared with the other transformer-based methods except RvT, our HiT effectively

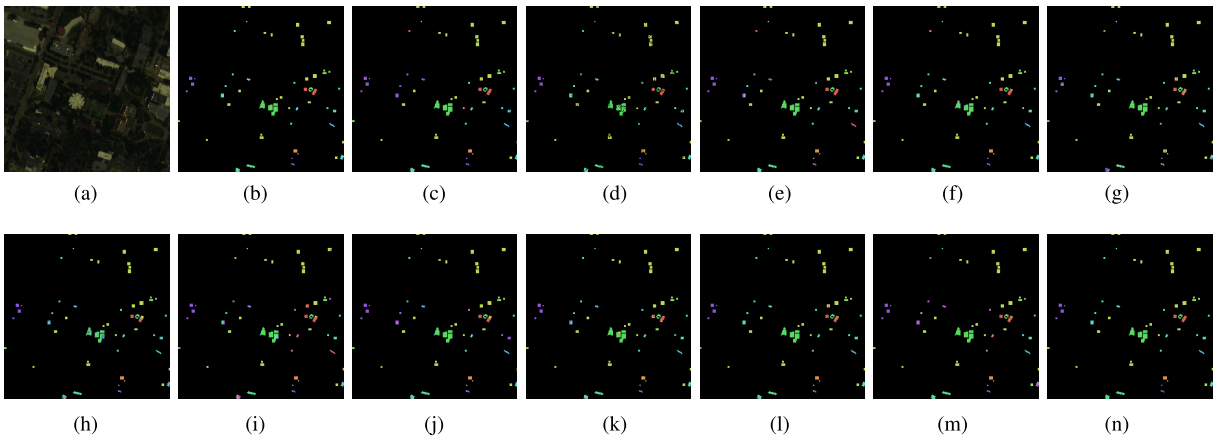


Fig. 7. Classification maps obtained by different methods on some areas of the Houston2013 dataset (with 10% training samples). (a) Input image. (b) Ground truth. (c) Mou. (d) Boulch. (e) 2-D-CNN. (f) R-2D-CNN. (g) 3-D-CNN. (h) He. (i) ViT. (j) Deep ViT. (k) T2T. (l) LeViT. (m) RvT. (n) HiT (Ours).

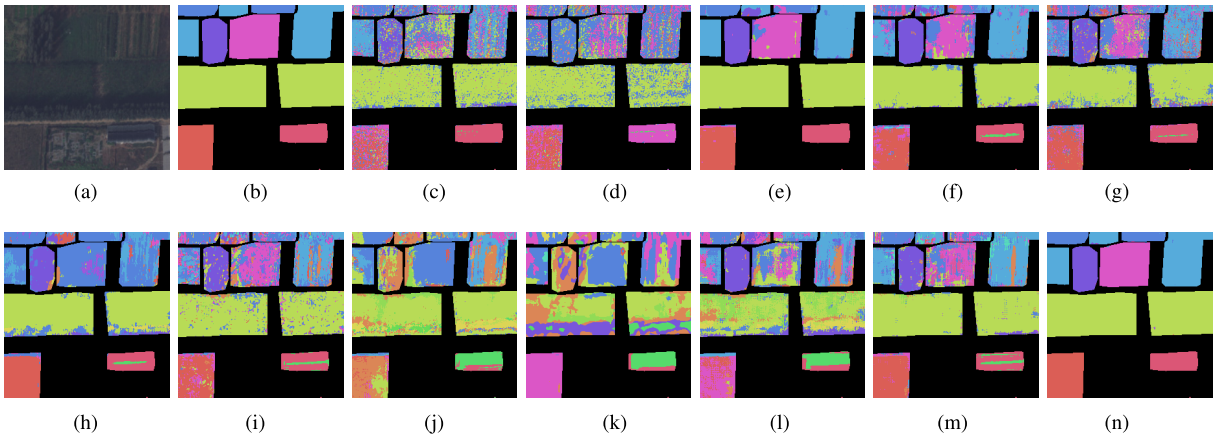


Fig. 8. Classification maps obtained by different methods on some areas of the XA dataset (with 1% training samples). (a) Input image. (b) Ground truth. (c) Mou. (d) Boulch. (e) 2-D-CNN. (f) R-2D-CNN. (g) 3-D-CNN. (h) He. (i) ViT. (j) Deep ViT. (k) T2T. (l) LeViT. (m) RvT. (n) HiT (Ours).

TABLE VI
COMPUTATIONAL COMPLEXITY OF ALL METHODS

Methods	Flops (GB)	Param (MB)	Tp (bimg/s)	Training time (s)	Testing time (s)
2D-CNN [24]	0.07	0.49	436	15.92	1.21
3D-CNN [24]	0.27	1.46	111	88.01	3.53
R-2D-CNN [24]	3.88	45.82	16	31.60	1.91
ViT [47]	0.14	2.60	115	62.87	2.09
Deep ViT [48]	2.71	52.21	31	110.31	6.60
LeViT [32]	1.81	16.94	13	148.56	7.06
RvT [31]	0.42	8.93	45	67.56	3.67
T2T [49]	5.95	730.18	8	479.91	6.79
HiT (Ours)	2.33	51.18	20	112.04	6.70

reduces both the training and testing times while producing the best performances (Tables II–V) on all datasets, e.g., compared with T2T, training time 112.04 s versus 479.91 s, and testing time 6.70 s versus 6.79 s. Specifically, compared with RvT (transformer-based method), the proposed HiT requires more training and testing times. Our HiT, however, achieves much better performances (Tables II–V) on all datasets. Especially, on the Houston2013 dataset, our HiT method improves the OA and κ by 97.73% and 97.37%, respectively. All aspects of comparisons demonstrate the best performances and satisfactory efficiency superiority of our HiT.

D. Ablation Studies

1) Ablation Study of the Proposed SACP Module: We first show that the SACP is important for the transformer-based HSI classification method. To demonstrate this argument, we adjust the image projection step by employing different projection modules (e.g., linear, Conv2D, Conv3D, and SACP) and keep the Conv-Permutators unchanged. We report the performance for different HiTs in Table VII. It is worth noting that the input size is $1 \times 200 \times 15 \times 15$ and the ViT [47] is chosen as the baseline method. According to the results,

TABLE VII
ABLATION STUDY OF THE PROPOSED SACP ON THE PAVIAU DATASET

Methods	Flops (GB)	Parameters (MB)	OA(%)
ViT [47]	0.14	2.60	86.41 **(-5.59%)**
HiT-linear	0.72	50.23	90.47 **(-1.53%)**
HiT-Conv2D	2.24	50.23	90.74 **(-1.26%)**
HiT-Conv3D	2.17	49.57	91.96 **(-0.04%)**
HiT-SACP	2.37	51.18	92.00

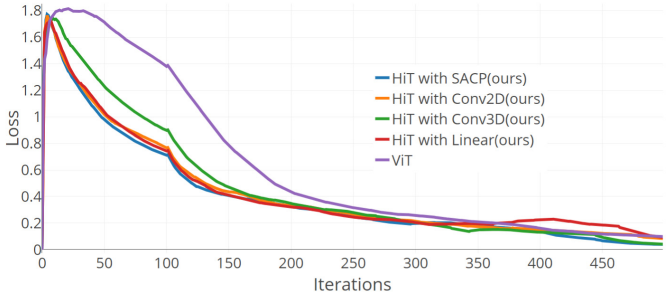


Fig. 9. Training loss of five different methods based on the PaviaU dataset.

HiT-SACP shows the best performance (92.00% in OA). Despite more parameters used in HiT-SACP, its efficiency still is the fastest (2.37 GB in Flops). This demonstrates that we can appropriately use more SACP to embed the input images to improve the performance. We further report the training loss of these five methods in Fig. 9. As shown in Fig. 9, such HiT with 3-D convolution projection modules (such as Conv3D and SACP) could improve the performance for HiT. This demonstrates that embedding much more spectral–spatial fusion representations does help in improving the proposed HiT performance but a drawback is that the parameters go increase a little.

Class activate map (CAM) [61] uses the principle of feature map weight overlapping to obtain the heatmaps, which helps us to have a good explanation of the network. We adopt CAM to explain why utilizing 3-D convolution projection module is necessary for the transformer-based networks in HSI classification tasks. We visualize the CAM of the proposed HiT (with different projection modules) in Fig. 10. It is worth noting that the darker the color is, the more attention the model focuses. We can observe that the network with SACP and Conv3D projection modules not only pays attention to the global region but also the local region closer to the central pixel. These attribute to the transformer-based architecture that could capture the global information, and the 3-D convolution projection module which could provide local spectral–spatial fusion information. Since the 2-D convolution projection module can extract the local spatial context information, the network with the Conv2D projection module is capable of focusing on the local region surrounding the central pixel. According to Fig. 10, the linear could offer sequence information to help the network for capturing the global region, however, resulting in the output features without the local information.

2) Ablation Study of the Proposed Modules: This ablation analysis will examine the effectiveness of these components in the proposed HiT, including SACP, height and width depthwise convolution operations, and channel-MLP. We conduct extensive ablation experiments on the XA dataset by adding different components. We summarize the results under different components in Table VIII. It is worth noting that there are some abbreviations in Table VIII, for example, SACP is the abbreviation of spectral-adaptive 3D convolution projection module. Height depthwise convolution (HDWC), width depthwise convolution (WDWC), and channel master limited partnership (CMLP) denote the height depthwise convolution module, WDWC, and channel-MLP, respectively.

TABLE VIII
ABLATION STUDY OF THE PROPOSED HiT WITH A COMBINATION OF DIFFERENT COMPONENTS ON THE XA DATASET

Methods	Components				Metric
	SACP	HDWC	WDWC	CMLP	OA (%)
ViT	×	×	×	×	86.17 **(-11.56%)**
ViT	√	×	×	×	88.10 **(-9.63%)**
HiT (Ours)	√	√	×	×	96.71 **(-1.02%)**
HiT (Ours)	√	×	√	×	96.16 **(-1.57%)**
HiT (Ours)	√	√	√	×	96.95 **(-0.78%)**
HiT (Ours)	√	√	√	√	**97.73**

It is clear that the classical ViT [47] without using any components performs the lowest classification results. This indicates that the classical ViT may not be applied to HSI classification tasks. However, the ViT outperforms the classical ViT when introducing the SACP in ViT to capture the local spectral information. In addition, we can see that joining either height information encoder or width information encoder leads to a better performance than the classic ViT (96.71% or 96.16% versus 86.17%). This demonstrates that local spatial information plays a crucial role in the HSI classification tasks. We can also observe that the joint exploitation of HDWC and WDWC can further improve the performance from 96.71% to 96.95%. By adding the CMLP module, the proposed HiT finally achieves the best classification accuracy of 97.73%.

3) Ablation Study of the Model Scale: Making the models deeper and wider for deep learning methods is always an effective way to improve the performance. Thereby, we learn the influence of the model scaling on the proposed HiT by increasing the number of layers and hidden dimension (Dim). We present the results of four different HiT with various layers and Dim in Table IX, including HiT-small/18, HiT-small/18 (with 512 Dim), HiT-medium/24, and HiT-large/36. We can observe that increasing the number of layers and Dim could lead to a better performance of the proposed HiT. HiT-small/18 (with 512 Dim) could improve the performance from 90.79% (obtained by HiT-small/18 (with 208 Dim) to 91.01%, but the parameters ("param" for short) only increase 6.65 MB (35.87 MB versus 29.22 MB). Further increasing the number of layers will result in achieving the better performance of 91.60% (obtained by HiT-larger/36).

4) Ablation Study of the Percentage of Training Samples: According to [47] and [48], transformer-based methods urgently require a large number of training samples. It is interesting to study that how the training samples affect the performance of the proposed HiT. We conduct extensive

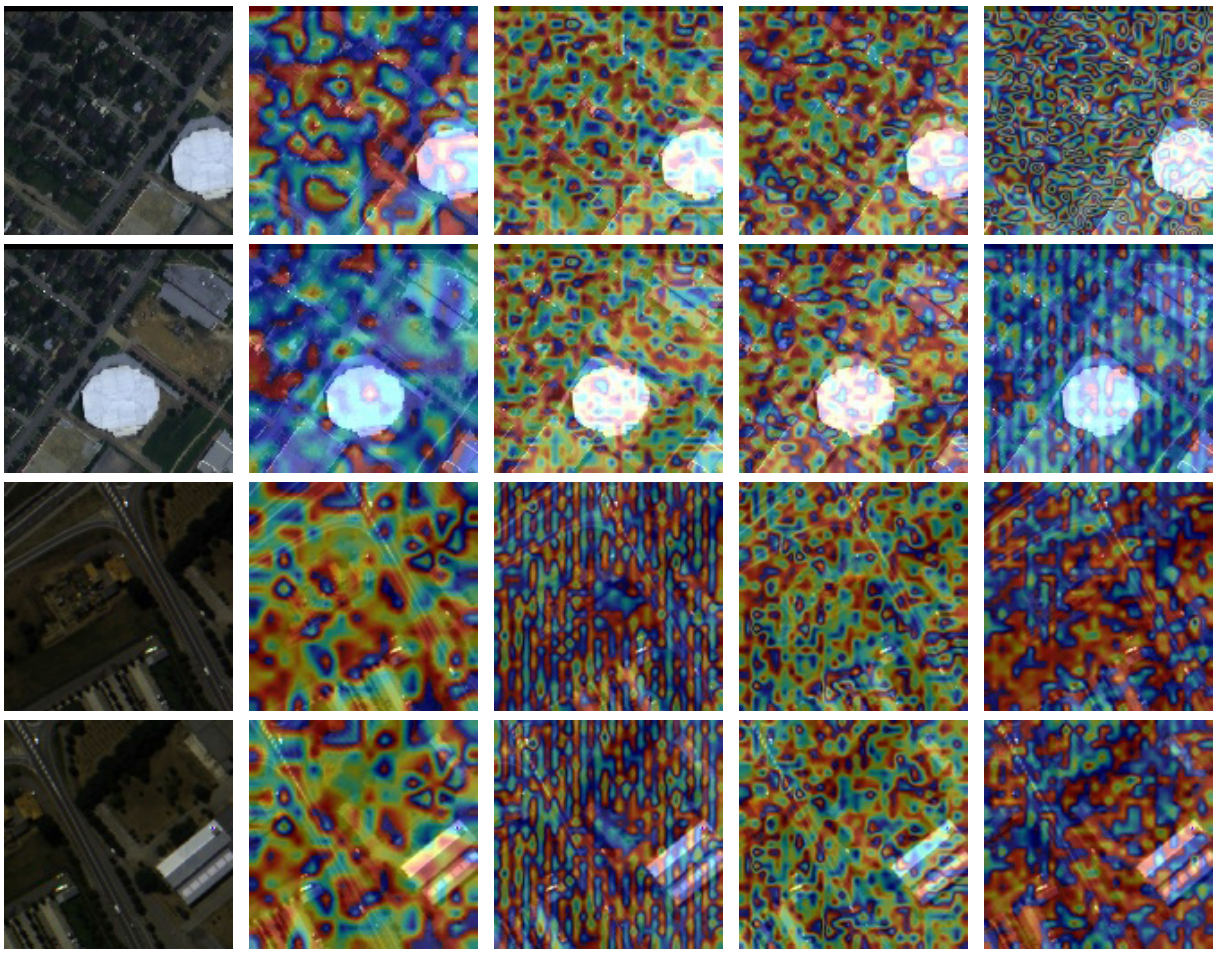


Fig. 10. Visualization of the CAM on four different input images. The first column denotes the input images; the second, third, and fourth columns are the CAMs of linear, Conv2D, Conv3D, and SACP (ours), respectively. It can be seen that the linear projection pays attention on a sequence of vectors, and the Conv2D focuses on the local spatial. For Conv3D and SACP, they focus on the spatial–spectral fusion information. It is worth noting that the darker the color, the more attention the model focus. (a) Inputs. (b) Linear. (c) Conv2D. (d) Conv3D. (e) SACP.

TABLE IX

ABLATION STUDY OF THE MODEL SCALE ON THE PAVIAU DATASET (10% TRAINING SAMPLES)

Methods	Layers	Dim	Flops (GB)	Param(MB)	OA(%)
HiT-small/18	18	208	1.09	29.22	90.79
HiT-small/18	18	512	1.80	35.87	91.01
HiT-medium/24	24	512	2.21	50.07	91.46
HiT-large/36	36	512	3.93	70.57	91.60

experiments on four benchmark HSI datasets varying the training samples from 10% to 50% at intervals of 10%. We run the proposed HiT for ten times. Fig. 11 reports the average results of the OA achieved by the proposed HiT. We can see that the classification performance gradually improves with varying the percentage of training samples from 10% to 50%. The OA has been obviously improved, particularly when increasing the training samples from 10% to 20%. This demonstrates that the number of training samples also affects the performance of the proposed HiT. It is noted that the OAs are tending to stabilize when varying the training samples from 40% to 50%, proving the stability of the proposed HiT.

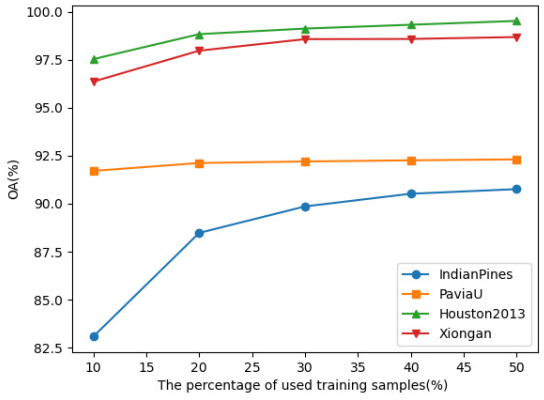


Fig. 11. Classification results (OA) achieved by the proposed HiT with a varying number of training samples on four benchmark datasets.

V. CONCLUSION

HSIs consist of spatial and spectral information that can be regarded not only as natural image data along the spatial dimension but also as a sequence of data. CNNs can perform

satisfactory results on HSI classification by extracting the local spatial context features. However, CNNs cannot mine the subtle spectral discrepancies. On the other hand, transformers have been proved that they have a powerful ability to capture global information from sequence data. Nevertheless, the existing transformer-based methods (e.g., ViT and Deep ViT) have a failure of capturing the subtle spectral discrepancies and fail in conveying the local spatial–spectral information from the shallow to deep layers. This article proposes a new transformer-based method, i.e., HiT, for HSI classification from a sequential perspective. The proposed HiT consists of two key modules for addressing the problems in HSI classification, i.e., an SACP module and a Conv-Permutator module. The SACP module captures the local spatial–spectral information (especially the subtle spectral discrepancies) using two spectral-adaptive 3-D convolution layers. The spectral-adaptive 3-D convolution layer consists of two branches: the local spatial branch to learn the spatial location-sensitive importance map and the global branch to adaptive capture the long-term spectral information. Conv-Permutators separately encode the spatial–spectral representations along the height, width, and spectral dimensions and convey the encoded spatial–spectral information to the next step. Extensive experiments on four benchmark HSI datasets demonstrate that the proposed HiT outperforms the other transformer-based methods and state-of-the-art CNN-based methods.

Our future work aims at integrating the advantages of CNNs and transformers, and improving the transformer architecture by introducing advanced techniques (e.g., transfer learning and self-supervised learning). Specifically, we will design a universal and lightweight transformer-based method that is more suitable for HSI classification. On this basis, we will establish a general HSI classification platform that integrates various state-of-the-art algorithms.

REFERENCES

[1] A. Lobo, "Image segmentation and discriminant analysis for the identification of land cover units in ecology," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 5, pp. 1136–1145, Sep. 1997.

[2] F. Tsai and W. D. Philpot, "A derivative-aided hyperspectral image analysis system for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 2, pp. 416–425, Feb. 2002.

[3] T. Zhan *et al.*, "TDSSC: A three-directions spectral–spatial convolution neural network for hyperspectral image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 377–388, 2020.

[4] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2018.

[5] B. Zhang, D. Wu, L. Zhang, Q. Jiao, and Q. Li, "Application of hyperspectral remote sensing for environment monitoring in mining areas," *Environ. Earth Sci.*, vol. 65, no. 3, pp. 649–658, Feb. 2012.

[6] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern trends in hyperspectral image analysis: A review," *IEEE Access*, vol. 6, pp. 14118–14129, 2018.

[7] Y. Wang, J. Peng, Q. Zhao, D. Meng, Y. Leung, and X.-L. Zhao, "Hyperspectral image restoration via total variation regularized low-rank tensor decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1227–1243, Apr. 2017.

[8] C. H. Lin, P. H. Tsai, K. H. Lai, and J. Y. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 232–241, Jan. 2012.

[9] M. Wang, Q. Wang, J. Chanussot, and D. Hong, "L_0-L_1 hybrid total variation regularization and its applications on hyperspectral image mixed noise removal and compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7695–7710, Sep. 2021.

[10] F. Luo, T. Guo, Z. Lin, J. Ren, and X. Zhou, "Semisupervised hypergraph discriminant learning for dimensionality reduction of hyperspectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4242–4256, 2020.

[11] F. Luo, Z. Zou, J. Liu, and Z. Lin, "Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[12] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[13] F. Luo, H. Huang, Z. Ma, and J. Liu, "Semisupervised sparse manifold discriminative analysis for feature extraction of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6197–6211, Oct. 2016.

[14] 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. 2018.

[15] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[16] 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, Aug. 2010.

[17] Y. Duan, H. Huang, and T. Wang, "Semisupervised feature extraction of hyperspectral image using nonlinear geodesic sparse hypergraphs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[18] Y. LeCun, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[19] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[21] T.-H. Chan, K. Jia, S. Gao, J. Lu, and Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[22] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 17, 2021, doi: 10.1109/TPAMI.2021.3059968.

[23] A. Boulch, N. Audebert, and D. Dubucq, "Autoencodeurs pour la visualisation d'images hyperspectrales," in *Proc. 25th Colloque Gretsi*, Juan-les-Pins, France, 2017, pp. 1–4.

[24] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[26] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Aistats*, vol. 15, 2011, p. 275.

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

[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[29] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2019.

[30] X. He, Y. Chen, and Z. Lin, "Spatial–spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, p. 498, Jan. 2021.

[31] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11936–11945.

[32] B. Graham *et al.*, "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12259–12269.

[33] A. B. Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[34] X. Yang *et al.*, "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 2033, Jun. 2020.

[35] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[36] J. Fan, T. Chen, and S. Lu, "Superpixel guided deep-sparse-representation learning for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3163–3173, Nov. 2018.

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

[38] 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.

[39] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[40] V. Sharma, A. Diba, T. Tuytelaars, and L. Van Gool, "Hyperspectral CNN for image classification & band selection, with application to face recognition," KU Leuven, ESAT, Leuven, Belgium, Tech. Rep. KUL/ESAT/PSI/1604, 2016.

[41] L. Ran, Y. Zhang, W. Wei, and T. Yang, "Bands sensitive convolutional network for hyperspectral image classification," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Aug. 2016, pp. 268–272.

[42] P. R. Lorenzo, L. Tulczyjew, M. Marcinkiewicz, and J. Nalepa, "Band selection from hyperspectral images using attention-based convolutional neural networks," 2018, *arXiv:1811.02667*.

[43] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, Feb. 2017.

[44] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[45] H. Lee and H. Kwon, "Contextual deep CNN based hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3322–3325.

[46] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, Jan. 2017.

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

[48] D. Zhou *et al.*, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[49] L. Yuan *et al.*, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 558–567.

[50] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jegou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 32–42.

[51] C.-F.-R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[52] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[53] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[54] X. Chu *et al.*, "Twins: Revisiting spatial attention design in vision transformers," 2021.

[55] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," 2021, *arXiv:2104.09864*.

[56] Z. Zhang, H. Zhang, L. Zhao, T. Chen, and T. Pfister, "Aggregating nested transformers," 2021, *arXiv:2105.12723*.

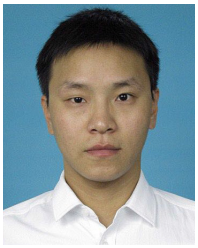
[57] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[58] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[59] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2020.

[60] Y. Cen *et al.*, "Aerial hyperspectral remote sensing classification dataset of Xiongan new area (Matiwan village)," *J. Remote Sens.*, vol. 24, no. 11, pp. 1299–1306, 2020.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.



Xiaofei Yang received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2014 and 2019, respectively.

He is currently a Post-Doctoral Researcher with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests are in the areas of semisupervised learning, deep learning, remote sensing, transfer learning, and graph mining.

Weijia Cao received the master's and Ph.D. degrees in computer science from the University of Macau, Macau, China, in 2013 and 2017, respectively.

She is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her main research interests revolve around multimedia encryption, machine learning, and remote sensing image processing.

Yao Lu received the B.S. degree in computer science and technology from Huaqiao University, Xiamen, China, in 2015, and the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Shenzhen, China, in 2020.

She was a Post-Doctoral Fellow with the University of Macau, Macau, China, from 2020 to 2021. She is currently an Assistant Professor with the Biocomputing Research Center, Harbin Institute of Technology. Her research interests include pattern recognition, deep learning, computer vision, and relevant applications.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA, in 2006 and 2010, respectively.

He is currently a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020 and 2021. He was a recipient of the Third Price of Macao Natural Science Award as a Sole Winner in 2020 and a co-recipient in 2014. He has been the Leading Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society since 2015. He serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.