remote sensing

MDPI

Article

SS-TMNet: Spatial–Spectral Transformer Network with Multi-Scale Convolution for Hyperspectral Image Classification

Xiaohui Huang [1], Yunfei Zhou [1], Xiaofei Yang [2,*], Xianhong Zhu [1] and Ke Wang [3]

[1] School of Information Engineering, East China Jiaotong University, Nanchang 330013, China
[2] The Department of Computer and Information Science, University of Macau, Macau 519000, China
[3] School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China
* Correspondence: xiaofeiyang@um.edu.mo

Abstract: Hyperspectral image (HSI) classification is a significant foundation for remote sensing image analysis, widely used in biology, aerospace, and other applications. Convolution neural networks (CNNs) and attention mechanisms have shown outstanding ability in HSI classification and have been widely studied in recent years. However, the existing CNN-based and attention mechanism-based methods cannot fully use spatial–spectral information, which is not conducive to further improving HSI classification accuracy. This paper proposes a new spatial–spectral Transformer network with multi-scale convolution (SS-TMNet), which can effectively extract local and global spatial–spectral information. SS-TMNet includes two key modules, i.e., multi-scale 3D convolution projection module (MSCP) and spatial–spectral attention module (SSAM). The MSCP uses multi-scale 3D convolutions with different depths to extract the fused spatial–spectral features. The spatial–spectral attention module includes three branches: height spatial attention, width spatial attention, and spectral attention, which can extract the fusion information of spatial and spectral features. The proposed SS-TMNet was tested on three widely used HSI datasets: Pavia University, IndianPines, and Houston2013. The experimental results show that the proposed SS-TMNet is superior to the existing methods.

Keywords: multi-scale 3D convolution; convolution neural network (CNN); attention mechanism; hyperspectral image (HSI) classification

Citation: Huang, X.; Zhou, Y.; Yang, X.; Zhu, X.; Wang, K. SS-TMNet: Spatial–Spectral Transformer Network with Multi-Scale Convolution for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 1206. https://doi.org/10.3390/rs15051206

Academic Editors: Sidike Paheding and Ashraf Saleem

Received: 18 January 2023
Revised: 17 February 2023
Accepted: 20 February 2023
Published: 22 February 2023

1. Introduction

Hyperspectral image classification is a significant application of remote sensing technology. The hyperspectral remote sensing image has many spectral bands, which provides rich information to achieve a more precise classification of the scene object. Each pixel is a high-dimensional vector with hundreds of wavebands in a hyperspectral image. The numerical value of each vector with hundreds of bands in a hyperspectral image, representing the spectral reflectance at the corresponding wavelengths [1]. HSI classification is the pixel-by-pixel classification of remote sensing scenes, which is extensively used in agriculture, aerospace, biology, and other fields [2,3].

In the past two decades, hyperspectral image classification has received significant attention as an essential application of remote sensing technology. Some traditional machine learning methods [4–6] were proposed for HSI classification tasks in the early years. For instance, the support vector machine (SVM) [4] and K-nearest neighbor (KNN) [5] were used to capture abundant spectral information in HSI classification. Li et al. [6] presented a multinomial logistic regression method to classify HSIs using semi-supervised learning of a posterior distribution. An extended morphological profiles (EMPs) method [7] was proposed in handing the spatial information in HSIs through multiple morphological operations. Although the above HSI classification methods have been proven effective in some cases, the classification effect is not satisfactory when the environment is very complex.

With the development of deep learning, CNNs have made significant breakthroughs in many image-related fields, such as image classification [8–10], object detection [11], and instance segmentation [12]. Owing to the numerous bands of hyperspectral images, the ability of ordinary classifiers will decrease with the increase of dimension, and the accuracy will also decrease. Therefore, the traditional classifier based on CNNs for RGB images cannot directly be used for HSI classification tasks. Researchers have conducted much work and proposed a series of methods. For instance, an HSI classification method based on 2D-CNN proposed by Song et al. [13] used multi-layer feature fusion and residual connection to build a network. Chen et al. [14] used a 3D-CNN-based method for HSI classification and proposed a method combining 3D-CNN and regularization to extract fused global characteristics. Due to the strong ability of CNNs to extract local spatial features, they have shown optimistic results. However, CNN-based methods can not pay sufficient attention to the representation of spectral features, resulting in low utilization of global spectral information and hindering the further improvement of model performance. Chen et al. [15] proposed a method based on stacked autoencoders (SAE) to classify HSIs through layer-by-layer training. Mou et al. [16] presented a new recurrent neural network (RNN) method for HSI classification, which takes image pixels as sequence data for analysis and processing. However, this method can not capture the long-range relationship between spectra, resulting in unsatisfactory classification results.

Recently, a method based on the self-attention mechanism was presented, named Transformer [17], which shows excellent performance in natural language processing tasks. Thereafter, many researchers [18,19] are committed to introducing Transformer into the field of computer vision. Dosovitskiy et al. [18] used Transformer for image recognition and proposed a method named Vision Transformer (ViT), which divided the image into fixed-size patches and added position coding to obtain tokens and finally put them into Transformer Encoder for training. Due to the excellent performance of Transformer and its powerful ability to process sequence information, many researchers have also applied Transformer to the hyperspectral field. He et al. [19] presented a bidirectional encoder Representation Transformer for HSI classification (HSI-BERT) to capture the correlation between spectra using bidirectional Transformer encode representation. However, the networks do not effectively employ the local spatial features of HSI.

In general, all of the above methods for HSI classification have some shortcomings, which are summarized as follows. For the CNN-based methods [20–23], it pays too much attention to local spatial correlation, resulting in the inability to capture long-range spectral correlation, which limits the use of high-dimensional bands of HSIs. Even in the adjacent spectral domain, it is hard for CNN-based methods to capture the subtle discrepancies between different spectra. For the RNN-based methods [16], due to the problems of gradient disappearance and gradient explosion, RNN-based methods cannot learn the long-term dependence of spectral data well. For Transformer-based methods [18,24], although it has certain advantages for establishing remote dependency, Transformer-based methods cannot effectively extract important spatial context information and fused spatial–spectral features. Some improved Transformer-based methods, such as [25–29], although the well-designed CNN is used for spatial feature extraction before Transformer processing, can not effectively capture fused spatial–spectral information. Some HSI classification methods based on graph convolution neural network, such as [30–34], have unsatisfactory results due to the large number of parameters and overfitting problems.

In order to solve the above problems, this work presents a spatial–spatial Transformer network with multi-scale convolution (SS-TMNet) for HSI classification, which can more effectively utilize local and global spatial–spectral information. SS-TMNet includes two key modules: multi-scale 3D convolution project module (MSCP) and spatial–spectral attention module (SSAM). Specifically, we utilize the MSCP module for initial feature mapping to capture the fused spatial–spectral features, and employ the SSAM module to encode the height dimension, width dimension, and spectral dimension features, respectively,

to capture the local and global dependencies of each dimension. The main contributions of this work are as follows.

- We design a new Transformer-based HSI classification method (SS-TMNet), which uses multi-scale convolution and spatial–spectral attention to extract local and global information efficiently.
- We design an MSCP module to extract the fused spatial–spectral features as the initial feature projection. This module uses multi-scale 3D convolutions and feature fusion to extract fused spatial–spectral features from multiple scales efficiently.
- We propose an SSAM module to encode the input features from the height, width, and spectral dimensions. We use multi-dimensional convolution and self-attention to extract more effective local and global spatial–spectral features.
- We have conducted extensive experiments based on three benchmark datasets. The experimental results show that the proposed SS-TMNet outperforms the state-of-the-art CNN-based and Transformer-based hyperspectral image classifiers.

The structure of the work is as follows. Section 2 introduces the related work. Section 3 introduces the proposed SS-TMNet architecture, and then introduces the proposed MSCP module and SSAM module in detail. Section 4 reports and analyzes the experimental results. Section 5 summarizes this work.

2. Related Work

Hyperspectral image classification technology is one of the essential technologies in the field of remote sensing. After years of research, researchers have presented many methods for HSI classification tasks [35–39]. This section mainly summarizes related work in three parts: traditional classification methods, CNN-based methods, and Transformer-based methods.

2.1. Traditional Classification Methods

Some kernel-based methods were proposed in the early stage of HSI classification research. For instance, Melgani et al. [4] applied the SVM method to achieve HSI classification. Unlike SVM, the multiple kernel learning (MKL) method proposed by Rakotomamonje et al. [40] aims to learn the kernel and related predictors simultaneously in a supervised learning environment. However, both methods focus only on the feature information of the spectral dimension and overlook the spatial dimension. Benediktsson et al. [7] proposed extended morphological profiles (EMPs) to study the spatial feature information of HSI. Extended attribute profiles and extended multi-attribute profiles (EMAP) are presented in [41] for capturing spatial information. In order to make better use of the spatial features in HSI, Li et al. [42] presented a generalized composite kernel (GCK) method to model spatial information from the extended multiattribute profiles. In addition, due to the high-dimensional characteristics of HSIs, many works specifically explore how to reduce dimension and extract features more effectively. For instance, Bandos et al. [43] presented a linear discriminant analysis (LDA) method, which can be utilized to solve related ill-posed problems for HSIs. Villa et al. [44] applied the Independent Component Analysis (ICA) method to HSI classification and presented the Independent Component Discriminant Analysis (ICDA) method, which calculates the density function of each independent component by using a nonparametric kernel density estimator. Furthermore, linear versus nonlinear PCA (NLPCA) proposed by Licciardi et al. [45] for HSI classification. There are other methods in the literature, such as DSML-FS based on multimodal learning, which was presented by Zhang et al. [46]. This method utilizes joint structure sparse regularization to explore the relationship between the intrinsic structure of the data and its different characteristics. Jouni et al. [47] proposed an HSI classification method based on tensor decomposition and mathematical morphology by modeling the data as a higher-order tensor. Additionally, Luo et al. [48] introduced a new dimension reduction method for HSI classification, known as local geometric structure Fisher analysis (LGSFA), which uses neighboring points and corresponding intra-class reconstruction points to enhance

intra-class compactness and inter-class separability. However, these methods are based on shallow feature representation, which can show unsatisfactory classification results in complex scenes.

2.2. CNN-Based Methods

With deep learning development, CNN performs excellently in extracting local spatial features. Therefore, numerous CNN-based methods have been presented for the HSI classification task. Hu et al. [49] introduced the CNN into the HSI classification task and proposed a five-tier 1D-CNN-based method. Compared with the traditional classification methods, the effect has been improved. Hao et al. [20] presented a 2D-CNN-based method to classify ground plants. In addition, Fang et al. [22] presented a 3D asymmetric inception network to extract spatial–spectral features and overcome the overfitting problem. Chang et al. [23] presented a novel 3D-CNN-based method to capture the joint spatial–spectral information by stacking layers of 3D-CNN and 2D-CNN. In order to capture fused spatial–spectral information more effectively, He et al. [21] used multi-scale 3D-CNN for HSI classification and presented a multi-scale 3D deep convolution neural network (M3D-DCNN). Although CNN-based methods perform well in HSI classification, capturing the long-range dependence between spectra is challenging. Furthermore, the excessive dependence of CNN on local spatial information makes it difficult to improve the classification accuracy further.

2.3. Transformer-Based Methods

Recently, due to the excellent performance of Transformer in the NLP field, many researchers have applied it to the image classification field. Dosovitskiy et al. [18] presented a ViT method based on Transformer for image classification. However, in the top-level feature representation of the deep ViT model the feature maps are similar, which leads to the incapability of the self-attention mechanism to learn the deeper feature representation. Zhou et al. [24] presented a ViT-based method that can effectively use the deep architecture, called DeepViT, which generates a new set of attention maps by aggregating multiple attention maps dynamically. Although spectral dependence is considered in these methods, the effect of spatial features is omitted. Considering the superior performance of CNN in extracting local spatial features, many researchers applied convolution on Transformer to obtain better performance. Graham et al. [50] re-examined the CNNs, applied it to ViT, and proposed a hybrid neural network of CNN and ViT for image classification, called LeViT. In order to extract multi-scale features from ViT, Chen et al. [51] presented a multi-scale Transformer using cross attention, called CrossViT, which uses multiple multi-scale encoders with two branches for feature extraction. Many researchers also introduced Transformer-based methods into the HSI classification field. For example, He et al. [25] presented an HSI classification method called spatial–spectral transformer (SST), which uses VGGNet [52] to capture basic spatial information and then inputs the Transformer to capture spectral information. Yang et al. [53] presented a novel Transformer-based method called HiT for HSI classification, which uses double branch 3D convolution as feature mapping, embeds the convolution in the encoder of Transformer architecture, and extracts feature information from different dimensions using convolution. However, these methods do not effectively use the advantages of convolution in the attention mechanism, making it impossible to improve the classification effect further. In this work, we propose a novel Transformer-based method called SS-TMNet, which can effectively employ the advantages of convolution and attention mechanisms to extract global and local spatial–spectral features. In the SS-TMNet, two modules, MSCP and SSAM, are proposed to extract multi-scale fused spatial–spectral information and construct cross-dimensional interactions between different dimensions, respectively.

3. The Proposed SS-TMNet Method

In this section, we introduce our SS-TMNet method in three aspects: the overall architecture of SS-TMNet, the MSCP module, and the encoder sequence module.

3.1. The Framework of the Proposed SS-TMNet

This work presents a novel HSI classification method called SS-TMNet based on Transformer. SS-TMNet consists of two key modules: the MSCP module and the SSAM module. MSCP is used for feature projection of the initial HSI image, where multi-scale 3D convolution is utilized to capture the fused multi-scale spatial–spectral information. SSAM is used to capture local and global spatial–spectral dependencies from different spatial and spectral dimensions. The encoder sequence includes four stages where a downsampling layer is added to reduce the dimensions after the second stage. Moreover, a global residual connection connects the input and the final output. Figure 1 shows the overall architecture of our SS-TMNet method.

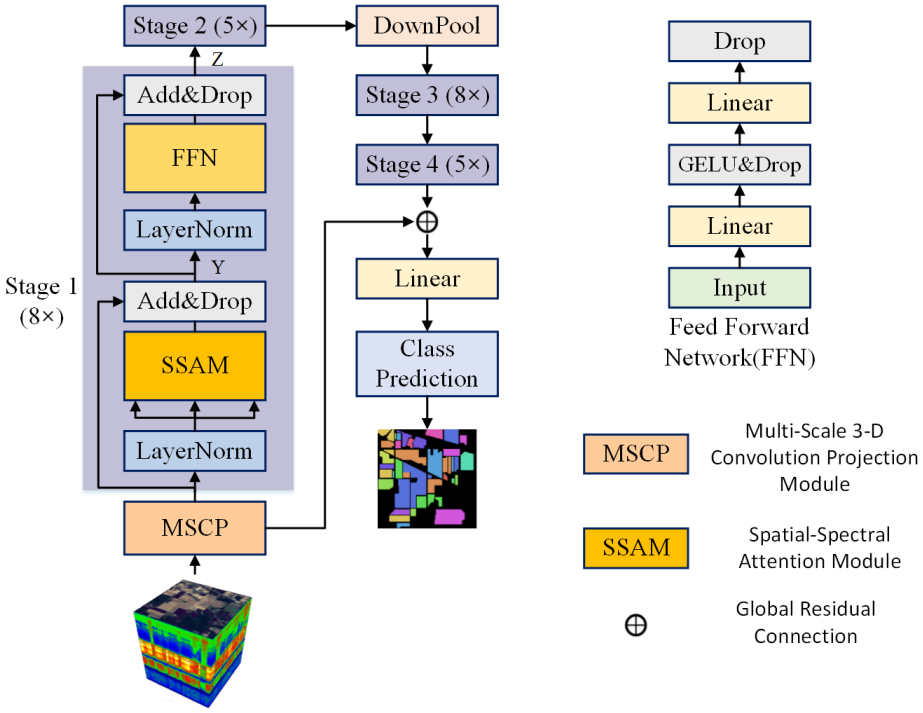


Figure 1. The overall architecture of the proposed SS-TMNet. The MSCP is a multi-scale 3D convolution projection module to extract the fused multi-scale spatial–spectral information. The extracted features are fed into the encoder sequence with four stages. Finally, a fully connected layer is used for category prediction.

3.2. MSCP Module

3.2.1. Multi-Scale 3D Convolution

Hyperspectral images differ from ordinary RGB images. Because of the high-dimensional characteristics of HSI, ordinary 2D convolution can not effectively capture the fused spatial–spectral information because it ignores the dependence between spectra. Meanwhile, 3D convolution can process the features from three dimensions, which can extract features more effectively. In general, HSI data can be represented by a tensor with the size of $C \times S \times H \times W$, where C represents the number of channels, S denotes the spectral domain, and H and W are the height and width in the spatial domain. Based on this, we can apply 3D convolution to the initial HSI data to extract more effective feature representation

for subsequent network learning. More specifically, the formula for 3D convolution is as follows:

$$v_{ij}^{xyz} = F\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \quad (1)$$

where m represents the feature map in the $(i-1)$th layer connected to the jth feature map, and P_i and Q_i are the height and width of the spatial convolution kernel, R_i is the size of the 3D kernel along the spectral dimension, w_{ijm}^{pqr} is the value at the (p,q,r)th position of the kernel connected to the mth feature map of the preceding layer. b_{ij} is the bias of the jth feature map in the ith layer. F represents the activation function.

We studied HSI's data characteristics and found that multi-scale 3D can perform feature mapping more effectively than ordinary 3D convolution. As shown in Figure 2, we developed a multi-scale 3D convolution to build the data mapping module and proposed a new feature mapping module called MSCP. The multi-scale convolution layer uses different sizes of 3D convolution to extract the feature map. From a global perspective, we extract features from the feature information of interest in the image to obtain new feature maps of different sizes and then fuse them to obtain the spatial–spectral feature map. The feature map obtained through the MSCP module has rich fused spatial–spectral information, which enhances the efficiency of feature extraction of subsequent networks.

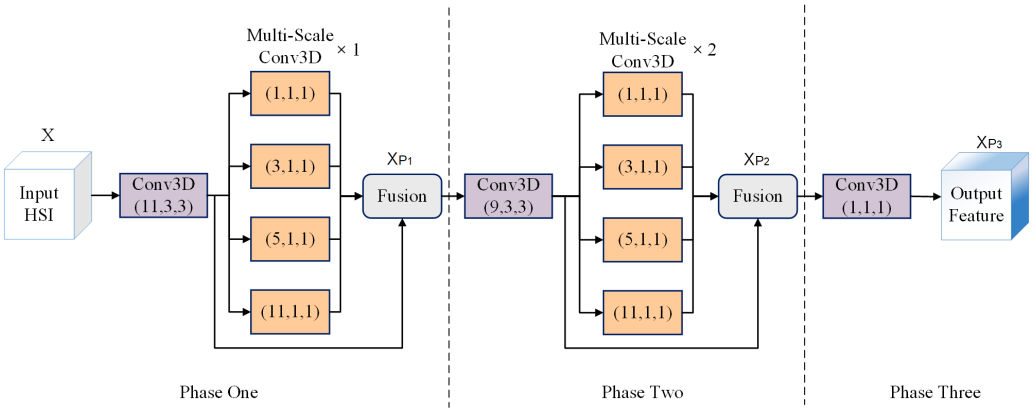


Figure 2. The overall architecture of the proposed MSCP module.

3.2.2. Module Composition

Figure 2 shows that the MSCP module comprises multiple multi-scale 3D convolution layers and feature fusion modules. MSCP processes input HSI data in three phases. Suppose $X \in \mathbb{R}^{C\times S\times H\times W}$ is a patch of the input data (in this paper, the input image is divided into several patches with the size of $H \times W$ for processing, and the values of H and W are 15 in the experiments). In the first phase P_1, the input data X are placed into a 3D convolution layer with ReLU operation to extract the spatial–spectral characteristics X_1, where the convolution kernel size is set to (11, 3, 3). Then, X_1 is fed into a multi-scale 3D convolution layer M_1 with four different convolution kernel sizes, mainly used to extract spectral characteristics of different scales. Then, we fuse the output multi-scale features with the addition operation. To prevent overfitting, we use the residual connection to link the fused multi-scale feature to the output of the first 3D convolution layer X_1. The BatchNorm and ReLU operations are then used to produce the first stage output X_{P_1}. The formula for feature mapping in the first stage is as follows:

$$\begin{aligned} X_1 &= ReLU(\text{Conv 3D}(X)), \\ X_{P_1} &= M_1(X_1) = ReLU(BN(X_1 \oplus \sum_{i=1}^{4} \text{Conv 3D}(X_1))), \end{aligned} \quad (2)$$

where *ReLU* represents the activation function, *BN* represents the BatchNorm operation, $\oplus$ represents the residual connection, and i represents the 3D convolution of different scales.

In the second stage P_2, we first feed the output X_{P_1} of the first stage into a 3D convolution layer with a ReLU operation whose convolution kernel size is (9, 3, 3) to further extract the spatial–spectral characteristics. The output features are placed in two successive multi-scale 3D convolution layers M_2 and M_3 with feature fusion and residual connection operations to extract deeper spectral features. Then, we perform BatchNorm and ReLU operations to obtain the output X_{P_2}. In the third stage, P_3, the activation function and 3D pointwise convolution operation are used to handle the features of the output further X_{P_2} in the second stage. Finally, the MSCP module outputs the final representation $X_{P_3} \in \mathbb{R}^{H \times W \times D}$ as the extracted features. The formula for the second and third stages is as follows:

$$\begin{aligned} X_{P_2} &= M_3(M_2(ReLU(\text{Conv 3D}(X_{P_1})))), \\ X_{P_3} &= F_{rescale}(\text{GELU}(\text{Conv 3D}(X_{P_2}))). \end{aligned} \tag{3}$$

Overall, the proposed approach employs multiple multi-scale 3D convolution layers to extract fused spatial–spectral feature information at multiple scales, as well as shallow local spatial–spectral dependences. To mitigate the issue of gradient disappearance, residual connections are used in multiple locations. The extracted fused spatial–spectral information provides an excellent feature representation for processing subsequent encoder sequences.

3.3. Encoder Sequence

3.3.1. Encoder

As shown in Figure 1, the encoder consists of two modules: SSAM and FFN modules. SSAM encodes features from height, width, and spectral dimensions to extract local and global spatial–spectral features. FFN consists of linear layers with the activation function GELU, which is used to transform features and extract deeper features. The encoder adds LayerNorm and residual connection operations to alleviate overfitting and gradient disappearance, and more effectively cooperates with the above two modules for feature extraction. Given an input embedding $X_{P_3} \in \mathbb{R}^{H \times W \times D}$, the formulas of the coding process are as follows:

$$\begin{aligned} Y &= X_{P_3} \oplus \text{SSAM}(\text{LayerNorm}(X_{P_3})), \\ Z &= Y \oplus \text{FFN}(\text{LayerNorm}(Y)), \end{aligned} \tag{4}$$

where $\oplus$ represents residual connection, Y represents the residual connection between X_{P_3} and the output of SSAM, and Z represents the FFN module's output. In general, the SS-TMNet has four stages, and each stage consists of an encoder sequence composed of a different number of encoders. The implementation details of SSAM will be introduced in the next section.

3.3.2. SSAM Module

Figure 3 shows the structure of the SSAM, which encodes the inputting feature along the height, width, and spectral dimensions to extract local and global spatial–spectral features more effectively. We feed X_{in} (X_{P_3} after layer normalization operation) into three branches for height-spatial coding, width-spatial coding, and spectral coding.

In the height branch L_H, we utilize a depthwise convolution layer with convolution kernel size (1,3) to the X_{in} to obtain local height spatial features, which will be fed into the hight spatial attention (HSA) to calculate the spatial self-attention and obtain the global dependent X_H. In the width branch L_W, we employ a depthwise convolution layer with a convolution kernel size of (3,1) to handle the X_{in} and obtain local width spatial characteristics. The width spatial attention (WSA) module is then used to derive a globally dependent X_W based on the local width spatial characteristics. In the spectral branch L_S, local spectral information is captured from the X_{in} using a pointwise convolution layer with convolution kernel size (1,1). Then, the spectral attention (SA) is utilized to obtain globally dependent X_S.

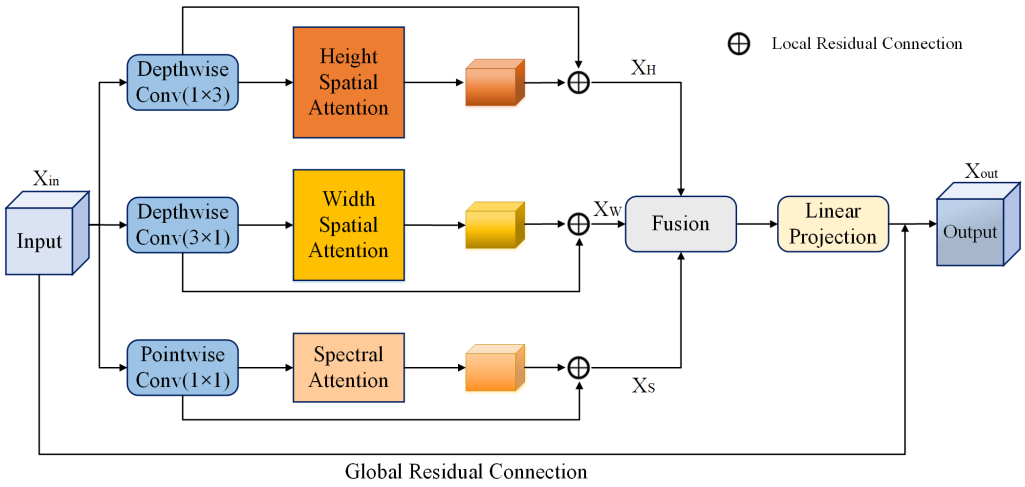


Figure 3. The architecture of the SSAM module.

Then, three local residuals connecting the input X_{in} with the outputs of hight spatial attention, width spatial attention and spectral attention are also added to alleviate gradient disappearance. It is worth noting that three learnable parameters γ_h, γ_w, and γ_s are used to adjust the proportion of learning to characteristics for each branch. Finally, we fuse the feature information of the three branches with the addition operation and linear projection, and join the global residual connection to the X_{in} to get the final output $X_{out} \in \mathbb{R}^{H\times W\times D}$. The calculation formula for SSAM is as follows:

$$\begin{aligned} X_H &= L_H(X_{in}) = \gamma_h \times \mathrm{HSA}(\mathrm{DepthConv}(X)) \oplus X_{in}, \\ X_W &= L_W(X_{in}) = \gamma_w \times \mathrm{WSA}(\mathrm{DepthConv}(X)) \oplus X_{in}, \\ X_S &= L_S(X_{in}) = \gamma_s \times \mathrm{SA}(\mathrm{PointConv}(X)) \oplus X_{in}, \\ X_{out} &= F(X_H + X_W + X_S) \oplus X_{in}, \end{aligned} \tag{5}$$

where γ_h, γ_w, and γ_s represent the learnable parameters, DepthConv represents a depthwise convolution layer, PointConv represents a pointwise convolution layer, $\oplus$ represents residual connection, and F denotes linear projection. Next, we will detail the spatial and spectral attention modules.

As shown in Figure 4, we introduce spatial attention for feature extraction to establish rich spatial feature dependency. We first reshape the $X_{in} \in \mathbb{R}^{H\times W\times D}$ to $X_{re} \in \mathbb{R}^{(H\times W)\times D}$, and then send it to three parallel linear layers for feature mapping to obtain the output $\{Q, K, V\} \in \mathbb{R}^{N\times D}$, where N equals H times W. The concrete procedure of spatial attention can be formulated as follows:

$$X_{out} = \mathrm{Linear}(\mathrm{Transpose}(\mathrm{softmax}\left(\frac{Q \otimes K^T}{\sqrt{d}}\right)) \otimes V), \tag{6}$$

where $\otimes$ denotes the operation of matrix multiplication, and d is the scale factor. Finally, a linear layer maps the feature and reshapes the dimension to obtain the final output $X_{out} \in \mathbb{R}^{H\times W\times D}$. Furthermore, our spectral attention part is similar to spatial attention. To simplify the calculation, our spectral attention discards the initial linear projection layer and uses the input features to calculate the self-attention.

In summary, the SSAM module uses depthwise convolution and pointwise convolution to map features from height, width, and spectral dimensions, respectively, and further extract features using spatial and spectral attention. To extract long-range relationship dependences of both spatial and spectral features, we utilize spatial and spectral self-attention mechanisms. Specifically, our SSAM module integrates convolution with self-attention mechanisms, extracting features from three dimensions and fusing them to obtain feature representations with both global and local dependencies.

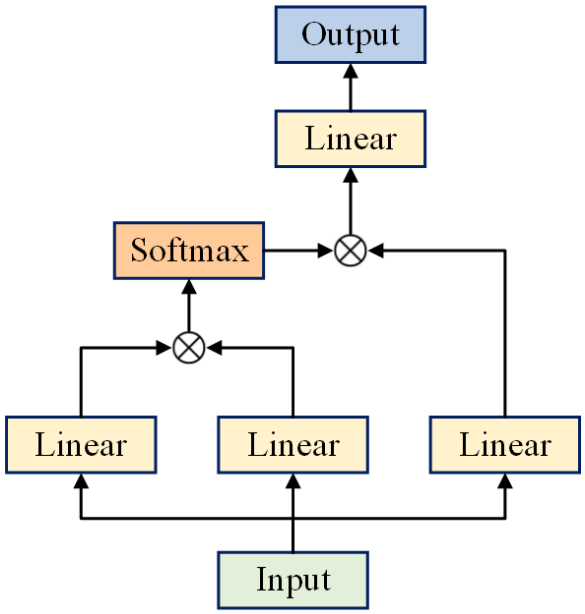


Figure 4. The structure of the height and width spatial attention modules. ⊗ denotes the operation of matrix multiplication.

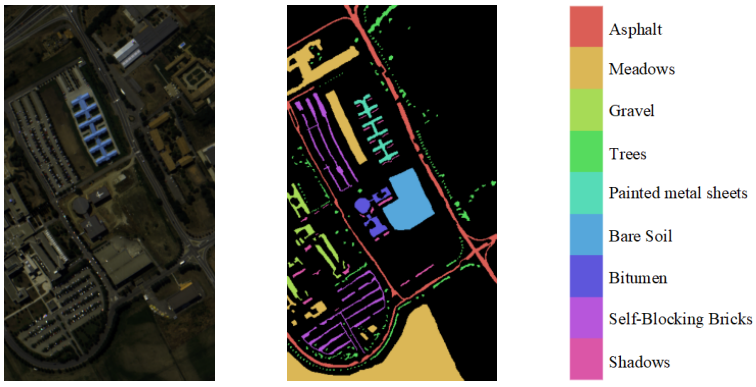
4. Experiments

This section introduces the HSI datasets used in the experiments, including the Pavia University, Indian Pines, and Houston2013 datasets. In addition, we introduce the parameter settings, evaluation metrics, and comparison models in experimental settings. Then, we show and analyze the results. Finally, the ablation experiment and model performance analysis are introduced.

4.1. Datasets

4.1.1. Pavia University Dataset

This dataset was obtained with the Reflective Optical Spectral Imaging System (ROSIS) sensor of the University of Pavia, Italy. The spatial size of the hyperspectral image is 610 × 340 pixels, the spectral bands range from 0.43 to 0.86 µm, a total of 103 bands, excluding 12 water absorption bands. The dataset has 9 classification categories. The dataset is shown in Figure 5.



(a) (b) (c)

Figure 5. The Pavia University dataset. (**a**) false-color composite image; (**b**) ground truth map; (**c**) label color bar.

4.1.2. Indian Pines Dataset

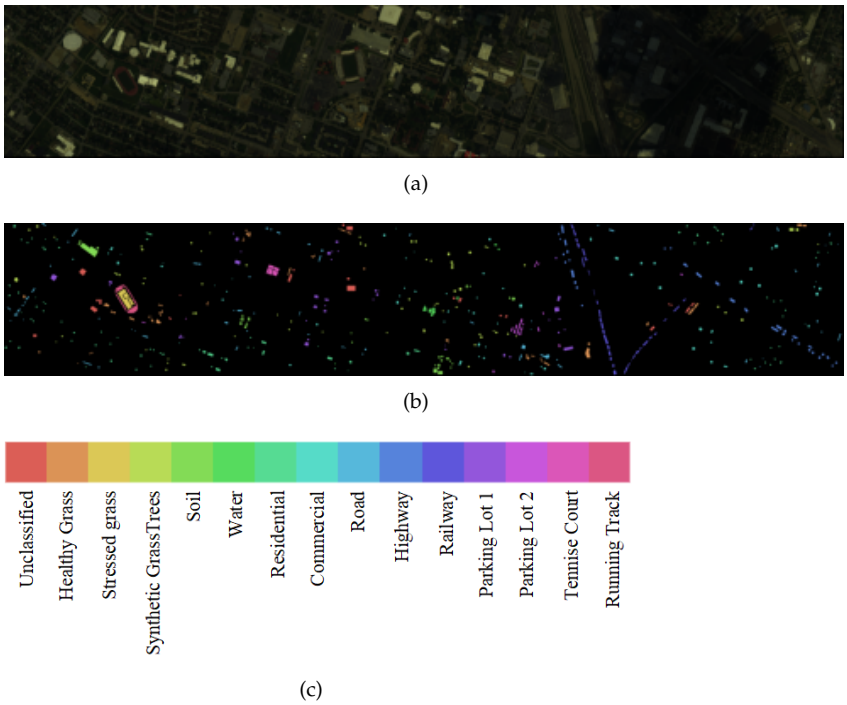
This dataset was collected in 1992 by the AVIRIS sensors in northwestern India, USA. The spatial size of the hyperspectral image in the dataset is 145 × 145 pixels, and spectral bands range from 0.4 µm to 2.5 µm. The total number of spectral bands is 200, excluding 20 water absorption bands. Available ground truths comprise 16 classes. The dataset is shown in Figure 6.

(a) (b) (c)

Figure 6. The Indian Pines dataset. (**a**) False-color composite image; (**b**) ground truth map; (**c**) label color bar.

4.1.3. Houston2013 Dataset

This dataset was captured by the CASI-1500 sensor over the University of Houston and its surroundings in Texas, USA. The spatial size of the image in the dataset is 949 × 1905 pixels, and the spectral dimension includes 144 bands. The dataset has 15 classification categories. The dataset is shown in Figure 7.



(a)

(b)

(c)

Figure 7. The Houston2013 dataset. (**a**) False-color composite image; (**b**) ground truth map; (**c**) label color bar.

4.2. Experimental Setup

4.2.1. Parameters Setting

The training samples for this work were set to 10% in three datasets, and the rest were used as test samples. It is noteworthy that the selection of training and testing samples was

random. To ensure the fairness of the comparative trials, we performed all the comparison models ten times and recorded the results as mean ± standard deviation to compare the performance of different models. The proposed SS-TMNet and the compared methods were implemented on a NVIDIA RTX 3080Ti GPU machine with the pytorch [54] platform. We used the Adam optimizer for gradient descent and set the initial learning rate to 1×10^{-4}. The mini-batch size was set to 32, and we set the epochs on these three benchmark datasets to 200.

4.2.2. Evaluation Metrics

Overall accuracy (OA) and Kappa coefficient (K) were chosen in our experiments to evaluate the results produced by different models in experiments. The OA is the average accuracy for each category. The Kappa measures whether the classification results are consistent with the actual underlying category. The formulas for calculating the above evaluation criteria are as follows:

$$\begin{aligned} OA &= \left(\frac{1}{n} \sum_{k} \left(\frac{TP + TN}{TP + TN + FN + FP} \right)_k \right), \\ K &= \frac{N \sum_{i=1}^{n} x_{ii} - \sum_{i=1}^{n} (x_{i+} \times x_{+i})}{N^2 - \sum_{i=1}^{n} (x_{i+} \times x_{+i})}, \end{aligned} \tag{7}$$

where TP represents the true positive value, TN is the true negative value, FP represents the false positive value, and FN represents the false negative value. n is the number of categories, and N is the total number of data samples. x_{ii} denotes the value on the diagonal line of the confusion matrix, x_{i+} and x_{+i} denote the total value of rows i and columns i of the confusion matrix, respectively.

4.2.3. Baselines

To validate the proposed SS-TMNet method, several representative baselines and the most advanced backbone methods are chosen for comparison, including RNN-based methods (such as Mou [16]), CNN-based methods (such as He [21], 3D-CNN [55], and HybridSN [56]), and Transformer-based methods (such as ViT [18], CrossViT [51], LeViT [50], RvT [57], and HiT [53]). A more detailed description is as follows:

- Mou [16]: An RNN-based method, which uses a recurrent layer containing multiple gated recurrent units. In addition, a fully connection layer and softmax layer are utilized to construct the network.
- He [21]: A 3D-CNN-based method is composed of 3D convolution layers and multi-scale 3D convolution layers. Each multi-scale 3D convolution layer consists of four sublayers.
- 3D-CNN [55]: Another 3D-CNN method includes three convolution blocks and two fully connection layers. Each convolution block includes a 3D convolution layer, a BatchNorm layer, and an average pooling layer.
- HybridSN [56]: A method integrating 2D and 3D convolution, including three 3D convolution layers, one 2D convolution layer, and two fully connection layers.
- ViT [18]: A classic Transformer-based method, which firstly splits the input image into 16 × 16 patches and then feed them into the Transformer encoder to learning the representation of the image.
- CrossViT [51]: A method based on dual-branch ViT architecture, where each branch contains a linear projection layer and a different number of Transformer encoders for processing different sized image patches.
- LeViT [50]: Another Transformer-based method, which includes four convolution layers and three stage codes, and each stage contains four multiple attention layers. We replicated the methods used for HSI classification according to this architecture.
- RvT [57]: Based on ViT, the RvT method uses a pooling layer to downsample the image and reduce the size of the images. We follow this architecture to design the network for the HSI classification tasks.

- HiT [53]: A method of embedding convolution into Transformer, which uses two proposed SACP layers based on 3D convolution to process the input image. Feature extraction is performed using a three-branch convolution layer based on transformer architecture.

4.3. Results and Analysis

This section will elaborate the experimental results and analysis, including results comparison and visualization of three datasets: Pavia University, Indian Pines, and Houston2013.

4.3.1. Experimental Analysis on Pavia University Dataset

Table 1 shows the experimental results produced by different comparison models with respects to OA and Kappa metrics on the Pavia University dataset. The table shows that our proposed SS-TMNet is superior to all comparison methods, with OA and Kappa reaching 91.74% and 89.44%. OA was 0.6%, 0.3%, 0.16% higher than RNN-based method Mou [16], CNN-based method HybridSN, and Transformer-based method LeViT, respectively. The possible reason is that SS-TMNet can more effectively capture local and global dependencies. Among all the mentioned methods, the original ViT method performed the worst, with 88.92% OA and 85.81% Kappa, which indicates that it is difficult for the original ViT network to perform the hyperspectral classification task. The reason may be that ViT lacks effective modeling capabilities for spatial characteristics. The methods using only 3D convolution, such as He [21] and 3D-CNN, which obtained 89.97% and 90.72% OA, respectively, did not perform well since these methods focus on only spatial characteristics and spectral correlation is not fully considered. Transformer-based methods, such as LeViT and HiT, their OA metrics were 91.58% and 91.28%, respectively. They are developed on the basis of ViT, which performs better than ViT-only and 3D convolution-only methods. This demonstrates that combining the convolution and Transformer networks can improve classification results.

Table 1. The comparative experimental results on Pavia University dataset (Bold numbers represent the best results for the corresponding category).

Class	Methods									
#	Mou	He	3D-CNN	HybridSN	ViT	CrossViT	LeViT	RvT	HiT	SS-TMNet
1	90.32 ± 0.41	93.34 ± 0.80	93.97 ± 0.74	95.30 ± 0.57	92.92 ± 0.58	94.67 ± 0.22	95.70 ± 0.37	94.63 ± 0.43	95.14 ± 0.28	**96.11 ± 0.24**
2	**95.77 ± 0.14**	92.11 ± 0.20	92.56 ± 0.10	92.65 ± 0.06	91.13 ± 0.21	92.13 ± 0.13	92.61 ± 0.10	91.95 ± 0.19	92.53 ± 0.08	92.67 ± 0.08
3	75.34 ± 0.69	84.99 ± 2.20	88.73 ± 1.39	90.68 ± 1.44	82.35 ± 1.41	87.63 ± 0.87	91.48 ± 1.04	87.41 ± 1.00	89.91 ± 1.29	**92.35 ± 0.66**
4	94.63 ± 0.46	97.08 ± 0.29	96.31 ± 0.40	**97.30 ± 0.24**	95.80 ± 0.47	96.86 ± 0.31	96.80 ± 0.26	96.92 ± 0.39	97.15 ± 0.17	96.46 ± 0.49
5	99.80 ± 0.15	99.77 ± 0.11	99.79 ± 0.16	99.93 ± 0.08	99.69 ± 0.21	99.89 ± 0.07	99.51 ± 0.77	**99.94 ± 0.06**	99.91 ± 0.07	99.66 ± 0.16
6	85.96 ± 0.53	97.66 ± 0.79	99.52 ± 0.29	99.77 ± 0.18	94.74 ± 0.86	98.26 ± 0.39	99.54 ± 0.14	97.61 ± 0.61	99.38 ± 0.23	**99.91 ± 0.09**
7	71.43 ± 2.52	91.48 ± 1.65	92.22 ± 1.68	96.51 ± 1.55	90.72 ± 1.34	95.05 ± 0.95	97.90 ± 1.08	95.92 ± 0.85	95.79 ± 1.50	**99.05 ± 0.57**
8	82.87 ± 0.67	94.39 ± 1.05	95.95 ± 1.26	97.25 ± 1.22	94.43 ± 0.58	96.69 ± 0.38	**98.84 ± 0.29**	96.44 ± 0.53	97.39 ± 0.57	98.31 ± 0.38
9	99.44 ± 0.20	98.97 ± 1.00	97.50 ± 1.63	99.61 ± 0.37	97.79 ± 0.96	99.74 ± 0.19	97.83 ± 2.12	**99.83 ± 0.19**	99.47 ± 0.25	98.02 ± 0.76
OA(%)	91.14 ± 0.21	89.97 ± 0.36	90.72 ± 0.37	91.44 ± 0.28	88.92 ± 0.31	90.70 ± 0.13	91.58 ± 0.17	90.55 ± 0.20	91.28 ± 0.21	**91.74 ± 0.12**
K(%)	88.19 ± 0.27	87.17 ± 0.47	88.13 ± 0.47	89.06 ± 0.35	85.81 ± 0.40	88.11 ± 0.17	89.24 ± 0.22	87.92 ± 0.25	88.85 ± 0.27	**89.44 ± 0.16**

The visualization experiment results produced by the comparison methods are shown in Figure 8. As shown in the red rectangle box in the figure, most methods produce much noise in the classification maps compared to our SS-TMNet method. It is worth noting that although the classification maps of the HybridSN and LeViT methods are similar to ours, there is still a small amount of noise, and from the evaluation metrics in Table 1, the results of the method we presented are still better. The possible reason is that compared with HybridSN and LeViT, SS-TMNet learns the fused local spatial–spectral features through the proposed MSCP module and more effective local and global feature representations through the SSAM module. The visualization proves that our proposed method can produce better results than most existing methods.

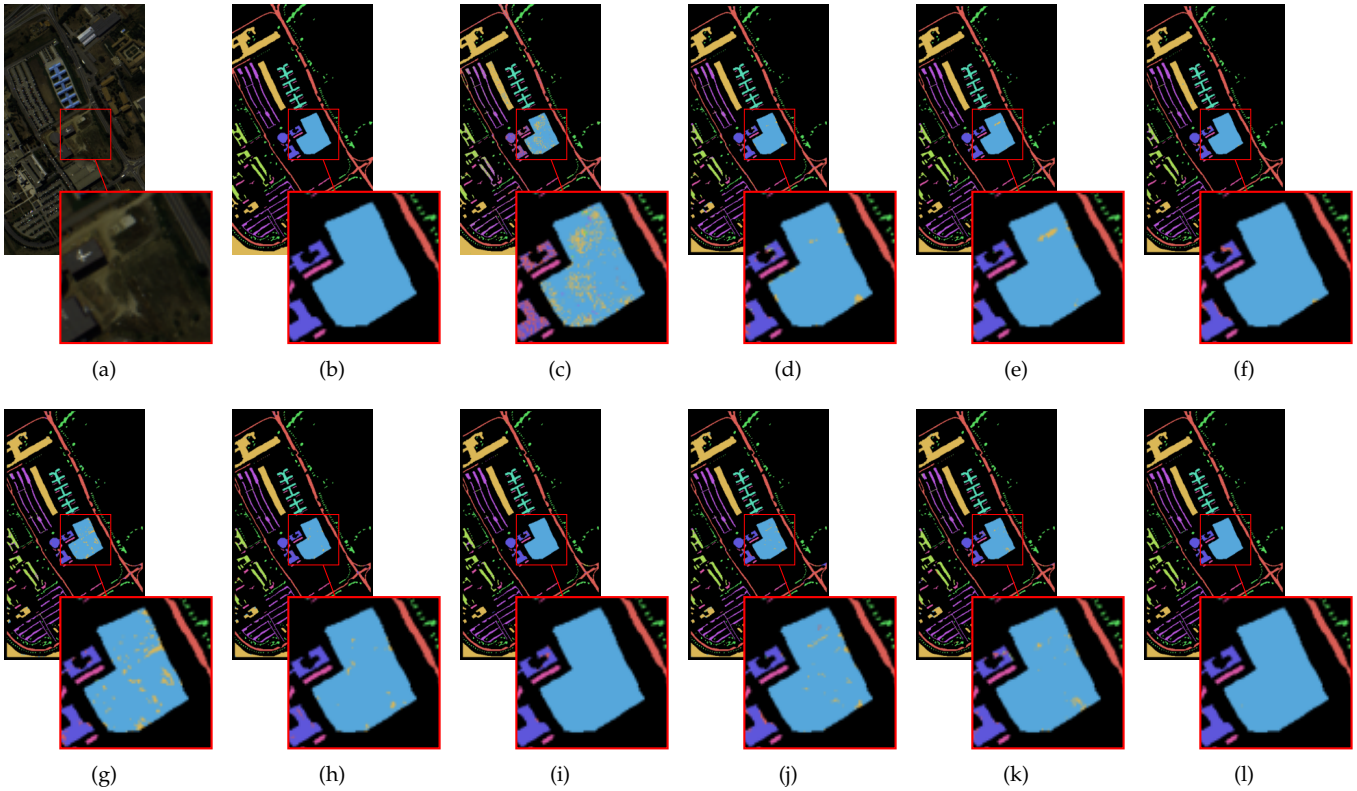


Figure 8. Visualization of the experimental results on Pavia University dataset. (**a**) Original image, (**b**) Ground truth, (**c**) Mou, (**d**) He, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) ViT, (**h**) CrossViT, (**i**) LeViT, (**j**) RvT, (**k**) HiT, (**l**) SS-TMNet(Ours).

4.3.2. Experimental Analysis on Indian Pines Dataset

Table 2 shows the evaluation results of our presented and compared models on the dataset. Our proposed SS-TSNet shows the best results, with OA and Kappa reaching 84.67% and 82.66%. The OA metric of our proposed method is 9.40% higher than RNN-based methods (i.e., Mou), 12.08% higher than CNN-based methods (i.e., 3D-CNN), and 1.04% higher than Transformer-based methods (i.e., LeViT). One possible reason is that we have improved the encoding of feature projections and spatial–spectral features to enable more efficient feature encoding.

Our method differs from the existing methods (i.e., CrossViT, LeViT, RvT, and HiT). We use MSCP to capture the spatial–spectral dependence of the fused multi-scale features. Meanwhile, the SSAM is presented to capture the local and global spatial–spectral information of multidimensional data. Thus, our proposed model can more effectively model the HSIs from spatial–spectral dependence and local–global features. Figure 9 shows the dataset's visualization results, which shows that our proposed model produces the classification map with the least noise and achieves satisfactory results. For example, as shown in the red rectangle in the figure, compared with other comparison methods, the SS-TMNet method generates the slightest noise in the classification map. The reason HiT does not perform well relative to our proposed method may be due to its ineffective integration of convolution into Transformer, leading to a lack of effective modeling of global feature dependencies. From the overall effect, our proposed method produces a classification map closer to the ground truth image than other methods, which proves the validity of our proposed method.

Table 2. The comparative experimental results on Indian Pines dataset (Bold numbers represent the best results for the corresponding category).

Class	Methods									
#	Mou	He	3D-CNN	HybridSN	ViT	CrossViT	LeViT	RvT	HiT	SS-TMNet
1	31.28 ± 11.84	70.82 ± 13.59	49.33 ± 19.93	34.68 ± 26.05	50.11 ± 10.28	62.57 ± 11.25	65.77 ± 11.61	51.24 ± 14.29	80.64 ± 8.22	**87.48 ± 8.15**
2	72.76 ± 1.68	62.45 ± 10.18	68.47 ± 4.06	67.37 ± 20.83	65.46 ± 2.57	62.14 ± 8.14	**88.59 ± 4.60**	76.48 ± 3.17	86.18 ± 4.40	88.56 ± 2.34
3	55.39 ± 2.30	48.96 ± 12.23	51.89 ± 7.60	44.89 ± 22.79	52.57 ± 2.58	41.82 ± 9.88	72.43 ± 3.22	65.42 ± 5.50	69.94 ± 4.99	**76.50 ± 3.23**
4	47.20 ± 6.37	48.86 ± 14.14	40.95 ± 10.33	34.47 ± 25.45	57.92 ± 7.50	65.34 ± 11.69	77.73 ± 3.51	77.96 ± 5.31	75.63 ± 5.39	**82.19 ± 3.92**
5	**85.59 ± 2.77**	62.62 ± 19.22	75.29 ± 4.92	55.75 ± 24.99	52.76 ± 5.08	55.28 ± 5.24	79.78 ± 2.11	50.33 ± 3.87	75.26 ± 2.71	81.71 ± 3.49
6	93.19 ± 0.92	91.35 ± 4.93	93.40 ± 3.30	81.17 ± 21.38	79.49 ± 2.52	88.64 ± 1.35	95.69 ± 1.53	86.43 ± 2.16	94.79 ± 1.60	**97.76 ± 0.92**
7	50.16 ± 17.71	46.70 ± 15.68	22.49 ± 17.72	13.61 ± 16.56	43.72 ± 16.49	38.62 ± 33.31	21.61 ± 25.55	62.93 ± 21.35	**73.03 ± 18.03**	72.09 ± 16.30
8	93.37 ± 0.81	92.52 ± 2.77	91.76 ± 1.40	75.92 ± 26.44	89.41 ± 2.32	89.45 ± 2.41	91.48 ± 1.89	92.02 ± 1.13	93.09 ± 0.78	**94.39 ± 0.47**
9	33.62 ± 14.20	63.57 ± 18.28	35.74 ± 23.68	32.78 ± 29.91	31.35 ± 13.48	13.99 ± 19.46	23.43 ± 26.35	47.50 ± 13.45	59.99 ± 16.58	**68.66 ± 17.26**
10	66.05 ± 1.98	62.03 ± 17.77	72.40 ± 4.64	52.51 ± 32.21	61.48 ± 2.95	59.01 ± 6.94	83.09 ± 3.11	73.70 ± 4.17	85.34 ± 3.43	**87.19 ± 1.98**
11	72.82 ± 1.14	75.50 ± 6.83	79.24 ± 2.16	80.16 ± 9.91	72.26 ± 1.33	70.54 ± 4.43	**92.85 ± 1.22**	79.74 ± 3.35	89.73 ± 2.47	90.70 ± 1.63
12	60.66 ± 2.77	50.03 ± 15.02	58.43 ± 8.37	49.05 ± 25.87	51.64 ± 2.99	40.73 ± 14.18	**83.30 ± 5.45**	66.83 ± 6.57	76.38 ± 7.70	81.85 ± 3.97
13	94.23 ± 2.25	92.96 ± 4.79	96.80 ± 2.83	70.88 ± 25.35	86.61 ± 3.46	87.18 ± 3.73	92.75 ± 5.96	88.69 ± 5.38	95.57 ± 1.90	**97.18 ± 3.02**
14	92.56 ± 0.75	92.44 ± 2.79	93.60 ± 1.34	89.57 ± 10.91	88.50 ± 1.37	90.59 ± 0.57	**97.09 ± 0.66**	89.64 ± 1.10	94.53 ± 1.25	96.21 ± 0.89
15	61.43 ± 3.35	48.79 ± 5.36	44.69 ± 9.52	29.38 ± 13.56	44.94 ± 3.84	47.55 ± 4.32	58.74 ± 6.96	48.06 ± 7.94	58.84 ± 7.65	**63.92 ± 3.78**
16	84.57 ± 2.65	55.79 ± 16.94	55.15 ± 15.24	30.91 ± 30.87	48.29 ± 12.06	27.14 ± 28.65	87.47 ± 10.02	**94.67 ± 3.59**	86.10 ± 6.24	87.73 ± 3.15
OA(%)	75.27 ± 0.77	69.25 ± 6.60	72.59 ± 2.80	67.26 ± 13.98	66.21 ± 0.89	65.71 ± 4.03	83.63 ± 1.13	73.98 ± 2.35	82.13 ± 2.65	**84.67 ± 1.25**
K(%)	71.57 ± 0.87	64.72 ± 8.00	68.69 ± 3.27	62.21 ± 16.96	61.65 ± 0.98	60.79 ± 4.76	81.55 ± 1.27	70.55 ± 2.65	79.77 ± 3.02	**82.66 ± 1.41**

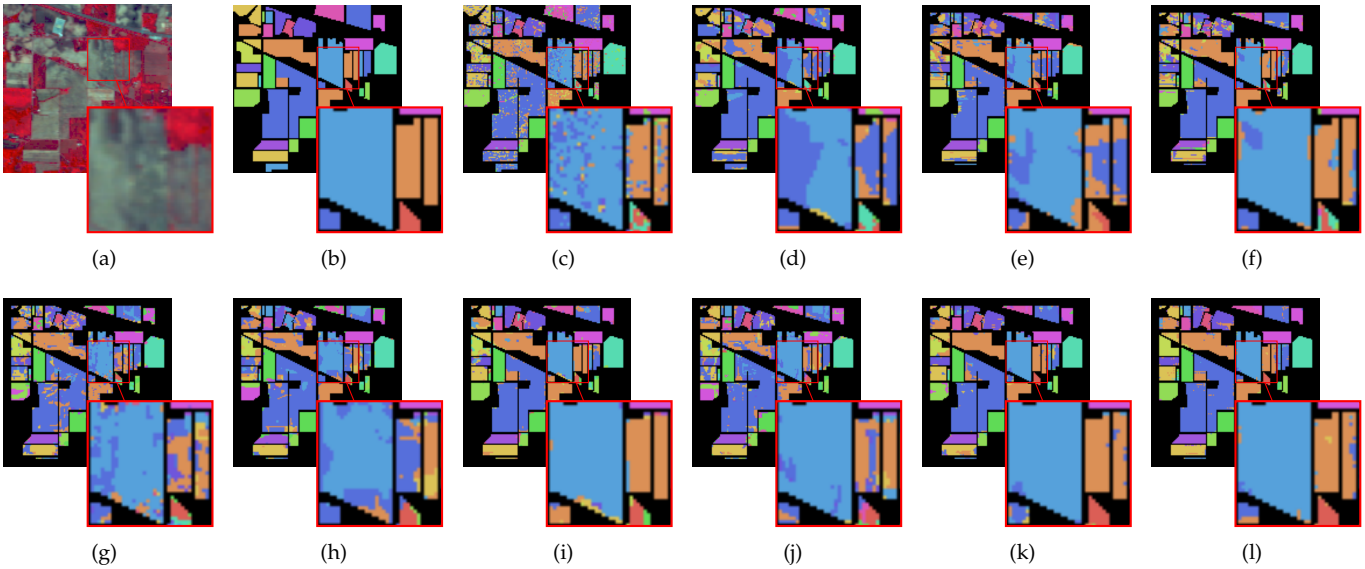


Figure 9. Visualization of the experimental results based on Indian Pines dataset. (**a**) Original image, (**b**) Ground truth, (**c**) Mou, (**d**) He, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) ViT, (**h**) CrossViT, (**i**) LeViT, (**j**) RvT, (**k**) HiT, (**l**) SS-TMNet(Ours).

4.3.3. Experimental Analysis on Houston2013 Dataset

The experimental results of our proposed SS-TSNet and compared models on the Houston2013 dataset are shown in Table 3. We can see that our model works best with OA and Kappa, reaching 96.22% and 96.22%, respectively. In addition, the standard deviation of our model is the smallest, only 0.12, indicating the stability of our model. It is worth noting that the LeViT performs much worse on this dataset than the other two datasets in the experiment with respect to OA and Kappa, only 87.36% and 86.34%, which indicates that the generalization capability of the LeViT model is relatively weak. Our model performs well on all three datasets, possibly because our SSAM models both local and global features effectively from three dimensions.

Figure 10 shows the visualization results of the experiment. To make it clearer to see the difference at the pixel level, we crop local details to show the classification map. As shown in the red rectangle in the figure, the classification map generated by our method is less noisy than comparison methods and closer to the ground truth image, which shows the superiority of our presented method. Other methods, such as HybridSN, may not

perform well because only the combination of 3D convolution and 2D convolution is used. Although it has a good model of local spatial characteristics, it lacks the dependence on the relationship between capturing long-range spectra. As for the CrossViT method, it only uses Transformer to build the network without considering the effect of convolution on classification results, which may result in unsatisfactory performance.

Table 3. The comparative experimental results on Houston2013 dataset (Bold numbers represent the best results for the corresponding category).

Class	Methods									
#	Mou	He	3D-CNN	HybridSN	ViT	CrossViT	LeViT	RvT	HiT	SS-TMNet
1	95.49 ± 0.92	95.45 ± 1.45	96.31 ± 1.91	**97.74 ± 0.72**	95.82 ± 0.84	94.36 ± 2.16	94.66 ± 1.51	97.50 ± 0.54	96.99 ± 0.87	97.60 ± 0.64
2	96.28 ± 0.68	97.04 ± 1.46	96.40 ± 1.51	97.54 ± 0.91	96.03 ± 0.98	94.91 ± 2.60	95.37 ± 2.58	98.42 ± 0.24	97.69 ± 0.52	**98.44 ± 0.56**
3	**99.97 ± 0.05**	99.03 ± 0.31	98.91 ± 0.94	99.21 ± 1.00	98.15 ± 0.65	98.59 ± 1.07	92.46 ± 8.68	99.70 ± 0.29	99.28 ± 0.69	99.50 ± 0.23
4	96.50 ± 0.97	95.59 ± 1.18	96.54 ± 1.52	**98.35 ± 0.84**	95.25 ± 0.79	97.10 ± 0.37	94.50 ± 1.50	98.32 ± 0.54	97.45 ± 0.64	97.26 ± 0.96
5	97.76 ± 0.71	95.07 ± 1.55	96.38 ± 0.63	96.72 ± 1.11	96.10 ± 0.87	96.74 ± 0.64	96.56 ± 1.46	97.63 ± 0.47	97.49 ± 0.61	**98.19 ± 0.33**
6	97.19 ± 2.86	74.68 ± 5.34	83.14 ± 5.14	**93.85 ± 2.61**	73.81 ± 5.75	88.76 ± 2.57	87.45 ± 3.28	93.30 ± 2.21	88.74 ± 3.62	93.67 ± 2.33
7	83.06 ± 0.99	90.96 ± 1.45	91.60 ± 1.81	93.78 ± 1.73	91.16 ± 1.34	94.86 ± 0.80	91.19 ± 4.61	**95.99 ± 1.41**	93.05 ± 1.08	94.54 ± 1.03
8	67.91 ± 1.94	82.25 ± 3.29	86.07 ± 2.16	90.60 ± 2.20	88.58 ± 1.38	89.61 ± 1.99	81.54 ± 8.63	94.48 ± 1.71	91.24 ± 1.97	**95.74 ± 1.35**
9	78.28 ± 1.83	83.92 ± 2.64	89.48 ± 1.81	89.02 ± 4.05	88.71 ± 1.77	92.63 ± 0.82	83.87 ± 8.50	92.34 ± 1.55	90.64 ± 1.84	**94.29 ± 1.32**
10	72.09 ± 2.47	86.58 ± 2.87	90.36 ± 1.50	92.31 ± 3.74	90.39 ± 1.23	89.33 ± 2.54	76.11 ± 12.47	94.44 ± 1.44	92.39 ± 1.92	**96.91 ± 0.81**
11	76.74 ± 1.04	85.83 ± 2.80	90.29 ± 2.17	91.84 ± 3.23	91.15 ± 1.53	91.29 ± 2.27	78.69 ± 12.96	93.75 ± 1.06	93.28 ± 1.55	**94.94 ± 0.72**
12	71.20 ± 2.04	82.39 ± 4.06	89.76 ± 2.29	91.47 ± 3.03	87.13 ± 1.52	88.22 ± 3.34	84.79 ± 7.95	93.37 ± 1.77	90.72 ± 2.25	**96.50 ± 1.00**
13	54.00 ± 5.40	83.31 ± 3.68	90.21 ± 4.11	92.38 ± 1.39	74.81 ± 4.09	80.82 ± 2.66	57.02 ± 31.62	85.68 ± 4.58	88.52 ± 2.33	**93.42 ± 1.61**
14	95.64 ± 1.02	95.41 ± 1.85	96.94 ± 2.92	96.06 ± 2.52	95.13 ± 1.43	95.03 ± 1.53	90.17 ± 7.81	99.12 ± 0.43	97.13 ± 1.62	**99.88 ± 0.19**
15	98.25 ± 0.40	96.28 ± 1.59	98.13 ± 0.83	96.02 ± 2.05	94.69 ± 2.24	97.65 ± 1.18	94.10 ± 4.05	98.20 ± 0.70	98.40 ± 1.06	**98.98 ± 0.58**
OA(%)	84.91 ± 0.51	89.61 ± 1.82	92.40 ± 1.30	93.90 ± 1.70	91.28 ± 0.69	92.61 ± 1.01	87.36 ± 5.97	95.28 ± 0.72	93.94 ± 1.02	**96.22 ± 0.35**
K(%)	83.68 ± 0.55	88.77 ± 1.97	91.79 ± 1.40	93.41 ± 1.84	90.58 ± 0.74	92.02 ± 1.09	86.34 ± 6.47	94.91 ± 0.78	93.45 ± 1.10	**95.92 ± 0.38**

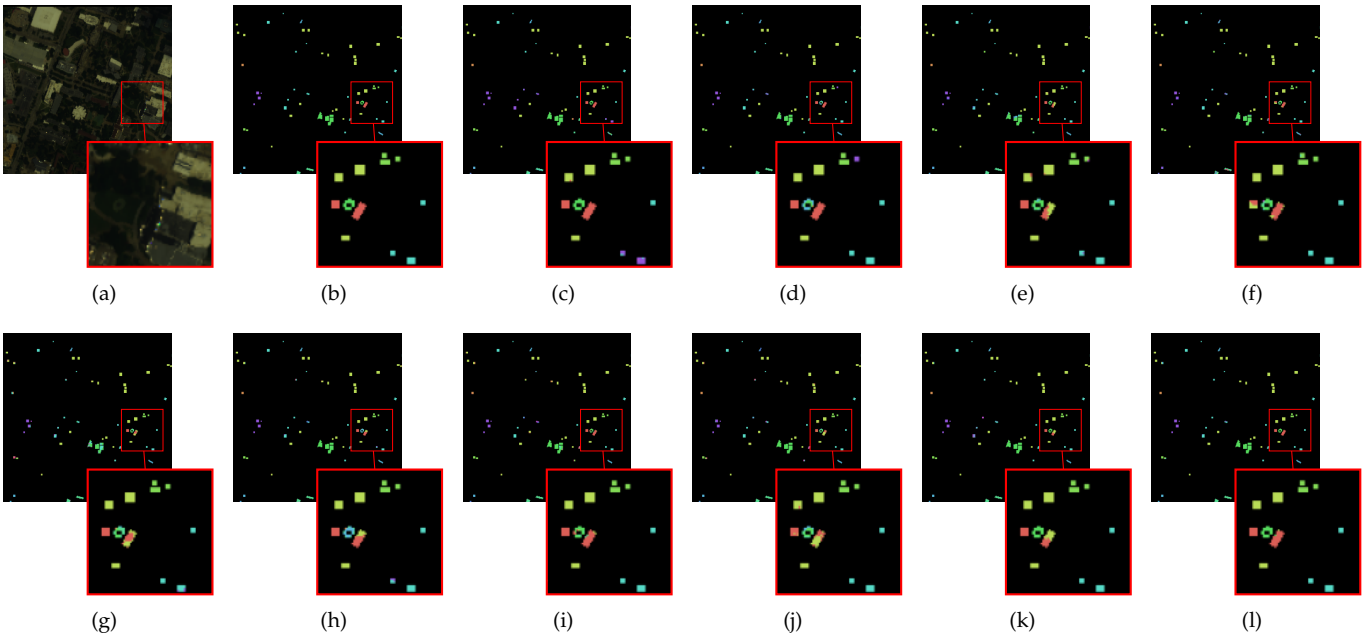


Figure 10. Visualization of the experimental results based on Houston2013 dataset. (**a**) Original image, (**b**) Ground truth, (**c**) Mou, (**d**) He, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) ViT, (**h**) CrossViT, (**i**) LeViT, (**j**) RvT, (**k**) HiT, (**l**) SS-TMNet(Ours).

4.3.4. Student's *t*-Test

We conducted a Student's *t*-test between our presented method and the compared methods with 10 times randomized initializations. We collected OA results produced by 10 randomized experiments on Pavia University, Indian Pines, and Houston2013 datasets using SS-TMNet and other comparative methods. Student's t-test method was employed to compute the *p*-value between our proposed methods and existing methods. When the

p-value is greater than 0.05, there is no significant difference between the two models. When the *p*-value is less than 0.05, the results of the two models are significantly different.

To make it easier to observe data differences, our experimental data are represented by scientific notation. As shown in Table 4, the *p*-value between the SS-TMNet and all the compared methods is less than 0.05 on the three datasets, which shows that our SS-TMNet method has significant advantages over other methods. For instance, on the Pavia University dataset, the *p*-values between the SS-TMNet method and HybridSN and HiT methods are 1.40×10^{-2} and 2.26×10^{-5}, respectively, which are less than 0.05, showing significant differences between methods.

Table 4. Student's *t*-test results between SS-TMNet and the compared methods.

Datasets	**Methods**								
#	**Mou**	**He**	**3D-CNN**	**HybridSN**	**ViT**	**CrossViT**	**LeViT**	**RvT**	**HiT**
Pavia University	7.12×10^{-7}	5.20×10^{-11}	7.45×10^{-6}	1.40×10^{-2}	1.02×10^{-11}	1.22×10^{-12}	3.68×10^{-2}	1.08×10^{-11}	2.26×10^{-5}
Indian Pines	1.77×10^{-13}	1.94×10^{-6}	4.05×10^{-8}	5.01×10^{-3}	2.80×10^{-18}	4.59×10^{-8}	4.48×10^{-2}	4.76×10^{-10}	1.78×10^{-2}
Houston2013	1.78×10^{-21}	1.16×10^{-6}	5.61×10^{-6}	2.57×10^{-3}	1.86×10^{-13}	7.50×10^{-9}	1.59×10^{-3}	2.50×10^{-3}	5.08×10^{-5}

4.4. Ablation Studies

We have performed ablation experiments on the main components of the SS-TMNet model. The results and analysis of the ablation experiments for the proposed MSCP and SSAM modules are described in the following two sections. The results in Tables 5 and 6 are the average of ten times experiments.

4.4.1. The Effectiveness of the MSCP Module

In order to verify the effectiveness of our proposed MSCP module, we used different projection methods (such as Linear, Conv2D, SACP [53], and MSCP) to project the image features without changing the subsequent module and network structure. Furthermore, SACP is the feature projection module in the HiT method. As shown in Table 5, we chose ViT as the baseline method. The experimental results show that our MSCP+SSAM showed the best performance (91.74% in OA and 89.44% in Kappa). The Mean and Std columns in the table represent the mean and standard deviation differences between our proposed SS-TMNet (MSCP+SSAM) and the comparison method. We can see that our presented method had the highest mean and lowest standard deviation, which shows that the MSCP module is more effective than the other feature extraction modules.

Table 5. Ablation study of the proposed MSCP on the Pavia University dataset (Bold numbers represent the best results).

Methods	**OA(%)**	**Mean(−)**	**Std(+)**	**Kappa(%)**	**Mean(−)**	**Std(+)**
ViT	88.92 ± 0.31	−2.28%	+0.19%	85.81 ± 0.40	−3.63%	+0.24%
Linear + SSAM	90.58 ± 0.24	−1.16%	+0.12%	87.95 ± 0.30	−1.49%	+0.14%
Conv2D + SSAM	91.53 ± 0.12	−0.21%	+0.00%	89.18 ± 0.16	−0.26%	+0.00%
SACP [53] + SSAM	91.59 ± 0.19	−0.15%	+0.03%	89.25 ± 0.24	−0.19%	+0.08%
MSCP + SSAM	**91.74 ± 0.12**	0%	0%	**89.44 ± 0.16**	0%	0%

4.4.2. The Effectiveness of the SSAM Module

In order to verify the effectiveness of the SSAM module, we took the ViT method as the baseline method and set up four groups of comparison experiments. In the experiment, the SSAM Module in our proposed method is replaced by the single linear layer connection (Linear), the convolution permutator module (ConvPermute) of the HiT method, and the ViP [58] method (ViP), respectively. The table shows that our MSCP+SSAM had the best performance (91.74 ± 0.12 in OA and 89.44 ± 0.16 in Kappa). Compared with the replaced SACP and ViP modules, our proposed method was 0.45% and 0.37% higher in the OA metric, respectively, which shows the effectiveness of our SSAM module for improving network performance.

Table 6. Ablation study of the proposed SSAM on the Pavia University dataset (Bold numbers represent the best results).

Methods	OA(%)	Mean(−)	Std(+)	Kappa(%)	Mean(−)	Std(+)
ViT	88.92 ± 0.31	−2.82%	+0.19%	85.81 ± 0.40	−3.63%	+0.24%
MSCP + Linear	90.15 ± 0.30	−1.59%	+0.18%	87.40 ± 0.38	−2.04%	+0.22%
MSCP + ConvPermute [53]	91.29 ± 0.17	−0.45%	+0.05%	88.86 ± 0.22	−0.58%	+0.06%
MSCP + ViP [58]	91.37 ± 0.40	−0.37%	+0.28%	88.97 ± 0.52	−0.47%	+0.36%
MSCP + SSAM	**91.74 ± 0.12**	0%	0%	**89.44 ± 0.16**	0%	0%

4.5. Scability

Due to the scarcity of hyperspectral image data, it is meaningful to study the influence of the number of training samples on the classification method. We changed the training samples from 10% to 50% on the Houston 2013 dataset to study the scability. Each model was run ten times, and the average value was taken as the final result. Table 7 reports the average OA of the proposed SS-TMNet and compared models. We can see that as the training samples change from 10% to 50%, the performance gradually improves, and our model always shows excellent results and high stability. It is worth noting that the experimental results of LeViT, when the training sample are 40% and 50%, are slightly higher than the model we proposed. However, LeViT performs poorly when the training samples are few, indicating its instability.

Table 7. The results of OA by the SS-TMNet method and comparison methods with different training samples on the Houston 2013 dataset (Bold numbers represent the best results for the corresponding category).

Training Sample	Methods									
#	Mou	He	3D-CNN	HybridSN	ViT	CrossViT	LeViT	RvT	HiT	SS-TMNet
10%	84.91 ± 0.51	89.6 1± 1.82	92.40 ± 1.30	93.90 ± 1.70	91.28 ± 0.69	92.61 ± 1.01	87.36 ± 5.97	95.28 ± 0.72	93.94 ± 1.02	**96.22 ± 0.35**
20%	87.77 ± 0.36	94.71 ± 1.05	95.84 ± 0.64	97.82 ± 0.28	95.59 ± 0.37	97.19 ± 0.15	97.70 ± 0.33	97.55 ± 0.22	96.96 ± 0.96	**97.98 ± 0.19**
30%	89.42 ± 0.40	96.38 ± 0.93	97.32 ± 0.28	97.92 ± 0.67	97.15 ± 0.25	98.19 ± 0.13	98.46 ± 0.17	98.27 ± 0.22	98.04 ± 0.26	**98.49 ± 0.15**
40%	90.53 ± 0.42	96.88 ± 0.90	97.88 ± 0.23	98.65 ± 0.41	97.78 ± 0.25	98.61 ± 0.16	**98.85 ± 0.11**	98.63 ± 0.11	98.43 ± 0.30	98.79 ± 0.11
50%	91.48 ± 0.35	97.59 ± 0.38	98.40 ± 0.15	98.76 ± 0.18	98.24 ± 0.26	98.84 ± 0.13	**98.98 ± 0.07**	98.82 ± 0.11	98.54 ± 0.29	98.88 ± 0.13

Moreover, to study the experimental results of our SS-TMNet method on several datasets that vary with the number of training samples, we tested SS-TMNet on three datasets. The experiment also adopted the average of 10 results as the final result. The experimental visualization results of the OA metric are shown in Figure 11. With the increase of training samples, OA gradually increases and eventually tends to be stable, which effectively proves the proposed method's stability.

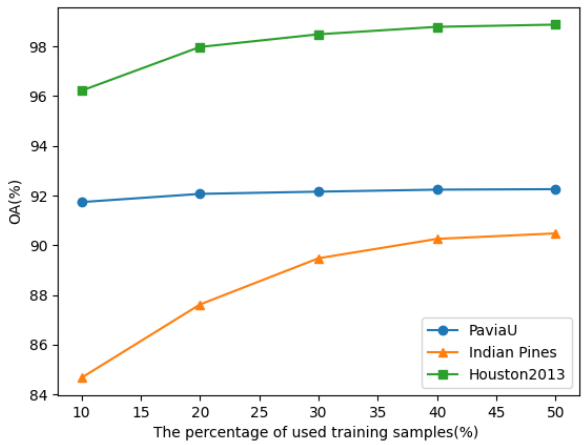


Figure 11. The OA results of the proposed SS-TMNet on three datasets with a varying number of training samples.

5. Conclusions

This work presents a novel HSI classification Transformer-based method (SS-TMNet) to improve HSI classification, which can fully use the spatial–spectral information in HSI data. SS-TMNet includes two key modules: the MSCP module and the SSAM module. The MSCP module uses multi-scale 3D convolution to extract the fused spatial–spectral features. The SSAM module extracts features through height dimension, width dimension, and spectral dimension, which can more effectively obtain local and global feature information. We compared our proposed method with the most advanced Transformer-based and CNN-based methods on three benchmark HSI datasets. Experimental results show that our SS-TMNet method performs the best overall accuracy on three datasets.

In future work, we plan to study more efficient HSI classification methods based on Transformer by embedding convolution neural networks into Transformer more effectively. For the scarcity problem of labeled HSI, we plan to study transfer learning and self-supervised learning based on SS-TMNet to improve the performance of classification of limited training samples.

Author Contributions: Conceptualization, review and editing, X.H.; write the original draft preparation and correct it, Y.Z.; methodology and correct this paper, X.Y.; data curation and correct this paper, X.Z. and K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant No. 62062033.

Data Availability Statement: The Pavia University and Indian Pines datasets used in our work are publicly available. The Pavia University dataset is available at: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_University_scene, accessed on 15 September 2022. The Indian Pines dataset is available at: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Indian_Pines, accessed on 15 September 2022. The Huston2013 dataset is provided by the Society for Geosciences and Remote Sensing Data Fusion Competition: https://hyperspectral.ee.uh.edu/?page_id=459, accessed on 15 September 2022. The Houston2013 dataset can be obtained through this competition.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
2. Zhan, T.; Song, B.; Sun, L.; Jia, X.; Wan, M.; Yang, G.; Wu, Z. TDSSC: A three-directions spectral–spatial convolution neural network for hyperspectral image change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 377–388. [CrossRef]
3. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.; Wu, X.; Yao, J.; Khan, A.M.; Mazzara, M.; Distefano, S.; Chanussot, J. Hyperspectral image classification—Traditional to deep models: A survey for future prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 968–999. [CrossRef]
4. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
5. Samaniego, L.; Bárdossy, A.; Schulz, K. Supervised classification of remotely sensed imagery using a modified *k*-NN technique. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2112–2125. [CrossRef]
6. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098. [CrossRef]
7. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
9. Algan, G.; Ulusoy, I. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl.-Based Syst.* **2021**, *215*, 106771. [CrossRef]
10. Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**.

11. Zhao, Z.Q.; Zheng, P.; Xu, S.t.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
12. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef] [PubMed]
13. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3173–3184. [CrossRef]
14. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
15. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2094–2107. [CrossRef]
16. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, Virtual, 3–7 May 2021.
19. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [CrossRef]
20. Hao, J.; Dong, F.; Li, Y.; Wang, S.; Cui, J.; Zhang, Z.; Wu, K. Investigation of the data fusion of spectral and textural data from hyperspectral imaging for the near geographical origin discrimination of wolfberries using 2D-CNN algorithms. *Infrared Phys. Technol.* **2022**, *125*, 104286. [CrossRef]
21. He, M.; Li, B.; Chen, H. Multi-Scale 3D Deep Convolutional Neural Network for Hyperspectral Image Classification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3904–3908.
22. Fang, B.; Liu, Y.; Zhang, H.; He, J. Hyperspectral Image Classification Based on 3D Asymmetric Inception Network with Data Fusion Transfer Learning. *Remote Sens.* **2022**, *14*, 1711. [CrossRef]
23. Chang, Y.L.; Tan, T.H.; Lee, W.H.; Chang, L.; Chen, Y.N.; Fan, K.C.; Alkhaleefah, M. Consolidated Convolutional Neural Network for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 1571. [CrossRef]
24. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. Deepvit: Towards Deeper Vision Transformer. *arXiv* **2021**, arXiv:2103.11886.
25. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
26. Yu, D.; Li, Q.; Wang, X.; Zhang, Z.; Qian, Y.; Xu, C. DSTrans: Dual-Stream Transformer for Hyperspectral Image Restoration. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 3739–3749.
27. Li, J.; Xing, H.; Ao, Z.; Wang, H.; Liu, W.; Zhang, A. Convolution-Transformer Adaptive Fusion Network for Hyperspectral Image Classification. *Appl. Sci.* **2023**, *13*, 492. [CrossRef]
28. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [CrossRef]
29. Wang, Y.; Jiang, S.; Xu, M.; Zhang, S.; Jia, S. A Center-Masked Convolutional Transformer for Hyperspectral Image Classification. In Proceedings of the 31st International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; Volume 3207, pp. 1–6.
30. Zhang, Y.; Wang, X.; Jiang, X.; Zhou, Y. Marginalized graph self-representation for unsupervised hyperspectral band selection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5516712. [CrossRef]
31. Ding, Y.; Zhang, Z.; Zhao, X.; Hong, D.; Cai, W.; Yu, C.; Yang, N.; Cai, W. Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification. *Neurocomputing* **2022**, *501*, 246–257. [CrossRef]
32. Zhang, Z.; Ding, Y.; Zhao, X.; Siye, L.; Yang, N.; Cai, Y.; Zhan, Y. Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification. *Expert Syst. Appl.* **2023**, *217*, 119508.
33. Zhang, Y.; Wang, Y.; Chen, X.; Jiang, X.; Zhou, Y. Spectral–Spatial Feature Extraction With Dual Graph Autoencoder for Hyperspectral Image Clustering. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8500–8511. [CrossRef]
34. Ding, Y.; Zhang, Z.; Zhao, X.; Hong, D.; Li, W.; Cai, W.; Zhan, Y. AF2GNN: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification. *Inf. Sci.* **2022**, *602*, 201–219. [CrossRef]
35. Ding, Y.; Zhang, Z.; Zhao, X.; Cai, W.; Yang, N.; Hu, H.; Huang, X.; Cao, Y.; Cai, W. Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536716. [CrossRef]
36. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5518615. [CrossRef]

37. He, X.; Chen, Y.; Li, Q. Two-Branch Pure Transformer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6015005. [CrossRef]
38. Feng, J.; Luo, X.; Li, S.; Wang, Q.; Yin, J. Spectral Transformer with Dynamic Spatial Sampling and Gaussian Positional Embedding for Hyperspectral Image Classification. In Proceedings of the International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 3556–3559.
39. Ding, Y.; Zhang, Z.; Zhao, X.; Cai, Y.; Li, S.; Deng, B.; Cai, W. Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536016. [CrossRef]
40. Rakotomamonjy, A.; Bach, F.; Canu, S.; Grandvalet, Y. SimpleMKL. *J. Mach. Learn. Res.* **2008**, *9*, 2491–2521.
41. Dalla Mura, M.; Atli Benediktsson, J.; Waske, B.; Bruzzone, L. Extended profiles with morphological attribute filters for the analysis of hyperspectral data. *Int. J. Remote Sens.* **2010**, *31*, 5975–5991. [CrossRef]
42. Li, J.; Marpu, P.R.; Plaza, A.; Bioucas-Dias, J.M.; Benediktsson, J.A. Generalized Composite Kernel Framework for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4816–4829. [CrossRef]
43. Bandos, T.V.; Bruzzone, L.; Camps-Valls, G. Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 862–873. [CrossRef]
44. Villa, A.; Benediktsson, J.A.; Chanussot, J.; Jutten, C. Hyperspectral Image Classification With Independent Component Discriminant Analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4865–4876. [CrossRef]
45. Licciardi, G.; Marpu, P.R.; Chanussot, J.; Benediktsson, J.A. Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data Based on the Extended Morphological Profiles. *IEEE Geosci. Remote Sens. Lett.* **2011**, *9*, 447–451. [CrossRef]
46. Zhang, Q.; Tian, Y.; Yang, Y.; Pan, C. Automatic spatial–spectral feature selection for hyperspectral image via discriminative sparse multimodal learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 261–279. [CrossRef]
47. Jouni, M.; Dalla Mura, M.; Comon, P. Hyperspectral image classification based on mathematical morphology and tensor decomposition. *Math.-Morphol.-Theory Appl.* **2020**, *4*, 1–30. [CrossRef]
48. Luo, F.; Huang, H.; Duan, Y.; Liu, J.; Liao, Y. Local geometric structure feature for dimensionality reduction of hyperspectral imagery. *Remote Sens.* **2017**, *9*, 790. [CrossRef]
49. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
50. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. Levit: A Vision Transformer in Convnet's Clothing for Faster Inference. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12259–12269.
51. Chen, C.F.R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
52. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
53. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral Image Transformer Classification Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [CrossRef]
54. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
55. Sharma, V.; Diba, A.; Tuytelaars, T.; Van Gool, L. *Hyperspectral CNN for Image Classification & Band Selection, with Application to Face Recognition*; Technical Report KUL/ESAT/PSI/1604, KU Leuven; ESAT: Leuven, Belgium, 2016.
56. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281. [CrossRef]
57. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11936–11945.
58. Hou, Q.; Jiang, Z.; Yuan, L.; Cheng, M.M.; Yan, S.; Feng, J. Vision Permutator: A Permutable MLP-Like Architecture for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1328–1334. [CrossRef]