Self- and Cross-Attention Enhanced Transformer for Visible and Thermal Infrared Hyperspectral Image Classification

Enyu Zhao, *Member, IEEE*, Yongfang Su, Nianxin Qu, *Student Member, IEEE*, Yulei Wang, *Member, IEEE*, Caixia Gao, *Member, IEEE*, and Jian Zeng

***Abstract*—Visible hyperspectral image (V-HSI) and thermal infrared hyperspectral image (TI-HSI) have been crucial data sources for land cover classification. V-HSI can directly provide information of land surface, such as shape, color, texture, and other features. TI-HSI contains rich long-wave spectral information, which can reflect the unique emission characteristics of ground objects in the thermal infrared spectral range. To fully leverage the advantages of V-HSI and TI-HSI while enhancing the classification accuracy, this article proposes a self- and cross-attention enhanced transformer network (SCAET), integrated with convolutional neural network (CNN) for HSI classification. Initially, the proposed method employs a dual-branch spatial-spectral CNN (SS CNN) to extract spectral convolution features from V-HSI and TI-HSI, respectively. Subsequently, a spectral feature mapping (SFM) module is proposed to perform feature transformation, extracting independent and interactive features of V-HSI and TI-HSI. Then, a self- and cross-attention interactive enhancement module is designed to extract deeper features and enhance the independent features by using the interactive features. In addition, a self-projection mixing module is formulated to promote feature interaction and improve the generalization capability of the model. To validate the effectiveness of the proposed network, extensive experiments are conducted on real-world datasets, and the results indicate that SCAET significantly outperforms current multisource fusion networks.**

***Index Terms*—Convolutional neural network (CNN), image classification, thermal infrared hyperspectral image (TI-HSI), transformer, visible hyperspectral image (V-HSI).**

I. Introduction

HYPERSPECTRAL image (HSI) consists of numerous spectral bands, offering highly precise and continuous spectral information for each individual pixel. The nanometer-scale spectral resolution enables the differentiation of ground objects based on variations in their spectral curves. Consequently, HSI classification has found widespread applications in various fields including environmental monitoring [1], mineral exploration [2], precision agriculture, and urban planning [3], [4], [5], [6].

The visible hyperspectral image (V-HSI) serves as a widely used data source for HSI classification, as it offers intricate details regarding shape and texture [7], [8]. However, the acquisition of V-HSI is heavily contingent upon the light and atmospheric conditions. The imaging efficiency of optical equipment significantly decreases in dark weather conditions, thereby compromising the reliability and accuracy of the captured spectral diagnostic information [9], [10]. Moreover, the spectral information in HSI may not accurately reflect object categories due to the variations in spectra curves within the same category and identical spectra curves corresponding to the objects from different categories. Therefore, relying solely on single-source V-HSI is insufficient for reliable land cover classification [11], [12], [13].

Currently, the integration of multisource remote sensing (RS) data are being explored to enhance classification performance by jointly utilizing the HSI information acquired from different sensors for more accurate and comprehensive earth observation [14]. The thermal infrared hyperspectral image (TI-HSI) serves as a supplementary data source for land cover classification due to its ability to provide thermal radiation emitted by objects and complement the reflectance information from the V-HSI with emissivity information [15], [16]. The combination of these two types of data effectively associates their respective spectral features, allowing for a more complete and comprehensive representation of target characteristics and ultimately improving the classification performance [17], [18], [19], [20].

With the exponential increase in data volume and the growing demand for the classification accuracy, traditional machine-learning methods are increasingly revealing their limitations [21], [22], [23]. Consequently, researchers are increasingly turning to deep learning techniques to learn and extract more complex feature information from HSI, which has the potential to significantly enhance classification accuracy [24], [25]. A multitude of deep learning networks have been employed in HSI classification, including convolutional neural networks (CNNs) and transformer architectures.

Received 10 February 2025; revised 9 April 2025 and 5 May 2025; accepted 15 May 2025. Date of publication 16 May 2025; date of current version 4 June 2025. This work was supported by National Nature Science Foundation of China under Grant 42271355. *(Corresponding authors: Yulei Wang; Jian Zeng.)*

Enyu Zhao, Yongfang Su, Nianxin Qu, and Yulei Wang are with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 11-6026, China (e-mail: zhaoenyu@dlmu.edu.cn; 18009241387syf@dlmu.edu.c-n; qvnx@dlmu.edu.cn; wangyulei@dlmu.edu.cn).

Caixia Gao is with the Key Laboratory of Quantitative Remote Sensing Information Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: gaocaixia@aoe.ac.cn).

Jian Zeng is with the China Centre for Resources Satellite Data and Application, Beijing 100094, China (e-mail: zengjian@spacechina.com).

Digital Object Identifier 10.1109/JSTARS.2025.3571226

CNN is one of the most widely utilized deep learning networks for HSI classification [26], [27]. By extracting local features and spatial information from HSI through convolution operations, subtle spectral and spatial differences can be effectively distinguished, resulting in accurate classification outcomes [28]. Hu et al. [29] proposed a CNN classification network that employed 1D CNN model on each spectral feature convolution layer to effectively extract the important spectral features. Zhao and Du [30] presented a 2D CNN model that first reduced the dimensionality of HSI before extracting both spectra and spatial features for classification purposes. Chen et al. [31] introduced an innovative 3D CNN approach to effectively leverage the spectral and spatial inherent in HSI achieving precise results in HSI classification.

Furthermore, CNN has made substantial advancements in multisource data classification research by employing two distinct branches to extract features from multisource RS data for effective classification [32], [33], [34]. Xu et al. [35] developed a novel dual-branch deep CNN architecture, the hyperspectral branch utilized dual-channel CNN to capture both spatial and spectral information, while the other branch implemented a cascaded network to derive the spatial information from LiDAR data. Fang et al. [36] proposed a spatial-spectral enhancement network that incorporated a spatial enhancement module (SAEM) and a spectral enhancement module (SEEM) into a dual-branch CNN framework, thereby enhancing the feature interaction among multisource RS data. Hong et al. [37] developed a depth encoder–decoder network aimed at fusing HSI and LiDAR data more compactly, which achieved feature fusion by reconstructing multimodal inputs.

However, CNN typically employs a fixed-size convolution kernel to convolve local regions, which imposes limitations on the ability to model long-distance dependencies. Furthermore, the convolution operation in CNN is not sensitive to the noise [38], often resulting in the extraction of inaccurate features. Different from CNN, transformer has a strong global dependency modeling capability, which can calculate the correlation between any two positions in the sequence through the attention mechanism, enhancing valuable features while reducing the influence of noise. In addition, the attention mechanism in the transformer facilitates parallel computation, making it easier to implement parallel processing and reduce processing time [39].

At present, transformer models have demonstrated excellent performance in the field of cross modal HSI classification. Xue et al. [40] developed a method that captured multimodal features and contextual information through deep structures within the vision transformer, effectively integrating features from both HSI and LiDAR data. Zhang et al. [41] employed a multimodal Transformer network (MTNet) to extract spatial and elevation information from HSI and LiDAR data, subsequently integrating these feature tokens into a new transformer encoder to explore shared spatial features more comprehensively. He et al. [42] proposed a multilevel attention dynamic-scale network, introducing a global-local cross-attention module. This module effectively reduces redundant information in long-range feature interactions by employing a distance-weighted operator, enabling deep-level multimodal feature integration. Tang et al. [43] presented a multiple information collaborative fusion network (MICF-Net), which leveraged spatial relationships and high-level semantic information from multimodal data to guide the feature extraction process. They innovatively developed a dual-branch cross-modal attention fusion Transformer (CMAFT). Wang et al. [44] proposed a convolutional interaction transformer network (CITNet) for the joint classification of hyperspectral and LiDAR data. The network designed a local-global transformer to extract joint local–global features from multimodal data, while also developing an optimized convolutional cross-attention module to achieve more efficient feature interaction.

The LiDAR data primarily provide spatial structural information with relatively low feature dimensions, exhibiting significant modal differences from the spectral information in HSI. Consequently, fusion strategies for LiDAR and HSI tend to focus on cross-modal feature collaboration and alignment optimization. In contrast, TI-HSI and V-HSI both belong to spectral imaging modalities, sharing higher spectral dimensionality and stronger spectral correlations. The high-dimensional nature presents a dual effect where it enables the extraction of richer spectral information with greater analytical potential while simultaneously risking feature redundancy and substantially increased computational complexity. Consequently, when fusing TI-HSI and V-HSI data, the fundamental challenge lies in developing fusion strategies that can effectively reduce redundancy while fully exploiting complementary information between the two modalities.

While transformer-based approaches have achieved great success, the existing models still have some limitations in feature interaction. Some methods overly rely on multilayer encoder stacks, which can enhance feature abstraction ability but easily lead to local information loss and significantly increase computational complexity. Other methods directly perform global attention calculations on multisource features during feature interaction, completely ignoring the independence of different modal features, resulting in serious information interference between modalities. In response to these issues, this article proposes an innovative approach to dynamically learn the interaction relationships between modalities by explicitly modeling independent and interactive features in multimodal data, and combining self- and cross-attention mechanisms. This method effectively preserves the independent information of each modality and achieves cross-modal context aware fusion, achieving significant improvements in information integrity and interaction flexibility.

Based on the aforementioned analysis, a novel network that integrates CNN and transformer for classification tasks involving V-HSI and TI-HSI is proposed. This model commences with employing a dual-branch spatial-spectral CNN (SS CNN) to extract shallow spectral convolution features of V-HSI and TI-HSI, respectively. To enhance the feature interaction between V-HSI and TI-HSI, a spectral feature mapping (SFM) module is implemented to the spectral convolution features to derive the independent and interactive tokens of V-HSI and TI-HSI. Subsequently, these independent and interactive features are enhanced through the self- and cross-attention interactive enhancement (SCIE) module. Finally, a self-projection mixing (SPM) module is designed to achieve the further feature interaction. The primary contributions of this article can be summarized as follows.

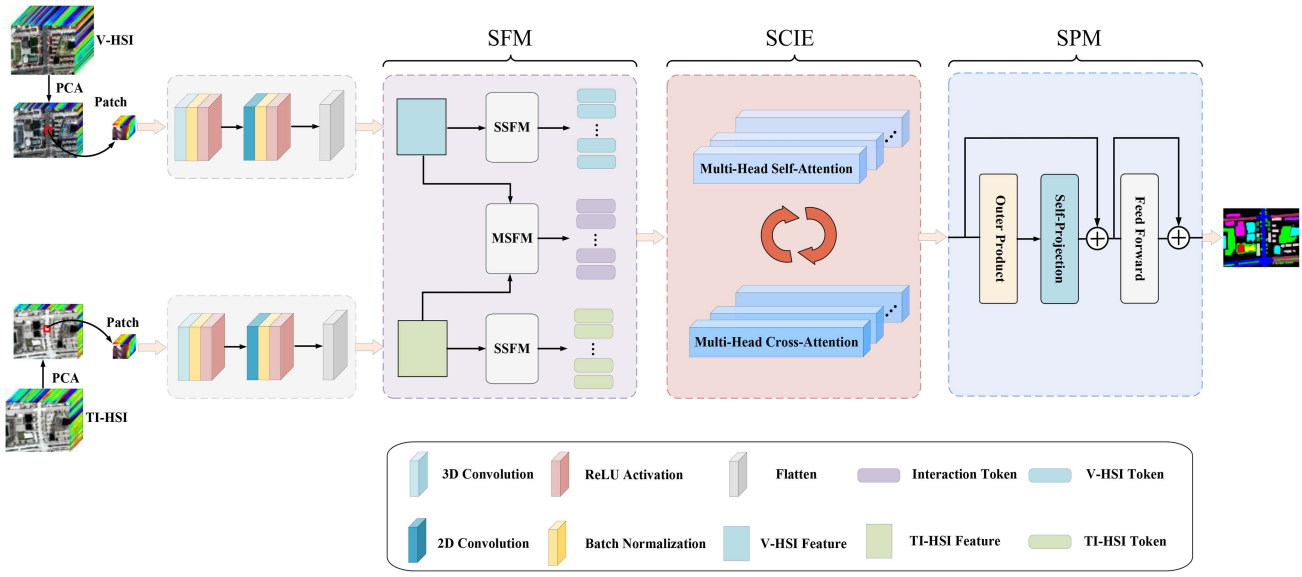


Fig. 1. Overall structure of the proposed SCAET.

1) To comprehensively extract the spectral features of V-HSI and TI-HSI, a dual-branch SS CNN is designed to operate on both types of HSIs independently, which can maintain the independence and integrity of spectral features and reduce redundant information.
2) To facilitate the feature interaction between V-HSI and TI-HSI, the SFM module is developed, which comprises two components: the single-source SFM (SSFM) module and the multisource SFM (MSFM) module. These modules are used to extract the independent and interactive feature tokens from both V-HSI and TI-HSI. Subsequently, the SCIE module is utilized to enhance these independent and interactive features.
3) To improve generalization and feature representation capabilities, the proposed method introduces an efficient SPM module that effectively integrates enhanced features from both V-HSI and TI-HSI.

The rest of this article is organized as follows: Section II describes the proposed self- and cross-attention enhanced transformer network (SCAET) model. Section III introduces the datasets used and experiments analysis. Section IV concludes the article.

II. Methodology

The SCAET model comprises four stages, as shown in Fig. 1. In the initial stage, V-HSI and TI-HSI are input into a dual-branch SS CNN, respectively. The second stage employs the SFM module to extract independent and interactive feature tokens of V-HSI and TI-HSI. During the third stage, these feature tokens are processed through the SCIE module for interaction and enhancement. Finally, the SPM module is utilized to integrate the features prior to their application in classification.

A. V-HSI and TI-HSI Data Processing

The dimension of V-HSI is denoted by $X_V \in \mathbb{R}^{H\times W\times C}$, and TI-HSI is denoted by $X_T \in \mathbb{R}^{H\times W\times D}$, where H and W represent the height and width of the spatial size, and C and D are the number of spectral bands. To reduce the computational demands and minimize the data redundancy, principal component analysis (PCA) is applied independently to V-HSI and TI-HSI, respectively [45], thereby extracting the most representative bands on a global scale. PCA demonstrates an effective dimension reduction capability while preserving the spatial structure's dimensionality. Afterward, the dimensions of V-HSI and TI-HSI are transformed into $X_V^{PCA} \in \mathbb{R}^{H\times W\times B}$ and $X_T^{PCA} \in \mathbb{R}^{H\times W\times B}$, where B is the number of bands. Subsequently, each pixel is segmented into patches, with each patch encompassing the spectral information of the central pixel along with its adjacent pixels. The dimensions of the segmented patches are $X_V^{\text{patch}} \in \mathbb{R}^{s\times s\times B}$ and $X_T^{\text{patch}} \in \mathbb{R}^{s\times s\times B}$, where $s \times s$ represents the size of the patch. All patches of V-HSI and TI-HSI form two sets Γ_V and Γ_T, respectively. The training set D_{train} and testing set D_{test}, which are selected from sets Γ_V and Γ_T, are obtained by using the following formula:

$$D_{\text{train}} = \left\{ \left(x_v^{\text{patch}}, x_t^{\text{patch}} \right)^{(i)}, y^{(i)} | i = 1, \ldots, n_1 \right\} \quad (1)$$

$$D_{\text{test}} = \left\{ \left(x_v^{\text{patch}}, x_t^{\text{patch}} \right)^{(j)}, y^{(j)} | j = 1, \ldots, n_2 \right\} \quad (2)$$

where x_v^{patch} and x_t^{patch} are randomly selected from the sets Γ_V and Γ_T, respectively, n_1 and n_2 represent the number of training and testing samples, and y represents the land cover category label.

B. Dual-Branch SS CNN

The proposed method uses SS CCN to extract spectral convolution features from V-HSI and TI-HSI, respectively. First, a Conv3D layer with a kernel size of $3 \times 3 \times 3$ is utilized for V-HSI, facilitating information exchange across different channels. Subsequently, a Conv2D layer with a kernel size of 3×3 is applied to capture detailed spectral features. The SS

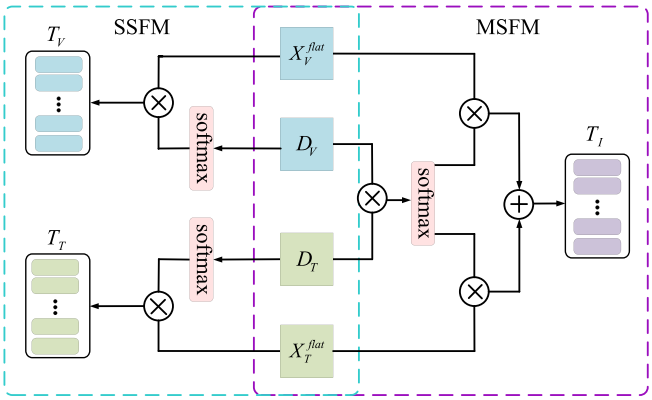


Fig. 2. Structure of SFM module.

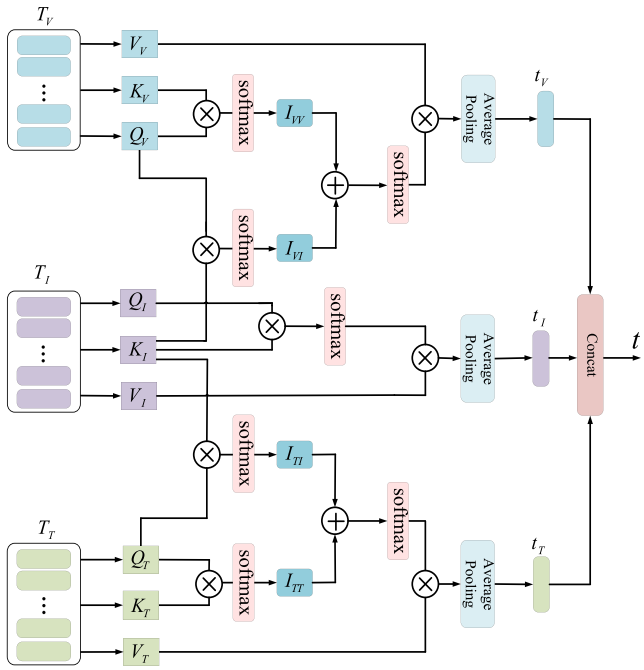


Fig. 3. Structure of SCIE module.

CNN extracts both spectral and spatial information from HSI by employing convolution kernels of varying dimensions, thereby enabling shallow feature extraction. Following the convolution layer, batch normalization is implemented for feature scaling, after which a rectified linear unit activation function is employed to enhance the nonlinear representation capabilities of the extracted features.

The convolution features $X_V^{\text{conv}} \in \mathbb{R}^{m\times n\times z}$ are derived from SS CNN applied to the V-HSI, where m and n represent the height and width of the convolution feature maps, respectively, and z denotes the number of feature channels. The convolution features X_V^{conv} undergoes a flattening operation to yield $X_V^{\text{flat}} \in \mathbb{R}^{mn\times z}$, X_V^{conv}, and X_V^{flat} can be expressed using the following formula:

$$X_V^{\text{conv}} = \text{Conv2D}\left(\text{Conv3D}\left(X_V^{\text{patch}}\right)\right) \tag{3}$$

$$X_V^{\text{flat}} = \text{Flatten}\left(X_V^{\text{conv}}\right). \tag{4}$$

For TI-HSI, the method also employs Conv3D layer with a $3 \times 3 \times 3$ convolution kernel and Conv2D layer with 3×3 convolution kernel. The resulting convolution features $X_T^{\text{conv}} \in \mathbb{R}^{m\times n\times z}$ are then flattened to facilitate $X_T^{\text{flat}} \in \mathbb{R}^{mn\times z}$, which are performed using the following equations:

$$X_T^{\text{conv}} = \text{Conv2D}\left(\text{Conv3D}\left(X_T^{\text{patch}}\right)\right) \tag{5}$$

$$X_T^{\text{flat}} = \text{Flatten}\left(X_T^{\text{conv}}\right). \tag{6}$$

C. Spectral Feature Mapping

As illustrated in Fig. 2, the proposed SFM module comprises SSFM and MSFM submodules, which are designed to extract independent and interactive feature tokens of V-HSI and TI-HSI based on spectral convolution features.

The SSFM submodule is employed to extract the independent feature tokens of V-HSI and TI-HSI. Two learnable parameter matrices, denoted as W_V^a and W_T^a, are utilized to perform linear transformations on the flattened spectral convolution features X_V^{flat} and X_T^{flat} resulting in the matrices D_V and D_T. These matrices are subsequently converted into weight matrices through the application of the softmax function. The resulting weight matrices are then multiplied by their corresponding flattened spectral convolution features X_V^{flat} and X_T^{flat} to derive the independent feature tokens $T_V \in \mathbb{R}^{w\times z}$ and $T_T \in \mathbb{R}^{w\times z}$ of V-HSI and TI-HSI, where w represents the number of tokens. The following equations can be used to calculate T_V and T_T:

$$T_V = \text{softmax}(D_V)X_V^{\text{flat}} \tag{7}$$

$$T_T = \text{softmax}(D_T)X_T^{\text{flat}}. \tag{8}$$

The MSFM submodule is employed to extract the interactive feature tokens of V-HSI and TI-HSI. The softmax function is utilized to transform the dot product of matrices D_V and D_T into a weight matrix. Subsequently, the weight matrix is multiplied by the flattened spectral features X_V^{flat} and X_T^{flat}, respectively, and summed together. The resulting interactive feature tokens $T_I \in \mathbb{R}^{w\times z}$ comprehensively integrate the interactive feature information of V-HSI and TI-HSI. The specific formula can be expressed as follows:

$$T_I = \text{softmax}(D_V D_T)X_V^{\text{flat}} + \text{softmax}(D_V D_T)X_T^{\text{flat}}. \tag{9}$$

D. Self- and Cross-Attention Interactive Enhancement

Self- and cross-attention mechanisms are essential methods for feature-level fusion within the transformer architecture. The transformer employs the attention mechanisms to compute the weights between each input position and all other positions, thereby integrating the global information [46]. Self-attention mechanism is used to address the relationships among positions within a single sequence, while the cross-attention mechanism focuses on capturing the relationships between two different sequences [47].

As shown in Fig. 3, the proposed module integrates the self-and cross-attention mechanisms to fully leverage their respective advantages. This approach enhances both independent and interactive features while facilitating the extraction of deeper feature information. For the three sets of the extracted tokens T_V, T_T, and T_I, used by the SFM module, we define three distinct learnable weight matrices W^q, W^k, and W^v. These learnable weight matrices are employed to map the tokens into three corresponding matrices, which include query Q, key K, and value V.

The proposed SCIE module uses the self-attention mechanism to enhance the interaction features T_I. First, the query matrix Q_I is multiplied by the transposed key matrix K_I corresponding to T_I. Subsequently, the softmax function is applied to convert the attention scores into attention weights. These attention weights are then multiplied with the value matrix V_I, and the specific formula is as follows:

$$T_I^w = \text{softmax}\left(\frac{Q_I K_I^T}{\sqrt{d_I}}\right) V_I \tag{10}$$

where d_I represents the dimension of K_I. T_I^w denotes the weighted version of T_I.

In order to facilitate the interaction between independent and interactive feature tokens, the SCIE module employs interactive feature tokens to enhance the independent feature tokens of V-HSI and TI-HSI. First, the self-attention mechanism is applied to calculate the dot product of the Q_V matrix with the transposed K_V matrix corresponding to T_V, followed by the application of the softmax function to convert these results into attention weight S_{VV}. Subsequently, another application of the softmax function transforms the attention score, which is derived from the dot product of the query matrix Q_V corresponding to T_V and the transposition of key matrix K_I into the attention weight S_{VI}. Next, S_{VV} and S_{VI} are multiplied by two learnable interactive weight matrices I_{VV} and I_{VI} for weighted aggregation. The resulting values undergo another round of softmax function processing to calculate attention weights that further enhance independent features. Finally, the weighted feature token T_V^w is obtained by multiplying attention weights with the value matrix V_V. Overall, this procedure can be summarized as follows:

$$S_{VV} = \text{softmax}\left(\frac{Q_V K_V^T}{\sqrt{d_V}}\right) \tag{11}$$

$$S_{VI} = \text{softmax}\left(\frac{Q_V K_I^T}{\sqrt{d_V}}\right) \tag{12}$$

$$T_V^w = \text{softmax}\left(\left(S_{VV} \odot I_{VV}\right) + \left(S_{VI} \odot I_{VI}\right)\right) V_V \tag{13}$$

where $\odot$ represents element-wise multiplication, and d_V represents the dimension of K_V.

Utilizing the similar steps mentioned above can enhance the independent features T_T of the TI-HSI based on self- and cross-attention, the weighted feature token T_T^w is computed through the following procedure:

$$S_{TT} = \text{softmax}\left(\frac{Q_T K_T^T}{\sqrt{d_T}}\right) \tag{14}$$

$$S_{TI} = \text{softmax}\left(\frac{Q_T K_I^T}{\sqrt{d_T}}\right) \tag{15}$$

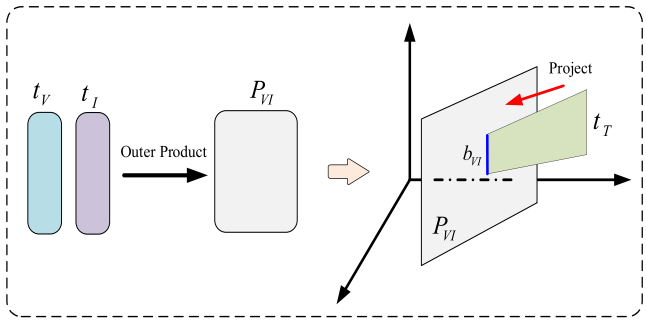


Fig. 4. Structure of SPM module.

$$T_T^w = \text{softmax}\left(\left(S_{TT} \odot I_{TT}\right) + \left(S_{TI} \odot I_{TI}\right)\right) V_T \tag{16}$$

where the values Q_T, K_T, and V_T represent the query, key, and value matrices corresponding to T_T. The matrix S_{TT} denotes the attention weights derived from self-attention applied to T_T. Meanwhile, S_{TI} refers to the weight matrix obtained through cross-attention between T_T and T_I. In addition, I_{TT} and I_{TI} are the interaction weight matrices. Finally, d_T represents the dimension of K_T.

The independent and interactive three groups feature tokens of V-HSI and TI-HSI are pooled averagely to obtain vectors$t_V \in \mathbb{R}^{1\times z}$, $t_T \in \mathbb{R}^{1\times z}$, and $t_I \in \mathbb{R}^{1\times z}$. Subsequently, they are concatenated into vector$t \in \mathbb{R}^{3\times z}$ through the concatenation operation. The formula is as follows:

$$t_V = \text{AP}\left(T_V^w\right), t_T = \text{AP}\left(T_T^w\right), t_I = \text{AP}\left(T_I^w\right)) \tag{17}$$

$$t = \text{Concat}\left(t_V, t_T, t_I\right) \tag{18}$$

where AP is average pooling function.

E. SPM and Classification

As illustrated in Fig. 4, the proposed method employs an SPM module to integrate vector t, which encompasses critical classification features. Subsequently, the linear layers are connected for the purpose of classification.

The vectors t_V, t_I, and t_T undergo outer product with each other to obtain matrices $P_{VI} \in \mathbb{R}^{z\times z}$, $P_{IT} \in \mathbb{R}^{z\times z}$, and $P_{TV} \in \mathbb{R}^{z\times z}$. The calculation process can be expressed as

$$P_{VI} = t_V \otimes t_I, P_{IT} = t_I \otimes t_T, P_{TV} = t_T \otimes t_V \tag{19}$$

where $\otimes$ represents the outer product operation.

The SPM module then utilizes the vector t_T to project onto the matrix P_{VI}, which is derived from the outer product of the other two vectors t_V and t_I. The obtained projection vector $b_{VI} \in \mathbb{R}^{1\times z}$ can be calculated according to the following formula:

$$b_{VI} = P_{VI}\left(P_{VI}{}^T P_{VI}\right)^{-1} P_{VI}{}^T t_T \tag{20}$$

Similarly, the projection vectors $b_{IT} \in \mathbb{R}^{1\times z}$ and $b_{TV} \in \mathbb{R}^{1\times z}$ can be obtained through analogous steps. By applying the outer product and projecting vectors, the relationships involving vector t can be effectively captured to achieve a mixed effect. Subsequently, the projection vectors b_{VI}, b_{IT}, and b_{TV} are concatenated into a single vector $b \in \mathbb{R}^{3\times z}$ using the concatenation

Algorithm 1: SCAET.

Input: input V-HSI is $X_V \in \mathbb{R}^{H\times W\times C}$, TI-HSI is $X_T \in \mathbb{R}^{H\times W\times D}$ and a ground truth is $Y \in \mathbb{R}^{H\times W}$.
Output: Classification results of testing dataset and visualization map.
1 Obtain X_V^{PCA}, X_T^{PCA} after PCA dimensionality reduction for V-HSI and TI-HSI, respectively.
2 Perform patch segmentation separately to obtain X_V^{patch} and X_T^{patch}.
3 Divide the data into a training set D_{train} and a testing set D_{test}.
4 **for** epoch in range (epochs)**:**
5 Perform dual-branch SS CNN for V-HSI and TI-HSI patches.
6 The SFM module is used to extract V-HSI and TI-HSI independent and interactive tokens T_V, T_T, and T_I.
7 Perform SCIE module to enhance interactive and independent features, then use average pooling on the tokens to obtain vector *t*.
8 Input the vector *t* to the SPM for features interaction to obtain vector *b*.
9 Connect the linear layer.
10 Use the softmax function to identify the labels.
11 **end for**
12 Use the testing set with the trained model to get predicted labels.

operation. The formulaic expression of the above operation is given as follows:

$$b = \text{Concat}(b_{VI}, b_{IT}, b_{TV}). \tag{21}$$

The entire process involves multilayer transfer and residual connection. Finally, the vector *b* undergoes an average pooling function, followed by a connection to a linear layer, ultimately utilizing a softmax function for classification. This entire process can be succinctly represented as

$$\text{Lable} = \text{softmax}(\text{Liner}(\text{AP}(b))). \tag{22}$$

III. Experiments and Analysis

This section introduces the V-HSI and TI-HSI datasets utilized in this study and provides comprehensive details regarding the implementation process. In addition, extensive experiments are conducted, including parameter sensitivity analyses and ablation studies, to evaluate the performances of the SCAET model. Furthermore, a comparison is made between the SCAET model and a variety of advanced methods with respect to the classification performance, computational time, and number of trainable parameters.

A. Datasets

This article utilizes a collection of airborne V-HSI and TI-HSI datasets from Hengdian Town, Dongyang City, Zhejiang Province, China. The V-HSI data encompasses 230 spectral bands, covering a range from 0.44 to 1 μm, with a spatial resolution of 1 m and a spectral resolution of 2.4 nm. The TI-HSI data consists of 110 spectral bands, spanning a range of 8.06–11.22 μm, with a spatial resolution of 1 m and a spectral resolution of 28.7 nm.

TABLE I
Number of Training and Testing Samples for the Urban Dataset

Class	Class Name	Training	Testing
1	Marble	9	911
2	Grass	117	11746
3	Asphalt road	93	9312
4	Synthetic	9	947
5	Water	57	5705
6	Cement road	55	5500
7	BuildingA	36	3664
8	BuildingB	8	888
9	BungalowA	11	1123
10	BungalowB	30	3056
11	BungalowC	9	969
-	Total	434	43821

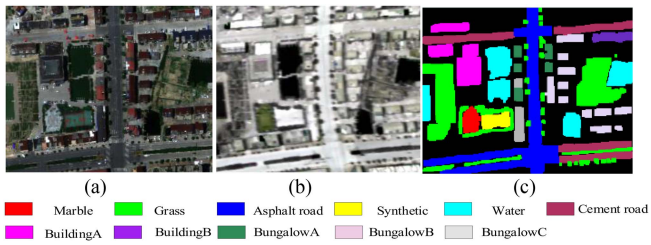


Fig. 5. Urban dataset. (a) False-color image of V-HSI. (b) False-color image of TI-HSI. (c) Ground truth.

The V-HSI and TI-HSI data mentioned above are divided into two datasets: Urban dataset and Suburb dataset. The Urban dataset primarily includes houses, roads, water areas, and grassland areas, comprising a total of 251 $\times$ 334 pixels, which translates to 44 255 sample pixels across 11 test categories. Table I presents the sample categories along with their respective quantities for both training set and testing set utilized in the experiments. Notably, only 1% of the total dataset is allocated as the training set for the Urban category. Fig. 5 illustrates the visualization outcomes of the Urban dataset, featuring false color maps for both V-HSI and TI-HSI data alongside a ground truth map.

The suburb dataset mainly includes water, grassland, and buildings, and consists 272 $\times$ 423 pixels, 72 976 sample pixels, and 16 test categories. Table II lists the sample categories and quantities of the training set and testing set applied in the experiments. The suburb dataset only uses 1% of the total as the training set. Fig. 6 shows the visualization outcomes of the suburb dataset, including the false color map of V-HSI data, the false color map of TI-HSI data, and the ground truth map.

B. Experimental Settings

This article employs three widely recognized evaluation indicators to assess the classification performance of the proposed model in comparison with existing classification models. These

TABLE II
NUMBER OF TRAINING AND TESTING SAMPLES FOR THE SUBURB DATASET

Class	Class Name	Training	Testing
1	Steel plate	4	456
2	Grass	182	18256
3	Water	232	23212
4	Dirt road	72	7278
5	Asphalt road	57	5760
6	Cement road	16	1623
7	Board	14	1455
8	Rock area	16	1619
9	BuildingA	21	2123
10	BuildingB	10	1033
11	BuildingC	9	926
12	BuildingD	13	1353
13	BungalowA	27	2773
14	BungalowB	21	2108
15	BungalowC	15	1518
16	BungalowD	7	767
-	Total	716	72260

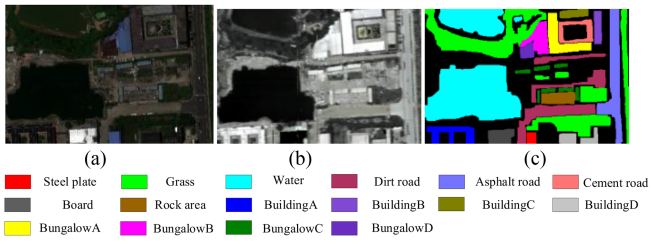


Fig. 6. Suburb dataset. (a) False-color image of V-HSI. (b) False-color image of TI-HSI. (c) Ground truth.

metrics include overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa). For each indicator, a higher value indicates superior classification performance. Each model undergoes ten repeated tests, and the average value along with the standard deviation of each test result is calculated to determine the final classification accuracy.

All experiments are conducted on the PyTorch platform, using NVIDIA GeForce RTX 3090 GPU and 24 GB VRAM. An Adam optimizer is adopted to optimize the network. In the training stage, batchsize and the number of training epoch are set to 64, and 300, respectively. The learning rate is set to 1e-3.

C. Parameter Sensitivity Analysis

This article presents a comprehensive analysis of the model's hyper-parameters, including patch size, learning rate, number of heads, and depth of SPM. Through detailed experiments, the influence of these hyper-parameters on classification performance can be examined. Ultimately, the most suitable hyper-parameters are identified and established for use in the subsequent experiments.

1) *Patch size:* The patch size is closely associated with the spatial and spectral information used by the samples in the classification process. In this study, the patch size is defined as [5], [7], [9], [11], [13], [15], [17], [19]. As illustrated in Figs. 7(a) and 8(a), when the patch sizes are set to be 13, both the urban and suburb datasets exhibit optimal classification accuracy.
2) *Learning rate:* As an important hyper-parameter in deep learning, the learning rate significantly influences the convergence behavior of the model. In the learning rate analysis experiment, the learning rate is set as [1e-2, 5e-3, 1e-3, 5e-4, 1e-4]. Subsequently, the experiments are conducted on the two datasets to evaluate and record the classification performance of the model under each specified learning rate configuration. It can be seen from Figs. 7(b) and 8(b) that the classification accuracies gradually increase with decreasing learning rate and ultimately stabilize at a value of 1e-3.
3) *The number of heads:* In the model training process, the number of heads is a crucial parameter in the multihead attention mechanism of transformer. Appropriately increasing the number of heads in attention can facilitate better information fusion across different feature spaces. The number of heads is set to [2], [4], [8], [16], [32], and the classification performance is evaluated for varying numbers of heads. As shown in Figs. 7(c) and 8(c), when the number of heads is set to 8, the model achieves optimal classification performance.
4) *The depth of SPM:* The SPM module integrates the enhanced features, and the depth of SPM significantly influences the model's feature expression capability. To determine the optimal depth of SPM, this article evaluates the depths as [2], [4], [6], [8], [10]. As illustrated in Figs. 7(d) and 8(d), it is evident that the classification performances for both the Urban dataset and Suburb dataset improve as the depths from 2 to 6 layers. The peak classification accuracy is attained at a depth of six layers. However, when the depth exceeds six layers, there is a gradual decline in classification accuracy indicating that an increased depth of SPM may not necessarily yield better performance.

D. Experimental Comparison With Competitive Methods

To validate the effectiveness and superior classification performance of the SCAET model, this article conducts a comparative analysis with several state-of-the-art HSI classification models on urban and suburb datasets. The comparison includes HSI classification models based on CNN architectures, such as S2ENet [36], EndNet [37], CCR-Net [48], MDL [49]; as well as those based on Transformer frameworks, including FusAtNet [50], DSHFNet [51], GLT-Net [52], MICF-Net [43], MTNet [41], and MFT [53].

1) In S2ENet, both the SAEM and SEEM are employed to improve the spectral features of V-HSI and TI-HSI.
2) In EndNet, identical encoder and decoder architectures are used to encode and decode the features of the V-HSI and TI-HSI. Subsequently, the input data are reconstructed from the integrated features.
3) In CCR-Net, CNN is used to extract the spectral features of V-HSI and TI-HSI, and then reconstruction strategy across modalities is utilized to facilitate feature fusion.

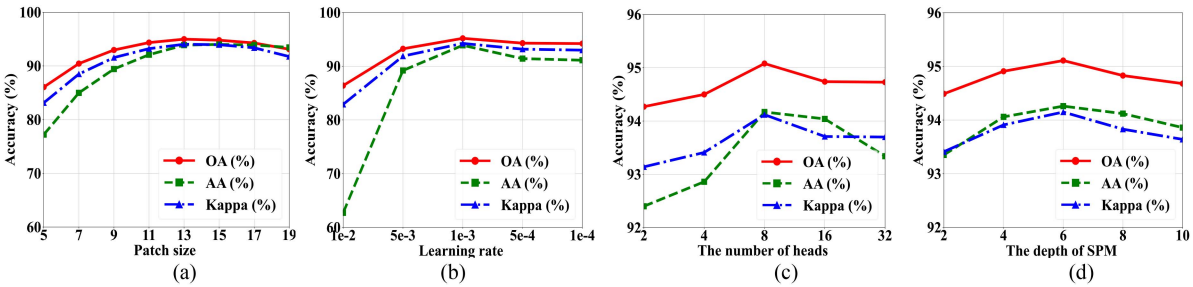


Fig. 7. Impact of hyper-parameter settings on classification performance for urban dataset. (a) Patch size. (b) Learning rate. (c) Number of heads. (d) Depth of SPM.

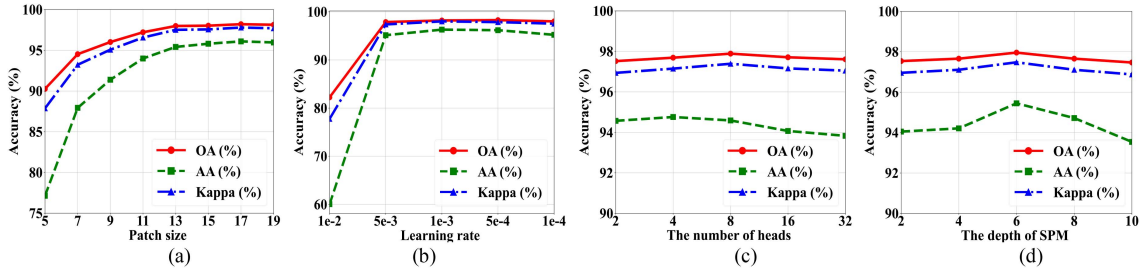


Fig. 8. Impact of hyper-parameter settings on classification performance for suburb dataset. (a) Patch size. (b) Learning rate. (c) Number of heads. (d) Depth of SPM.

4) In MDL, the fusion architecture of cross-modal learning is used to achieve feature interaction between V-HSI and TI-HSI.
5) FusAtNet leverages self- and cross-attention mechanisms to extract features of V-HSI and TI-HSI, with the combined features being applied for land cover classification.
6) DSHFNet dynamically selects features from multiscale V-HSI and TI-HSI, and then uses different attention modules to achieve hierarchical fusion.
7) GLT-Net extracts multiscale local spatial features of V-HSI and TI-HSI through CNN, models global spectral dependencies using transformer, and finally combines local and global features for joint classification, fully mining the spectral spatial complementary information of multimodal data.
8) MICF-Net uses convolutional networks to extract features from V-HSI and TI-HSI, and then uses CMAFT module to achieve feature interaction, fully mining complementary information.
9) MTNet performs a token transformation on the features of V-HSI and TI-HSI independently before employing a transformer encoder network for feature enhancement. In addition, a novel transformer encoder network is introduced to investigate the shared features across different modalities.
10) MFT converts the convolution features of V-HSI and TI-HSI extracted by CNN into tokens individually, which are then inputted into the transformer encoder network for feature interaction.

Tables III and IV present the classification performance of various classification methods applied to the urban and suburb datasets. To ensure the experimental fairness, an identical number of training samples are utilized for conducting ten training and tests for each method under investigation. The average results along with their standard deviation are recorded as the final classification outcomes. From the results displayed in the tables, it is evident that CNN models have demonstrated commendable classification performance, attributed to their robust feature representation capabilities. In CNN models, MDL and S2ENet achieve the best classification performance on the two datasets, respectively, with OA scores of 91.98% and 94.41%. Besides, owing to its global learning capacity, transformer is skilled in capturing long-range dependencies and contextual information. The classification accuracy of GLT-Net on both datasets reaches 94.03% and 97.47%, passing that of traditional CNN models. Compared with CNN models and transformer models, the proposed SCAET model demonstrates superior classification performance by effectively extracting deeper independent and interactive features for categorization. To intuitively illustrate the classification performance of various models, Figs. 9 and 10, respectively, present the ground truth and classification maps derived from different models across the two experimental datasets. The SCAET model consistently demonstrates high accuracy in all category classification tasks. Even within the urban dataset, which encompasses a diverse range of bungalows and buildings, it effectively distinguishes between different ground categories. In the suburb dataset, while the transformer models have achieved commendable results in classifying individual categories, overall, the SCAET model continues to exhibit superior classification performance.

In order to further analyze the classification performances of the SCAET model in scenarios involving small sample sizes, this section randomly selects [0.1, 0.3, 0.5, 0.7, 0.9] % samples

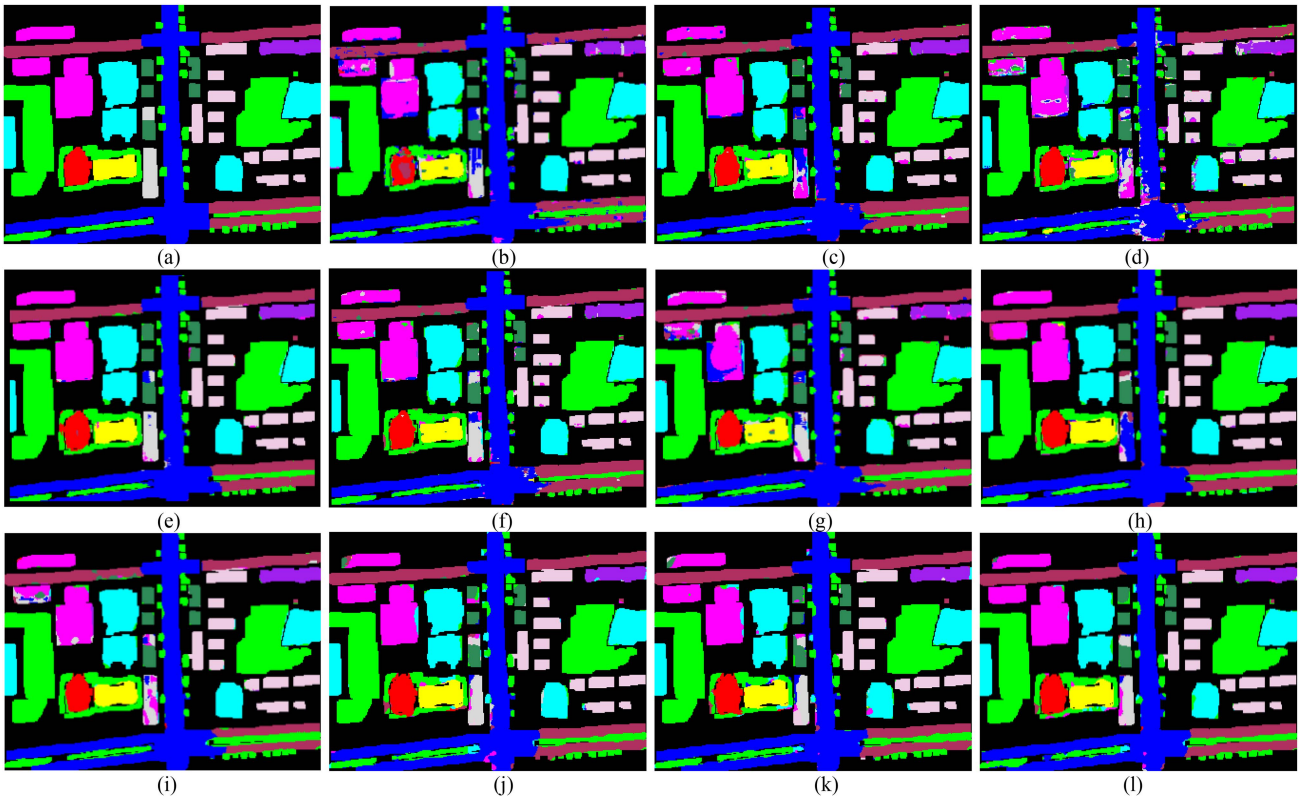


Fig. 9. Classification result maps on urban dataset. (a) Ground truth. (b) S2ENet. (c) EndNet. (d) CCR-Net. (e) MDL. (f) FusAtNet. (g) DSHFNet. (h) GLT-Net. (i) MICF-Net. (j) MTNet. (k) MFT. (l) SCAET.

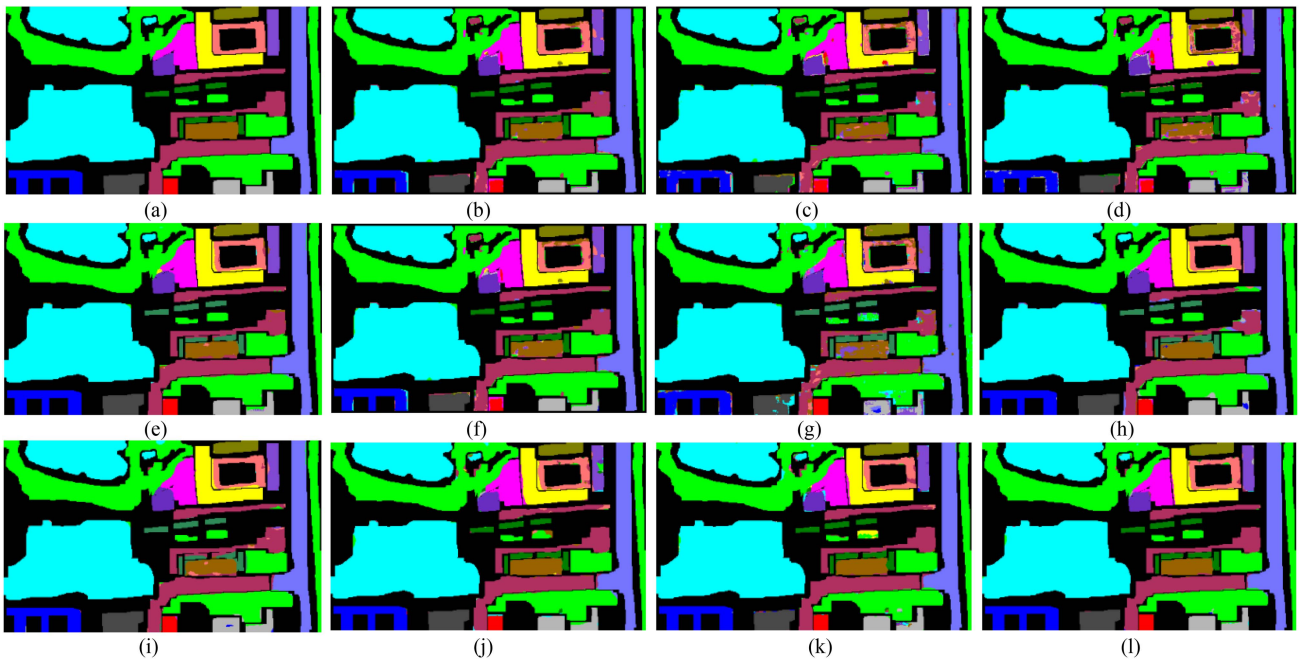


Fig. 10. Classification result maps on suburb dataset. (a) Ground truth. (b) S2ENet. (c) EndNet. (d) CCR-Net. (e) MDL. (f) FusAtNet. (g) DSHFNet. (h) GLT-Net. (i) MICF-Net. (j) MTNet. (k) MFT. (l) SCAET.

TABLE III
CLASSIFICATION ACCURACIES OF DIFFERENCE METHODS OF URBAN DATASET

Class	CNN				Transformer						
No.	S2ENet	EndNet	CCR-Net	MDL	FusAtNet	DSHFNet	GLT-Net	MICF-Net	MTNet	MFT	SCAET
1	89.08±1.17	92.10±1.26	90.98±0.07	94.24±0.38	90.15±1.17	84.38±0.41	91.73±0.03	93.85±0.54	98.61±1.35	98.85±0.81	**98.94±0.57**
2	92.86±0.09	90.88±0.58	90.79±0.13	91.89±0.15	93.12±0.22	94.39±0.08	92.05±0.09	**96.96±0.27**	91.53±1.04	90.81±1.41	91.13±0.50
3	94.93±0.06	89.92±0.80	89.49±0.06	91.58±0.18	94.76±0.11	97.52±0.05	**98.23±0.06**	95.98±1.90	95.21±0.74	97.06±0.76	97.76±0.59
4	85.45±0.98	83.68±0.40	85.77±6.87	95.29±0.36	88.66±1.17	93.09±0.03	**99.26±0.18**	95.24±0.57	98.95±0.69	97.61±1.46	99.20±0.15
5	90.26±0.05	87.13±1.91	88.94±1.60	89.88±0.04	90.60±0.12	92.35±0.02	**99.51±0.05**	98.11±0.63	96.96±0.28	96.05±1.79	98.05±0.50
6	92.40±0.46	80.59±1.28	82.41±1.60	88.76±0.02	91.67±0.42	90.91±0.54	95.75±0.14	80.27±1.76	95.64±0.62	96.41±0.38	**96.54±0.84**
7	93.57±0.17	89.57±0.73	89.92±0.24	**98.23±0.07**	92.83±0.24	66.59±0.18	92.70±0.26	88.95±0.50	92.60±0.53	91.62±1.02	94.55±0.47
8	94.56±0.37	86.68±2.99	89.42±6.64	87.39±0.01	94.71±2.70	81.85±0.06	**99.90±0.44**	99.61±0.06	91.88±3.23	94.05±3.29	93.75±3.00
9	93.70±0.08	91.26±1.18	92.26±1.88	95.82±0.20	90.38±8.83	90.65±0.01	98.14±0.01	95.72±2.01	97.91±0.49	98.40±0.68	**99.67±0.14**
10	30.22±9.17	40.29±8.95	65.13±9.23	**81.05±0.82**	64.34±9.52	56.95±0.55	54.41±0.08	63.44±3.75	69.10±3.65	64.47±3.94	68.27±3.32
11	91.17±0.44	69.29±7.90	87.39±4.35	97.43±0.09	91.78±0.83	92.81±0.85	98.80±0.03	**98.83±0.19**	88.74±2.42	95.95±1.78	95.66±1.03
OA(%)	91.35±0.27	87.40±9.55	88.79±1.08	91.98±0.01	92.24±0.26	90.51±0.34	94.03±0.04	93.44±0.04	93.70±0.15	93.17±0.25	**94.50±0.18**
AA(%)	86.13±1.07	81.87±2.78	86.59±2.64	91.57±0.01	89.50±1.16	85.59±0.03	92.57±0.06	91.52±0.80	90.46±0.61	92.34±0.57	**93.94±0.33**
Kappa(%)	89.64±0.31	84.98±1.16	86.67±1.37	90.48±0.02	90.74±0.31	88.56±0.04	92.69±0.05	92.14±0.05	92.47±0.18	92.01±0.78	**93.51±0.22**

The best performance is highlighted in bold.

TABLE IV
CLASSIFICATION ACCURACIES OF DIFFERENCE METHODS OF SUBURB DATASET

Class	CNN				Transformer						
No.	S2ENet	EndNet	CCR-Net	MDL	FusAtNet	DSHFNet	GLT-Net	MICF-Net	MTNet	MFT	SCAET
1	71.79±1.92	58.80±4.50	72.97±4.48	71.01±0.25	65.32±3.18	**99.90**±0.01	95.18±0.09	87.61±0.09	95.82±3.72	98.51±1.81	99.41±0.45
2	84.98±5.72	94.89±0.18	94.91±0.11	89.18±0.07	95.13±0.11	96.19±0.23	98.85±0.03	98.71±0.04	98.30±0.30	97.33±0.45	**99.05±0.11**
3	87.68±0.15	84.74±1.70	83.57±1.11	95.82±0.02	88.33±0.43	99.80±0.01	99.73±0.02	**99.81±0.05**	99.06±1.15	98.28±0.59	99.64±0.50
4	97.13±0.11	96.39±1.33	96.07±0.92	96.39±0.27	97.34±0.19	84.14±0.03	92.47±0.29	96.11±0.71	96.44±1.06	95.81±0.93	**97.46±0.15**
5	96.89±0.07	96.60±0.25	96.73±0.09	95.41±0.14	96.71±0.04	98.05±0.02	98.07±0.11	98.24±0.01	99.56±0.06	**99.63±0.07**	99.56±0.05
6	**93.16±0.26**	88.69±1.22	89.75±1.68	82.22±0.31	92.37±0.15	62.72±0.12	85.76±0.17	76.08±0.09	91.18±1.70	89.76±0.84	91.29±0.14
7	95.39±0.08	92.33±0.76	90.02±0.78	80.60±0.00	96.05±0.42	86.79±0.16	**99.79±0.02**	99.35±0.44	97.05±0.43	96.16±1.78	99.23±0.19
8	95.97±0.18	95.25±0.83	95.72±0.22	92.62±1.84	96.41±0.12	78.53±0.32	92.44±0.08	75.63±0.37	**98.44±0.28**	98.11±0.46	98.41±0.21
9	90.39±1.06	87.15±2.92	95.25±1.30	86.66±0.01	88.66±0.40	90.81±0.05	98.15±0.11	**98.92±0.04**	88.02±7.44	88.38±2.96	91.81±2.50
10	**98.81±0.08**	94.14±2.26	98.27±0.24	95.75±0.34	98.57±0.38	95.30±0.12	98.63±0.08	92.42±0.42	97.28±1.27	96.95±2.16	98.29±0.36
11	90.37±0.03	85.83±5.25	88.75±0.81	71.51±0.13	90.49±0.76	93.47±0.03	99.47±0.02	99.68±0.11	**100.0±0.00**	**100.0±0.00**	99.67±0.56
12	83.86±1.44	38.64±9.64	68.78±8.41	59.20±0.12	72.77±5.40	56.66±1.20	**95.75±0.12**	89.45±0.46	79.44±3.20	81.94±4.30	81.75±2.19
13	78.40±0.29	66.01±2.30	71.02±1.40	**99.98±0.04**	78.14±0.28	95.78±0.03	99.45±0.01	99.68±0.07	90.85±1.68	87.57±3.20	95.31±1.52
14	90.95±0.85	79.20±7.30	88.05±1.70	88.12±0.28	90.01±1.50	83.88±0.09	88.87±0.06	88.82±0.99	**92.76±2.97**	91.21±2.77	92.63±1.68
15	85.50±0.64	79.93±3.03	78.39±2.50	90.00±0.38	85.55±0.99	99.02±0.01	90.56±0.02	81.86±0.03	**99.61±0.21**	98.30±1.38	99.58±0.28
16	89.24±0.77	73.14±5.48	78.50±5.0	**98.41±0.27**	84.28±0.51	94.83±0.07	98.23±0.01	96.70±0.71	94.07±1.88	92.49±3.45	92.43±0.57
OA(%)	94.41±0.04	91.34±0.27	92.69±0.45	91.90±0.01	94.06±0.14	93.71±0.05	97.47±0.03	96.81±0.6	97.48±0.25	97.01±0.27	**98.14±0.19**
AA(%)	90.03±0.04	80.55±2.04	85.66±2.18	86.42±0.02	88.52±0.55	88.50±0.07	**95.71±0.04**	92.44±0.23	94.86±0.83	94.43±0.36	95.07±0.47
Kappa(%)	93.15±0.05	89.38±0.34	91.03±0.57	90.14±0.01	92.72±0.17	92.23±0.06	96.87±0.04	96.06±0.75	96.88±0.32	96.31±0.34	**97.70±0.23**

The best performance is highlighted in bold.

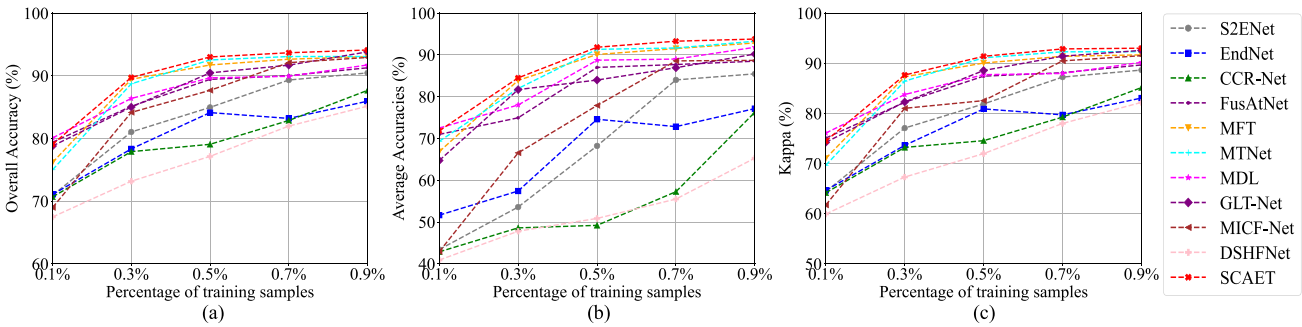


Fig. 11. Classification performance with different percentages of training samples on urban dataset. (a) OA (%). (b) AA (%). (c) Kappa (%).

from the urban and suburb datasets for training purposes, while reserving the remainder for testing. The results obtained are compared with those from other models and are illustrated in Figs. 11 and 12. The experimental results indicate that the SCAET model demonstrates a significant advantage in classification accuracy over other models when faced with limited training samples. With the continuous increase in the number of samples, the classification accuracy of the transformer architectures first increases, then gradually stabilizes and remains stable. Consequently, these experimental findings and analyses robustly affirm the superiority of the SCAET model regarding generalization performance.

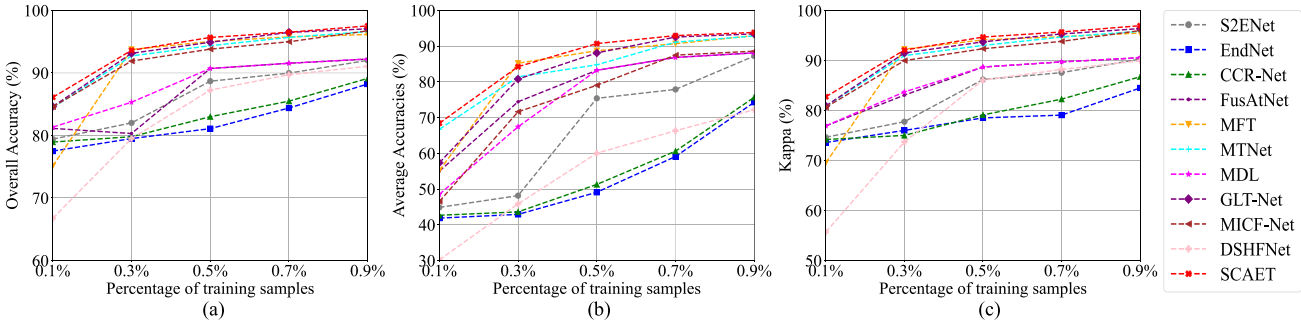


Fig. 12. Classification performance with different percentages of training samples on suburb dataset. (a) OA (%). (b) AA (%). (c) Kappa (%).

TABLE V
ABLATION STUDY OF THE DUAL-BRANCH SS CNN MODULE FOR URBAN AND SUBURB DATASETS

Dataset	Single-branch SS CNN	Dual-branch 3D CNN	Dual-branch 2D CNN	Dual-branch SS CNN	OA(%)	AA(%)	Kappa(%)
Urban Dataset	√	×	×	×	93.88±0.27	92.13±0.27	92.67±0.31
	×	√	×	×	94.14±0.72	93.44±2.50	93.00±0.88
	×	×	√	×	93.81±0.30	93.11±1.91	92.60±0.37
	×	×	×	×	92.44±0.27	92.58±0.79	90.97±0.33
	×	×	×	√	**94.63±0.36**	**94.36±1.84**	**93.58±0.41**
Suburb Dataset	√	×	×	×	95.36±0.18	89.63±0.70	94.26±0.46
	×	√	×	×	97.32±0.15	**95.10±0.16**	96.69±0.18
	×	×	√	×	96.78±0.29	94.19±0.55	96.02±0.36
	×	×	×	×	95.66±0.13	90.59±1.09	94.63±0.17
	×	×	×	√	**97.93±0.17**	95.07±0.47	**97.45±0.21**

The best performance is highlighted in bold.

TABLE VI
ABLATION STUDY OF SFM, SCIE, AND SPM MODULES FOR URBAN AND SUBURB DATASETS

Dataset	SFM	SCIE	SPM	OA(%)	AA(%)	Kappa(%)
Urban Dataset	×	√	√	92.97±1.12	90.75±2.81	91.57±1.37
	√	×	√	92.89±0.14	91.11±0.22	91.47±0.30
	√	√	×	93.23±0.35	92.14±0.53	91.87±0.77
	√	√	√	**94.63±0.36**	**94.36±1.84**	**93.58±0.41**
Suburb Dataset	×	√	√	95.72±0.63	91.52±1.75	94.72±0.77
	√	×	√	96.66±0.36	94.13±0.58	95.88±0.65
	√	√	×	97.28±0.49	**95.08±0.79**	94.01±0.02
	√	√	√	**97.93±0.17**	95.07±0.47	**97.45±0.21**

The best performance is highlighted in bold.

E. Ablation Study

To thoroughly evaluate the contribution of each module in the proposed model, this article conducts a series of systematic ablation experiments. These experiments not only systematically analyze the performance of each module within the overall framework but also provide comprehensive comparative analyses with existing feature extraction and interaction modules. The experimental results demonstrate that the proposed modules exhibit significant advantages in classification performance, validating their effectiveness.

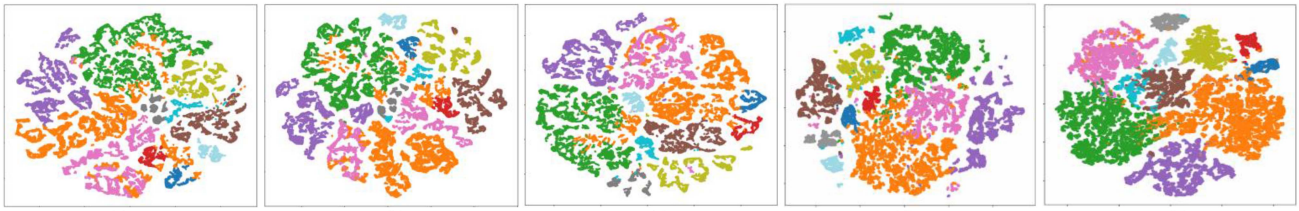
1) Effectiveness of Dual-Branch SS CNN: This section presents ablation experiments on the dual-branch SS CNN module. While keeping the SFM, SCIE, and SPM modules unchanged, this part systematically compares the performance differences among single-branch SS CNN, a dual-branch module using only 3D convolution, and a dual-branch module using only 2D CNN for feature extraction. To ensure consistent feature dimensions, layers are introduced during the experiments. As shown in Table V, the dual-branch SS CNN module outperforms other configurations, which validates the effectiveness of the dual-branch structure in extracting features from V-HSI and T-HSI. The synergistic effect of 3D CNN and 2D CNN captures more diverse feature information, thereby improving classification performance. In contrast, the single-branch SS CNN mixes V-HSI and T-HSI in the spectral dimension, which may lead to mutual interference between the two modalities during feature extraction, making it difficult for the model to effectively distinguish and utilize their unique characteristics, ultimately degrading classification performance.

2) Effectiveness of SFM: This section conducts ablation experiments on the SFM module, whose primary function is to transform convolutional features from V-HSI and T-HSI into independent and interactive features. To verify its effectiveness,

TABLE VII
FURTHER ABLATION STUDY OF SCIE MODULE FOR URBAN AND SUBURB DATASETS

Dataset	Self-attention	Cross-attention	Self- and cross-attention	SCIE	OA(%)	AA(%)	Kappa(%)
Urban Dataset	√	×	×	×	93.78±0.60	93.96±1.49	92.58±0.72
	×	√	×	×	93.70±0.74	93.93±1.73	92.47±0.89
	×	×	√	×	93.47±0.50	93.26±0.70	92.20±0.60
	×	×	×	×	92.89±0.14	91.11±0.22	91.47±0.30
	×	×	×	√	**94.63±0.36**	**94.36±1.84**	**93.58±0.41**
Suburb Dataset	√	×	×	×	96.75±0.62	93.30±1.63	95.98±0.77
	×	√	×	×	96.28±0.46	92.24±1.74	95.39±0.58
	×	×	√	×	97.18±0.12	92.82±0.54	96.51±0.15
	×	×	×	×	96.66±0.36	94.13±0.58	95.88±0.65
	×	×	×	√	**97.93±0.17**	**95.07±0.47**	**97.45±0.21**

The best performance is highlighted in bold.



(a) (b) (c) (d) (e)

Fig. 13. Feature distributions of different methods, which are counted on urban dataset using the t-SNE algorithm. (a) MFT. (b) MTNet. (c) MICF-Net. (d) GLT-Net. (e) SCAET.

TABLE VIII
RUNNING TIME (S) AND PARAMETER SIZE OF DIFFERENT METHODS

Dataset	Metrics	S2ENet	EndNet	CCR-Net	MDL	FusAtNet	DSHFNet	GLT-Net	MICF-Net	MTNet	MFT	SCAET
Urban	Parameters (M)	0.57	0.22	0.16	**0.13**	36.84	0.35	0.73	0.24	0.59	0.29	0.44
	Testing time (s)	9.22	**3.56**	4.27	5.39	14.84	28.47	23.95	15.65	11.34	7.97	14.68
	Training time (s)	0.39	**0.10**	0.13	0.37	0.98	0.75	0.90	0.32	0.29	0.26	0.31
Suburb	Parameters (M)	0.57	0.22	0.16	**0.13**	36.84	0.35	0.73	0.24	0.59	0.29	0.44
	Testing time (s)	5.61	**5.14**	7.91	7.23	21.47	45.36	35.94	25.07	15.37	10.21	25.49
	Training time (s)	0.61	**0.16**	**0.16**	0.57	1.48	1.27	1.03	0.57	0.42	0.39	0.48

The best performance is highlighted in bold.

this part replaces it with a simple feature concatenation method, where the concatenated convolutional features are then split into three feature blocks for processing. As shown in Table VI, while keeping the other structures unchanged, the use of the SFM module significantly improves the model's classification performance, demonstrating its critical role in feature transformation and interaction, and highlighting its superiority in capturing meaningful feature relationships.

3) Effectiveness of SCIE: The SCIE module serves as the core component of the proposed method, playing a pivotal role in feature interaction. Ablation experiments are conducted to validate its effectiveness through two comparative approaches. From Table VI, it can be seen that replacing the SCIE module with a linear connection results in a significant deterioration of classification performance, substantiating its critical contribution.

Besides, this part compares the proposed method with conventional feature interaction modules, particularly self- and cross-attention mechanisms, and reveals that the SCIE module consistently achieves superior performance across both datasets, as evidenced in Table VII. These results demonstrate that the SCIE module facilitates more profound interaction between heterogeneous features, consequently optimizing the overall performance of the network.

4) Effectiveness of SPM: The SPM module further fuses the features enhanced by the SCIE module, effectively improving the model's classification performance. As shown in Table VI, the SPM module contributes significantly to classification accuracy. Taking OA as an example, the introduction of the SPM module improves OA by 1.0% and 0.6% on urban and suburb datasets, respectively, demonstrating its crucial role in feature fusion and classification performance optimization.

F. Time and Parameter Comparison Analysis

This article presents a comparative analysis of the running time and the quantity of trainable parameters associated with the proposed method, in relation to other methodologies. This comparison encompasses both training and testing durations.

Fig. 14. Feature distributions of different methods, which are counted on suburb dataset using the t-SNE algorithm. (a) MFT. (b) MTNet. (c) MICF-Net. (d) GLT-Net. (e) SCAET.

To ensure fairness in evaluation, all experiments are conducted within the same environment.

As illustrated in Table VIII, it is evident that the SCAET model does not excel in terms of operational efficiency or the number of trainable parameters. The main reason is that the processes involved in calculating both the SFM and SCIE modules place a greater emphasis on independent features and interaction features derived from V-HSI and TI-HSI. This focus aims to improve classification accuracy but may inadvertently increase the computational burden of the model.

G. Feature Visualization

To further validate the effectiveness of the proposed SCAET method, t-distributed stochastic neighbor embedding (t-SNE) [54] is employed to visualize the learned feature representations. As shown in Figs. 13 and 14, SCAET generates feature distribution maps urban dataset and suburb dataset and compares them with transformer-based classification methods. Through comparative analysis, it is evident that our SCAET method generates more distinct category boundaries and achieves higher intraclass compactness. These visualization results clearly demonstrate SCAET's advantages in feature representation learning, which contributes significantly to improving final classification accuracy.

IV. Conclusion

A SCAET model, which integrates CNN and the attention mechanisms from the transformer to extract deeper independent and interactive features, is proposed in this article for the classification utilizing V-HSI and TI-HSI datasets. Initially, a dual-branch SS CNN is applied to the reduced-dimensional V-HSI and TI-HSI datasets. Following this, the proposed SFM module is employed to extract independent and interactive feature tokens of V-HSI and TI-HSI datasets. In the SCIE module, the independent features are further augmented through interaction with other features via self- and cross-attention mechanisms. Ultimately, an SPM module is designed to effectively blend these features, thereby improving the model's feature representation capability.

Extensive experiments have been conducted across the urban and suburb datasets, including parameter analysis experiments aimed at identifying optimal values for patch size, learning rate, number of heads in attention layers, and the depth of the SPM module. In addition, ablation studies have been performed to validate the effectiveness of each proposed module. The performance of the proposed model has been compared against several advanced CNN-based and transformer-based models to demonstrate superior efficacy. Concurrently, an analysis of running time and parameters indicates that further optimization regarding training duration and resource utilization remains necessary.

Although the proposed SCAET model has yielded promising experimental results, we intend to explore additional optimizations that enhance information interaction among different modal features while also reducing execution time. Furthermore, we aspire to utilize more diverse datasets in future work to improve the generalization capabilities of our model.

References

[1] S. Peyghambari and Y. Zhang, "Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: An updated review," *J. Appl. Remote Sens.*, vol. 15, no. 3, Jul. 2021, Art. no. 031501.

[2] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, Jul. 2019, Art. no. 3071.

[3] L. Ravikanth, D. S. Jayas, N. D. White, P. G. Fields, and D.-W. Sun, "Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products," *Food Bioprocess Technol.*, vol. 10, no. 1, pp. 1–33, Nov. 2016.

[4] J. Yuan, S. Wang, C. Wu, and Y. Xu, "Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: A case study of Wuhan," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3972–3991, 2022.

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

[6] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[7] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[8] Y. Wang, X. Chen, F. Wang, M. Song, and C. Yu, "Meta-learning based hyperspectral target detection using Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5527913.

[9] F. Cao and W. Guo, "Cascaded dual-scale crossover network for hyperspectral image classification," *Knowl.-Based Syst.*, vol. 189, Feb. 2020, Art. no. 105122.

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

[11] M. E. Paoletti, O. Mogollon-Gutierrez, S. Moreno-Álvarez, J. C. Sancho, and J. M. Haut, "A comprehensive survey of imbalance correction techniques for hyperspectral data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5297–5314, 2023.

[12] B. Kumar, O. Dikshit, A. Gupta, and M. K. Singh, "Feature extraction for hyperspectral image classification: A review," *Int. J. Remote Sens.*, vol. 41, no. 16, pp. 6248–6287, Jun. 2020.

[13] Y. Zhong, T. Jia, J. Zhao, X. Wang, and S. Jin, "Spatial-spectral-emissivity land-cover classification fusing visible and thermal infrared hyperspectral imagery," *Remote Sens.*, vol. 9, no. 9, Sep. 2017, Art. no. 910.

[14] Y. Wang, H. Wang, E. Zhao, M. Song, and C. Zhao, "Tucker decomposition-based network compression for anomaly detection with large-scale hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 10674–10689, 2024.
[15] E. Zhao et al., "Thermal infrared hyperspectral band selection via graph neural network for land surface temperature retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5003414.
[16] X. Liu et al., "Local temperature responses to actual land cover changes present significant latitudinal variability and asymmetry," *Sci. Bull.*, vol. 68, no. 22, pp. 2849–2861, Nov. 2023.
[17] L. He et al., "Non-symmetric responses of leaf onset date to natural warming and cooling in northern ecosystems," *PNAS Nexus*, vol. 2, no. 9, Sep. 2023, Art. no. pgad308.
[18] L. Mei et al., "GTMFuse: Group-attention transformer-driven multiscale dense feature-enhanced network for infrared and visible image fusion," *Knowl.-Based Syst.*, vol. 293, Jun. 2024, Art. no. 111658.
[19] Y. Yang et al., "Spectral-enhanced sparse transformer network for hyperspectral super-resolution reconstruction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 17278–17291, 2024.
[20] E. Zhao et al., "An operational land surface temperature retrieval methodology for Chinese second-generation huanjing disaster monitoring satellite data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1283–1292, 2022.
[21] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.
[22] 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.
[23] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.
[24] H. Lee and H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
[25] M. Ahmad, U. Ghous, M. Usama, and M. Mazzara, "Waveformer: Spectral–spatial wavelet transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5502405.
[26] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.
[27] X. Wang, K. Tan, P. Du, B. Han, and J. Ding, "A capsule-vectored neural network for hyperspectral image classification," *Knowl.-Based Syst.*, vol. 268, May 2023, Art. no. 110482.
[28] 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.
[29] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jan. 2015, Art. no. 258619.
[30] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Apr. 2016.
[31] 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.
[32] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.
[33] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.
[34] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, May 2023.
[35] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
[36] S. Fang, K. Li, and Z. Li, "S2ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6504205.
[37] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5500205.
[38] Y. Wang, X. Chen, E. Zhao, C. Zhao, M. Song, and C. Yu, "An unsupervised momentum contrastive learning based transformer network for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9053–9068, 2024.
[39] E. Zhao, N. Qu, Y. Wang, and C. Gao, "Spectral reconstruction from thermal infrared multispectral image using convolutional neural network and transformer joint network," *Remote Sens.*, vol. 16, no. 7, Apr. 2024, Art. no. 1284.
[40] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.
[41] Y. Zhang et al., "Multimodal transformer network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514317.
[42] Y. He et al., "Multilevel attention dynamic-scale network for HSI and LiDAR data fusion classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5529916.
[43] X. Tang, Y. Zou, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Multiple information collaborative fusion network for joint classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5525416.
[44] M. Wang, Y. Sun, J. Xiang, and Y. Zhong, "CITNet: Convolution interaction transformer network for hyperspectral and LiDAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5535918.
[45] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.
[46] Y. Fan et al., "MslaeNet: Multiscale learning and attention enhancement network for fusion classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 10041–10054, 2022.
[47] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8002005.
[48] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.
[49] 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 2021.
[50] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 416–425.
[51] Y. Feng, L. Song, L. Wang, and X. Wang, "DSHFNet: Dynamic scale hierarchical fusion network based on multiattention for hyperspectral image and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522514.
[52] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.
[53] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.
[54] S. Arora, W. Hu, and P. K. Kothari, "An analysis of the t-SNE algorithm for data visualization," in *Proc. Conf. Learn. Theory*, 2018, pp. 1455–1462.

Enyu Zhao (Member, IEEE) was born in Dalian, Liaoning Province, China, in 1987. He received the Ph.D. degree in cartography and geographic information system from the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, in 2017.

He was a joint Ph.D. Student with Engineering Science, Computer Science and Imaging Laboratory, University of Strasbourg, Strasbourg, France, from 2014 to 2016. He is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. His research interests include quantitative remote sensing and hyperspectral image processing.

Yongfang Su was born in Shangqiu, Henan Province, China, in 2000. He received the B.S. degree in automation from Xi'an University of Technology, Xi'an, China, in 2023. He is currently working toward the M.S. degree in computer science and technology with Dalian Maritime University, Dalian, China.

His research interests include hyperspectral image classification and deep learning.

Nianxin Qu (Student Member, IEEE) was born in Liaoyang, Liaoning Province, China, in 1999. He received the M.S. degree in computer science and technology, in 2024, from Dalian Maritime University, Dalian, China, where he is currently working toward the Ph.D. degree in computer science and technology.

His research interests include hyperspectral image processing and deep learning.

Yulei Wang (Member, IEEE) was born in Yantai, Shandong Province, China, in 1986. She received the B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively.

She was a joint Ph.D. student with Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County in 2011–2013. From 2011 to 2013, she was a Research Assistant with the Shock, Trauma and Anesthesiology Research Organized Research Center, School of Medicine, University of Maryland. She is currently an Associate Professor and doctoral supervisor in Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her current research interests include hyperspectral image processing, multisource remote sensing fusion, and vital signs signal processing.

Caixia Gao (Member, IEEE) received the B.S. degree in electronic and information engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2006,, the M.S. degree in computer science from the Academy of Opto-Electronics, Chinese Academy of Sciences, Beijing, China, in 2009, and the Ph.D. degree in cartography and geography information system from the University of Chinese Academy of Sciences, Beijing, in 2012.

She is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include in-orbit calibration and validation of optical sensors and the retrieval of surface temperature and emissivity.

Jian Zeng received the M.S. degree in cartography and geographic information system from Beijing Forestry University, Beijing, China, in 2019.

He is currently a middle Engineer with China Centre for Resources Satellite Data and Application, Beijing. His research interests include radiometric calibration of optical and thermal infrared sensors for on-orbit satellites, surface reflectance, and temperature retrieval from remote sensing data.