remote sensing

MDPI

Article

Spectral-Swin Transformer with Spatial Feature Extraction Enhancement for Hyperspectral Image Classification

Yinbin Peng [1], Jiansi Ren [1,2,*], Jiamei Wang [1] and Meilin Shi [1]

[1] School of Computer Science, China University of Geosciences, Wuhan 430078, China
[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China
* Correspondence: renjsv@cug.edu.cn

Abstract: Hyperspectral image classification (HSI) has rich applications in several fields. In the past few years, convolutional neural network (CNN)-based models have demonstrated great performance in HSI classification. However, CNNs are inadequate in capturing long-range dependencies, while it is possible to think of the spectral dimension of HSI as long sequence information. More and more researchers are focusing their attention on transformer which is good at processing sequential data. In this paper, a spectral shifted window self-attention based transformer (SSWT) backbone network is proposed. It is able to improve the extraction of local features compared to the classical transformer. In addition, spatial feature extraction module (SFE) and spatial position encoding (SPE) are designed to enhance the spatial feature extraction of the transformer. The spatial feature extraction module is proposed to address the deficiency of transformer in the capture of spatial features. The loss of spatial structure of HSI data after inputting transformer is supplemented by proposed spatial position encoding. On three public datasets, we ran extensive experiments and contrasted the proposed model with a number of powerful deep learning models. The outcomes demonstrate that our suggested approach is efficient and that the proposed model performs better than other advanced models.

Keywords: transformer; shifted window; spatial feature extraction (SFE); spatial position encoding (SPE); hyperspectral image (HSI) classification

Citation: Peng, Y.; Ren, J.; Wang, J.; Shi, M. Spectral-Swin Transformer with Spatial Feature Extraction Enhancement for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 2696. https://doi.org/10.3390/rs15102696

Academic Editor: Gwanggil Jeon

Received: 15 April 2023
Revised: 14 May 2023
Accepted: 20 May 2023
Published: 22 May 2023

1. Introduction

Because of the rapid advancement of hyperspectral sensors, the resolution and accuracy of hyperspectral images (HSI) have also increased greatly. HSI contains a wealth of spectral information, collecting hundreds of bands of electron spectrum at each pixel. Its rich information allows for excellent performance in classifying HSI, and thus its application has great potential in several fields such as precision agriculture [1] and Jabir et al. [2] used machine learning algorithm for weed detection, medical imaging [3], object detection [4], urban planning [5], environment monitoring [6], mineral exploration [7], dimensionality reduction [8] and military detection [9].

Numerous conventional machine learning methods have been used to the classification of HSI in the past decade or so, such as K-nearest neighbors (KNN) [10], support vector machines (SVM) [11–14], random forests [15,16]. Navarro et al. [17] used neural network for hyperspectral image segmentation. However, as the size and complexity of the training set increases, the fitting ability of traditional methods can show weakness for the task, and the performance often encounters bottlenecks. Song et al. [18] proposed a HSI classification method based on the sparse representation of KNN, but it cannot effectively apply the spatial information in HSI. Guo et al. [19] used a fused SVM of spectral and spatial features for HSI classification, but it is still difficult to extract important features from high-dimensional HSI data. Deep learning have developed rapidly in recent years, and their powerful fitting ability can extract features from multivariate data. Inspired by

this, the designed deep learning models have proposed in HSI classification tasks, such as recurrent neural network (RNN) [20–22], convolutional neural network (CNN) [23–28], graph convolutional network (GCN) [29,30], capsule network (CapsNet) [31,32], long short term memory (LSTM) networks [33–35]. Although these deep learning models show good performance in several different domains, they have certain shortcomings in HSI classification tasks.

For CNNs, which are good at natural image tasks, Its benefit is that the image's spatial information can be extracted during the convolution operation. HSI-CNN [36] stacks multi-dimensional data from HSI into two-dimensional data and then extracts features efficiently. 2D-CNN [37] can capture spatial features in HSI data to improve classification accuracy. However, HSI has rich information in the spectral dimension, and if it is not exploited, the performance of the model is bound to be difficult to break through. Although the advent of 3D-CNN [38–41] enables the extraction of both spatial and spectral features, the convolution operation is localized, so the extracted features lack the mining and representation of the global Information.

Recently, transformer has evolved rapidly and shown good performance when performing tasks like natural language processing. Based on its self-attention mechanism, it is very good at processing long sequential information and extracting global relations. Vision transformer (ViT) [42] makes it perform well in several vision domains by dividing images into patches and then inputting them into the model. Swin-transformer [43] enhances the capability of local feature extraction by dividing the image into windows and performing multi-head self-attention (MSA) separately within the windows, and then enabling the exchange of information between the windows by shifting the windows. It improves the accuracy in natural image processing tasks and effectively reduces the computational effort in the processing of high-resolution images. Due to transformer's outstanding capabilities for natural image processing, more and more studies are applying it to the classification of HSI [44–50]. However, if ViT is applied directly to the HSI classification, there will be some problems that will limit the performance improvement, specifically as follows.

(1) The transformer performs well at handling sequence data(spectral dimension information), but lacks the use of spatial dimension information.
(2) The multi-head self-attention (MSA) of transformer is adept at resolving the global dependencies of spectral information, but it is usually difficult to capture the relationships for local information.
(3) Existing transformer models usually map the image to linear data to be able to input into the transformer model. Such an operation would destroy the spatial structure of HSI.

HSI can be regarded as a sequence in the spectral dimension, and the transform is effective at handling sequence information, so the transformer model is suitable for HSI classification. The research in this paper is based on tranformer and considers the above mentioned shortcomings to design a new model, called spectral-swin transformer (SSWT) with spatial feature extraction enhancement, and apply it in HSI classification. Inspired by swin-transformer and the characteristics of HSI data which contain a great deal of information in the spectral dimension, we design a method of dividing and shifting windows in the spectral dimension. MSA is performed within each window separately, aiming to improve the disadvantage of transformer to extract local features. We also design two modules to enhance model's spatial feature extraction. In summary, the following are the contributions of this paper.

(1) Based on the characteristics of HSI data, a spectral dimensional shifted window multi-head self-attention is designed. It enhances the model's capacity to capture local information and can achieve multi-scale effect by changing the size of the window.
(2) A spatial feature extraction module based on spatial attention mechanism is designed to improve the model's ability to characterize spatial features.
(3) A spatial position encoding is designed before each transformer encoder to deal with the lack of spatial structure of the data after mapping to linear.

(4) Three publicly accessible HSI datasets are used to test the proposed model, which is compared with advanced deep learning models. The proposed model is extremely competitive.

The rest of this paper is organized as follows: Section 2 discusses the related work on HSI classification using deep learning, which includes transformer. Section 3 describes the proposed model and the design method for each component. Section 4 presents the three HSI datasets, as well as the experimental setup, results, corresponding analysis. Section 5 concludes with a summary and outlook of the full paper.

2. Related Work

2.1. Deep-Learning-Based Methods for HSI Classification

Deep learning has developed quickly, more and more researchers are using deep learning methods(e.g., RNNs, CNNs, GCNs, CapsNet, LSTM) to the classification tasks of HSI [20,22,23,29–31,33,34]. Mei et al. [51] constructed a network based on bidirectional long short-term memory (Bi-LSTM) for HSI classification. Zhu et al. [52] proposed an end-to-end residual spectral–spatial attention network (RSSAN) for HSI classification, which consists of spectral and spatial attention modules for spectral band and spatial information adaptive selection. Song et al. [53] created a deep feature fusion network (DFFN) to solve the negative effects of excessively increasing network depth.

Due to CNN's excellent capability of taking the local spatial context information and it's outstanding capabilities in natural picture processing, many CNN-based HSI classification models have emerged. For example, Hang et al. [54] proposed two CNN sub-networks based on the attention mechanism for extracting the spectral and spatial features of HSI, respectively. Chakraborty et al. [55] designed a wavelet CNN that uses layers of wavelet transforms to display spectral features. Gong et al. [56] proposed a hybrid model that combines 2D-CNN and 3D-CNN in order to include more in-depth spatial and spectral features while using fewer learning samples. Hamida et al. [57] introduced a new 3-D DL method that permits the processing of both spectral and spatial information simultaneously.

However, each of these deep learning approaches has some respective drawbacks that can limit the model performance when processing HSI classification tasks. For CNN, it is good at handling two-dimensional spatial features, but since the data of HSI is stereoscopic and contains a large amount of information in the spectral dimension. It's possible that CNN will have trouble extracting the spectral features. Moreover, although CNNs have achieved good results by relying on their local feature focus, the inability to deal with global dependencies limits their performance when processing spectral information in the form of long sequences. These shortcomings will be addressed in the transformer.

2.2. Vision Transformers for Image Classification

With the increasing use of transformers in computer vision, researchers have begun to consider images in terms of sequential data, such as ViT [42] and Swin-transformer [43] etc. Fang et al. [58] proposed MSG-Transformer, which presents a specialized token in each region as a messenger (MSG). Information can be transmitted flexibly among areas and computational cost is decreased by manipulating these MSG tokens. Guo et al. [59] proposed CMT, which combines the advantages of CNN and ViT, a new hybrid transformer-based network that captures long-range dependencies using transformers and extracts local information using CNN. Chen et al. [60] designed MobileNet and transformer in parallel, connected in the middle by a two-way bridge. This structure benefits from MobileNet for local processing and Transformer for global communication.

An increasing number of researchers are applying transformer to HSI classification tasks. Hong et al. [44] proposed a model called SpectralFormer (SF) for HSI classification, which divides neighboring bands into the same token for learning features and connects encoder blocks across layers, but the spatial information in HSI was not considered. Sun et al. [45] proposed the Spectral-Spatial Feature Tokenization Transformer (SSFTT) to capture high-level semantic information and spectral-spatial features, resulting in a large

performance improvement. Ayas et al. [61] designs a spectal-swin module in front of the swin transformer, which extracts spatial and spectral features and fuses them with Conv 2-D operation and Conv 3-D operation, respectively. Mei et al. [47] proposed the Group-Aware Hierarchical Transformer (GAHT) to restrict the MSA to a local spatial-spectral range by using a new group pixel embedding module, which enables the model to have improved capability of local feature extraction. Yang et al. [46] proposed a hyperspectral image transformer (HiT) classification network that captures subtle spectral differences and conveys local spatial context information by embedding convolutional operations in the transformer structure, however it is not effective in capturing local spectral features. Transformer is increasingly used in the field of HSI classification and we believe it has great potential for the future.

3. Methodology

In this section, we will introduce the proposed spectral-swin transformer (SSWT) with spatial feature extraction enhancement, which will be described in four aspects: the overall architecture, spatial feature extraction module(SFE), spatial position encoding(SPE), and spectral swin-transformer module.

3.1. Overall Architecture

In this paper, we design a new transformer-based method SSWT for the HSI classification. SSWT consists of two major Components for solving the challenges in HSI classification, namely, spatial feature extraction module(SFE) and spectral swin(S-Swin) transformer module. An overview of the proposed SSWT for the HSI classification is shown in Figure 1. The input to the model is a patch of HSI. the data is first input to SFE to perform initial spatial feature extraction, the module consists of convolution layers and spatial attention. In Section 3.2, it is explained in further detail. The data is then flattened and entered into the s-swin transformer module. A spatial position encoding is added in front of each s-swin transformer layer to add spatial structure to the data. This part will be described in Section 3.3. The s-swin transformer module uses the spectral-swin self attention, which will be introduced in Section 3.4. The final classification results are obtained by linear layers.

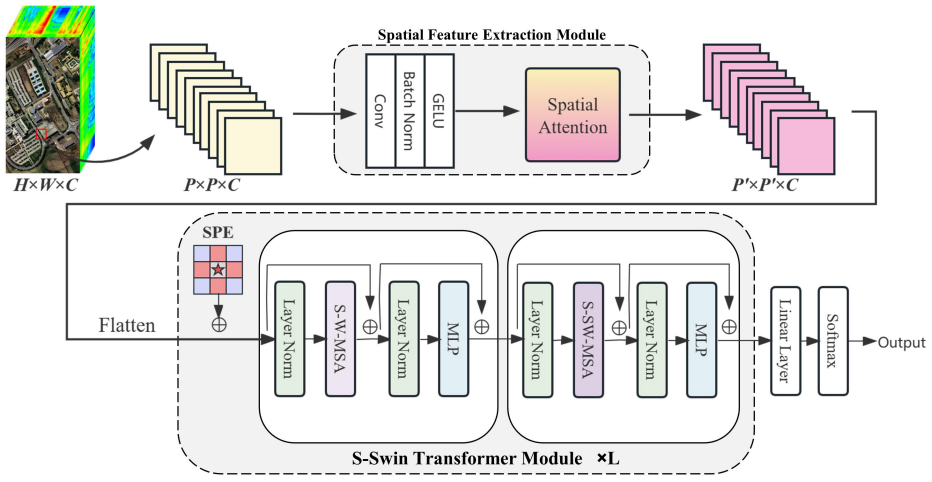


Figure 1. Overall structure of the proposed SSWT model for HSI classification.

3.2. Spatial Feature Extraction Module

Due to transformer's lack of ability in handling spatial information and local features, we designed a spatial feature extraction (SFE) module to compensate. It consists of two parts, the first one consists of convolutional layers to preliminary extraction of spatial features and batch normalization to prevent overfitting. The second part is a spatial

attention mechanism, which aims to enable the model to learn the important spatial locations in the data. The structure of SFE is shown in Figure 1.

For the input HSI patch cube $I \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ is the spatial size and C is the number of spectral bands. Each pixel space in I consists of C spectral dimensions and forms a one-hot category vector $S = [s_1, s_2, s_3, \cdots, s_n] \in \mathbb{R}^{1 \times 1 \times n}$, where n is the number of ground object classes.

Firstly, the spatial features of HSI are initially extracted by CNN layers, and the formula is shown as follows:

$$X = GELU\Big(BN\Big(Conv(I)\Big)\Big) \tag{1}$$

where $Conv(\cdot)$ represents the convolution layer. $BN(\cdot)$ represents batch normalization. $GELU(\cdot)$ denotes the activation function. The formula for the convolution layer is shown below:

$$Conv(I) = \mathop{||}_{j=0}^{J} (I * W_j^{r1 \times r2} + b_j) \tag{2}$$

where I is the input, J is the number of convolution kernels, $W_j^{r1 \times r2}$ is the jth convolution kernel with the size of $r1 \times r2$, and b_j is the jth bias. $||$ denotes concatenation, and $*$ is convolution operation.

Then, the model may learn important places in the data thanks to a spatial attention mechanism (SA). The structure of SA is shown in Figure 2. For an intermediate feature map $X \in \mathbb{R}^{H' \times W' \times C}$ ($H' \times W'$ is the spatial size of X), the process of SA is shown in the following formula:

$$S_M = MaxPooling(X) \tag{3}$$

$$S_A = AvgPooling(X) \tag{4}$$

$$X_{SA} = \sigma\Big(Conv\Big(Concat\big(S_M, S_A\big)\Big)\Big) \otimes X \tag{5}$$

MaxPooling and AvgPooling are global maximum pooling and global average pooling along the channel direction. Concat denotes concatenation in the channel direction. σ is activation function. $\otimes$ denotes the elementwise multiplication.

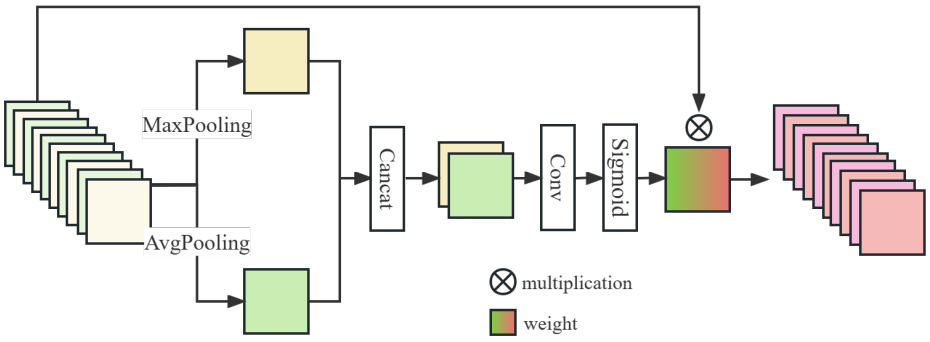


Figure 2. The structure of the spatial attention in SFE.

3.3. Spatial Position Encoding

The HSI of the input transformer is mapped to linear data, which can damage the spatial structure of HSI. To describe the relative spatial positions between pixels and to maintain the rotational invariance of samples, a spatial position encoding (SPE) is added before each transformer module.

The input to HSI classification is a patch of a region, but only the label of the center pixel is the target of classification. The surrounding pixels can provide spatial information for the classification of center pixel, and their importance tends to decrease with the distance

to the center. SPE is to learn such a center-important position encoding. The pixel positions of a patch is defined as follows.

$$pos(x_i, y_i) = |x_i - x_c| + |y_i - y_c| + 1 \tag{6}$$

where (x_c, y_c) denotes the coordinate of central position of the sample, that is the pixel to be classified. (x_i, y_i) denotes the coordinates of other pixels in the sample. The visualization of SPE when the spatial size of the sample is 7×7 can be seen in Figure 3. The pixel in the central position is unique and most important, and the other pixels are given different position encoding depending on the distance from the center.

To flexibly represent the spatial structure in HSI, the learnable position encoding are embedded in the data:

$$Y = X + spe(P) \tag{7}$$

where X is the HSI data, and P represents the position matrix (like Figure 3) constructed according to Equation (6). $spe(\cdot)$ is a learnable array that takes the position matrix as a subscript to get the final spatial position encoding. Finally, the position encoding is added to the HSI data.

7	6	5	4	5	6	7
6	5	4	3	4	5	6
5	4	3	2	3	4	5
4	3	2	1	2	3	4
5	4	3	2	3	4	5
6	5	4	3	4	5	6
7	6	5	4	5	6	7

Figure 3. SPE in a sample with the spatial size is 7×7.

3.4. Spectral Swin-Transformer Module

The structure of the spectral swin-transformer (S-SwinT) module is shown in Figure 1. Transformer is good at processing long dependencies and lacks the ability to extract local features. Inspired by swin-transformer [43], window-based multi-head self-attention (MSA) is used in our model. Because the input of HSI is a patch which is usually small in spatial size, it cannot divide the window in space as Swin-T does. Considering the rich data of HSI in the spectral dimension, a window of spectral shift was designed for MSA, called spectral window multi-head self-attention (S-W-MSA) and spectral shifted window multi-head self-attention (S-SW-MSA). MSA within windows can effectively improve local feature capturing, and window shifting allows information to be exchanged in the neighboring windows. MSA can be expressed by the following formula:

$$Z = Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{8}$$

$$\psi = Concat(Z_1, Z_2, \cdots, Z_h)W \tag{9}$$

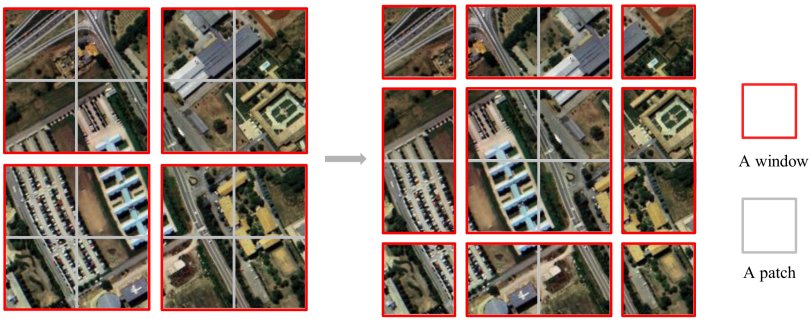
Q, K, V are matrices mapped from the input matrices called queries, keys and values. d_K is the dimension of K. The attention scores are calculated from Q and K. h is the head number of MSA, W denotes the output mapping matrix., and ψ represents the output of MSA.

As shown in Figure 4, the size of input is assumed to be $H \times W \times C$, where $H \times W$ is the space size and C is the number of spectral bands. Given that all windows' size is set to

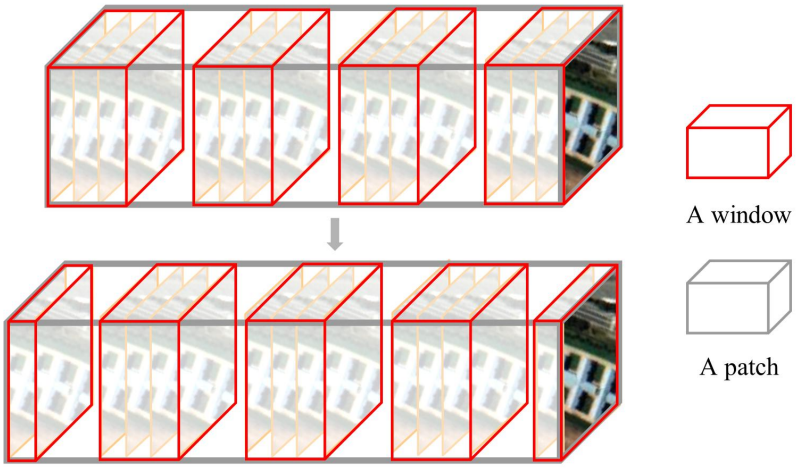
$C/4$, the window is divided uniformly for the spectral dimension. The size of each window after division is $[C/4, C/4, C/4, C/4]$. Then MSA is performed in each window. Next the window is moved half a window in the spectral direction, The size of each window at this point is $[C/8, C/4, C/4, C/4, C/8]$. MSA is again performed in each window. Wherefore, the process of S-W-MSA with m windows is:

$$Y^{(m)} = \left[\psi(y^{(1)}) \oplus \psi(y^{(2)}) \oplus, \cdots, \oplus \psi(y^{(m)})\right] \tag{10}$$

where $\oplus$ means concat, $y^{(i)}$ is the data of the i-th window.



(a)



(b)

Figure 4. The structure of (**a**) S(W)-MSA of SwinT and (**b**) S-(S)W-MSA of SSWT (ours).

Compared to SwinT, the other components of the S-SwinT module remain the same except for the design of the window, such as MLP, layer normalization (LN) and residual connections. Figure 1 describes two nearby S-SwinT modules in each stage, which can be represented by the following formula.

$$\hat{Y}^{l} = \text{S-W-MSA}(\text{LN}(Y^{l-1})) + Y^{l-1} \tag{11}$$

$$Y^{l} = \text{MLP}(\text{LN}(\hat{Y}^{l})) + \hat{Y}^{l} \tag{12}$$

$$\hat{Y}^{l+1} = \text{S-SW-MSA}(\text{LN}(Y^{l})) + Y^{l} \tag{13}$$

$$Y^{l+1} = \text{MLP}(\text{LN}(\hat{Y}^{l+1})) + \hat{Y}^{l+1} \tag{14}$$

where S-W-MSA and S-SW-MSA denote the spectral window based and spectral shifted window based MSA, $\hat{Y}^l$ and Y^l are the outputs of S-(S)W-MSA and MLP in block l.

4. Experiment

In this section, we conducted extensive experiments on three benchmark datasets to demonstrate the effectiveness of the proposed method, including Pavia University (PU), Salinas (SA) and Houston2013 (HU).

4.1. Dataset

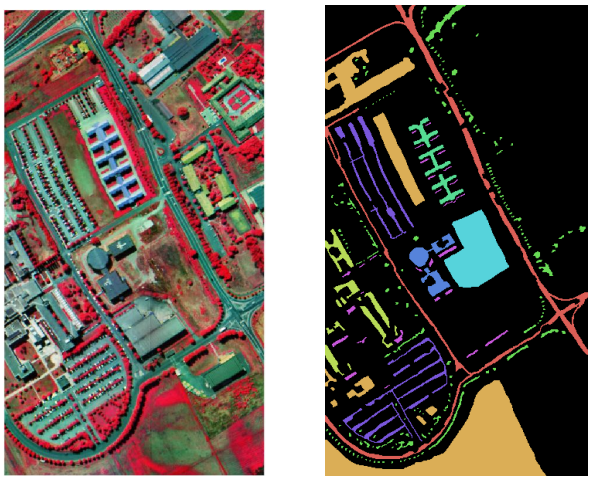
The three datasets that utilised in the experiments are detailed here.

(1) Pavia University:The Reflective Optics System Imaging Spectrometer (ROSIS) sensor acquired the PU dataset in 2001. It comprises 115 spectral bands with wavelengths ranging from 380 to 860 nm. Following the removal of the noise bands, there are now 103 open bands for investigation. The image measures 610 pixels in height and 340 pixels in width. The collection includes 42,776 labelled samples of 9 different land cover types.

(2) Salinas: The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor acquired the SA dataset in 1998. The 224 bands in the original image have wavelengths between 400 and 2500 nm. 204 bands are used for evaluating after the water absorption bands have been removed. The data has 512 and 217 pixels of height and width, respectively. There are 16 object classes represented in the dataset's 54,129 marked samples.

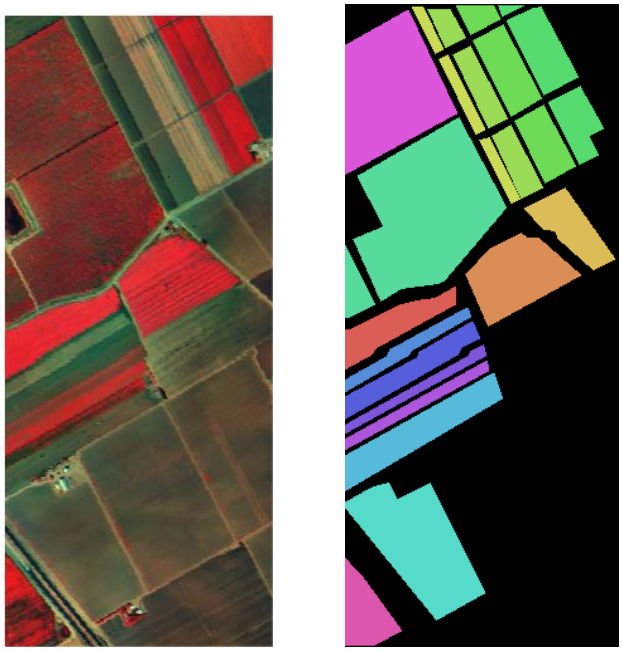
(3) Houston2013: The Hyperspectral Image Analysis Group and the NSF-funded Airborne Laser Mapping Center (NCALM) at the University of Houston in the US provided the Houston 2013 dataset. The 2013 IEEE GRSS Data Fusion Competition used the dataset initially for scientific research. It has 144 spectral bands with wavelengths between 0.38 and 1.05 m. This dataset contains 15 classes and measures 349 $\times$ 1905 pixels with a 2.5 m spatial resolution.

We divided the label samples in different ways for each dataset. Tables 1–3 provide specifics on the number of each class for the three dataset training, validation, and testing sets. False-color map and ground-truth map of three datasets are shown in Figures 5–7.



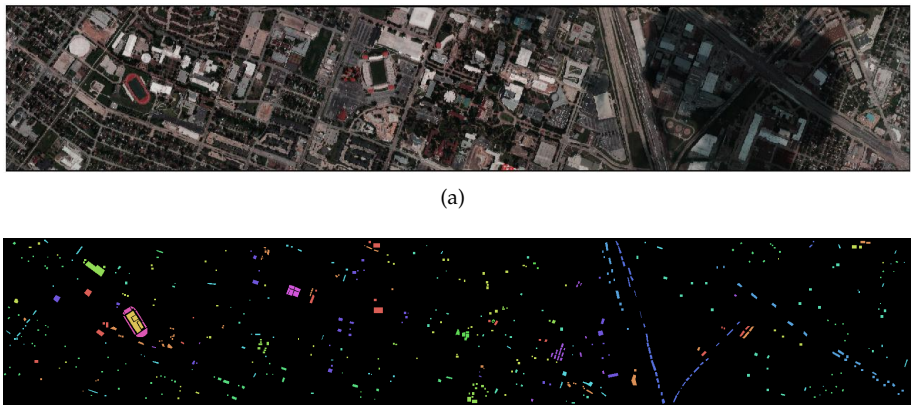
(a) (b)

Figure 5. Visualization of PU Datasets. (**a**) False-color map. (**b**) Ground-truth map.



(a) (b)

Figure 6. Visualization of SA Datasets. (**a**) False-color map. (**b**) Ground-truth map.



(a)

(b)

Figure 7. Visualization of HU Datasets. (**a**) False-color map. (**b**) Ground-truth map.

Table 1. Number of training, validation and testing samples for the PU dataset.

No.	Name	Train.	Val.	Test.
1	Asphalt	83	83	6465
2	Meadows	233	233	18,183
3	Gravel	26	26	2047
4	Trees	38	38	2987
5	Painted metal sheets	17	17	1311
6	Bare Soil	63	63	4903
7	Bitumen	17	17	1297
8	Self-Blocking Bricks	46	46	3590
9	Shadows	12	12	923
-	Total	535	535	41,706

Table 2. Number of training, validation and testing samples for the SA dataset.

No.	Name	Train.	Val.	Test.
1	Brocoli_green_weeds_1	25	25	1959
2	Brocoli_green_weeds_2	47	47	3633
3	Fallow	25	25	1927
4	Fallow_rough_plow	17	17	1358
5	Fallow_smooth	33	33	2611
6	Stubble	49	49	3860
7	Celery	45	45	3490
8	Grapes_untrained	141	141	10,989
9	Soil_vinyard_develop	78	78	6048
10	Corn_senesced_green_weeds	41	41	3196
11	Lettuce_romaine_4wk	13	13	1041
12	Lettuce_romaine_5wk	24	24	1879
13	Lettuce_romaine_6wk	11	11	893
14	Lettuce_romaine_7wk	13	13	1043
15	Vinyard_untrained	91	91	7086
16	Vinyard_vertical_trellis	23	23	1762
-	Total	676	676	52,775

Table 3. Number of training, validation and testing samples for the HU dataset.

No.	Name	Train.	Val.	Test.
1	Healthy grass	31	31	1188
2	Stressed grass	31	31	1191
3	Synthetic grass	17	17	662
4	Trees	31	31	1182
5	Soil	31	31	1180
6	Water	8	8	309
7	Residential	32	32	1205
8	Commercial	31	31	1182
9	Road	31	31	1189
10	Highway	31	31	1166
11	Railway	31	31	1173
12	Parking Lot 1	31	31	1171
13	Parking Lot 2	12	12	446
14	Tennis Court	11	11	407
15	Running Track	17	17	627
-	Total	376	376	14,278

4.2. Experimental Setting

(1) Evaluation Indicators: To quantitatively analyse the efficacy of the suggested method and other methods for comparison, four quantitative evaluation indexes are introduced: overall accuracy (OA), average accuracy (AA), kappa coefficient (κ), and the classification accuracy of each class. A better classification effect is indicated by a higher value for each indicator.

(2) Configuration: All verification experiments for the proposed technique were performed in the PyTorch environment using a desktop computer with an Intel(R) Core(TM) i7-10750H CPU, 16GB of RAM, and an NVIDIA Geforce GTX 1660Ti 6-GB GPU. The learning rate was initially set to 1×10^{-3} and the Adam optimizer was selected as the initial optimizer. The size of each training batch was set to 64. Each dataset received 500 training epochs.

4.3. Parameter Analysis

4.3.1. Influence of Patch Size

Patch size is the spatial size of the input patches, which determines the spatial information that the model can utilize when classifying HSIs. Therefore, The model's performance is influenced by the patch size. A too large patch size will increase the computational burden of the model. In this section we compare a set of patch sizes $\{3, 5, 7, 9, 11, 13\}$ to explore the effect of patch size on the model. The experimental results about patch size on the three datasets are shown in Figure 8. A similar trend was observed in all three datasets, OA first increased and then stabilized with increasing patch size. Specifically, the highest value of OA is achieved when the patch size is 9 in the PU and HU datasets, and the highest value of OA is achieved when the patch size is 11 in the SA dataset.

The size of patch is positively correlated with the spatial information contained in the patch. Increasing the patch means that the model can learn more spatial information, which will be beneficial to improve OA. And when the patch increases to a certain size, the distance between the pixels in the newly region and the center pixel is too far, and the spatial information that can be provided is of little value. So the improvement of OA is not much, and the OA will tend to be stable at this time.

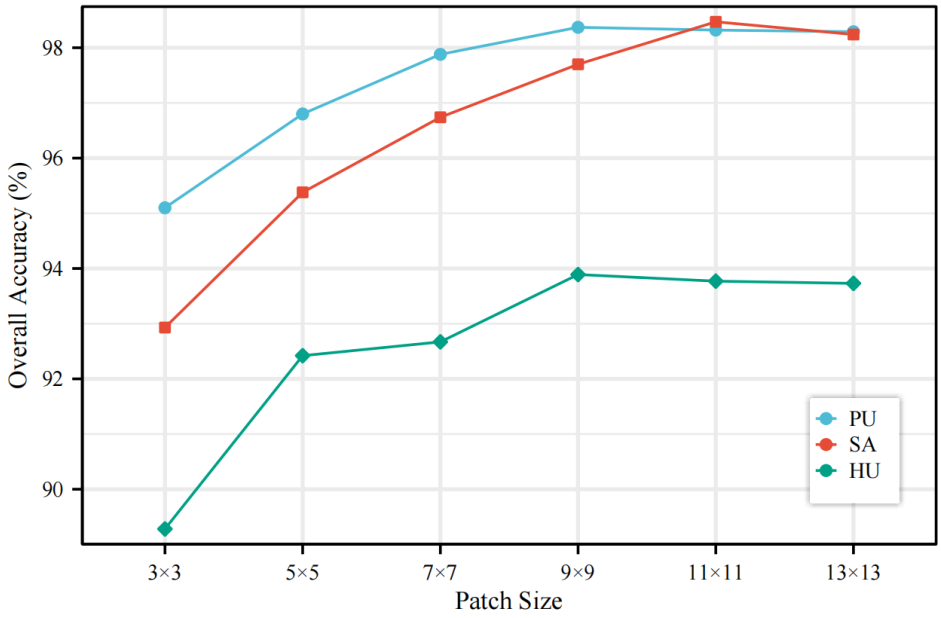


Figure 8. Overall accuracy(%) with different patch sizes on the three datasets. The window numbers in transformer layers is set to [1, 2, 2, 4].

4.3.2. Influence of Window Number

In proposed S-SW-MSA, the number of windows is a parameter that can be set depending on the characteristics of the dataset. Moreover, the number of windows can be different for each transformer layer in order to extract multiple scales of features. We set up six sets of experiments, the model contains four transformer layers in the first four sets, and five transformer layers in the last two sets. the numbers in [] indicate the number of windows of S-SW-MSA in each transformer layer. The experimental results on the three datasets are shown in Table 4. According to the experimental results, the best OA for each dataset was obtained for different window number settings, and the best OA was obtained for the PU, SA and HU datasets in the 4th, 2nd and 6th group settings, respectively. We also found that increasing the number of transformer layers does not necessarily increase the performance of the model. For example, the best OA is achieved when the number of transformer layers is 4 for the PU and SA datasets and 5 for the HU dataset. Because the features of each dataset are different, the parameter settings will change accordingly.

Table 4. Overall accuracies (%) of proposed model with different number of windows in transformer layers on SA, PU and HU datasets. The patch size is set to 9.

Windows Size	PU	SA	HU
[1, 1, 2, 2]	97.05	97.56	93.24
[1, 2, 2, 4]	97.86	**97.80**	93.35
[2, 2, 4, 4]	98.33	96.93	93.31
[2, 2, 4, 8]	**98.37**	97.70	93.58
[1, 1, 2, 4, 8]	98.20	96.25	93.38
[2, 2, 4, 4, 8]	98.25	96.31	**93.69**

4.4. Ablation Experiments

To sufficiently demonstrate that proposed method is effective, we conducted ablation experiments on the Pavia University dataset. With ViT as the baseline, the components of the model are added separately: S-Swin, SPE and SFE. In total, there are 5 combinations. The experimental results are shown in the Table 5. The classification overall accuracy of ViT without any improvement was 84.43%. SPE, SFE and S-Swin are proposed improvements for the ViT backbone network, which can respectively increase classification overall accuracy of 1.69%, 7.21% and 7.87% after adding into the model. The classification overall accuracy of applying the two improvements to the model together can reach 93.78%, which is higher than baseline by 9.35%. It is considered to be a great result for the improved pure transformer, but it's a little lower than our final result. After the SFE was added to the model, the classification overall accuracy improved by 4.59%, eventually reaching 98.37.

Table 5. Ablation experiments in PU.

Method	Module (%)			Metric (%)		
	S-Swin	SPE	SFE	OA(%)	AA(%)	$\kappa \times 100$ (%)
ViT(Baseline)	✗	✗	✗	84.43	78.06	78.95
ViT	✗	✓	✗	86.12	80.18	81.31
ViT	✗	✗	✓	91.64	90.43	88.97
SSWT(Ours)	✓	✗	✗	92.30	89.58	89.75
SSWT(Ours)	✓	✓	✗	93.78	91.17	91.74
SSWT(Ours)	✓	✓	✓	98.37	97.25	97.84

4.5. Classification Results

The proposed model's outcomes are compared with those of the advanced deep learning models: a LSTM based network (Bi-LSTM) [51], a 3-D CNN-based deep learning network (3D-CNN) [57], a deep feature fusion network (DFFN) [53], a RSSAN [52], and some transformer based model include a Vit, Swin-transformer (SwinT) [43], a SpectralFormer (SF) [44], a Hit [46] and a SSFTT [45].

Tables 6–8 show the OA, AA, κ and the accuracy of each category for each model's classification on the three public datasets. Each result is the average of repeating the experiment five times. The best results are shown in bold. As the results show, proposed SSWT performs the best. On the PU dataset, SSWT is 1.02% higher than SSFTT, 3.85% higher than HiT, 9.01% higher than SwinT and 1.51% higher than RSSAN in terms of OA. Moreover, SSWT outperforms other models in terms of AA and κ. SSWT achieved the highest classification accuracy in 7 out of 9 categories. On the SA dataset, the advantage of SSWT is more prominent. SSWT is 3.22% higher than SSFTT, 3.99% higher than HiT, 7.10% higher than SwinT, 2.64% higher than RSSAN, and 3.01% higher than DFFN in terms of OA. The same advantage was achieved for SSWT in AA and κ. SSWT achieved the highest classification accuracy in 11 out of 16 categories. Similar results can be observed in HU dataset, where SSWT achieved significant advantages in all three metrics of OA, AA and κ. SSWT achieved the highest classification accuracy in 6 out of 15 categories.

Table 6. Classification results of the PU dataset.

Class	Bi-LSTM	3D-CNN	RSSAN	DFFN	Vit	SwinT	SF	Hit	SSFTT	SSWT
1	91.67 ± 0.83	95.16 ± 1.56	97.12 ± 0.57	96.66 ± 0.81	87.96 ± 1.80	93.05 ± 5.32	89.41 ± 2.23	93.72 ± 1.44	97.31 ± 1.12	**98.06 ± 0.24**
2	96.96 ± 1.60	98.31 ± 0.96	99.46 ± 0.11	99.05 ± 0.51	96.56 ± 3.00	96.98 ± 1.43	97.22 ± 0.76	98.66 ± 0.48	99.37 ± 0.26	**99.91 ± 0.08**
3	70.65 ± 9.73	36.91 ± 6.18	85.74 ± 5.05	70.37 ± 12.56	53.18 ± 19.35	29.49 ± 23.08	77.28 ± 3.19	80.42 ± 7.56	87.25 ± 5.43	**94.59 ± 2.40**
4	92.88 ± 2.78	95.52 ± 1.58	96.92 ± 1.32	94.22 ± 3.16	89.76 ± 2.25	92.09 ± 1.41	90.80 ± 1.92	94.74 ± 1.84	97.59 ± 1.15	**97.70 ± 1.05**
5	99.10 ± 0.60	99.83 ± 0.34	99.86 ± 0.17	99.97 ± 0.06	**100.00 ± 0.00**	99.16 ± 0.59	**100.00 ± 0.00**	99.95 ± 0.04	99.95 ± 0.06	99.85 ± 0.27
6	67.03 ± 14.76	49.91 ± 12.17	97.00 ± 1.09	95.07 ± 3.03	51.97 ± 7.05	88.47 ± 5.03	82.13 ± 6.02	95.54 ± 2.05	97.00 ± 1.60	**98.37 ± 1.63**
7	82.67 ± 3.31	46.74 ± 14.05	84.15 ± 5.66	74.68 ± 7.86	47.59 ± 8.36	45.18 ± 31.47	52.80 ± 6.23	75.17 ± 8.05	91.43 ± 3.70	**91.95 ± 5.61**
8	83.17 ± 3.25	89.73 ± 3.00	92.49 ± 1.51	87.38 ± 4.34	78.79 ± 8.88	92.76 ± 1.65	81.81 ± 4.44	85.83 ± 4.19	93.81 ± 1.51	**95.35 ± 5.61**
9	98.94 ± 0.51	98.66 ± 0.62	98.37 ± 0.96	99.57 ± 0.22	96.71 ± 1.00	76.85 ± 12.09	96.32 ± 1.38	97.16 ± 1.29	**99.72 ± 0.20**	99.48 ± 0.88
OA(%)	89.52 ± 1.91	86.63 ± 1.43	96.86 ± 0.36	94.74 ± 1.40	84.43 ± 1.56	89.36 ± 3.14	90.16 ± 0.89	94.52 ± 1.03	97.35 ± 0.45	**98.37 ± 0.24**
AA(%)	87.01 ± 1.97	78.97 ± 2.05	94.57 ± 0.84	90.77 ± 2.46	78.06 ± 2.56	79.34 ± 7.46	85.31 ± 1.20	91.24 ± 1.94	95.94 ± 0.73	**97.25 ± 0.64**
$\kappa \times 100$	85.94 ± 2.63	81.79 ± 2.04	95.84 ± 0.48	93.00 ± 1.87	78.95 ± 2.03	85.82 ± 4.23	86.87 ± 1.21	92.74 ± 1.37	96.49 ± 0.60	**97.84 ± 0.32**

Table 7. Classification results of the SA dataset.

Class	Bi-LSTM	3D-CNN	RSSAN	DFFN	Vit	SwinT	SF	Hit	SSFTT	SSWT
1	79.24 ± 39.63	97.09 ± 1.46	99.58 ± 0.48	97.12 ± 0.89	90.19 ± 2.51	72.30 ± 1.87	95.05 ± 1.49	98.69 ± 2.05	99.44 ± 0.90	**99.79 ± 0.43**
2	98.94 ± 0.55	**99.90 ± 0.08**	99.36 ± 0.87	99.58 ± 0.14	98.05 ± 1.17	97.24 ± 1.92	99.32 ± 0.18	99.32 ± 0.35	99.80 ± 0.34	99.80 ± 0.17
3	85.20 ± 12.23	88.23 ± 4.35	97.01 ± 1.63	95.01 ± 3.54	87.52 ± 1.83	89.31 ± 2.96	92.89 ± 1.29	95.51 ± 2.29	98.41 ± 1.04	**98.48 ± 1.54**
4	97.79 ± 1.21	98.22 ± 1.10	98.56 ± 0.70	96.67 ± 1.39	94.11 ± 1.43	96.12 ± 1.50	94.05 ± 2.02	98.82 ± 0.51	**99.59 ± 0.56**	98.53 ± 1.23
5	96.40 ± 1.22	93.41 ± 2.41	96.06 ± 1.37	96.87 ± 1.04	82.59 ± 2.93	97.68 ± 0.76	93.24 ± 1.83	96.03 ± 2.17	98.28 ± 0.77	**98.74 ± 0.80**
6	99.46 ± 0.37	99.79 ± 0.32	99.36 ± 1.00	99.84 ± 0.30	99.44 ± 0.64	98.89 ± 1.29	99.68 ± 0.36	**99.99 ± 0.02**	99.98 ± 0.02	99.96 ± 0.06
7	98.84 ± 0.36	99.47 ± 0.23	99.28 ± 0.40	99.62 ± 0.28	98.05 ± 0.71	97.79 ± 0.92	98.81 ± 0.47	98.88 ± 0.62	99.44 ± 0.46	**99.72 ± 0.42**
8	83.66 ± 3.85	82.53 ± 2.36	90.93 ± 2.87	89.16 ± 1.74	82.79 ± 1.93	87.64 ± 1.38	85.03 ± 2.46	88.55 ± 1.73	90.08 ± 4.06	**95.87 ± 1.47**
9	97.84 ± 1.34	98.51 ± 1.11	99.66 ± 0.26	98.88 ± 0.80	96.38 ± 0.57	99.16 ± 0.63	98.05 ± 0.64	99.62 ± 0.37	99.53 ± 0.24	**99.92 ± 0.06**
10	81.10 ± 8.62	89.40 ± 2.50	95.58 ± 2.48	95.39 ± 1.01	75.44 ± 3.81	89.52 ± 3.74	91.23 ± 2.28	93.74 ± 2.38	95.73 ± 2.58	**97.07 ± 1.88**
11	83.59 ± 6.83	73.95 ± 4.65	93.37 ± 5.75	92.56 ± 5.81	70.47 ± 15.29	83.99 ± 14.49	89.86 ± 4.74	91.16 ± 6.19	94.66 ± 4.66	**95.64 ± 4.52**
12	98.84 ± 0.61	99.21 ± 0.56	99.36 ± 0.79	**99.97 ± 0.03**	98.67 ± 1.31	95.76 ± 0.75	98.45 ± 1.46	99.30 ± 0.64	99.80 ± 0.28	99.78 ± 0.45
13	94.78 ± 2.72	99.66 ± 0.07	98.92 ± 0.99	**99.98 ± 0.04**	96.28 ± 2.05	94.92 ± 6.31	98.61 ± 0.92	98.99 ± 1.12	99.06 ± 1.66	99.87 ± 0.18
14	90.20 ± 2.51	97.24 ± 1.05	96.63 ± 0.57	98.52 ± 0.76	96.51 ± 1.38	94.47 ± 1.04	95.03 ± 2.32	97.16 ± 0.77	95.61 ± 2.88	**99.23 ± 0.55**
15	78.87 ± 9.66	73.91 ± 2.47	86.60 ± 3.27	87.97 ± 2.81	72.03 ± 5.50	86.75 ± 6.26	79.87 ± 3.00	81.79 ± 3.34	81.36 ± 6.09	**94.10 ± 2.05**
16	90.27 ± 9.62	92.36 ± 1.46	96.67 ± 1.27	95.16 ± 2.32	91.57 ± 0.75	92.77 ± 3.30	95.35 ± 0.99	96.79 ± 1.67	97.20 ± 1.02	**98.40 ± 1.08**
OA(%)	89.66 ± 3.03	90.22 ± 0.70	95.16 ± 0.35	94.79 ± 0.80	87.58 ± 0.37	90.70 ± 2.38	91.81 ± 0.73	93.81 ± 0.56	94.58 ± 0.41	**97.80 ± 0.25**
AA(%)	90.94 ± 3.31	92.68 ± 0.71	96.68 ± 0.49	96.39 ± 0.57	89.38 ± 0.51	90.02 ± 3.99	94.03 ± 0.48	95.90 ± 0.24	96.75 ± 0.26	**98.43 ± 0.35**
$\kappa \times 100$	88.49 ± 3.39	89.11 ± 0.77	94.61 ± 0.39	94.20 ± 0.89	86.17 ± 0.41	89.63 ± 2.67	90.89 ± 0.81	93.10 ± 0.62	93.97 ± 0.46	**97.55 ± 0.28**

Table 8. Classification results of the HU dataset.

Class	Bi-LSTM	3D-CNN	RSSAN	DFFN	Vit	SwinT	SF	Hit	SSFTT	SSWT
1	84.09 ± 4.77	89.90 ± 6.62	95.05 ± 2.77	94.71 ± 5.79	90.72 ± 6.21	94.56 ± 2.55	95.05 ± 5.10	93.37 ± 4.54	93.96 ± 4.32	**95.13 ± 4.45**
2	90.60 ± 7.71	81.28 ± 6.08	98.05 ± 1.19	97.75 ± 1.06	83.93 ± 9.70	93.93 ± 5.83	93.53 ± 3.77	97.78 ± 0.87	98.71 ± 1.11	**98.77 ± 1.18**
3	75.14 ± 17.70	91.81 ± 4.04	98.67 ± 0.81	99.49 ± 0.74	88.01 ± 8.50	96.68 ± 1.98	97.19 ± 2.01	98.64 ± 0.91	**99.52 ± 0.89**	99.46 ± 0.67
4	90.83 ± 3.70	91.91 ± 0.35	94.06 ± 1.91	91.34 ± 0.74	85.63 ± 3.35	94.42 ± 2.77	89.54 ± 1.79	95.35 ± 1.99	**96.65 ± 2.55**	95.75 ± 1.55
5	92.86 ± 2.93	95.97 ± 1.83	98.29 ± 0.77	98.44 ± 0.74	95.86 ± 1.75	97.99 ± 0.74	96.97 ± 0.92	98.69 ± 0.98	99.54 ± 0.49	**99.93 ± 0.08**
6	52.43 ± 31.32	72.69 ± 2.15	80.58 ± 6.70	86.15 ± 6.72	6.93 ± 7.34	71.20 ± 14.20	63.88 ± 5.20	81.49 ± 2.85	90.42 ± 6.32	**92.62 ± 5.67**
7	72.93 ± 9.32	84.15 ± 2.50	87.09 ± 3.56	84.60 ± 3.98	64.32 ± 11.11	71.84 ± 14.62	74.67 ± 4.06	81.16 ± 5.29	86.22 ± 5.43	**88.70 ± 4.61**
8	55.74 ± 5.24	55.87 ± 6.14	78.88 ± 3.64	79.10 ± 3.82	66.84 ± 6.80	73.69 ± 9.90	76.31 ± 2.76	78.85 ± 2.03	82.79 ± 2.81	**85.08 ± 3.38**
9	73.05 ± 5.75	81.90 ± 2.13	81.77 ± 4.72	84.24 ± 4.75	66.24 ± 5.56	73.28 ± 2.75	72.94 ± 6.60	83.62 ± 5.81	**89.96 ± 4.24**	87.47 ± 3.31
10	39.43 ± 20.49	48.10 ± 12.51	89.76 ± 0.52	90.22 ± 5.12	63.29 ± 5.92	78.56 ± 2.66	81.13 ± 5.79	86.14 ± 5.11	93.60 ± 1.29	**96.05 ± 3.71**
11	66.55 ± 10.85	60.66 ± 2.63	82.85 ± 4.35	82.46 ± 3.64	58.67 ± 3.08	76.21 ± 0.37	68.80 ± 6.54	79.52 ± 4.94	86.36 ± 2.82	**87.55 ± 5.08**
12	67.21 ± 9.90	58.29 ± 10.86	92.13 ± 2.73	93.10 ± 2.00	61.69 ± 6.32	87.50 ± 3.52	85.02 ± 4.18	90.96 ± 3.22	88.95 ± 5.90	**97.83 ± 1.12**
13	19.96 ± 14.65	59.10 ± 10.82	71.21 ± 8.17	**92.47 ± 1.57**	40.09 ± 16.86	71.60 ± 2.70	50.85 ± 9.67	79.28 ± 2.86	92.33 ± 2.81	90.76 ± 3.15
14	89.93 ± 8.82	93.12 ± 3.76	92.38 ± 3.91	94.74 ± 2.65	77.49 ± 4.47	89.03 ± 7.12	78.28 ± 3.02	93.96 ± 2.88	**96.46 ± 2.24**	94.55 ± 3.68
15	90.91 ± 8.78	**99.39 ± 0.77**	95.82 ± 2.62	98.88 ± 0.88	91.48 ± 3.12	96.65 ± 1.75	95.15 ± 2.84	98.47 ± 1.03	98.66 ± 1.13	98.63 ± 1.56
OA(%)	72.60 ± 3.03	76.73 ± 1.69	89.76 ± 0.39	90.62 ± 0.79	72.80 ± 1.54	84.78 ± 2.39	82.97 ± 0.99	89.16 ± 1.03	92.47 ± 0.97	**93.69 ± 1.07**
AA(%)	70.78 ± 4.49	77.61 ± 1.80	89.11 ± 0.61	91.18 ± 0.86	69.41 ± 0.73	84.48 MSA 1.67	81.29 ± 1.16	89.15 ± 0.88	92.94 ± 1.01	**93.89 ± 1.07**
$\kappa \times 100$	70.34 ± 3.29	74.84 ± 1.83	88.93 ± 0.42	89.86 ± 0.85	70.58 ± 1.64	83.55 MSA 2.58	81.58 ± 1.07	88.28 ± 1.11	91.86 ± 1.05	**93.18 ± 1.16**

We visualized the prediction results of each model on the samples to compare the performance of the models, and the visualization results of each model on the three datasets are shown in Figures 9–11 Proposed SSWT has less noise in all three datasets compared to other models, and the classification result of SSWT are closest to the ground truth. In the

PU dataset, the blue area in the middle is misclassified by many models, and the SSWT result in the fewest errors. In the SA dataset, the pink area and the green area on the top left show a number of errors in the classification results of other models, and the SSWT classification results are the smoothest. A similar situation is observed in the HU dataset. The superiority of proposed model is further demonstrated.

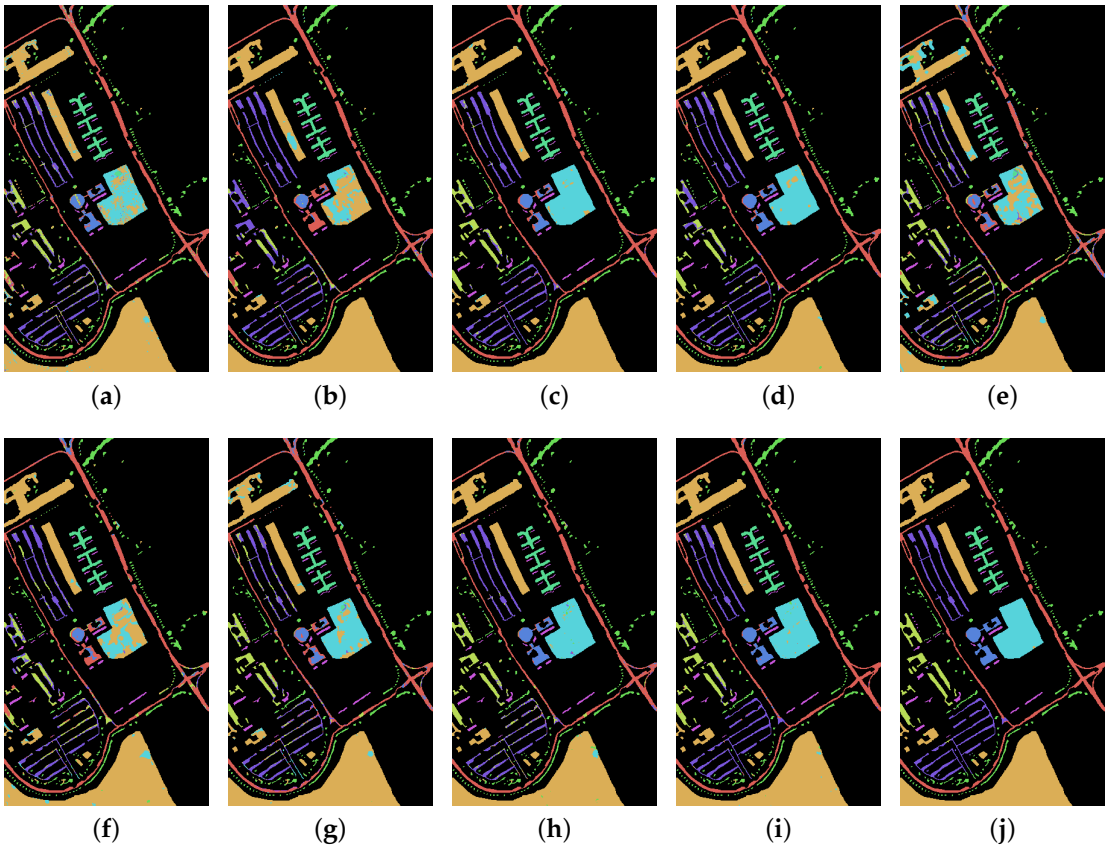


Figure 9. Classification maps of different methods in PU dataset. (**a**) Bi-LSTM. (**b**) 3D-CNN. (**c**) RSSAN. (**d**) DFFN. (**e**) Vit. (**f**) SwinT. (**g**) SF. (**h**) Hit. (**i**) SSFTT. (**j**) Proposed SSWT.

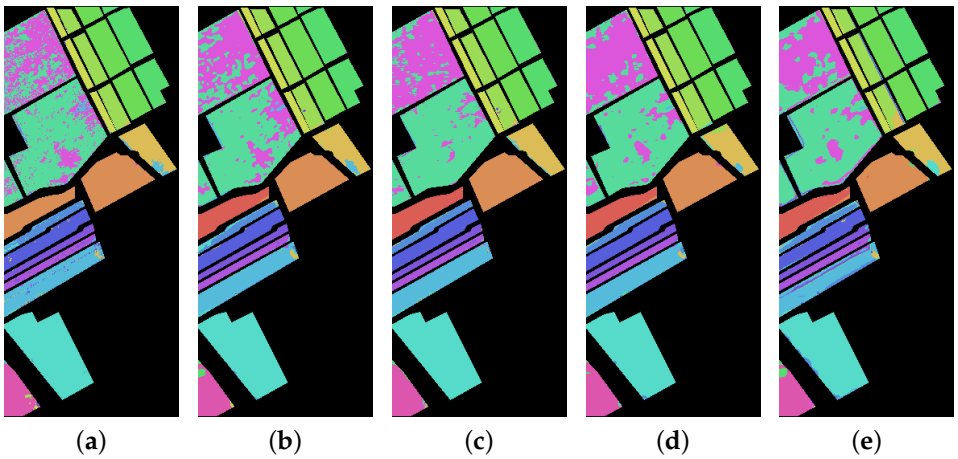
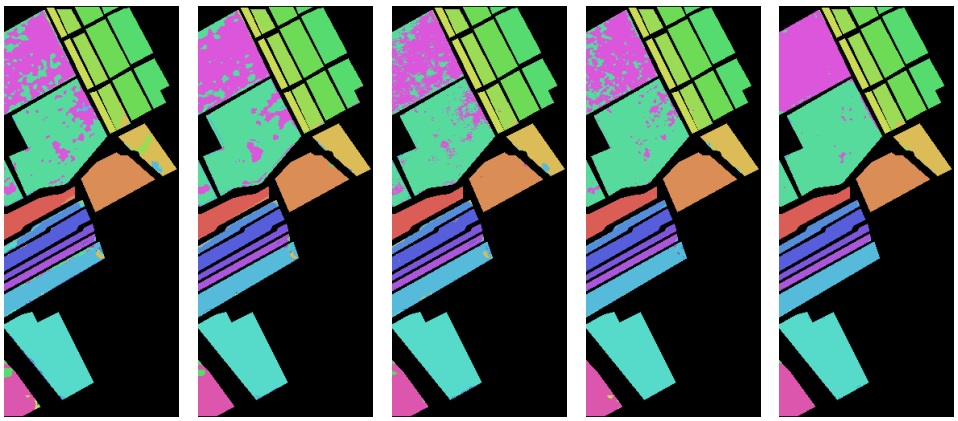
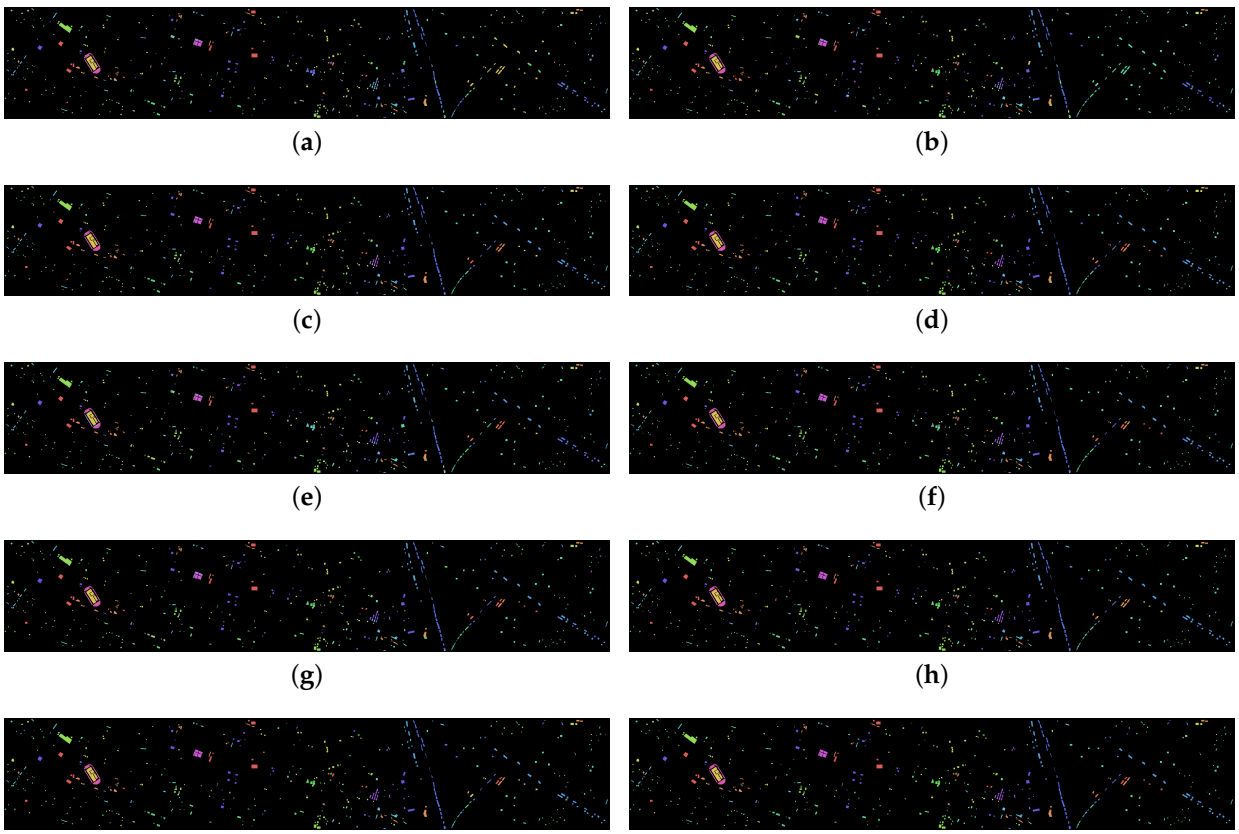


Figure 10. *Cont.*



(**f**) (**g**) (**h**) (**i**) (**j**)

Figure 10. Classification maps of different methods in SA dataset. (**a**) Bi-LSTM. (**b**) 3D-CNN. (**c**) RSSAN. (**d**) DFFN. (**e**) Vit. (**f**) SwinT. (**g**) SF. (**h**) Hit. (**i**) SSFTT. (**j**) Proposed SSWT.



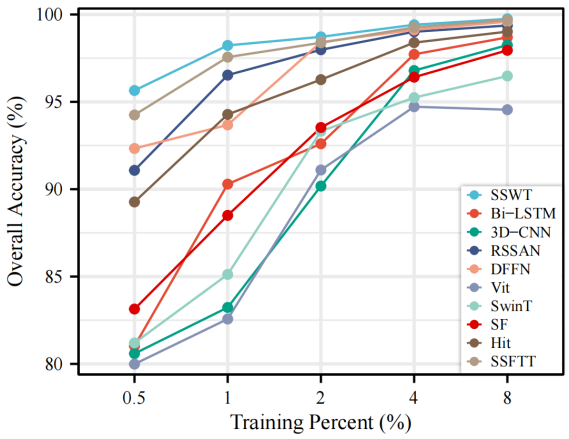
(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**)

Figure 11. Classification maps of different methods in HU dataset. (**a**) Bi-LSTM. (**b**) 3D-CNN. (**c**) RSSAN. (**d**) DFFN. (**e**) Vit. (**f**) SwinT. (**g**) SF. (**h**) Hit. (**i**) SSFTT. (**j**) Proposed SSWT.

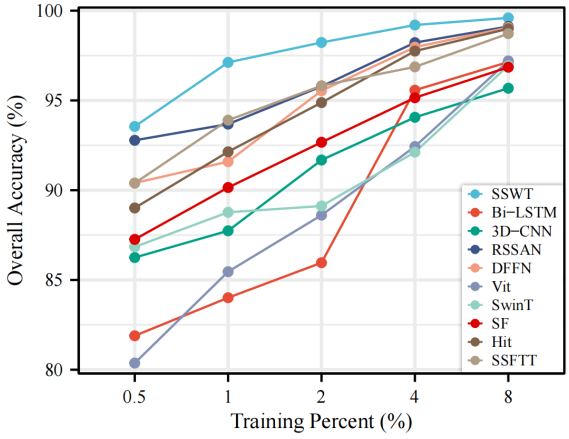
4.6. Robustness Evaluation

In order to evaluate the robustness of the proposed model, we conducted experiments with the proposed model and other models under different numbers of training samples. Figure 12 shows the experimental results on three datasets, we selected 0.5%, 1%, 2%, 4%, and 8% of the samples in turn as training data for the PU and SA dataset , while 2%, 4%, 6%, 8% and 10% for the HU dataset. It can be observed that the proposed SSWT is performing best in every situation, especially in the case of few training samples. The robustness

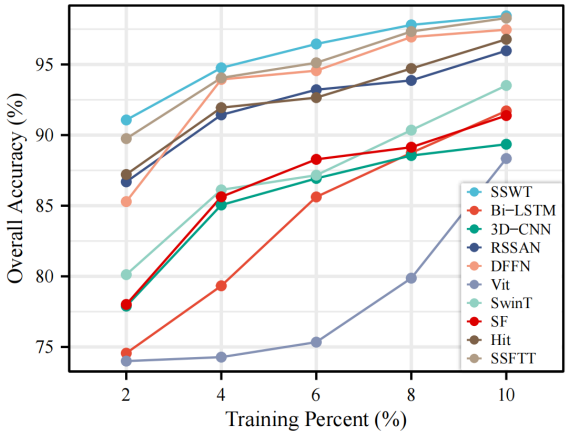
of proposed SSWT and its superiority in the case of small samples can be demonstrated. Taking the PU dataset as an example, most of the models achieve high accuracy at 8% of the training percent, with SSWT having a small advantage. And as the training percent decreases, SSWT has higher accuracy compared to other models. Similar results were found on the SA and HU datasets, where SSWT showed excellent performance for all training percents.



(**a**)



(**b**)



(**c**)

Figure 12. Classification results in different training percent of samples on the three datasets. (**a**) PU. (**b**) SA. (**c**) HU.

5. Conclusions

In this paper, we summarize the shortcomings of the existing ViT for HSI classification tasks. For the lack of ability to capture local contextual features, we use the self-attentive mechanism of shifted windows. The corresponding design is made for the characteristics of HSI, i.e., the spectral shifted window self-attention, which effectively improves the local feature extraction capability. For the insensitivity of ViT to spatial features and structure, we designed the spatial feature extraction module and spatial position encoding to compensate. The superiority of the proposed model has been verified by experimental results across three public HSI datasets.

In future work, we will improve the calculation of S-SW-MSA to reduce its time complexity. In addition, we will continue our research based on the transformer and try to achieve higher performance with a model of pure transformer structure.

Author Contributions: All the authors made significant contributions to the work. Y.P., J.R. and J.W. designed the research, analyzed the results, and accomplished the validation work. M.S. provided advice for the revision of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gevaert, C.M.; Suomalainen, J.; Tang, J.; Kooistra, L. Generation of spectral–temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2015**, *8*, 3140–3146. [CrossRef]
2. Jabir, B.; Falih, N.; Rahmani, K. Accuracy and Efficiency Comparison of Object Detection Open-Source Models. *Int. J. Online Biomed. Eng.* **2021**, *17*, 165–184. [CrossRef] [CrossRef]
3. Lu, G.; Fei, B. Medical hyperspectral imaging: A review. *J. Biomed. Opt.* **2014**, *19*, 010901. [CrossRef]
4. Lone, Z.A.; Pais, A.R. Object detection in hyperspectral images. *Digit. Signal Process.* **2022**, *131*, 103752. [CrossRef] [CrossRef]
5. Weber, C.; Aguejdad, R.; Briottet, X.; Avala, J.; Fabre, S.; Demuynck, J.; Zenou, E.; Deville, Y.; Karoui, M.S.; Benhalouche, F.Z.; et al. Hyperspectral imagery for environmental urban planning. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: New York, NY, USA, 2018; pp. 1628–1631.
6. Li, N.; Lü, J.S.; Altermann, W. Hyperspectral remote sensing in monitoring the vegetation heavy metal pollution. *Spectrosc. Spectr. Anal.* **2010**, *30*, 2508–2511. [CrossRef]
7. Saralıoğlu, E.; Görmüş, E.T.; Güngör, O. Mineral exploration with hyperspectral image fusion. In Proceedings of the 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, Turkey, 16–19 May 2016; IEEE: New York, NY, USA, 2016; pp. 1281–1284.
8. Ren, J.; Wang, R.; Liu, G.; Feng, R.; Wang, Y.; Wu, W. Partitioned relief-F method for dimensionality reduction of hyperspectral images. *Remote Sens.* **2020**, *12*, 1104. [CrossRef] [CrossRef]
9. Ke, C. Military object detection using multiple information extracted from hyperspectral imagery. In Proceedings of the 2017 International Conference on Progress in Informatics and Computing (PIC), Nanjing, China, 15–17 December 2017; IEEE: New York, NY, USA, 2017; pp. 124–128.
10. Cariou, C.; Chehdi, K. A new k-nearest neighbor density-based clustering method and its application to hyperspectral images. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: New York, NY, USA, 2016; pp. 6161–6164.
11. Ren, J.; Wang, R.; Liu, G.; Wang, Y.; Wu, W. An SVM-based nested sliding window approach for spectral–spatial classification of hyperspectral images. *Remote Sens.* **2020**, *13*, 114. [CrossRef] [CrossRef]
12. Yaman, O.; Yetis, H.; Karakose, M. Band Reducing Based SVM Classification Method in Hyperspectral Image Processing. In Proceedings of the 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 26–27 May 2020; IEEE: New York, NY, USA, 2020; pp. 21–25.
13. Chen, Y.; Zhao, X.; Lin, Z. Optimizing subspace SVM ensemble for hyperspectral imagery classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2014**, *7*, 1295–1305. [CrossRef] [CrossRef]
14. Shao, Z.; Zhang, L.; Zhou, X.; Ding, L. A novel hierarchical semisupervised SVM for classification of hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1609–1613. [CrossRef] [CrossRef]
15. Zhang, Y.; Cao, G.; Li, X.; Wang, B. Cascaded random forest for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 1082–1094. [CrossRef] [CrossRef]

16. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef] [CrossRef]
17. Navarro, A.; Nicastro, N.; Costa, C.; Pentangelo, A.; Cardarelli, M.; Ortenzi, L.; Pallottino, F.; Cardi, T.; Pane, C. Sorting biotic and abiotic stresses on wild rocket by leaf-image hyperspectral data mining with an artificial intelligence model. *Plant Methods* **2022**, *18*, 45. [CrossRef] [CrossRef] [PubMed]
18. Song, W.; Li, S.; Kang, X.; Huang, K. Hyperspectral image classification based on KNN sparse representation. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: New York, NY, USA, 2016; pp. 2411–2414. [CrossRef]
19. Guo, Y.; Yin, X.; Zhao, X.; Yang, D.; Bai, Y. Hyperspectral image classification with SVM and guided filter. *EURASIP J. Wirel. Commun. Netw.* **2019**, *2019*, 56. [CrossRef] [CrossRef]
20. Wu, H.; Prasad, S. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sens.* **2017**, *9*, 298. [CrossRef] [CrossRef]
21. Luo, H. Shorten spatial-spectral RNN with parallel-GRU for hyperspectral image classification. *arXiv* **2018**, arXiv:1810.12563. [CrossRef]
22. 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] [CrossRef]
23. Lee, H.; Kwon, H. Contextual deep CNN based hyperspectral classification. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: New York, NY, USA, 2016; pp. 3322–3325.
24. Chen, Y.; Zhu, L.; Ghamisi, P.; Jia, X.; Li, G.; Tang, L. Hyperspectral images classification with Gabor filtering and convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2355–2359. [CrossRef] [CrossRef]
25. Zhao, X.; Tao, R.; Li, W.; Li, H.C.; Du, Q.; Liao, W.; Philips, W. Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7355–7370. [CrossRef] [CrossRef]
26. 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] [CrossRef]
27. 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; IEEE: New York, NY, USA, 2017; pp. 3904–3908.
28. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; IEEE: New York, NY, USA, 2015; pp. 4959–4962.
29. Wan, S.; Gong, C.; Zhong, P.; Du, B.; Zhang, L.; Yang, J. Multiscale dynamic graph convolutional network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3162–3177. [CrossRef] [CrossRef]
30. Mou, L.; Lu, X.; Li, X.; Zhu, X.X. Nonlocal graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8246–8257. [CrossRef] [CrossRef]
31. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J.; Pla, F. Capsule networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2145–2160. [CrossRef] [CrossRef]
32. Yin, J.; Li, S.; Zhu, H.; Luo, X. Hyperspectral image classification using CapsNet with well-initialized shallow layers. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1095–1099. [CrossRef] [CrossRef]
33. Zhou, F.; Hang, R.; Liu, Q.; Yuan, X. Hyperspectral image classification using spectral-spatial LSTMs. *Neurocomputing* **2019**, *328*, 39–47. [CrossRef] [CrossRef]
34. Gao, J.; Gao, X.; Wu, N.; Yang, H. Bi-directional LSTM with multi-scale dense attention mechanism for hyperspectral image classification. *Multimed. Tools Appl.* **2022**, *81*, 24003–24020. [CrossRef]
35. Xu, Y.; Du, B.; Zhang, L.; Zhang, F. A band grouping based LSTM algorithm for hyperspectral image classification. In *Computer Vision: Second CCF Chinese Conference, CCCV 2017, Tianjin, China, 11–14 October 2017, Proceedings, Part II*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 421–432.
36. Luo, Y.; Zou, J.; Yao, C.; Zhao, X.; Li, T.; Bai, G. HSI-CNN: A novel convolution neural network for hyperspectral image. In Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; IEEE: New York, NY, USA, 2018; pp. 464–469. [CrossRef]
37. Haut, J.M.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Li, J. Hyperspectral image classification using random occlusion data augmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1751–1755. [CrossRef] [CrossRef]
38. Sun, K.; Wang, A.; Sun, X.; Zhang, T. Hyperspectral image classification method based on M-3DCNN-Attention. *J. Appl. Remote Sens.* **2022**, *16*, 026507. [CrossRef]
39. Xu, H.; Yao, W.; Cheng, L.; Li, B. Multiple spectral resolution 3D convolutional neural network for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 1248. [CrossRef] [CrossRef]
40. Li, W.; Chen, H.; Liu, Q.; Liu, H.; Wang, Y.; Gui, G. Attention mechanism and depthwise separable convolution aided 3DCNN for hyperspectral remote sensing image classification. *Remote Sens.* **2022**, *14*, 2215. [CrossRef]
41. Sellami, A.; Abbes, A.B.; Barra, V.; Farah, I.R. Fused 3-D spectral-spatial deep neural networks and spectral clustering for hyperspectral image classification. *Pattern Recognit. Lett.* **2020**, *138*, 594–600. [CrossRef]

42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
43. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2013**, arXiv:2103.14030. [CrossRef]
44. 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*, 1–15. [CrossRef] [CrossRef]
45. 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] [CrossRef]
46. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
47. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
48. Xue, Z.; Xu, Q.; Zhang, M. Local transformer with spatial partition restore for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 4307–4325. [CrossRef]
49. Hu, X.; Yang, W.; Wen, H.; Liu, Y.; Peng, Y. A lightweight 1-D convolution augmented transformer with metric learning for hyperspectral image classification. *Sensors* **2021**, *21*, 1751. [CrossRef] [CrossRef] [PubMed]
50. Qing, Y.; Liu, W.; Feng, L.; Gao, W. Improved transformer net for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 2216. [CrossRef] [CrossRef]
51. Mei, S.; Li, X.; Liu, X.; Cai, H.; Du, Q. Hyperspectral image classification using attention-based bidirectional long short-term memory network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
52. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual spectral-spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 449–462. [CrossRef] [CrossRef]
53. 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] [CrossRef]
54. Hang, R.; Li, Z.; Liu, Q.; Ghamisi, P.; Bhattacharyya, S.S. Hyperspectral image classification with attention-aided CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2281–2293. [CrossRef] [CrossRef]
55. Chakraborty, T.; Trehan, U. Spectralnet: Exploring spatial-spectral waveletcnn for hyperspectral image classification. *arXiv* **2021**, arXiv:2104.00341. [CrossRef]
56. Gong, H.; Li, Q.; Li, C.; Dai, H.; He, Z.; Wang, W.; Li, H.; Han, F.; Tuniyazi, A.; Mu, T. Multiscale information fusion for hyperspectral image classification based on hybrid 2D-3D CNN. *Remote Sens.* **2021**, *13*, 2268. [CrossRef] [CrossRef]
57. Hamida, A.B.; Benoit, A.; Lambert, P.; Amar, C.B. 3-D deep learning approach for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4420–4434. [CrossRef]
58. Fang, J.; Xie, L.; Wang, X.; Zhang, X.; Liu, W.; Tian, Q. MSG-transformer: Exchanging local spatial information by manipulating messenger tokens. *arXiv* **2022**, arXiv:2105.15168. [CrossRef]
59. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. *arXiv* **2022**, arXiv:2103.14030. [CrossRef]
60. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. *arXiv* **2022**, arXiv:2108.05895. [CrossRef]
61. Ayas, S.; Tunc-Gormus, E. SpectralSWIN: A spectral-swin transformer network for hyperspectral image classification. *Int. J. Remote Sens.* **2022**, *43*, 4025–4044. [CrossRef] [CrossRef]