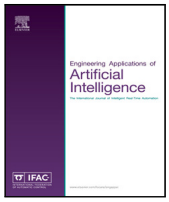
Engineering Applications of Artificial Intelligence 127 (2024) 107351

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Dual attention transformer network for hyperspectral image classification

Zhenqiu Shu *, Yuyang Wang, Zhengtao Yu

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, China

ARTICLE INFO

Keywords:
Hyperspectral image classification
CNN
Transformer
Spatial–spectral
Local
Global

ABSTRACT

Hyperspectral image classification (HSIC) has been a significant topic in the field of remote sensing in the past few years. Convolutional neural networks have shown promising performance in HSIC applications due to their strong local feature extraction ability. However, they struggle to extract global information from HSIs, thereby resulting in classification performance limitations. Recently, vision transformers have been used to solve HSIC problems, and its advantage is to adopt the multi-head self-attention mechanism to explore global dependencies. Nevertheless, the extracted features using MHSA usually exhibit over-dispersion due to the abundance of band information hidden in HSIs. In this work, we propose a novel method, called dual attention transformer network (DATN), for HSIC problems. It consists of two types of modules, namely the spatial–spectral hybrid transformer (SSHT) module and the spectral local-conv block (SLCB) module. Specifically, the SSHT module aims to utilize the MHSA to capture spatial and spectral feature information. Therefore, it can effectively utilize global spatial–spectral features and embed the local spatial information, simultaneously. Besides, we design a SLCB module to extract the local spectral information of HSIs effectively. Then the SSHT and SLCB modules are integrated into an end-to-end framework. Finally, the global and local spatial–spectral features extracted from this framework are input into the fully connected layer, and then classification results of HSIs are obtained. A series of experiments on three HSI datasets have demonstrated that our DATN approach outperforms several state-of-the-art HSIC approaches.

1. Introduction

At present, remote sensing technology has been applied to many areas of our lives (Meedeniya et al., 2020; Jayanetti et al., 2017). Hyperspectral images (HSIs) are usually obtained from satellites or airborne vehicles. In contrast to conventional RGB images, HSIs can offer more abundant spectral information to compensate for the limited spectral resolution (Li et al., 2019; Paoletti et al., 2019). Over the few decades, HSI classification (HSIC) technology has demonstrated great potential in various real applications, such as agriculture (Zhu et al., 2020a), geological exploration (Peyghambari and Zhang, 2021), environmental monitoring (Zhang and Liu, 2010), medical imaging (Khan et al., 2021). However, there are some significant challenges in HSIC applications due to the high dimensionality of HSIs and the fusion of spatial–spectral information.

In general, the goal of HSIC technology is to identify the land cover for each pixel of HSIs. Over the last few years, many approaches have been developed for HSIC tasks. In the early stages, various traditional machine learning approaches, such as k-nearest neighbor (k-NN) (Ma et al., 2010), sparse representation (Cariou and Chehdi, 2015), morphological profile (MP) (Benediktsson et al., 2005), support vector machine (SVM) (Melgani and Bruzzone, 2004), random forest (RF) (Ham et al., 2005), are used to cope with the HSIC tasks. However, these approaches aim to statistically analyze prior information from training data. Therefore, it implies that these methods are susceptible to noises or outliers, and thus cannot guarantee their robustness and stability.

With the significant progress of deep learning in numerous real-world problems (Yang et al., 2023; Xian et al., 2022a; Li et al., 2021; Himeur et al., 2022, 2021), deep neural networks (DNNs) have been widely coped with classification problems. Among them, convolutional neural networks (CNNs) have dominated this field in the past decade. In Meedeniya et al. (2022), a CNN-based deep learning model was designed to classify the satellite images and perform geospatial analysis. The Ref. Mahakalanda et al. (2022) utilized a fully convolutional network to classify the standage and land use of rubber plantations. In the past few years, various CNN methods have been also applied to HSIC due to their powerful ability to extract spatial–spectral features of HSIs. Hu et al. (2015) utilized a 1D CNN architecture with five convolutional layers to extract spectral information and selected individual pixels from HSIs to train the model. However, it can only consider the spectral information of HSIs and ignores the spatial relationships between pixels. In Zhao and Du (2016), a 2D CNN-based network was

* Corresponding author.
E-mail address: shuzhenqiu@163.com (Z. Shu).

https://doi.org/10.1016/j.engappai.2023.107351
Received 11 July 2023; Received in revised form 20 September 2023; Accepted 22 October 2023
Available online 31 October 2023

designed for HSIC tasks. This approach fully utilizes the high spectral resolution and strong spatial dependencies to realize the joint analysis of spatial–spectral information. Song et al. (2018) introduced a deep feature fusion network (DFFN) based on 2D CNNs, which alleviates the problems of overfitting and gradient vanishing. Besides, it fuses the features learned at different levels to further enhance the overall classification performance. Considering the three-dimensional structure of HSIs, Chen et al. (2016) proposed a deep network by stacking 3D CNNs for extracting both spatial and spectral information of HSIs. Li et al. (2017) introduced a spatial–spectral feature extraction method based on 3D CNNs, which eliminates the need for pre-possessing or post-processing. Zhong et al. (2017) developed a 3D spatial–spectral residual network (SSRN) for extracting the spatial–spectral features using multiple connected residual blocks. Since there are spatial correlations and inter-spectral correlations between adjacent pixels and bands, and thus HSIs contain significant redundant information. However, the aforementioned CNNs treat all bands of HSIs equally in classification tasks, and thus the useless bands lead to unsatisfactory classification results. To address this problem, the attention mechanism was introduced into the CNN-based HSIC model. Shu et al. (2022) introduced a spatial–spectral split attention residual network that integrates three different attention mechanisms to highlight the useful information in spectral and spatial dimensions. Li et al. (2020) put forward a dual-branch dual-attention classification network (DBDA) for HSIC problems. This method attempts to employ two distinct attention mechanisms within two separate branches and captures the spectral and spatial information in HSIs independently.

Despite CNN-based HSIC methods achieving promising results in most cases, they still encounter the following limitations:

(1) CNN is a vector-based approach that treats the input of HSIs as a collection of pixel vectors. However, the equal treatment of all channels can lead to CNN-based HSIC methods being incredibly insensitive to the spectral sequence information, resulting in feature maps that cannot effectively represent the spectral sequence information from HSIs.

(2) Due to the commonly used fixed patch size, CNN-based HSIC methods cannot accommodate larger scales and more complex convolutional forms. Therefore, there is a trend to select smaller convolution kernel sizes for performing the convolutional operation. However, the size limitation of the receptive field cannot effectively capture the spatial information of HSIs, leading to the loss of global contextual information in HSI patches. Moreover, CNN-based methods are unable to depict the long-range dependencies of the pixels and the bands, thus leading to unsatisfactory classification performance.

To tackle the aforementioned challenges, some studies have attempted to apply transformer networks to capture long-range information. As a key component of the Transformer, self-attention is a variant of the attention mechanism that operates on the feature vectors of each position in an input sequence rather than the elements in the sequence. It uses a weighted average of all feature vectors in a sequence to represent the importance of each position. As a result, various HSIC methods based on transformer network have been developed in the past few years. Hong et al. (2021) introduced a transformer-based HSIC framework, termed spectralformer. It preserves the local sequence spectral information by groupwise spectral embedding and fuses multi-layer features through cross-layer skip connection. However, many finer local spatial features cannot be extracted compared with CNNs. To alleviate this defect, many studies (Sun et al., 2022)-(Mei et al., 2022) attempt to combine CNN and transformer to capture both local and global feature information from HSIs. Zhang et al. (2022) designed a convolution transformer mixer (CTMixer) network for HSIC. Firstly, it applies the CNN blocks to extract the local spatial–spectral information. Then CNN is introduced into a multi-headed attention mechanism (MHSA) to achieve a more elegant combination of CNN and transformer. Song et al. (2022) designed a bottleneck spatial–spectral transformer (BS2T) network that uses a dual-branch structure to learn the spatial and spectral information. Specifically, it designs a multi-head spatial–spectral self-attention (MHS2 A) structure to replace the convolution operation, which integrates the spectral and spatial position information into the multi-head attention mechanism.

However, the existing transformer-based HSIC methods still suffer from the following limitations:

(1) Both the linear projection and the flattening operation in the transformer may destroy the local spatial–spectral information and the position information, thereby causing the loss of useful local information in the classification procedure.

(2) Although the MHSA can effectively model long-term dependencies and extract global information, it is unreasonable to establish the context relationships for all pixels equally. This is because nearby pixels in HSI are always more relevant than distant ones.

(3) Unlike traditional RGB images with only three channels, HSIs contain hundreds of bands, which makes the extracted information too scattered in the global scope by applying MHSA to HSIs. Therefore, it is inappropriate to generate a single token from the spectral or spatial perspective.

To address the above-mentioned issues, this work proposes a dual attention transformer network (DATN) for HSIC tasks. In DATN, we design a squeeze-excitation-convolutional (SE-conv) block by using 1×1 convolutions with channel attention, which splits the input feature map into isolated pixels sequences as the embedding for the input of the transformer. 2D CNN is employed to change the number of channels in the middle layer feature map and reduce computational costs. The channel attention is also used to highlight the useful information in HSIs. Meanwhile, we adopt a novel Spatial–Spectral Hybrid Transformer (SSHT) to model the long short-range dependencies of the HSI features. It can effectively capture the local spatial and global spatial–spectral information of HSIs. To capture local spectral information ignored by the transformer branch, we design a spectral local convolutional block (SLCB) that is connected with the self-attentive branch in a parallel manner. To effectively explore the correlations and complementarity of local and global HSI features, the feature representation is enhanced by fusing the outputs from both SSHT and SLCB. The proposed DATN method can fully merge the global and local spatial–spectral information of HSIs while considering the local information of pixels. The classification evaluation experiments on various HSI datasets have clearly demonstrated the superiority of our DATN method.

The main contributions of this work are summarized as follows:

1. We propose a novel HSIC method, namely dual attention transformer network (DATN), which can integrate the merits of transformer and CNN to capture global and local spatial–spectral feature information of HSIs.
2. To exploit the three-dimensional structure of HSIs, a novel spatial–spectral hybrid transformer (SSHT) module is designed to generate the tokens from the spectral and spatial dimensions, which can model long-range dependencies in the spatial and spectral spaces separately. The proposed SSHT module splits the original patches into smaller patches and then feeds them into MHSA. Therefore, it overcomes the limitation of MHSA that cannot effectively extract local information.
3. We design a simple yet powerful SLCB module to exploit local spectral information and partially preserve the spatial information of HSIs, simultaneously. The proposed SLCB module can convey original spatial information from shallow to deep layers, resulting in a significant improvement in classification performance.
4. Extensive experiments on various HSI datasets are conducted to verify the effectiveness of the DATN method in HSIC problems. Besides, several ablation experiments are carried out to confirm the usefulness and necessity of each module in our DATN method.

The remainder of this paper is organized as follows: In Section 2, we briefly review the relative works. In Section 3, we give the detailed structure of our designed network. Section 4 presents the experimental results of the proposed method on three benchmark datasets. Finally, we draw the conclusions in Section 5.

2. Related work

2.1. CNN-based HSIC methods

Over the past decade, various CNNs have been proposed to cope with classification tasks in remote sensing and can achieve remarkable performances in many cases. In general, the simple 1D CNN only focuses on the individual pixel information of HSIs, thereby ignoring a large amount of useful spatial and positional information. Besides, 2D CNN tends to pay more attention to the extraction of spatial information without concerning the spectral information of HSIs. Compared to these two methods, 3D CNN are better equipped to handle the 3D structure of HSIs. Nevertheless, the simple stacking of 3D CNN blocks may lead to extremely high computational costs. To address the aforementioned issues, an octave 3D CNN (Xu et al., 2020) was proposed to decompose the mixed feature map by their frequency, which can effectively reduce the parameter number of the network. In Cui et al. (2021), the authors proposed a 3D CNN-based lightweight network that replaces the convolution operation into point-wise and depth-wise convolution. This approach has achieved remarkable classification performance while maintaining network efficiency. To learn the direct spectral continuity information of HSIs, a long short-term memory network (LSTM) based on band grouping was designed in Graves and Graves (2012). It can extract the spatial features from HSIs by integrating a CNN sub-network. Due to the complex acquisition process of HSIs, there is limited availability of large-scale publicly accessible HSI datasets, making it challenging to obtain a sufficient number of labeled samples. To address this limitation, Cao et al. (2020) designed a classification framework based on autonomous learning (AL) and CNN, in which a limited number of samples were required to train CNN. The deep cross-domain few-sample method (Li et al., 2022) was put forward to solve the few-sample cross-domain HSIC task with few labeled pixels.

Until now, various CNN models have been utilized for extracting spatial–spectral features from HSIs and can achieve encouraging results in HSIC problems. However, the limited receptive field of CNNs restricts the modeling ability of long-range dependencies, thus hindering further improvements in classification performance.

2.2. Attention mechanism model

The attention mechanism is considered as a signal-processing technology and has recently played an important role in deep learning. It aims at assigning different weights to various parts of the input data, allowing the network to focus more effectively on critical components. Many studies have shown that the attention mechanism not only effectively enhances model performance, but also provides stronger interpretability for the decision-making process. Hu et al. proposed (Hu et al., 2018) a squeeze-and-excitation network (SENet) to focus on the channel relationships of feature maps. It operates on the feature maps to obtain the attention parameters, and then fuses these parameters with the original feature maps. In Woo et al. (2018), a convolutional block attention module (CBAMBlock) was designed by combining spatial attention and channel attention. In this method, the attention module incorporates multiple stacked max or average pooling operations instead of a single operation. Moreover, the aforementioned methods can improve classification performance without significantly increasing computational complexity. Zhu et al. (2020b) proposed an end-to-end residual spatial–spectral attention network (RSSAN) by applying spatial–spectral attention to HSIC. Mei et al. (2019) further developed a novel spectral–spatial attention block based on SSRN to enhance classification performance.

The attention mechanism is mainly used to calculate an attention weight separately for each part in a set of input data. It usually requires additional reference information to calculate the attention weights, such as context vector, query vector, etc. Self-attention is a variant of the attention mechanism that operates on the feature vectors of each position in an input sequence, rather than the elements in the sequence. As a result, the self-attention mechanism is more adept at quickly capturing internal correlations within a sequence (Sun et al., 2023). Based on the non-local mean filtering operation, a non-local neural network (Wang et al., 2018) was developed by incorporating self-attention into the computer vision domain. Therefore, it can effectively model long-range dependencies. Compared to continuously stacking CNNs, the non-local operation achieves a more efficient optimization procedure. The transformer network (Vaswani et al., 2017) was designed with a large amount of self-attention mechanisms, in which the self-attention model is used as a standalone structure without adding convolutional layers to extract the features.

2.3. Transformer-based HSIC methods

As a network structure composed of self-attention mechanisms, the transformer network has swept the natural language processing (NLP) field with powerful learning long-range dependencies. Recently, the vision transformer (ViT) model has achieved superior results in image classification tasks in comparison to CNN-based models[(Xian et al., 2022b)-(Long et al., 2023)]. In recent years, various variants of transformer have also emerged for HSIC applications (Qi et al., 2023). Bai et al. (2022) designed a multi-branch transformer structure by taking into account both spatial and spectral attentions, and designed a mask prediction (MP) model that can run in parallel with the classification prediction branch. Therefore, this network focuses on all pixel categories in hyperspectral images. Xu et al. (2022) designed a dual-branch network comprising a grouped bidirectional long short-term memory (GBiLSTM) network and a multi-stage fusion convolutional transformer (MFCT). The GBiLSTM network segments the input feature vector into groups and integrates the spectral features extracted from each group. The MFCT module designs a convolutional visual transformer block and uses a residual approach to fuse multi-layers of features together. In order to leverage the spatial–spectral information of HSIs, Wang et al. (2022) developed a bilateral classification network referred to as an efficient spatial–spectral transformer. In this approach, a dual-branch architecture is implemented in the transformer network. Tang et al. (2023) presented an approach that considers both global spectral information and local spatial information, and uses a cross-layer CLS token fusion scheme to fuse multilayer features. Yang et al. (2022) adopted both series and parallel methods to fuse CNN and transformer to achieve better HSI classification performance. Xue et al. (2022) developed a novel HSIC model by combining the neural architecture search (NAS) algorithm with the transformer for the first time. Dang et al. (2023) adopted two-branch transformer structure to extract global spatial–spectral information separately. Pan et al. (2023) captured shallow-to-deep HSI features by fusing the two transformer subnetworks. Unlike the aforementioned methods, our work aims at designing a spatial–spectral transformer module that can effectively explore the global spatial–spectral information of HSIs.

3. Proposed method

A comprehensive introduction to the proposed DATN model is presented in this section. Fig. 1 illustrates the overall architecture of the DATN model. This architecture adopts a four-stage hierarchical framework as the main branch, where the squeeze-excitation-convolutional (SE-Conv) block and the spatial–spectral hybrid transformer (SSHT) are concatenated in each stage. Additionally, the proposed spectral

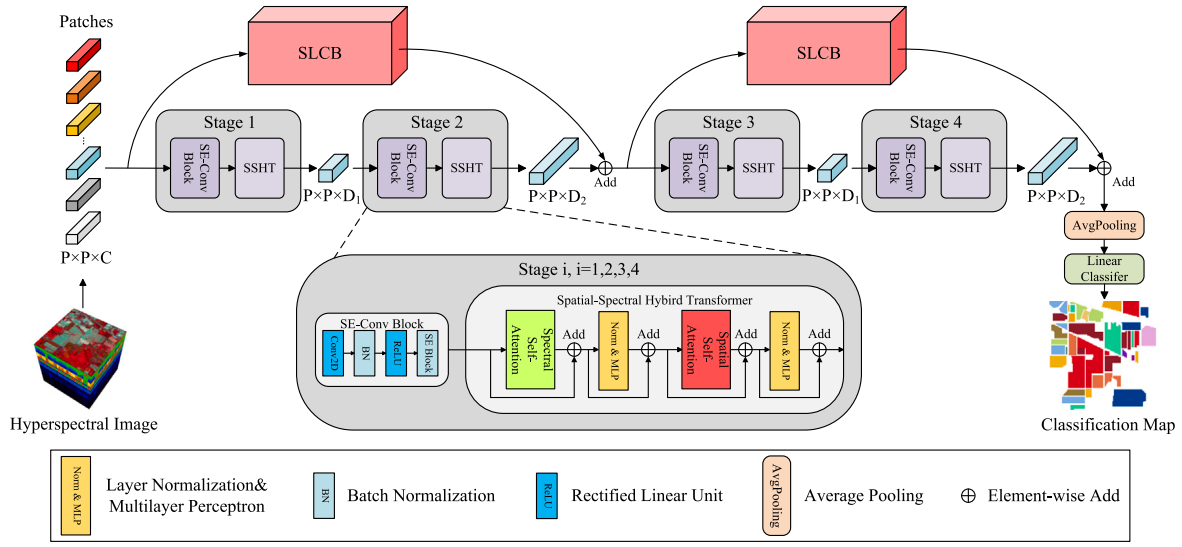


Fig. 1. The framework structure of the proposed DATN method.

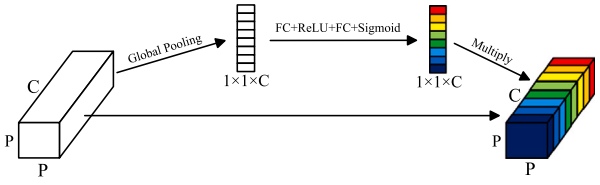


Fig. 2. SE Block.

local-convolutional block (SLCB) module is integrated into the main branch in a skip connections manner, and the feature maps extracted from stages 2 and 4 are fused with the output of SLCB, respectively. Finally, we compress the spatial information by performing a global average pooling operation. Subsequently, a linear classifier is employed to classify the pixels, and the classification information can be obtained.

3.1. SE-Conv block

To control the number of feature map channels and carry out initial feature map processing before the SSHT module, a SE-conv block is incorporated before each SSHT module. The SE-Conv block consists of a 2D CNN with simple 1×1 convolution and a SE module. Fig. 2 depicts the structure of the SE module. The HSI data is denoted $X \in \mathbb{R}^{H \times W \times C}$, where H and W represent spatial dimensions, respectively, and C is the number of spectral bands. The model input is a $P \times P \times C$ patch block split from the original image, where P is the patch size. To reduce the parameter number and computational cost, we change the channel number in the feature maps by feature mapping, and adopt a bottleneck structure for the input channels of these four SSHT modules. The feature map sizes input for these four layers are $P \times P \times D_1$, $P \times P \times D_2$, $P \times P \times D_1$ and $P \times P \times D_2$, respectively. Meanwhile, the SE module can preliminarily emphasize the channels in the feature maps that are useful for classification. Due to the relatively smaller size of HSI compared to traditional images, we adopt pixel embedding instead of image patches. After the SE module is performed on the feature map, we split the patches of HSIs into smaller patches, and then feed them into the SSHT module as tokens.

3.2. Spatial–spectral hybrid transformer

It is crucial to extract global contextual information for many computer vision tasks. Recently, the transformer structure is used to construct long-range spatial dependencies by using the self-attention mechanism, rather than relying solely on the multi-layer down-sampling in CNNs to obtain a global receptive field. Dosovitskiy et al. (2021) directly applied a pure transformer to sequences of image patches and performed very well on image classification tasks. It splits an image into fixed-size patches and feeds them to a standard Transformer encoder. The transformer encode can be represented by the following steps.

Assuming that the sequence of tokens is split by the SE-conv block is $X = \{X_1, X_2, \ldots, X_s\}$, where the dimension of each token is d, and s is the number of tokens. In the original transformer encode block, X needs to be multiplied by three different learnable weights matrices W^Q, W^K, W^V to obtain three different vectors Q, K and V. Then Q, K, and V are respectively divided into n parts along the channel dimension as follows:

$$\begin{aligned} Q &= \{Q_1, Q_2, \ldots, Q_i, \ldots, Q_n\} \\ K &= \{K_1, K_2, \ldots, K_i, \ldots, K_n\} \\ V &= \{V_1, V_2, \ldots, V_i, \ldots, V_n\} \end{aligned} \tag{1}$$

where n is the number of heads in multi-head attention. Next, we use the dot product operation between each Q_i and the transpose of each K_i to calculate the attention, which represents the similarity between Q_i and K_i. Since the value of $Q_i \times K_i$ increases with the increase of the dimension d of Q_i and K_i, we need to divide by $\sqrt{d_k}$ to control the problem of gradient disappearance. The attention matrix can be normalized by using the softmax function. Finally, we multiply the attention matrix by V_i to obtain the output of a single head as follows:

$$h_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{2}$$

Finally, the output of multiple heads is concatenated to obtain the output of MHSA as follows:

$$\text{Attention}(Q, K, V) = \text{Concat}(h_1, h_2, \ldots, h_n) W^o \tag{3}$$

where W^o is the output projection matrix. Following the MHSA module in the transformer encoder, a MLP block is added to extract feature information. It consists of two fully connected layers implemented as

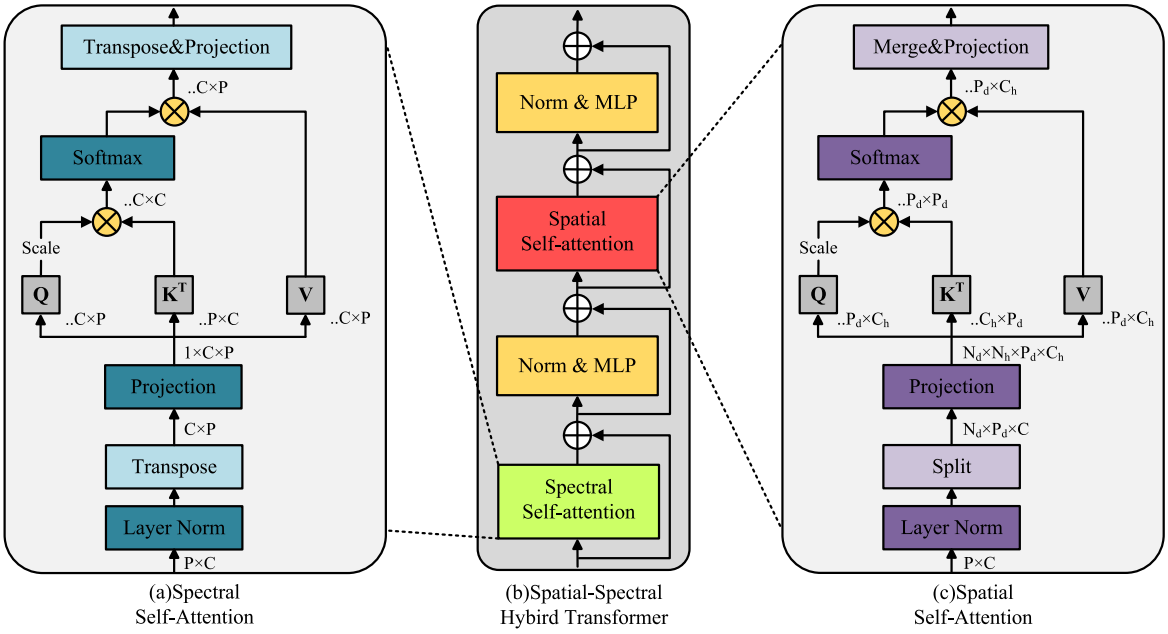


Fig. 3. The overall structure of the proposed SSHT module.

a linear projection layer (FC) and a Gaussian error linear unit (GELU). Here, the MLP layer can be represented as follows:

$$\mathrm{MLP}(X) = \mathrm{FC}_2\Big(\mathrm{GELU}\big(\mathrm{FC}_1(X)\big)\Big) \tag{4}$$

In addition, the LayerNorm operation is performed before the MHSA module and the MLP layers, and the residual connection is applied between the two modules. The transformer encoder can be expressed as follows:

$$\tilde{X} = \mathrm{MHSA}\big(\mathrm{LayerNorm}(X)\big) + X \tag{5}$$

$$O = \mathrm{MLP}\left(\mathrm{LayerNorm}\left(\tilde{X}\right)\right) + \tilde{X} \tag{6}$$

where $\tilde{X}$ and O represent the output features of the MHSA module and MLP layer, respectively.

However, most of the existing image-based MHSA mechanisms cannot effectively cope with the 3D structure of HSIs, thereby losing a significant amount of spectral information. To address this issue, we seek to improve the existing MHSA based on the characteristics of HSIs. Considering the "spatial tokens" commonly used to define image patches, we introduce the spectral tokens by utilizing self-attention in the spectral dimension. For spatial tokens, the spatial dimension limits the range of tokens and the spectral dimension represents the dimension of token feature. For spectral tokens, there is a opposite situation: the spectral dimension limits the range of tokens and the spatial dimension represents the dimension of token feature. Consequently, each spectral token is global in the spatial dimension and contains an abstract representation of the entire HSI patches. Compared to traditional self-attention methods that model long-range dependencies based on local patches, spectral attention takes all spatial positions into account while calculating spectral global attention based on spectral tokens. Therefore, spectral self-attention can be regarded as a dynamic feature fusion across a sequence of spatial abstract representations of the entire patches.

In addition, we also modify the self-attention based on spatial tokens to focus on local spatial information interaction. Due to the excellent extraction of global spatial information using the aforementioned spectral self-attention, we construct a SSHT module based on spectral self-attention and spatial self-attention to effectively explore global spatial–spectral feature information of HSIs and local spatial information of HSI patches. Specifically, spectral attention not only extracts attention information between bands by calculating across global spectral tokens, but also provides a global receptive field in the spatial dimension. The spatial attention refines the local spatial fine-grained information by chunking the input. Therefore, it is clear to know that these two self-attentions complement each other. The structure of our SSHT module is depicted in Fig. 3. Here, we provide a detailed introduction to two distinct self-attention mechanisms.

3.2.1. Spatial self-attention

In traditional transformer-based HSIC methods, all pixel tokens in HSI patches are fed into the transformer block together to extract global spatial features. However, this extraction approach fails to fully incorporate local spatial information and the positional relationship between pixels. Actually, the close pixels are always more relevant and the adjacent pixels in homogeneous regions usually convey similar semantic information. Therefore, we uniformly divide each HSI patch into non-overlapping blocks instead of feeding all pixels into MHSA. Supposing there are N_d different blocks with each block containing P_d pixels, the number of pixels in the original patches is $P = N_d * P_d$. If the block size cannot be evenly divided into the patches, we perform the padding operations before division. Thus, the spatial self-attention can be represented as follows:

$$A_{spatial}(Q, K, V) = \{\mathrm{Attention}(Q_i, K_i, V_i)\}_{i=0}^{N_d} \tag{7}$$

where Q_i, K_i, $V_i \in \mathbb{R}^{P_d \times D_s}$ are the queries, keys, and values of the local block. The structure of spatial self-attention is illustrated in Fig. 3(c). Although the feature map obtained in this manner preserves the local spatial information and positional information of HSIs, the partial global modeling ability may be lost by performing self-attention on the local blocks. Therefore, we proposed spectral self-attention to address this issue of global spatial information loss.

3.2.2. Spectral self-attention

Due to the inherent structural characteristics of HSIs, we design a spectral self-attention in the SSHT module. Assume that the segmented feature map $X \in \mathbb{R}^{N \times C}$ after splitting pixels by SE-conv block. Here, $N = P \times P$ is the number of pixels in each patch, where P denotes patch spatial size. After performing the 2D CNN layer, the feature map contains C channels. To derive the spectral token, the feature map is

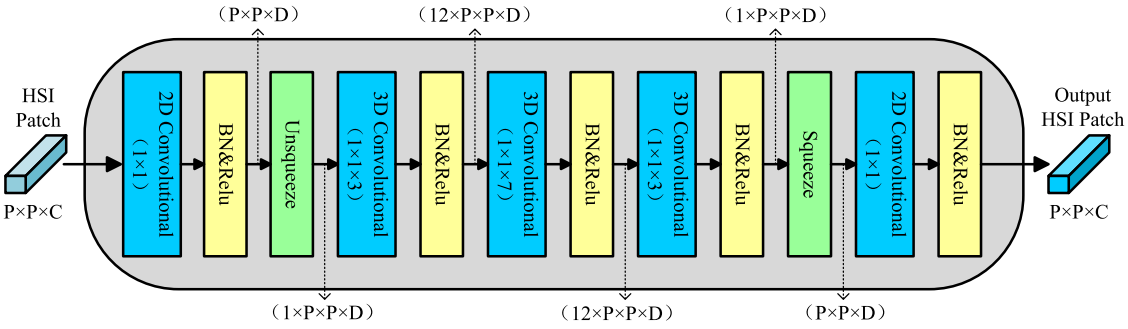


Fig. 4. The framework structure of the proposed SLCB module.

transposed and then send it to the self-attention module. At this time, the feature is represented as $X \in \mathbb{R}^{C\times N}$, where C denotes the number of the spectral tokens. When the number of heads in the self-attention module is set to 1, each spectral token contains the global spatial information. Therefore, the spectral self-attention can be expressed as follows:

$$A_{spectral}(Q, K, V) = \text{Attention}(Q_i, K_i, V_i) \tag{8}$$

where Q_i, K_i, $V_i \in \mathbb{R}^{C\times N}$ are the queries, keys, and values of the transposed spectral-wise feature maps. Since we switch the tokens in the self-attention, the scaling factor $\sqrt{d_k}$ and the projection matrix W^o are also calculated based on the spatial dimension N. By setting the number of heads to 1, we eliminate the concern of spatial pixel count being indivisible by the number of attention heads. The proposed spectral self-attention module can be easily integrated into other models based on different patch sizes. Fig. 3(a) describes the structure of spectral self-attention.

In our model, we design a spatial self-attention and a spectral self-attention instead of using the MHSA in the transformer, and then combine them to further construct a SSHT block. It enables the two attention mechanisms to work alternately for better extracting global spatial–spectral features and fine-grained local spatial features. The detailed algorithm flow is shown in Algorithm 1.

Algorithm 1: The Proposed Spatial-Spectral Hybrid Transformer Method

Input: Input feature map $X \in \mathbb{R}^{N\times C}$; Number of heads h; size of the divided block s

Transpose feature map from $\mathbb{R}^{N\times C}$ to $\mathbb{R}^{C\times N}$;
Perform a Spectral Self-Attention by Eq.(8);
Transpose feature map from $\mathbb{R}^{C\times N}$ to $\mathbb{R}^{N\times C}$;
Perform a skip connect by Eq.(5);
Perform a Multi-Layer Perceptron by Eq.(4);
Calculate the feature map by Eq.(6);
while *s cannot divide the feature map evenly* **do**
  perform the padding operations before division;
Divide feature map from $\mathbb{R}^{N\times C}$ to $\mathbb{R}^{N_d\times P_d\times C}$;
Perform a Spatial Self-Attention by Eq.(7);
Merge divided feature map from $\mathbb{R}^{N_d\times P_d\times C}$ to $\mathbb{R}^{N\times C}$;
Perform a skip connect by Eq.(5);
Perform a Multi-Layer Perceptron by Eq.(4);
Calculate the feature map by Eq.(6);

Output: Output feature map $X \in \mathbb{R}^{N\times C}$

3.3. Spectral local-convolutional block

Although the SSHT module can model the long-range dependencies in both spatial and spectral dimensions and preserve local spatial information and positional information, simultaneously, it still ignores the local spectral information of HSIs. Therefore, the accuracy of classification results may be significantly harmed by neglecting this vital inter-band information in HSIs. To resolve this problem effectively, we integrate the spectral local-convolutional block (SLCB) into the proposed framework to preserve the local spectral information.

The SLCB includes two 2D CNN blocks and three 3D CNN blocks. After each CNN layer, we adopt a batch normalization (BN) layer and a rectified linear unit (ReLU) activation function to process the feature map. The overall structure of the proposed SLCB module is illustrated in Fig. 4. Assuming that the input patch block is $X \in \mathbb{R}^{P\times P\times C}$, where P denotes the patch size and C represents the original band number of HSIs. To process the input patch block, we first use a 2D CNN with a 1×1 convolution kernel to extract the spectral information and increase the dimensionality of the feature map. The extracted feature is reshaped to $X \in \mathbb{R}^{P\times P\times D}$, where D is the number of convolution kernels of 2D CNN. The activation value at the spatial position (x, y) in the jth feature map of the ith layer is represented as $v_{i,j}^{x,y}$ in 2D convolution. Therefore, we have

$$v_{i,j}^{x,y} = f\left(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho} \times v_{i-1,\tau}^{x+\sigma,y+\rho}\right) \tag{9}$$

where $f(.)$ is the activation function. The bias parameter for the jth feature map in the ith layer is denoted as $b_{i,j}$. The number of feature map in $(l-1)$-th layer is represented as d_{l-1}, and d_{l-1} is also the depth of kernel $w_{i,j}$ for the jth feature map of the ith layer. The width and height of kernel are denoted as $2\gamma+1$ and $2\delta+1$, respectively, where $w_{i,j}$ stands for the weight parameter value for the jth feature map of the ith layer. Then the feature map is convolved with a set of 3D convolutional kernels with two $1\times 1\times 3$ and a $1\times 1\times 7$ 3D convolution kernels used alternately to extract local spectral information with different receptive fields. We perform the stride and padding operations in the 3D CNN to maintain the dimensionality of the data. In 3D convolution, $v_{i,j}^{x,y,z}$ is the activation value at spatial position (x, y, z) in the jth feature map of the ith layer. Therefore, we have

$$v_{i,j}^{x,y,z} = f\left(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\lambda=-\eta}^{\eta} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho,\gamma} \times v_{i-1,\tau}^{x+\sigma,y+\rho,z+\lambda}\right) \tag{10}$$

Here, the depth of the kernel along the spectral dimension is set to $2\eta+1$, and the other parameters remain the same setting as in (9). Finally, we adjust the number of channels to match the attention branch using a 2D CNN with 1×1 convolutional kernel, so that the final feature map is represented as $X \in \mathbb{R}^{P\times P\times D}$. In the overall model, we add two SLCBs to capture local spectral features in different layers, and then fuse the extracted features with the main branch. By setting the size of 3D convolutional kernel to 1 in the spatial dimension, the kernel only slides along the spectral dimension, which can effectively preserve the initial spatial features of the input feature map. Subsequently, the features that have been processed by SSHT are combined with local spectral features in a skip connection fashion, and the final feature map is employed for classification purposes.

Table 1
Sample division of the Indian Pines (IP) dataset.

No.	Name	Training	Validation	Test
1	Alfalfa	5	5	36
2	Corn-notill	25	25	1378
3	Corn-mintill	25	25	780
4	Corn	25	25	187
5	Grass-pasture	25	25	433
6	Grass-trees	25	25	680
7	Grass-pasture-mowed	5	5	18
8	Hay-windrowed	25	25	428
9	Oats	5	5	10
10	Soybean-notill	25	25	922
11	Soybean-mintill	25	25	2405
12	Soybean-clean	25	25	543
13	Wheat	25	25	155
14	Woods	25	25	1215
15	Building-Grass-Trees-Drives	25	25	336
16	Stone-Steel-Towers	10	10	73
Total		325	325	9599

Table 2
Sample division of the Houston 2013 (HU) dataset.

No.	Name	Training	Validation	Test
1	Healthy grass	40	40	1171
2	Stressed grass	40	40	1174
3	Synthetic grass	10	10	677
4	Tree	40	40	1164
5	Soil	40	40	1162
6	Water	10	10	305
7	Residential	40	40	1188
8	Commercial	40	40	1164
9	Road	40	40	1172
10	Highway	40	40	1147
11	Railway	40	40	1155
12	Parking lot1	40	40	1153
13	Parking lot2	10	10	449
14	Tennis court	10	10	408
15	Running track	10	10	640
Total		450	450	14129

Table 3
Sample division of the Salinas Valley (SV) dataset.

No.	Name	Training	Validation	Test
1	Brocoli_green_weeds_1	40	40	1929
2	Brocoli_green_weeds_2	60	60	3606
3	Fallow	40	40	1896
4	Fallow_rough_plow	20	20	1354
5	Fallow_smooth	50	50	2578
6	Stubble	60	60	3839
7	Celery	60	60	3459
8	Grapes_untrained	60	60	11151
9	Soil_vinyard_develop	40	40	6123
10	Corn_senesced_green_weeds	20	20	3238
11	Lettuce_romaine_4wk	10	10	1048
12	Lettuce_romaine_5wk	10	10	1907
13	Lettuce_romaine_6wk	10	10	896
14	Lettuce_romaine_7wk	20	20	1030
15	Vinyard_untrained	10	10	7248
16	Vinyard_vertical_trellis	20	20	1767
Total		580	580	52969

4. Experiments

To validate the effectiveness of the proposed approach, we conducted comprehensive evaluation experiments on three widely used multi-view datasets by comparing several state-of-the-art MVC methods.

4.1. Datasets description

(1) Indian Pines (IP): Indian Pines is a relatively early HSI dataset, which was acquired in June 1992 in Indiana, USA. It consists of a 145 × 145-pixel area that was selected and annotated from images captured by the airborne visible/infrared imaging spectrometer (AVIRIS). Its wavelength range spans from 400 to 2500 nm with a spatial resolution of approximately 20 meters and 220 contiguous spectral bands. Twenty water absorption bands were removed, and 200 bands were left for training. The IP dataset includes 16 land cover types, such as Corn, Oats, Wheat, Woods, etc.

(2) Houston 2013 (HU): The Houston dataset was obtained using the ITERS CASI-1500 sensor in June 2012 in Houston, Texas. Its spatial resolution is 2.5 meters and the size is 349 × 1905 pixels with 144 spectral bands ranging from 364–1046 nm. There are 15 and-cover classes in the Houston dataset, which have been widely utilized to evaluate the effectiveness of HSIC methods.

(3) Salinas Valley (SV): The Salinas dataset was acquired by the AVIRIS sensor in 1998 covering the Salinas Valley region of California. This dataset contains 224 contiguous spectral bands with a spatial resolution of 512 × 217 pixels and its spatial resolution is 3.7 m. We removed twenty water absorption bands, thereby reducing the number of bands to 204. In this dataset, there are 16 different land cover classes.

4.2. Experimental settings

To ensure the fairness in comparison, all HSIC methods were run in the same configuration environment. The operating system is Ubuntu 20.04.5 LTS, and the configured virtual environment includes Python 3.9.16, PyTorch 1.13.1, CUDA 11.6, etc. The hardware configuration includes a 12th Gen Intel(R) Core(TM) i9-12900KF CPU and an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory.

The available pixels were randomly divided into training, validation, and testing sets. Different proportions of training samples on different datasets were chosen to train the algorithm. It is worth noting that the proportion setting of the selected training set and the validation set is consistent. Tables 1–3 present the sample sizes of the training, validation, and testing sets of three HSI datasets. The training proportions of the IP, HU, and SV datasets were set to approximately 3%, 3%, and 1%, respectively.

The performance evaluation of various HSIC methods was quantified using the following metrics: overall accuracy (OA), average accuracy (AA), and kappa coefficient (KAPPA). In general, the values of these three metrics are close to 1, indicating better classification performance. All HSIC algorithms were trained independently for 100 epochs in each trial. The model with the highest OA on the validation set was selected for testing. Then the OA, AA, and KAPPA were obtained on the testing set as the experimental results. Each experiment was repeated ten times for each model, and the highest OA was selected as the final classification performance of the model. Finally, we generated the classification maps of these methods accordingly.

4.3. Experiment parameter tuning

The classification performances of DNNs are closely related to the hyperparameters setting experimentally, and thus tuning the parameters is an essential step in achieving optimal model performance. The learning rate is a crucial hyperparameter in deep learning networks, and its role is to control the step size in each iteration and enable the loss function to converge toward its minimum value. A larger or smaller learning rate may affect the classification performance of DNNs. Fig. 8 shows the classification results of the proposed model on the three datasets with different learning rates. The results demonstrate that our model achieves the best performance on three datasets when the learning rate is set to 0.008. Therefore, we set a fixed learning rate of 0.008 for all experiments.

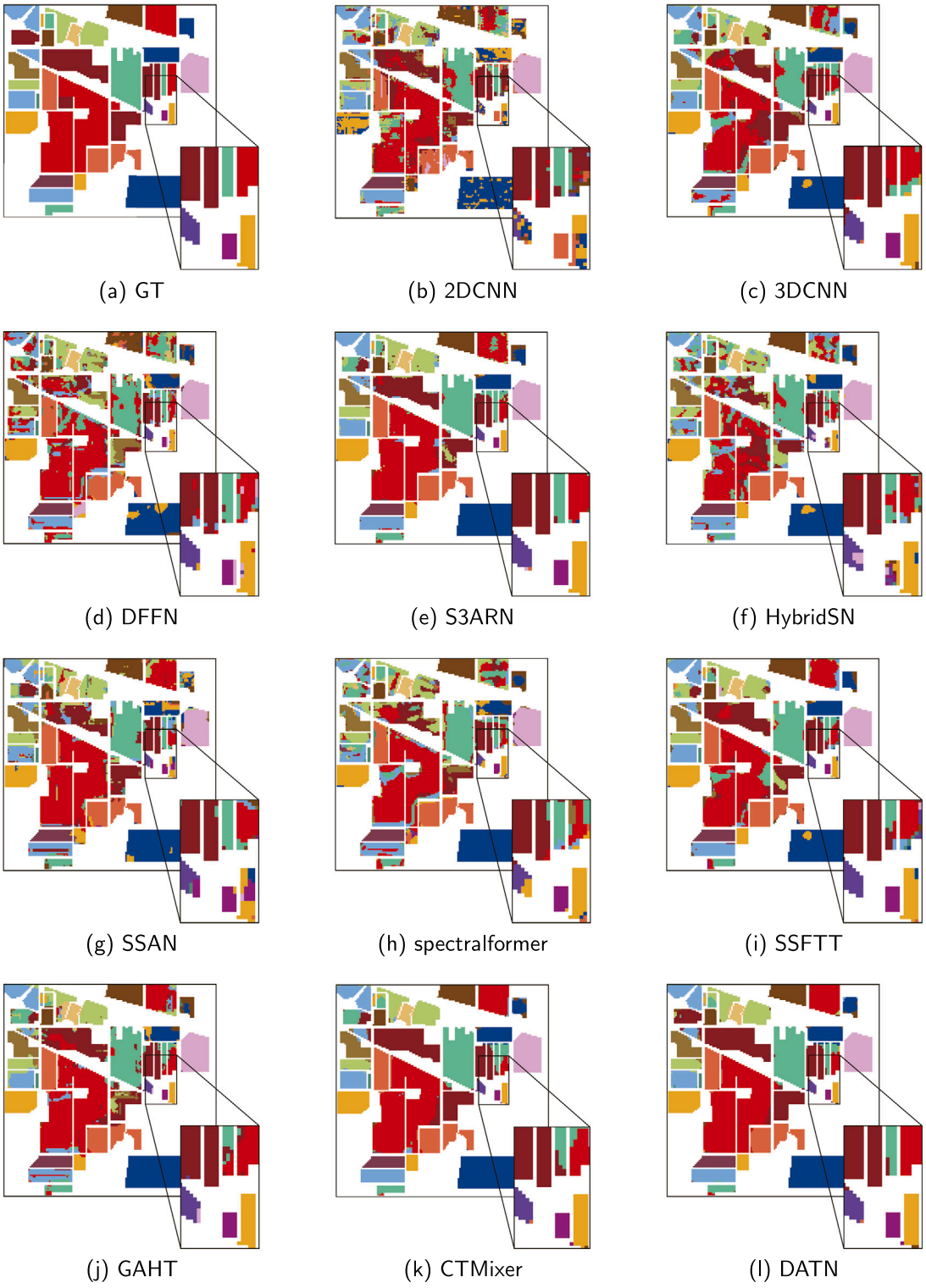


Fig. 5. Classification maps of different models on the IP dataset.

The patch size setting can also affect the experimental results. Generally, a patch with a larger size can encompass more information. The proposed spatial self-attention partitions the patches in the spatial dimension. It is worth noting that this operation cannot be performed by setting a small patch size, and the effectiveness of the module is also affected if the pixel information in each block is too small. On the other hand, by setting a larger patch size, it may contain rich spatial information. However, these edge pixels also affect the classification performance of the central pixel to a certain extent. The powerful local information extraction ability of spatial self-attention further amplifies this effect. Thus, we set the patch size to 11×11 in our experiments. We also established a batch size of 64 and conducted 100 epochs with a momentum of 0.9 and a weight decay of 0.0001.

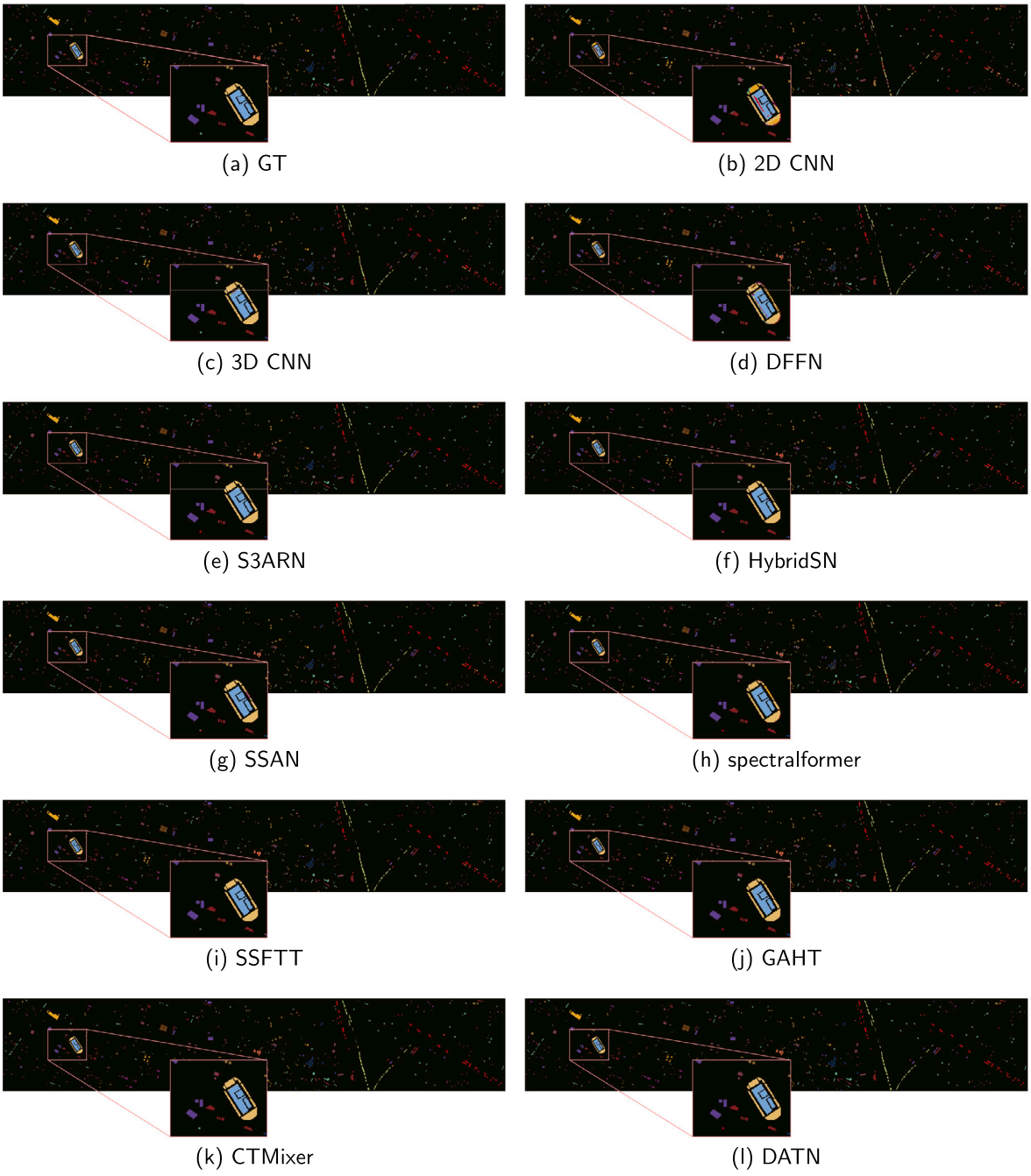


Fig. 6. Classification maps of different models on the HU dataset.

4.4. Parameter analysis

The SSHT module proposed in this study involves two crucial hyperparameters: the block size and the number of heads. To investigate the effects of these parameters on performance across different HSI datasets, we conducted two experiments in this subsection.

In the experiments, the embedding dimensions of the SSHT module in four stages were set to 64, 256, 64, and 256, respectively. To test the influence of different sizes of the division block, we correspondingly fixed the number of heads to 8 and 4, when the embedding dimensions were set to 256 and 64, respectively. We experimented with different sizes of the division block, ranging from 2 to 6, to find the optimal size for each dataset. Fig. 9 illustrates the classification results of our DATN approach across three datasets. On the HU and SV datasets, the optimal division size was set to 3. On the IP dataset, the optimal division size was set to 4.

Then, we fixed the optimal size of the dividing block and experimented with different numbers of heads. In addition, when the embedding dimension was 256, the number of heads was set to 4, 8, and 16, respectively. When the embedding dimension was 64, the number of heads was set to 2, 4, and 8, respectively. Fig. 10 shows the classification performances of different head numbers when the embedding dimensions are 256 and 64, respectively. We notice that the proposed method is notably influenced by the block size and the number of heads on the IP dataset. This is because the pixel distribution in the IP dataset is less smooth compared to the HU and SV datasets and the value change between adjacent pixels is to some extent discontinuous.

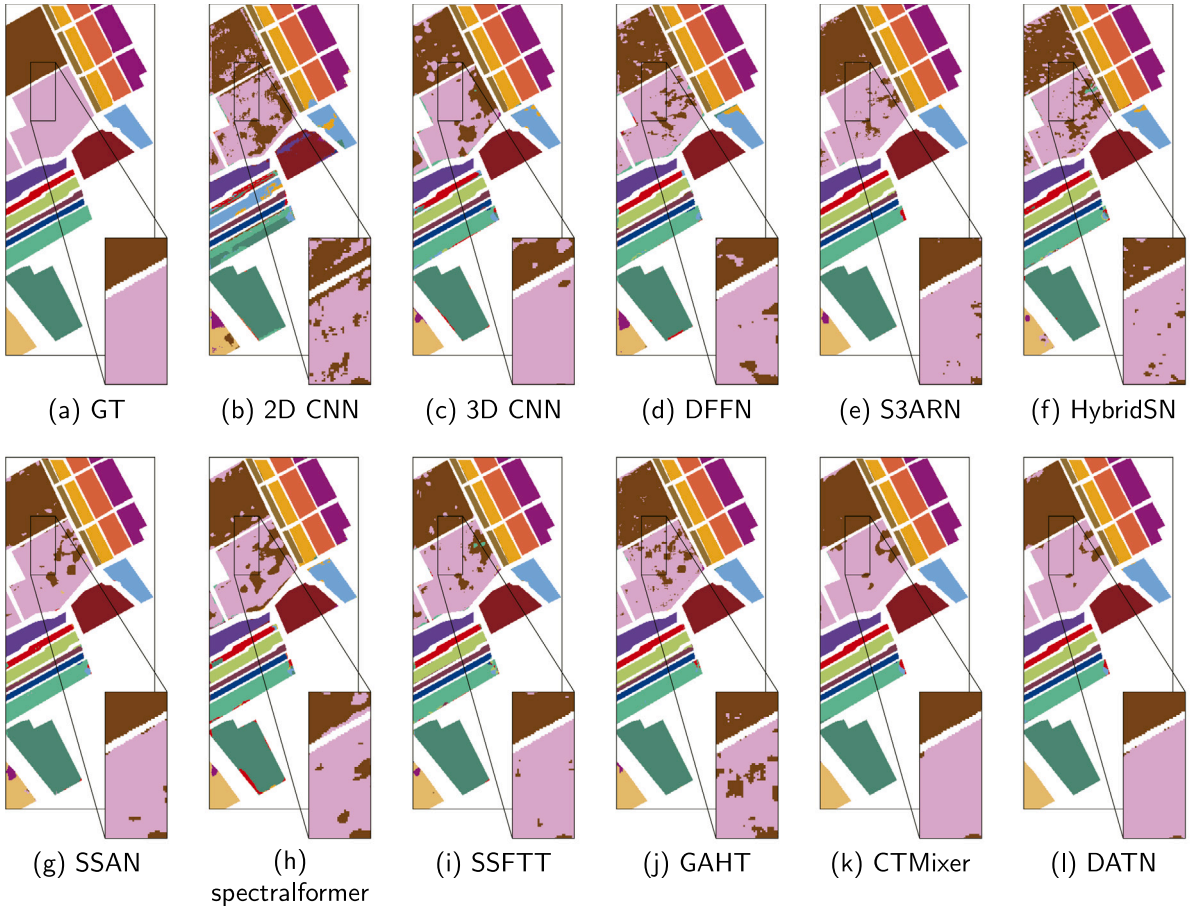


Fig. 7. Classification maps of different models on the SV dataset.

Table 4
Classification results of different models on the IP dataset.

No.	2D CNN	3D CNN	DFFN	HybridSN	SSAN	S3ARN	spectralformer	GAHT	SSFTT	CTMixer	DATN
1	41.18	100.00	91.89	87.50	82.76	100.00	65.62	91.43	83.33	100.00	100.00
2	62.52	68.75	56.38	62.09	82.52	92.96	75.25	83.99	87.07	91.19	94.73
3	71.56	73.33	59.94	53.52	72.82	88.12	59.29	79.53	85.84	92.49	98.24
4	60.44	64.94	34.33	65.76	68.63	88.32	44.12	64.52	74.48	91.46	91.92
5	45.09	86.90	70.26	83.30	70.29	99.06	76.96	88.98	90.22	98.38	98.60
6	77.25	97.92	89.98	98.25	87.38	98.10	93.95	96.30	95.14	95.30	97.37
7	00.00	51.43	100.00	100.00	33.33	100.00	58.62	94.74	69.23	94.74	94.12
8	85.48	98.83	84.51	96.79	98.20	98.83	95.96	97.90	92.95	97.24	98.84
9	45.45	45.45	22.22	75.00	16.39	100.00	31.25	47.62	58.82	90.91	71.43
10	71.51	61.93	52.92	55.95	86.14	82.91	58.36	77.47	77.81	82.17	94.38
11	79.64	78.79	68.64	77.11	92.30	94.59	77.76	88.08	91.91	96.32	96.09
12	45.21	68.68	47.49	42.54	55.09	76.18	41.89	67.90	70.16	94.56	90.96
13	71.23	96.20	89.53	93.94	86.29	94.48	76.50	91.18	93.33	96.27	97.48
14	86.69	99.54	98.19	98.55	96.26	99.48	98.96	99.81	98.62	98.98	98.74
15	70.03	74.04	72.46	83.08	68.47	80.34	73.01	72.51	80.24	87.01	86.75
16	58.97	77.42	65.77	75.00	57.60	75.00	59.17	57.26	75.26	82.76	84.71
OA(%)	70.52	78.83	67.36	72.65	81.99	91.64	72.78	84.96	87.47	93.42	**95.69**
AA(%)	60.76	77.76	69.03	78.02	72.15	91.77	67.91	81.20	82.77	93.11	**93.39**
KAPPA(%)	66.59	75.98	62.97	69.14	79.53	90.45	69.22	82.85	85.76	92.48	**95.05**

4.5. Experimental results

In this subsection, extensive experiments on three benchmark datasets were conducted to evaluate the efficacy of our proposed DATN method against several state-of-the-art approaches, such as 2D CNN (Chen et al., 2016), 3D CNN (Li et al., 2017), deep feature fusion network (DFFN) (Song et al., 2018), HybridSN (Roy et al., 2019), spatial–spectral attention networks (SSAN) (Mei et al., 2019), S3ARN (Shu et al., 2022), spectralformer (Hong et al., 2021), group-aware hierarchical transformer (GAHT) (Mei et al., 2022), spatial–spectral feature tokenization transformer (SSFTT) (Sun et al., 2022), and CT-mixer (Zhang et al., 2022). Tables 4–6 present the classification results of these methods on the three benchmark datasets.

(1) Quantitative Evaluation: It can be seen that the OA, AA, and KAPPA of 2D CNN are 71.53%, 77.50%, and 69.14% on the HU dataset, respectively. Therefore, it obtains the worst results among all HSIC methods. The main reason is that it ignored the information

Table 5
Classification results of different models on the HU dataset.

No.	2D CNN	3D CNN	DFFN	HybridSN	SSAN	S3ARN	spectralformer	GAHT	SSFTT	CTMixer	DATN
1	92.46	93.12	92.79	95.89	90.17	92.46	95.90	91.89	94.36	93.33	91.31
2	71.92	90.89	86.39	91.33	90.49	96.68	85.45	98.39	91.09	90.25	97.79
3	100.00	100.00	97.35	99.70	94.36	98.39	97.19	100.00	99.55	100.00	100.00
4	85.46	98.00	96.21	98.82	96.69	98.19	95.30	98.53	98.27	98.90	97.97
5	76.00	97.73	96.42	97.75	97.61	99.31	92.26	97.96	99.22	99.31	100.00
6	93.85	98.35	100.00	99.15	9245	100.00	87.63	88.67	92.83	100.00	99.64
7	74.77	92.60	82.95	90.93	86.96	96.82	92.08	95.51	94.99	95.05	97.48
8	71.56	96.44	81.16	94.69	97.18	98.13	88.66	98.59	93.84	98.51	99.81
9	65.61	81.68	85.63	89.68	89.51	90.21	85.37	88.14	88.08	91.80	94.41
10	49.83	84.04	95.32	89.74	89.96	96.09	85.44	91.05	95.53	92.37	94.77
11	59.75	77.12	78.62	92.63	82.68	96.73	87.19	93.77	94.61	98.58	94.37
12	48.81	79.06	68.36	86.22	93.11	88.17	78.42	91.74	86.04	98.84	95.69
13	91.89	91.67	75.00	93.04	85.81	94.73	92.09	92.72	93.41	92.94	91.76
14	92.36	97.60	76.98	95.75	99.51	96.17	88.12	96.42	95.59	100.00	96.22
15	88.35	92.85	93.03	93.66	95.15	96.37	94.39	97.11	96.96	94.53	97.25
OA(%)	71.53	89.82	86.15	93.25	91.64	95.46	88.99	94.72	93.90	95.83	**96.41**
AA(%)	77.50	91.40	87.08	93.93	92.10	95.89	89.69	94.70	94.29	96.29	**96.56**
KAPPA(%)	69.14	89.00	85.01	92.71	90.96	95.09	88.08	94.29	93.41	95.50	**96.12**

Table 6
Classification results of different models on the SV dataset.

No.	2D CNN	3D CNN	DFFN	HybridSN	SSAN	S3ARN	spectralformer	GAHT	SSFTT	CTMixer	DATN
1	82.90	99.84	99.53	100.00	100.00	100.00	100.00	100.00	99.43	100.00	100.00
2	98.98	100.00	99.75	99.94	100.00	100.00	98.73	100.00	99.59	100.00	100.00
3	44.70	88.74	93.67	95.80	96.13	99.36	94.59	98.68	97.51	96.42	96.01
4	91.59	99.63	97.69	98.62	98.10	98.98	97.97	99.41	94.36	98.61	98.54
5	84.64	95.37	93.24	96.48	99.84	98.27	93.14	99.61	99.72	99.65	99.15
6	97.52	99.97	99.97	100.00	99.95	100.00	99.74	99.95	99.58	100.00	100.00
7	89.48	95.57	92.10	97.54	94.35	96.21	95.93	97.57	97.07	99.91	100.00
8	86.42	90.48	93.25	91.08	97.43	98.12	92.97	97.31	97.04	98.40	98.34
9	78.26	96.18	98.23	98.36	99.87	99.46	96.85	99.61	99.38	99.25	99.71
10	82.45	89.89	87.75	94.27	97.53	99.52	97.17	99.37	94.02	99.74	99.84
11	88.61	80.61	88.42	90.11	95.53	92.83	65.19	95.94	94.76	93.63	94.37
12	00.00	98.04	97.24	96.70	95.43	92.65	97.62	91.89	94.23	99.74	99.32
13	98.75	95.62	97.08	99.22	98.28	99.34	88.09	99.73	88.97	99.89	100.00
14	94.91	98.51	97.39	98.16	98.73	99.13	99.70	99.42	98.65	99.04	99.61
15	62.71	71.83	82.62	76.64	84.72	87.11	74.13	81.87	81.94	88.38	91.78
16	96.68	100.00	97.14	98.36	98.27	99.82	97.19	99.07	98.94	100.00	100.00
OA(%)	80.36	91.10	93.39	93.16	96.02	96.74	91.71	95.84	95.06	97.49	**98.07**
AA(%)	79.91	93.76	94.69	95.70	97.13	97.54	93.06	97.46	95.94	98.29	**98.54**
KAPPA(%)	78.14	90.11	92.65	92.40	95.57	96.37	90.79	95.38	94.51	97.21	**97.84**

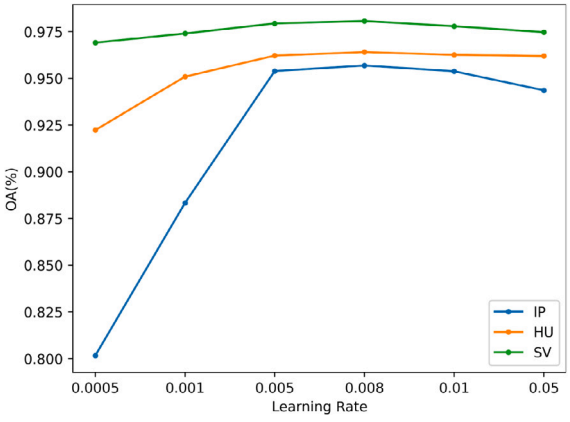


Fig. 8. The classification performance of the proposed DATN method under different learning rate.

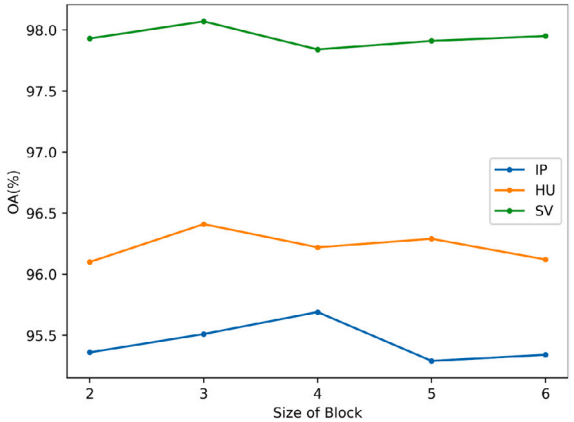


Fig. 9. The classification performance of the proposed DATN method under different size of division block.

contained in the large number of bands in HSIs. Although 3D CNN can achieve better classification performance (89.82%, 91.40%, and 89%) on this dataset than 2D CNN, the sole stacking of 3D convolutional blocks causes a significant redundancy of spectral information, and thus cannot achieve satisfactory classification performance. It is noteworthy that the HybridSN method obtains significant performance improvement by integrating 2D CNN and 3D CNN. Compared to 3D CNN, the OA of HybridSN increased by 3.43% and 2.06% on the HU and SV datasets, respectively. It indicates that the reasonable integration

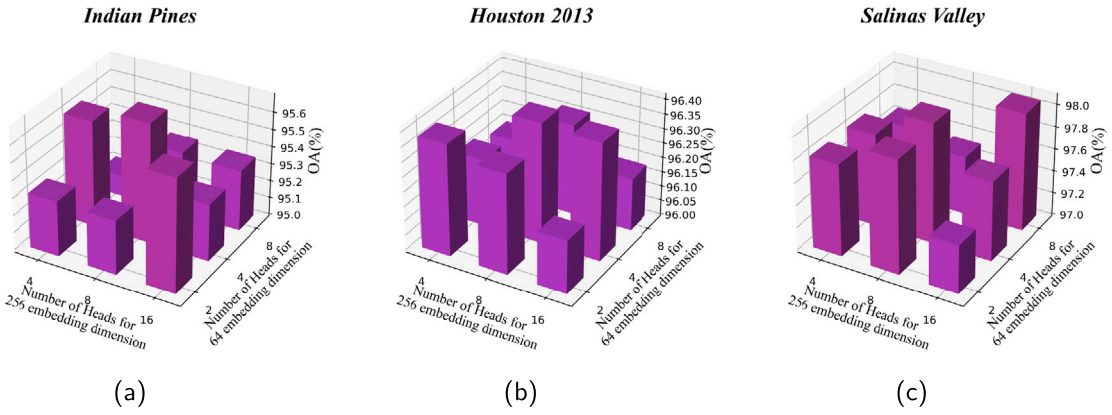


Fig. 10. The classification performance of the proposed DATN method under different number of heads.

of 2D CNN and 3D CNN can leverage the local contextual information extraction capabilities of CNNs in HSIC problems. Both SSAN and S3ARN integrate the attention mechanism into CNN to further improve the performance. Especially for S3ARN, the OA values of the classification results reach 91.64%, 95.46%, and 96.74% on the IP, HU, and SV datasets, respectively. The results demonstrate that the effective utilization of the attention mechanism can enhance the classification performance of HSIC methods.

In general, the transformer aims at modeling global dependencies and thus achieves excellent performance. However, it can be seen that the spectralformer model cannot outperform the CNN-based HSIC methods on three datasets in some cases. This is because it captures the feature information solely through self-attention without considering the local contextual information in CNNs, and the lack of inductive biases built-in CNNs makes it difficult for spectralformer to deal with the position relationships between pixels in HSIs. Compared to spectralformer, the classification performances of both SSFTT and GAHT have been significantly improved on different datasets. In addition, the experimental results reveal that CTmixer achieves the second-best classification performance among transformer-based methods on the three datasets. This is because these HSIC methods leverage the advantages of both CNN and self-attention and thus effectively extract the global and local feature information from HSIs.

Notably, our proposed DATN approach outperforms other CNN-based and transformer-based HSIC methods on all three datasets. For OA evaluation indicators, our DATN method achieves 96.41% on the HU dataset, outperforming 2D CNN, 3D CNN, DFFN, HybridSN, SSAN, S3ARN, spectralformer, GAHT, SSFTT, and CTMixer by 24.88%, 6.59%, 10.26%, 3.16%, 4.77%, 0.95%, 7.42%, 1.69%, 2.51% and 0.58%, respectively. Since the HU dataset contains a large number of unlabeled pixels, the pixels available for classification are mostly discrete. Nevertheless, our method still achieves considerable classification results, especially in the categories “Soil” and “Commercial”. This is because the SSHT module in DATN not only effectively extracts both global spatial and spectral information of HSIs but also utilizes chunked spatial tokens as the input to obtain local spatial contextual information. Moreover, the SLCB module is further integrated into our framework to preserve the local spectral information of HSIs. Therefore, our proposed DATN approach exhibits superior performance in extracting both global and local spatial–spectral feature information compared to other competitors.

It is worth noting that the performances of most HSCI methods significantly decrease on the IP dataset due to the insufficient number of training samples, especially 3D CNN-based and pure ViT methods that require a large number of training samples. However, even with a limited number of training samples, the proposed DATN method can achieve satisfactory classification performance with the highest OA (95.69%), AA (93.39%), and Kappa (95.05%), improving the OA values by 2.27% (ctmixer), 8.22% (SSFTT), 10.73% (GAHT), and 22.91% (spectralformer), 25.17% (2DCNN), 16.86% (3D CNN), 13.70% (SSAN), 23.04% (HybridSN), and 4.05% (S3ARN). It can be seen that the proposed method possesses excellent generalization ability and requires less training data to fully learn the feature distribution of HSIs. Since there are sufficient samples to train the models on the HU and SV datasets, the performance improvement of our model on these two datasets is relatively less significant compared to the IP dataset. Nevertheless, our proposed DATN method still achieves the best classification performance on larger datasets.

(2) Visual Evaluation: Figs. 5–7 present the visualization classification results of various HSIC methods on the three datasets. The results clearly demonstrate that our proposed method has the fewest misclassified pixels and achieves the best classification performance. Note that, the HU dataset contains fewer bands, and the pixel distribution is more scattered. Most methods exhibit more classification errors in the “Soil” category, while the proposed method still maintains a lower level of noise in the classification map. Additionally, due to the relatively unsmooth pixel distribution on the IP dataset, CNN-based HSIC methods achieve poor performance in classifying edge pixels. By fusing global–local features, the networks combining CNN and transformer obtain better edge pixel classification results. Especially, our proposed DATN method still exhibits stronger classification ability than other methods.

4.6. Ablation study

To validate the efficacy of each module in our proposed model, we conducted four ablation experiments with the optimal parameters obtained from the previous experiments. The ablation results of our proposed approach are reported in Table 7.

In the first experiment, we connected the four layers of MHSA and aimed to verify the superiority of the proposed module.

In the second experiment, we aimed to verify the contribution of the SSHT module in HSIC. Specifically, we removed this module from the proposed network and connected the remaining two spectral local-conv blocks. It aims to verify the impact of the global modeling ability using the SSHT module on the classification results. We can know that the performance degradation of the algorithm can be caused by the abandonment of the SSHT module.

In the third experiment, we removed the SLCB module from the proposed model to verify whether the SSHT module without the local information extraction module could achieve the same classification performance. It can be observed that the proposed DATN method outperforms this variant without the SLCB module on three datasets. This demonstrates the effectiveness of the SLCB module in extracting local information and its contribution to the overall classification performance. Meanwhile, it highlights the significance of integrating local and global information to enhance the accuracy of HSIC applications.

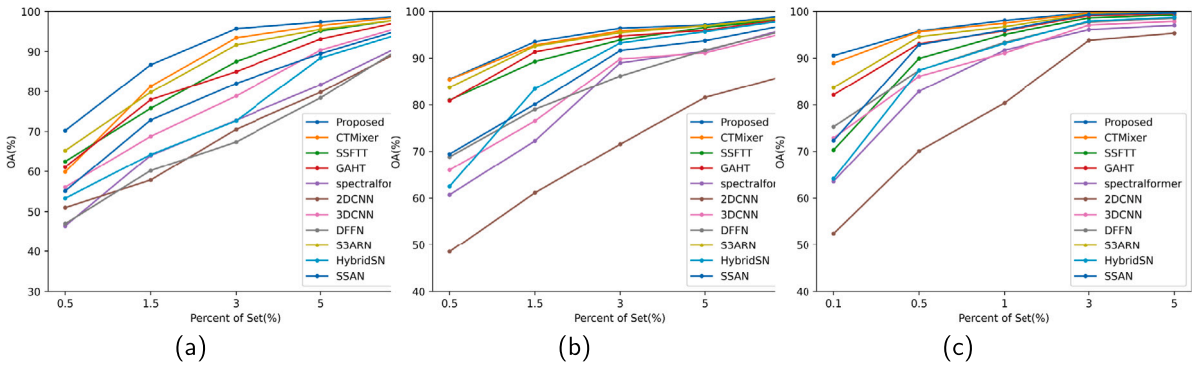


Fig. 11. The results of all models under different training set ratios. (a) IP (b) HU (c) SV.

Table 7
Ablation results of the proposed model on three datasets.

Case	Component			IP			HU			SV		
	SSHT	SLCB	Attention branch	OA(%)	AA(%)	KAPPA(%)	OA(%)	AA(%)	KAPPA(%)	OA(%)	AA(%)	KAPPA(%)
1	×	×	✓	86.53	81.42	84.67	94.19	94.97	93.72	95.72	96.90	95.23
2	×	✓	×	95.20	93.19	94.50	96.05	96.37	97.67	97.67	98.22	97.41
3	✓	×	✓	95.00	94.23	94.27	96.04	96.23	95.72	97.20	97.95	96.89
4	×	✓	✓	91.84	86.31	90.67	95.57	95.83	95.21	97.10	97.85	96.78
5	✓	✓	✓	95.69	93.39	95.05	96.41	96.56	96.12	98.07	98.54	97.84

In the fourth experiment, we evaluated the contribution of the proposed SSHT module in our proposed framework. Therefore, one SSHT module was replaced by two MHSA modules. For a fair comparison, we set the number of heads to be consistent. The experiments on three HSI datasets have demonstrated that the proposed method using the SSHT module outperforms the method using the MHSA block in terms of classification performance. This is because the proposed SSHT can more effectively utilize the spatial–spectral information of HSIs, thereby improving classification performances.

Overall, the ablation results in Table 7 consistently demonstrate that our proposed method achieves better performance than these three variants on different HSI datasets. It indicates that each module of our proposed approach makes a substantial contribution to enhancing the performance of HSIC tasks.

4.7. Influence of training set size

To assess the generalization performance of the proposed network, we conducted classification experiments using different training set ratios. Specifically, we randomly selected 0.5%, 1.5%, 3%, 5% and 10% the samples from the IP and HU datasets, and 0.1%, 0.5%, 1%, 3% and 5% samples for the SV dataset as the training set, respectively. In addition, the same number of samples was also selected as the validation set for each experiment. Under the same experimental configuration, we also run ten experiments for each method and took the best OA as the classification result. The classification performances of different methods are shown in Fig. 11.

Fig. 11 illustrates that the proposed method achieves superior performance compared to other HSIC methods across different ratios of training samples. Besides, we can see that the performance of all HSIC methods steadily improves as the number of training samples increases. Notably, our proposed model exhibits significant superiority over other algorithms, particularly when the training set ratio is relatively small.

5. Conclusion

In this paper, we propose a novel HSIC method, namely dual attention transformer network (DATN), which mainly consists of four SSHT modules and two SLCB modules. Specifically, the self-attention mechanism is applied to both spectral and spatial dimensions of HSIs in our proposed SSHT module, thereby allowing the network to describe the long-range dependencies of both spatial and spectral features separately. Moreover, it also effectively learns the local spatial and position information from HSIs. Additionally, the proposed SLCB module is used to explore the local spectral information and can effectively compensate for the shortcomings of feature maps extracted by multi-layer SSHT. Finally, the learned features are fused and then fed into the classifier, and the classification results of HSIs can be obtained. The experimental results obtained from three mainstream HSI datasets have demonstrated that our DATN approach surpasses several state-of-the-art HSIC methods. However, although applying MHSA to the spectral dimension is a good way to model global information, its remarkable performance comes with heavy computational costs and a huge amount of parameters. In the future, we will design a lightweight network to reduce the parameters and complexity of the model.

CRediT authorship contribution statement

Zhenqiu Shu: Formulation or evolution of overarching research goals and aims, Writing – review & editing. **Yuyang Wang:** Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection, Designing computer programs, Preparation, creation and/or presentation of the published work, specifically writing the initial draft. **Zhengtao Yu:** Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant Nos. 61603159, 62162033, U21B2027], Yunnan Provincial Major Science and Technology Special Plan Projects, China [Grant Nos. 202002AD080001, 202103AA080015], Yunnan Foundation Research Projects, China [Grant Nos. 202101AT070438, 202101BE070001-056], Yunnan Xingdian Talent Support Plan Project, China. All authors approved the final version of manuscript to be published.

References

Bai, Jing, Wen, Zheng, Xiao, Zhu, Ye, Fawang, Zhu, Yongdong, Alazab, Mamoun, Jiao, Licheng, 2022. Hyperspectral image classification based on multibranch attention transformer networks. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Benediktsson, Jón Atli, Palmason, Jón Aevar, Sveinsson, Johannes R., 2005. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. 43 (3), 480–491.

Cao, Xiangyong, Yao, Jing, Xu, Zongben, Meng, Deyu, 2020. Hyperspectral image classification with convolutional neural network and active learning. IEEE Trans. Geosci. Remote Sens. 58 (7), 4604–4616.

Cariou, Claude, Chehdi, Kacem, 2015. Unsupervised nearest neighbors clustering with application to hyperspectral images. IEEE J. Sel. Top. Sign. Proces. 9 (6), 1105–1116.

Chen, Yushi, Jiang, Hanlu, Li, Chunyang, Jia, Xiuping, Ghamisi, Pedram, 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 54 (10), 6232–6251.

Cui, Benlei, Dong, Xue-Mei, Zhan, Qiaoqiao, Peng, Jiangtao, Sun, Weiwei, 2021. LiteDepthwiseNet: A lightweight network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Dang, Lanxue, Weng, Libo, Hou, Yane, Zuo, Xianyu, Liu, Yang, 2023. Double-branch feature fusion transformer for hyperspectral image classification. Sci. Rep. 13 (1), 272.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, Houlsby, Neil, 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale.

Graves, Alex, Graves, Alex, 2012. Long short-term memory. Supervised sequence labelling with recurrent neural networks 37–45.

Ham, Jisoo, Chen, Yangchi, Crawford, Melba M., Ghosh, Joydeep, 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43 (3), 492–501.

Himeur, Yassine, Ghanem, Khalida, Alsalemi, Abdullah, Bensaali, Faycal, Amira, Abbes, 2021. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Appl. Energy 287, 116601.

Himeur, Yassine, Rimal, Bhagawat, Tiwary, Abhishek, Amira, Abbes, 2022. Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. Inf. Fusion 86, 44–75.

Hong, Danfeng, Han, Zhu, Yao, Jing, Gao, Lianru, Zhang, Bing, Plaza, Antonio, Chanussot, Jocelyn, 2021. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Hu, Wei, Huang, Yangyu, Wei, Li, Zhang, Fan, Li, Hengchao, 2015. Deep convolutional neural networks for hyperspectral image classification. J. Sensors 2015, 1–12.

Hu, Jie, Shen, Li, Sun, Gang, 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Jayanetti, J.A.A.M, Meedeniya, D.A., Dilini, M.D.N., Wickramapala, M.H., Madushanka, J.H., 2017. Enhanced land cover and land use information generation from satellite imagery and foursquare data. In: Proceedings of the 6th International Conference on Software and Computer Applications. ICSCA '17, Association for Computing Machinery, New York, NY, USA, pp. 149–153.

Khan, Uzair, Paheding, Sidike, Elkin, Colin P., Devabhaktuni, Vijaya Kumar, 2021. Trends in deep learning for medical hyperspectral image analysis. IEEE Access 9, 79534–79548.

Li, Zhixin, Lin, Lan, Zhang, Canlong, Ma, Huifang, Zhao, Weizhong, Shi, Zhiping, 2021. A semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 17 (1), 1–23.

Li, Zhaokui, Liu, Ming, Chen, Yushi, Xu, Yimin, Li, Wei, Du, Qian, 2022. Deep cross-domain few-shot learning for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Li, Shutao, Song, Weiwei, Fang, Leyuan, Chen, Yushi, Ghamisi, Pedram, Benediktsson, Jon Atli, 2019. Deep learning for hyperspectral image classification: An overview. IEEE Trans. Geosci. Remote Sens. 57 (9), 6690–6709.

Li, Ying, Zhang, Haokui, Shen, Qiang, 2017. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. Remote Sens. 9 (1), 67.

Li, Rui, Zheng, Shunyi, Duan, Chenxi, Yang, Yang, Wang, Xiqi, 2020. Classification of hyperspectral image based on double-branch dual-attention mechanism network. Remote Sens. 12 (3), 582.

Long, Yaqian, Wang, Xun, Xu, Meng, Zhang, Shuyu, Jiang, Shuguo, Jia, Sen, 2023. Dual self-attention swin transformer for hyperspectral image super-resolution. IEEE Trans. Geosci. Remote Sens..

Ma, Li, Crawford, Melba M., Tian, Jinwen, 2010. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 48 (11), 4099–4109.

Mahakalanda, Indra, Demotte, Piyumal, Perera, Indika, Meedeniya, Dulani, Wijesuriya, Wasana, Rodrigo, Lakshman, 2022. Deep learning-based prediction for stand age and land utilization of rubber plantation. In: Khan, Mohammad Ayoub, Khan, Rijwan, Ansari, Mohammad Aslam (Eds.), Application of Machine Learning in Agriculture. Academic Press, pp. 131–156, (Chapter 7).

Meedeniya, D.A., Jayanetti, J.A.A.M., Dilini, M.D.N., Wickramapala, M.H., Madushanka, J.H., 2020. Land-use classification with integrated data. Mach. Vis. Insp. Syst. Image Process. Concepts Methodol. Appl. 1, 1–36.

Meedeniya, D.A., Mahakalanda, I., Lenadora, D.S., Perera, I., Hewawalpita, S.G.S., Abeysinghe, C., Nayak, Soumya Ranjan, 2022. Prediction of paddy cultivation using deep learning on land cover variation for sustainable agriculture. In: Poonia, Ramesh Chandra, Singh, Vijander, Nayak, Soumya Ranjan (Eds.), Deep Learning for Sustainable Agriculture. In: Cognitive Data Science in Sustainable Computing, Academic Press, pp. 325–355 (Chapter 13).

Mei, Xiaoguang, Pan, Erting, Ma, Yong, Dai, Xiaobing, Huang, Jun, Fan, Fan, Du, Qinglei, Zheng, Hong, Ma, Jiayi, 2019. Spectral-spatial attention networks for hyperspectral image classification. Remote Sens. 11 (8), 963.

Mei, Shaohui, Song, Chao, Ma, Mingyang, Xu, Fulin, 2022. Hyperspectral image classification using group-aware hierarchical transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Melgani, Farid, Bruzzone, Lorenzo, 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 42 (8), 1778–1790.

Pan, Zhaojie, Ding, Sunjinyan, Sun, Genyun, Zhang, Aizhu, Jia, Xiuping, Fu, Hang, 2023. Multi-scale spectral-spatial dual-transformer network for hyperspectral image classification. Int. J. Remote Sens. 44 (7), 2480–2494.

Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A., 2019. Deep learning classifiers for hyperspectral imaging: A review. ISPRS J. Photogramm. Remote Sens. 158, 279–317.

Peyghambari, Sima, Zhang, Yun, 2021. Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: an updated review. J. Appl. Remote Sens. 15 (3), 031501.

Qi, Wenchao, Huang, Changping, Wang, Yibo, Zhang, Xia, Sun, Weiwei, Zhang, Lifu, 2023. Global-local three-dimensional convolutional transformer network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens.

Roy, Swalpa Kumar, Krishna, Gopal, Dubey, Shiv Ram, Chaudhuri, Bidyut B., 2019. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 17 (2), 277–281.

Shu, Zhenqiu, Liu, Zigao, Zhou, Jun, Tang, Songze, Yu, Zhengtao, Wu, Xiao-Jun, 2022. Spatial–spectral split attention residual network for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 419–430.

Song, Ruoxi, Feng, Yining, Cheng, Wei, Mu, Zhenhua, Wang, Xianghai, 2022. BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Song, Weiwei, Li, Shutao, Fang, Leyuan, Lu, Ting, 2018. Hyperspectral image classification with deep feature fusion network. IEEE Trans. Geosci. Remote Sens. 56 (6), 3173–3184.

Sun, Genyun, Pan, Zhaojie, Zhang, Aizhu, Jia, Xiuping, Ren, Jinchang, Fu, Hang, Yan, Kai, 2023. Large kernel spectral and spatial attention networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens..

Sun, Le, Zhao, Guangrui, Zheng, Yuhui, Wu, Zebin, 2022. Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Tang, Ping, Zhang, Meng, Liu, Zhihui, Song, Rong, 2023. Double attention transformer for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 20, 1–5.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Xiaolong, Girshick, Ross, Gupta, Abhinav, He, Kaiming, 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, Wenxuan, Liu, Leiming, Zhang, Tianxiang, Shen, Jiachen, Wang, Jing, Li, Jiangyun, 2022. Hyper-ES2T: efficient spatial–spectral transformer for the classification of hyperspectral remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 113, 103005.

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, Kweon, In So, 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Xian, Tiantao, Li, Zhixin, Tang, Zhenjun, Ma, Huifang, 2022a. Adaptive path selection for dynamic image captioning. IEEE Trans. Circuits Syst. Video Technol. 32 (9), 5762–5775.

Xian, Tiantao, Li, Zhixin, Zhang, Canlong, Ma, Huifang, 2022b. Dual global enhanced transformer for image captioning. Neural Netw. 148, 129–141.
Xu, Qin, Xiao, Yong, Wang, Dongyue, Luo, Bin, 2020. CSA-MSO3DCNN: Multiscale octave 3D CNN with channel and spatial attention for hyperspectral image classification. Remote Sens. 12 (1), 188.
Xu, Qin, Yang, Chao, Tang, Jin, Luo, Bin, 2022. Grouped bidirectional LSTM network and multistage fusion convolutional transformer for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–14.
Xue, Xizhe, Zhang, Haokui, Fang, Bei, Bai, Zongwen, Li, Ying, 2022. Grafting transformer on automatically designed convolutional neural network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–16.
Yang, Xiwei, Li, Zhixin, Zhong, Xinfang, Zhang, Canlong, Ma, Huifang, 2023. Mining graph-based dynamic relationships for object detection. Eng. Appl. Artif. Intell. 126, 106928.
Yang, Liming, Yang, Yihang, Yang, Jinghui, Zhao, Ningyuan, Wu, Ling, Wang, Liguo, Wang, Tianrui, 2022. FusionNet: a convolution–transformer fusion network for hyperspectral image classification. Remote Sens. 14 (16), 4066.
Zhang, Weixing, Liu, Sanchao, 2010. Applications of the small satellite constellation for environment and disaster monitoring and forecasting. Int. J. Disaster Risk Sci. 1, 9–16.
Zhang, Junjie, Meng, Zhe, Zhao, Feng, Liu, Hanqiang, Chang, Zhenhui, 2022. Convolution transformer mixer for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Zhao, Wenzhi, Du, Shihong, 2016. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. IEEE Trans. Geosci. Remote Sens. 54 (8), 4544–4554.
Zhong, Zilong, Li, Jonathan, Luo, Zhiming, Chapman, Michael, 2017. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. IEEE Trans. Geosci. Remote Sens. 56 (2), 847–858.
Zhu, Min, Huang, Dan, Hu, Xin-Jun, Tong, Wen-Hua, Han, Bao-Lin, Tian, Jian-Ping, Luo, Hui-Bo, 2020a. Application of hyperspectral technology in detection of agricultural products and food: A review. Food Sci. Nutr. 8 (10), 5206–5214.
Zhu, Minghao, Jiao, Licheng, Liu, Fang, Yang, Shuyuan, Wang, Jianing, 2020b. Residual spectral–spatial attention network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 59 (1), 449–462.

Zhenqiu Shu received his Ph.D. degree in control science and engineering at Nanjing University of Science and Technology in 2015. He was a visiting scholar with the School of Information and Communication Technology, Griffith University, QLD, Australia, from November 2014 to May 2015. In February 2021, he joined the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, where he is currently an associate professor. Before joining Kunming University of Science and Technology, he worked as a postdoctoral at Jiangnan University for four years. He has published over 80 research papers in refereed international journals and conferences. His research interests include pattern recognition, computer vision, and machine learning.

Yuyang Wang is currently pursuing toward Master degree at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include remote sensing and environmental informatics.

Zhengtao Yu received his Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor at the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language processing, information retrieval, and machine learning.