








# D<sup>2</sup>S<sup>2</sup>BoT: Dual-Dimension Spectral-Spatial Bottleneck Transformer for Hyperspectral Image Classification

Lan Zhang , Yang Wang , Linzi Yang , Jianfeng Chen , Zijie Liu , Lifeng Bian ,  
and Chen Yang , *Member, IEEE*

**Abstract**—Hyperspectral image (HSI) classification has become a popular research topic in recent years, and transformer-based networks have demonstrated superior performance by analyzing global semantic features. However, using transformers for pixel-level HSI classification has two limitations: ineffective capture of spatial-spectral correlations and inadequate exploitation of local features. To address these challenges, we propose a dual-dimension self-attention (D<sup>2</sup>SA) mechanism that fully exploits HSI's high spectral-spatial correlation by using two separate branches to model the global dependence of features from the spectral and spatial dimensions. Additionally, we develop a multilayer residual convolution module that extracts local features and introduces shallow-deep feature interactions to obtain more discriminative representations. Based on these components, we propose a dual-dimension spectral-spatial bottleneck transformer (D<sup>2</sup>S<sup>2</sup>BoT) framework for HSI classification that simultaneously models the local interactions and global dependencies of HSI pixels to achieve high-precision classification. By virtue of the D<sup>2</sup>SA mechanism, the introduced D<sup>2</sup>S<sup>2</sup>BoT framework can produce competitive classification results with a limited number of training samples on three well-known datasets, which we hope will provide a strong baseline for future research on transformers in the field of HSI.

**Index Terms**—Convolutional neural network (CNN), dual-dimension self-attention (D<sup>2</sup>SA) mechanism, hyperspectral image (HSI) classification, remote sensing, transformer.

## I. INTRODUCTION

**H**YPERSPECTRAL imagery (HSI) is a widely utilized form of satellite remote sensing data, characterized by

Manuscript received 15 October 2023; revised 23 November 2023; accepted 28 November 2023. Date of publication 14 December 2023; date of current version 10 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62065003, in part by Guizhou Provincial Science and Technology Projects—ZK [2022] Key-020, General-105, and in part by Renjihe of Guizhou University (2012). (*Corresponding authors: Lifeng Bian; Chen Yang.*)

Lan Zhang, Yang Wang, Linzi Yang, Jianfeng Chen, and Zijie Liu are with the Power Systems Engineering Research Center, Ministry of Education, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China (e-mail: 3689485@qq.com; y.wang.gzu@foxmail.com; 812394230@qq.com; 1414606919@qq.com; 897822862@qq.com).

Lifeng Bian is with the Frontier Institute of Chip and System, Fudan University, Shanghai 200433, China (e-mail: lfbian@fudan.edu.cn).

Chen Yang is with the Power Systems Engineering Research Center, Ministry of Education, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China, and also with the China State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China (e-mail: eliot.c.yang@163.com).

Codes are released at: <https://github.com/HyperSystemAndImageProc/D2S2BoT>.

Digital Object Identifier 10.1109/JSTARS.2023.3342461

pixels containing multiple continuous narrow spectral bands. This attribute enables more accurate identification of material information on the Earth's surface. Consequently, HSI has found extensive applications in diverse fields such as anomaly detection [1], military reconnaissance [2], and vegetation disease analysis [3]. To expand the application scope of HSI, researchers have extensively investigated various data processing techniques, including unmixing [4], super-resolution [5], semantic segmentation [6], and classification [7]. HSI classification has emerged as a prominent research area.

Over the years, researchers have developed several HSI classification methods, aiming to achieve high-accuracy results. These methods fall into two categories: traditional and deep learning-based approaches. Initially, traditional methods employed machine learning techniques such as k-nearest neighbors [7], the Bayesian estimation [8], and support vector machines [9], [10], which used spectral information as input. Later, researchers integrated spectral-spatial information into classification methods. For example, in 2015, Li et al. [11] proposed an integrated learning framework, allowing for the joint utilization of multiple features. In 2017, Lu et al. [12] developed a fusion framework that combined subpixels, pixels, and super-pixels features. Despite their high performance, traditional methods rely heavily on hand-crafted descriptors, which can compromise classification robustness [13].

Compared with traditional methods, DL-based methods are more robust in automatic extraction and representation of high-level image features. In recent years, many computer vision tasks have benefited from DL and made significant breakthroughs, such as natural language processing (NLP) [14], image segmentation [15], [16], and image classification [17], [18]. HSI classification is a typical image classification task, and DL-based methods are widely used in this field [19], [20], [21]. In 2014, Chen et al. [22] proposed a stacked autoencoder-based HSI classification model, introducing deep learning methods to the task. Subsequently, many DL-based backbone networks have been successfully applied to the HSI classification task, such as deep belief network [23], graph neural network [24], convolutional neural network (CNN) [25], [26], recurrent neural network [27], [28], capsule networks [29], and graph convolutional network [30], [31], with CNN emerging as the mainstream.

CNNs are a powerful tool for capturing local correlations due to their shared weights and local connections, making them

well-suited for HSI classification tasks. Similar to traditional methods, the researchers explored CNN-based classification methods from the spectral information of HSI. Hu et al. [32] proposed a 1D CNN with five convolutional layers to extract the spectral features of HSI. Li et al. [33] proposed an HS image classification method based on a pixel-level CNN framework, which can automatically extract hierarchical features from HS pixels. Later, 2D-CNN and 3D-CNN-based classification methods were developed to integrate spatial and spectral information. Xu et al. [34] introduced a dual-branch framework for HSI classification, in which the 1-D CNN branch and 2-D CNN branch are used to explore HSI's spectral and spatial information, respectively. Li et al. [35] proposed an HSI classification framework, which realized the joint extraction of spatial-spectral features using 3D-CNN. In subsequent years, researchers introduced numerous CNN-based structural frameworks to enhance classification performance. Roy et al. [36] proposed the HybridSN framework for HSI classification, which uses 3D-CNN to extract spectral-spatial features and subsequently uses 2D-CNN to learn high-level spatial representations further. Zhong et al. [37] proposed the spectral-spatial residual network (SSRN), wherein the residual spectral block and the residual spatial block sequentially learn features from the spectral and spatial information of HSI, effectively enhancing feature utilization. Li et al. [38] proposed the dual-branch dual-attention mechanism network (DBDA), where 3D-CNNs with different receptive fields separately extracted spatial and spectral features, followed by the application of attention modules in both branches to highlight features.

Recently, a model called transformer [39] has been proposed for NLP, demonstrating promising results in analyzing global long-range dependencies of input data with its unique self-attention mechanism. Given its success in language tasks, researchers have increasingly explored the expansion of transformer-based models in computer vision, with a particular emphasis on applications in hyperspectral image fields. Consequently, numerous variants of the transformer architecture for HSI classification have been proposed. For example, He et al. [40] proposed a spatial-spectral transformer that combines CNNs for extracting HSI spatial features and a modified transformer to capture spectral sequence relationships. Liu et al. [41] proposed a deep spectral-spatial transformer that studies transformer classification results along spatial and spectral dimensions. Sun et al. [42] proposed a spectral-spatial feature Tokenization transformer that utilizes a Gaussian-weighted feature tokenizer to exploit the deep semantic properties of the spectral-spatial features. Roy et al. [43] proposed the Morphormer network, which integrates a self-attention mechanism with morphological operation to better learn spatial-spectral features. Yao et al. [44] proposed an extended vision transformer (ViT) with a parallel architecture, incorporating a cross-modal attention module to effectively fuse spectral-spatial features from HSI and geometric information from Lidar data, enhancing complementary feature learning. However, the self-attention mechanism in the transformer model, which excels at conducting comprehensive analysis and establishing global dependencies within input sequences, encounters limitations when it comes to attending to local information between adjacent sequences

[45]. This constraint hinders its effectiveness in capturing local features among neighboring pixels in HSI data—a crucial consideration in remote sensing HSI data where adjacent pixels often correspond to the same feature [42]. Furthermore, the conventional self-attention mechanism faces a challenge in effectively handling the redundant spectral information present in HSI data, as it extensively analyzes all spectral components and incorporates their global dependencies. This comprehensive analysis poses difficulties in discerning valuable spectral information and accurately extracting discriminative features.

To address these challenges, a transformer-based framework is proposed in this article with the aim of achieving high-accuracy HSI classification. The framework comprises of two key components, namely dual-dimension spectral-spatial bottleneck transformer ( $D^2S^2BoT$ ) and a multilayer residual convolution module. The former is designed to explore the global correlation of HSI to predict classification results, while the latter captures and provides local features. Specifically, in  $D^2S^2BoT$ , a unique dual-dimension self-attention ( $D^2SA$ ) mechanism is introduced to model the spectral-spatial correlation of HSI. Two independent branches are applied to refine the long-range dependencies of the HSI data on the two dimensions, resulting in better joint exploitation of spatial and spectral global information. An efficient linear projection classifier is then introduced to summarize the features learned by  $D^2SAs$  for determining classification results. The multilayer residual convolution module is developed to preprocess HSI data to enhance  $D^2SA$ 's local feature perception. Two residual blocks are used to extract HSI spectral and spatial local features differentially, and shallow-deep local feature interactions are employed to obtain more discriminative representations. By developing local-global features, the  $D^2S^2BoT$  framework effectively models the spectral-spatial interrelationships of HSI, producing competitive classification results on three well-known hyperspectral datasets.

The main contributions of this research are summarized as follows.

- The  $D^2S^2BoT$  framework is proposed for HSI classification, employing a progressive feature extraction strategy from local to global. The framework utilizes CNNs to learn multilayer local features and incorporates a unique bottleneck transformer (BoT) structure to effectively capture and adaptively fuse global spectral and spatial dependencies. Experimental results on three well-known datasets demonstrate that the proposed framework achieves competitive performance even with limited training samples.
- A  $D^2SA$  mechanism is proposed. Through differentiated mapping of HSI features, this mechanism effectively captures the long-term dependencies in both channel and spatial dimensions, while employing an adaptive fusion strategy to enhance information interaction.
- A multilayer residual convolution module is introduced for capturing local features, where a double parallel block structure is employed to extract HSI spectral and spatial features differentially, and shallow-deep feature interactions are further applied to enhance the multilayer feature representation.

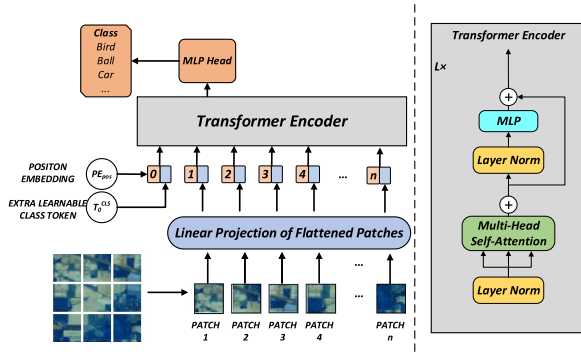


Fig. 1. Architecture of ViT.

The rest of the sections of this article are structured as follows. Section II provides a concise overview of the ViT and BoT. In Section III, our proposed framework is elaborated in detail. Experimental description and analytical results are presented in Section IV. Finally, Section V concludes this article.

## II. RELATED WORK

In this section, we present a succinct overview of two transformer models employed in the domain of computer vision, including the ViT and the BoT.

### A. Vision Transformer

With the popularity in the NLP field, researchers have also applied the transformer model to image classification tasks and have proposed many transformer-based image classification models. In 2020, Dosovitskiy et al. [46] first applied the transformer model to the image classification task, which is called the “vision transformer.” As shown in Fig. 1, the input image of ViT is divided into multiple blocks, and flattened after passing through the patch embedding layer to obtain several vectors called tokens. After concatenated with position information, the tokens are then fed into the transformer encoder to simulate the deep relationships among them. Finally, the features learned by the encoder are sent to a multilayer perceptron (MLP) head to complete the classification.

The multihead self-attention (MHSA) layer is the key component to the encoder of classical ViT architecture, which aims to leverage global information to learn the interrelations between tokens. As shown in Fig. 2, the MHSA layer consists of  $n$  identical head self-attentions, each with three learnable matrices defined in advance: the query matrix  $W^Q$ , the key matrix  $W^K$  and the value matrix  $W^V$ . These input tokens  $X$  are linearly mapped into 3D invariant matrices, including the query matrix  $Q = XW^Q$ , the key matrix  $K = XW^K$ , and the value matrix  $V = XW^V$ . The dot products are used to calculate the  $Q$  with all  $K$ , and then the weights on the  $V$  is calculated by using *softmax* function. Output by the  $i$ th head self-attention can be defined by the

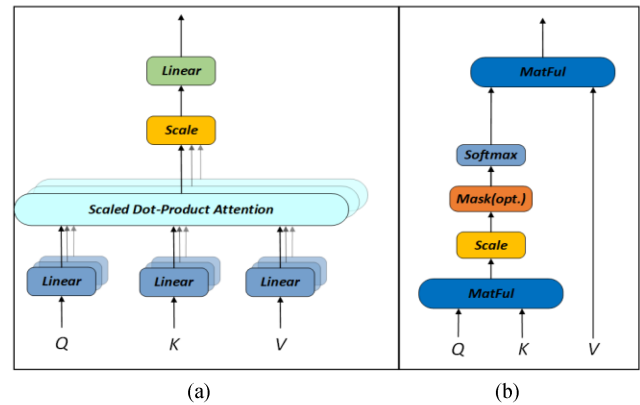


Fig. 2. Attention mechanisms in the transformer encoder, i.e., (a) MHSA and (b) self-attention.

following formula:

$$\begin{aligned} \text{head}_i &= \text{Attention}(Q, K, V) \\ &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \end{aligned} \quad (1)$$

where  $d$  is the dimension of  $K$ .

Then, the output of the MHSA layer can be obtained by concatenating the self-attention results of independent head. Formula (2) can represent this process.

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)W \quad (2)$$

where  $W$  is the parameter matrix and  $\text{Concat}(\cdot)$  is the concatenation operation.

Recently, some ViT-based HSI classification models have been introduced by researchers. In 2020, He et al. [47] proposed a bidirectional encoder representation from transformers that can efficiently analyze global dependencies between HSI pixels. In 2021, He et al. [48] proposed a spectral-spatial HSI classification framework in which a DenseTransformer was used to capture the spectral relationships of HSI. In 2022, Hong et al. [49] proposed the SpectralFormer network architecture based on ViT, which rethinks the HSI classification in terms of the sequential properties of the spectra. Although these HSI classification models based on ViT structure has good performance, they ignore the fundamental differences between sequence-based NLP tasks and image-based visual tasks. For instance, in ViT, the flattening of image blocks into vectors for processing results in the loss of 2D structure and local spatial information [45]. Additionally, ViT exhibits limitations in extracting low-resolution and multiscale features [45], thereby posing a significant challenge to further enhance the classification accuracy of these models.

### B. Bottleneck Transformer

In 2016, the bottleneck structure was first introduced to ResNet by He et al. [50] As shown in Fig. 3(a), the classical bottleneck block consists of three convolutional layers: a  $(1 \times 1)$  convolution for dimensionality reduction, a  $(3 \times 3)$

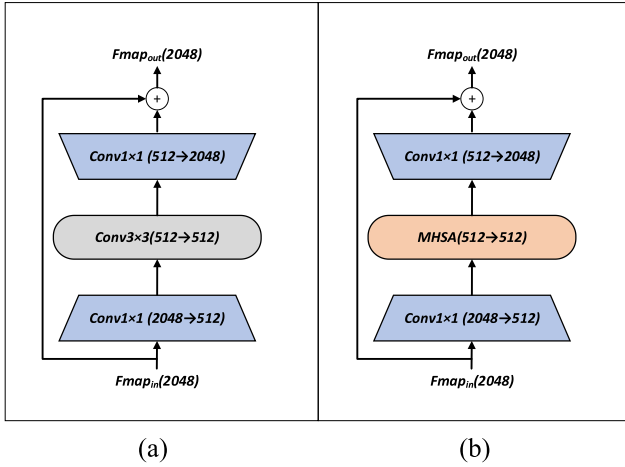


Fig. 3. Different bottleneck blocks. (a) ResNet bottleneck block. (b) BoT block.

convolution for spatial feature extraction, and a  $(1 \times 1)$  convolution for dimensionality expansion. The bottleneck structure can effectively reduce the computation of the convolution layer without reducing the performance of model, and has therefore been widely used in the design of neural networks and further refined in subsequent researches. In [51], the channels of each convolutional layer in the bottleneck block are further increased to improve feature extraction ability. In [52], the middle spatial layer of the classical bottleneck structure is replaced by grouped convolution to aggregate richer image features. Lately, as attention mechanisms have evolved and demonstrated good utility for improving various types of networks, their hybrid structures with the bottleneck block have been researched. In [53], the bottleneck block was extended by a channel attention branch, which aims to improve the representation of the network by modeling the relationship between channels of the feature map. In [54], a multiscale bottleneck structure is introduced, in which branches with different receptive field sizes are automatically assigned weights by the cross-channel attention module, effectively improving the adaptability of the model. Furthermore, the MHSA is a unique attention mechanism from the transformer [39], which has powerful global information modeling capability. Therefore, Srinivas et al. [55] migrated it in the bottleneck structure and named it the BoT.

BoT performs well in visual recognition tasks owing to the combination of the global information modeling capability of the self-attention mechanism and the lightweight attribute of bottleneck structure. As the BoT detailed in Fig. 3(b), the  $(3 \times 3)$  spatial convolution layer in the classical bottleneck structure is replaced by the MHSA from the transformer. Benefiting from its architecture, BoT processes and aggregates the information in the 2D feature map by exploiting the global self-attention from MHSA layer, thus better modeling the long-range correlation between pixels. In addition, it is not necessary for BoT to reshape 2D feature maps into 1D token sequences, which facilitates the maintenance of 2D spatial adjacency information of the input image for further exploitation.

The attention of the MHSA layer in BoT be defined by the following formula:

$$Attention(Q, K, V) = softmax(QP + QK^T)V \quad (3)$$

where  $Q$  is the query,  $K$  is the key, and  $V$  is the value.  $P$  represents the position encoding, and  $R_h$  and  $R_w$  represent the height and width of 2D feature map, respectively.

While MHSA can help networks focus on globally useful features from the spatial domain, the BoT is not designed for HSI data with a three-dimensional stereoscopic structure, which is challenging in balancing the analysis of spectral and spatial features. To further explore the long-distance dependence of spectral and spatial features, we design the dual-domain self-attention ( $D^2SA$ ) mechanism with a parallel structure for application scenarios of HSIs. This mechanism comprises a channel global attention branch and a spatial global attention branch. The former projects input features into three adaptable vectors, facilitating the complementary learning of spectral long-range dependencies by incorporating vectors from both the global channel and local space. The spatial attention branch, derived from the original MHSA, enhances location correlation through two-dimensional relative position encoding. The outputs of these branches are adaptively fused to prioritize discriminative features essential for classification. This dual attention structure enables the collaborative utilization of global features in both spectral and spatial dimensions, effectively capturing long-range interactions among HSI pixels.

### III. METHODS

Fig. 4 illustrates the overall framework for HSI classification, consisting of two components: a multilayer residual convolution module and  $D^2S^2BoT$ .

First, the multilayer residual convolution module extracts local information and includes two residual blocks that focus on spectral and spatial features. A shallow-deep feature interaction mechanism enhances the representation of HSI multilayer features.

Then, the  $D^2S^2BoT$  receives the local feature map to establish the global correlation of HSI pixels and predict the classification result. It includes stacked dual-dimension encoders ( $D^2Encoder$ ) and a linear projection classifier. The encoders analyze local feature maps and simulate the long-range correlation of HSI pixels using the introduced  $D^2SA$  structure. The linear projection classifier summarizes the output of the encoder to make predictions for classification result.

Algorithm 1 presents the pseudocode of this framework, while subsequent sections offer a comprehensive explanation of these components.

#### A. Multilayer Residual Convolution Module

To exploit the feature extraction capability of CNNs, the multilayer residual convolution module is first introduced to refine the local features of HSI, which provides rich local information for the subsequent  $D^2S^2BoT$ . As shown in Fig. 5, the



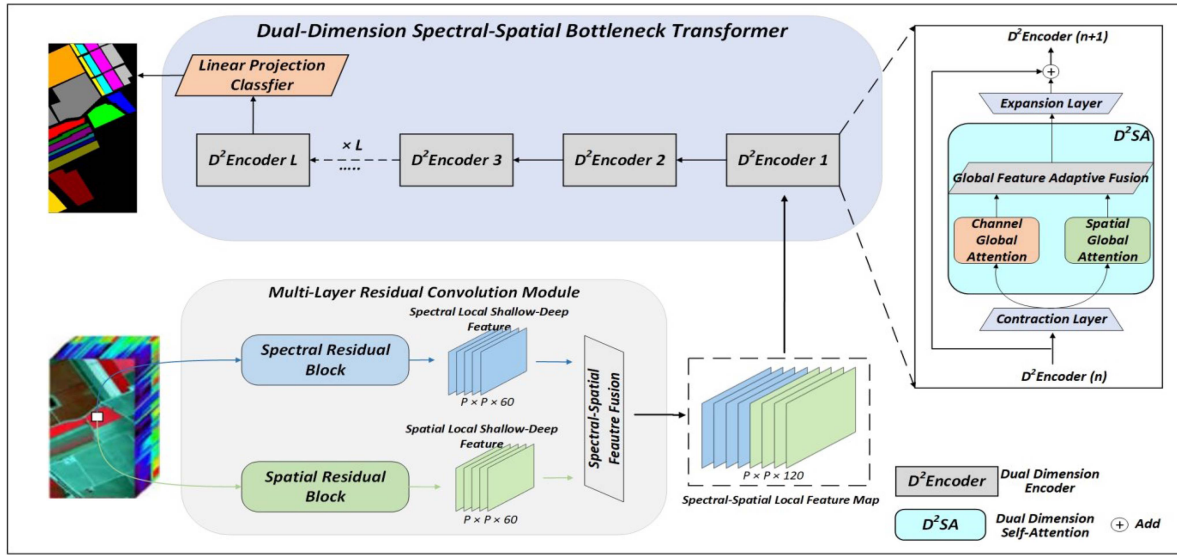


Fig. 4. Overall structure of proposed framework for HSI classification. It consists of two main components, i.e., multilayer residual convolution module for local feature extraction, and D<sup>2</sup>S<sup>2</sup>BoT for global correlation analysis and predict the classification results.

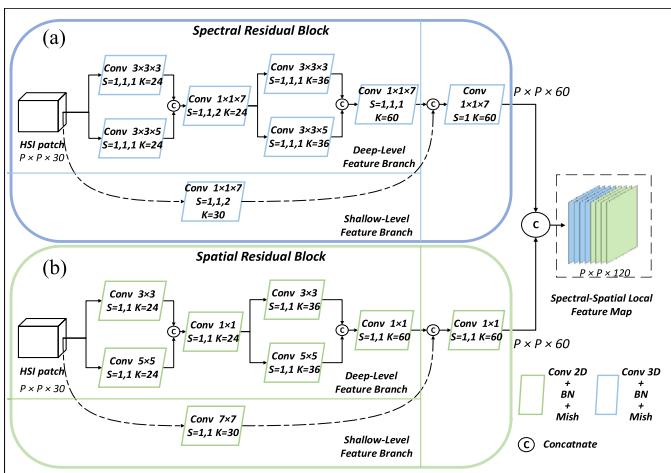


Fig. 5. Multilayer residual convolution module. Where  $S$  and  $K$  represent the strides and size of the convolution kernel, respectively. (a) Spectral Residual Block and (b) Spatial Residual Block.

HSI's spectral dimension is initially reduced by principal component analysis. Then, the dimensionality-reduced data patch is passed through the spatial and spectral residual blocks to extract the corresponding features, employing shallow-deep features interaction to capture more useful local information. Finally, the extracted features are fused by a channel concatenation operation to further obtain feature maps containing rich local spectral-spatial information.

1) *Spectral Residual Block*: The spectral residual block employs two independent branches, i.e., a deep-level feature and a shallow-level feature branch, as shown in Fig. 5(a).

In the deep-level feature branch, multiscale perception is combined to build convolutional layers, which are stacked twice to refine the multiscale features and further learn more abstract representations. In each layer, multiscale features are extracted through two sets of convolution kernels with receptive fields of  $(3 \times 3 \times 3)$  and  $(3 \times 3 \times 5)$ , respectively. The resulting

feature maps are then concatenated along the channel dimension, followed by a  $(1 \times 1 \times 7)$  convolutional kernel to further aggregate information. This can be represented by the following formula:

$$\begin{aligned} X_{spe(d)}^i = & Conv3d_{(1 \times 1 \times 7)} \\ & \times \left[ Concat \left( \begin{array}{l} Conv3d_{(3 \times 3 \times 3)} \left( X_{spe(d)}^{(i-1)} \right) \\ Conv3d_{(3 \times 3 \times 5)} \left( X_{spe(d)}^{(i-1)} \right) \end{array} \right) \right] \end{aligned} \quad (4)$$

where  $X_{spe(d)}^i$  denotes the output of the  $i$ th ( $i = 1, 2$ ) layer of spectral residual block;  $Concat(\cdot)$  denotes the concatenation operation, and  $Conv3d(\cdot)$  denotes convolution operation, which consists of 3D convolution operation, batch normalization layer, and mish activation function [56].

In contrast, the shallow-level feature branch employs a single-layer convolutional structure with a  $(1 \times 1 \times 7)$  kernel, which is well-suited for extracting coarse-grained features of the HSI. The resulting output of the shallow feature branch can be represented by the following formula:

$$X_{spe(s)} = Conv3d_{(1 \times 1 \times 7)}(X) \quad (5)$$

where  $X$  represents dimension-reduced HSI patch.

Next, we introduce the shallow-deep feature interaction operation, which aims to fuse the output of the two branches and enhance the residual block's ability to capture and refine useful information from the multilayer features. This operation consists of a concatenation operation and a convolutional layer with a size of  $(1 \times 1 \times 7)$ . The resulting output of the spectral residual block can be expressed as follows:

$$X_{spe} = Conv3d_{(1 \times 1 \times 1)} \left[ Concat \left( \begin{array}{l} X_{spe(s)} \\ X_{spe(d)} \end{array} \right) \right] \quad (6)$$

where  $X_{spe}$  represents the output from the spectral residual block.

2) *Spatial Residual Block*: To explore the spatial features of HS images, we construct the spatial residual block using 2D

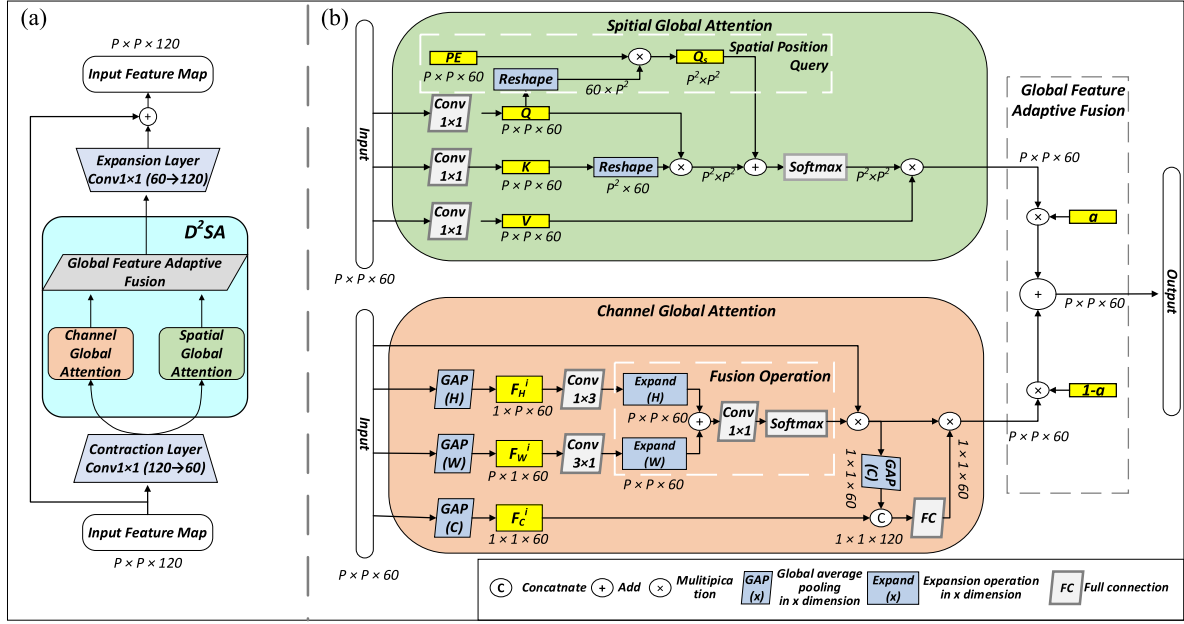


Fig. 6. (a) D<sup>2</sup>Encoder and (b) D<sup>2</sup>SA. The encoder consists of three layers: a contraction layer, a D<sup>2</sup>SA layer, and an expansion layer. The D<sup>2</sup>SA is implemented by two branches, i.e., the spatial global attention branch and the channel global attention branch. Where  $H$ ,  $W$ , and  $C$  represent the feature map's heights, widths, and channels, respectively.

convolution with a similar concept to the spectral residual block. The block is composed of two branches, extracting deep and shallow features independently, as depicted in Fig. 5(b). The shallow-depth feature interaction operation is then employed to learn and refine these features, involving a concatenation operation and a 2D convolution operation with  $(1 \times 1)$  kernel size. The output of the spatial residual block can be expressed as follows:

$$\mathbf{X}_{spa} = \text{Conv}2d_{(1 \times 1)} \left[ \text{Concat} \begin{pmatrix} \mathbf{X}_{spa(s)} \\ \mathbf{X}_{spa(d)}^2 \end{pmatrix} \right] \quad (7)$$

$$\mathbf{X}_{spa(d)}^i = \text{Conv}2d_{(1 \times 1)} \left[ \text{Concat} \begin{pmatrix} \text{Conv}2d_{(3 \times 3)}(\mathbf{X}_{spa(d)}^{(i-1)}) \\ \text{Conv}2d_{(5 \times 5)}(\mathbf{X}_{spa(d)}^{(i-1)}) \end{pmatrix} \right] \quad (8)$$

$$\mathbf{X}_{spa(s)} = \text{Conv}2d_{(7 \times 7)}(\mathbf{X}) \quad (9)$$

where  $\mathbf{X}_{spa}$  represents the output of the spatial residual block,  $\mathbf{X}_{spa(s)}$  and  $\mathbf{X}_{spa(d)}$  represent the outputs of shallow and deep feature branch, respectively.  $\text{Conv}2d(\cdot)$  denotes convolution operation, which consists of 2D convolution operation, batch normalization layer, and Mish activation function.

3) *Spectral and Spatial Feature Fusion*: Using spectral and spatial residual blocks, feature maps are obtained for both spectral and spatial dimensions, which are then fused using concatenation operations instead of element summation to prevent cross-interference of features from different dimensions. This process can be represented by the following formula:

$$\mathbf{X}_{out} = \text{Concat} \begin{pmatrix} \mathbf{X}_{spa} \\ \mathbf{X}_{spe} \end{pmatrix} \quad (10)$$

where  $\mathbf{X}_{out}$  represents the output of the multilayer residual convolution module, which serves as the input of the subsequent transformer.

## B. Dual-Dimension Spectral-Spatial BoT

Given the strong interdependence between the spectral and spatial dimensions of HSI data, we have developed a D<sup>2</sup>S<sup>2</sup>BoT that can efficiently learn local spectral-spatial features and analyze long-range dependencies to achieve accurate land cover classification. The D<sup>2</sup>S<sup>2</sup>BoT consists of a stacked set of D<sup>2</sup>Encoders and a linear projection classifier. The encoders are specifically designed to capture global correlations of HSI pixels across both spectral and spatial dimensions, with each layer of the stacked encoders learning spectral-spatial features at increasing depths. The deep HSI features learned by the encoders are then aggregated by the linear projection classifier to produce the final classification results.

1) *Dual-Dimension Encoder*: As shown in Fig. 6(a), the proposed D<sup>2</sup>Encoder consists of three layers: a contraction layer, a D<sup>2</sup>SA layer, and an expansion layer. The contraction and expansion layers utilize  $(1 \times 1)$  convolutions to reduce and increase the number of feature map channels, respectively. The D<sup>2</sup>SA layer is the core of modeling HSI global relationships, with the ability to analyze and learn from long-range dependencies of spatial and channel dimensions in feature maps. The D<sup>2</sup>SA layer is implemented through two separate branches, including the channel global-attention and the spatial global-attention.

As illustrated in Fig. 6(b), the channel global-attention branch utilizes averaging pooling operations to aggregate information from the input feature map  $\mathbf{F}^i (i = 1, 2, 3, \dots, L)$ . Specifically, these operations are applied along the height, width, and channel dimensions, resulting in the extraction of three distinct feature maps, denoted as  $\mathbf{F}_h^i$ ,  $\mathbf{F}_w^i$ , and  $\mathbf{F}_c^i$ , respectively. The first two are applied for modeling HSI neighborhood pixel interactions, and the last for capturing long-range channel correlations.

**Algorithm 1:** D<sup>2</sup>S<sup>2</sup>BoT Framework.

- Input:** Input an HS image data  $I \in R^{H \times W \times C}$  and ground-truth  $Y \in R^{H \times W}$ ; PCA band number  $B = 30$ ; HS patch size  $S = 11$ ; training ratio  $T\%$ .
- Output:** Predicted labels of the test HS dataset.
- 1: Set the batch size of the model to 64, the optimizer to Adam (learning rate  $1e^{-4}$ ), the epoch number  $E$  to 50, and the D<sup>2</sup>S<sup>2</sup>BoT encoder layers  $L$  to 3.
  - 2: Obtain HS image data  $I_{pca}$  after PCA transform and divide it into training and test sets. Generate training loader and test loader.
  - 3: **For**  $n = 1$  to  $E$  **do**
  - 4: Perform the multilayer residual convolution module to obtain a feature map  $X_{out}$ .
  - 5: Input  $X_{out}$  to the Transformer module's Layer 1 encoder.
  - 6: **For**  $m = 1$  to  $L$  **do**
  - 7: Perform the dual dimension encoder on the feature map of layer  $m$ .
  - 8: **End for**
  - 9: Perform linear projection classifier to identify labels
  - 10: **End for**
  - 11: Use the trained model to make predictions on the test set.

First, the  $F_h^i$  and  $F_w^i$  are fed to the 2D convolution layers of size  $(3 \times 1)$  and  $(1 \times 3)$ , respectively, to obtain  $\hat{F}_h^i$  and  $\hat{F}_w^i$ , which are fused and multiplied with the input  $F^i$  to obtain the attention map  $\hat{F}_{channel}^i$ . These processes can be expressed by the following formulas:

$$\hat{F}_{channel}^i = F^i * f(\hat{F}_h^i, \hat{F}_w^i) \quad (11)$$

$$\hat{F}_h^i = Conv2d_{(1 \times 3)}(F_h^i) \quad (12)$$

$$\hat{F}_w^i = Conv2d_{(3 \times 1)}(F_w^i) \quad (13)$$

where  $f(\cdot)$  represents fusion operation, which consists of expansion operations, element-wise addition, a convolution of size  $(1 \times 1)$ , and a sigmoid activation function.

Subsequently, the attention map  $\hat{F}_{channel}^i$  is processed along the channel dimension to perform the global average pooling operation, which is concatenated with  $F_c^i$ , followed by a fully connected layer to learn each channel correlation and obtain the channel attention vector  $V_c^i$ . Finally, the channel attention branch output  $F_{channel}^i$  is derived by multiplying the channel attention vector  $V_c^i$  with the attention map  $\hat{F}_{channel}^i$ , as described by the following formulas:

$$F_{channel}^i = V_c^i * \hat{F}_{channel}^i \quad (14)$$

$$V_c^i = FC \left[ Concat \left( V_c^i, GP_c \left( \hat{F}_{channel}^i \right) \right) \right] \quad (15)$$

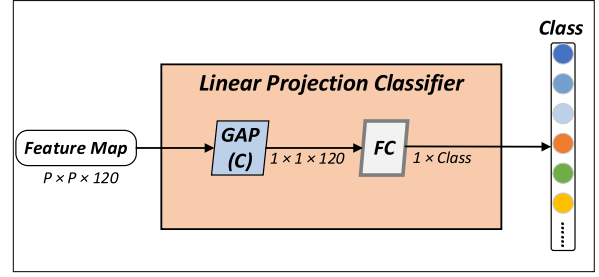


Fig. 7. Linear projection classifier in D<sup>2</sup>S<sup>2</sup>BoT.

where  $F_{channel}^i$  represents the channel global-attention branch output of layer  $I$  encoder.  $Concat(\cdot)$  represents the concatenation operation.  $GP_c(\cdot)$  represents the global average pooling operation performed by the channel dimension.

For the spatial global-attention branch, the self-attention mechanism is used to extract the global correlation in the spatial dimension. As shown in Fig. 6(b), 2D convolution operations are performed on the input feature map to create three trainable matrices:  $Q$ ,  $K$ , and  $V$ , which are used to analyze the global context information of spatial dimensions. The process can be expressed as follows:

$$Q, K, V = Conv2d_{(1 \times 1)}(F^i) \quad (16)$$

where  $Q$ ,  $K$ , and  $V$  are trainable matrices of query, key, and value, respectively.

To better align the feature information in the spatial dimension of the feature map with location information, we introduce a spatial position query matrix  $Q_s$ . This is realized by implementing 2D relative location self-attention [57]. These four matrices are then used to analyze the global context information of spatial dimensions. This process can be expressed as follows:

$$F_{spatial}^i = Attention(Q, K, V) \\ = softmax(Q_s + QK^T)V \quad (17)$$

$$Q_s = Q * (P_h + P_w) \quad (18)$$

where  $F_{spatial}^i$  represents the spatial global-attention branch output of layer  $I$  encoder.  $Q_s$  is the spatial location query matrix, which contains relative height and width information ( $P_h$  and  $P_w$ , respectively) on the spatial dimension of the feature map.

Then, the outputs of the two branches are fused, where a learnable parameter  $\gamma$  is introduced to adaptively adjust the weights of each branch. The  $I$ th layer encoder output can be expressed as follows:

$$F^{i+1} = Encode(F^i) = a * F_{spatial}^i \\ + (1 - a) * F_{channel}^i \quad (19)$$

where  $Encode(\cdot)$  represents D<sup>2</sup>S<sup>2</sup>BoT encoding operation, and  $a \in (0, 1)$  is a learnable parameter.

2) *Linear Projection Classifier*: The linear projection classifier layer receives the global features learned by encoders and further completes the classification. As shown in Fig. 7, the classifier is constructed by global average pool operation and

TABLE I  
NUMBER OF SAMPLES IN THREE DATASETS

Order number	Datasets		
	IP	SV	BS
1	46	1969	270
2	1428	3652	101
3	830	1938	251
4	237	1368	215
5	483	2626	269
6	730	3881	269
7	28	3509	259
8	478	11047	203
9	20	6079	314
10	972	3214	248
11	2455	1048	305
12	593	1889	181
13	205	898	268
14	1265	1050	98
15	386	7124	
16	93	1771	
<b>Total</b>	10249	53063	3248

full connection layer instead of the MLP layer in the original transformer. The classification process can be expressed as follows:

$$Y = FC(GP(F^L)) \quad (20)$$

where  $Y$  represents the classification result predicted by the  $D^2S^2BoT$ , and  $F^L$  represents the last ( $L$ th) encoder output.

#### IV. EXPERIMENT

We conduct various experiments on three HSI datasets to validate the proposed  $D^2S^2BoT$  framework. First, we provide an overview of these datasets. Then, we elaborate on the experimental setup, including the implementation details, and conduct several ablation experiments to demonstrate the validity of the different modules. Finally, we compare the classification results of the  $D^2S^2BoT$  framework with some advanced algorithms to demonstrate its superiority.

##### A. Overview of Datasets

This study involved conducting experiments on three hyperspectral image datasets, namely the Indian Pine (IP), Salinas Valley (SV), and Botswana (BS) datasets. Table I shows the detailed information of each dataset.

**IP:** The IP dataset, acquired by the AVIRIS spectrometer in northwest Indiana, encompasses 200 spectral bands spanning wavelengths ranging from 0.4 to 2.5  $\mu\text{m}$ . This dataset comprises a total of  $145 \times 145$  pixels with a resolution of 20 m/pixel and encompasses a diverse set of 16 classes. Fig. 8(a) visually presents the corresponding ground-truth map.

**SV:** The SV dataset was acquired using AVIRIS sensors and is situated in California, USA. It encompasses 204 spectral bands spanning wavelengths from 0.4 to 2.5  $\mu\text{m}$ . This dataset comprises a spatial resolution of 3.7 m with dimensions of  $512 \times 217$  pixels, encompassing a total of 16 land cover classes. The ground truth for the SV dataset is depicted in Fig. 9(a).

**BS:** The BS dataset was acquired by the EO-1 HS sensor in Okavango Delta, California, USA in 2001. Following the removal of noisy bands, a total of 144 spectral bands were selected for the experiment, encompassing wavelengths ranging from 0.38 to 1.05  $\mu\text{m}$ . The dataset comprises  $1476 \times 256$  pixels with a spatial resolution of 30 m and encompasses 14 distinct classes. The corresponding ground-truth map can be observed in Fig. 10(a).

##### B. Experimental Setting

This article compares several well-known CNN-based and Transformer-based models, including HybridSN [36], SSRN [37],  $A^2S^2K$  [58], DBMA [59], DBDA [38], BS2T [60], Morphformer [61], and CTmixer [62]. All models are executed on a server with an i5-3470 CPU and a K40c 12 GB GPU for fairness, with an input patch size of  $(11 \times 11)$ . Each experiment is trained for 50 epochs.

To quantify the experimental results, three evaluation indexes are introduced, namely overall accuracy (OA), average accuracy (AA), and kappa coefficient (K). The HS image dataset is split into three subsets (training, validation, and test) with fewer training samples to better distinguish performance differences among algorithms. For the IP dataset, 2.5% of the samples are randomly assigned to the training set and another 2.5% to the validation set. For the SV and BS datasets, 0.5% and 1.5% of the samples are used for training, respectively, while another 0.5% and 1.5% are used for validation. The remaining samples are reserved for testing.

##### C. Classification Results

**1) Classification Results for the IP Dataset:** The low spatial resolution (20 m) and mixed pixel phenomenon of the IP dataset pose challenges in distinguishing land covers, thus we adopted a training ratio of 2.5%. Our proposed framework outperformed other algorithms, achieving superior classification results with an OA of 95.96%, AA of 94.58%, and a kappa coefficient of 0.9539, as elaborated in Table II. Comparative analyses with CNN-based networks such as HybridSN and SSRN reveal that our proposed method exhibited superior classification accuracy across most classes, with notable improvements, particularly for class 12 and class 16, surpassing the performance of the HybridSN network by 35.98% and 17.36%, respectively. Moreover, compared to the SSRN network, our method improved the classification accuracy for these classes by 16.88% and 19.94%, respectively. Additionally, in contrast to attention-based methods like  $A^2S^2K$ , DBMA, and DBDA, our proposed method excelled in analyzing the global relationships of HS pixels, resulting in significant accuracy enhancements, particularly for classes with limited samples. For example, in class 1, our method's accuracy



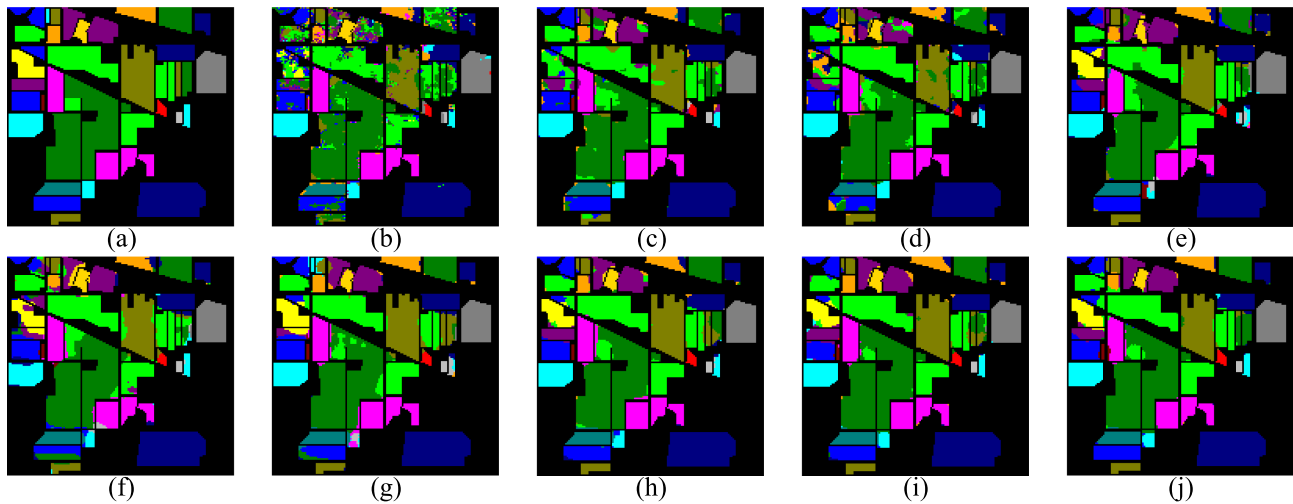


Fig. 8. Map of classification results for the IPs dataset. (a) Ground-truth map. (b) HybridSN. (c) SSRN. (d) A<sup>2</sup>S<sup>2</sup>K. (e) DBMA. (f) DBDA. (g) BS2T. (h) MorphFormer. (i) CTmixer. (j) PROPOSED WORK.

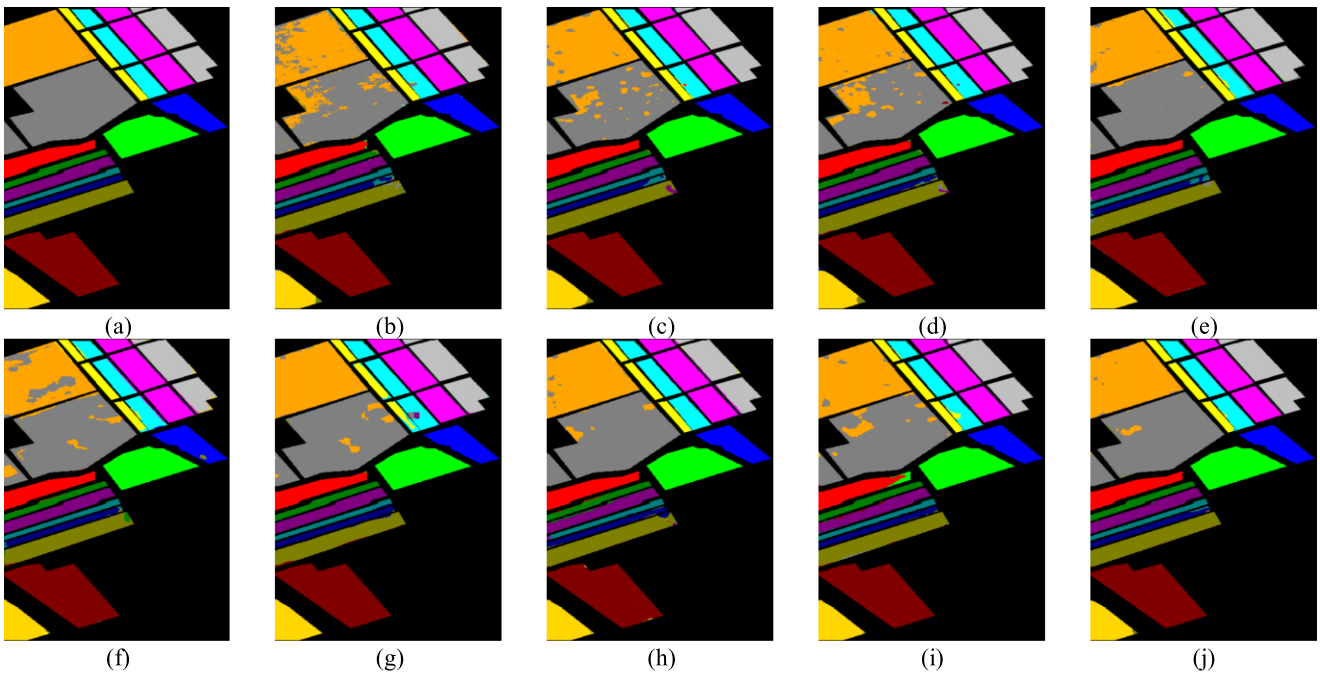


Fig. 9. Classification maps of the SV dataset. (a) Ground-truth map. (b) HybridSN. (c) SSRN. (d) A<sup>2</sup>S<sup>2</sup>K. (e) DBMA. (f) DBDA. (g) BS2T. (h) MorphFormer. (i) CTmixer. (j) PROPOSED WORK.

surpasses the mentioned methods by 54.15%, 3.64%, and 6.32%, respectively. By leveraging the interdependence of HSI features in both dimensions, our method outperformed Transformer-based approaches such as BS2TMorphformer and CTmixer in terms of OA metrics, achieving a 1.09% improvement over the suboptimal Morphformer, which achieved an OA of 94.87%.

2) *Classification Results for the SV Dataset:* Table III presents the three metrics for all evaluated methods on the SV dataset, with the best-performing results highlighted in bold. The SV dataset ground truth and the classification maps predicted by different methods are displayed in Fig. 9.

For the SV dataset, training is executed using only 0.5% of the samples, as the majority of land classes are continuous and linearly separable. The framework exhibits a prominent classification effect, with outstanding OA, AA, and kappa coefficients reaching 98.46%, 98.29%, and 99.12%, respectively, as detailed in Table III. Notably, the introduced method attains more than 90% accuracy for each ground class, with the highest accuracy observed for classes 5, 8, 9, 11, and 14 compared to other methods. It is noteworthy that, while the Morphformer and CTmixer methods also achieve commendable classification results (OA = 97.81% and 97.71%, respectively), the proposed method generates fewer noisy points and

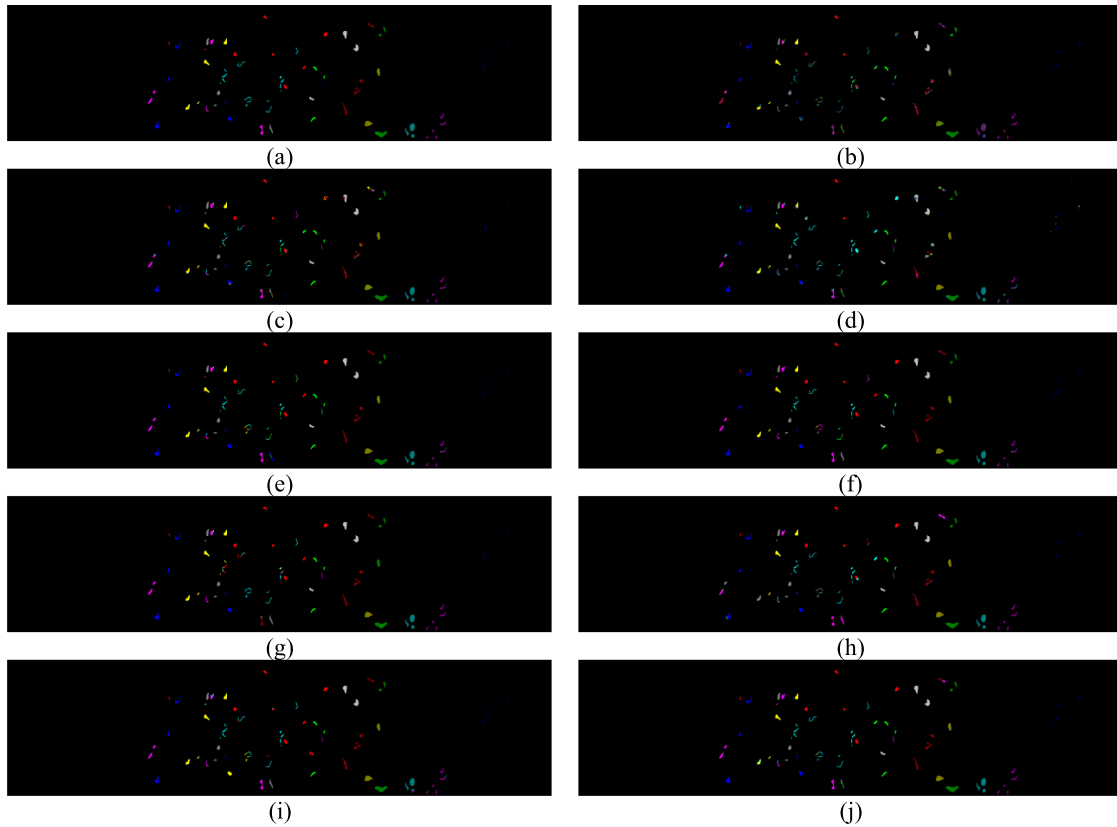


Fig. 10. Map of classification results for the BS dataset. (a) Ground-truth map. (b) HybridSN. (c) SSRN. (d) A<sup>2</sup>S<sup>2</sup>K. (e) DBMA. (f) DBDA. (g) BS2T. (h) MorphFormer. (i) CTMIXER. (j) PROPOSED WORK.

exhibits smoother boundary regions during the classification process.

3) *Classification Results for the BS Dataset*: Table IV presents the three metrics for all methods evaluated on the BS dataset, with the top-performing results highlighted in bold. The ground truth maps for the BS dataset and the prediction result maps generated by the different algorithms are displayed in Fig. 10.

The BS dataset contains only 3248 labeled samples, with 1.5% allocated for training. The proposed framework achieves outstanding classification results, with an OA of 96.69%, AA of 96.29%, and kappa of 0.9461. The introduced method also attains over 80% accuracy for all land-cover classes, including seven classes with the highest accuracy among all methods. Moreover, the introduced method achieves 100% accuracy for classes 7, 10, 12, 13, and 14, leveraging local and global contexts to develop HSI patch features. These results demonstrate the proposed method's capability in capturing deep global relationships between different objects, achieving high-precision classification on the BS dataset, which contains a substantial number of discrete and local samples.

#### D. Parameter Studies

In this section, we analyze the influence of two key network parameters, i.e., the size of the input patch and the number of encoder layers of D<sup>2</sup>S<sup>2</sup>BoT. Two sets of experiments are then

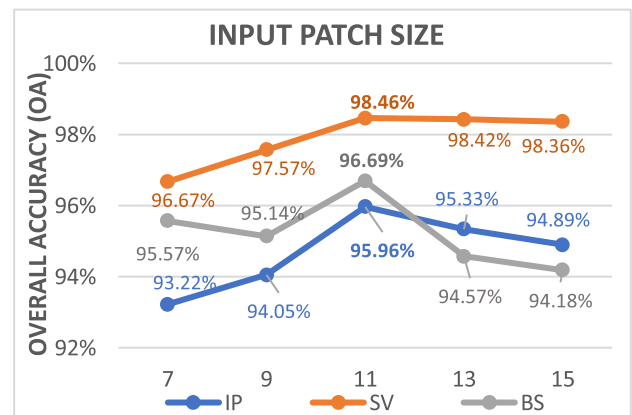


Fig. 11. Effect of different input patch size on the OA.

performed on three datasets to compare the complexity of all compared methods and their performance at different training ratios.

1) *Input Spatial Size*: The OA of the proposed framework with different input sizes is illustrated in Fig. 11. As the input size increases, the model can effectively incorporate more spatial-spectral neighborhood information. Optimal classification results are achieved when the input spatial size expands to (11 × 11). However, further increasing the spatial size counter-intuitively leads to a decrease in classification accuracy due to

TABLE II  
RESULTS FOR IP DATASET OBTAINED BY DIFFERENT CLASSIFICATION METHODS

Class No.	HYBRIDSN	SSRN	A <sup>2</sup> S <sup>2</sup> K	DBMA	DBDA	BS2T	MORORPH FORMER	CTMIXER	PROPOSED METHOD
1	78.28	95.28	55.15	96.52	93.68	<b>100</b>	99.04	<b>100</b>	<b>100</b>
2	65.60	83.22	83.70	80.69	76.00	91.86	93.83	89.61	<b>95.75</b>
3	72.34	99.42	88.41	94.37	93.44	92.77	95.01	95.90	<b>97.11</b>
4	76.22	89.51	73.19	84.54	79.75	82.73	95.87	95.68	<b>96.51</b>
5	88.89	59.74	97.96	96.08	97.85	92.15	99.63	<b>99.66</b>	99.31
6	91.05	99.06	93.04	95.17	<b>97.58</b>	87.50	95.68	93.22	95.96
7	81.82	59.74	91.61	21.47	49.30	30.86	<b>86.51</b>	81.18	86.21
8	92.20	99.06	86.87	96.57	99.15	98.42	<b>99.42</b>	99.57	96.89
9	68.37	67.50	47.00	41.35	49.77	<b>88.63</b>	81.48	97.78	77.27
10	64.77	86.25	82.60	89.83	86.55	85.32	93.42	92.72	<b>95.20</b>
11	72.35	89.39	90.37	90.76	86.57	95.65	93.97	<b>98.38</b>	96.68
12	56.64	75.74	86.46	78.99	79.87	<b>97.08</b>	93.26	96.32	92.62
13	95.81	98.65	92.01	95.04	97.26	89.92	<b>98.02</b>	89.09	92.49
14	84.12	90.70	93.95	91.45	<b>97.08</b>	95.85	95.87	93.64	94.84
15	82.72	69.29	85.80	92.54	91.34	91.51	95.09	92.59	<b>99.01</b>
16	80.08	77.50	81.32	87.12	87.56	<b>98.73</b>	91.48	87.12	97.44
K × 100	84.81	82.88	85.95	86.86	85.81	90.78	94.14	93.93	<b>95.39</b>
OA(%)	75.15	85.08	87.68	88.48	87.58	91.90	94.87	94.67	<b>95.96</b>
AA(%)	78.20	80.74	83.09	83.28	85.17	89.77	94.42	93.90	<b>94.58</b>

Best results in bold.

TABLE III  
RESULTS FOR SV DATASET OBTAINED BY DIFFERENT CLASSIFICATION METHODS

Class no.	HYBRIDSN	SSRN	A <sup>2</sup> S <sup>2</sup> K	DBMA	DBDA	BS2T	MORPH FORMER	CTMIXER	PROPOSED METHOD
1	99.89	99.60	99.94	99.98	98.81	<b>100</b>	99.19	<b>100</b>	<b>100</b>
2	99.23	<b>100</b>	99.27	99.02	99.97	99.94	99.95	98.91	99.95
3	99.34	<b>100</b>	98.56	99.71	<b>100</b>	99.92	99.71	99.74	<b>100</b>
4	98.99	96.71	97.96	90.36	96.50	<b>99.01</b>	94.41	98.69	98.09
5	98.45	99.80	99.63	99.49	98.60	94.47	99.78	97.05	<b>99.92</b>
6	98.23	<b>100</b>	99.73	99.17	99.89	99.70	98.33	99.99	99.41
7	99.80	99.88	97.98	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	99.74	<b>100</b>
8	93.21	96.65	96.56	90.63	97.17	97.56	98.63	98.35	<b>99.94</b>
9	99.25	99.88	99.09	99.75	99.83	98.72	99.87	98.88	<b>99.91</b>
10	95.31	97.83	96.93	99.19	97.11	<b>99.22</b>	98.45	99.16	99.21
11	99.02	100	99.98	99.34	94.86	99.43	92.47	98.98	<b>100</b>
12	97.70	95.08	98.43	99.43	<b>100</b>	98.39	97.45	99.59	<b>100</b>
13	90.73	83.74	73.37	90.56	83.90	<b>99.88</b>	93.37	91.28	99.24
14	97.73	99.12	89.24	98.81	98.82	95.45	91.24	94.52	<b>99.43</b>
15	77.77	86.72	82.18	94.47	95.48	89.66	<b>93.62</b>	92.01	90.87
16	98.86	100	99.87	99.15	99.55	<b>100</b>	99.41	<b>100</b>	<b>100</b>
K × 100	93.46	96.01	94.82	95.98	96.28	96.98	97.57	97.45	<b>98.29</b>
OA(%)	94.12	97.07	95.34	96.39	96.66	97.29	97.81	97.71	<b>98.46</b>
AA(%)	96.47	95.54	95.54	97.44	97.16	98.21	97.24	97.93	<b>99.12</b>

Best results in bold.

unnecessary information introduced by excessive input space, resulting in over-fitting of the model.

2) *Number of Encoder Layers of D<sup>2</sup>S<sup>2</sup>BoT*: Fig. 12 shows the proposed model's classification results with varying numbers of D<sup>2</sup>S<sup>2</sup>BoT encoder layers. Classification accuracy initially increases with the number of layers but reaches its peak at three layers, with further increases causing slight fluctuations. Stacking encoders enhances the HSI feature representation, but excessive encoders weaken the correlation between layers, leading to unstable performance. To balance model complexity and performance, subsequent experiments set the number of encoder layers to 3.

3) *Parameters Numbers and Running Time*: Table V summarizes the time consumption and number of parameters for

nine methods on IP, SV, and BS datasets. The outcomes indicate that the method achieved the highest classification accuracy with moderate parameters and time consumption, demonstrating its effectiveness in accomplishing the HSI classification task.

4) *Training Ratios Experiment*: The experimental results obtained by employing various comparison methods on three datasets with different training sample ratios are presented in Table VI. Specifically, for the IP dataset, training ratios of 1%, 2%, 5%, 8%, and 10% were employed. For the BS dataset, training ratios of 1%, 2%, 5%, 7%, and 8% were employed. Considering the abundance of samples in the SV dataset, smaller training ratios are used, namely 0.1%, 0.2%, 0.5%, 0.8%, and 1%.

The results in Table VI show that there is a positive relationship between training rate and classification accuracy. It

TABLE IV  
RESULTS FOR BS DATASET OBTAINED BY DIFFERENT CLASSIFICATION METHODS

Class no.	HYBRIDSN	SSRN	A <sup>2</sup> S <sup>2</sup> K	DBMA	DBDA	BS2T	MORPH FORMER	CTMIXER	PROPOSED METHOD
1	72.67	71.02	85.49	98.01	98.40	89.33	90.79	88.67	<b>98.82</b>
2	57.33	71.42	87.27	69.89	73.17	77.82	91.84	<b>97.90</b>	80.33
3	76.18	92.78	90.46	96.94	96.66	<b>100</b>	99.76	96.77	96.50
4	86.10	75.28	76.29	88.66	94.78	94.04	<b>98.41</b>	80.30	97.70
5	70.72	81.69	39.01	88.25	76.26	94.05	<b>98.08</b>	90.67	91.33
6	57.46	77.27	87.48	94.87	<b>97.04</b>	90.84	89.46	88.53	96.89
7	72.44	91.83	97.51	<b>100</b>	95.76	<b>100</b>	99.37	88.87	<b>100</b>
8	75.63	88.76	74.09	95.23	96.42	<b>99.80</b>	91.38	87.89	93.78
9	52.97	72.10	89.30	97.03	<b>99.57</b>	79.14	97.71	96.09	98.71
10	57.13	71.18	90.96	94.12	96.55	<b>100</b>	98.60	98.49	<b>100</b>
11	62.32	93.59	88.64	92.01	96.23	79.46	92.13	<b>97.94</b>	94.08
12	47.53	87.45	61.63	<b>100</b>	<b>100</b>	97.44	<b>100</b>	97.18	<b>100</b>
13	60.40	88.06	56.86	96.39	91.09	99.38	87.50	90.73	<b>100</b>
14	60.15	96.66	97.14	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	99.74	<b>100</b>
K × 100	59.09	79.87	68.95	93.40	92.01	90.61	94.36	91.41	<b>96.41</b>
OA(%)	62.10	81.40	71.32	93.91	92.62	91.34	94.80	92.08	<b>96.69</b>
AA(%)	64.93	82.79	80.15	93.67	93.71	92.95	95.35	92.84	<b>96.29</b>

Best results in bold.

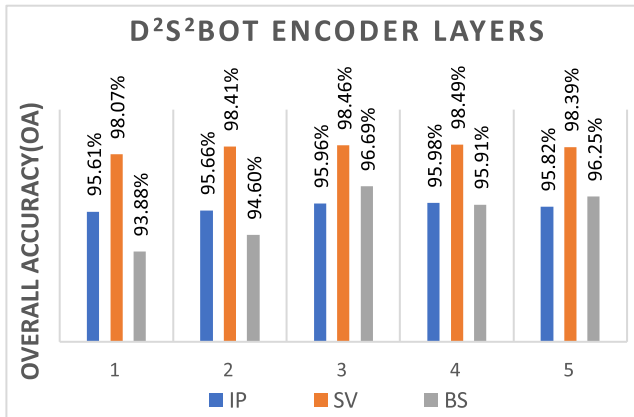


Fig. 12. Effect of different D<sup>2</sup>S<sup>2</sup>BoT encoders layer (number of encoders) on the OA.

is worth noting that all models show excellent performance when provided with sufficient training samples (10% for IP dataset, 8% for BS dataset, and 1% for SV dataset). However, the introduced method outperforms other models in all training ratios, especially when training samples are extremely scarce. For example, for the IP and SV datasets, the OA of the proposed method reaches 86.28% and 94.19% at 1% training rate, which are 1.64% and 1.46% higher than the suboptimal CTmixer method, respectively. For the BS dataset, the OA of the proposed method reaches 92.69% with the training rate of 0.1%. Moreover, the introduced framework can achieve similar or even better classification results with a smaller training rate compared to other methods. For example, the introduced framework achieves a 98.23% OA metric on the IP dataset with a training rate of 5%, which is significantly higher than the HybirdSN (98.01%), SSRN (97.54%), A<sup>2</sup>S<sup>2</sup>K (98.11%), and DBMA (97.87%) methods with a training rate of 10%. These results demonstrate the excellent potential of our proposed method in terms of labor and cost savings, given the high cost of labeled datasets.

TABLE V  
COMPARISON OF RUNNING TIME AND PARAMETERS OF DIFFERENT METHOD

DATASETS	METHODS	TRAINING		FLOPS (G)
		TIME /EPOCH(s)	PARAMETERS	
IP	HybirdSN	0.31	534,896	0.05
	SSRN	18.73	364,204	0.23
	A <sup>2</sup> S <sup>2</sup> K	8.93	162,181	0.13
	DBMA	2.77	609,791	0.36
	DBDA	3.64	606,906	0.36
	BS2T	3.89	383,356	0.16
	MORPHFORMER	1.01	132,192	0.05
	CTMIXER	1.07	663,071	0.13
	PROPOSED	2.21	375,574	0.19
SV	HybirdSN	0.31	534,896	0.05
	SSRN	18.78	370,348	0.24
	A <sup>2</sup> S <sup>2</sup> K	8.98	165,065	0.13
	DBMA	2.99	621,407	0.37
	DBDA	3.81	618,522	0.37
	BS2T	8.03	390,625	0.16
	MORPHFORMER	0.95	132,192	0.05
	CTMIXER	1.02	663,071	0.13
	PROPOSED	2.09	375,574	0.19
BS	HybirdSN	0.10	534,896	0.05
	SSRN	2.41	278,138	0.17
	A <sup>2</sup> S <sup>2</sup> K	1.19	121,781	0.09
	DBMA	0.53	446,925	0.26
	DBDA	0.59	444,040	0.26
	BS2T	0.68	280,970	0.11
	MORPHFORMER	0.31	132,192	0.05
	CTMIXER	0.34	663,071	0.13
	PROPOSED	0.63	375,574	0.19

### E. Ablation Studies

Ablation experiments are conducted on three datasets to validate the effectiveness of two key components in our proposed method: the multilayer residual convolution module and the D<sup>2</sup>SA. First, we evaluate the performance impact of



TABLE VI  
EXPERIMENTAL RESULTS OF TRAINING RATIOS ON THREE DATASETS UTILIZING DIFFERENT METHODS

DATASET	TRAINING RATIO	METHODS(OA%)								
		HYBRIDSN	SSRN	A <sup>2</sup> S <sup>2</sup> K	DBMA	DBDA	BS2T	MORPH FORMER	CTMIXER	PROPOSED METHOD
IP	1%	67.19	72.85	75.17	83.19	80.72	79.03	81.16	84.64	86.28
	2%	78.57	83.85	86.88	88.19	88.33	90.52	92.60	91.93	94.72
	5%	94.45	95.63	95.49	94.53	95.74	92.76	97.33	97.05	98.23
	8%	97.62	97.05	97.59	97.47	97.92	95.85	98.51	98.47	98.92
	10%	98.01	97.54	98.11	97.87	98.57	98.33	99.08	99.04	99.22
BS	1%	60.89	78.66	67.05	87.19	90.94	91.03	89.09	92.73	94.19
	2%	81.97	83.95	72.52	94.59	94.33	95.52	95.92	94.21	96.90
	5%	94.56	97.83	87.68	97.96	97.50	92.76	98.45	98.31	99.19
	7%	98.19	98.05	98.19	98.47	98.74	99.19	99.20	99.01	99.40
	8%	98.45	98.54	98.41	98.87	99.06	99.33	99.61	99.47	100.00
SV	0.1%	85.04	85.71	81.68	86.27	92.12	90.36	92.47	92.16	92.69
	0.2%	92.67	93.71	85.04	93.14	95.00	93.69	97.12	96.09	96.61
	0.5%	94.12	97.07	95.34	96.39	96.66	97.29	97.81	97.71	98.46
	0.8%	98.98	98.16	98.79	98.69	98.48	99.31	98.78	99.27	99.47
	1%	99.07	98.99	99.12	99.29	99.47	99.48	99.41	99.76	99.81

TABLE VII  
ACCURACY ANALYSIS FOR DIFFERENT COMPONENTS COMBINATIONS IN THE PROPOSED METHOD

Multi-layer Residual Convolution Module		Transformer block	IP			SV			BS		
Spectral Residual Block	Spatial Residual Block	BoT/D <sup>2</sup> S <sup>2</sup> BoT	K × 100	OA(%)	AA(%)	K × 100	OA(%)	AA(%)	K × 100	OA(%)	AA(%)
√	×	BoT	90.01	91.23	88.83	96.44	96.80	97.17	88.35	89.26	90.25
×	√		91.93	92.92	90.66	97.03	97.33	97.74	90.35	91.09	93.16
√	√		92.80	93.70	93.46	97.87	98.09	98.26	92.72	93.29	94.42
√	×	D <sup>2</sup> S <sup>2</sup> BoT	92.41	93.35	91.55	97.64	97.88	97.78	92.18	92.78	93.64
×	√		93.24	94.07	90.31	97.88	98.10	97.27	94.06	94.53	95.70
√	√		94.58	95.39	95.96	99.12	98.29	98.46	96.29	96.41	96.69

TABLE VIII  
ACCURACY ANALYSIS FOR DIFFERENT ATTENTION MECHANISM

Attention mechanism	IP			SV			BS		
SA/D <sup>2</sup> SA	K × 100	OA(%)	AA(%)	K × 100	OA(%)	AA(%)	K × 100	OA(%)	AA(%)
SA	93.01	93.87	93.13	97.65	97.89	98.11	93.08	93.61	94.92
D <sup>2</sup> SA-S	93.71	94.50	92.99	98.57	98.50	98.10	94.69	95.10	94.69
D <sup>2</sup> SA-C	93.96	94.72	93.43	97.89	98.10	98.43	93.24	93.76	94.07
D <sup>2</sup> SA	94.58	95.39	95.96	99.12	98.29	98.46	96.29	96.41	96.69

the two residual blocks (spectral and spatial) in the multi-layer residual convolution module on the two Transformer modules through two sets of comprehensive experiments. Additionally, we compare the traditional self-attention mechanism with the introduced D<sup>2</sup>SA mechanism. We evaluate the performance of each branch of the D<sup>2</sup>SA mechanism based on classification accuracy, and detailed results are presented in Tables VII and VIII.

To validate the effectiveness of the multilayer residual convolution module, we undertake an exploration of three combinations of residual blocks to gauge their impact on the performance of two transformer modules—the original BoT module and the introduced D<sup>2</sup>S<sup>2</sup>BoT module. As illustrated in Table VII, the framework’s classification accuracy reaches its lowest point when only the spectral residual block is activated. This is attributed to the inherent spectral specificity in HSI data, where

the same land class may manifest varying spectral features, or different land classes may exhibit analogous spectral features [4]. Therefore, the coactivation of the two residual blocks is necessary (lines three and six), allowing the spectral and spatial features to be jointly extracted and enabling the subsequent Transformer module to capture discriminative features from different land covers, thus improving overall performance.

On the other hand, Table VI also reveals that the introduced D<sup>2</sup>S<sup>2</sup>BoT module consistently outperforms the original BoT module in all three residual block combinations, especially when only one spectral residual block is activated (rows one and four). The OA of the three datasets is improved by 2.12%, 1.29%, and 3.62%, respectively, which is attributed to the ability of D<sup>2</sup>S<sup>2</sup>BoT to extract features from two dimensions, adaptively encode global features, and efficiently fuse and summarize them for exploring deep interactions.

Furthermore, we conduct a comprehensive comparative analysis to validate the efficacy of our introduced D<sup>2</sup>SA mechanism in comparison with the original self-attention mechanism, as shown in Table VIII. Specifically, we examine the independent activation of the spatial global attention branch (D<sup>2</sup>SA-S) and channel global attention branch (D<sup>2</sup>SA-C) within D<sup>2</sup>SA. The results demonstrate that the introduced D<sup>2</sup>SA better attends to the global features of HSI compared to the traditional self-attention mechanism. Specifically, the use of relative spatial location coding in D<sup>2</sup>SA-S significantly enhances the model's ability to process location information of local spatial features. In contrast, D<sup>2</sup>SA-C focuses on learning and summarizing global dependencies from the channel dimension of the HSI feature map, allowing for better mining of deeper features. The synergistic activation of D<sup>2</sup>SA-S and D<sup>2</sup>SA-C unleashes dynamic adaptations that enable D<sup>2</sup>SA to learn the interdependencies of features between the two domains. This adaptation leads to a more efficient analysis of joint features in spectral space, significantly improving classification performance with average accuracies of 95.32%, 98.29%, and 96.41% on the three datasets, respectively.

## V. CONCLUSION

In order to effectively extract the spectral-spatial correlations of 3D stereo-structured HSI data, we developed a D<sup>2</sup>SA mechanism, which formed the basis of our high-performance D<sup>2</sup>S<sup>2</sup>BoT classification framework. To explore the local-global features of HSI, we first introduced a multilayer residual convolution module for extracting local features. This module utilized two parallel residual blocks to extract spatial and spectral features, respectively. We then introduced a D<sup>2</sup>S<sup>2</sup>BoT to receive local features. D<sup>2</sup>S<sup>2</sup>BoT included a critical D<sup>2</sup>SA mechanism, which was devised to effectively capture the spectral-spatial correlation of the HSI data by modeling the long-term dependence of spatial and spatial-dimension features through two independent global attention branches. Finally, we introduced a linear projection classifier to summarize the features learned by the D<sup>2</sup>SAs and predict the classification results.

The experimental results on three HSI datasets demonstrate that the D<sup>2</sup>S<sup>2</sup>BoT framework significantly improves the performance of the transformer model. Specifically, the D<sup>2</sup>SA mechanism designed for HSI classification effectively explores both global spectral and spatial features, thereby capturing the spectral-spatial correlation accurately. In our future research endeavors, we aim to optimize the multimodal feature-awareness capability of D<sup>2</sup>SA mechanisms by incorporating advanced techniques such as complementary learning of multimodal features. Additionally, we plan to enhance the spectral feature extraction capabilities of framework in complex HSI scenes by integrating strategies for mitigating spectral variability, such as augmented linear mixture modeling [4].

## ACKNOWLEDGMENT

The authors would like to thank the State Key Laboratory of Public Big Data, Guizhou University for the computing support.

## REFERENCES

- [1] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5513412, doi: [10.1109/TGRS.2023.3279834](https://doi.org/10.1109/TGRS.2023.3279834).
- [2] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019, doi: [10.1109/MGRS.2019.2902525](https://doi.org/10.1109/MGRS.2019.2902525).
- [3] C. Wang et al., "A review of deep learning used in the hyperspectral image analysis for agriculture," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 5205–5253, Oct. 2021, doi: [10.1007/s10462-021-10018-y](https://doi.org/10.1007/s10462-021-10018-y).
- [4] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019, doi: [10.1109/TIP.2018.2878958](https://doi.org/10.1109/TIP.2018.2878958).
- [5] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 5527812, doi: [10.1109/TGRS.2023.3324497](https://doi.org/10.1109/TGRS.2023.3324497).
- [6] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.
- [7] C. Cariou and K. Chehdi, "A new k-nearest neighbor density-based clustering method and its application to hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 6161–6164, doi: [10.1109/IGARSS.2016.7730609](https://doi.org/10.1109/IGARSS.2016.7730609).
- [8] Y. E. SahIn, S. Arisoy, and K. Kayabol, "Anomaly detection with Bayesian Gauss background model in hyperspectral images," in *Proc. IEEE 26th Signal Process. Commun. Appl. Conf.*, 2018, pp. 1–4, doi: [10.1109/SIU.2018.8404293](https://doi.org/10.1109/SIU.2018.8404293).
- [9] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004, doi: [10.1109/TGRS.2004.831865](https://doi.org/10.1109/TGRS.2004.831865).
- [10] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015, doi: [10.1109/TGRS.2014.2381602](https://doi.org/10.1109/TGRS.2014.2381602).
- [11] J. Li et al., "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015, doi: [10.1109/TGRS.2014.2345739](https://doi.org/10.1109/TGRS.2014.2345739).
- [12] T. Lu, S. Li, L. Fang, X. Jia, and J. A. Benediktsson, "From subpixel to superpixel: A novel fusion framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4398–4411, Aug. 2017, doi: [10.1109/TGRS.2017.2691906](https://doi.org/10.1109/TGRS.2017.2691906).
- [13] A. Wang, S. Xing, Y. Zhao, H. Wu, and Y. Iwahori, "A hyperspectral image classification method based on adaptive spectral spatial kernel combined with improved vision transformer," *Remote Sens.*, vol. 14, no. 15, Aug. 2022, Art. no. 3705, doi: [10.3390/rs14153705](https://doi.org/10.3390/rs14153705).
- [14] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: [10.1109/MCI.2018.2840738](https://doi.org/10.1109/MCI.2018.2840738).
- [15] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi: [10.1007/s10278-019-00227-x](https://doi.org/10.1007/s10278-019-00227-x).
- [16] C. Chen et al., "Deep learning for cardiac image segmentation: A review," *Front. Cardiovasc. Med.*, vol. 7, Mar. 2020, Art. no. 25, doi: [10.3389/fcvm.2020.00025](https://doi.org/10.3389/fcvm.2020.00025).
- [17] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4550–4568, Oct. 2018, doi: [10.1109/TNNLS.2017.2766168](https://doi.org/10.1109/TNNLS.2017.2766168).
- [18] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020, doi: [10.1109/JS-TARS.2020.3005403](https://doi.org/10.1109/JS-TARS.2020.3005403).
- [19] 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, doi: [10.1109/MGRS.2019.2912563](https://doi.org/10.1109/MGRS.2019.2912563).

- [20] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9646–9660, Nov. 2021, doi: [10.1109/TGRS.2020.3046756](https://doi.org/10.1109/TGRS.2020.3046756).
- [21] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, Feb. 2023, doi: [10.1109/TIP.2023.3243853](https://doi.org/10.1109/TIP.2023.3243853).
- [22] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014, doi: [10.1109/JSTARS.2014.2329330](https://doi.org/10.1109/JSTARS.2014.2329330).
- [23] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015, doi: [10.1109/JSTARS.2015.2388577](https://doi.org/10.1109/JSTARS.2015.2388577).
- [24] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: [10.1109/TNNLS.2022.3185795](https://doi.org/10.1109/TNNLS.2022.3185795).
- [25] X. Mei et al., "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, Apr. 2019, Art. no. 963, doi: [10.3390/rs11080963](https://doi.org/10.3390/rs11080963).
- [26] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021, doi: [10.1109/TGRS.2020.3007921](https://doi.org/10.1109/TGRS.2020.3007921).
- [27] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, Mar. 2017, Art. no. 298, doi: [10.3390/rs9030298](https://doi.org/10.3390/rs9030298).
- [28] S. Hao, W. Wang, and M. Salzmann, "Geometry-aware deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2448–2460, Mar. 2021, doi: [10.1109/TGRS.2020.3005623](https://doi.org/10.1109/TGRS.2020.3005623).
- [29] M. E. Paoletti et al., "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019, doi: [10.1109/TGRS.2018.2871782](https://doi.org/10.1109/TGRS.2018.2871782).
- [30] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021, doi: [10.1109/TGRS.2020.3015157](https://doi.org/10.1109/TGRS.2020.3015157).
- [31] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, Jun. 2023, doi: [10.1109/TNNLS.2021.3109872](https://doi.org/10.1109/TNNLS.2021.3109872).
- [32] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deepwise convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, 2015, Art. no. 258619, doi: [10.1155/2015/258619](https://doi.org/10.1155/2015/258619).
- [33] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017, doi: [10.1109/TGRS.2016.2616355](https://doi.org/10.1109/TGRS.2016.2616355).
- [34] 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, doi: [10.1109/TGRS.2017.2756851](https://doi.org/10.1109/TGRS.2017.2756851).
- [35] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, Jan. 2017, Art. no. 67, doi: [10.3390/rs9010067](https://doi.org/10.3390/rs9010067).
- [36] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020, doi: [10.1109/LGRS.2019.2918719](https://doi.org/10.1109/LGRS.2019.2918719).
- [37] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018, doi: [10.1109/TGRS.2017.2755542](https://doi.org/10.1109/TGRS.2017.2755542).
- [38] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, Feb. 2020, Art. no. 582, doi: [10.3390/rs12030582](https://doi.org/10.3390/rs12030582).
- [39] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. Accessed: Jul. 23, 2022. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [40] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, Jan. 2021, Art. no. 498, doi: [10.3390/rs13030498](https://doi.org/10.3390/rs13030498).
- [41] B. Liu, A. Yu, K. Gao, X. Tan, Y. Sun, and X. Yu, "DSS-TRM: Deep spatial-spectral transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 103–114, Dec. 2022, doi: [10.1080/22797254.2021.2023910](https://doi.org/10.1080/22797254.2021.2023910).
- [42] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral-spatial feature Tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522214, doi: [10.1109/TGRS.2022.3144158](https://doi.org/10.1109/TGRS.2022.3144158).
- [43] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral-spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503615, doi: [10.1109/TGRS.2023.3242346](https://doi.org/10.1109/TGRS.2023.3242346).
- [44] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415, doi: [10.1109/TGRS.2023.3284671](https://doi.org/10.1109/TGRS.2023.3284671).
- [45] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.
- [46] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [47] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020, doi: [10.1109/TGRS.2019.2934760](https://doi.org/10.1109/TGRS.2019.2934760).
- [48] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, Jan. 2021, Art. no. 498, doi: [10.3390/rs13030498](https://doi.org/10.3390/rs13030498).
- [49] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5518615, doi: [10.1109/TGRS.2021.3130716](https://doi.org/10.1109/TGRS.2021.3130716).
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [51] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
- [52] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995, doi: [10.1109/CVPR.2017.634](https://doi.org/10.1109/CVPR.2017.634).
- [53] D. Li, A. Zhou, and A. Yao, "HBONet: Harmonious bottleneck on two orthogonal dimensions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3315–3324, doi: [10.1109/ICCV.2019.00341](https://doi.org/10.1109/ICCV.2019.00341).
- [54] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519, doi: [10.1109/CVPR.2019.00060](https://doi.org/10.1109/CVPR.2019.00060).
- [55] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16514–16524, doi: [10.1109/CVPR46437.2021.01625](https://doi.org/10.1109/CVPR46437.2021.01625).
- [56] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*.
- [57] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 464–468, doi: [10.18653/v1/N18-2074](https://doi.org/10.18653/v1/N18-2074).
- [58] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral-spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021, doi: [10.1109/TGRS.2020.3043267](https://doi.org/10.1109/TGRS.2020.3043267).
- [59] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, Jun. 2019, Art. no. 1307, doi: [10.3390/rs11111307](https://doi.org/10.3390/rs11111307).
- [60] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial-spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5532117, doi: [10.1109/TGRS.2022.3185640](https://doi.org/10.1109/TGRS.2022.3185640).
- [61] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral-spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503615, doi: [10.1109/TGRS.2023.3242346](https://doi.org/10.1109/TGRS.2023.3242346).
- [62] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 6014205, doi: [10.1109/LGRS.2022.3208935](https://doi.org/10.1109/LGRS.2022.3208935).