A Center-Masked Transformer for Hyperspectral Image Classification

Sen Jia, *Senior Member, IEEE*, Yifan Wang, Shuguo Jiang, and Ruyan He

***Abstract*— Convolutional neural networks (CNNs) are widely used in hyperspectral image (HSI) classification. However, the fixed receptive field of CNN-based methods limits their capability to extract global features. In recent years, transformer has been introduced into networks to tackle this limitation, but it brings other challenges, including a significant increase in model size, the number of labeled training samples required, and the limited effectiveness of sample encoding-reconstruction pretraining methods for HSI classification. To address these issues, a center-masked transformer (CMT) approach is proposed to improve the HSI classification accuracy from two perspectives. On one hand, a local-to-global token embedding (L2GTE) framework coupled with a multiscale convolutional token embedding (MCTE) module is used, which is well-designed to obtain local and global embedding tokens. This effectively reduces the number of model parameters. On the other hand, a regularized center-masked pretraining (RCPT) task is proposed and first introduced into the transformer-based network, which enables the network to learn the dependencies between central ground objects and neighboring objects without labels during the pretraining process. The experimental results conducted on five public HSI datasets demonstrate that our CMT approach outperforms other state-of-the-art methods for HSI classification when training samples are insufficient.**

***Index Terms*— Convolutional transformer, deep learning (DL), hyperspectral image (HSI) classification, mask autoencoder.**

NOMENCLATURE

$\mathcal{I}$	Original HSI data.
$\mathcal{H}$	HSI data after dimensionality reduction through PCA.
$\mathcal{X}$	Input sample patch.
h, w	Height and width of the input sample patch.
$\mathcal{X}'$	Partitioning result of the input sample patch.
$\mathcal{X}_i^{(i-1)s+1,\text{is}}$	Subband with the band interval between $(i-1)s+1$ and is.
s	Number of spectral bands in each subband.
$\mathbf{T}_i^{\text{local}}$	Local feature extracted from the i^{th} subband.
$\mathbf{T}^{\text{local}}$	Local feature sets of the subbands for an input sample patch.
$\mathbf{T}^{\text{global}}$	Global feature obtained by concatenating $\mathbf{T}^{\text{local}}$ along the feature dimension.
$\mathcal{E}_i^{\text{small}}$,$\mathcal{E}_i^{\text{mid}}$, $\mathcal{E}_i^{\text{large}}$	Embedding processes at the small, medium, and large scales, respectively.
$\widetilde{\mathbf{T}}$	Embedding sequence of $\mathbf{T}^{\text{global}}$ in the RCPT task.
$\widetilde{\mathbf{T}}'$	Masked sequence obtained by the sample reconstruction subtask.
$\overline{\mathcal{X}}$	Pixel reconstruction result corresponding to the $\widetilde{\mathbf{T}}_i$.

I. INTRODUCTION

HYPERSPECTRAL remote sensing has the unique advantage of simultaneously acquiring images and spectra of ground objects. The acquired data, known as hyperspectral images (HSIs), usually consist of dozens to hundreds of bands, which contain rich spectral information and enable HSIs to have more specialized applications than RGB images [1], [2], [3], [4]. As supporting research for many applications, HSI classification has received much attention. The goal is to assign a class label to each pixel in HSIs, which is similar to semantic segmentation in the field of computer vision (CV). In early research, traditional classifiers such as logistic regression and support vector machine (SVM) are commonly used for HSI classification [5], [6], [7], [8]. Meanwhile, considering that the high dimensionality of HSIs may lead to the Hughes phenomenon [9], principal component analysis (PCA) [10], independent component analysis (ICA) [11], and linear discriminant analysis (LDA) [12] are widely adopted for spectral information extraction in the data preprocessing stage. However, the above-mentioned methods only use the spectral information of input samples and neglect the spatial information of sample neighborhood. To extract spatial features, a series of morphology-based methods have been proposed, including morphological profile (MP) [13], extended MP (EMP) [14], and extended multiattribute profile (EMAP) [15]. Moreover, several filtering-based methods [16],

Manuscript received 7 March 2023; revised 14 July 2023 and 29 October 2023; accepted 6 February 2024. Date of publication 23 February 2024; date of current version 8 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290; in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005; and in part by the Research Team Cultivation Program of ShenZhen University under Grant 2023JCT002. *(Corresponding author: Ruyan He.)*

Sen Jia, Yifan Wang, and Ruyan He are with the College of Computer Science and Software Engineering, the Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities, and the Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, Ministry of Natural Resources, Shenzhen University, Shenzhen 518060, China (e-mail: 2070276050@szu.edu.cn).

Shuguo Jiang is with the School of Computer Science, Wuhan University, Wuhan 430071, China (e-mail: senjia@szu.edu.cn; shuguoj@foxmail.com; luckhry106@163.com).

Digital Object Identifier 10.1109/TGRS.2024.3369075

[17] and coding-based strategies [18], [19] are also developed to incorporate both spatial features and spectral features to effectively improve the performance of HSI classification.

With the development of deep learning (DL), many excellent DL-based networks have been successfully applied to HSI classification, such as stacked autoencoders (SAEs), deep belief networks (DBNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs) [20]. Among them, AE is widely used in self-supervised and semi-supervised learning [21], [22], [23], which endows the encoder with certain feature extraction capabilities by encoding and reconstructing samples without using labels. DBN is a probabilistic generation model. Chen et al. [21] first introduced DL into HSI classification and used SAEs for spectral feature extraction. After that, Chen et al. [24] developed a DBN-based deep framework to extract spectral–spatial features hierarchically. RNN has the ability to process pixels of HSI as a sequence-based data and performs classification of ground objects via network reasoning [25], [26], [27], [28], [29]. Mou et al. [25] first proposed a RNN framework for HSI classification. Zhou et al. [28] designed a two-branch long short-term memory network (LSTM) to extract spectral feature and spatial feature, respectively. GAN is trained in an adversarial manner with a generative model and a discriminative model [30], and Zhu et al. [31] explored the effectiveness of GAN for the first time and achieved good performance for HSI classification.

CNN, as a relatively special network, has made a major breakthrough in the field of CV. Despite the constant emergence of new networks, CNNs can achieve competitive results in various visual tasks, which may be attributed to the continuous in-depth study on its network structure and the combination with other networks. The inductive bias in CNNs is well-suited for image processing tasks and is good at extracting spatial–spectral features from HSIs. Some researchers have explored the feasibility of applying CNN-based networks to HSI classification. For example, Hu et al. [32] made a preliminary attempt to stack several 1-D convolutional layers to extract local spectral information and improved the classification accuracy. But the shortcoming is that the spatial information of HSIs is not fully used. Roy et al. [33] proposed a hybrid-CNN by combining the advantages of 3-D-CNN and 2-D-CNN to achieve hierarchical spatial-spectral feature learning. This method reduced the computational complexity of their model compared with the standalone 3-D-CNN model. Yu et al. [34] designed a lightweight 2-D-CNN network that used multiple 2-D convolution kernels with a kernel size of 1×1, and the designed network demonstrated satisfactory performance for HSI classification. With the popularity of CNN, a variety of convolutional kernels and network structures have been designed, which can be combined with other networks to form excellent feature extraction modules to maintain the competitiveness of CNN-based methods.

Due to the proposed frameworks of ResNet [35] and DenseNet [36], skip connection and dense connection have had a great influence on the connection modes of the existing deep network structures. Subsequently, these connection methods have been incorporated into CNN-based networks. Zhong et al. [37] adopted skip connection for every two 2-D-CNN layers and explored the impact of network depth and batch normalization on classification performance of the model. Based on the residual block of PyramidNet's pyramid structure [38], Paoletti et al. [39] designed a novel deep CNN to enhance the diversity of high-level spatial–spectral features. Wang et al. [40] proposed a fast dense spatial–spectral convolution network which applied the dense spectral block and dense spatial block to separately extract spectral and spatial features. In addition, the attention mechanisms [41], [42], [43], [44], including soft attention, hard attention, and self-attention, have been used in combination with CNNs for classification tasks and have achieved competitive results [45], [46], [47]. Mou and Zhu [48] designed a spectral attention module to improve model performance of feature extraction by discriminating the importance of different spectral bands. Hang et al. [49] developed a spectral attention subnetwork and a spatial attention subnetwork to assist the traditional CNN models in extracting the prior information of input samples and obtained superior performance. Zhu et al. [50] introduced a feedback attention module to improve the perception of attention maps. However, it is difficult for CNN-based models to capture the global relationship between features due to the fixed size of receptive field.

In recent years, the transformer, which is initially proposed in the field of natural language processing (NLP) based on attention mechanisms, has received great attention. Researchers have also introduced the transformer into HSI classification. Hong et al. [51] analyzed the difference between transformer and some classical neural networks in detail and developed a vision transformer (ViT)-based SpectralFormer for spectral information learning. Zhong et al. [52] proposed a spatial–spectral transformer (SST) network and a model structure search framework for HSI classification. Sun et al. [53] designed a Gaussian weighted feature tokenizer in the spectral–spatial feature tokenization transformer (SSFTT) to transform the features extracted by CNN. Xue et al. [54] proposed a local transformer network based on spatial partitioning and reconstruction, which mainly consists of a spatial partition restore (SPR) module. Yu et al. [55] proposed a multilevel spectral–spatial transformer network (MSTNet) and aggregated multilevel features through a well-designed decoder. Tu et al. [56] developed a transformer-based framework called local semantic feature aggregation-based transformer (LSFAT) for long-range dependencies representation of multiscale features. Zou et al. [57] designed a local-enhanced spectral–spatial transformer (LESSFormer) to acquire adaptive spectral–spatial tokens and enhance representation capabilities. Song et al. [58] proposed a novel bottleneck SST (BS2T) to capture the long-range global dependencies of HSI pixels and extract the local–global features. The above-mentioned models demonstrate the effectiveness of transformer-based networks for HSI classification. However, transformer-based models still face challenges, including a large number of model parameters and the high requirement for training samples. In addition, current models lack sufficient exploration of self-supervised methods in the pretraining task to fully use the information from unlabeled samples.

In this article, we aim to develop a center-masked transformer (CMT) approach for improving the performance of transformer-based networks in the HSI classification task with limited samples. The CMT approach updates the embedding method and adds a self-supervised pretraining task into the transformer for the first time. On one hand, inspired by the divide-and-conquer algorithm, we design a local-to-global token embedding (L2GTE) framework for HSI datasets, which captures better embedding features. Based on this framework, a multiscale embedding module is built and operated in parallel to obtain finer local embedding tokens for each subband. These local embedding tokens are then concatenated to achieve the global embedding tokens. On the other hand, a regularized center-masked pretraining (RCPT) task is introduced by combining a generative self-supervised learning method based on sample encoding-reconstruction with a mask image model (MIM) [59], which is used to aid the network in effectively using the information of unlabeled samples. More specifically, the main contributions of this article are summarized as follows.

1) First, considering that each spectral interval has unique spatial–spectral characteristics, a L2GTE framework is designed in our CMT approach for HSI embedding to extract the better local representation using an individual embedding module for each subband. Specifically, the feature embedding task in the transformer-based network is divided into several embedding subtasks for the subbands of HSI and each subband adopts its own embedding module to learn local spatial–spectral feature, which not only improves the embedding quality but also reduces the requirements in model size and training sample. Then, to obtain the global embedding of samples, we concatenate the local embedding results of the subbands based on the spatial dimension.
2) Second, to meet the feature extraction of HSI datasets with different spatial distribution and resolutions, a multiscale convolutional token embedding (MCTE) module is proposed for each subband, which consists of three embedding branches at the small, medium, and large scales. The MCTE can obtain better local embedding tokens of the subbands to improve the robustness of the proposed CMT approach and make the model performance more stable.
3) Third, to fully use the information of a large number of unlabeled samples in the HSI datasets, an RCPT is proposed. We use a learnable vector to mask the center token and reconstruct the spectrum of the center pixel in the decoder to effectively learn the contextual relationship between the center pixel and its neighbors without labeled samples. Furthermore, an auxiliary task based on sample encoding-reconstruction is used to prevent model collapse and ensure model stability in the pretraining process.
4) Finally, a series of comparison and ablation experiments are designed and conducted to demonstrate the effectiveness of the proposed CMT approach on five public HSI datasets. The results of the comparison experiments show that our CMT approach achieves excellent performance for HSI classification with insufficient training samples, and the results of the ablation experiments demonstrate the effectiveness of our proposed backbone network and pretraining method.

The rest of this article is organized as follows. Section II briefly introduces the related works of ViT and self-supervised learning methods. Section III provides a detailed description of the proposed CMT approach. The used HSI datasets and the experimental result analysis are presented in Section IV. Finally, the conclusions are summarized in Section V.

II. Related Works

In this section, we introduce the background of the ViT and self-supervised learning methods, which are related to our proposed CMT approach.

A. Vision Transformer

Before the advent of the transformer, RNNs are widely used to process sequence data because their network structure can record short-term memory and excavate the contextual information of the sequence effectively. However, RNNs not only have a disadvantage in handling long sequences but also are complex to perform efficient parallel computations. To a certain extent, the LSTM network alleviates the problem in RNNs of capturing long-range dependencies, but the shortcomings of computational complexity and time-consuming still exist in some scenarios. With the emergence of transformer, transformer networks greatly tackle the above-mentioned problems and dominate tasks in the field of NLP, and the characteristic without inductive bias endows the networks with great potential and have a profound influence on the development of CV.

ViT is the first work to introduce the transformer into CV, which divides the image into equal-sized square regions and achieves the tokens via a linear embedding, eliminating the need for deep CNNs to capture the global receptive field. After that, the transformer models have been applied in the study of HSI classification by various researchers, such as the SpectralFormer by Hong et al. [51], the SSFTT by Sun et al. [53], the MSTNet by Yu et al. [55], the LESSFormer by Zou et al. [57], the global–local 3-D convolutional transformer (GTCT) by Qi et al. [60], and the SST by He et al. [61]. Although some good achievements have been made in these studies, there are still challenges in achieving better classification performance using transformer models. Many of these models adopt a general framework that combines CNN and transformer with a large number of parameters to strike a balance between local and global feature extraction. In addition, an adequate amount of training samples is relied on to achieve better results. In contrast, our work mainly focuses on the design of lightweight structures and the use of limited samples to achieve excellent performance in HSI classification. Therefore, the proposed CMT approach takes a step further in the embedding framework and effectively reduces the number of parameters in the feature extraction module for limited samples, which is more suitable for real-world application scenarios.

B. Self-Supervised Learning Method

It is well-known that the performance of an HSI classification model is closely related to the size of the training samples, and a model with more labeled samples usually achieves better training results. However, sample annotation in HSI is time-consuming and labor-intensive. To address this issue, effectively capturing and fully using the features of unlabeled samples can be considered in the pretraining process, and self-supervised learning provides a good solution for this. Self-supervised learning can construct pseudolabels based on the data itself, which can be used to establish a pretext to learn feature representation of samples without labels. So far, the methods of self-supervised learning, such as mask language model (BERT) and generative pretraining (GPT), have made breakthrough progress in NLP, and the development in the CV domain is slightly behind that in NLP.

In recent years, the most mainstream self-supervised learning includes contrastive learning (CL) and MIM. The representative CL-based methods are the Moco series [62], [63], [64] and the Simclr series [65], [66], and the classical MIM-based methods mainly involve MAE [59], BEIT [67], etc. For HSI classification, CL can directly use the input data to construct positive samples and negative samples by data augmentation and adopt a pretext task to automatically extract the discriminative features of samples for pretraining, but there is no very practical and universal augmentation strategy for HSI datasets due to the unrich spatial information of HSI compared with that of RGB images. In contrast, MIM is a more advantageous framework because HSI has the characteristics of images and the pseudolabels of MIM can be constructed by dividing the input data into several different subsets. Therefore, based on the MIM framework, we design a self-supervised pretraining task to effectively capture the features from unlabeled samples in HSI datasets, and an auxiliary task based on sample reconstruction (SR) is used as a regularization term to prevent model collapse during the pretraining process. Our work represents a pioneering exploration of transformer-based self-supervised learning methods for HSI classification.

III. Methodology

In this section, the proposed CMT approach for HSI classification is depicted and shown in Fig. 1, which improves the transformer-based methods from two aspects: a multiscale convolutional transformer for embedding task and a regularized center-mask pretraining (RCPT) task for self-supervised pretraining. Specifically, CMT mainly consists of an L2GTE framework, an MCTE module, the standard transformer encoder (TE), and an RCPT task. The L2GTE framework primarily divides the sample embedding task into several subtasks to acquire multiscale local embedding tokens individually using the MCTE module (Fig. 2), which are then combined through concatenation to obtain global embedding tokens. The RCPT task, including a center pixel reconstruction (CRP) subtask and an SR subtask, can effectively learn the dependencies between the central object and its neighboring objects. The summary of key mathematical symbols used for the CMT approach is presented in nomenclature.

A. L2GTE Framework

HSIs are quite different from RGB images in both spectral and spatial distributions. When the RGB image embedding method in the transformer is directly applied to extract features of HSIs for classification, the model performance is usually poor, and even worse than that of some state-of-the-art CNN-based methods. This may be because there are fewer labeled samples for training and HSIs have less spatial information compared with RGB images. Therefore, a more suitable embedding method is needed for HSI classification, especially in the case of insufficient labeled samples. In this article, an L2GTE framework for HSI datasets is proposed. First, we introduce a partitioning strategy in the L2GTE framework to divide the feature embedding task into several equivalent subtasks based on the HSI spectral dimensions. Then, the subbands with equal number of spectral bands are fed into embedding modules to obtain the corresponding local spatial–spectral representation, i.e., the local token. Finally, the local tokens extracted by the parallel embedding modules are concatenated together to achieve the global token.

In particular, let $\mathcal{I} \in \mathbb{R}^{H\times W\times D}$ denote the original HSI, where H is the height, W is the width, and D is the number of spectral bands. Considering that the feature extraction will lead to an increase in the number of model parameters and the spectral redundancy phenomenon is common in the original HSI, so principal components analysis (PCA) is adopted to reduce computational load and spectral dimension. The obtained dimension reduction result is denoted as $\mathcal{H} \in \mathbb{R}^{H\times W\times K}$, where K is the number of channels after dimensionality reduction. For an input sample patch $\mathcal{X} \in \mathbb{R}^{h\times w\times K}$, where h and w are the height and width of the input sample patch, the partitioning strategy is used to evenly divide the spectral channels of the input patch into n consecutive subbands. Then, the achieved result can be denoted as $\mathcal{X}' = [\mathcal{X}_1^{1,s}, \mathcal{X}_2^{s+1,2s}, \ldots, \mathcal{X}_n^{(n-1)s+1,ns}] \in \mathbb{R}^{n\times h\times w\times s}$, where s is the number of channels contained in each subband, and $\mathcal{X}_i^{(i-1)s+1,is} \in \mathbb{R}^{h\times w\times s}$ is the subband with the channel interval between $(i-1)s+1$ and is $(i = 1, 2, \ldots, n)$, and $n = K/s$. Finally, the local tokens are extracted from these subbands by the corresponding embedding modules. This process can be formulated as follows:

$$\mathbf{T}_i^{\text{local}} = \text{Emb}_i\left(\mathcal{X}_i^{(i-1)s+1,is}\right) \quad (i = 1, 2, \ldots, n) \tag{1}$$

where $\mathbf{T}_i^{\text{local}} \in \mathbf{R}^{h\times w\times z}$ represents the local feature extracted from the ith subband, and $\text{Emb}_{\text{i}}(\cdot)$ is the feature extraction process in the ith embedding module, which is performed by parallel embedding modules.

Thus, the local features of the subbands in a patch $\mathcal{X}'$ can be described as $\mathbf{T}^{\text{local}} = [\mathbf{T}_1^{\text{local}}, \mathbf{T}_2^{\text{local}}, \ldots, \mathbf{T}_n^{\text{local}}]$, $\mathbf{T}^{\text{local}} \in \mathbb{R}^{n\times hw\times z}$. The global token $\mathbf{T}^{\text{global}} \in \mathbb{R}^{hw\times \text{embdim}}$ is obtained by concatenating these local tokens along the feature dimension, which is represented as follows:

$$\mathbf{T}^{\text{global}} = \text{UnFold}\left(\text{Concat}\left(\mathbf{T}_1^{\text{local}}, \mathbf{T}_2^{\text{local}}, \ldots, \mathbf{T}_n^{\text{local}}\right)\right) \tag{2}$$

where $\text{Concat}(\cdot)$ is the concatenation operation along the feature dimension, $\text{UnFold}(\cdot)$ denotes a spatial dimensional spreading operation, and embdim is the feature dimension of the global token.

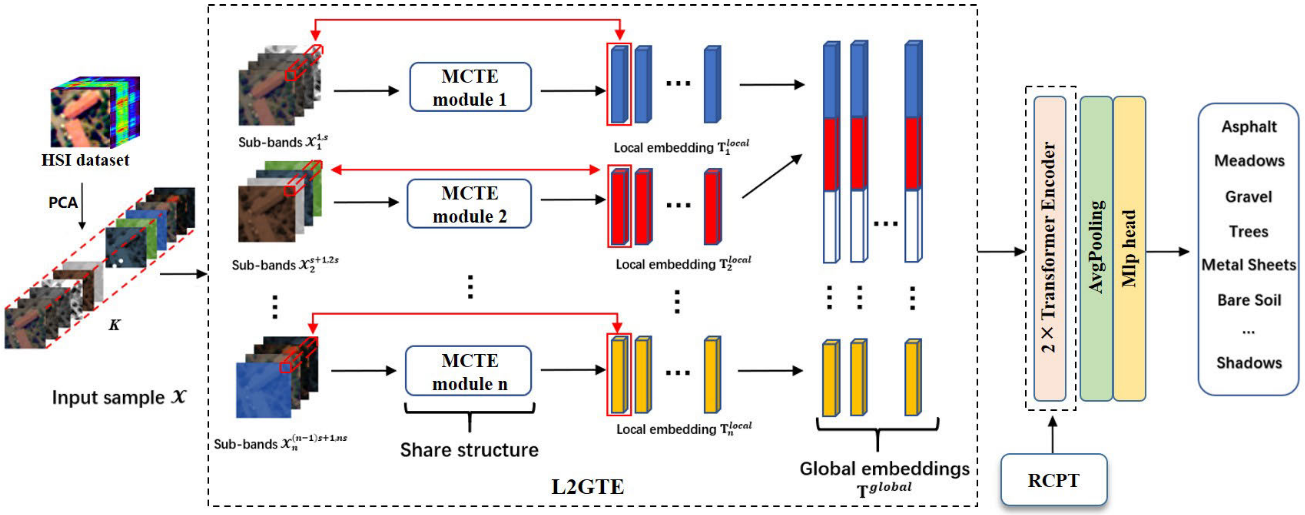

HSI dataset
PCA
K
Input sample $\mathcal{X}$
Sub-bands $\mathcal{X}_1^{1,s}$
Sub-bands $\mathcal{X}_2^{s+1,2s}$
Sub-bands $\mathcal{X}_n^{(n-1)s+1,ns}$
MCTE module 1
MCTE module 2
MCTE module n
Share structure
Local embedding T_1^{local}
Local embedding T_2^{local}
Local embedding T_n^{local}
Global embeddings T^{global}
L2GTE
2 × Transformer Encoder
AvgPooling
Mlp head
RCPT
Asphalt
Meadows
Gravel
Trees
Metal Sheets
Bare Soil
...
Shadows

Fig. 1. Flowchart of the proposed CMT approach for HSI classification.

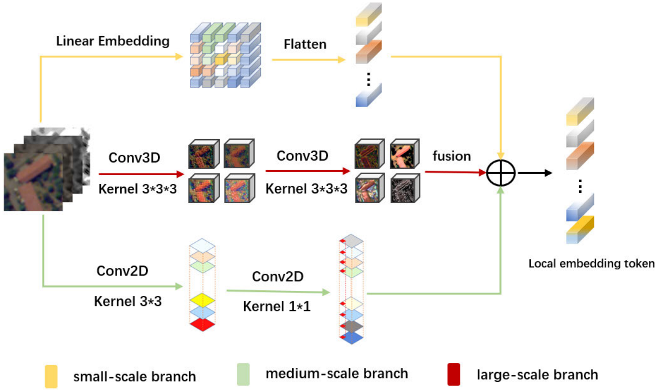


Fig. 2. Structure of the designed MCTE module.

B. MCTE Module

To extract better spatial–spectral features from HSIs with insufficient training samples, we develop an MCTE module for the subband, which includes three branches of linear embedding, 2-D convolutional embedding, and 3-D convolutional embedding at different spatial scales (Fig. 2). Specifically, the linear embedding is applied to extract spectral features at the small scale, the 2-D convolutional layers with kernel size 3 and kernel size 1 are mainly used for spatial information extraction at the medium scale, and two 3-D convolutional layers are also used to obtain spatial information at the large scale.

For an input subband $\mathcal{X}_i^{(i-1)s+1,is} \in \mathbb{R}^{h\times w\times K}$, the spectral information of each pixel is first encoded using a linear embedding at the small scale in the MCTE module, and the process is shown in the following equation:

$$\mathcal{E}_i^{\text{small}} = \text{Linear}^i\left(\mathcal{X}_i^{(i-1)s+1,is}\right) \tag{3}$$

where $\mathcal{E}_i^{\text{small}} \in \mathbb{R}^{h\times w\times z}$ and $\text{Linear}^i(\cdot)$ represents the linear embedding function corresponding to the ith subband at the small scale.

For the medium scale in the MCTE module, the process of feature embedding is shown as follows:

$$\mathcal{F}_i^{\text{spa}} = \text{Conv2D}_1^i\left(\mathcal{X}_i^{(i-1)s+1,is}\right) \tag{4}$$

$$\mathcal{E}_i^{\text{midium}} = \text{Conv2D}_2^i\left(\mathcal{F}_i^{\text{spa}}\right) \tag{5}$$

where $\mathcal{F}_i^{\text{spa}} \in \mathbb{R}^{h\times w\times m}$, $\mathcal{E}_i^{\text{midium}} \in \mathbb{R}^{h\times w\times z}$, m denotes the number of channels of the feature, and $\text{Conv2D}_1^i(\cdot)$ and $\text{Conv2D}_2^i(\cdot)$ denote 2-D convolution operations with the kernel sizes of 3 × 3 and 1 × 1 on the ith subband, respectively.

For the large scale in the MCTE module, the spatial–spectral embedding feature is obtained by performing two 3-D convolution operations. The process can be expressed as follows:

$$\mathcal{F}_i^{\text{ss}} = \text{BN}\left(\text{ReLU}\left(\text{Conv3D}_1^{\text{i}}\left(\mathcal{X}_{\text{i}}^{(\text{i}-1)\text{s}+1,is}\right)\right)\right) \tag{6}$$

$$\mathcal{F}_i^{\text{ss}\prime} = \text{BN}\left(\text{ReLU}\left(\text{Conv3D}_2^{\text{i}}\left(\mathcal{F}_{\text{i}}^{\text{ss}}\right)\right)\right) \tag{7}$$

$$\mathcal{E}_i^{\text{large}} = \text{ReLU}\left(\text{Conv2D}^{\text{i}}\left(\mathcal{F}^{\text{ss}\prime}_{\text{i}}\right)\right)) \tag{8}$$

where $\mathcal{E}_i^{\text{large}} \in \mathbb{R}^{h\times w\times z}$, $\text{Conv3D}^i(\cdot)$ and $\text{Conv2D}^i(\cdot)$ represent the 3-D convolution operation and 2-D convolution operation with the kernel sizes of 3 × 3 × 3 and 1 × 1 on the ith subband, respectively, $\text{ReLU}(\cdot)$ represents the ReLU activation function, and $\text{BN}(\cdot)$ denotes the batch normalization function.

Finally, the three embedding features are fused to obtain the local token corresponding to the ith subband, and the fusion process is shown as follows:

$$\mathcal{E}_i^{\text{mix}} = \text{Conv}_{1\times 1}\left(\mathcal{E}_i^{\text{small}} + \mathcal{E}_i^{\text{midium}} + \mathcal{E}_i^{\text{large}}\right) \tag{9}$$

$$\mathbf{T}_i^{\text{local}} = \text{UnFold}\left(\text{BN}\left(\text{ReLU}\left(\mathcal{E}_{\text{i}}^{\text{mix}}\right)\right)\right) \tag{10}$$

where the definitions of $\text{Conv}_{1\times 1}(\cdot)$, $\text{ReLU}(\cdot)$, $\text{BN}(\cdot)$, and $\text{UnFold}(\cdot)$ are consistent with the above.

C. Transformer Encoder

A standard TE is shown in Fig. 3 [68], which is mainly composed of the position encoding, multihead attention, and feedforward layers. The calculation process can be expressed as follows:

$$\mathbf{E} = \text{Emb}(\mathbf{X}) \tag{11}$$

$$\mathbf{z}_0 = [x_{\text{class}}; \mathbf{E}] + \mathbf{E}_{\text{pos}} \tag{12}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1} \quad \ell = 1, \ldots, \text{L} \tag{13}$$

$$\mathbf{z}_\ell = \text{MLP}\left(\text{LN}\left(\mathbf{z}'_\ell\right)\right) + \mathbf{z}'_\ell \quad \ell = 1, \ldots, \text{L} \tag{14}$$

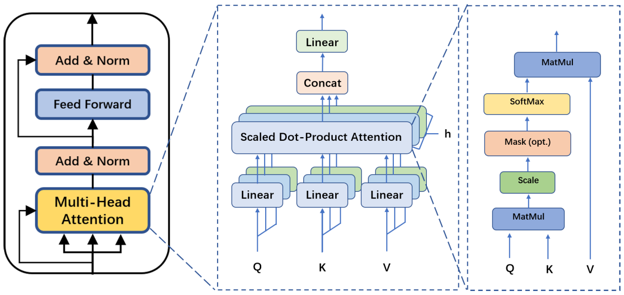


Fig. 3. Detailed structure of a standard TE.

$$y = \mathrm{LN}\left(\mathbf{z}_{\mathrm{L}}^{0}\right) \tag{15}$$

where $\mathbf{X}$ denotes the input, $\mathrm{Emb}(\cdot)$ denotes the embedding function, which in ViT is the linear projection, $\mathrm{LN}(\cdot)$ denotes the LayerNorm function, $\mathrm{MSA}(\cdot)$ denotes the multihead attention function, and $\mathrm{MLP}(\cdot)$ denotes the multilayer perceptron function. It is worth noting that the position encoding and learnable class tokens are removed in our TE to reduce the model burden in the case of small samples, and the multiscale embedding features which are extracted by the convolution operation are directly fed into the module.

As shown in Fig. 3, the multihead attention extends the network ability to focus on the different locations with each head corresponding to a subrepresentation, which has a stronger feature modeling capability than a single head. Specifically, the scaled dot product attention is used in the transformer, and the computational process can be represented as follows:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{16}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are the learnable parameter matrices, $\mathrm{Softmax}(\cdot)$ is the activation function, and $(d_k)^{1/2}$ is the parameter used to keep the gradient smooth.

The calculation process of the multihead attention can be expressed as follows:

$$\mathbf{head}_i = \mathrm{Attention}\left(\mathbf{Q}\mathbf{W}_i^{Q}, \mathbf{K}\mathbf{W}_i^{K}, \mathbf{V}\mathbf{W}_i^{V}\right) \tag{17}$$

$$\mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Concat}(\mathbf{head}_1, \ldots, \mathbf{head}_{\mathrm{h}})\mathbf{W}^{O} \tag{18}$$

where $\mathbf{W}_i^{Q} \in \mathbb{R}^{d_{\text{models}} \times d_k}$, $\mathbf{W}_i^{K} \in \mathbb{R}^{d_{\text{models}} \times d_k}$, $\mathbf{W}_i^{V} \in \mathbb{R}^{d_{\text{models}} \times d_v}$, and $\mathbf{W}^{O} \in \mathbb{R}^{d_{\text{models}} \times \mathrm{h}d_v}$ are parameter matrices, and h is the number of heads and the value is set to 4 in this article. d_{models} is the dimension of input embedding, and the value of d_v and d_k equals to (d_{models}/h).

After that, an average pooling layer is applied to avoid information redundancy and achieve feature compression. Finally, the HSI classification results are obtained through a three-layer MLP head.

D. RCPT Task

Fig. 4 shows the flowchart of our proposed RCPT task. In the pretraining process, the RCPT task consists of two subtasks of the SR and the CRP. Considering the spatial distribution characteristics of HSI samples, the embedding token is generated from the center pixel of the training samples using the masking method. The encoder is adopted basing on our proposed CMT network without the classification layer, and the used decoder is composed of two standard TEs. In general, our RCPT is a multitask reconstruction method, which can reconstruct the center pixel as efficiently as possible and take into account the reconstruction of the whole sample. The CPR subtask can make the model better learn the relationship between the center pixel and adjacent pixels without labels. The SR subtask can be used as a regularization item and encourage the model to capture the global features of HSI samples, which can effectively prevent model collapse.

More specifically, given an input sample $\mathcal{X}$ with a center pixel vector P_{center}, the embedding token of the input sample is $\mathbf{T}^{\text{global}}$. Let $\widetilde{\mathbf{T}} = [\widetilde{T}_1, \widetilde{T}_2, \ldots, \widetilde{T}_{\text{center}}, \ldots, \widetilde{T}_m]$ denote the sequence of embeddings after spatial dimensional flattening of $\mathbf{T}^{\text{global}}$, where $\widetilde{\mathbf{T}} \in \mathbb{R}^{m \times d}$, $\widetilde{T}_i \in \mathbb{R}^{1 \times d}$, and $m = h \times w$. For the SR subtask, the $\widetilde{\mathbf{T}}_{\text{center}}$ is replaced by a learnable vector V_{learn}, resulting in a new sequence $\widetilde{\mathbf{T}}' = [\widetilde{T}_1, \widetilde{T}_2, \ldots, V_{\text{learn}}, \ldots, \widetilde{T}_m]$. Then, the masked sequence $\widetilde{\mathbf{T}}'$ is fed into the decoder. For the CRP subtask, an MLP head is performed to obtain the reconstruction result $\overline{\mathcal{X}} = [\overline{P}_1, \overline{P}_2, \ldots, \overline{P}_{\text{center}}, \ldots, \overline{P}_m]$, where $\overline{P}_i$ is the pixel reconstruction result corresponding to $\widetilde{\mathbf{T}}_i$. Finally, the loss function is defined to make the input P_{center} and the output $\overline{P}_{\text{center}}$ as similar as possible, and it is expressed as follows:

$$L_{\text{center}} = \mathrm{MSE}\left(\mathrm{P}_{\text{center}}, \overline{\mathrm{P}}_{\text{center}}\right) \tag{19}$$

$$L_{\text{sample}} = \mathrm{MSE}\left(\mathcal{X}, \overline{\mathcal{X}}\right) \tag{20}$$

$$L_{\text{total}} = L_{\text{center}} + L_{\text{sample}}. \tag{21}$$

where $\mathrm{MSE}(\cdot)$ denotes the mean squared error function, L_{center} is the value of the central pixel reconstruction loss, and L_{sample} is the value of the SR loss.

IV. Experiments

A. Hyperspectral Datasets

To evaluate the effectiveness and robustness of our proposed CMT method, a series of comparative and ablation experiments are conducted on the five public HSI datasets which were acquired from different scenes. The detailed description of the datasets is as follows.

1) Indian Pines Dataset: The Indian Pines dataset was collected by the AVIRIS sensor in Northwest Indiana, USA, in 1992, which contains 10 249 labeled samples with 16 land-cover classes. The spatial resolution is 20 m for each pixel and the spatial dimension is 145 × 145 pixels. The wavelengths range from 0.4 to 2.5μ m, including 224 spectral bands. After removing 24 noisy bands and water absorption bands, 200 bands are retained in the experiment. Fig. 5 shows the false color image and ground-truth map of the dataset, and the detailed information is presented in Table I.

2) Houston 2013 Dataset: The Houston 2013 dataset was acquired in the area of the University of Houston, Texas, USA, by the ITRES CASI-1500 sensor. It includes 15 land-cover classes and 15 029 labeled samples with a 2.5-m spatial resolution. There are 144 spectral bands ranging from 0.38 to 1.05 μm, and each band covers 349 × 1905 pixels. The false color image and ground-truth map of the dataset are shown

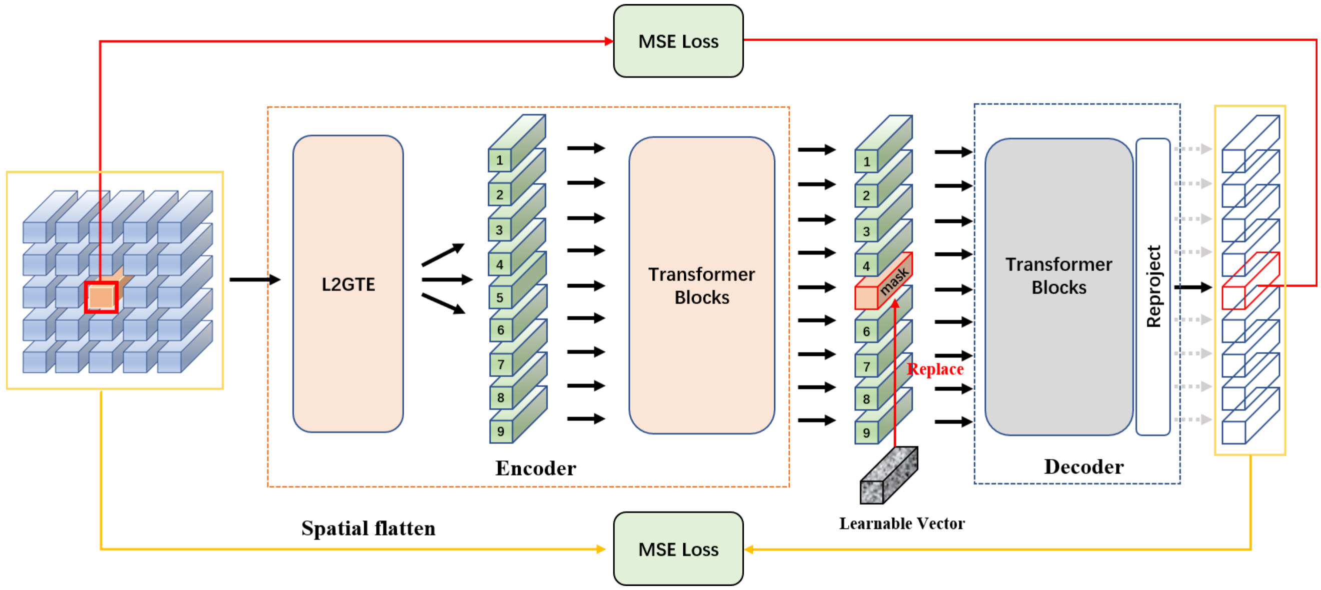


Fig. 4. Flowchart of the proposed RCPT task.

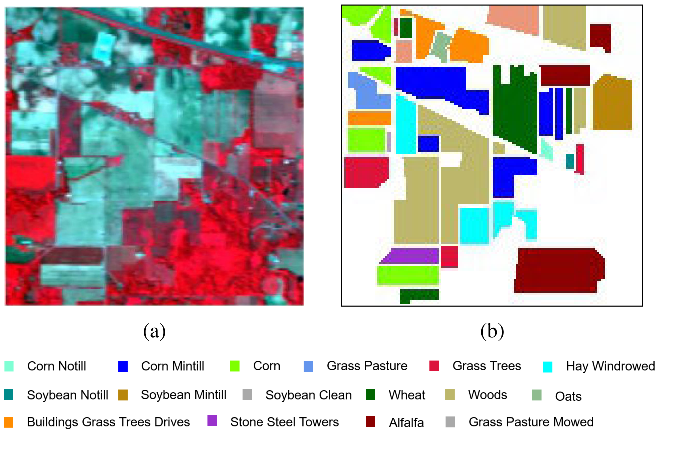


Fig. 5. (a) False color image. (b) Ground-truth map. (c) Labels of the Indian Pines dataset.

in Fig. 6, and the specific information is given in detail in Table II.

3) Pavia University Dataset: The Pavia University dataset was gathered by the ROSIS-03 sensor over the Pavia University, Northern Italy, in 2003. This dataset consists of 42 776 labeled pixels from nine land-cover classes. The spatial dimension is 610 × 340 pixels with a high spatial resolution of 1.3 m/pixel. After 12 noisy bands are discarded from the original 115 spectral bands, the remaining 103 bands are used for classification. Fig. 7 shows the false color image and ground-truth map of the dataset, and the detailed information is presented in Table III.

4) YRE Dataset: The yellow river estuary (YRE) dataset was captured by the Gaofen-5 satellite over the YRE region of Shandong Province, China. The dataset contains 77 937 labeled samples with 20 land-cover classes, most of which are wetland plants. The image has a spatial dimension of 1400 × 1400 pixels and a spatial resolution of 30 m for each pixel. After removing noisy bands, 180 bands from the original spectral bands are selected and processed in the experiment. The false color image and ground-truth map of the dataset are shown in Fig. 8, and the specific information is given in detail in Table IV.

TABLE I

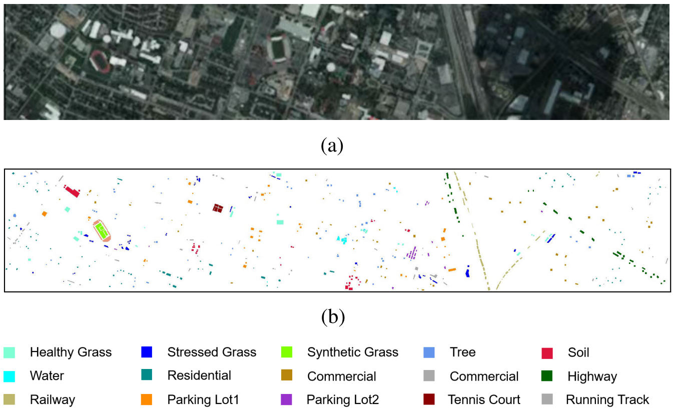
NUMBER OF TRAINING AND TESTING SAMPLES ON THE INDIAN PINES DATASET

Class	Class Name	Training	Testing
1	Corn Notill	5	1429
2	Corn Mintill	5	829
3	Corn	5	229
4	Grass Pasture	5	492
5	Grass Trees	5	742
6	Hay Windrowed	5	484
7	Soybean Notill	5	963
8	Soybean Mintill	5	2463
9	Soybean Clean	5	609
10	Wheat	5	207
11	Woods	5	1289
12	Buildings Grass Trees Drives	5	375
13	Stone Steel Towers	5	90
14	Alfalfa	5	49
15	Grass Pasture Mowed	5	21
16	Oats	5	15
	Total	80	10,286

5) Salinas Dataset: The Salinas dataset was collected by the AVIRIS sensor in the Salinas Valley, California, USA. There are 54 129 labeled samples classified into 16 land-cover classes. The image consists of 512 × 217 pixels with a spatial resolution of 3.7 m/pixel and 224 spectral bands. After removing noisy bands, 204 bands are reserved in the experiment. Fig. 9 shows the false color image and ground-truth map of the dataset, and the detailed information is presented in Table V.

B. Experimental Setup

1) Hyperparameters Setting: In this article, the number of channels (K) after PCA dimensionality reduction is set to 80 which captures most of the information in the HSI datasets. Regarding the number of subband and patch size



(a)

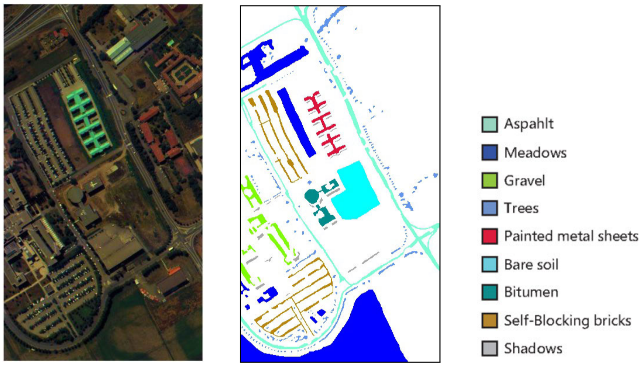
(b)

(c)

Fig. 6. (a) False color image. (b) Ground-truth map. (c) Labels of the Houston 2013 dataset.

TABLE II
NUMBER OF TRAINING AND TESTING SAMPLES ON THE HOUSTON 2013 DATASET

Class	Class Name	Training	Testing
1	Healthy Grass	5	1,246
2	Stressed Grass	5	1,249
3	Synthetic Grass	5	692
4	Tree	5	1,239
5	Soil	5	1,237
6	Water	5	320
7	Residential	5	1,263
8	Commercial	5	1,239
9	Road	5	1,247
10	Highway	5	1,222
11	Railway	5	1,230
12	Parking Lot1	5	1,228
13	Parking Lot2	5	464
14	Tennis Court	5	423
15	Running Track	5	655
	Total	75	14,954



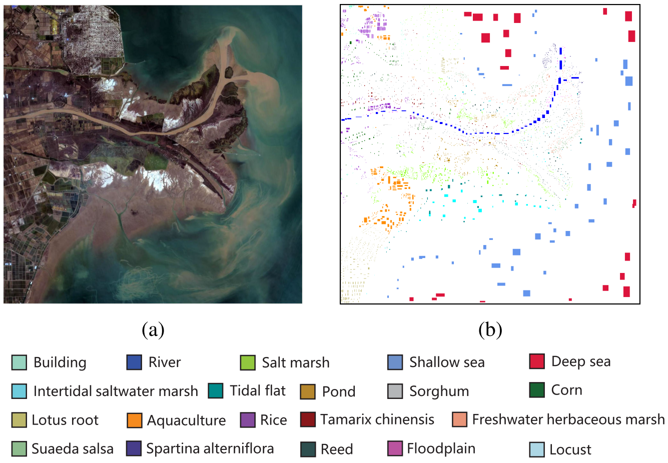
(a) (b) (c)

Fig. 7. (a) False color image. (b) Ground-truth map. (c) Labels of the Pavia University dataset.

for an input sample, a series of quantitative experiments are conducted to study parameter sensitivity, and the overall classification accuracy is shown in Fig. 10. In general, a large input sample size would increase the computational load, while a small size may not contain the sufficient information. Therefore, the parameters of subband number and patch size are determined by balancing the computational complexity and the spectral–spatial information. The patch size of height (h) and width (w) is set to 13, the number of subbands (n) is set to 8, and each subband contains ten spectral channels. The embedding size of each token is set to 128, the hidden layer

TABLE III
NUMBER OF TRAINING AND TESTING SAMPLES ON THE PAVIA UNIVERSITY DATASET

Class	Class Name	Training	Testing
1	Asphalt	5	6,626
2	Meadows	5	18,644
3	Gravel	5	2,094
4	Trees	5	3,059
5	Painted metal sheets	5	1,340
6	Bare Soil	5	5,024
7	Bitumen	5	1,325
8	Self-Blocking Bricks	5	3,677
9	Shadows	5	942
	Total	45	42,731



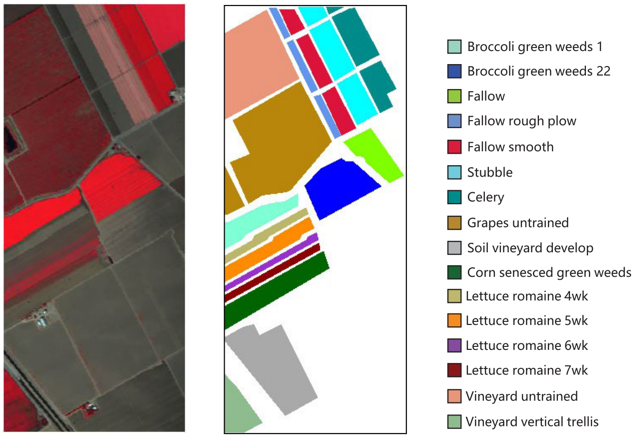
(a) (b)

(c)

Fig. 8. (a) False color image. (b) Ground-truth map. (c) Labels of the YRE dataset.

TABLE IV
NUMBER OF TRAINING AND TESTING SAMPLES ON THE YRE DATASET

Class	Class Name	Training	Testing
1	Building	10	523
2	River	10	5,366
3	Salt Marsh	10	4,985
4	Shallow Sea	10	17,540
5	Deep Sea	10	18,667
6	Intertidal Saltwater Marsh	10	2,333
7	Tidal Flat	10	1,782
8	Pond	10	1,777
9	Sorghum	10	636
10	Corn	10	1499
11	Lotus Root	10	2,709
12	Aquaculture	10	8,009
13	Rice	10	5,498
14	Tamarix Chinensis	10	1,210
15	Freshwater Herbaceous Marsh	10	1,407
16	Suaeda Salsa	10	864
17	Spartina Alterniflora	10	570
18	Reed	10	1,960
19	Floodplain	10	337
20	Locus	10	65
	Total	200	77,737



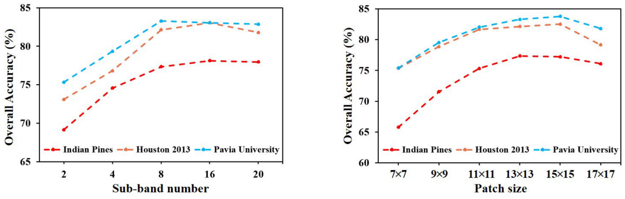
(a) (b) (c)

Fig. 9. (a) False color image. (b) Ground-truth map. (c) Labels of the Salinas dataset.

TABLE V

NUMBER OF TRAINING AND TESTING SAMPLES ON THE SALINAS DATASET

Class	Class Name	Training	Testing
1	Brocoli green weeds 1	5	2,004
2	Brocoli green weeds 2	5	3,721
3	Fallow	5	1,971
4	Fallow rough plow	5	1,389
5	Fallow smooth	5	2,673
6	Stubble	5	3,954
7	Celery	5	3,574
8	Grapes untrained	5	11,266
9	Soil vinyard develop	5	6,198
10	Corn senesced green weeds	5	3,273
11	Lettuce romaine 4wk	5	1,063
12	Lettuce romaine 5wk	5	1,922
13	Lettuce romaine 6wk	5	911
14	Lettuce romaine 7wk	5	1,065
15	Vinyard untrained	5	7,263
16	Vinyard vertical trellis	5	1,802
	Total	80	54,129



(a) (b)

Fig. 10. Classification accuracy for parameter sensitivity of (a) subband number and (b) patch size.

dimension is 64, the number of attention heads is four, and the number of encoders is two. Taking into account of the training stability and convergence speed, the learning rate for model training is set to 0.001, and Adam is adopted as the gradient optimizer.

2) Evaluation Metrics and Running Platforms: To quantitatively compare the effectiveness of the proposed CMT approach and other methods, three metrics of overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) are calculated to evaluate the classification results of different methods. Moreover, to reduce the errors caused by random sample selection, each experiment is repeated ten times. All the experiments are conducted on the computer with an Intel Xeon Platinum 8260 CPU, a 64-GB RAM, and an NVIDIA Tesla P100-16GB GPU.

3) Comparison Methods: To comprehensively evaluate the performance of our proposed method for HSI classification, three CNN-based models and three transformer-based models are selected for comparative experiments, which are the state-of-the-art methods and described as follows.

1) *CNNHSI:* The CNN architecture used several 2-D convolution layers and learned features automatically with the limited training samples for HSI classification [34].
2) *HybridSN:* The 3-D-CNN and 2-D-CNN were designed to extract hierarchical spatial–spectral features to obtain high-level sample representations [33].
3) *SPRN:* A series of parallel CNNs were adopted to extract spectral–spatial features from different subbands of HSI spectral bands [69].
4) *SpectralFomer:* This is a highly flexible backbone network with the ability to both pixelwise and patchwise inputs, and a soft residual structure was designed to effectively reduce the information loss during feature transformation [51].
5) *SSFTT:* The CNN embedding module was applied for shallow spectral–spatial feature extraction, and a Gaussian weighted feature tokenizer was introduced to transform the shallow features into deep semantic features [53].
6) *SPRLT:* The local transformer with a spatial partitioning restore module was devised to dynamically obtain the spatial attention weights of samples by measuring the similarity between pixels [54].

It is worth noting that the model structures and parameter settings of the compared methods follow their open source codes or the corresponding original papers.

C. Ablation Analysis

To verify the effectiveness of our proposed method and the rationality of the model structure design, the ablation experiments are carried out on the L2GTE framework, the MCTE module, and the RCPT task on five public HSI datasets. The specific analysis is as follows.

1) L2GTE Framework: To verify the effectiveness of L2GTE framework, we compare the model performance with and without the L2GTE framework for feature extraction by the same MCTE module and RCPT task. The model embeddings for two experiments are as follows: all the bands of HSI dataset versus subbands of the HSI dataset obtained using the L2GTE framework. The classification results of L2GTE ablation experiment are shown in Table VI. Our proposed L2GTE framework with subband embedding effectively improves the model performance on each dataset, especially on the Indian Pines and Pavia University datasets with an increase of 3.18% and 3.92% on the classification accuracy, respectively. In addition, the accuracy on the other three datasets improves by 1%–2%. The results show that compared with all the bands' embedding, more spatial–spectral information can be captured from multiple subbands by the MCTE module to

TABLE VI
ABLATION RESULTS OF THE L2GTE FRAMEWORK ON THE FIVE DATASETS

Dataset	All-bands Embedding			Sub-bands L2GTE		
	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ
Indian Pines	70.67	83.29	67.08	73.85	85.07	70.68
Houston 2013	78.02	80.84	76.25	79.78	82.34	78.15
Pavia University	77.52	77.58	70.28	81.44	80.47	75.26
YRE	89.13	84.61	87.45	90.22	85.52	88.68
Salinas	89.34	95.20	88.16	90.31	95.61	90.02

obtain the discriminative features of ground objects for HSI classification, illustrating the superiority and reasonableness of subband embedding in the L2GTE framework.

2) MCTE Module: In the ablation experiments of the MCTE module, we compare and evaluate the model performance of single-scale and multiscale embedding modules for feature extraction under the same L2GTE framework and RCPT task. This includes three individual single-branch experiments conducted at small, medium, and large scales, and an experiment using our proposed MCTE module. The experimental results are shown in Table VII. The comparison results show that the classification performance of the single branch at the three scales is inconsistent on the different datasets. For the small-scale branch, the model performs well on the Indian Pines, Pavia University, and Salinas datasets, while the accuracy is relatively poor on the Houston 2013 and YRE datasets. On the contrary, the medium-scale branch achieves better performance on the Houston 2013 and YRE datasets, and the large-scale branch performs better on the Houston 2013, YRE, and Salinas datasets. Our proposed multibranch MCTE module well integrates the advantages of each branch and achieves higher classification accuracy than the single-branch module, i.e., an increase of 3.22% on the Houston 2013 dataset and an improvement of 4.07% on the Pavia University dataset, which proves that our MCTE module is effective and necessary for HSI classification.

3) RCPT Task: Our designed RCPT task includes two subtasks of the SR and the CPR. To verify the effectiveness of the RCPT task, the SR subtask and the CPR subtask are performed independently to evaluate the contribution to HSI classification accuracy under the same L2GTE framework and MCTE module. All the experiments are conducted using five training samples per class for fine-tuning, and the results are listed in Table VIII. It can be seen that when the SR subtask is used individually, the classification accuracy has an increase of 2.18% on the Pavia University dataset, while the accuracy on the YRE dataset is not significantly increased. For the CPR subtask, the accuracy improves by about 2% on the Indian Pines and Houston 2013 datasets, but the improvement is insignificant on the Salinas dataset. However, the classification performance of our proposed RCPT task is significantly improved on the five datasets, especially on the Indian Pines dataset with an increase of 3.49%. This may be attributed to the fact that our proposed pretraining task can effectively capture the relationship between the central ground objects and their neighbors in self-supervised learning, leading to an effective performance improvement for HSI classification.

TABLE VII
ABLATION RESULTS OF THE MCTE MODULE ON THE FIVE DATASETS (NOTE: *s* IS FOR THE SMALL SCALE, *m* IS FOR THE MEDIUM SCALE, AND *l* IS FOR THE LARGE SCALE)

Dataset	Branch			Metric		
	s	m	l	OA (%)	AA (%)	κ
Indian Pines	✓			71.34	82.38	67.74
		✓		69.76	81.39	65.93
			✓	67.90	81.54	64.15
	✓	✓	✓	**73.85**	**85.07**	**70.68**
Houston 2013	✓			75.09	78.34	73.09
		✓		76.56	79.75	74.67
			✓	76.39	79.58	74.51
	✓	✓	✓	**79.78**	**82.34**	**78.15**
Pavia University	✓			77.37	79.18	70.43
		✓		75.49	72.81	67.50
			✓	74.66	74.55	67.01
	✓	✓	✓	**81.44**	**80.47**	**75.26**
YRE	✓			88.22	84.56	86.39
		✓		89.90	85.36	88.32
			✓	89.58	84.75	87.95
	✓	✓	✓	**90.22**	**85.52**	**88.68**
Salinas	✓			89.74	95.27	88.60
		✓		88.98	95.11	87.77
			✓	89.56	95.17	88.39
	✓	✓	✓	**90.31**	**95.61**	**90.02**

TABLE VIII
ABLATION RESULTS OF THE RCPT TASK ON THE FIVE DATASETS

Dataset	Pretrain Task		Metric		
	SR	CPR	OA (%)	AA (%)	κ
Indian Pines	✗	✗	73.85	85.07	70.68
	✓	✗	75.24	85.73	72.24
	✗	✓	76.06	**86.07**	73.17
	✓	✓	**77.34**	**86.07**	**74.56**
Houston 2013	✗	✗	79.78	82.34	78.15
	✓	✗	81.43	83.83	79.97
	✗	✓	81.92	**84.40**	80.47
	✓	✓	**82.14**	84.19	**80.70**
Pavia University	✗	✗	81.44	80.47	75.26
	✓	✗	83.52	83.59	77.99
	✗	✓	82.89	83.60	77.37
	✓	✓	**83.70**	**83.78**	**78.21**
YRE	✗	✗	90.22	85.52	88.68
	✓	✗	90.40	86.24	88.89
	✗	✓	90.75	86.56	89.30
	✓	✓	**91.43**	**86.60**	**90.08**
Salinas	✗	✗	90.22	85.52	88.68
	✓	✗	91.03	**96.17**	90.04
	✗	✓	90.33	95.62	89.24
	✓	✓	**91.82**	95.38	**90.90**

D. Classification Results of Comparative Experiments

With respect to the division of training set and testing set, a small fixed number of training samples is randomly selected to train all the models, and the remaining samples

TABLE IX
CLASSIFICATION ACCURACY (%) AND KAPPA MEASURE FOR THE INDIAN PINES DATASET

Class	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
1	98.29	82.20	96.34	79.51	**99.76**	98.78	97.32
2	60.04	35.64	54.70	34.03	54.03	45.56	**61.49**
3	45.47	22.65	44.08	17.47	48.82	**51.99**	46.96
4	80.43	53.88	88.62	22.50	**88.66**	75.09	85.69
5	69.58	63.89	66.00	16.53	**70.25**	66.05	68.20
6	97.61	79.63	97.93	45.23	90.10	94.08	**98.34**
7	**100.00**	93.04	**100.00**	83.91	**100.00**	**100.00**	**100.00**
8	99.22	93.07	98.88	90.19	99.83	99.75	**100.00**
9	84.00	83.33	**100.00**	99.33	98.00	**100.00**	**100.00**
10	32.04	22.45	31.83	26.95	43.55	63.31	**81.44**
11	52.83	32.85	68.06	61.49	64.55	52.71	**71.44**
12	75.58	36.36	60.77	18.76	53.33	42.60	**77.35**
13	98.85	97.05	**100.00**	94.10	99.55	99.65	99.15
14	93.98	50.92	93.60	75.73	**94.71**	84.23	92.44
15	85.20	45.30	88.87	16.64	95.35	89.37	**97.35**
16	**100.00**	99.89	**100.00**	77.50	99.43	**100.00**	**100.00**
OA (%)	67.31	44.66	69.40	46.54	69.59	65.32	**77.34**
AA (%)	79.57	62.01	80.61	53.74	81.25	78.95	**86.07**
κ	63.14	38.87	65.04	38.97	65.64	61.13	**74.56**

are the testing set. For the YRE dataset, ten samples are selected for each category as the training set. For the other four datasets, the number of training samples for each class is five. The specific information is shown in Tables I–V. In addition, the classification maps of the Houston 2013 and YRE datasets show the entire scene of the HSIs, and the other three datasets, Indian Pines, Pavia University, and Salinas, display the classification results of labeled regions in the HSIs. The detailed result analysis is as follows.

1) Indian Pines Dataset: Table IX exhibits the classification accuracy of the Indian pines dataset, and our proposed CMT approach has strong competitiveness and the best classification performance in the case of small samples. There are ten classes of ground objects which achieved the highest classification accuracy by our CMT method, and the OA and AA are 77.34% and 86.07% which are 7.75%–32.68% and 4.82%–32.33% higher than the compared ones, respectively. Among transformer-based methods, the SSFTT model has the second highest accuracy, while the SF model has the worst classification accuracy, which may be caused by the limited training samples. Among the CNN-based methods, the classification performance of CNNHSI and SPRN is relatively similar, and the classification accuracy is higher than that of the HybirdSN model. Fig. 11 shows the classification maps of seven methods on the Indian Pines dataset, and our CMT approach is more advantageous than the others, which better shows the distribution of ground objects and is closer to the ground truth. The classification results of HybridSN and SF are the worst and many ground objects have misclassification errors, which may be related to the poor feature learning ability of the models in the case of small samples. In addition, our CMT approach successfully identifies the wheat type (class 10) in most areas, while other methods generally misclassify the wheat class as the woods type (class 11) in the center of the classification map.

2) Houston 2013 Dataset: The Houston 2013 dataset is a large scene dataset, and only some areas are labeled as ground truth [Fig. 12(a)]. The classification results and classification maps of the Houston 2013 dataset are shown in Table X and Fig. 12. Our approach achieves the highest classification accuracy on eight types of ground objects and has the best performance with the OA of 82.14%, AA of 84.19%, and κ of 80.70. The OA values of the compared methods vary from 62.55% to 77.70%, and the κ values range from 59.60 to 75.90. The HybridSN and SF have the worst classification accuracy, which may be due to the fact that the model performance cannot be fully used in the case of small samples. The classification maps show that our proposed CMT approach better exhibits the spatial distribution of urban pattern, especially on the types of highway and railway, while those regions identified by other methods usually expand the real areas occupied by such ground objects, resulting in relatively large misclassifications.

3) Pavia University Dataset: The classification accuracy and classification maps for the Pavia University dataset are shown in Table XI and Fig. 13. Our CMT approach demonstrates the superior performance and achieves the best accuracy with an OA of 83.28%, which is 2.77%–19.21% higher than other methods. The SSFTT obtains the second best

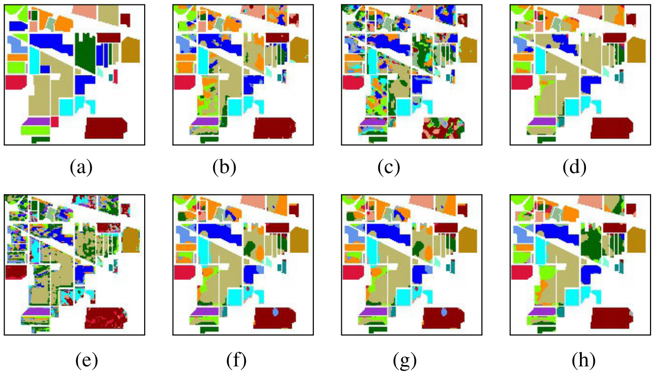

(a) (b) (c) (d)
(e) (f) (g) (h)

Fig. 11. Classification maps on the Indian Pines dataset of (a) ground truth, (b) CNNHSI (67.31%), (c) HybridSN (44.60%), (d) SPRN (69.40%), (e) SF (46.54%), (f) SSFTT (69.59%), (g) SPRLT (65.32%), and (h) ours (77.34%).

TABLE X
CLASSIFICATION ACCURACY (%) AND KAPPA MEASURE FOR THE HOUSTON 2013 DATASET

Class	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
1	82.61	71.71	**85.62**	83.72	75.13	80.51	79.47
2	64.37	64.95	**79.33**	59.46	70.22	67.45	79.08
3	**100.00**	96.69	99.80	92.30	98.86	99.12	**100.00**
4	92.68	70.36	94.14	95.41	95.18	97.39	**98.51**
5	99.38	78.91	98.79	86.87	99.39	96.64	**100.00**
6	85.75	86.50	86.37	44.38	86.91	84.66	**87.53**
7	70.18	46.37	**79.54**	26.21	70.37	78.78	77.00
8	39.45	43.32	39.04	44.00	40.44	45.49	**49.77**
9	61.28	46.05	74.30	46.55	59.27	**76.05**	72.19
10	65.36	61.95	46.80	62.36	89.21	36.93	**94.51**
11	79.72	70.98	77.63	42.89	84.07	64.05	**85.59**
12	52.05	56.76	55.26	60.31	62.43	**70.01**	59.63
13	82.75	62.59	85.51	25.47	81.49	**96.81**	80.06
14	99.29	98.20	99.92	83.76	99.88	99.81	**100.00**
15	**100.00**	99.97	99.90	87.30	93.92	96.41	99.57
OA (%)	74.86	66.09	76.96	62.55	77.70	75.65	**82.14**
AA (%)	78.32	70.35	80.14	62.73	80.45	79.34	**84.19**
κ	72.83	63.41	75.12	59.60	75.90	73.73	**80.70**

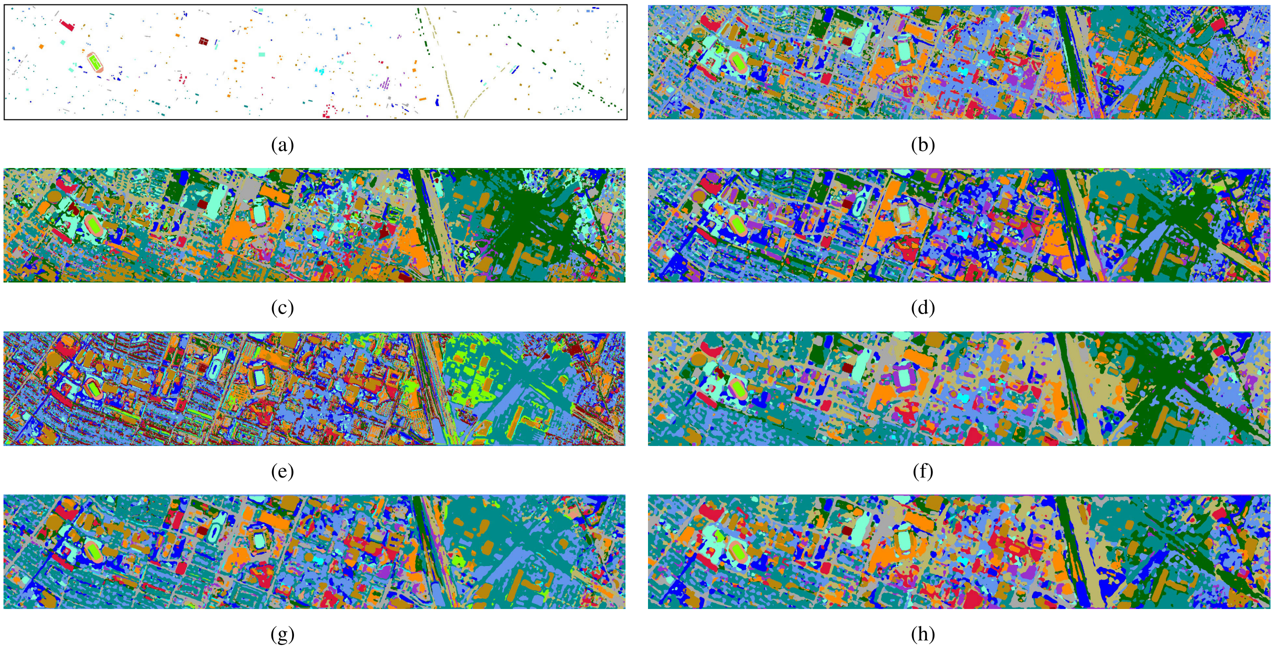

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 12. Classification maps on the Houston 2013 dataset of (a) ground truth, (b) CNNHSI (74.86%), (c) HybridSN (66.09%), (d) SPRN (76.96%), (e) SF (62.55%), (f) SSFTT (77.70%), (g) SPRLT (75.65%), and (h) ours (82.14%).

TABLE XI
CLASSIFICATION ACCURACY (%) AND KAPPA MEASURE FOR THE PAVIA UNIVERSITY DATASET

Class	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
1	66.46	46.74	**81.63**	53.68	66.13	78.09	76.94
2	78.21	70.47	87.54	89.67	90.95	84.67	**93.70**
3	**91.66**	89.30	60.35	30.09	81.28	53.47	81.10
4	80.72	66.49	**86.56**	68.43	75.62	75.09	78.42
5	**100.00**	**100.00**	99.74	99.99	99.72	99.83	**100.00**
6	57.01	51.76	34.27	29.95	**67.47**	52.40	64.24
7	81.59	88.41	85.18	66.73	92.88	78.74	**96.27**
8	46.55	42.61	78.57	67.65	60.16	**84.35**	58.58
9	99.67	59.27	**99.98**	99.84	93.50	99.60	98.16
OA (%)	73.27	64.07	78.77	70.57	80.51	78.23	**83.28**
AA (%)	77.99	68.34	79.31	67.34	80.85	78.47	**83.06**
κ	65.78	54.71	71.84	60.82	74.29	71.55	**77.74**

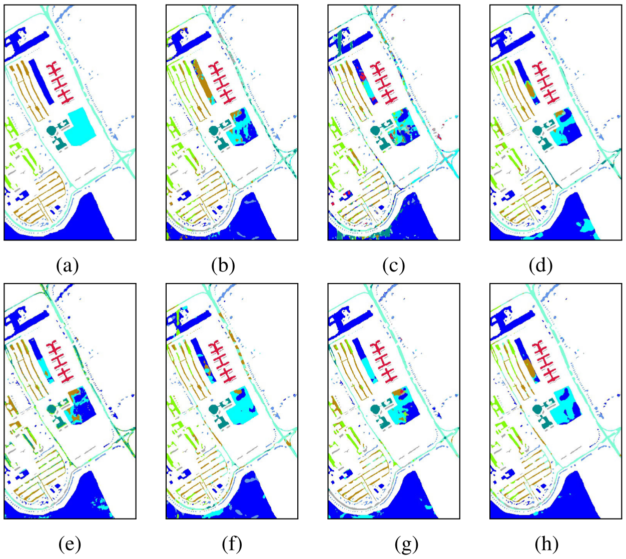

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 13. Classification maps on the Pavia University dataset by (a) ground truth, (b) CNNHSI (73.27%), (c) HybridSN (64.07%), (d) SPRN (78.77%), (e) SF (70.57%), (f) SSFTT (80.51%), (g) SPRLT (78.23%), and (h) ours (83.28%).

classification results, followed by SPRN and SPRLT, while the HybridSN has the worst classification accuracy. Specifically, the CMT approach achieves the best results in types of meadows, painted metal sheets, and bitument, and clearly restores the outline of asphalt. Similarly, the CNNHSI has the highest accuracy in the identification of gravel, trees, and painted metal sheets. In general, most methods can better identify painted metal sheets and shadows, but the classification accuracy for bare soil is relatively poor, ranging from 29.95% of SF to 67.47% of SSFTT. In addition, there is a large performance difference in distinguishing self-blocking bricks, i.e., the accuracy of 84.35% and 42,61% for SPRLT and HybridSN, respectively.

4) YRE Dataset: The YRE dataset is also a large scene dataset, and Table XII and Fig. 14 show the specific classification accuracy and classification maps. Except for CNNHSI and HyBridSN, the classification accuracy of other methods is higher than 87.00%. Our CMT approach has the best accuracy with an OA of 91.38%, which is 2.31%–18.42% higher than that of the compared methods. The transformer-based methods achieve more satisfactory classification results than the CNN-based methods, i.e., the OA of 89.07% of SSFTT and the OA of 72.96% of CNNHSI, which may be attributed to the powerful context capture ability of the transformer module to better distinguish the ground objects in the complex scenes. From the classification maps, we can see that the results of the SPRN and SPRLT methods show the smoother feature distribution of ground objects, followed by the maps of SSFTT and our approach. However, the obvious smearing phenomenon occurs in the classification maps of the CNNHSI and HyBridSN methods, which may be related to the CNN-based structure.

TABLE XII
CLASSIFICATION ACCURACY (%) AND KAPPA MEASURE FOR THE YRE DATASET

Class	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
1	73.13	73.10	85.95	66.96	80.76	**90.61**	88.83
2	98.91	99.39	99.49	**99.65**	99.46	98.60	99.09
3	60.89	63.83	81.12	82.79	**89.74**	80.13	85.27
4	64.17	65.84	89.75	90.43	91.84	93.49	**95.79**
5	70.94	76.11	98.91	**98.94**	94.60	97.23	98.40
6	82.00	81.30	82.36	84.59	**87.06**	83.34	86.38
7	49.24	53.39	57.23	63.93	57.77	**64.97**	59.19
8	74.03	71.46	83.20	82.80	81.37	**85.58**	84.67
9	87.24	86.38	84.10	69.56	85.94	83.88	**88.63**
10	**93.38**	78.79	83.03	76.94	82.89	81.67	91.21
11	79.57	83.79	70.67	83.07	**92.92**	81.42	92.47
12	79.10	79.43	83.68	73.97	80.98	77.28	**85.04**
13	75.31	83.64	88.37	84.24	91.49	87.26	**92.38**
14	77.43	72.18	88.30	**88.33**	79.60	86.76	83.12
15	80.42	62.96	**92.93**	92.86	70.18	88.63	63.28
16	89.64	88.24	90.32	77.82	84.09	87.70	**90.66**
17	93.76	92.49	97.96	86.28	94.44	84.65	**97.98**
18	56.53	55.43	**75.96**	66.32	69.16	57.62	69.09
19	74.82	81.22	78.58	61.99	82.97	75.79	**84.21**
20	**95.14**	81.23	71.85	77.69	90.46	72.31	85.85
OA (%)	72.96	74.86	89.05	87.78	89.07	88.57	**91.38**
AA (%)	77.78	76.51	84.19	80.46	84.39	82.95	**86.08**
κ	69.30	71.54	87.32	85.86	87.37	86.78	**90.02**

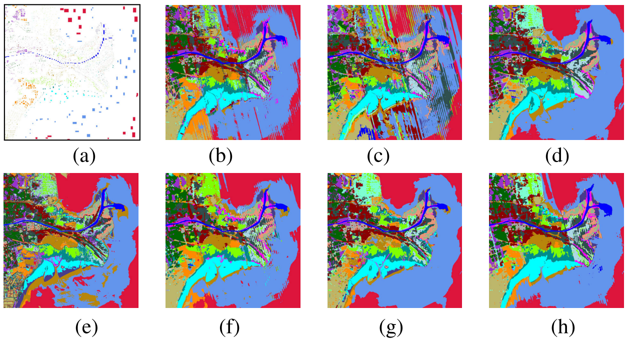

(a) (b) (c) (d)
(e) (f) (g) (h)

Fig. 14. Classification maps on the YRE dataset by (a) ground truth, (b) CNNHSI (72.96%), (c) HybridSN (74.86%), (d) SPRN (89.05%), (e) SF (87.78%), (f) SSFTT (89.07%), (g) SPRLT (88.57%), and (h) ours (91.38%).

TABLE XIII
CLASSIFICATION ACCURACY (%) AND KAPPA MEASURE FOR THE SALINAS DATASET

Class	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
1	99.89	98.81	91.73	79.28	**100.00**	85.94	**100.00**
2	98.01	93.69	99.02	86.09	**100.00**	91.44	99.70
3	79.97	94.73	87.65	77.57	99.70	85.95	**99.93**
4	96.40	94.74	99.48	98.78	**99.87**	99.55	98.25
5	96.05	92.17	94.90	91.50	**98.28**	97.90	96.55
6	99.81	94.95	99.84	99.69	98.51	99.54	**100.00**
7	99.98	98.98	**100.00**	99.53	**100.00**	**100.00**	**100.00**
8	73.57	61.84	73.31	70.63	**84.72**	55.59	82.34
9	99.73	86.32	98.76	91.68	**100.00**	97.35	99.81
10	90.51	88.89	**95.02**	66.52	94.07	87.60	93.58
11	98.79	94.73	96.88	77.61	98.31	93.95	**100.00**
12	96.84	86.81	99.32	99.75	95.59	**99.98**	96.83
13	84.89	85.12	**99.41**	97.56	81.30	96.10	90.30
14	95.19	91.80	98.54	98.78	98.78	**99.31**	95.14
15	73.39	62.07	61.37	53.16	68.97	74.47	**79.43**
16	98.65	87.04	96.44	80.24	97.06	89.93	**98.83**
OA (%)	88.61	81.66	87.46	80.38	91.07	83.95	**91.82**
AA (%)	92.61	88.29	93.23	85.52	94.70	90.91	**95.38**
κ	87.34	79.76	86.04	78.20	90.47	82.24	**90.89**

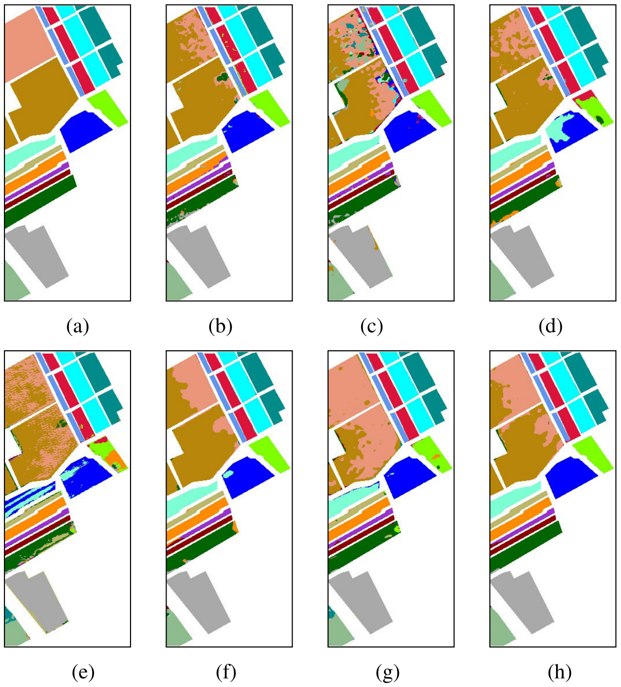
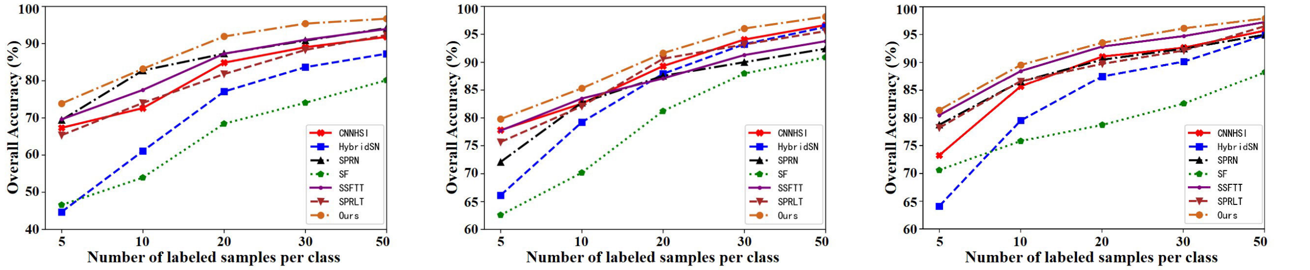

(a) (b) (c) (d)
(e) (f) (g) (h)

Fig. 15. Classification maps on the Salinas dataset by (a) ground truth, (b) CNNHSI (88.61%), (c) HybridSN (81.66%), (d) SPRN (87.46%), (e) SF (80.38%), (f) SSFTT (91.07%), (g) SPRLT (83.95%), and (h) ours (91.82%).

5) Salinas Dataset: The classification results and classification maps for the Salinas dataset are shown in Table XIII and Fig. 15. The classification accuracy of all the methods is higher than 80.00%, probably because the ground objects in the Salinas dataset are mainly distributed in blocks with less interference between different classes, and the spectra of the same ground objects have high similarity. Our CMT approach still achieves the best accuracy with an OA of 91.82% and seven types of ground objects have the highest classification accuracy, followed by the methods of SSFTT, CNNHSI, and SPRN. However, the accuracy of the other three methods of HybridSN, SF, and SPRLT has a large gap with that of our approach, and the OAs are 10.16%, 11.44%, and 7.85% lower compared with our OA, which may be because the models struggle to handle the small samples for HSI classification. From the classification maps, all the methods have a certain degree of misclassification on the ground objects of grapes untrained (class 8) and vineyard untrained (class 15), since there are great similarities between those two objects, and the methods are difficult to clearly distinguish their discriminative features when the training samples are insufficient.

E. Complexity Analysis

To further demonstrate the superiority and effectiveness of our proposed CMT approach, we present the computational complexity of the comparative experiments. The evaluation metrics include the running time and the number of model parameters. For the running time, we aggregate the model training and testing times as an evaluation metric, considering the limited number of training samples used for each model. The results are displayed in Table XIV. In terms of running



(a) (b) (c)

Fig. 16. Classification accuracy versus different percentages of training samples per class. (a) Indian Pines, (b) Houston 2013, and (c) Pavia University.

TABLE XIV

RUNNING TIME (S) AND MODEL PARAMETERS OF DIFFERENT METHODS

Dataset	Metrics	CNNHSI	HybridSN	SPRN	SF	SSFTT	SPRLT	Ours
Indian Pines	Running time	1.2 s	1.8 s	3.3 s	3.8 s	2.9 s	6.6 s	2.1 s
	Model parameters	35,024	533,753	205,396	126,775	130,496	839,728	103,872
Pavia University	Running time	2.0 s	3.3 s	4.2 s	5.1 s	5.2 s	17.4 s	4.1 s
	Model parameters	22,153	533,753	184,633	106,640	130,041	827,081	102,969
Houston 2013	Running time	1.6 s	2.4 s	4.5 s	5.8 s	3.8 s	9.5 s	3.2 s
	Model parameters	27,791	534,527	193,275	114,128	130,431	832,527	103,743
Salinas	Running time	2.7 s	3.4 s	6.2 s	7.3 s	9.8 s	23.7 s	6.7 s
	Model parameters	35,536	534,656	205,396	128,413	130,496	840,240	103,872
YRE	Running time	6.7 s	4.0 s	6.3 s	10.3 s	12.4 s	35.4 s	10.1 s
	Model parameters	45,524	535,172	169,336	153,198	130,756	850,100	104,388

time, the CNN-based methods generally require less time compared with the transformer-based methods. Specifically, the CNNHSI method has the shortest running time, ranging from 1.2 to 6.7 s, while the SPRLT method exhibits the longest running time, ranging from 6.6 to 35.4 s. Among the transformer-based methods, our CMT approach achieves the shortest running time. For example, on the YRE dataset, it reduces the running time compared with the SF, SSFTT, and SPRLT methods by 0.2, 2.3, and 25.3 s, respectively. Concerning model parameters, our CMT approach boasts relatively fewer parameters, approximately 103 500, excluding the CNNHSI method. The parameters count is significantly lower than that of the HybridSN (around 534 000 parameters) and SPRLT (around 837 000 parameters) methods. Overall, our CMT approach not only achieves high classification accuracy but also exhibits relatively lower computational complexity.

F. Generalization Performance

To illustrate the effectiveness and robustness of our CMT approach, the generalization experiments are conducted on the Indian Pines, Houston 2013, and Pavia University datasets, because the classification accuracy can be greatly improved with the increase in the number of training samples. The number of training samples for each class is set to 5, 10, 20, 30, and 50 for all the methods. Considering that the total number of samples for some of the categories in Indian Pines is smaller than the number of samples to be sampled, the experiment will keep five samples per category as test samples and the rest as training samples. The experimental results are shown in Fig. 16. It can be seen that the accuracy of all the methods improves with the increase in the number of training samples, and the model accuracy reaches the maximum when the training sample size is 50. Our approach not only has an obvious advantage in the case of small samples but also significantly outperforms the compared methods in the case of any number of samples. Some CNN-based methods have achieved excellent improvement. For example, the classification accuracy of CNNHSI and HybridSN methods on the Houston 2013 dataset is higher than 95%. Similarly, in the transformer-based methods, the SSFTT and SPRLT perform well on those three datasets, while the performance of SF is relatively poor.

V. CONCLUSION

In this article, we propose a lightweight and efficient transform-based network named CMT for HSI classification with limited training samples. Specifically, the L2GTE framework and the MCTE module are designed for the feature extraction of HSI, and the RCPT task is first proposed in combination with the MIM method, which can make full use of the unlabeled samples to learn the relationship between central ground objects and their surrounding objects and can effectively improve the classification performance in the case of small samples. To demonstrate the robustness and effectiveness of our proposed CMT approach, a series of ablation experiments and comparative experiments are conducted on the five public datasets. The experimental results show that our approach achieves a significant performance advantage in the case of small samples and outperforms other state-of-the-art CNN-based and transformer-based methods. For reproducibility, the core code of the proposed approach can be found at https://github.com/rookie-YIFAN/CMT.

REFERENCES

[1] R. N. Sahoo, S. Ray, and K. Manjunath, “Hyperspectral remote sensing of agriculture,” *Current Sci.*, vol. 108, pp. 848–859, Mar. 2015.

[2] C. Weber et al., “Hyperspectral imagery for environmental urban planning,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1628–1631.

[3] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, “A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges,” *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[4] S. H. S. Basha et al., “RCCNet: An efficient convolutional neural network for histological routine colon cancer nuclei classification,” in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1222–1227.

[5] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Apr. 2013.

[6] H. Z. M. Shafri, A. Suhaili, and S. Mansor, "The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis," *J. Comput. Sci.*, vol. 3, no. 6, pp. 419–423, Jun. 2007.

[7] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2017.

[8] M. Chi, R. Feng, and L. Bruzzone, "Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem," *Adv. Space Res.*, vol. 41, no. 11, pp. 1793–1799, 2008.

[9] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[10] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 192–195, Apr. 2005.

[11] S. Moussaoui et al., "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomputing*, vol. 71, nos. 10–12, pp. 2194–2208, Jun. 2008.

[12] S. Jia et al., "Flexible Gabor-based superpixel-level unsupervised LDA for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10394–10409, Dec. 2021.

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

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

[15] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[16] S. Jia, L. Shen, and Q. Li, "Gabor feature-based collaborative representation for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 1118–1129, Feb. 2015.

[17] S. Jia et al., "Gradient feature-oriented 3-D domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5505517.

[18] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[19] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, "Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1638–1652, May 2019.

[20] 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.

[21] 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.

[22] S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, Jun. 2019.

[23] S. Zhang, M. Xu, J. Zhou, and S. Jia, "Unsupervised spatial–spectral CNN-based feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524617.

[24] 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.

[25] 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, Jul. 2017.

[26] 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.

[27] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral–spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, p. 1330, Dec. 2017.

[28] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral–spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, Feb. 2019.

[29] W.-S. Hu, H.-C. Li, L. Pan, W. Li, R. Tao, and Q. Du, "Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4237–4250, Jun. 2020.

[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[31] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[32] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jan. 2015.

[33] 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.

[34] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.

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

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[37] Z. Zhong, J. Li, L. Ma, H. Jiang, and H. Zhao, "Deep residual networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 1824–1827.

[38] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6307–6315.

[39] 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.

[40] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, p. 1068, Jul. 2018.

[41] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[42] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[43] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[44] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[45] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[46] J. Feng et al., "Attention multibranch convolutional neural network for hyperspectral image classification based on adaptive region search," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5054–5070, Jun. 2021.

[47] 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.

[48] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

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

[50] 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.

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

[52] 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.

[53] 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.

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

[55] 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.

[56] B. Tu, X. Liao, Q. Li, Y. Peng, and A. Plaza, "Local semantic feature aggregation-based transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536115.

[57] J. Zou, W. He, and H. Zhang, "LESSFormer: Local-enhanced spectral–spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535416.

[58] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532117.

[59] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[60] W. Qi, C. Huang, Y. Wang, X. Zhang, W. Sun, and L. Zhang, "Global–local 3-D convolutional transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510820.

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

[62] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[63] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[64] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised visual transformers," 2021, *arXiv:2104.02057*.

[65] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[66] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

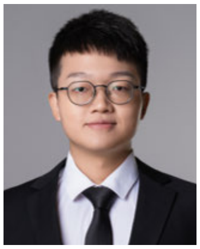
[67] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

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

[69] X. Zhang, S. Shang, X. Tang, J. Feng, and L. Jiao, "Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5507714.

Sen Jia (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include remote sensing image processing, signal and image processing, and machine learning.

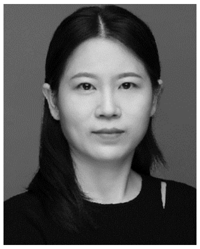


Yifan Wang received the B.S. degree in computer science and technology from Guangdong University of Technology, Guangzhou, China, in 2019, and the M.E. degree in computer technology from Shenzhen University, Shenzhen, China, in 2023.

His research interests include hyperspectral image processing, data analysis, and deep learning.

Shuguo Jiang received the B.E. degree from Xiamen University of Technology, Xiamen, China, in 2020, and the M.E. degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China.

His research interests include remote sensing classification and multimodalities' interpretation.



Ruyan He received the B.S. degree from Liaoning Technical University, Fuxin, China, in 2012, and the Ph.D. degree in photogrammetry and remote sensing from the China University of Mining and Technology at Beijing, Beijing, China, in 2019.

She was a Visiting Scholar with the University of California (UC) at Davis, Davis, CA, USA, from 2015 to 2017. She is currently an Associate Research Fellow with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include remote sensing image processing and deep learning.