Spectral–Spatial Score Fusion Attention Network for Hyperspectral Image Classification With Limited Samples

Shun Cheng, Zhaohui Xue, *Member, IEEE*, Ziyu Li, Aijun Xu, and Hongjun Su, *Senior Member, IEEE*

***Abstract*—Convolutional neural network (CNN) and transformer-based models have been widely used in hyperspectral image (HSI) classification due to their excellent local and global modeling capabilities. In addition, attention mechanism is widely embedded in these models due to the effective enhancement of features learning. However, it is difficult to learn adaptive weights that effectively enhance features and most of existing methods lack transitional processing of shallow features. To overcome the above issues, a lightweight spectral–spatial score fusion attention network (S3FAN) with dual architecture is proposed for HSI classification with limited samples. Different from the regular dual branch models, S3FAN first performs pixel-level interaction and spatial feature extraction, then the obtained two sets of features are weighted and fused. In addition, we designed a spectral–spatial score fusion attention mechanism to enhance dynamic attention to spectral–spatial features. We also propose a spectral transition block to enhance model adaptability. Performance evaluation experiments conducted on five HSI datasets demonstrate that S3FAN has higher accuracy and generalization capabilities compared to existing advanced CNN and Transformer-based methods, with improvements in terms of OA around 3.18%–34.3% for Indian Pines, 5.87%–28.58% for University of Pavia, 2.57%–15.37% for Salinas, 1.64%–8.95% for Yellow River Delta, 2.87%–11.33% for WHU-Hi-LongKou, under ten samples per class.**

***Index Terms*—Hyperspectral image (HSI) classification, limited samples, spectral–spatial attention (SSA), transformer.**

I. Introduction

HYPERSPECTRAL image (HSI) can not only provide spatial information on the surface of the Earth, but also capture richer spectral information. HSIs capture hundreds of continuous bands, covering visible and near-infrared region to short-wave infrared region [1]. Benefiting from the extremely high spectral resolution of HSI, HSIs can obtain subtle changes in surface objects in different wavelength ranges. Therefore, HSI classification is widely used for mineral survey [2], environmental assessment [3], precision agriculture [4], forestry monitoring [5], target detection [6], and other research fields, as the basis for data analysis.

In the past decades, machine learning and pattern recognition methods have greatly promoted the development of HSI classification, including but not limited to support vector machine [7], random forest [8], and representation learning [9]. However, the focus on low-level handcrafted features greatly limits further exploration of the above methods.

In recent years, deep learning (DL) has been widely used in HSI classification due to its powerful ability to learn high-level features. Some typical DL frameworks are stacked autoencoders [10], deep belief networks [11], recurrent neural networks [12], convolutional neural network (CNN) [13], long short-term memory networks [14], generative adversarial network (GAN) [15], etc. Recently, Feng et al. [16] designed multi-complementary GANs with contrastive learning, prompting two groups of GANs to generate different multiscale samples to cope with the complex sample distribution in HSIs.

In the above DL framework, CNN is the most widely used due to its excellent ability for spectral and spatial feature extraction. CNN can be roughly divided into 1-DCNN [17], 2-DCNN [18], 3-DCNN [19], [20], [21], [22], and hybrid CNN methods [23], [24], [25]. Under the condition of sufficient training samples, these CNN-based methods achieve relatively good performance. However, due to the fixed receptive field, these methods are difficult to take into account both coarse-grained and fine-grained feature structures [26]. Therefore, some improved residual networks and deeper networks are proposed to enhance the ability to capture deep discriminative features and promote regularization [27], [27], [29].

As the depth of the network continues to increase, the computational complexity of the model continues to increase, and the efficiency of feature extraction is greatly reduced. In this case, in order to extract more valuable features and weaken the weight of invalid information to improve the performance of the model, the attention mechanism is introduced into the DL framework. Attention mechanisms are usually embedded into the network as modular network modules, and can be roughly divided into spectral attention, spatial attention, spatial-spectral attention, and self-attention.

1) *Spectral attention:* Spectral attention is generally a simple and effective multiplication mechanism for spectral bands,

Manuscript received 4 June 2024; revised 22 July 2024; accepted 3 August 2024. Date of publication 8 August 2024; date of current version 26 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42271324, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20221506. *(Corresponding author: Zhaohui Xue.)*

Shun Cheng, Ziyu Li, and Aijun Xu are with the School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China.

Zhaohui Xue and Hongjun Su are with the College of Geography and Remote Sensing, Hohai University, Nanjing 211100, China (e-mail: zhaohui.xue@hhu.edu.cn).

The code is available online at https://github.com/ZhaohuiXue/S3FAN.

Digital Object Identifier 10.1109/JSTARS.2024.3440254

focusing on the channel characteristics of the input data. Mou and Zhu [30] proposed a spectral attention module to learn and recalibrate strengths of different spectral bands, selectively emphasize useful bands, and suppress less informative ones. Li et al. [31] designed a novel dual-channel attention method for improving the spectral feature learning capability of the classifiers based on nonlocal and global interchannel correlations. To enhance the difference of hyperspectral data between categories, Liu et al. [32] designed a classification method based on group spectral attention with "squeeze-and-excitation" module [33] as reference.

2) *Spatial attention:* Spatial attention can calculate independent weights for heterogeneous pixels and obtain global spatial information, which may weaken spatial homogeneity and heterogeneity in HSI [34]. In [35], spatial attention module has been used to establish the spatial correlation of different features. Zhang et al. [36] proposed a novel spatial attention module to alleviate the degradation of performance as the input patch size increases by capturing the homogeneous pixels in the input patch. Xue et al. [37] added a spatial attention module to model discriminative and representative features.
3) *Spectral–Spatial attention:* The single use of spectral attention or spatial attention in the process of extracting discriminative features makes the model focus on spectral sequence or space, which inevitably leads to the neglect of the other. In this context, more and more methods use the two together, or propose spectral–spatial joint attention. Dong et al. [38] designed an attention module consisting of spectral and spatial axes, by which the salient spectral–spatial features will be emphasized. In [39], a spectral attention that implicitly implements band selection for HSI data and a spatial attention that adaptively selects spatial information from different pixels in the field are proposed and embedded in the residual block. This way of stacking the attention mechanism with the residual block greatly enhances the ability to refine features. Roy et al. [40] proposed an adaptive spectral–spatial kernel attention module to learn selective 3-D convolutional kernels for HSI classification, and Lu et al. [41] proposed a 3-D channel and spatial attention effectively express the region of interest and conducive to information flow within the network. Xie et al. [42] adaptively guided the basic CNN to focus on different spectral channels and spatial positions through global spectral–spatial attention incorporated into each convolutional layer, achieving the capture of discriminative features from shallow to deep layers.

Although the introduction of the attention mechanism has greatly promoted the performance improvement of the above-mentioned CNN-based methods, it essentially does not escape the disadvantages of CNN focusing on local features and lacks pixel-level interaction of spectral sequence features. Information is lost during the downsampling process, and deep networks consume a lot of computing resources [43]. This characteristic makes CNN more capable of extracting local spatial features than modeling the global dependencies of spectral sequences.

Recently, transformer architecture [44] formed by combining the self-attention mechanism with the multilayer perceptron (MLP) and the demonstration of Vision Transformer (ViT) [45] promotes the introduction of this excellent modeling ability of long-distance dependence into HSI classification to overcome the shortcomings of CNNs in global perception. However, transformer-based methods exhibit robust capabilities for global spectral feature modeling but struggle to extract local spatial features effectively [46]. To give more consideration to the hierarchy of feature extraction and further improve the performance of the transformer, many transformer-based methods incorporating convolution have been proposed.

1) *Transformer with traditional structure:* In [47], a transformer with dense connection is designed to capture spectral sequence relationships. From a sequence perspective, Hong et al. [48] implemented learning spectral local sequence information from adjacent bands of HSIs. To capture much more spectral–spatial information, Yang et al. [49] proposed to encode the input representations along the height, width, and spectral dimensions. Xu et al. [50] reconstructed Swin Transformer [51] in 1-D space for patchwise HSI classification. Tang et al. [52] designed a newly double-attention transformer encoder, which fuses the local spatial information with global spectral features, to maximize spectral–spatial information fusion.
2) *Transformer combined with CNN:* Sun et al. [53] used CNN to extract shallow spectral–spatial features before extracting deep semantic features. In [54], grouped convolution was used to extract discriminative spatial-spectral features from nonoverlapping subchannels. Ouyang et al. [55] utilized the local representation ability of CNN to contribute richer image features to Transformer model. In [56], CNN kernel is used in the spectral transformer to refine the spectral value. Roy et al. [57] used spectral and spatial morphological convolution combined with attention mechanisms to improve the interaction between structural and shape information. Zhang et al. [58] combined multiattention and transformer to select bands and spatial areas to pay more attention to the key areas. In [59], a hierarchical attention designed to replace self-attention, promoting the effective combination of CNN and transformer.

Although all the above methods have good performance when the training samples are sufficient, these deep models are prone to overfitting problems when the training samples are severely constrained [60]. In order to deal with few labeled samples, the most challenging issue in hyperspectral classification scenarios [61], some transfer learning [62], [63], [64], active learning [65], [66], [67], and few-shot learning (FSL) [54] methods have been proposed. In transfer learning, Yang et al. [68] proposed an effective transfer learning method that used hierarchical deep neural networks for shallow feature transfer and deep feature

classification to retain source domain features. In [69], a generative domain adaptation method was proposed to make the discriminant boundary of the classifier more suitable for the target domain. Qin et al. [70] designed a novel cross domain method based on feature disentanglement to preserve discriminative information from the heterogeneous data space. In active learning, an iterative semisupervised CNN framework [71] was proposed through active learning and superpixel segmentation techniques. Zhao et al. [72] proposed an adaptive superpixel segmentation active learning framework to select important samples for the transformer model to cope with limited labeled samples tasks. As for FSL, Xi et al. [73] proposed a deep prototypical network with hybrid residual attention, which can effectively investigate the spectral–spatial information in the HSI. Zeng et al. [74] designed a multistage relational network with dual metric to effectively represent the class distribution with fewer labeled samples. In [75], a spectral–spatial siamese network (S3Net) was proposed to resolve the challenge of overfitting in DL models.

The above DL methods have all achieved excellent performance in HSI classification. However, these methods still have many limitations when dealing with limited samples.

1) The CNN-based methods have strong ability to capture local spatial relationships benefit from the excellent local extraction capabilities and parameter sharing mechanism, but lack the ability to model long-distance relationships between pixels, and are difficult to ensure the continuity of spectral sequence features.
2) The transformer-based methods can ensure the integrity of the spectral sequence and the modeling of long-distance dependencies. However, the lack of hierarchical feature extraction results in models that often rely on deeper encoders and lack the ability to represent local spatial features. Although some models utilize CNN to extract shallow features, the lack of feature transition processing leads to frequent linear mapping that destroys feature correlation.
3) Existing transfer learning, active learning, and FSL methods for limited samples are overly reliant on complex training techniques and data preprocessing, such as data augmentation, domain adaptation, and metric learning. While enriching the number of trainable samples, they inevitably increase the complexity of training.
4) Although the existing attention mechanism can focus on spectral or spatial tasks, the lack of feature similarity learning leads to insufficient intermediate feature representation capabilities, making it difficult to provide adaptive weights for the feature extraction process.

Very recently, a new white-box transformer is proposed [76], in which a multihead subspace self-attention (MSSA) operator is designed to replace the self-attention mechanism in traditional ViTs as the gradient descent step for compressing the token sets, and achieves better performance. MSSA greatly reduces the number of parameters, which creates excellent conditions for feature fusion of two-stream network architecture.

In this article, we proposed a dual architecture lightweight spectral–spatial score fusion attention network (S3FAN) that achieves efficient weighted extraction and feature recalibration and fuses spectral–spatial features and spectral interaction features. First, after dimensionality reduction using principal component analysis (PCA), spectral score fusion attention (SpeSFA) is designed to achieve adaptive weighting of spectral features, and use lightweight spectral ResNet to realize feature dimensions first enlarging and then compressing. Second, we designed a spectral transition (ST) block, which combined fully connection and convolution together to perform feature recalibration on the spectral features compressed by spectral ResNet. Third, MSSA operator is adopted to form transformer encoder to capture spectral interaction features and spatial score fusion attention (SpaSFA) and spatial ResNet are designed to obtain spectral–spatial features. Finally, the spectral interaction features and spectral–spatial features are aligned in the feature dimension and weighted fused, and global average pooling (GAP) is used for classification.

The main contributions of our work can be summarized as follows.

1) Score fusion attention, including SpeSFA and SpaSFA of dynamically updated weights and fully considered feature similarities are designed to obtain fused attention scores. With these two attentions, adaptive attention weights for spectral and spatial extraction are calculated to enhance feature extraction capabilities.
2) ST block is designed to achieve reweighting of spectral features and perform global and local feature recalibration on the extracted spectral features to adapt to the subsequent feature extraction process. Experiments show that this transitional processing of extracted features has a significant effect on further improving performance.
3) Spectral ResNet is constructed to achieve pre-extraction of spectral features. MSSA is introduced to match MLP to achieve pixel-level spectral sequence interaction of self-correlation, and the resulting spectral interactive features are weighted fused with spatial features extracted by spatial ResNet. The proposed S3FAN shows advantages over other advanced methods in its performance on multiple datasets for HSI classification with limited samples.

II. Proposed Methodology

An overview of the proposed S3FAN is shown in Fig. 1. First, PCA dimensionality reduction is performed on the input HSI dataset, and then the patch $X \in \mathbb{R}^{h\times w\times C}$ centered on the target pixel is fed to the spectral weighted extraction and transition module (SWETM). Second, through SpeSFA and spectral ResNet in SWETM, high-dimensional intermediate features are extracted and compressed to varying degrees to obtain dual-branch output features. And then, ST blocks are used to reweight spectral features. Third, the two branch features are input to spectral subspace interaction module (SSIM) and

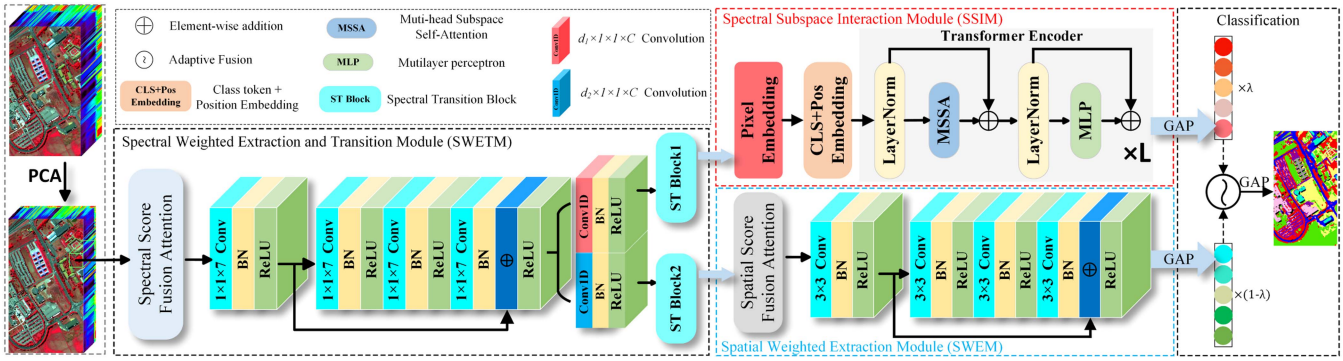


Fig. 1. Graphical illustration of the proposed S3FAN.

TABLE I
ARCHITECTURAL DETAILS OF THE PROPOSED S3FAN

Module	Layer Specification		Input size	Output size
SWETM	SpeSFA	3×1×1×7 Conv3-D $C/2$×1×1×7 Conv3-D 1×1×7 Conv3-D	$h \times w \times C$	$h \times w \times C$
	Spectral ResNet	10×1×1×7 Conv3-D* 4 d_i×1×1×C Conv3-D	$h \times w \times C$	$h \times w \times d_i (i = 1, 2)$
	ST Block	1×3 Conv1D Linear(dim = C)	$h \times w \times d_i$	$h \times w \times d_i$
SSIM	Pixel Embedding	Linear(dim=63)	$h \times w \times d_1$	$h^* w \times d_1$
	Transformer Encoder	MSSA(heads = 3, headdim = 21) MLP(dim = 63)		
SWEM	SpaSFA	d_2×3×3×1 Conv3-D $d_2/2$×3×3×1 Conv3-D 1×1×7 Conv3-D	$h \times w \times d_2$	$h \times w \times d_2$
	Spatial ResNet	3×3 Conv2-D * 4	$h \times w \times d_2$	$h \times w \times d_2$
Classification	Dimension Alignment	GAP	$h^* w \times d_1$,$h \times w \times d_2$	$1 \times n$
	Prediction	GAP		

SWETM: Spectral weighted extraction and transition module; SSIM: Spectral subspace interactive module; SWEM: Spatial weighted extraction module.

spatial weighted extraction module (SWEM), respectively. And two deep discriminant features obtained are first dimensionally reduced through GAP and then weighted and fused to achieve the complementarity of spectral interactive features and spectral–spatial features, the resulting fusion features are presented in the form of low-dimensional vectors. Finally, the fused features are classified through GAP, a parameterless classification head, and the classification result will be used as the predicted label of the central pixel of the original patch. Layer specification as well as the input and output size of each module in the proposed S3FAN are presented in Table I.

A. Spectral Weighted Extraction and Transition Module

1) Spectral Score Fusion Attention: As shown in Fig. 2, the input patch $x^l \in \mathbb{R}^{h \times w \times C}$ first enters a 1-D convolution layer (this module is used as a 1-D convolution by setting the spatial size of the 3-D convolution to 1) to simply extract spectral discriminative features, then performs residual connection to obtain intermediate features $x^{l+1} \in \mathbb{R}^{h \times w \times C}$, the formulas are as follows:

$$x^{l+\frac{1}{2}} = f^1(x^l) + x^l \tag{1}$$

$$f_1(x^l) = \text{ReLU}\left(\text{BN}\left(\text{Conv}(x^l)\right)\right) \tag{2}$$

$$\text{Conv}(x^l) = W_l \cdot x^l + b_l \tag{3}$$

$$\text{BN}(x) = \frac{x - \text{mean}(x)}{\sqrt{\text{var}(x) + \epsilon}} \cdot \gamma + \beta \tag{4}$$

$$\text{ReLU}(x) = \max(x, 0) \tag{5}$$

where $f^i(\cdot)$ is a complete convolution layer, including a 1-D convolution Conv$(\cdot)$, batch normalization layer BN$(\cdot)$, and ReLU activation function ReLU$(\cdot)$. In (3), W_l is the weight matrix and b_l is the bias. In (4), γ and β are learnable parameter vectors and ϵ is a parameter for numerical stability.

Then, use 2-D average pooling and 2-D maximum pooling, respectively, on the intermediate features $x^{l+\frac{1}{2}} \in \mathbb{R}^{h \times w \times C}$ to obtain two representation vectors $v_{\text{avg}}, v_{\max} \in \mathbb{R}^{1 \times 1 \times C}$:

$$v_{\text{avg}} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} x^{l+\frac{1}{2}}(i, j) \tag{6}$$

$$v_{\max} = \max\left(x^{l+\frac{1}{2}}(i, j)\right) \tag{7}$$

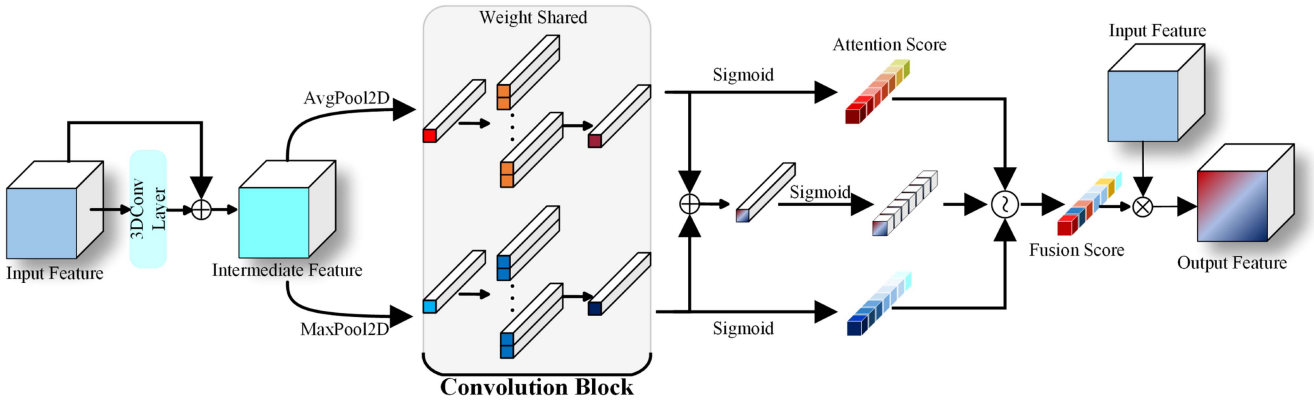


Fig. 2. Graphical illustration of SpeSFA.

In this way, the most significant features and global features of each HSI band can be extracted to form two sets of feature vectors.

To achieve dimension scaling of the feature vector to enrich the features and extract them, vectors are input into the convolution block. This convolution block is composed of two convolutional layers.

$$[v'_{\text{avg}}, v'_{\text{max}}] = \text{ReLU}\left(\text{Conv}^1([v_{\text{avg}}, v_{\text{max}}])\right) \tag{8}$$

$$v'_{\text{avg}}, v'_{\text{max}} \in \mathbb{R}^{m\times 1\times 1\times C}$$

$$[v^*_{\text{avg}}, v^*_{\text{max}}] = \text{ReLU}\left(\text{Conv}^2([v'_{\text{avg}}, v'_{\text{max}}])\right)$$

$$v^*_{\text{avg}}, v^*_{\text{max}} \in \mathbb{R}^{1\times 1\times C} \tag{9}$$

where v'_{avg}, v'_{max} are the output features of the first convolutional layer, m is the dimension of feature vectors, and v^*_{avg} and v^*_{max} are the output discriminant vectors. The weights of the convolutions in this process are shared to learn the similarity of the two feature vectors, and help reduce the number of parameters and accelerate the convergence of the model.

To achieve dynamic attention to spectral features, first, v^*_{avg} and v^*_{max} are summed to promote information interaction and enrich feature expression to obtain v_{sum}. Second, v^*_{avg}, v^*_{max}, and v_{sum} are adaptively weighted and summed after calculating the attention scores through sigmoid to obtain the attention of score fusion, this treats each set of features independently to introduce more nonlinear features. Finally, these three scores are multiplied with the input features as follows:

$$v_{\text{sum}} = v^*_{\text{avg}} + v^*_{\text{max}} \tag{10}$$

$$s_e = r_1 \times \sigma(v_{\text{sum}}) + r_2 \times \sigma(v^*_{\text{avg}}) + r_3 \times \sigma(v^*_{\text{max}}) \tag{11}$$

$$x^{l+1} = s_e \cdot x^l \tag{12}$$

where s_e is the spectral attention score, r_1, r_2, and r_3 are the adaptive normalized weights, and x^{l+1} is the output feature map of spectral attention.

2) Spectral ResNet: The lightweight spectral ResNet we designed has been presented in Fig. 1. After performing a convolutional layer to extract features and expand the feature dimension, insert the residual connection and perform subsequent extraction

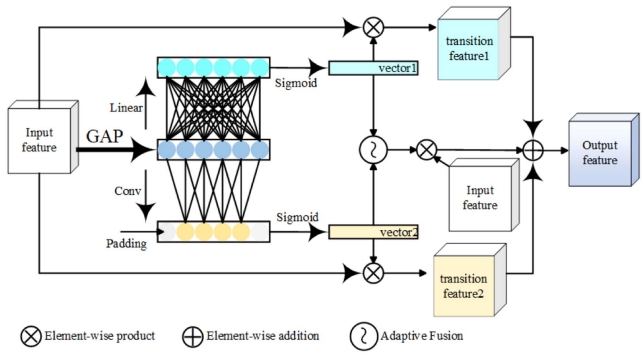


Fig. 3. Graphical illustration of ST block.

work. The formula is expressed as follows:

$$x^{l+2} = f^2(x^{l+1}) \tag{13}$$

$$x^{l+2} = [x_1^{l+2}, x_2^{l+2}, \ldots, x_{10}^{l+2}] \in \mathbb{R}^{10\times h\times w\times C}$$

where x^{l+2} is the high-dimensional feature map after the first convolutional layer. To ensure sufficient spectral features for spectral semantic extraction and reduce spectral intervention during spatial information extraction, differential output is adopted for spectral ResNet as follows:

$$F^i_{\text{SpeRes}}(x^{l+1}) = f_i^5\left(\text{ReLU}\left(x^{l+2} + BN\left(f^4(f^3(x^{l+2}))\right)\right)\right)$$

$$x_i^{l+3} = F^i_{\text{SpeRes}}(x^{l+1}) \in \mathbb{R}^{h\times w\times d_i} \tag{14}$$

where $x_i^{l+3} \in \mathbb{R}^{h\times w\times d_i}$ is the output feature of spectral ResNet. We believe that spectral features need to be treated differently for semantic extraction and spatial extraction. The feature of the upper branch is compressed into $x_1^{l+3} \in \mathbb{R}^{h\times w\times C}(d_1 = C)$ to ensure the continuity and completeness of the spectral sequence. The lower branch is $x_2^{l+3} \in \mathbb{R}^{h\times w\times KS}(d_2 = KS < C)$ to weaken the impact of spectral dimensions on spatial extraction.

3) ST Block: As shown in Fig. 3, ST block efficiently combines global and local interactive sequence information. Specifically, GAP is first used on the input feature $x_i^{l+3} \in \mathbb{R}^{h\times w\times d_i}$ to obtain the spectral representation vector. Linear mapping and

1-D convolution are used to obtain spectral weight values ω_i^1 and ω_i^2, thereby achieving complete and local modeling of spectral sequence correlation. The input feature is then reweighted according to the spectral dimension. In order to further realize the global and local interaction of the spectral sequence and realize spectral feature recalibration, the two weight vectors are weighted and summed and the input features are reweighted. Finally, the three reweighted features are summed. The corresponding formulas are as follows:

$$\begin{aligned}
&\text{Conv1}d(x, W^*) = \sum_{i=1}^{k} x_i \cdot W_i^* \\
&\omega_i^1 = \sigma\left(\text{Conv1}d\left(\frac{1}{h \times w}\sum_{k=1}^{h}\sum_{j=1}^{w} x_i^{l+3}(k,j), W^*\right)\right) \\
&\omega_i^2 = \sigma\left(\text{Linear}(\frac{1}{h \times w}\sum_{k=1}^{h}\sum_{j=1}^{w} x_i^{l+3}(k,j), W_1)\right) \\
&x_i^{l+4} = (r_1' \times \omega_i^{1\tilde{}T} + r_2' \times \omega_i^{2\tilde{}T}) \cdot x_i^{l+3} + \omega_i^{1\tilde{}T} \cdot x_i^{l+3} \\
&\qquad + \omega_i^{2\tilde{}T} \cdot x_i^{l+3}
\end{aligned} \tag{15}$$

where W^* is the weight of 1-D convolution convolution. Linear$(\cdot)$ represents linear transformation. $x_i^{l+4} \in \mathbb{R}^{h \times w \times d_i}$ denotes the output of ST block. Specifically, the output of the upper branch (ST block1) is $x_1^{l+4} \in \mathbb{R}^{h \times w \times C}$, whereas the output of the lower branch (ST block2) is $x_2^{l+4} \in \mathbb{R}^{h \times w \times KS}$.

B. Spectral Subspace Interactive Module

As shown in Fig. 1, the output feature of ST block1 $x_1^{l+4} \in \mathbb{R}^{h \times w \times C}$ serves as input into SSIM to achieve spectral sequence interaction of pixels. Before entering the transformer encoder, $x_1^{l+4} \in \mathbb{R}^{h \times w \times C}$ performs preprocessing. To highlight the central pixel of the patch and reduce the weakening of the central pixel by the spatial information brought by the surrounding pixels, so that the transformer can focus more on realizing the global interaction of the spectral sequence, we use pixel embedding in the tokens mapping. First, patch-level features are stretched into pixel-level tokens $x_1^{l+4} = [y_1^{l+4}, y_2^{l+4}, \ldots, y_{h^*w}^{l+4}] \in \mathbb{R}^{h^*w \times C}$ through pixel embedding. $y_i^{l+4} \in \mathbb{R}^{1 \times C} (i = 1, 2, \ldots, h^*w)$ denote pixel-level spectral vectors. Then, tokens are mapped to the hidden layer $x_1^{l+4} \in \mathbb{R}^{h^*w \times C} \rightarrow X \in \mathbb{R}^{h^*w \times D}$ by linear projection (D is the dimension of the hidden layer), and then, class token and position embedding are performed, its expression is as follows:

$$\begin{aligned}
&X_i = y_i^{l+4} \cdot \text{E} \in \mathbb{R}^{1 \times D}, X_{[\text{CLS}]} \in \mathbb{R}^{1 \times D}, \text{PE}_{\text{pos}} \in \mathbb{R}^{(h^*w+1) \times D} \\
&X = [X_{[\text{CLS}]}, X_1, X_2, \ldots, X_{h^*w}] + \text{PE}_{\text{pos}}
\end{aligned} \tag{16}$$

where E represents the linear projection. $X_{[\text{CLS}]}$ and PE_{pos} denote learnable classification token and positional information, respectively.

Next, we iteratively perform MSSA and MLP, and the following represents the formula for the lth layer:

$$Z_{\frac{l}{2}} = \text{MSSA}(\text{LN}(X_l)) + \text{LN}(X_l) \tag{17}$$

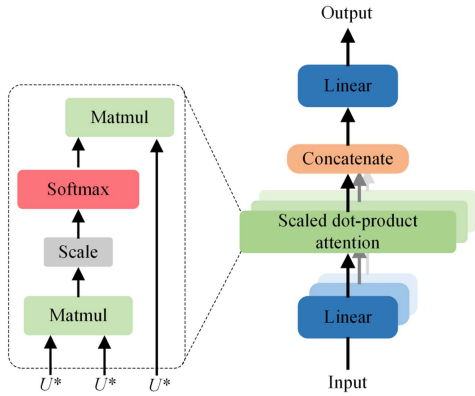


Fig. 4. Graphical illustration of MSSA operator.

$$Z_l = \text{MLP}(\text{LN}(Z_{\frac{l}{2}})) + \text{LN}(Z_{\frac{l}{2}}) \tag{18}$$

where $Z_{\frac{l}{2}}$ and Z_l denote the output features of MSSA module and MLP module, respectively. LN$(\cdot)$ is layer normalization. As in (17) and (18), residual connection prevents gradient disappearance. The combination of MSSA and MLP enables the global correlation between sequences to be fully considered, and the spectral sequence features to fully interact. MSSA compresses the sequence composed of input spectral vectors, extracts global relationships, and aggregates information. MLP projects aggregated information into a specific semantic space. More details are given below.

1) Layer Normalization (LN): LN is an important means to ensure stable training and faster convergence of the model and effectively avoid gradient explosion. LN is applied over each input feature as follows:

$$\text{LN}(X) = \frac{X - \mu}{\delta} \cdot \Gamma + \boldsymbol{\beta} \tag{19}$$

where μ and δ are the mean and standard deviation of X, while Γ and $\boldsymbol{\beta}$ are learnable affine transform parameters.

2) Mutihead Subspace Self-Attention (MSSA): To further learn the pixel-level global dependence of spectral sequences, we use MSSA [76], a self-attention mechanism that is different from MHSA. Fig. 4 shows the MSSA operator process. Sufficient correlation calculations are performed on the pixel-level spectral sequence features in the subspace to achieve global modeling of the spectral sequence. In detail, the linear operators of value, key, and query are all set to be the same as the subspace basis U^*, which greatly reduces the number of parameters. The counting process of MSSA can be formulated as follows:

$$\begin{aligned}
&V = K = Q = U^* \\
&\text{SSA} = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \\
&\text{MSSA}(Q, K, V) = \text{Concat}(\text{SSA}_1, \text{SSA}_2, \ldots, \text{SSA}_h)W_0
\end{aligned} \tag{20}$$

where h is the number of heads and W_0 is the parameter matrix.

3) MLP: MLP excels in modeling long-range dependencies of markers in sequences, consisting of linear projection and

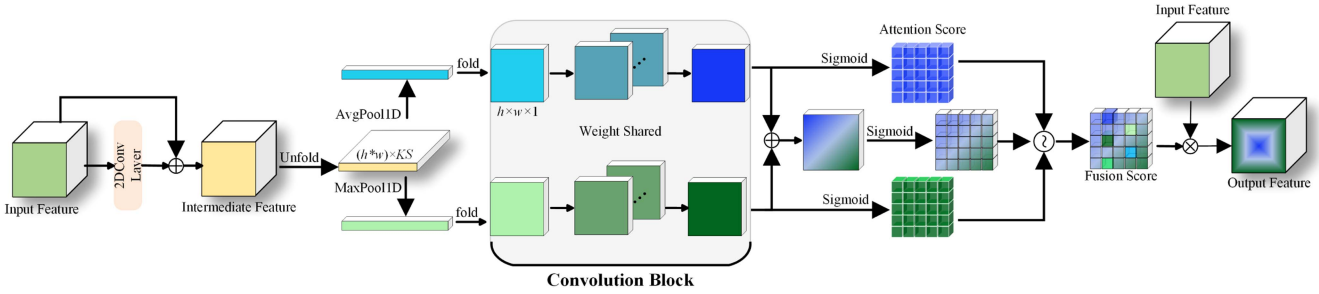


Fig. 5. Graphical illustration of SpaSFA.

GELU activation function. The specific formulas are as follows:

$$\mathrm{GELU}(x) = 0.5x\left(1+\mathrm{Tanh}\left(\sqrt{\frac{2}{\pi}}(x+0.044715x^3)\right)\right) \tag{21}$$

$$\mathrm{MLP}(z) = \mathrm{Linear}\left(\mathrm{GELU}(\mathrm{Linear}(z,W')),W''\right) \tag{22}$$

where x and z denote input feature maps.

C. Spatial Weighted Extraction Module

1) Spatial Score Fusion Attention: As shown in Fig. 5, SpaSFA is structurally similar to SpeSFA, but changes have been made in specific components to adapt to the spatial dimension and weaken the influence of spectral dimension. In the pre-extraction and convolution block, 2-D convolution that is more focused on extracting spatial features and spectral dimension are used. In addition, after unfolding the spatial domain, a 1-D pooling operation was performed on the spectral domain to achieve the purpose of weakening the spectral dimension and obtain two sets of representation matrices of spatial features. In this way, the size of the spatial dimension will not be lost due to pooling, avoiding the destruction of spatial information. The details are described below.

The output feature of ST block2 is $x_2^{l+4} \in \mathbb{R}^{h\times w\times KS}$. First, perform a convolution pre-extraction and residual connection on it to get x_2^{l+5}, and then unfold it: $x_2^{l+5} \in \mathbb{R}^{h\times w\times KS} \rightarrow x_2^{l+6} \in \mathbb{R}^{h^*w\times KS}$. Next, 1-D average pooling and 1-D maximum pooling are applied to it, respectively, to obtain two spatial matrix S_{avg} and S_{max}. Then, input the two matrix into the weight-shared convolution block to obtain the deep features and learn the spatial similarity. Then, use sigmoid on the obtained deep features, respectively, and use sigmoid on the summed features and perform weighted fusion to obtain the fused spatial score S_a. Finally, the spatial fraction is inner product along the spectral dimension of the input feature as follows:

$$x_2^{l+4} = [X_1^{l+4}, X_2^{l+4}, \ldots, X_{KS}^{l+4}], X_i^{l+4} \in \mathbb{R}^{h\times w\times 1}$$

$$Z^i = X_i^{l+4} \odot S_a = \begin{bmatrix} X_{11}^i S_{11} & X_{12}^i S_{12} & \cdots & X_{1\text{-}w}^i S_{1\text{-}w} \\ X_{21}^i S_{21} & X_{22}^i S_{22} & \cdots & X_{2\text{-}w}^i S_{2\text{-}w} \\ \vdots & \vdots & \ddots & \vdots \\ X_{h1}^i S_{h1} & X_{h2}^i S_{h2} & \cdots & X_{hw}^i S_{hw} \end{bmatrix} \tag{23}$$

$$z^{l+4} = \mathrm{Concat}(Z^1, Z^2, \ldots, Z^{KS}) \in \mathbb{R}^{h\times w\times KS} \tag{24}$$

where $\odot$ represents Hadamard product and z^{l+4} denotes the output feature of spatial attention.

2) Spatial ResNet: As shown in Fig. 1, Spatial ResNet consists of a reasonable combination of four 2-D convolutional layers and a residual connection to simply and effectively extract spatial features. In order to ensure the focus on extracting local spatial information, the feature dimension remains unchanged during this process to avoid unnecessary information redundancy. The formula is as follows:

$$z^{l+5} = F_{\mathrm{SpaRes}}(z^{l+4}) \in \mathbb{R}^{h\times w\times KS} \tag{25}$$

where z^{l+5} is the output feature map.

D. Adaptive Feature Fusion and Classification

The complementary information between features of different scales can be aggregated to continuously enhance the representation ability of features [77]. To achieve the complementarity between the global pixel-level spectral features obtained by SSIM and the local patch-level spatial features captured by SWEM and obtain the most discriminative features, thereby further improving classification performance, we perform weighted fusion after dimensionally aligning the spectral interative features $Z_L \in \mathbb{R}^{(h^*w+1)\times D}$ obtained by SSIM and the spatial-spectral features $z^{l+5} \in \mathbb{R}^{h\times w\times KS}$ obtained by SWEM, the details are described below.

First, class token $X'_{[\mathrm{CLS}]}$ employed to represent features for subsequent classification are extracted from spectral interactive features $Z_L \in \mathbb{R}^{(h^*w+1)\times D}$. After identity mapping of the class token, 1-D average pooling is used to achieve parameter-free dimensionality reduction to obtain $\mathtt{X}_{\mathrm{se}} \in \mathbb{R}^{1\times c}$. At the same time, 1-D average pooling is also used to the reshaped spectral–spatial features $z^{l+5} \in \mathbb{R}^{h^*w\times KS}$ to obtain $\mathtt{X}_{\mathrm{sa}} \in \mathbb{R}^{1\times c}$. Then, the two discriminant features are weighted and fused to obtain spectral–spatial fusion features $\mathtt{X}_{\mathrm{fused}} \in \mathbb{R}^{1\times c}$. Finally, parameter-free 1-D average pooling is still used as the classification head for classification. The formula represent feature fusion is as follows:

$$\mathtt{X}_{\mathrm{fused}} = \lambda \times \mathtt{X}_{\mathrm{se}} + (1-\lambda) \times \mathtt{X}_{\mathrm{sa}} \tag{26}$$

where λ is the adaptive normalized weight.

In the entire process of feature dimension alignment and classification, the reason why parameter-free 1-D average pooling is

Algorithm 1: S3FAN.

Input: HSI data $\hat{X} \in \mathbb{R}^{H\times W\times B}$ and ground truth image $Y \in \mathbb{R}^{H\times W}$, number of PCs C, training epoch for S3FAN $epoch$, depth of transformer encoder L.
Output: Predicted labels for HSI.

Reduce the demension of HSI by PCA.
Split training set and testing set, build dataloaders.
for i=1 to $epoch$ **do**
  I. SWETM
  Obtain weighted features $\hat{x}$ through SpeSFA;
  Apply (14) on $\hat{x}$ to get output features $\hat{x}_1 \in \mathbb{R}^{h\times w\times C}$ and $\hat{x}_2 \in \mathbb{R}^{h\times w\times KS}$;
  Apply ST Blocks on $\hat{x}_1$ and $\hat{x}_2$ to get reweighted features $\hat{x}_1^1$ and $\hat{x}_2^1$.
  II. SSIM and SWEM
  While using pixel embedding and (16) for $\hat{x}_1^1$, use SpaSFA for $\hat{x}_2^1$ to get $\hat{x}_1^2$ and $\hat{x}_2^2$.
  for j=1 to L **do**
  Apply (17) and (18) on $\hat{x}_1^2$.
  end
  Apply (25) on $\hat{x}_2^2$.
  III. Adaptive Feature Fusion and Classification
  For the output of SSIM and SWEM, after aligning the dimensions, using (26) to get fused feature;
  Use GAP to obtain classification vectors;
  Get the predicted label using arg max;
  Calculate $L_{\text{cre}-\text{LS}}$ using (28) with Adam optimizer.
end

used as a means of dimensionality reduction and classification is because the extraction of deep discriminative features has been completed in the previous process. Parameter-free pooling can reduce the overall parameter amount of the model and avoid overfitting.

Label smoothing cross-entropy loss function is used to calculate training loss and update model weights, the formula can be represented as follows:

$$L_{\text{cre}} = -\sum_{i=0}^{N} y_i \log(\hat{y}_i) \tag{27}$$

$$L_{\text{cre}-LS} = (1-\epsilon_0)L_{\text{cre}} - \frac{\epsilon_0}{N}\sum_{i=0}^{N}\log(\hat{y}_i) \tag{28}$$

where N is the number of categories, y_i and $\hat{y}_i$ are the true label and the predicted class probabilities of the ith sample. ϵ_0 represents label smoothing factor.

At the end of this section, the pseudocode of S3FAN algorithm is shown in Algorithm 1.

III. Experimental Results

A. Hyperspectral Datasets

We select five publicly available and popular datasets including Indian Pines, University of Pavia, Salinas, Yellow River Delta, and WHU-Hi-LongKou to verify the proposed method.

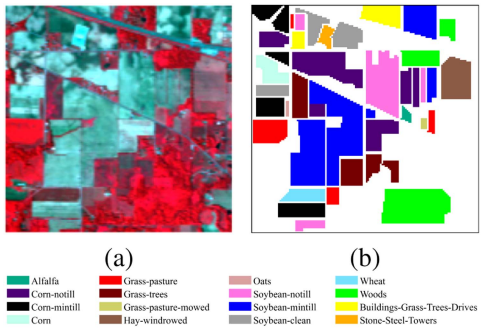


Fig. 6. Indian Pines. (a) False-color composite image (R: 50, G: 27, B: 17). (b) Ground-truth map.

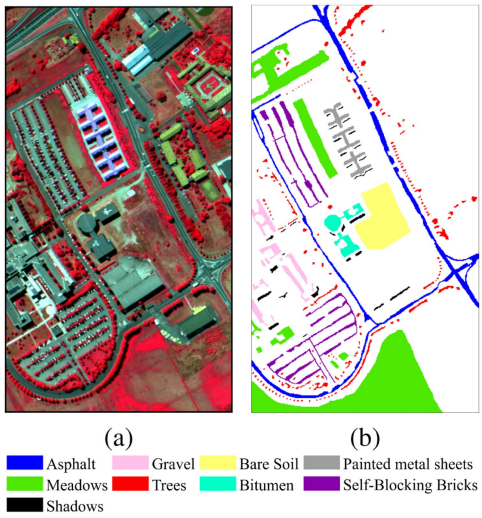


Fig. 7. University of Pavia. (a) False-color composite image (R: 102, G: 56, B: 31). (b) Ground-truth map.

All of them are used for parameter sensitive analysis, ablation study, and to evaluate the classification performance. The number of labeled samples corresponding to different categories in different datasets is presented in Table II. And the false-color composite images and ground-truth maps corresponding to these five datasets are shown in Figs. 6–10, respectively.

B. Experimental Settings

1) Hyperparameters: For training epoch, the training process would be stopped when the training loss stabilizes. In the experiment for five datasets, the training epoch was uniformly set to 50, and the training loss is stable under this setting. For all datasets, the batch size is set to 32. In addition, other hyperparamerters, including the number of principal component, patch size, learning rate, depth of transformer encoder, and the input dimensions of ST block2 are analyzed in the experiment of parameter sensitive analysis.

2) Training Samples: To evaluate the classification performance of the model under limited sample conditions, we select ten labeled samples per class for training randomly, and the rest of samples are constructed for testing, in the experiment

TABLE II
NUMBER OF LABELED SAMPLES PER CLASS FOR FIVE DATASETS

Class	Indian Pines		University of Pavia		Salinas		Yellow River Delta		WHU-Hi-LongKou	
	Name	Number	Name	Number	Name	Number	Name	Number	Name	Number
1	Alfalfa	46	Asphalt	6631	Broccoli green weeds 1	2009	Reed	310	Corn	34 511
2	Corn-notill	1428	Meadows	18 649	Broccoli green weeds 2	3726	Spartina alterniflora	187	Cotton	8374
3	Corn-mintill	830	Gravel	2099	Fallow	1976	Salt filter pond	247	Sesame	3031
4	Corn	237	Trees	3064	Fallow rough plow	1394	Salt evaporation pond	300	Broad-leaf soybean	63 212
5	Grass-pasture	483	Painted metal sheets	1345	Fallow smooth	2678	Dry pond	140	Narrow-leaf soybean	4151
6	Grass-trees	730	Bare soil	5029	Stubble	3959	Tamarisk	127	Rice	11 854
7	Grass-pasture-moved	28	Bitumen	1330	Celery	3579	Salt pan	306	Water	67 056
8	Hay-windrowed	478	Self-Blocking Bricks	3682	Grapes untrained	11 271	Seepweed	218	Roads and houses	7124
9	Oats	20	Shadows	947	Soil vineyard develop	6203	River	584	Mixed weed	5229
10	Soybeans-notill	972			Corn senesced green weeds	3278	Sea	4694		
11	Soybeans-mintill	2455			Lettuce romaine 4wk	1068	Mudbank	14		
12	Soybeans-clean	593			Lettuce romaine 5wk	1927	Tidal creek	67		
13	Wheat	205			Lettuce romaine 6wk	916	Fallow land	459		
14	Woods	1265			Lettuce romaine 7wk	1070	Ecological restoration pond	310		
15	Bldg-grass-tree-drivers	386			Vineyard untrained	7268	Robinia	111		
16	Stone-streel-towers	93			Vineyard vertical trellis	1807	Fishpond	124		
17							Pit pond	128		
18							Building	398		
19							Bare land	87		
20							Paddyfield	508		
21							Cotton	332		
22							Soybean	71		
23							Corn	103		
	Total	10 249	Total	42 776	Total	54 129	Total	9825	Total	204 542

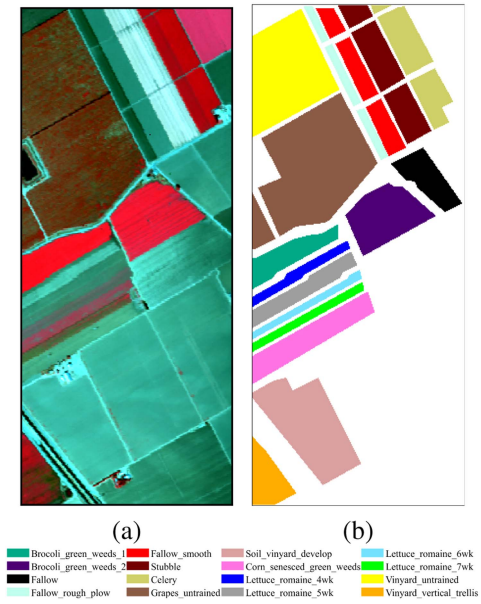


Fig. 8. Salinas. (a) False-color composite image (R: 57, G: 27, B: 17). (b) Ground-truth map.

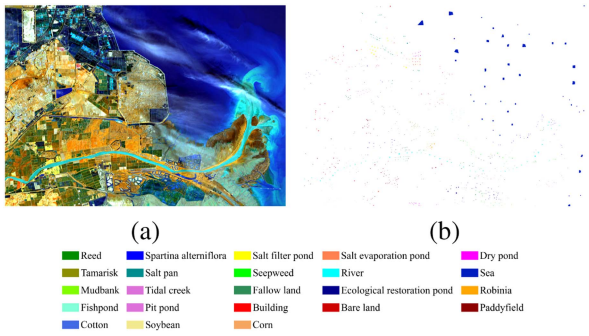


Fig. 9. Yellow River Delta. (a) False-color composite image (R: 55, G: 28, B: 8). (b) Ground-truth map.

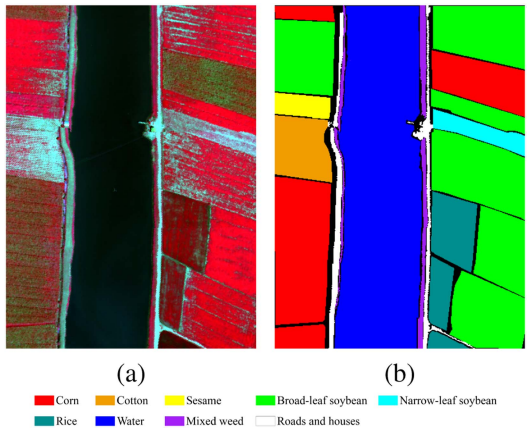


Fig. 10. WHU-Hi-LongKou. (a) False-color composite image (R: 174, G: 118, B: 57). (b) Ground-truth map.

of ablation study and performance comparison. If there are few labeled samples for a certain class that the selected samples are less than half of the total labeled samples, we select half of the labeled samples for the train set, and the remaining samples for the test set. In addition, 20 labeled samples per class are randomly selected for the parameter sensitive analysis. However, for the generalization performance experiment, we set the labeled samples per class in the range of [5, 10, 15, 20, 30, 40, 50].

3) Performance Comparison: We compare the proposed method with other advanced DL methods, including six CNN-based methods and five transformer-based methods. Specifically, the CNN-based methods include three common DL methods, 3-DCNN [19], HybridSN [25], and spectral–spatial residual network (SSRN) [28], and the other three methods are residual spectral–spatial attention network (RSSAN) [39], which inserts spectral–spatial attention into residual network, A^2S^2K ResNet [40], which has a spectral–spatial kernel attention, there

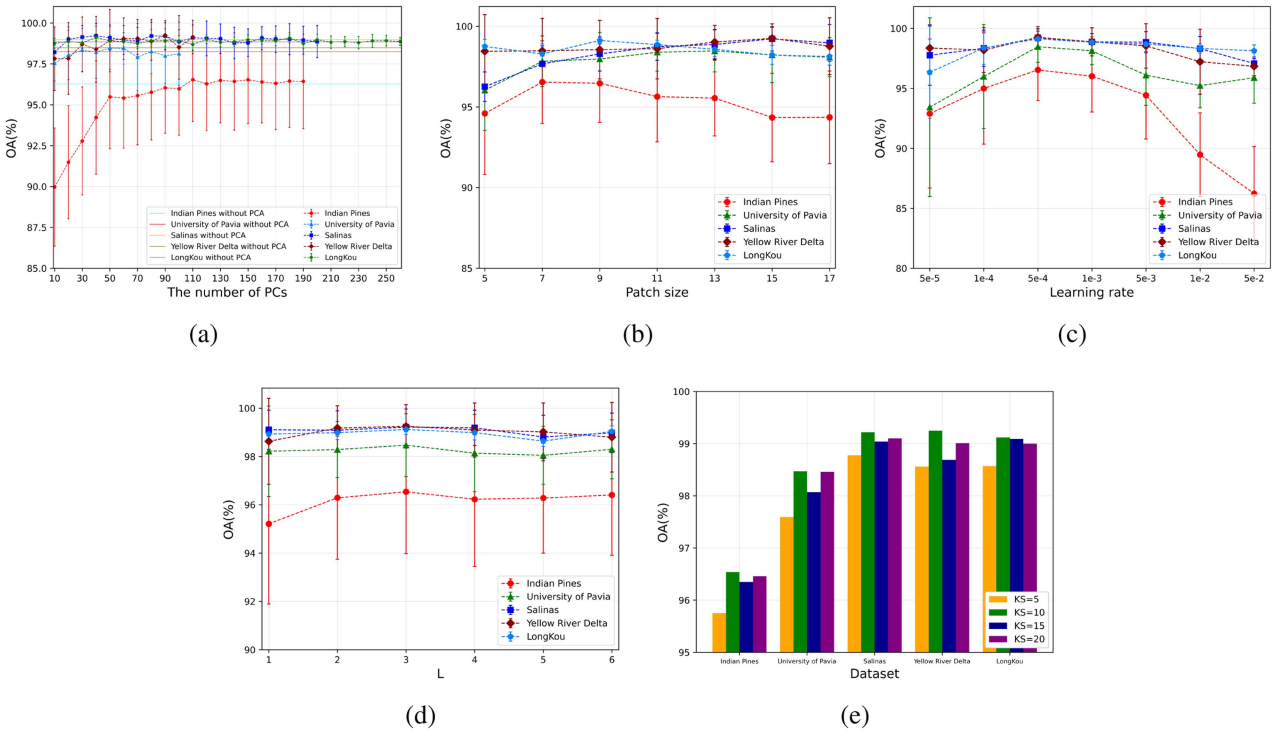


Fig. 11. Evolution of OA as a function of (a) number of PCs, (b) patch size, (c) learning rate, (d) transformer depth L, and (e) output dimension KS.

are also two FSL methods, including S3Net [75] and multi-stage relation network with dual-metric (DM-MRN) [74]. Five transformer-based methods include SpectralFormer (SF) [48], hyperspectral image transformer (HiT) [49], spectral–spatial tokenization transformer (SSFTT) [53], group-aware hierarchical transformer (GAHT) [54], and local transformer with spatial partition restore (SPRLT) [34]. The parameter settings of these methods are the same as those presented in the corresponding articles.

4) Running Platforms and Metrics: All experiments were run with 12th generation Intel (R) Core (TM) i7-12700, NVIDIA GeForce RTX 4070 (12 GB), and 64 GB RAM, based on PyTorch 2.1.1 and CUDA 12.1. For the purpose of improving the stability of experimental results, we present the average result of ten independent experiments. The evaluation metrics include three commonly used metrics, namely, overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ). Furthermore, to evaluate the complexity of the model, training time (s), test time (s), and the amount of parameters are used as evaluation indicators.

C. Parameter Sensitive Analysis

1) Dimensionality Reduction Using PCA: To find the optimum reduced-dimension of PCA, Fig. 11(a) shows the evolution of OA with the number of PCs. We can easily see that for Indian Pines and University of Pavia, the effect of PCA for dimensionality reduction is not significant, but their OA performance is the best when the number of PCs is set to 110 and 50, respectively. For the other three datasets, most of the performance after dimensionality reduction has been significantly improved compared to that without PCA dimensionality reduction. For the other three datasets, most of the performance after dimensionality reduction has been significantly improved compared to that without PCA dimensionality reduction. For Salinas, Yellow River Delta, and WHU-Hi-LongKou, the highest OA can be achieved when the number of PCs is set to 40, 90, and 40, respectively. Therefore, the above configurations are used for subsequent experiments.

2) Patch Size: To select the optimum patch size for each dataset, we set patch size in the range of [5, 7, 9, 11, 13, 15, 17]. As shown in Fig. 11(b), for Indian Pines, OA achieves the best performance when the patch size is set to 7×7. For University of Pavia, when patch is set to 13×13, the highest OA can be achieved. Both Salinas and Yellow River Delta achieve the highest OA when the patch is set to 17×17. For WHU-Hi-LongKou, when the patch is adjusted to 9×9, OA performs best. Therefore, according to the above experimental results, patches are set for these five datasets.

3) Learning Rate: To find the optimum learning rate for each dataset, we set it in the range of [0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05]. As shown in Fig. 11(c), for all five datasets, when learning rate is set to 0.0005, OA achieves the highest level. Therefore, learning rate is kept at 0.0005 in all other experiments.

4) Depth of Transformer Encoder: In order to evaluate the influence of transformer encoder at different depth on model performance, as shown in Fig. 11(d), depth L is set in the range [1, 2, 3, 4, 5, 6]. As can be seen, for all five datasets, OA performance is relatively low when L is larger or smaller, that is, when transformer encoder depth is deeper or shallower. Besides when L is set to 3, the proposed model achieves optimal performance in all datasets.

TABLE III
ABLATION ANALYSIS OF SCORE FUSION ATTENTION AND ST BLOCK

SpeSFA	SpaSFA	ST Block1	ST Block2	Indian Pines	University of Pavia	Salinas	Yellow River Delta	LongKou
×	×	✓	✓	91.00±5.47	94.04±3.21	97.26±1.00	97.70±2.54	96.69±3.36
✓	×	✓	✓	91.86±4.85	96.52±2.79	97.94±0.90	98.37±1.90	97.27±1.03
×	✓	✓	✓	91.23±5.18	94.79±4.54	97.46±1.00	98.06±2.24	97.91±0.80
✓	✓	×	×	88.32±7.08	91.83±6.78	96.41±2.88	94.52±4.58	97.31±1.38
✓	✓	✓	×	90.79±5.95	95.97±3.73	97.02±1.81	95.21±6.25	97.98±1.12
✓	✓	×	✓	91.69±6.41	93.70±5.36	96.71±2.05	96.28±5.03	97.94±1.07
✓	✓	✓	✓	**93.64±4.70**	**97.07±2.10**	**98.41±0.98**	**98.82±1.32**	**98.41±0.49**

SpeSFA: Spectral Score Fusion Attention; SpaSFA: Spatial Score Fusion Attention; ST Block: Spectral Transition Block. The bold values represent the best OAs.

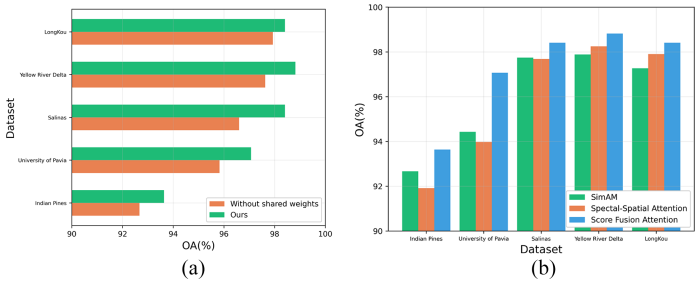


Fig. 12. Ablation analysis of score fusion attention. (a) Shared weights. (b) Attention mechanisms.

TABLE IV
ABLATION ANALYSIS OF ST BLOCK

Datasets	ECA [80]	SRM [81]	ST Block
Indian Pines	92.63±3.71	92.10±4.19	**93.64±4.70**
University of Pavia	95.66±3.92	95.18±3.24	**97.07±2.10**
Salinas	97.81±0.90	97.27±0.98	**98.41±0.98**
Yellow River Delta	97.97±2.28	98.27±2.29	**98.82±1.32**
LongKou	98.16±0.83	98.18±0.65	**98.41±0.49**

The bold values represent the best OAs.

5) Feature Dimension of SWEM: The feature dimension of SWEM represents the feature dimension size for spatial extraction, which is actually controlled by the number of output channels of the last convolution layer of spectral ResNet. In order to find the compressed feature dimension size that is optimum for the spatial extraction module, we set the output dimension of lower branch (KS) of spectral ResNet in the range [5, 10, 15, 20]. As shown in Fig. 11(e), We can clearly see that when KS is set to 10, the performance of the proposed model achieves optimal.

D. Ablation Study

1) Effectiveness of Score Fusion Attention (SFA): We removed the proposed SpeSFA and SpaSFA, respectively, to conduct experiments, and compared the accuracy achieved by the complete S3FAN to verify the function of SpeSFA and SpaSFA. It can be clearly seen from Table III that when only SpeSFA or SpaSFA is retained, the accuracy that the model can achieve will be better than eliminating both at the same time. After only adding SpeSFA, OA has the most obvious growth on University of Pavia, with an increase of 2.48% compared to the no-attention mechanism. By only adding SpaSFA, the OA of all datasets also increased. Among them, the most significant improvement was LongKou, which increased by 1.23%. Furthermore, when the proposed SpeSFA and SpaSFA are added at the same time, as shown in Table III, the performance achieved by the model is significantly improved compared to the no-attention mechanism, especially on Indian Pines and University of Pavia, which increased by 2.64% and 3.03%, respectively. In addition, to prove the effectiveness of shared weights in SFA in improving model performance, we removed the shared weights mechanism in SpeSFA and SpaSFA, and compared the accuracy obtained with the accuracy under the complete SFA mechanism. As shown in Fig. 12(a), SFA (Ours) achieves higher OA than SFA (without shared weights). Specifically, the classification accuracy of the model on the five datasets decreases by 0.48%–1.81%, when SFA removes shared weights. To further demonstrate the improvement of SFA on model performance, we replaced other advanced attention mechanisms, including SimMA [78], and spectral–spatial attention (SSA) module [79], to compare the accuracy. Fig. 12(b) shows the accuracy comparison of the model under three attention mechanisms. Obviously, the model using SFA has improved performance compared with the other two attention methods, especially on University of Pavia, which is improved by at least 2.64%.

2) Effectiveness of ST Block: In the proposed S3FAN, the ST blocks realize effective transition from spectral weighted extraction to subsequent tasks and achieve the spectral feature recalibration and the enhancement of the representation ability of features. This strategy of reweighting the extracted features has significantly improved model performance. To confirm the effectiveness of ST blocks, the ST blocks are all removed first, and then added successively. As shown in Table III, the embedding of ST blocks obviously improves the performance of OA, especially for Indian Pines, University of Pavia, and Yellow River Delta. In addition, compared with the complete S3FAN and the S3FAN without ST blocks, OAs are greatly improved on all datasets. As shown in Table III, for University of Pavia and Indian Pines, OAs have the most dramatic improvement, increasing by 5.32% and 5.24%, respectively. Besides, for Salinas, Yellow River Delta, and LongKou, the improvement of OAs are also significant, with OAs increasing by 2%, 4.3%, and 1.1%, respectively. To further demonstrate the improvement of ST block on model performance, we compared the two feature recalibration methods, ECA [80] and SRM [81]. As shown in Table IV, compared with the other two methods, the accuracy of the model using ST block is improved on the five datasets. Especially on Indian Pines, University of Pavia, and Salinas, the accuracy is 1.01%, 1.41%, and 0.6% higher than the suboptimal ECA method, respectively.

TABLE V
ABLATION ANALYSIS OF MODULES OF S3FAN

SWETM	SSIM	SWEM	Indian Pines	University of Pavia	Salinas	Yellow River Delta	LongKou
×	✓	×	47.18±11.78	40.28±8.99	82.76±8.71	56.28±24.63	62.06±16.87
×	×	✓	50.22±9.98	51.74±6.17	91.67±3.46	76.60±10.33	77.31±8.67
×	✓	✓	48.86±10.57	51.35±8.48	92.22±3.53	76.89±11.94	77.94±6.72
✓	×	×	81.90±7.54	86.82±5.85	93.51±4.37	88.19±4.39	88.46±9.93
✓	×	✓	84.02±9.38	95.59±3.43	97.21±1.51	97.80±2.14	97.83±0.56
✓	✓	×	91.08±4.82	94.60±3.94	98.04±0.89	94.07±6.10	97.83±0.64
✓	✓	✓	**93.64±4.70**	**97.07±2.10**	**98.41±0.98**	**98.82±1.32**	**98.41±0.49**

SWETM: Spectral Weighted Extraction and Transition Module; SSIM: Spectral Subspace Interactive Module; SWEM: Spatial Weighted Extraction Module. The bold values represent the best OAs.

TABLE VI
ABLATION STUDY OF WEIGHTED FUSION FOR FEATURE FUSION

Datasets	Evaluation	Conv2-D [82]	PAFM [83]	AF [84]	Weighted Fusion
Indian Pines	OA (%)	90.42±6.23	91.21±5.90	91.04±4.64	**93.64±4.70**
	AA (%)	86.50±8.33	86.69±6.42	87.27±5.48	**92.06±5.89**
	$\kappa\times100$	89.11±7.07	89.99±6.67	89.83±5.24	**92.79±5.25**
	Parameters	208 101	208 098	208 101	168 321
University of Pavia	OA (%)	96.17±3.09	91.57±3.81	92.65±5.43	**97.07±2.10**
	AA (%)	94.85±3.43	87.70±4.73	89.29±6.78	**95.82±2.19**
	$\kappa\times100$	94.97±4.01	88.98±4.95	90.45±6.72	**96.13±2.77**
	Parameters	92 241	92 238	92 241	68 751
Salinas	OA (%)	97.29±0.99	96.35±2.76	97.45±1.15	**98.41±0.98**
	AA (%)	98.38±0.64	96.91±4.41	97.96±0.93	**98.71±0.87**
	$\kappa\times100$	96.98±1.10	95.95±3.06	97.16±1.28	**98.23±1.09**
	Parameters	80 631	80 628	80 631	61 490
Yellow River Delta	OA (%)	98.16±2.03	95.70±4.97	95.84±6.34	**98.82±1.32**
	AA (%)	95.78±3.85	88.62±7.62	91.07±9.80	**96.59±3.43**
	$\kappa\times100$	97.57±2.66	94.34±6.44	94.54±8.20	**98.43±1.75**
	Parameters	160 681	160 678	160 681	135 755
WHU-Hi-LongKou	OA (%)	98.40±0.48	95.87±2.82	98.07±0.83	**98.41±0.49**
	AA (%)	94.58±1.50	89.68±4.72	94.43±2.36	**94.80±1.40**
	$\kappa\times100$	97.91±0.63	94.64±3.61	97.48±1.08	**97.93±0.63**
	Parameters	80 631	80 628	80 631	52 274

The bold values indicate the best-performing results.

3) Effectiveness of Modules of S3FAN: In the proposed S3FAN, a total of three modules can be divided into SWETM, SSIM, and SWEM. To verify the effectiveness of these three modules, SWETM, SSIM, and SWEM are removed and added in order. Since the discriminative nature of spectral information is the basis of pixel-level HSI classification, the removal of SWETM will inevitably lead to a decrease in feature discrimination ability. As shown in Table V, after eliminating SWETM, even if SSIM and SWETM are retained at the same time, the accuracy results are not good. In the case of only retaining SWETM, although the accuracy is improved compared to the former, it is not outstanding enough. Next, SWETM-SSIM and SWETM-SWEM are compared to verify the role of fusion of spectral interaction features and spectral–spatial features. For Indian Pines and Salinas, the OA performance of SWETM-SSIM is superior, while for the other datasets, SWETM-SWEM achieves better OA performance. However, after weighted fusion of the features extracted by these two feature extraction strategies, for all datasets, OA has been significantly improved. Specifically, compared to SWETM-SWEM and SWETM-SSIM, OAs increase 9.62% and 2.56%, respectively, for Indian Pines, 1.48% and 2.47%, respectively, for University of Pavia, 1.2% and 0.37%, respectively, for Salians, 1.02% and 4.75%, respectively, for Yellow River Delta, 0.58% and 0.58%, respectively, for LongKou. To achieve the complementarity of spectral and spatial features captured by SWETM-SSIM and SWETM-SWEM, weighted fusion is used in feature fusion stage. To validate the superiority of weighted fusion, we substitute weighted fusion with other feature fusion methods, including Conv2-D [82], pooled activation fusion module (PAFM) [83], and AF Operation (AF) [84], and then conduct comparative experiments. As is reported in Table VI, weighted fusion achieves the highest accuracy among all methods, with fewer model parameters than Conv2-D, PAFM, and AF.

TABLE VII
ABLATION ANALYSIS OF PCA FOR DATA DIMENSIONALITY REDUCTION

Methods	Indian Pines	University of Pavia	Salinas
ICA	93.43±4.34	95.74±2.83	97.37±1.26
SPCA	86.76±4.73	96.19±3.30	97.75±0.75
LDA	92.39±2.54	95.62±2.71	98.05±0.57
SVD	93.56±3.09	96.33±3.27	97.90±0.85
PCA	**93.64±4.70**	**97.07±2.10**	**98.41±0.98**

The bold values represent the best OAs.

4) PCA for Data Dimensionality Reduction: The dimensionality disaster caused by the spectral information redundancy of HSIs is not conducive to subsequent information extraction and classification tasks. In this experiment, we compare the OAs of different dimensionality reduction methods on Indian Pines, University of Pavia, and Salinas, including independent component analysis (ICA), sparse principal component analysis (SPCA), linear discriminant analysis (LDA), singular value decomposition (SVD), and the widely used PCA. As shown in Table VII, when PCA is employed for dimensionality reduction,

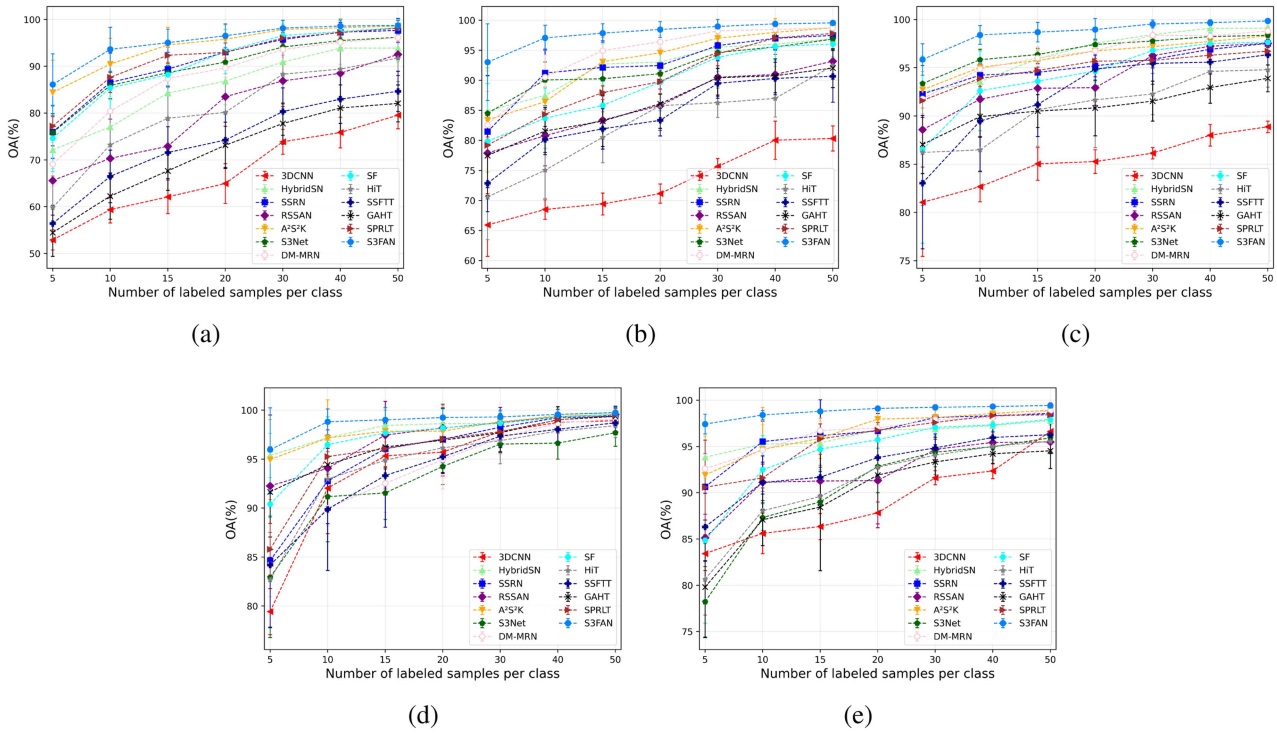


Fig. 13. Evolution of OA as a function of training samples per class. (a) Indian Pines. (b) University of Pavia. (c) Salinas. (d) Yellow River Delta. (e) WHU-Hi-LongKou.

TABLE VIII
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS IN INDIAN PINES (10 LABELED SAMPLES PER CLASS FOR TRAINING)

Class	CNN-Based							Transformer-Based					**S3FAN (Ours)**
	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	
1	31.96±3.90	43.21±11.74	90.71±18.58	40.31±9.74	88.5±18.27	**100.00±0.00**	99.17±1.27	86.55±24.80	58.46±13.87	58.75±18.46	36.34±17.00	97.91±5.42	96.81±5.07
2	49.65±7.89	70.39±5.40	85.23±8.58	63.94±15.08	90.93±4.37	80.65±7.84	75.36±6.87	72.28±9.28	56.41±9.70	55.24±7.96	53.82±9.07	87.28±0.78	**91.94±6.54**
3	30.40±2.43	66.81±9.26	67.44±19.75	59.31±9.31	88.50±8.23	81.99±8.93	74.94±10.19	72.60±7.87	45.38±6.06	43.75±5.69	40.04±8.09	87.81±9.34	**89.20±12.57**
4	35.38±6.78	62.05±8.57	80.39±17.26	48.52±18.38	86.11±17.12	**96.70±4.46**	96.52±2.59	64.22±15.33	30.55±7.09	45.92±10.42	43.98±7.88	70.89±6.15	86.51±9.13
5	56.38±7.59	87.57±3.64	94.07±5.10	75.11±15.07	96.20±2.75	87.91±4.76	85.77±7.49	94.47±5.04	66.92±13.19	69.79±7.60	75.82±13.57	99.10±0.41	**99.33±1.06**
6	95.57±2.19	82.71±5.78	96.50±2.97	82.71±11.35	92.47±4.25	96.46±1.69	96.04±2.83	93.85±6.41	86.29±11.36	88.90±5.91	85.42±6.74	90.73±3.17	**96.53±2.67**
7	17.58±4.20	43.51±13.05	71.26±21.30	55.02±11.37	88.92±24.91	**100.00±0.00**	**100.00±0.00**	56.58±22.32	50.17±10.84	54.18±18.77	40.37±17.02	80.48±5.61	86.49±7.86
8	98.68±0.82	91.37±2.94	99.69±0.81	89.31±13.76	96.80±7.32	99.64±1.02	99.87±0.22	96.66±4.15	94.27±3.41	98.96±1.67	98.34±1.11	98.16±0.76	**99.98±0.06**
9	27.35±6.26	43.50±16.08	81.68±25.84	25.15±10.49	62.23±31.40	**100.00±0.00**	**100.00±0.00**	69.47±27.33	16.71±6.07	31.18±10.99	21.97±13.65	35.37±2.85	72.29±34.12
10	46.37±8.58	72.40±7.48	87.17±5.67	57.96±9.43	77.20±7.65	82.39±5.05	81.93±5.00	70.90±6.83	45.87±9.66	57.41±6.11	47.59±5.35	73.39±2.26	**88.84±6.96**
11	69.42±4.45	87.28±6.32	88.03±7.67	79.14±5.89	93.80±4.07	75.91±11.07	63.68±11.32	87.12±6.05	75.97±6.07	74.63±6.62	71.47±7.16	91.35±0.97	**96.08±2.43**
12	36.12±4.34	69.92±8.02	**94.57±7.73**	58.55±10.93	92.67±15.18	79.04±7.08	79.66±8.65	69.17±5.27	46.00±8.34	39.50±5.39	35.22±4.31	86.45±9.25	94.30±5.19
13	88.82±14.39	85.33±17.29	95.35±4.32	89.38±9.15	95.61±12.35	99.08±1.21	**99.79±0.34**	88.47±7.76	71.77±12.77	89.22±13.17	75.96±8.08	81.34±3.33	95.68±6.26
14	86.71±1.62	93.87±1.80	97.85±1.68	90.86±4.40	96.40±2.04	**98.85±1.61**	92.19±5.12	96.98±2.83	89.96±4.92	92.18±3.81	91.05±4.40	98.20±0.54	97.86±1.85
15	56.30±5.93	57.91±5.86	82.73±6.33	67.28±7.65	91.74±5.76	**97.13±5.21**	90.77±7.65	80.08±5.55	59.39±8.74	66.57±8.21	64.04±8.35	78.07±2.89	89.68±7.47
16	92.70±4.44	48.97±7.57	90.55±4.88	80.04±14.29	93.37±3.46	**98.31±3.54**	99.40±0.81	80.01±5.32	45.73±16.26	73.45±24.78	53.73±17.74	71.22±5.43	91.37±4.87
OA (%)	59.34±2.82	77.02±5.90	86.55±6.31	70.33±8.39	90.46±5.56	85.89±2.81	80.44±3.08	81.57±6.68	62.33±6.96	66.49±5.61	62.27±4.94	87.63±3.23	**93.64±4.70**
AA (%)	57.46±3.18	69.18±6.83	87.70±6.80	66.41±8.45	89.47±8.98	**92.13±1.21**	89.69±1.22	79.96±9.55	58.74±6.64	64.98±8.64	58.45±5.93	82.98±3.63	92.06±5.89
$\kappa\times100$	54.35±3.16	74.17±6.52	84.75±7.06	66.67±9.21	89.22±6.16	84.06±3.12	78.02±3.35	79.20±7.48	57.92±7.61	62.28±6.34	57.64±5.48	85.98±3.56	**92.79±5.25**

The bold values indicate the best-performing results.

the model achieves the best performance on the three datasets, and the OA accuracy is 0.08%, 0.74%, and 0.36% higher than the suboptimal method, respectively.

E. Generalization Performance

To evaluate the generalization performance of the proposed S3FAN, the classification results of different advanced methods under few training samples in the range of [5, 10, 15, 20, 30, 40, 50] from five different datasets are compared in this experiment. Details are shown in Fig. 13, S3FAN achieved the highest accuracy under all conditions of training samples on all datasets, which shows that our proposed method has better generalization performance than other methods. The proposed S3FAN not only has the highest OA than other methods when the number of labeled samples per class for training is 5, but as the number of samples increases, its performance still exceeds other advanced methods.

F. Classification Results

In order to verify the superiority of the proposed S3FAN under few labeled samples, we randomly select 10 labeled samples per class for training and the rest are used for testing to compare the classification performance with other methods. The classification accuracy result of different methods for Indian Pines, University of Pavia, Salinas, Yellow River Delta, and WHU-Hi-LongKou are reported in Tables VIII–XII. To more

TABLE IX
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS IN UNIVERSITY OF PAVIA (10 LABELED SAMPLES PER CLASS FOR TRAINING)

Class	CNN-Based							Transformer-Based					S3FAN (Ours)
	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	
1	91.25±1.09	87.02±4.12	95.36±7.47	91.74±4.24	87.74±7.79	92.23±6.16	91.34±4.91	87.44±4.37	88.13±4.14	91.28±2.52	91.66±3.56	90.52±1.69	**97.44±1.85**
2	87.44±2.34	97.67±1.37	97.77±1.01	93.56±3.19	98.08±0.34	86.46±7.31	85.92±8.38	95.72±1.90	90.32±2.21	94.55±1.95	90.61±3.12	93.96±1.72	**99.43±1.17**
3	47.16±2.42	77.00±8.95	73.30±9.10	53.62±13.06	54.42±6.21	88.65±9.26	87.08±5.25	62.39±6.25	47.59±6.98	54.83±6.90	51.11±6.77	48.10±0.68	**95.73±2.59**
4	58.32±4.84	73.08±10.14	84.10±6.98	92.98±8.18	92.97±3.80	94.83±1.93	**97.30±1.69**	91.44±4.70	82.12±6.26	85.00±5.47	81.21±12.93	86.08±1.72	91.42±2.02
5	96.49±0.44	96.49±4.53	99.87±0.22	96.12±4.93	99.62±0.70	99.63±0.44	**99.91±0.17**	96.50±2.58	92.17±4.56	93.18±4.32	97.94±2.69	99.37±0.14	98.64±1.16
6	36.51±3.15	87.96±6.78	85.72±8.68	54.62±10.75	82.64±6.16	91.07±6.60	**97.81±3.50**	74.16±5.64	55.01±13.32	54.60±2.83	48.35±6.30	88.62±2.94	95.25±5.50
7	41.00±2.17	80.65±13.30	86.68±8.92	60.80±10.01	71.92±5.55	99.66±0.92	99.82±0.33	59.80±12.27	43.01±9.32	49.84±7.56	49.61±7.19	79.50±2.53	**99.97±0.07**
8	68.41±3.76	73.72±5.58	82.63±12.69	81.08±5.85	65.12±1.02	90.97±7.46	89.97±6.52	67.92±5.43	73.53±7.80	76.82±1.86	76.92±3.60	58.34±0.72	**93.68±5.80**
9	99.48±1.09	71.72±12.71	98.55±2.45	92.11±6.60	82.45±5.73	97.20±3.54	**98.77±1.22**	67.03±19.22	59.81±12.52	97.16±1.39	89.19±8.55	76.57±3.77	90.79±6.07
OA (%)	68.49±1.64	87.48±2.37	91.20±4.02	80.78±5.00	86.43±2.47	90.05±3.00	90.53±3.71	83.69±5.01	75.05±4.69	80.17±2.16	76.37±2.06	84.35±0.87	**97.07±2.10**
AA (%)	69.56±1.60	82.81±3.00	89.33±4.11	79.63±4.76	81.66±2.18	93.41±1.22	94.21±1.23	78.04±6.19	70.19±3.47	77.47±2.64	75.18±2.50	80.12±0.89	**95.82±2.19**
$\kappa\times100$	60.66±2.06	83.68±3.03	88.54±5.11	75.29±6.08	82.36±3.10	87.17±3.17	87.85±4.49	78.83±6.43	67.96±5.39	74.55±2.81	69.71±2.36	79.49±1.29	**96.13±2.77**

The bold values indicate the best-performing results.

TABLE X
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS IN SALINAS (10 LABELED SAMPLES PER CLASS FOR TRAINING)

Class	CNN-Based							Transformer-Based					S3FAN (Ours)
	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	
1	86.58±5.20	98.77±1.71	99.92±0.25	99.19±1.60	99.77±0.68	99.76±0.52	99.65±0.69	99.10±2.31	86.50±9.16	94.02±4.09	96.03±3.44	**100.00±0.00**	**100.00±0.00**
2	98.82±0.29	99.36±0.88	99.75±0.29	98.44±1.68	99.14±1.15	98.49±2.39	**99.99±0.01**	98.10±3.32	95.31±7.67	99.23±0.26	99.88±0.09	99.90±0.13	99.97±0.08
3	89.14±3.01	97.72±2.07	97.63±1.23	97.03±1.84	99.03±1.50	**100.00±0.00**	**100.00±0.00**	99.50±1.13	96.23±2.62	95.26±3.67	96.10±1.93	**100.00±0.00**	**100.00±0.00**
4	99.15±1.53	78.75±7.31	97.07±2.12	98.29±2.44	96.73±2.97	**99.67±0.65**	98.48±2.52	95.89±2.65	92.77±4.49	98.73±1.93	99.49±0.86	89.79±2.32	92.30±4.81
5	91.39±3.30	96.04±3.99	99.26±0.60	95.80±9.09	98.69±0.99	96.41±3.62	97.44±1.82	96.58±3.91	97.99±2.24	98.20±2.56	98.65±1.65	99.33±0.05	**99.79±0.28**
6	**100.00±0.00**	97.61±4.99	99.61±0.95	99.88±0.17	99.77±0.63	98.89±1.01	99.85±0.35	99.33±1.61	97.34±1.00	99.68±0.23	99.98±0.01	98.49±0.91	99.84±0.23
7	94.34±1.70	97.28±3.12	99.15±1.20	98.11±1.57	**99.99±0.30**	99.97±0.05	99.86±0.20	97.04±7.66	95.53±2.65	96.71±1.75	97.63±1.18	99.98±0.01	99.04±1.15
8	71.50±2.67	94.23±3.46	89.91±1.78	87.97±9.86	90.90±2.15	88.53±4.26	86.23±3.97	85.98±6.26	84.72±4.81	80.18±4.88	83.41±2.62	92.60±1.25	**98.55±1.25**
9	95.91±0.80	99.65±0.35	98.88±1.35	99.33±0.59	99.46±0.30	99.94±0.08	**99.99±0.02**	98.69±2.19	97.24±2.46	98.07±1.47	98.92±0.92	99.64±0.15	99.98±0.03
10	85.52±2.28	97.76±2.77	92.56±2.74	95.76±2.73	98.49±0.94	95.87±2.60	95.96±2.93	94.57±3.16	86.23±4.77	86.49±4.93	89.42±5.17	91.20±0.77	**98.88±1.35**
11	58.29±5.25	91.64±8.13	97.80±1.98	92.13±3.08	98.27±3.89	99.59±0.40	**99.98±0.04**	97.65±3.38	83.81±9.32	69.47±7.29	80.85±9.16	98.43±3.59	98.74±2.18
12	94.84±2.20	98.25±2.31	99.08±1.13	98.71±1.06	98.86±3.09	98.27±1.98	98.86±0.64	95.21±10.98	97.16±1.33	97.33±1.48	97.31±2.93	**99.95±0.15**	99.92±0.18
13	90.74±3.93	92.36±3.72	99.29±1.55	88.62±6.51	97.92±3.98	98.09±2.16	99.09±1.14	94.88±9.21	90.04±5.14	92.40±7.30	95.04±6.38	99.77±0.39	**99.88±0.18**
14	92.53±1.54	88.74±16.71	95.60±4.37	95.22±3.46	97.73±2.92	97.52±4.38	97.34±3.37	93.29±9.94	94.57±3.00	97.65±2.01	91.29±2.10	97.89±3.14	**98.34±4.32**
15	54.32±2.05	82.64±7.24	81.65±6.03	71.84±9.06	82.67±6.50	92.48±7.07	87.97±4.43	79.19±7.20	62.54±6.34	66.09±5.47	70.04±5.20	77.17±3.99	**94.13±4.46**
16	90.40±3.05	96.70±5.69	99.71±0.39	99.20±1.28	99.48±1.43	98.81±1.88	99.39±1.15	98.47±3.07	91.13±3.30	93.31±6.21	91.58±2.68	**100.00±0.00**	**100.00±0.00**
OA (%)	82.68±1.56	93.55±1.78	94.23±1.18	91.73±3.86	95.04±1.93	95.84±1.07	94.87±1.11	92.62±4.14	86.49±3.72	87.95±2.58	89.93±2.13	93.89±1.17	**98.41±0.98**
AA (%)	87.09±1.83	94.22±1.64	96.68±0.73	94.72±2.62	97.31±1.79	97.70±0.85	97.45±0.52	95.22±4.55	90.57±2.21	91.43±2.65	92.85±2.19	96.51±0.89	**98.71±0.87**
$\kappa\times100$	80.74±1.72	92.84±1.97	93.59±1.31	90.79±4.36	94.49±2.13	95.37±1.19	94.29±1.24	91.80±4.59	85.06±4.05	86.62±2.84	88.82±2.34	93.21±1.29	**98.23±1.09**

The bold values indicate the best-performing results.

TABLE XI
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS IN YELLOW RIVER DELTA (10 LABELED SAMPLES PER CLASS FOR TRAINING)

Class	CNN-Based							Transformer-Based					S3FAN (Ours)
	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	
1	76.32±4.96	92.12±8.12	91.37±5.54	90.93±6.10	95.58±1.17	85.97±6.53	79.90±9.30	97.47±2.77	90.41±3.63	78.79±2.92	90.39±8.09	91.46±3.95	**98.03±4.86**
2	98.34±1.72	95.79±5.37	88.22±8.04	97.96±1.53	97.12±4.87	95.87±2.24	93.50±6.29	99.11±1.98	88.32±9.92	97.25±4.35	98.43±3.69	**100.00±0.00**	99.20±0.52
3	90.68±9.46	**100.00±0.00**	99.71±0.46	94.64±6.66	99.33±2.01	99.91±0.26	96.96±4.98	95.55±13.35	84.24±9.03	81.96±6.81	98.03±4.17	98.55±3.09	**100.00±0.00**
4	95.80±1.51	95.56±5.85	96.11±2.96	96.40±5.02	96.82±4.78	98.32±3.47	94.62±5.41	97.82±3.92	87.39±9.74	87.24±8.14	97.31±3.82	99.42±0.72	**99.62±0.10**
5	94.25±5.74	99.12±2.65	93.38±4.45	97.33±3.31	97.87±1.74	97.76±3.52	98.15±2.82	95.68±2.98	92.16±5.76	95.65±4.46	99.30±1.61	99.03±0.48	**100.00±0.00**
6	79.15±6.48	91.21±3.05	77.45±13.12	86.72±4.38	90.13±4.53	91.25±3.29	86.58±7.16	86.65±7.92	81.03±7.19	76.74±7.10	86.44±5.64	87.47±0.56	**91.31±5.16**
7	**100.00±0.00**	**100.00±0.00**	99.01±1.20	99.49±1.54	99.87±0.40	**100.00±0.00**	99.43±1.72	99.27±1.35	83.12±8.24	99.36±1.91	99.67±0.60	**100.00±0.00**	99.61±1.17
8	97.73±0.87	98.47±2.24	98.26±2.02	98.20±2.15	98.91±1.52	98.42±1.81	94.81±3.36	98.89±1.85	95.08±5.23	95.33±4.18	**99.14±1.32**	96.50±0.51	99.04±1.30
9	99.36±0.13	99.95±0.08	99.95±0.16	**100.00±0.00**	99.69±0.19	99.95±0.16	99.63±0.58	98.49±3.04	98.48±1.55	99.07±2.21	99.31±0.70	98.33±0.69	99.97±0.07
10	99.65±0.26	**99.97±0.03**	99.77±0.21	99.92±0.14	99.94±0.03	86.09±9.18	85.91±4.62	99.78±0.26	99.43±0.67	99.44±0.97	99.78±0.27	99.94±0.17	99.97±0.05
11	95.28±7.64	90.25±20.27	76.70±21.83	96.53±7.28	94.67±16.00	**100.00±0.00**	**100.00±0.00**	89.95±20.46	66.66±22.12	66.83±28.87	91.12±18.90	**100.00±0.00**	92.80±21.60
12	91.55±9.45	59.94±29.75	22.04±10.06	68.77±15.90	61.26±32.08	90.58±6.69	92.81±6.44	36.05±11.70	**91.91±10.78**	87.10±14.89	79.73±13.59	25.54±5.47	77.63±19.90
13	93.64±2.11	99.18±1.64	99.27±2.10	95.21±1.55	99.21±1.56	95.68±4.78	91.94±10.83	97.94±1.67	91.85±4.21	92.91±1.33	98.80±0.95	99.03±0.25	**99.28±0.87**
14	72.94±13.67	86.21±13.28	92.25±4.70	72.46±27.43	**98.38±2.25**	93.08±4.90	92.53±7.72	89.64±4.61	79.46±21.88	77.64±9.38	72.79±21.05	82.91±0.63	95.26±4.65
15	73.70±9.15	99.44±1.68	95.95±4.67	91.07±14.02	96.04±11.24	**100.00±0.00**	**100.00±0.00**	93.26±9.47	81.98±12.20	56.35±6.10	90.45±16.95	99.80±0.39	96.79±3.63
16	87.11±5.77	**100.00±0.00**	81.98±19.15	73.85±21.31	89.65±13.80	**100.00±0.00**	98.25±2.22	88.13±8.52	80.71±11.36	88.38±8.24	94.52±2.86	94.29±9.73	96.51±10.46
17	65.02±29.34	91.52±7.24	93.72±4.51	95.55±5.79	97.49±6.52	97.52±2.90	95.25±4.49	97.81±6.56	91.66±6.80	63.96±25.76	61.35±17.60	**99.22±2.34**	97.73±2.28
18	94.96±4.17	98.95±0.69	98.83±1.34	95.21±7.60	98.90±1.44	95.98±3.19	94.77±4.61	95.96±1.45	93.23±3.61	92.22±6.32	95.16±3.90	89.08±3.96	**99.05±1.57**
19	77.56±12.30	77.47±4.25	86.83±14.59	77.77±10.18	93.67±7.90	98.61±2.15	**99.48±1.19**	85.50±4.14	77.07±8.77	63.98±8.63	82.89±11.08	83.77±2.38	85.39±4.55
20	88.42±1.13	**98.37±2.66**	97.90±1.41	96.13±5.23	97.38±4.03	97.22±3.20	92.07±3.41	98.11±0.47	93.93±3.96	87.73±1.56	91.39±2.11	93.58±1.06	98.32±3.63
21	80.40±5.67	96.61±1.39	90.81±4.63	93.38±7.76	94.06±7.16	87.54±6.36	87.33±6.63	95.23±3.76	87.15±11.78	77.06±3.07	94.75±6.10	97.51±0.47	**98.30±2.39**
22	45.70±6.74	93.30±11.27	77.39±16.89	67.58±16.28	91.32±17.44	99.29±2.14	**99.67±0.66**	89.94±14.12	61.07±13.49	36.61±6.04	54.38±13.47	74.95±1.73	97.68±4.64
23	98.72±1.03	93.37±3.85	94.51±4.82	99.14±1.15	89.86±1.73	98.64±1.11	98.82±0.32	99.89±0.33	93.09±6.02	93.64±7.29	96.31±4.37	92.68±0.75	**100.00±0.00**
OA (%)	92.03±4.64	97.18±2.11	94.12±3.07	94.07±3.92	97.16±3.90	91.16±4.60	89.99±1.79	96.48±2.87	92.98±2.81	89.87±6.24	94.43±4.57	95.27±1.79	**98.82±1.32**
AA (%)	86.79±4.33	93.77±3.86	89.19±2.82	90.62±3.50	94.66±5.54	95.99±0.77	94.45±1.02	92.44±4.80	86.50±4.22	82.40±4.59	90.06±5.66	91.44±1.41	**96.59±3.43**
$\kappa\times100$	89.65±5.69	96.26±2.77	92.35±3.81	92.31±4.89	96.31±4.95	88.70±5.48	87.16±2.14	95.36±3.67	90.74±3.65	86.94±7.32	92.74±5.78	93.76±2.28	**98.43±1.75**

The bold values indicate the best-performing results.

TABLE XII
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS IN WHU-HI-LONGKOU (10 LABELED SAMPLES PER CLASS FOR TRAINING)

Class	CNN-Based							Transformer-Based					S3FAN (Ours)
	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	
1	97.54±1.11	98.25±0.89	95.52±3.17	96.84±1.06	94.58±5.76	95.73±2.82	99.36±0.44	96.53±1.31	97.53±1.78	96.69±0.56	93.68±2.54	90.84±0.94	**99.76±0.13**
2	66.05±10.93	84.58±5.65	79.87±8.31	70.78±10.64	80.30±6.53	**98.63±1.81**	97.05±1.91	71.40±8.87	45.53±9.88	75.25±7.60	71.88±9.89	70.89±1.28	93.33±5.72
3	48.55±18.34	97.18±1.78	94.11±8.27	41.69±14.19	94.01±17.49	97.83±4.69	**98.82±0.77**	65.96±15.72	33.74±10.66	41.77±12.32	28.11±10.18	64.80±5.80	96.09±2.40
4	97.65±1.36	98.92±0.53	98.89±1.40	97.93±0.94	99.05±1.03	89.36±4.75	89.11±4.54	99.22±0.89	98.49±1.29	97.87±0.64	98.20±0.62	99.64±0.31	**99.93±0.05**
5	47.06±16.96	73.98±13.14	69.14±15.96	44.95±9.22	76.91±19.85	**99.49±0.46**	98.47±1.20	62.02±14.16	40.49±10.99	53.99±13.72	39.06±7.17	71.80±3.01	91.27±6.70
6	94.84±5.54	95.31±3.67	98.76±1.63	96.97±2.47	99.39±0.25	94.81±3.90	97.54±2.36	94.00±4.69	97.38±2.20	92.85±4.28	81.13±12.62	98.17±0.42	**99.57±0.17**
7	**99.97±0.04**	99.78±0.21	99.77±0.63	99.94±0.05	98.87±2.20	79.01±6.78	97.57±0.68	97.57±3.05	98.99±0.54	99.88±0.15	99.84±0.28	98.18±0.84	99.88±0.18
8	85.85±4.83	83.68±5.55	92.69±4.02	87.43±5.59	89.16±11.49	76.47±5.20	83.22±8.42	79.02±6.52	78.50±7.17	84.43±3.45	84.15±7.46	80.31±1.38	**94.80±2.40**
9	43.69±14.54	60.97±10.35	80.71±12.00	64.93±10.65	75.68±7.23	77.50±10.48	**91.52±5.40**	75.28±11.44	59.20±7.46	60.16±7.23	60.44±10.08	54.63±3.38	78.56±8.62
OA (%)	89.47±4.51	95.26±2.10	95.54±2.22	91.14±1.98	94.69±4.55	87.32±1.59	94.62±1.42	92.50±3.70	88.05±2.11	91.11±2.88	87.08±2.78	91.62±1.10	**98.41±0.49**
AA (%)	75.69±6.40	88.07±3.89	89.94±3.14	77.94±3.53	89.77±5.30	89.87±1.14	94.74±1.01	82.33±5.77	72.20±3.28	78.10±4.00	72.94±3.25	81.03±1.64	**94.80±1.40**
$\kappa\times100$	86.55±5.54	93.86±2.66	94.21±2.85	88.60±2.48	93.15±5.75	83.81±1.96	93.04±1.80	90.33±4.72	84.70±2.61	88.59±3.61	83.50±3.42	89.21±1.41	**97.93±0.63**

The bold values indicate the best-performing results.

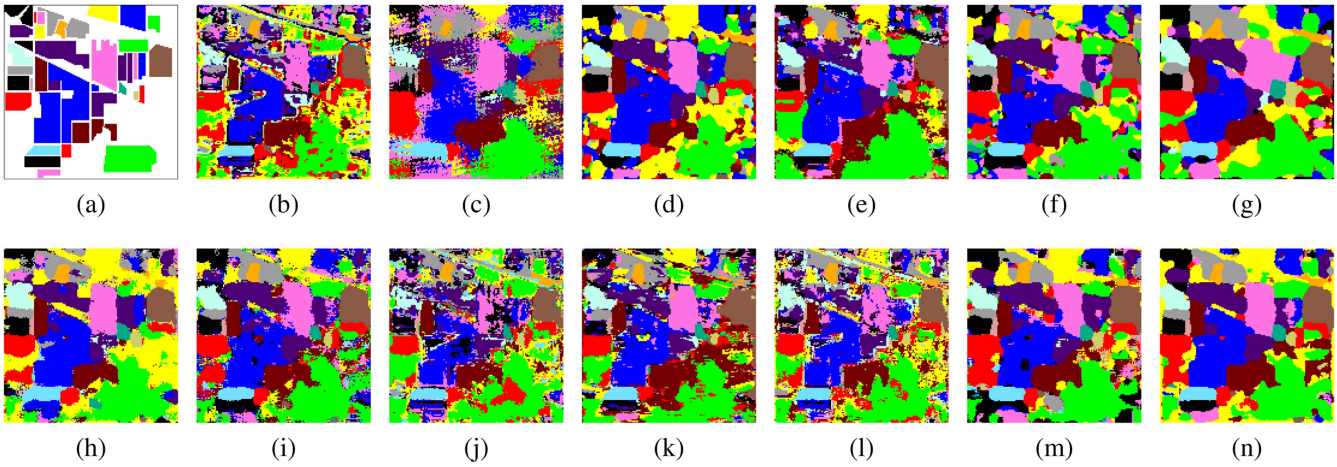


Fig. 14. Classification maps obtained by different methods in Indian Pines (10 labeled samples per class for training). (a) Ground truth, (b) 3-DCNN (59.34%), (c) HybridSN (77.02%), (d) SSRN (86.55%), (e) RSSAN (70.33%), (f) A2S2K (90.46%), (g) S3Net (85.89%), (h) DM-MRN (80.44%), (i) SF (81.57%), (j) HiT (62.33%), (k) SSFTT (66.49%), (l) GAHT (62.27%), (m) SPRLT (87.63%), and (n) S3FAN (93.64%).

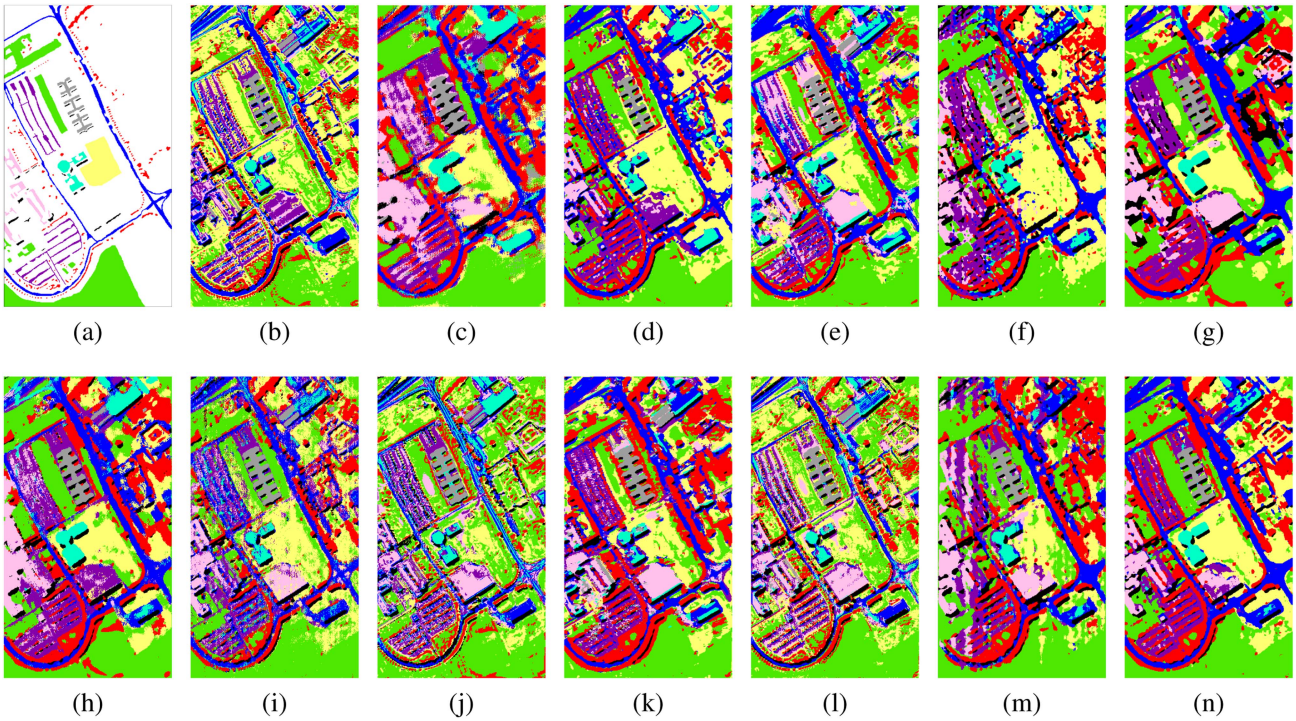


Fig. 15. Classification maps obtained by different methods in University of Pavia (10 labeled samples per class for training). (a) Ground truth, (b) 3-DCNN (68.49%), (c) HybridSN (87.48%), (d) SSRN (91.20%), (e) RSSAN (80.78%), (f) A2S2K (86.43%), (g) S3Net (90.05%), (h) DM-MRN (90.53%), (i) SF (83.69%), (j) HiT (75.05%), (k) SSFTT (80.17%), (l) GAHT (76.37%), (m) SPRLT (84.35%), and (n) S3FAN (97.07%).

intuitively and visually compare and reflect the classification performance based on S3FAN and other methods, we display the ground truth of the five datasets and the classification results of S3FAN and other methods on the five datasets in Figs. 14–18.

1) Indian Pines: The classification accuracy results of Indian Pines are reported in Table VIII. It can be seen that the average OA and $\kappa \times 100$ achieved by S3FAN are significantly higher than other methods. Specifically, OA achieves 93.64%, which is better than other methods by 34.3% to 3.18%. And kappa achieved 92.79%, which was an improvement of 38.44% to 3.57% compared with other methods. It is noted that S3Net performs best in AA, considering that S3Net uses a differentiated sample pairing strategy as a FSL method to provide the model with more learnable features, thereby enabling the model to perform better on categories with fewer learnable samples, such as *Grass-pasture-moved*, *Oats,* and *Stone-streel-towers*. DM-MRN also performs very well on categories with fewer total samples, because the sample recombination strategy increases

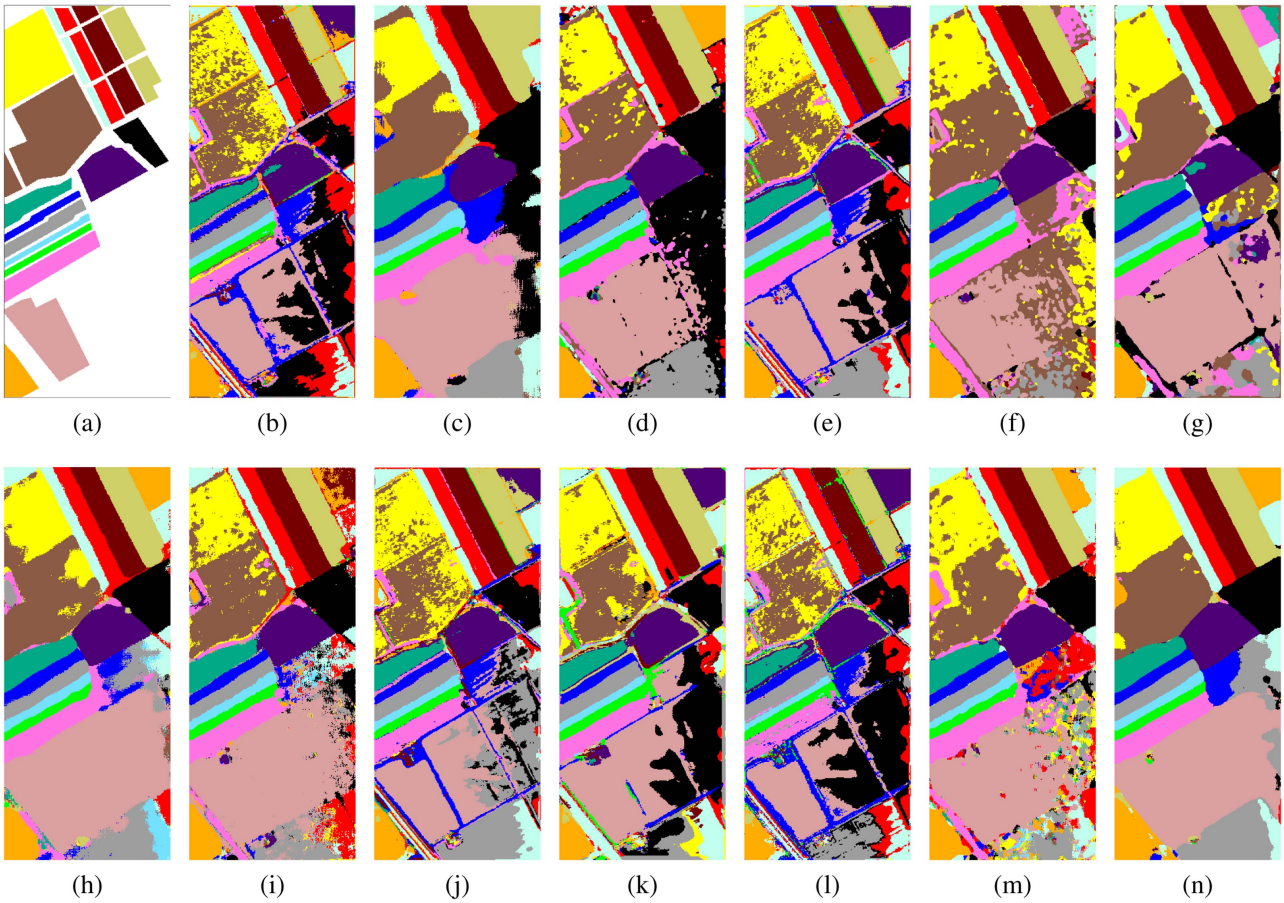


Fig. 16. Classification maps obtained by different methods in Salinas (10 labeled samples per class for training). (a) Ground truth, (b) 3-DCNN (82.68%), (c) HybridSN (93.55%), (d) SSRN (94.23%), (e) RSSAN (91.73%), (f) A2S2K (95.04%), (g) S3Net (95.84%), (h) DM-MRN (94.85%), (i) SF (92.62%), (j) HiT (86.49%), (k) SSFTT (87.95%), (l) GAHT (89.93%), (m) SPRLT (93.89%), and (n) S3FAN (98.41%).

the variety of classification tasks in the training period. In addition, S3FAN performs best than other methods in seven out of 16 land cover categories. Especially for *Corn notill*, *Corn mintill,* and *Soybeans mintill* categories, most other methods perform mediocrely in these three categories, whereas S3FAN has achieved significant improvements. The corresponding classification maps generated by different models are shown in Fig. 14. It can be clearly seen that the classification map of S3FAN is smoother and less noisy than other models. Compared with the most competitive S3Net, for the four categories of *Corn notill*, *Corn mintill*, *Soybeans mintill,* and *Bldg grass tree drivers*, which have a large number of total samples, the classification processing of S3FAN is obviously better. Besides, S3FAN can classify samples in unlabeled areas smoothly and accurately.

2) University of Pavia: The classification accuracy results of University of Pavia are reported in Table IX. Obviously, S3FAN achieved the best classification performance under the three accuracy indicators of OA, AA, and $\kappa \times 100$, with values of 97.07 ± 2.10, 95.82 ± 2.19, and 96.13 ± 2.77, respectively. Specifically, these three indicators outperform the suboptimal performances by 5.87%, 1.61%, and 7.59% respectively. At the same time, S3FAN achieves the highest classification accuracy in five of nine land cover categories. Especially for *Self-Blocking Bricks*, in which S3FAN has achieved a more significant accuracy improvement compared to other models. These performances exceed other advanced models by a huge margin. The corresponding classification maps produced by different models are shown in Fig. 15. The classification map generated by S3FAN exhibit better homogeneity within areas representing the same land cover categories. For *Asphalt*, *Bare soil,* and *Meadows*, three categories with large total sample numbers, the classification results of S3FAN are significantly smoother than other models. Compared with the two FSL methods, S3Net and DM-MRN, the cases of misclassifying *Meadows* into *Trees* and *Bare Soil* are significantly less. In addition, for the classification of samples in unlabeled areas, the result of S3FAN exhibits obviously fewer fragmented regions.

3) Salinas: The classification accuracy results of Salinas are reported in Table X, and S3FAN achieves the best classification performance across three indicators OA, AA, and $\kappa \times 100$, reaching values of 98.41 ± 0.98, 98.71 ± 0.87, and 98.23 ± 1.09, respectively. Furthermore, even though other methods perform well on this dataset, the results of these indicators surpass

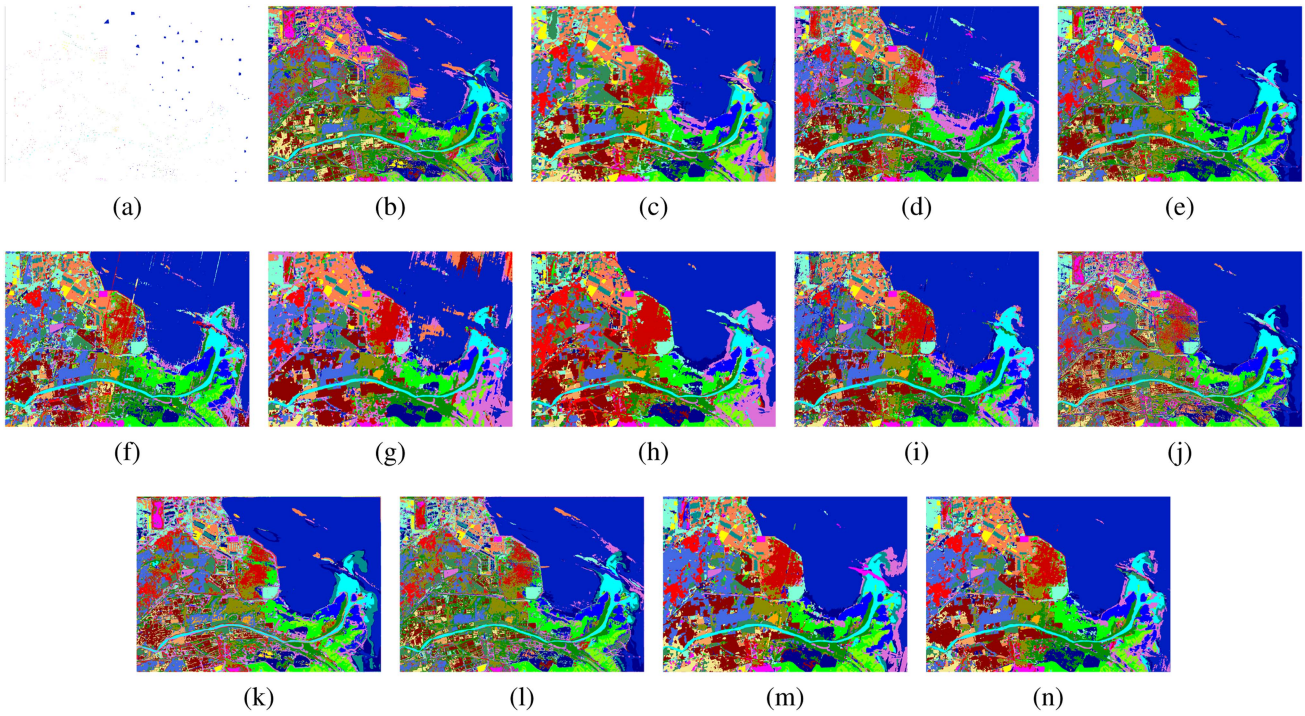


Fig. 17. Classification maps obtained by different methods in Yellow River Delta (10 labeled samples per class for training). (a) Ground truth, (b) 3-DCNN (92.03%), (c) HybridSN (97.18%), (d) SSRN (94.12%), (e) RSSAN (94.07%), (f) A2S2K (97.16%), (g) S3Net (91.16%), (h) DM-MRN (89.99%), (i) SF (96.48%), (j) HiT (92.98%), (k) SSFTT (89.87%), (l) GAHT (94.43%), (m) SPRLT (95.27%), and (n) S3FAN (98.82%).

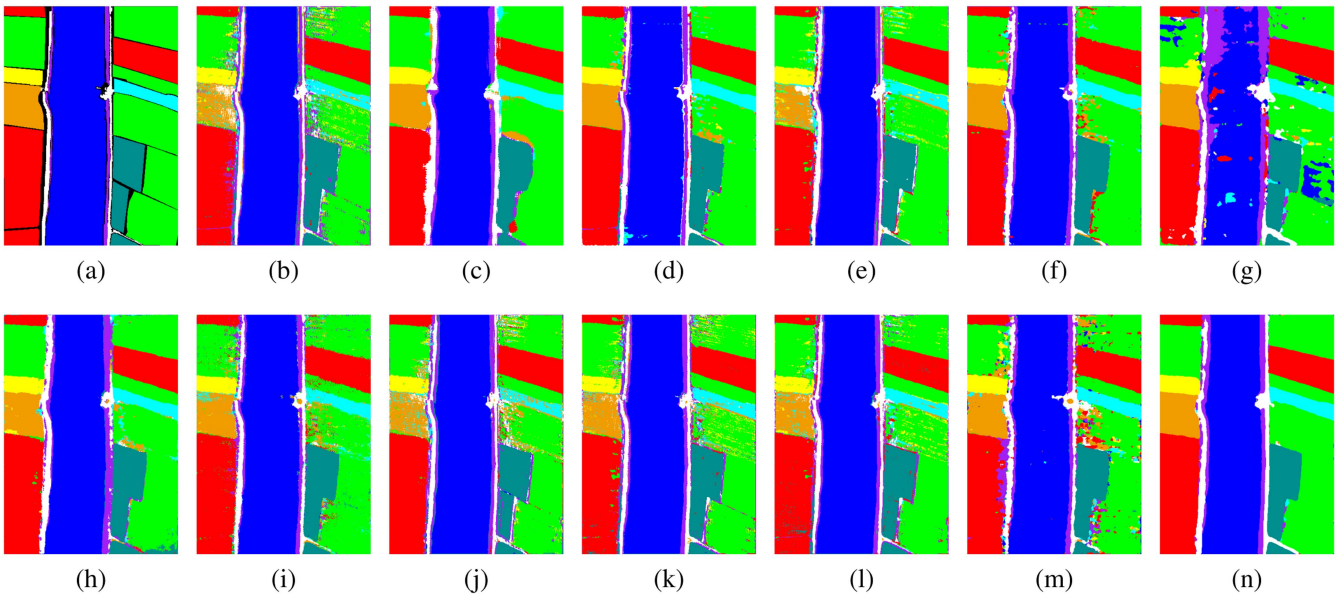


Fig. 18. Classification maps obtained by different methods in WHU-Hi-LongKou (10 labeled samples per class for training). (a) Ground truth, (b) 3-DCNN (89.47%), (c) HybridSN (95.26%), (d) SSRN (95.54%), (e) RSSAN (91.14%), (f) A2S2K (94.69%), (g) S3Net (87.32%), (h) DM-MRN (94.62%), (i) SF (92.50%), (j) HiT (88.05%), (k) SSFTT (91.11%), (l) GAHT (87.08%), (m) SPRLT (91.62%), and (n) S3FAN (98.41%).

other methods by significant margins, with improvements ranging from 15.73% to 2.57% for OA, 11.62% to 1.01% for AA, and 17.49% to 2.86% for $\kappa \times 100$. In addition, S3FAN achieves the highest averaged OA among the ten categories of all categories. Especially for *Grapes untrained*, other CNN-based methods or transformer-based methods perform mediocrely on this type of land cover category, and S3FAN has achieved significant improvement. The corresponding classification maps produced by different models are shown in Fig. 16. It is obvious that the classification map of S3FAN is far superior to other

methods. Specifically, in the classification map of S3FAN, the phenomenon of unclear boundaries of land cover categories is greatly improved compared with other methods, the transition of blocky land cover categories is smooth, especially for *Vineyard untrained*, *Fallow rough plow,* and *Fallow smooth*. At the same time, there are many fewer fragmented areas in the map.

4) Yellow River Delta: The classification accuracy results of Yellow River Delta are shown in Table XI. The classification performance of S3FAN is superior to other methods on three accuracy indicators. In detail, OA, AA, and $\kappa \times 100$ reach 98.82%, 96.59%, and 98.43%, respectively, which is a significant improvement compared with other methods. At the same time, the corresponding standard deviations of the three indicators also reach relatively low levels. In addition, among the 23 land cover categories, S3FAN achieves the highest classification accuracy in nine of them and the second highest classification accuracy in six of them, especially for *Reed* and *Cotton* categories, the classification performance of S3FAN is outstanding in these two categories. The corresponding classification maps generated by different methods are shown in Fig. 17. Compared with other classification methods, the classification map generated by S3FAN has significantly fewer fragments, less noise, and is smoother and clearer. Besides, for the classification of samples in unlabeled areas, S3FAN rarely misclassifies *Sea* into other categories, due to the powerful feature learning ability of S3FAN. While for the categories on land, the boundaries of land cover categories using S3FAN are clearer and smoother.

5) WHU-Hi-LongKou: The classification accuracy results of WHU-Hi-LongKou are shown in Table XII. For this data set, the performance of S3FAN is significantly improved compared to other methods. The three accuracy indicators exceed the results of other methods by a large margin. Specifically, the three indicators achieve 98.41%, 94.80%, and 97.9%, respectively, with improvements ranging from 11.33% to 2.87% for OA, 22.6% to 4.86% for AA, and 14.43% to 3.72% for $\kappa \times 100$. In addition, S3FAN has the highest accuracy in four of the nine land cover categories in WHU-Hi-LongKou dataset. Especially for the category *Roads and houses*, while other methods perform mediocrely in this category, S3FAN achieves significant improvements. The corresponding classification maps generated by different methods are shown in Fig. 18. Obviously, the classification map of S3FAN is superior to that of the other models. The classification map of S3FAN is visually smoother and clearer. Specifically, for *Broad leaf soybean*, there are many fewer dots or blocks of color fragments compared to other methods. The boundaries between *Roads and houses* and *Mixed weed* are also clearer and more separable. At the same time, it is extremely rare for S3FAN to mistakenly classify *Corn* and *Cotton* into *Broad leaf soybean*.

IV. Discussion

A. Complexity Analysis

To analyze the complexity of S3FAN, we compare the running time and parameters size of CNN-based methods and transformer-based methods under ten labeled samples in five datasets. As shown in Table XIII, HiT has the largest number of parameters compared to other methods, mainly because it uses a feature mapping method that combines convolution and MLP, and contains multiple stages. In terms of training time, S3FAN has obvious advantages over FSL methods that are specifically designed for limited samples, such as S3Net and DM-MRN. S3Net takes longer to train, which is attributed to its differentiated input strategy, which embeds patches with different sizes into the dual branches of the Siamese network, while DM-MRN takes longer to train due to the sample recombination strategy. In terms of parameter quantity, due to the rational use of dimension reduction, lightweight design of the model, and full utilization of discriminant features, the parameter quantity of S3FAN on the five datasets remains at a low level, which greatly avoids the risk of model overfitting. Therefore, S3FAN achieves shorter training time and testing time and a smaller number of parameters. In summary, S3FAN achieves excellent performance for few labeled samples HSI classification while avoiding consuming too much computing resources.

B. Feature Separability

We visualize the feature separability by using t-distributed stochastic neighbor embedding, to verify the feature separability of the proposed S3FAN. As shown in Fig. 19, we conduct experiments on Indian Pines, University of Pavia, Salinas, and Yellow River Delta datasets. For Indian Pines, samples are more clustered on the map produced by S3FAN (Ours), and can better separate *Soybean notill* and *Soybean clean*, *Corn notill* and *Soybean mintill*. For University of Pavia, the separability of *Meadows* and *Trees* is better than the other two variants. From Fig. 19(g)–(i), S3FAN (Ours) can clearly separate *Grape untrained* and *Vineyard untrained* compared to S3FAN (without SWEM) and S3FAN (without sematic extraction). For Yellow River Delta, S3FAN (without SWEM) is similar to S3FAN (Ours), both are better than S3FAN (without SSIM), but a sample representing *Fallow land* in S3FAN (without SWEM) is not separated from *Bare land*.

V. Conclusion

In this article, we propose a lightweight S3FAN for HSI classification with limited samples. It mainly includes three modules, SWETM, SSIM, and SWEM. SWETM weights the spectral features to extract and reweight them, and then the features enter SSIM and SWEM, respectively. The two sets of output features obtained by SSIM and SWEM are weighted, fused, and classified. The performance of S3FAN is further improved and its superiority is demonstrated on five datasets.

Specifically, SpeSFA in SWETM and SpaSFA in SWEM can calculate weighted fusion attention scores through rich discriminant features. In SFA, the weight sharing of the convolutional layer allows learning the similarity between feature vectors while keeping the number of parameters small, and the adaptive fusion of attention scores further enhances the representation ability of intermediate vectors or matrices for spectral sequences or spatial maps, thereby realizing dynamic attention that adapts

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

Datasets	Metrics	3DCNN [19]	HybridSN [24]	SSRN [27]	RSSAN [38]	A2S2K [39]	S3Net [65]	DM-MRN [74]	SF [45]	HiT [46]	SSFTT [49]	GAHT [50]	SPRLT [33]	**S3FAN (Ours)**
Indian Pines	Training time (s)	12.92	31.33	271.17	40.97	436.29	220.55	118.40	260.13	135.15	8.21	25.32	285.59	30.63
	Testing time (s)	3.47	9.35	129.43	23.18	191.11	43.98	28.71	29.99	49.91	3.60	10.27	125.13	10.39
	Parameters	45196	5122176	364168	265659	373184	155008	183062	342649	51231305	148488	1038160	846604	168321
University of Pavia	Training time (s)	4.77	55.09	57.35	32.32	123.38	197.98	102.93	57.55	58.75	6.60	15.71	155.43	22.17
	Testing time (s)	8.98	121.25	196.37	99.15	432.14	135.59	33.84	284.18	171.52	15.86	42.13	520.81	42.15
	Parameters	35039	5121273	216537	157704	221976	91517	181782	164393	43455602	148033	927113	833957	68751
Salinas	Training time (s)	15.73	90.71	174.19	44.16	429.27	312.10	111.64	278.34	205.06	7.86	26.38	274.78	47.80
	Testing time (s)	22.57	151.40	445.52	139.67	1049.69	144.59	107.43	1034.41	420.78	18.24	56.48	660.01	84.08
	Parameters	45196	5122176	370312	267219	379474	155008	183062	352405	51655943	148488	972624	847116	61490
Yellow River Delta	Training time (s)	11.43	42.47	33.09	37.83	425.27	576.55	556.22	34.70	116.52	21.40	60.58	59.17	102.87
	Testing time (s)	2.25	7.62	8.45	13.88	13.83	262.07	18.01	10.05	26.43	4.75	15.59	18.64	19.43
	Parameters	43523	5123079	241463	293554	247486	198019	183062	186855	44435576	148943	1342231	829591	135755
WHU-Hi-LongKou	Training time (s)	8.64	22.35	27.04	14.21	153.52	576.62	89.39	154.02	47.88	13.42	26.62	27.24	22.39
	Testing time (s)	73.33	193.35	298.86	210.35	609.26	532.95	181.75	689.67	583.63	139.70	570.78	389.44	224.37
	Parameters	35669	5121273	471513	236872	247136	58749	183062	540644	60104990	148033	1514121	848457	52274

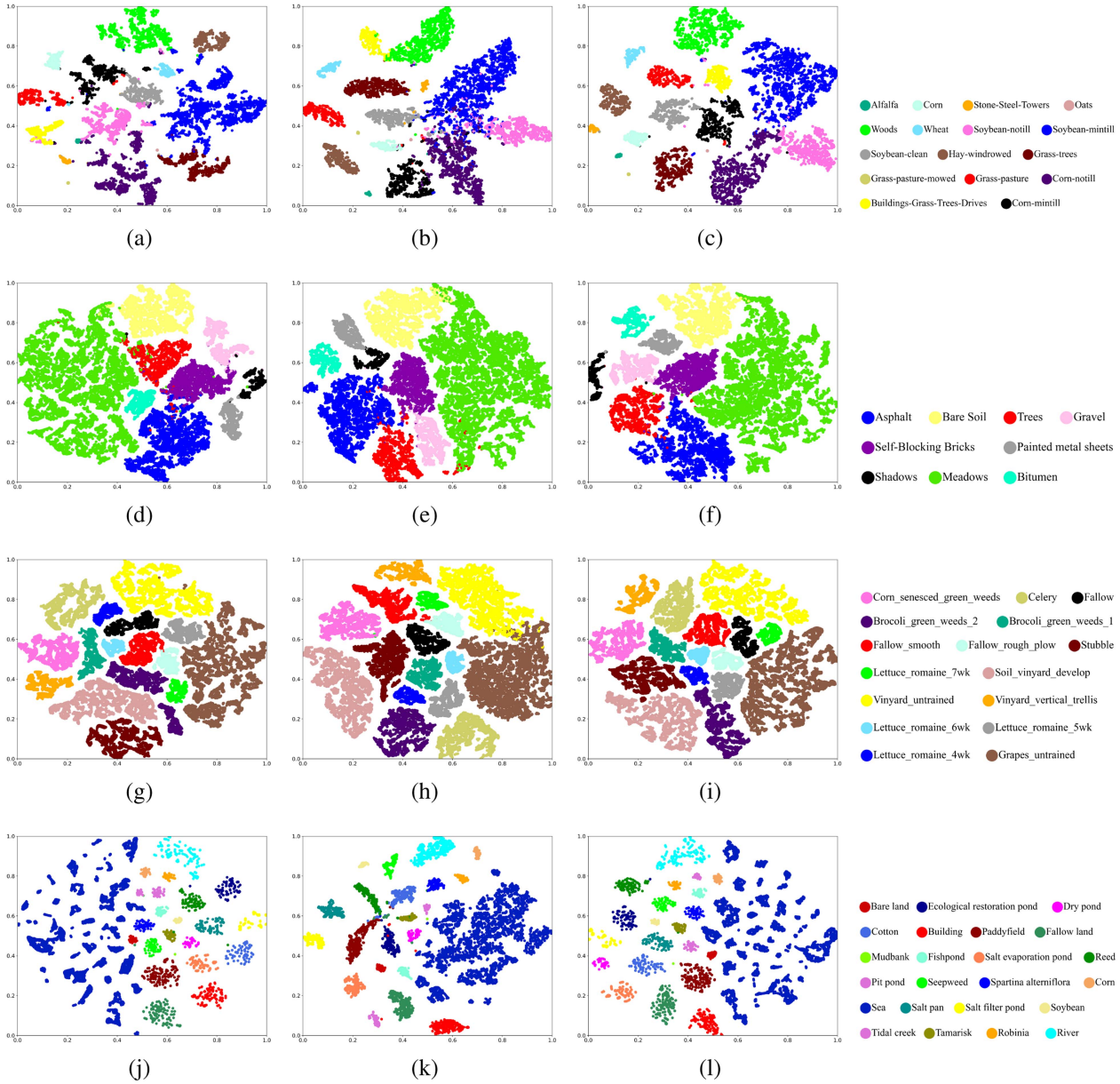


Fig. 19. Feature separability for different variants of S3FAN (10 labeled samples per class for training). (a) S3FAN (without SWEM), (b) S3FAN (without SSIM), (c) S3FAN (Ours) for Indian Pines, (d) S3FAN (without SWEM), (e) S3FAN (without SSIM), (f) S3FAN (Ours) for University of Pavia, (g) S3FAN (without SWEM), (h) S3FAN (without SSIM), (i) S3FAN (Ours) for Salinas, (j) S3FAN (without SWEM), (k) S3FAN (without SSIM), and (l) S3FAN (Ours) for Yellow River Delta.

to features. ST blocks embedded in SWETM realize the recalibration of extracted spectral features, making the shallow spectral features adapt to the subsequent deep extraction process, thus avoiding structural damage to the extracted features. In addition, SWETM provides high-quality spectral features for subsequent SSIM and SWEM, and the weighted fusion of features from SSIM and SWEM enriches feature representation and enhances feature separability.

Although the experiments of S3FAN show its excellent classification ability and feature separation ability under a small number of samples, the spectral subspace interaction and spatial weighted extraction rely on the pre-extraction of spectral features, and there is still much room for optimization in differential processing of spectral features. Furthermore, to ensure the lightweight of the model, the proposed S3FAN divides and conquers the spectral and spatial processing, the fused features are not further explored. In the future, we will explore the interaction between spectral and spatial semantic features and improve the joint representation ability of the attention mechanism for spectral and spatial features.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for making the Airborne Visible/Infrared Imaging Spectrometer Indian Pines hyperspectral dataset available to the community, Prof. P. Gamba for providing the Reflective Optics Spectrographic Imaging System data over Pavia, Italy, and Prof. W. W. Sun for sharing the Yellow River Delta hyperspectral dataset.

The authors would also like to thank Dr. A. Ben Hamida for sharing the code of 3DCNN, Dr. S. K. Roy for HybridSN and A^2S^2K ResNet, Dr. Z. Zhong for SSRN, Prof. L. Jiao for RSSAN, Prof. Z. Xue for S3Net and SPRLT, Dr. J. Zeng for DM-MRN, Prof D. Hong for SF, Prof Y. Zhou for HiT, Prof. L. Sun for SSFTT, and Prof. S. Mei for GAHT, respectively.

REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[2] F. D. v. d. Meer et al., "Multi- and hyperspectral geologic remote sensing: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 14, no. 1, pp. 112–128, 2012.

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

[4] P. Singh et al., "Hyperspectral remote sensing in precision agriculture: Present status, challenges, and future trends," in *Hyperspectral Remote Sensing*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 121–146.

[5] T. Adão et al., "Hyperspectral imaging: A review on UAV-Based sensors, data processing and applications for agriculture and forestry," *Remote Sens.*, vol. 9, no. 11, 2017, Art. no. 1110.

[6] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Lab. J.*, vol. 14, no. 1, pp. 79–116, 2003.

[7] 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.

[8] C. Zhao et al., "Identifying mangroves through knowledge extracted from trained random forest models: An interpretable mangrove mapping approach (IMMA)," *ISPRS J. Photogrammetry Remote Sens.*, vol. 201, pp. 209–225, 2023.

[9] W. Li and Q. Du, "A survey on representation-based classification and detection in hyperspectral remote sensing imagery," *Pattern Recognit. Lett.*, vol. 83, pp. 115–123, Nov. 2016.

[10] 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.

[11] 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.

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

[13] W. Hu et al., "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jul. 2015 , Art. no. 258619.

[14] Q. Liu et al., "Bidirectional-convolutional LSTM based spectral–spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, Dec. 2017, Art. no. 1330.

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

[16] J. Feng, Z. Gao, R. Shang, X. Zhang, and L. Jiao, "Multi-complementary generative adversarial networks with contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5520018.

[17] J. Yue et al., "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, May 2015.

[18] J. Yue, S. Mao, and M. Li, "A deep learning framework for hyperspectral image classification using spatial pyramid pooling," *Remote Sens. Lett.*, vol. 7, no. 9, pp. 875–884, 2016.

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

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

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

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

[23] S. Ghaderizadeh, D. A.-Moghadam, A. Sharifi, N. Zhao, and A. Tariq, "Hyperspectral image classification using a hybrid 3D-2D convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7570–7588, 2021.

[24] H. Gong et al., "Multiscale information fusion for hyperspectral image classification based on hybrid 2D-3D CNN," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2268.

[25] 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.

[26] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[27] M. E. Paoletti, J. M. Haut, R. F.-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[28] 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.

[29] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

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

[31] X. Li, M. Ding, and A. Pižurica, "Spectral feature fusion networks with dual attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5508614.

[32] Q. Liu et al., "Spectral group attention networks for hyperspectral image classification with spectral separability analysis," *Infrared Phys. Technol.*, vol. 108, Aug. 2020, Art. no. 103340.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

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

[35] Q. Hong et al., "SATNet: A spatial attention based network for hyperspectral image classification," *Remote Sens.*, vol. 14, 2022, Art. no. 5902.

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

[37] Z. Xue, M. Zhang, Y. Liu, and P. Du, "Attention-based second-order pooling network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9600–9615, Nov. 2021.

[38] Z. Dong, Y. Cai, Z. Cai, X. Liu, Z. Yang, and M. Zhuge, "Cooperative spectral–spatial attention dense network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 866–870, May 2021.

[39] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[40] 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.

[41] Z. Lu, B. Xu, L. Sun, T. Zhan, and S. Tang, "3-D channel and spatial attention based multiscale spatial–spectral residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4311–4324, 2020.

[42] Z. Xie, J. Hu, X. Kang, P. Duan, and S. Li, "Multilayer global spectral-spatial attention network for wetland hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518913.

[43] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532513.

[44] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[45] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, Jun. 2021.

[46] Y. Zhou, X. Huang, X. Yang, J. Peng, and Y. Ban, "DCTN: Dual-branch convolutional transformer network with efficient interactive self-attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5508616.

[47] 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.

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

[49] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[50] Y. Xu et al., "Spatial–Spectral 1DSwin transformer with groupwise feature tokenization for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516616.

[51] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.

[52] P. Tang, M. Zhang, Z. Liu, and R. Song, "Double attention transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5502105.

[53] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–Spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

[54] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

[55] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When multi-granularity meets spatial-spectral attention: A. hybrid transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401118.

[56] Y. Ding et al., "Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 223, 2023, Art. no. 119858.

[57] 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, 2023, Art. no. 5503615.

[58] B. Zhang, Y. Chen, Y. Rong, S. Xiong, and X. Lu, "MATNet: A combining multi-attention and transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506015.

[59] T. Arshad and J. Zhang, "Hierarchical attention transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5504605.

[60] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4080–4090.

[61] S. Jia et al., "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, 2021.

[62] 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.

[63] N. Wambugu et al., "Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102603.

[64] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501618.

[65] R. Thoreau, V. Achard, L. Risser, B. Berthelot, and X. Briottet, "Active learning for hyperspectral image classification: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 9, pp. 256–278, Sep. 2022.

[66] M. Xu, Q. Zhao, and S. Jia, "Multiview spatial–spectral active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512415.

[67] X. Di, Z. Xue, and M. Zhang, "Active learning-driven Siamese network for hyperspectral image classification," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 752.

[68] B. Yang, S. Hu, Q. Guo, and D. Hong, "Multisource domain transfer learning based on spectral projections for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3730–3739, 2022.

[69] J. Feng et al., "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509617.

[70] B. Qin, S. Feng, C. Zhao, W. Li, R. Tao, and W. Xiang, "Cross-domain few-shot learning based on feature disentanglement for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5514215.

[71] J. Yao et al., "Semi-active convolutional neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537915.

[72] C. Zhao et al., "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, 2023.

[73] B. Xi et al., "Deep prototypical networks with hybrid residual attention for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3683–3700, 2020.

[74] J. Zeng, Z. Xue, L. Zhang, Q. Lan, and M. Zhang, "Multistage relation network with dual-metric for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510017.

[75] Z. Xue, Y. Zhou, and P. Du, "S3Net: Spectral–spatial siamese network for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531219.

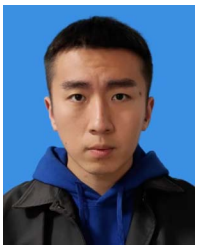
[76] Y. Yu et al., "White-box transformers via sparse rate reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 9422–9457.

[77] Y. Tang et al., "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600213.

[78] L. Yang et al., "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*. PMLR, 2021, pp. 11 863–11 874.

[79] T. Lu, M. Liu, W. Fu, and X. Kang, "Grouped multi-attention network for hyperspectral image spectral-spatial classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507912.

[80] Q. Wang, B. Wu, P. Zhu, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 531–11 539.
[81] H. J. Lee, H. E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1854–1862.
[82] C. Wu, L. Tong, J. Zhou, and C. Xiao, "Spectral-spatial large kernel attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519915.
[83] C. Shi, S. Yue, and L. Wang, "A dual branch multiscale transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5504520.
[84] B. Zhang, Y. Chen, Z. Li, S. Xiong, and X. Lu, "SANet: A self-attention network for agricultural hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol 62, 2023, Art. no. 5501315.



Shun Cheng received the B.S. degree in remote sensing science and technology from Nanjing Tech University, Nanjing, China, in 2023. He is currently working toward the M.E. degree in mapping and remote sensing engineering with Hohai University, Nanjing.

His research interests include hyperspectral image classification and machine learning.

Zhaohui Xue (Member, IEEE) received the B.S. degree in geomatics engineering from Shandong Agricultural University, Tai'an, China, in 2009, the M.E. degree in remote sensing from the China University of Mining and Technology, Beijing, China, in 2012, and the Ph.D. degree in cartography and geographic information system with Nanjing University, Nanjing, China, in 2015.

He is currently a Full Professor (Ph.D. Supervisor) with the College of Geography and Remote Sensing, Hohai University, Nanjing. He has authored more than 60 scientific papers including more than 30 Science Citation Index (SCI) papers. His research interests include hyperspectral image classification, time-series image analysis, pattern recognition, and machine learning.

Dr. Xue was the recipient of the National Scholarship for Doctoral Graduate Students granted by the Ministry of Education of the People's Republic of China in 2014, and the Best Reviewer award for the IEEE Geoscience and Remote Sensing Society. He is an Editorial Board Member in NATIONAL REMOTE SENSING BULLETIN (2020–2024). He has been a Reviewer for more than ten famous remote sensing journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Remote Sensing of Environment*, and *ISPRS Journal of Photogrammetry and Remote Sensing*.

Ziyu Li received the B.S. degree in remote sensing science and technology from Nanjing University of Information Science and Technology, Nanjing, China, in 2022. He is currently working toward the M.E. degree in surveying and mapping at Hohai University, Nanjing.

His research interests include hyperspectral image classification and cross-view remote sensing image geolocalization.

Aijun Xu received the B.S. degree from School of Marine Technology and Geomatics, Jiangsu Ocean University, Lianyungang, China, in 2023. He is currently working toward the M.E. degree in mapping and remote sensing engineering with Hohai University, Nanjing, China.

His research interests include hyperspectral image semantic segmentation.

Hongjun Su (Senior Member, IEEE) received the Ph.D. degree in cartography and geography information systems from the Key Laboratory of Virtual Geographic Environment (Ministry of Education), Nanjing Normal University, Nanjing, China, in 2011.

He is currently a Full Professor with the College of Geography and Remote Sensing, Hohai University, Nanjing. His main research interests include hyperspectral remote sensing dimensionality reduction, classification, and spectral unmixing.

Dr. Su was the recipient of the 2016 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society. He is an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.