electronics

MDPI

Article

Multi-Scale Residual Spectral–Spatial Attention Combined with Improved Transformer for Hyperspectral Image Classification

Aili Wang [1], Kang Zhang [1], Haibin Wu [1], Yuji Iwahori [2] and Haisong Chen [3,*]

[1] Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; aili925@hrbust.edu.cn (A.W.); 2220610126@stu.hrbust.edu.cn (K.Z.); woo@hrbust.edu.cn (H.W.)
[2] Computer Science, Chubu University, Kasugai 487-8501, Japan; iwahori@isc.chubu.ac.jp
[3] School of Undergraduate Education, Shenzhen Polytechnic University, Shenzhen 518115, China
* Correspondence: hschen@szpt.edu.cn

Abstract: Aiming to solve the problems of different spectral bands and spatial pixels contributing differently to hyperspectral image (HSI) classification, and sparse connectivity restricting the convolutional neural network to a globally dependent capture, we propose a HSI classification model combined with multi-scale residual spectral–spatial attention and an improved transformer in this paper. First, in order to efficiently highlight discriminative spectral–spatial information, we propose a multi-scale residual spectral–spatial feature extraction module that preserves the multi-scale information in a two-layer cascade structure, and the spectral–spatial features are refined by residual spectral–spatial attention for the feature-learning stage. In addition, to further capture the sequential spectral relationships, we combine the advantages of Cross-Attention and Re-Attention to alleviate computational burden and attention collapse issues, and propose the Cross-Re-Attention mechanism to achieve an improved transformer, which can efficiently alleviate the heavy memory footprint and huge computational burden of the model. The experimental results show that the overall accuracy of the proposed model in this paper can reach 98.71%, 99.33%, and 99.72% for Indiana Pines, Kennedy Space Center, and XuZhou datasets, respectively. The proposed method was verified to have high accuracy and effectiveness compared to the state-of-the-art models, which shows that the concept of the hybrid architecture opens a new window for HSI classification.

Keywords: hyperspectral image classification; multi-scale feature extraction; residual spectral–spatial attention; transformer

Citation: Wang, A.; Zhang, K.; Wu, H.; Iwahori, Y.; Chen, H. Multi-Scale Residual Spectral–Spatial Attention Combined with Improved Transformer for Hyperspectral Image Classification. *Electronics* **2024**, *13*, 1061. https://doi.org/10.3390/electronics13061061

Academic Editor: Silvia Liberata Ullo

Received: 25 January 2024
Revised: 3 March 2024
Accepted: 8 March 2024
Published: 13 March 2024

1. Introduction

Hyperspectral images (HSIs) contain both a high spatial resolution and continuous spectral bands of different objects at the same time, with the characteristics of "spectral image unity" [1,2]. They have been applied in a wide variety of applications, such as urban management [3], geological exploration [4], and military surveys [5].

HSI classification is a foundation component in Earth-monitoring applications, with the main goal of assigning each pixel in the HSI to specific land cover classes, thus achieving precise identification and classification of surface cover. Initially, HSI classification mainly used traditional machine learning methods to extract features. Typically, machine learning methods first adopted some dimension reduction methods to reduce spectral redundancy, such as principal component analysis (PCA) [6] and linear discriminant analysis (LDA) [7]. Traditional machine learning methods then employed classifiers such as the K-nearest neighbor method [8], support vector machine [9], random forest [10], decision tree [11], and other methods to classify the extracted features. Although traditional machine learning-based methods have made progress in improving classification performance, they often rely on hand-crafted features for HSI classification. With the rapid development of deep learning and practical progress in the task of HSI classification, deep learning-based methods fully

absorb the early experience of HSI classification, combining spectral and spatial information to complete the classification task, and can directly extract effective deep features from the original image [12]. Chen et al. [13] firstly introduced deep learning into the field of HSI classification, and used the unsupervised deep feature-learning model stacked autoencoder (SAE) to extract the features from the original image, which improves the accuracy of HSI classification.

Due to the existence of spectral and spatial heterogeneity in HSI, it is difficult to accurately identify land cover types using only spectral information. Therefore, the model used to jointly extract spectral–spatial features from HSIs for classification has become a research focus. Convolutional neural networks (CNNs) have been widely used in the field of HSI classification to realize the joint extraction of spectral and spatial features [14]. Among these, the three-dimensional convolutional neural network (3D-CNN) achieves direct end-to-end deep spectral–spatial feature extraction on HSIs, providing a robust and reliable feature extraction mechanism [15,16]. Considering the importance of multi-scale information for improving network performance, Song et al. [17] proposed a deep feature fusion strategy that is able to effectively fuse multi-scale feature representations by creating interconnections between different layers of information. Zhong et al. [18] proposed the spectral–spatial residual network (SSRN), which sequentially uses spectral residual blocks and spatial residual blocks to learn deep features from HSIs. Roy et al. [19] proposed the attention-based adaptive spectral–spatial kernel improved residual network (A^2S^2K-ResNet) with spectral attention to capture discriminative spectral–spatial features in an end-to-end training approach. In addition, attention mechanisms are widely used for HSI classification. Zhou et al. [20] designed a Cross-Attention Fusion module in an Attention Multihop Graph and Multiscale Convolutional Fusion Network (AMGCFN) to highlight important information and enhance feature fusion in different subnets. Gou et al. [21] proposed a global spatial feature representation model to learn global spatial features based on an encoder–decoder structure with channel attention and spatial attention. The CNN-based approaches improved the local perception of the model by point-wise operations with pixels around the image, but were limited by the kernel size and the number of network layers, which results in an insufficient ability to capture global contextual feature information.

In recent years, some studies have introduced a transformer, which extracts features by convolution, and then used the transformer to obtain contextual information [22]. Dosovitskiy et al. [23] proposed the Vision Transformer (ViT) with a dynamic and global sensory field, which ensures the model performs well in image classification tasks and can learn the dependencies of different positions of the output image. ViT learns features mainly using the multi-head attention mechanism, which can extract global information from the non-overlapping parts of the image. Therefore, ViT can effectively capture the long-range dependencies form the input images, enabling the network to parse the information from a global perspective, and thus effectively assisting in describing the local semantic information [24]. For the task of feature classification in HSIs, applying the ViT to sequence data is more effective and flexible in analyzing the spectral data of HSIs [25]. Sun et al. [26] proposed the Spectral–Spatial Feature Tokenisation Transformer (SSFTT) to obtain spectral–spatial features and high-level semantic information.

However, when applying ViT to the HSI classification task, a prominent issue is that the computational burden of the self-attention mechanism grows quadratically with input, and its computational amount hinders the inference speed of the model. Additionally, unlike CNNs that can be expanded to deeper layers to improve performance, the performance of ViT saturates rapidly when expanded to deeper layers, the expansion difficulty is mainly due to the collapse in attention, and the feature maps generated in deeper structures tend to be the same. Therefore, to address the problem of computational burden, Zhang et al. [27] proposed a lightweight transformer (LiT) that achieved a balance between high computational efficiency and significant performance. Liu et al. [28] proposed the Swin-Transformer to use a shift window to capture the global features. Meanwhile, Lin et al. [29]

proposed the Cross-Attention in Vision Transformer (CAT), which used Cross-Attention to focus on capturing local information inside the feature map patches, and captured global information between the feature map patches in a single channel. Both methods made the original square-growth computation become linear, which significantly reduced the computation of the transformer.

To address the problem of attention collapse, researchers from AI Lab proposed the Re-Attention mechanism, which regenerated the feature maps between layers to enhance the diversity between layers, avoiding the problem of feature maps converging to be the same in deeper layers [30]. Hybrid architectures combining the transformer and convolutions have garnered widespread attention in building lightweight, high-performance models. Some works have proposed a hybrid structure of a CNN and a transformer after analyzing the working principle of the CNN and the transformer in detail, where shallow features are extracted by the CNN and the extracted features are fed into a semantic tagger to tag the global semantic information [31–33].

Based on the above-mentioned analysis, in this paper, an efficient multi-scale residual spectral–spatial attention combined with an improved transformer (RSSAT) is proposed for HSI classification. In RSSAT, we designed a multi-scale residual spectral–spatial feature extraction module to improve the discriminative power of extracted features and adaptively fuse the acquired spectral and spatial information. In addition, we designed improved transformers to fully extract high-level semantic features and model long-range feature dependencies in HSI multidimensional datasets. Overall, our approach constructs a shallow-to-deep feature-learning model that effectively reduces misclassification of small target samples. The main contributions of this paper are summarized as follows:

1. In order to fully extract HSI high-level semantic features as well as to enhance the effective representation of global contextual information, this paper combines the respective representational features of a CNN and transformer, and proposes a new HSI classification method called RSSAT. RSSAT has strong advantages in discriminative feature extraction and capturing long-range dependencies with the best classification performance.
2. By investigating the characteristics of HSIs, a multi-scale residual spectral–spatial feature extraction module was designed. The module fully exploits the local information of HSIs in a two-layer cascade structure and selectively aggregates the information between spectral bands and spatial pixels to highlight discriminative information. The module alleviates the information loss in feature flow and retains more spectral and spatial information, reducing misclassification of small target samples and discrete samples.
3. In order to accurately capture long-range feature dependencies in HSI multidimensional datasets, we propose an improved transformer. For the transformer, we design the Cross-Re-Attention mechanism as an alternative to Self-Attention in the traditional transformer. The innovative strategy significantly enhances the model's ability to learn high-level semantic features by introducing a learnable matrix that dynamically generates new attention mappings between each layer.
4. According to the experimental results, RSSAT significantly outperforms other state-of-the-art deep learning methods in terms of classification performance, especially when dealing with uneven samples, and achieves an excellent improvement in its classification accuracy.

2. Materials and Methods

Figure 1 demonstrates the framework of the RSSAT model. In general, the model architecture mainly includes a multi-scale residual spectral–spatial feature extraction module and an improved transformer module. The model skillfully integrates the advantages of the CNN and transformer to enable feature extraction from shallow to deep, which enables the model to fully utilize the rich spectral–spatial information in HSIs, further improving the performance and robustness of the model. In the model training process, first, after

removing the spectral redundant bands by principal component analysis (PCA), the HSI data are fed into the convolution module to learn low-order features. Then, for the purpose of enhancing the spectral–spatial feature representation capability and robustness of the RSSAT model, residual spectral–spatial attention is embedded in the multi-scale residual feature-learning part. The multi-scale residual spectral–spatial feature extraction module re-adjusts and optimizes the extracted features through a two-level cascaded residual structure to highlight discriminative information. Meanwhile, the model can effectively establish channel connections between feature maps at different stages to enhance the convergence ability of the RSSAT. Finally, we propose the improved transformer to obtain long-distance dependencies of the sequential spectral features. The obtained discriminative spectral–spatial features are employed to obtain the classification results.

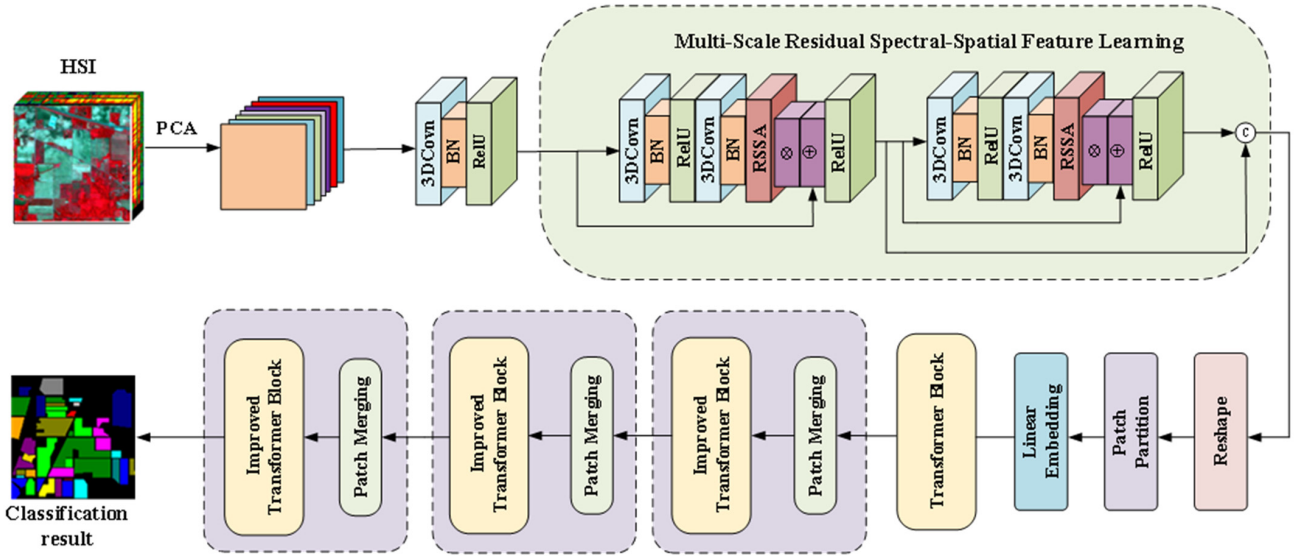


Figure 1. Framework of the proposed RSSAT model for HSI classification.

2.1. Residual Spectral–Spatial Attention

According to HSI pixel-level classification, there are two principles of joint extraction of spectral and spatial information [34]:

Principle 1: Spectral information is the basis of HSI pixel-level classification and is the most discriminative information.

Principle 2: Effective spatial information for HSI pixel-level classification refers to the information carried by neighboring pixels that are similar to the center pixel.

Based on the above two principles, this paper embeds the residual spectral–spatial attention module into the multi-scale feature extraction part to achieve the realignment and optimization of the spectral and spatial features to highlight the discriminative information, thereby improving the accuracy and efficiency of the HSI classification. Figure 2 illustrates the structure of the proposed residual spectral–spatial attention module.

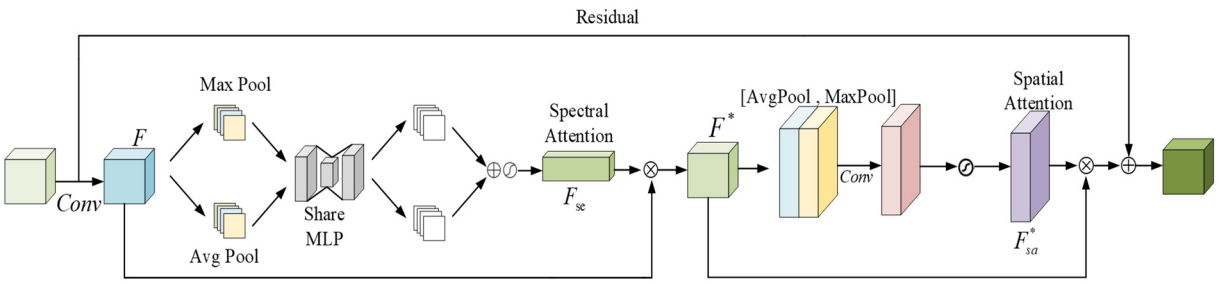


Figure 2. Residual spectral–spatial attention module.

In the paper, we introduce the spectral–spatial attention module [35] and combine it with residual operations to create the residual spatial–spectral attention module to enhance the feature extraction ability of RSSAT. First, we introduce the spectral attention module, which achieves the selection of specific spectral bands from the input HSI. The module highlights those bands that are useful for the classification task and reduces the influence of irrelevant bands. Next, we introduce the spatial attention module, which achieves fine extraction of spatial information by adaptively strengthening neighboring pixels that are the same category as the center pixel or weakening pixels of different categories. The two attention modules are arranged in a specific order. Based on the given input or intermediate features, spectral attention weights are computed and applied to the relevant features. Then, the obtained results are used as inputs to the spatial attention module.

Spectral Attention: The core purpose of the spectral attention module is to highlight those spectral features that are critical for HSI classification. To realize the refinement and selection of features, the spectral feature map is generated using the relationship between the spectral information of the features. The structure of spectral attention module is given in Figure 3.

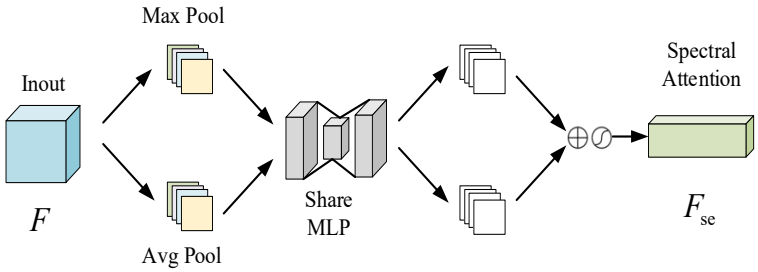


Figure 3. Spectral attention module.

To aggregate information and infer finer spectral attention, an average pooling layer and maximum pooling layer are employed. The two different feature descriptions are obtained for the feature mapping based on the different pooling schemes. The kth element of the output is calculated by Equation (1) and the kth channel of the output is calculated by Equation (2).

$$y_{avg}^{se} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} y_k(i,j) \tag{1}$$

$$y_{max}^{se} = max(y_k) \tag{2}$$

where $y_k(i,j)$ is the value at position (i,j) of the kth channel. y_{avg}^{se} and $y_{\max}^{se}$ denote the output of the average pooling and maximum pooling, respectively. H and W denote the height and width, respectively.

For the purpose of fully understanding the interrelationships between different spectral bands and to improve the generalization ability of the model, the outputs of the average pooling layer and maximum pooling layer are directly fed into a shared MLP, which contains two fully connected (FC) layers. A new weight is assigned to each pixel through the SoftMax function. The output of the module is given as follows:

$$F_{se} = Sigmoid(MLP(y_{avg}^{se}) + MLP(y_{max}^{se})) \tag{3}$$

where F_{se} denotes the output of the spectral attention module.

Spatial Attention: The main purpose of the spatial attention module aims to enhance the spatial information of neighboring pixels that have the same class label as the center pixel, and to weaken the spatial information of pixels that have different category labels. The spatial attention module is given in Figure 4.

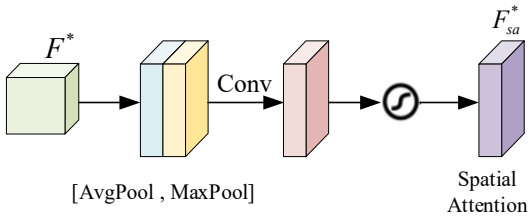


Figure 4. Spatial attention module.

To fully aggregate the spatial information, an average pooling layer and a maximum pooling layer are used to mine the target features. The spatial attention module takes the output of the spectral attention module and passes it through the maximum pooling and average pooling operations to obtain two new feature maps. Then, the information carried by the two feature maps is horizontally concatenated and input into the 7×7 convolution operation. Finally, the weight of attention is assigned to each pixel using a sigmoid function. The mathematical expressions are shown as follows:

$$y^{sa}_{avg} = \frac{1}{C}\sum_{C=1}^{C} y^*_k(i,j) \tag{4}$$

$$y^{sa}_{max} = max(y^*_k) \tag{5}$$

$$F_{sa} = Sigmoid(Conv_{7\times7}(Concat(y^{sa}_{avg}, y^{sa}_{max}))) \tag{6}$$

where y^{sa}_{avg} and y^{sa}_{max} denote the output denote the output of the average pooling and maximum pooling, respectively. F_{sa} denotes the output of the spatial attention module.

2.2. Multi-Scale Residual Spectral-Spatial Feature Extraction Module

The multi-scale information enables the effective enhancement of the robustness and increases the classification accuracy of the model [36]. Therefore, in this work, we designed a two-tier cascaded multi-scale residual spectral–spatial feature extraction module to refine the multi-scale information to obtain enhanced discriminative spectral–spatial features. Figure 5 illustrates the structure of the module.

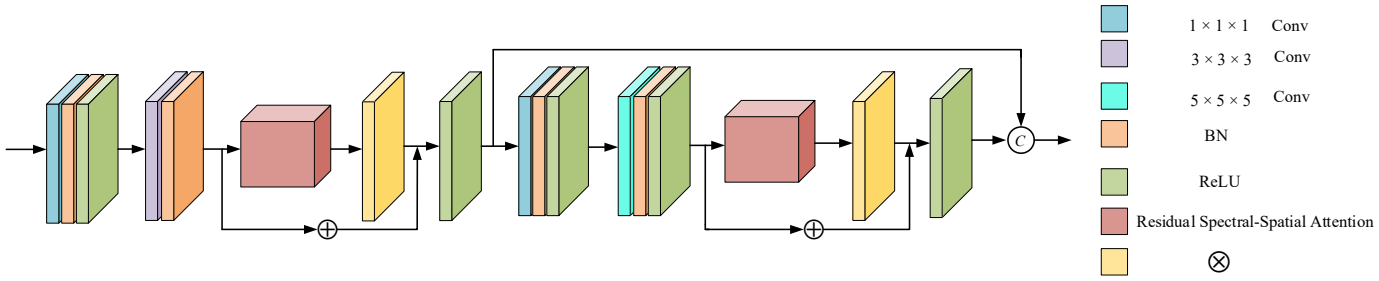


Figure 5. Multi-scale residual spectral–spatial feature extraction module.

The module uses convolution kernels of different sizes to obtain a better representation of the image to enhance the feature extraction capability of the model. In this work, we employed the $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$ convolution kernels. $1 \times 1 \times 1$ convolution is employed to extract the global information of the image, and $3 \times 3 \times 3$ and $5 \times 5 \times 5$ can provide the local information of the image under different receptive fields. The proposed model uses a 3D convolution layer after the residual spectral–spatial attention module so that the spectral–spatial features extracted from the previous residual block achieve feature fusion by 3D convolution. Based on this approach, the following residual spectral–spatial attention module can acquire both the base features and the optimized features, which is conducive to better learning of feature information by the model. Meanwhile, in order to obtain more in-depth feature information extracted by each residual spectral–spatial attention module and enrich the learning hierarchy of the network, we use residual learning

outside each residual spectral–spatial attention module to achieve the effective transfer of features and take full advantage of the independence of different features to complete the global fusion of the features obtained from different residual blocks. Finally, the features and information at different scales are fused using the Concat stitching operation to make the acquired spectral and spatial features more comprehensive.

2.3. Improved Transformer

For the purpose of further obtaining the long-distance relationship of sequential spectra, this work uses the transformer to enable the model to parse semantic information from a global perspective. However, when applied to HSI classification tasks, the transformer mainly suffers from the following two problems:

(1) Transformer architectures require large quantities of data and computational resources for training and optimization. The computational complexity of the MHSF in transformers shows quadratic growth with the input size. Therefore, using the transformer module to investigate high-resolution images can lead to reduced computational efficiency and slower model inference speed. The formula of computation can be expressed as:

$$FLOP_{MHSF} = 4HWC^2 + 2H^2W^2C \tag{7}$$

where H denotes the height of the input, W denotes the width of the input, and C denotes the number of channels in the input.

(2) Unlike CNNs, which can enhance performance by stacking additional convolutional layers, the performance of the transformer is quickly saturated when scaling to deeper layers. The difficulty of scaling the transformer is mainly caused by the attention collapse problem. As the number of transformer layers increases, the attention maps gradually become similar, and even after certain layers, the attention maps are basically the same. This situation suggests that MHSF may not be able to efficiently learn useful feature representations in deep transformer structures, resulting in the model failing to obtain the desired performance gains [30].

Based on the above two points, this paper proposes a Cross-Re-Attention mechanism to alleviate the problems of attention collapse and the huge computational burden. Generating new feature maps between the layers of the transformer enhances the diversity of each layer to avoid similarity in feature maps at deep layers. Meanwhile, considering the contextual information extraction and communication, an attention processing method on a single-channel feature map is used. The computation is significantly reduced compared to that for attention on all channels. Figure 6 illustrates the framework of improved transformer block.

Patch merging is employed to an input that is down-sampled twice, and is used to diminish the resolution and adjust the number of channels. The Cross-Re-Attention block is composed of an Inner-Patch-Re-Attention (IPRA) block and a Cross-Patch-Re-Attention (CPRA) block. By stacking IPRA blocks and CPRA blocks, the module efficiently extracts and integrates features between pixels in a patch and between patches in a feature map. The IPRA part performs pixel-by-pixel Re-Attention computation within each patch to obtain information. Attention computation is performed pixel by pixel within each patch pixel, aiming to capture and utilize the relationship between pixels within the patch to obtain global information. This strategy not only significantly reduces the computational burden, but also greatly enhances the inference efficiency of the model. The mathematical expression of computation is as follows:

$$FLOP_{IPRA} = 4HWC^2 + 2N^2HWC \tag{8}$$

where N denotes the size of the patch in IPRA. Compared to the MHSA in the standard transformer, the computational complexity is reduced from quadratic correlation to linear correlation.

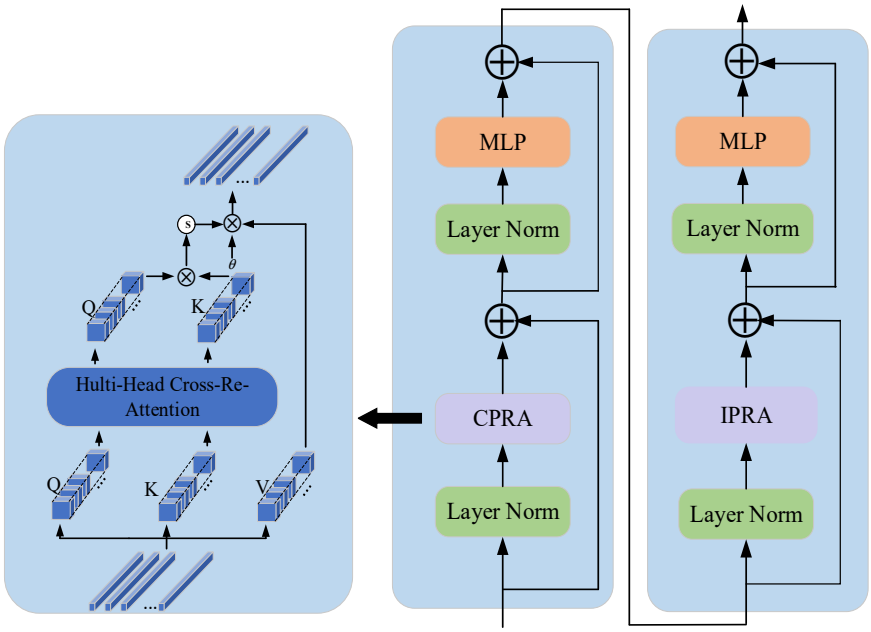


Figure 6. The internal structure of the improved transformer block.

In CNN-based networks, although the perceptual field can be expanded by stacking convolutional kernels, its sparse connectivity restricts its global dependency capture and makes it difficult to expand the perceptual field to the global range. However, in a transformer, the feature map having a single channel inherently encompasses global information. CPRA partially takes an individual channel as one of the group inputs. Re-Attention is performed in one group to cross the information of different patches to obtain global semantic information. Meanwhile, the attention maps are regenerated in layers of the transformer to enhance their diversity on different layers.

By virtue of the Cross-Re-Attention mechanism, the existing transformer model can be trained to obtain deep transformer models with linear growth in computation. Specifically, the method is based on the head-generated attention maps and generates new attention maps through dynamic aggregation. A learnable matrix, $\boldsymbol{\theta}$, is defined. This matrix is then used to map attention to a regenerated new matrix, which is multiplied with the $\boldsymbol{V}$ matrix in the transformer as follows:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Norm(\boldsymbol{\theta}^T(Softmax(\frac{\boldsymbol{\theta}\boldsymbol{K}^T}{\sqrt{d}})))\boldsymbol{V} \tag{9}$$

where d indicates the dimension of $\boldsymbol{K}$. The Norm function is employed to reduce the layer-wise variance. The SoftMax function is employed to compute the weights on the values. $\boldsymbol{Q}$ (Query), $\boldsymbol{K}$ (Key), and $\boldsymbol{V}$ (Value) are the projections of tokens, which are the matrices obtained by multiplying the input vectors with the weight matrices obtained after training.

3. Results

For the purpose of validating the performance of the RSSAT model, three public HSI datasets were selected, namely Indian Pines, Kennedy Space Center (KSC), and XuZhou datasets. To better understand the RSSAT structure, we used ablation experiments to investigate the validity of each component of the model by removing different modules. Meanwhile, we visualized the HSI classification maps to compare the feature extraction capabilities of the proposed RSSAT and other SOTA methods.

3.1. Dataset Description and Experiment Design

(1) The Indian Pines dataset was imaged by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) in 1992. A region of size 145 × 145 was selected for annotation

and used as a test dataset for HSI classification. The imaging wavelength range of the dataset is 0.4–2.5 µm, and it can continuously provide images at 220 hyperspectral bands with a spatial resolution of 20 m/pixel. Since bands 104–108, 150–163, and 220 cannot be reflected by water, these 20 bands are typically excluded from the research process, leaving only 200 bands for analysis. The dataset contains 10,249 labeled samples and 16 vegetation classes. It is worth mentioning that the number of samples for these 16 classes of ground objects is unevenly distributed, making the dataset prone to mixed pixels, which poses a challenge for classification.

(2) The KSC dataset was acquired by the National Aeronautics and Space Administration (NASA) Airborne Visible/Infrared Imaging Spectrometer (VIRIS) instrument. The dataset covers 224 spectral bands and 13 classes, and an area of 512 × 614 pixels has been specially selected for detailed labeling to ensure the accuracy and usefulness of the data. It has an excellent spectral resolution of 10 nm, ranging from 0.4 to 2.5 µm, capturing subtle spectral differences and providing strong support for analysis. Meanwhile, the 18 m spatial resolution ensures the spatial accuracy of the image, fully demonstrating the spatial characteristics of the features.

(3) The XuZhou dataset was acquired in November 2014 in XuZhou City, Jiangsu Province, China. The test area size of 500 × 260 pixel with 436 bands was selected for labeling to ensure classification accuracy. The test area, which is located near a coal mining area, has been categorized into nine classes of ground objects.

Tables 1–3 report the information of the classes and number of available samples. The Indian Pines and KSC datasets have 10,249 and 5211 labeled samples, respectively, while the XuZhou dataset has 68,877 labeled samples. Compared to the Indian Pines and KSC datasets, the XuZhou dataset has a larger number of samples. Therefore, this work designed different proportions of labeled samples as training strategies for all datasets and used different numbers of training samples to validate the performance of the RSSAT method. On the Indian pines and KSC datasets, 20% of labeled samples were randomly selected for training, 10% of labeled samples for validation, and 70% of labeled samples as the testing set. For the XuZhou dataset, 10%, 10%, and 80% of the labeled pixels were randomly selected as the training set, validation set, and test set, respectively.

Table 1. Details of the Indian Pines dataset.

No.	Class	Color	Sample Numbers	False-Color Map	Ground-Truth Map
1	Alfalfa		46		
2	Corn-notill		1428		
3	Corn-mintill		830		
4	Corn		237		
5	Grass-pasture		483		
6	Grass-trees		730		
7	Grass-pasture-mowed		28		
8	Hay-windrowed		478		
9	Oats		20		
10	Soybean-notill		972		
11	Soybean-mintill		2455		
12	Soybean-clean		593		
13	Wheat		205		
14	Woods		1265		
15	Buildings-Grass-Trees-Drives		386		
16	Stone-Steel-Towers		93		
	Total		10,249		

Table 2. Details of the KSC dataset.

No.	Class	Color	Sample Numbers	False-Color Map	Ground-Truth Map
1	Scrub		1997		
2	Willow		3726		
3	Palm		1976		
4	Pine		1394		
5	Broadleaf		2678		
6	Hardwood		3979		
7	Swap		3579		
8	Graminoid		11,213		
9	Spartina		6197		
10	Cattail		3249		
11	Salt		1058		
12	Mud		1908		
13	Water		909		
	Total		5211		

Table 3. Details of the XuZhou dataset.

No.	Class	Color	Sample Numbers	False-Color Map	Ground-Truth Map
1	Bareland-1		26,396		
2	Lakes		4027		
3	Coals		2783		
4	Cement		5214		
5	Crops-1		13,184		
6	Trees		2436		
7	Bareland-2		6990		
8	Crops-2		4777		
9	Red-tiles		3070		
	Total		68,877		

3.2. Experiment Configuration

For a fair comparison, our experiments were conducted on an Intel (R) Xeon (R) CPU E5–2620 v4 @ 2.10 GHz processor, 128 GB RAM, and an NVIDIA GeForce RTX 2080Ti (GPU), Window10, using the PyTorch framework and the Python 3.7 compiler. In order to minimize the errors and contingencies of the experiments, all experimental results are the average of 10 experiments. For model training, all experiments used batch processing. We set the training batch size to 32 × 32. Meanwhile, the Adam optimizer was employed to learn the weights and the original learning rate was set to 0.003. To ensure that the model can be adequately trained and perform optimally, we set the maximum iteration time to 200 epochs. We set an early stopping strategy to avoid the overfitting problem.

3.3. Experiment Comparison and Analysis

In the study, the classification performance of the proposed RSSAT was verified by comparison with the SVM [37], 3D-CNN [14], residual neural network (ResNet) [38], Multi-Attention Fusion Network (MAFN) [39], Spectral–Spatial Feature Tokenisation Transformer (SSFTT) [26], SSRN [18], and Dual-View Spectral and Global Spatial Feature Fusion Network (DSGSF) [21] methods. SVM is a traditional image classification method. The remaining models are deep learning-based algorithms, which utilize deep neural networks

to process HSI classification tasks. Among these, the experimental results were quantitatively evaluated by three metrics: overall accuracy (OA), average accuracy (AA), and Kappa (K) coefficient. OA represents the percentage of correctly classified pixels in all pixel classification results. The Kappa coefficient is employed to test uniformity and determine whether the prediction results of the method are identical to the true results. The value of the Kappa coefficient ranges from -1 to 1, where a positive value indicates superior classification performance, a negative value indicates poor classification performance, and a value close to 0 indicates average classification performance.

Quantitative classification results for evaluation indicators and the accuracy of each class are given in Tables 4–6, respectively (the standard deviation of ten runs was taken as the experimental result). Overall, it could be observed from all experimental results of the datasets that our proposal yields the best accuracy and relatively low standard deviations. Specifically, on the Indian Pines dataset, RSSAT achieves 97.31% in terms of AA; at the same time SVM, 3D-CNN, ResNet, MAFN, SSFTT, and SSRN achieve 79.76%, 76.95%, 92.64%, 96.65%, 96.08%, and 92.63%, respectively. On the Indian Pines, KSC, and XuZhou datasets, compared to SSRN, the increases in OA of our proposal are 0.30%, 0.44%, and 0.09%. The proposed RSSAT method consistently demonstrates superiority in performance compared to SSRN, which is a strong argument for the superiority of our method in improving the representation of specific spectral–spatial features by readjusting the high spatial correlation contexts over spectral bands. Moreover, in the Indian Pines dataset, 16 classes of samples are unevenly distributed in terms of quantity. For example, there are only 20 labeled samples for the 9th class (Oats), while the 11th class (Soybean-mintill) contains 2455 labeled samples. The uneven sample distribution presents a serious challenge to the HSI classification. For the accuracy of the 9th class, SSFTT (67.22 ± 7.29), SSRN (58.94 ± 48.22), and other methods with better performance still fail to provide a good solution.

Table 4. Quantitative classification performance of different methods on the Indian Pines dataset.

Methods	SVM	3D-CNN	ResNet	MAFN	SSFTT	SSRN	RSSAT
1	59.23 ± 19.05	81.73 ± 7.04	95.17 ± 10.06	95.57 ± 6.10	99.76 ± 0.73	77.21 ± 31.4	93.66 ± 9.42
2	71.14 ± 1.39	66.34 ± 7.96	90.65 ± 6.68	98.17 ± 2.11	94.85 ± 1.15	98.51 ± 1.36	98.80 ± 0.82
3	74.59 ± 1.81	72.20 ± 15.21	91.16 ± 5.82	96.68 ± 5.50	99.24 ± 0.47	97.81 ± 1.24	98.63 ± 0.75
4	60.65 ± 7.43	83.69 ± 7.94	92.25 ± 9.29	97.94 ± 2.62	99.30 ± 1.08	99.37 ± 0.84	98.64 ± 1.19
5	89.19 ± 3.15	94.91 ± 1.20	98.41 ± 1.17	98.58 ± 1.05	98.78 ± 0.94	95.91 ± 1.76	98.18 ± 1.60
6	88.63 ± 1.56	96.51 ± 1.37	97.05 ± 2.96	98.56 ± 1.88	99.37 ± 0.39	98.78 ± 1.43	99.22 ± 0.84
7	85.71 ± 8.31	88.56 ± 3.68	80.91 ± 31.15	94.94 ± 7.81	98.40 ± 3.20	70.01 ± 48.2	97.66 ± 7.00
8	90.19 ± 1.64	99.10 ± 0.85	96.08 ± 2.93	99.71 ± 0.49	99.79 ± 0.45	98.46 ± 2.27	99.82 ± 0.53
9	74.12 ± 13.49	64.01 ± 17.76	86.00 ± 24.67	81.89 ± 13.98	67.22 ± 7.29	58.94 ± 48.22	85.32 ± 13.57
10	75.28 ± 2.04	82.57 ± 5.35	91.40 ± 6.44	96.13 ± 3.54	97.54 ± 0.89	97.54 ± 1.61	98.12 ± 1.09
11	78.16 ± 1.19	64.15 ± 8.16	92.31 ± 5.22	98.61 ± 1.94	99.22 ± 0.35	99.22 ± 0.49	98.96 ± 0.73
12	72.68 ± 3.94	81.93 ± 5.41	90.8 ± 7.81	96.59 ± 3.26	95.96 ± 1.09	98.52 ± 0.95	98.09 ± 1.41
13	92.23 ± 2.85	99.02 ± 0.45	95.76 ± 4.37	98.92 ± 1.41	98.86 ± 0.71	98.67 ± 4.04	98.48 ± 2.19
14	91.66 ± 0.95	89.61 ± 4.56	93.38 ± 4.52	99.45 ± 0.36	99.38 ± 0.88	99.17 ± 1.01	99.39 ± 0.44
15	74.67 ± 6.95	87.44 ± 4.07	94.99 ± 5.48	97.94 ± 2.65	98.01 ± 1.21	99.22 ± 0.88	98.32 ± 1.62
16	98.10 ± 2.53	93.97 ± 5.14	95.58 ± 3.71	95.63 ± 3.64	91.69 ± 5.79	94.79 ± 6.14	95.68 ± 4.34
OA (%)	79.92 ± 0.65	77.15 ± 2.69	92.40 ± 1.81	97.98 ± 0.38	98.07 ± 0.39	98.41 ± 0.44	98.71 ± 0.28
AA (%)	79.76 ± 2.43	76.95 ± 2.54	92.64 ± 3.40	96.65 ± 1.10	96.08 ± 1.44	92.63 ± 6.20	97.31 ± 1.07
K $\times$ 100	76.99 ± 0.75	73.76 ± 2.96	91.31 ± 2.08	97.70 ± 0.43	97.85 ± 0.50	98.22 ± 0.51	98.53 ± 0.32

In our proposed RSSAT method, we use a two-tier cascaded multi-scale residual spectral–spatial feature-learning module by introducing a spectral–spatial attention mechanism. Meanwhile, we strategically embed ResBlock to enhance the nonlinear representation capability. The module mitigates the information loss in the feature stream, preserves more spatial information, and better addresses the challenge of scale diversity under different land cover types. Therefore, RSSAT (85.32 ± 13.57)% achieves the best classification results

on the 9th class. At the same time, we obtained the best classification performance in terms of overall evaluation metrics, and obtained the closest classification maps to the ground truth.

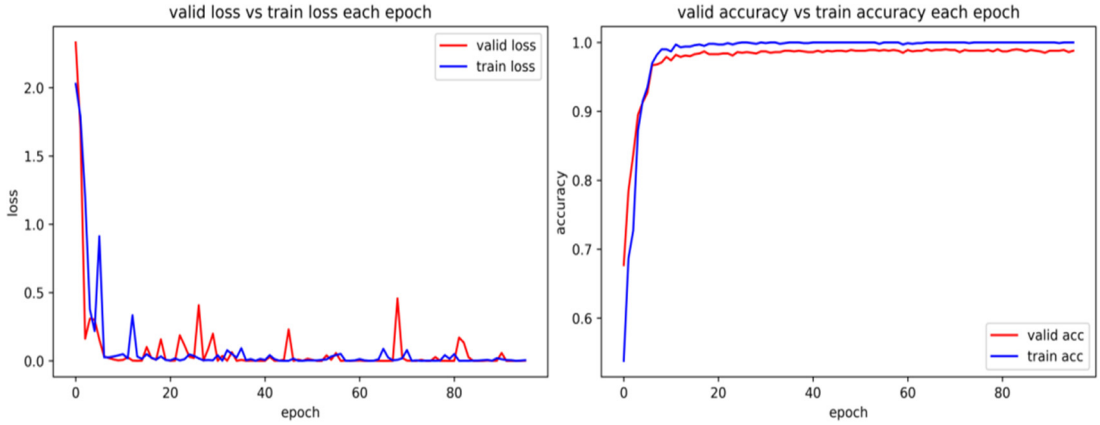
Table 5. Quantitative classification performance of different methods on the KSC dataset.

Method	SVM	3D-CNN	ResNet	MAFN	SSFTT	SSRN	RSSAT
1	92.81 ± 0.79	94.29 ± 1.81	94.27 ± 7.83	97.85 ± 2.41	99.91 ± 0.14	98.95 ± 0.14	99.72 ± 0.56
2	86.62 ± 5.10	91.52 ± 4.65	91.07 ± 9.95	93.82 ± 4.88	93.89 ± 4.30	95.97 ± 9.53	94.21 ± 10.00
3	73.33 ± 8.35	88.81 ± 8.04	84.39 ± 10.48	84.62 ± 10.1	98.04 ± 2.82	97.32 ± 3.07	98.85 ± 1.94
4	54.48 ± 8.64	80.55 ± 9.36	74.18 ± 10.06	75.65 ± 15.45	98.21 ± 1.64	97.63 ± 2.26	98.57 ± 1.78
5	60.22 ± 12.13	81.06 ± 5.16	66.39 ± 20.92	81.01 ± 9.92	98.30 ± 2.36	98.63 ± 2.76	97.88 ± 4.08
6	65.47 ± 8.34	85.64 ± 8.41	86.83 ± 21.62	88.64 ± 21.18	99.88 ± 0.22	100 ± 0.00	99.94 ± 0.16
7	76.21 ± 3.83	92.76 ± 13.2	80.35 ± 28.52	88.19 ± 15.34	99.77 ± 0.68	83.60 ± 33.76	98.52 ± 3.43
8	86.60 ± 5.03	95.88 ± 1.95	92.10 ± 11.31	97.20 ± 2.56	99.82 ± 0.42	99.76 ± 0.21	99.74 ± 0.35
9	88.45 ± 2.66	97.39 ± 1.31	91.30 ± 12.74	94.32 ± 6.49	99.97 ± 0.07	99.85 ± 0.35	100.00 ± 0.00
10	96.30 ± 4.94	99.91 ± 0.20	98.51 ± 1.67	99.18 ± 0.91	99.96 ± 00.09	100 ± 0.00	100.00 ± 0.00
11	96.16 ± 1.52	98.12 ± 1.86	97.45 ± 7.33	99.76 ± 0.44	99.91 ± 0.26	99.63 ± 1.09	100.00 ± 0.00
12	93.61 ± 2.67	98.12 ± 1.86	93.50 ± 12.89	96.12 ± 2.70	99.89 ± 0.30	98.94 ± 2.46	99.87 ± 0.25
13	99.68 ± 0.68	99.97 ± 0.57	99.36 ± 0.77	99.68 ± 0.85	100 ± 0.00	100 ± 0.00	100.00 ± 0.00
OA (%)	87.94 ± 1.57	95.49 ± 0.25	91.37 ± 4.92	94.09 ± 4.00	99.24 ± 0.14	98.89 ± 1.52	99.33 ± 0.57
AA (%)	82.30 ± 2.49	92.01 ± 0.50	88.44 ± 7.97	92.41 ± 4.34	98.58 ± 0.22	97.79 ± 3.43	99.02 ± 0.76
K × 100	86.57 ± 1.75	94.86 ± 0.28	90.36 ± 5.54	93.44 ± 4.42	99.11 ± 0.15	98.77 ± 1.69	99.25 ± 0.64

Table 6. Quantitative classification performance of different methods on the XuZhou dataset.

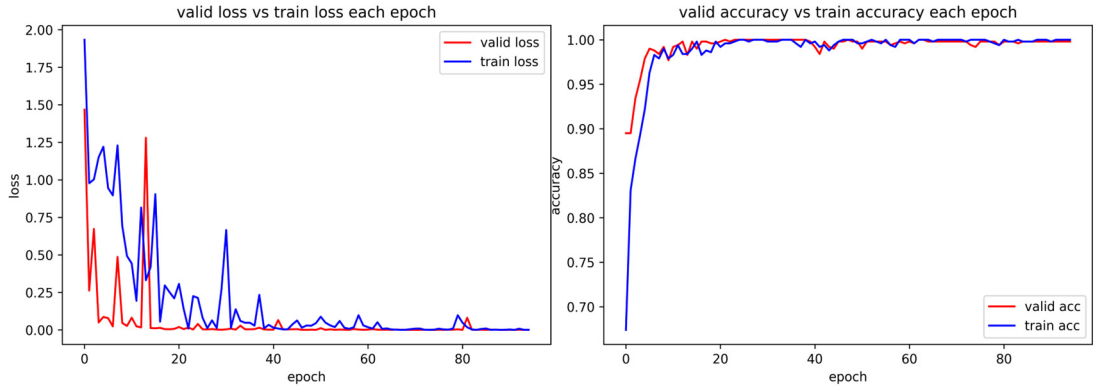
Methods	SVM	3D-CNN	ResNet	MAFN	SSFTT	SSRN	RSSAT
1	96.98 ± 0.24	88.01 ± 4.82	99.75 ± 0.12	99.15 ± 0.51	99.59 ± 0.21	99.59 ± 0.29	99.91 ± 0.04
2	99.91 ± 0.05	94.96 ± 9.80	99.94 ± 0.06	99.46 ± 0.64	99.98 ± 0.03	99.91 ± 0.05	99.98 ± 0.02
3	95.47 ± 0.45	92.13 ± 5.09	99.96 ± 0.06	95.95 ± 6.68	99.54 ± 0.25	99.04 ± 0.47	100.00 ± 0.00
4	97.79 ± 0.69	89.12 ± 2.53	99.56 ± 0.21	97.41 ± 3.31	99.92 ± 0.08	99.82 ± 0.16	99.98± 0.02
5	96.14 ± 0.19	92.92 ± 2.04	98.39 ± 0.78	97.71 ± 1.35	99.56 ± 0.20	99.92 ± 0.10	99.19± 0.26
6	89.81 ± 0.62	89.31 ± 1.92	96.77 ± 2.27	94.65 ± 3.98	99.64 ± 0.17	99.78 ± 0.10	99.69 ± 0.23
7	90.27 ± 0.41	84.24 ± 1.47	97.21 ± 2.78	97.75 ± 1.38	99.90 ± 0.05	99.64 ± 0.41	99.81 ± 0.11
8	98.68 ± 0.17	92.64 ± 4.94	95.91 ± 9.54	98.11 ± 2.09	99.94 ± 0.13	98.23 ± 0.44	99.62 ± 0.13
9	98.38 ± 0.33	99.36 ± 0.33	99.22 ± 0.84	95.85 ± 7.96	99.62 ± 0.16	99.47 ± 0.42	99.33± 0.25
OA (%)	96.23 ± 0.08	90.06 ± 2.67	98.72 ± 0.93	98.01 ± 1.34	99.68 ± 0.06	99.63 ± 0.09	99.72 ± 0.05
AA (%)	95.94 ± 0.16	87.57 ± 1.77	98.52 ± 1.19	97.34 ± 2.32	99.72 ± 0.04	99.49 ± 0.15	99.73 ± 0.05
K ×100	95.21 ± 0.11	87.51 ± 3.22	98.38 ± 1.18	97.48± 1.68	99.60 ± 0.09	99.51± 0.12	99.64 ± 0.07

In addition, Figures 7–9 show the learning curves of the proposed method. On the learning curves of these three datasets, as the number of epochs increases, both the loss values and accuracy tend to have a smoothed output. The maximum fluctuation in loss values is less than 0.5, effectively demonstrating the excellent convergence performance of the model. Meanwhile, the gradual fitting of the accuracy curves in the figure visually demonstrates the remarkable generalization ability of the model. Based on the above analysis, our method exploits complementary hybrid blocks to enable the efficient characterization of the deep spectral–spatial features.



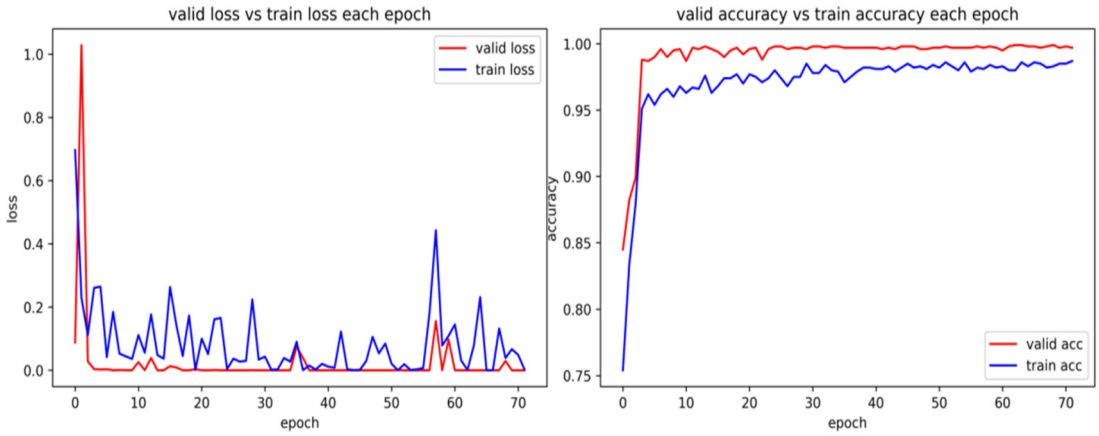
(**a**) (**b**)

Figure 7. Learning curves for the Indian Pines dataset. (**a**) Valid loss vs. train loss in each epoch. (**b**) Valid accuracy vs. train accuracy in each epoch.



(**a**) (**b**)

Figure 8. Learning curves for the KSC dataset. (**a**) Valid loss vs. train loss in each epoch. (**b**) Valid accuracy vs. train accuracy in each epoch.



(**a**) (**b**)

Figure 9. Learning curves for the XuZhou dataset. (**a**) Valid loss vs. train loss in each epoch. (**b**) Valid accuracy vs. train accuracy in each epoch.

3.4. Visualization of Classification Maps

In order to visually demonstrate the effectiveness of the RSSAT method, we analyzed the classification results over the Indian Pines, KSC, and XuZhou datasets, as shown in

Figures 10–12. These classification maps display that RSSAT has fewer misclassified pixels and cleaner boundaries than other SOTA models. Therefore, we can conclude that the RSSAT method outperforms all the methods for classification.

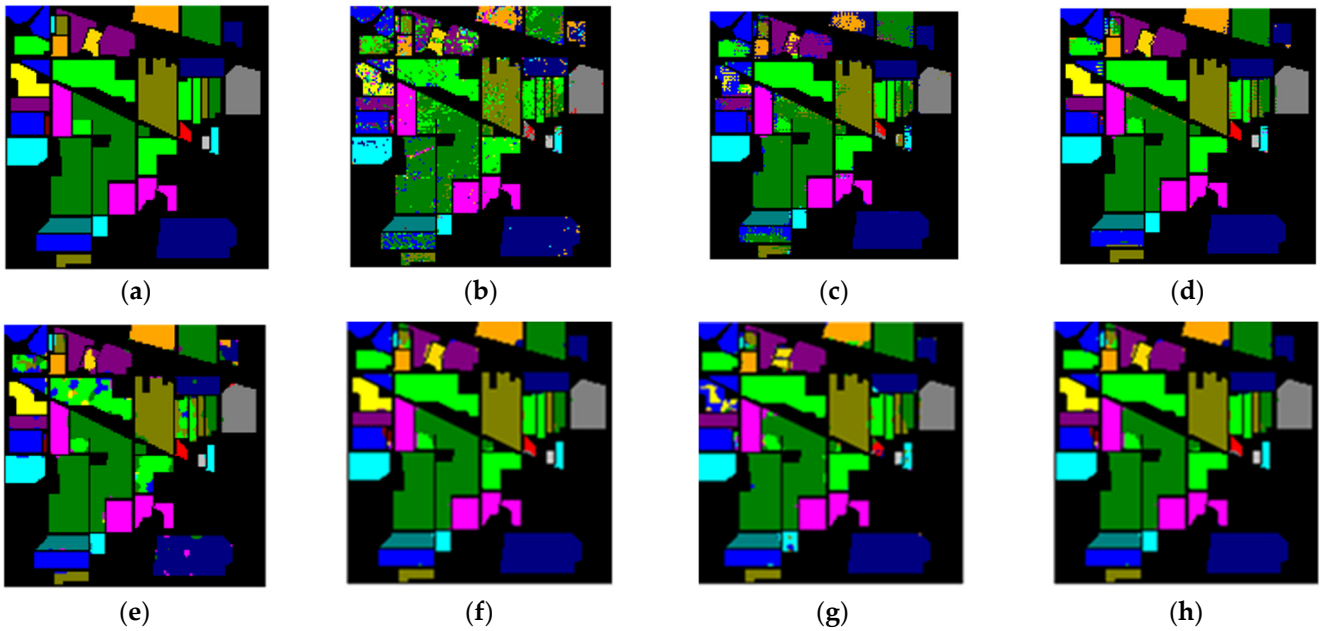


Figure 10. Classification maps of the Indian Pines dataset. (**a**) Ground truth. (**b**) SVM. (**c**) CNN. (**d**) ResNet. (**e**) MAFN. (**f**) SSFTT. (**g**) SSRN. (**h**) RSSAT.

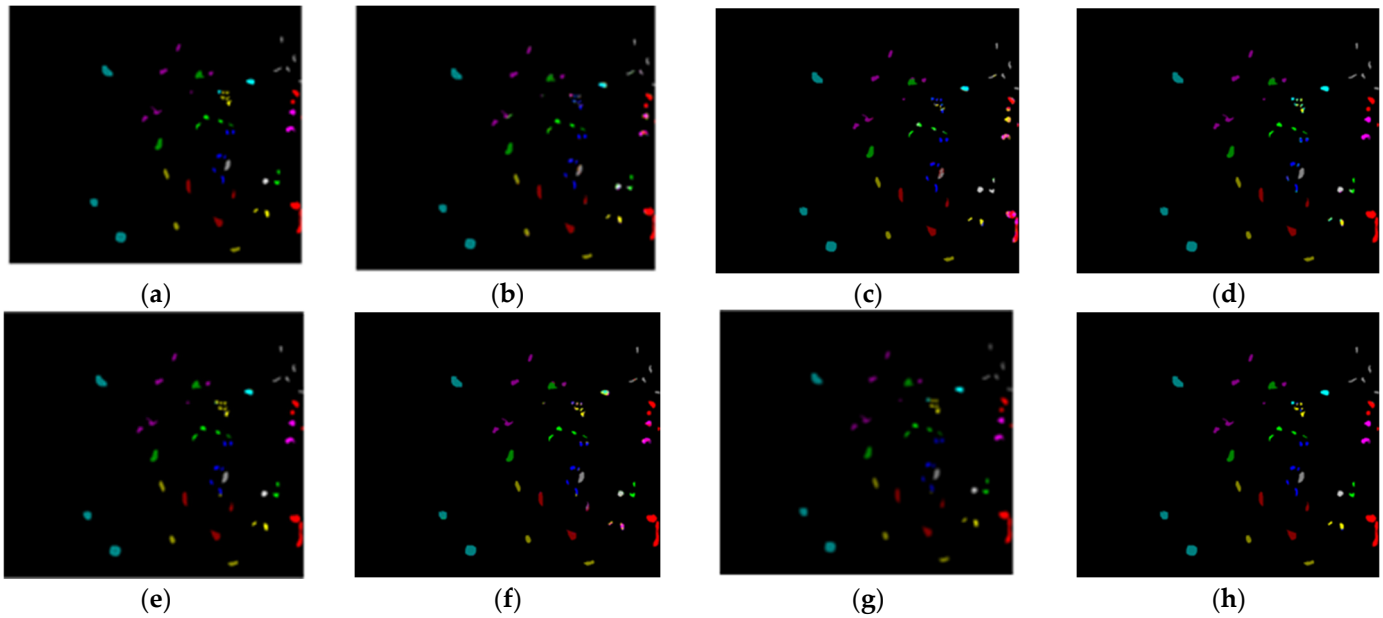


Figure 11. Classification maps of the KSC dataset. (**a**) Ground truth. (**b**) SVM. (**c**) CNN. (**d**) ResNet. (**e**) MAFN. (**f**) SSFTT. (**g**) SSRN. (**h**) RSSAT.

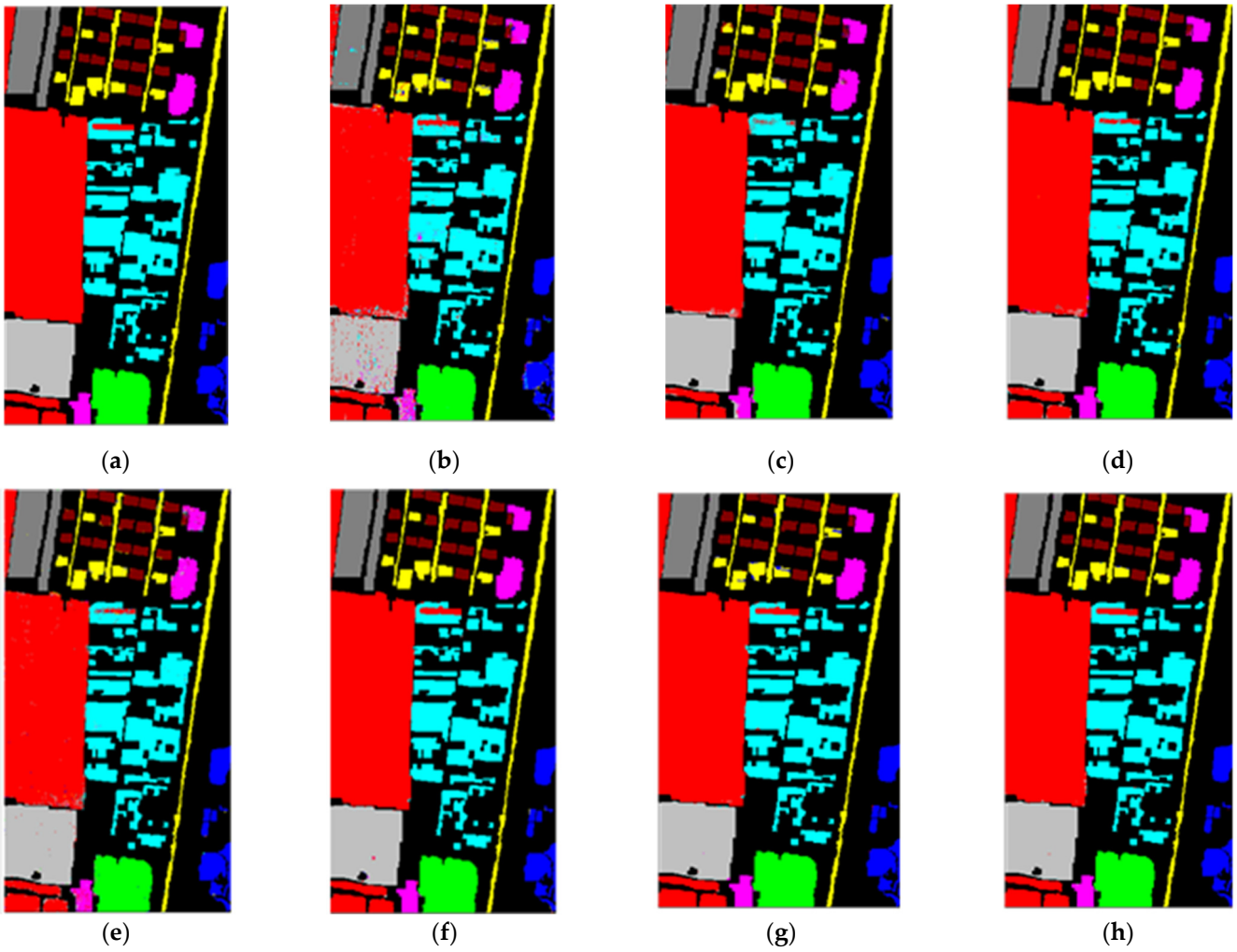
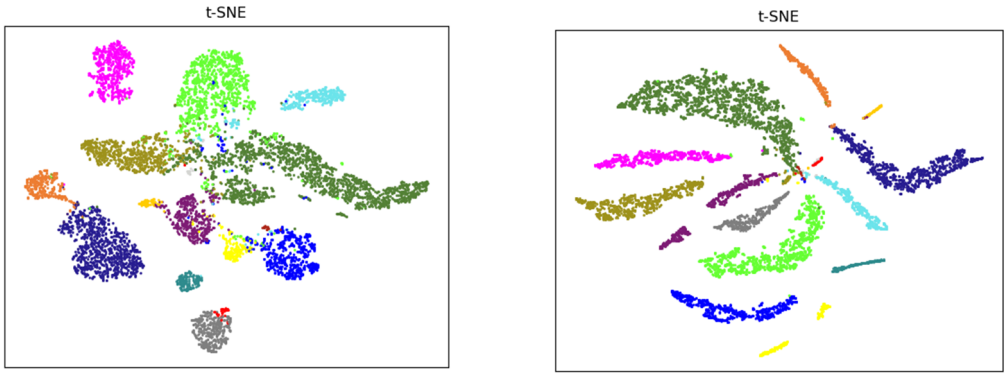


Figure 12. Classification maps of the XuZhou dataset. (**a**) Ground truth. (**b**) SVM. (**c**) CNN. (**d**) ResNet. (**e**) MAFN. (**f**) SSFTT. (**g**) SSRN. (**h**) RSSAT.

4. Discussions

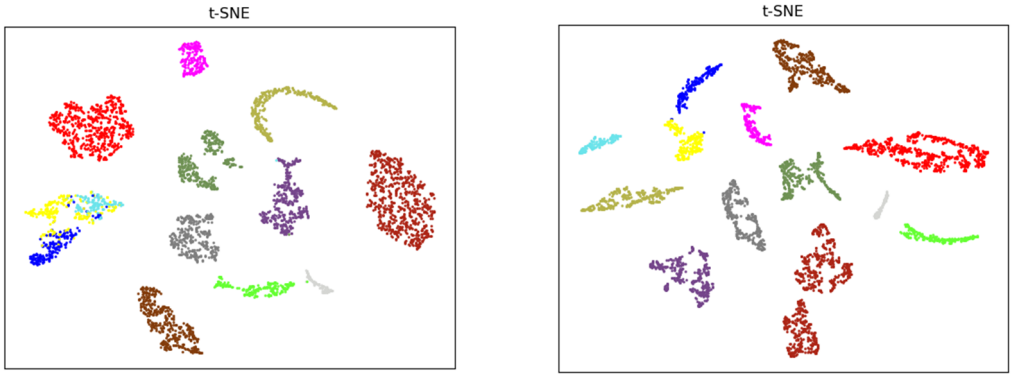
4.1. Feature Visualization Analysis

For the purpose of investigating the feature representation capability of RSSAT, the t-distributed stochastic neighborhood embedding (t-SNE) algorithm [40] was used to visualize and compare the features extracted by ResNet and RSSAN in 2D space. As shown in Figures 13–15, the samples belonging to the same class are clearly clustered into a group in the figures, while samples of different classes are easily separated from each other. From the visualization results, the RSSAT method is more significant and effective in clustering the features, which further proves that the method gains the abstract representation of spectral–spatial features for HSIs.



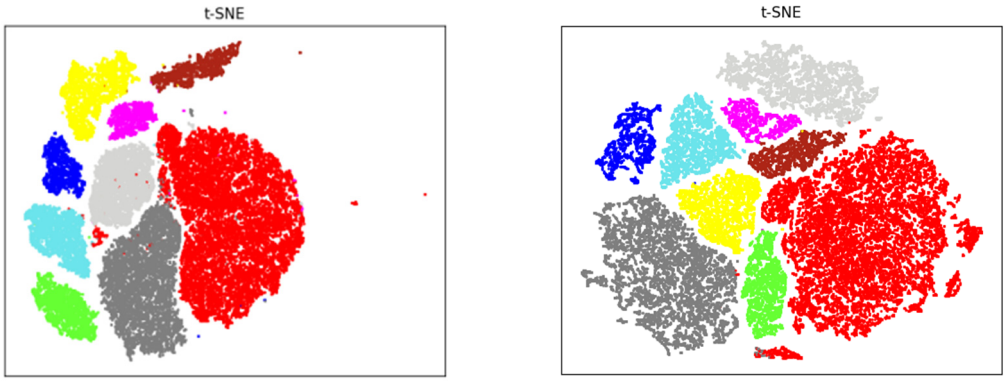
(**a**) (**b**)

Figure 13. Visualization of the 2D spectral–spatial features for the samples in the Indian Pines dataset via t-SNE. (**a**) ResNet. (**b**) RSSAT.



(**a**) (**b**)

Figure 14. Visualization of the 2D spectral–spatial features for the samples in the KSC dataset via t-SNE. (**a**) ResNet. (**b**) RSSAT.



(**a**) (**b**)

Figure 15. Visualization of the 2D spectral–spatial features for the samples in the XuZhou dataset via t-SNE. (**a**) ResNet. (**b**) RSSAT.

4.2. Time Cost Comparison

In order to comprehensively evaluate the efficiency of different methods in the HSI classification task, the running time and computational cost of each method are recorded in detail in Table 7. As seen from the data in the table, the training time of RSSAT is slightly longer compared to that of 3D-CNN, SSFTT, and SSRN. This is mainly attributed to the complexity of the RSSAT model design, which contains more layers, thus increasing the length of the training process to some extent. However, it is worth noting that RSSAT

exhibits a significant advantage in classification accuracy. This performance enhancement, especially in the accurate classification of small target samples, compensates for its minor shortfall in training time. This balance between performance and efficiency of RSSAT is reasonable considering that classification accuracy is often a crucial metric in practical applications. Meanwhile, RSSAT shows significant advantages in both efficiency and performance compared to ResNet and MAFN. This further demonstrates that RSSAT is able to achieve superior classification performance with moderate computational cost, providing an efficient and feasible solution for the HSI classification task.

Table 7. Training time in minutes (m) and test time in seconds (s) between the comparison methods and the RSSAT method for three datasets.

Methods	Indian Pines			KSC			XuZhou		
	Params (M)	Training (m)	Test (s)	Params (M)	Training (m)	Test (s)	Params (M)	Training (m)	Test (s)
3D-CNN	0.26	1.87	4.43	0.14	1.21	2.21	1.38	2.71	3.02
ResNet	83.58	15.36	13.06	83.29	8.08	6.41	86.39	28.65	7.78
MAFN	2.11	12.45	6.22	1.88	7.62	5.94	2.99	19.21	12.15
SSFTT	0.61	3.47	3.18	0.42	2.64	8.41	1.06	5.12	5.47
SSRN	1.39	11.50	4.67	1.25	5.09	2.26	2.77	16.34	5.41
RSSAT	1.61	12.07	3.31	1.45	8.91	7.10	2.85	18.55	9.27

Overall, although RSSAT may not be the optimal choice from the perspectives of execution time and computational cost, its high-precision overall classification performance and its ability to accurately recognize small target samples make up for these shortcomings.

4.3. Different Numbers of Training Samples

In order to be closer to real-world application scenarios and to test the generalization ability of the model under limited data, we reduced the ratio of training samples to validation samples. The experimental results are shown in Table 8. Specifically, we randomly selected 5% of the samples in the Indian Pines dataset as the training set, 5% of the samples as the validation set, and the remaining samples as the test set. From the experimental results, the performance of each method shows a different degree of degradation as the number of samples is reduced. Compared with other methods, RSSAT still has obvious advantages with fewer samples, which proves that RSSAT has superior generalization ability.

Table 8. Classification performance under 5% training samples for Indian Pines dataset.

Method	ResNet	MAFN	SSFTT	SSRN	DSGSF	RSSAN
OA (%)	92.87	95.75	95.26	95.14	97.68	97.61
AA (%)	87.44	94.34	95.87	76.30	94.29	95.03
K × 100	91.83	95.14	89.86	94.45	97.36	97.38

4.4. Ablation Experiments Analysis

In this experiment, we still used the three datasets as examples to perform ablation experiments to investigate the gain in each component when using our RSSAT by removing different modules. The relevant results are reported in Table 9.

(1) In this work, we employed the SSRN model with multi-scale information integration as the basic model architecture (the experimental model was defined as Base).
(2) For the purpose of verifying the validity of the residual spectral–spatial attention module over RSSAT, the experiment only increased the improved transformer based on Base (the experimental model was defined as Base+IT).
(3) For the purpose of verifying the validity of the improved transformer module over RSSAT, the experiment only increased the multi-scale residual spectral–spatial attention module based on Base (the experimental model was defined as Base+RSS).

Table 9. Ablation experiments for each component.

Dataset	Index	Base	Base+IT	Base+RSS	Base+RSS+IT
Indian Pines	OA (100%)	98.36 ± 0.47	98.53 ± 0.20	98.62 ± 0.51	98.71 ± 0.28
	AA (100%)	95.24 ± 2.79	97.25 ± 0.54	97.53 ± 1.32	97.31 ± 1.07
	K × 100	98.13 ± 0.53	98.32 ± 0.33	98.42 ± 0.58	98.53 ± 0.32
KSC	OA (100%)	98.68 ± 0.53	99.25 ± 0.60	99.11 ± 0.35	99.33 ± 0.57
	AA (100%)	98.19 ± 2.23	98.82 ± 1.10	98.71 ± 0.32	99.02 ± 0.76
	K × 100	98.53 ± 2.04	99.17 ± 0.67	99.01 ± 0.39	99.25 ± 0.64
XuZhou	OA (100%)	99.13 ± 0.63	99.70 ± 0.07	99.29 ± 0.03	99.72 ± 0.05
	AA (100%)	98.72 ± 1.14	99.56 ± 0.14	98.99 ± 0.04	99.73 ± 0.05
	K × 100	99.01 ± 0.81	99.62 ± 0.08	99.10 ± 0.04	99.64 ± 0.07

Specifically, Base+IT increased OA by 0.17%, 0.57%, and 0.57% over the Indian Pines, KSC, and XuZhou datasets, respectively, which showed that the transformer adequately captured contextual information, enabling the network to parse semantic information from a global perspective. Base+RSS improved OA by 0.26%, 0.43%, and 0.16% on different datasets, demonstrating that the multi-scale residual spectral–spatial feature extraction module helped the architecture to adaptively learn the important features of each spectral–spatial domain while emphasizing the information-rich features and suppressing less useful features.

5. Conclusions

In the paper, a novel hybrid architecture is examined for HSI classification. Specifically, the proposed RSSAT method improves the representational ability of extracted features and captures relationships within a long range in the spectral domain by combining the strengths of a transformer and a CNN. For the RSSAT method, the residual spectral–spatial attention mechanism is embedded in the multi-scale feature-learning part for the joint extraction of spectral and spatial features on the selected multi-scale feature maps to highlight the discriminative information. For the characteristics of the HSI spectral approximation continuation, we propose the Cross-Re-Attention mechanism to improve the formal transformer to achieve deeper ViT training, which effectively alleviates the ViT attention collapse problem and computational volume problem. Overall, RSSAT successfully extracts discriminative features in complex regions and significantly enhances remote contextual information in the spectral domain. The classification performance is evaluated on three challenging datasets. The overall accuracy of the RSSAT model was 98.71%, 99.33%, and 99.72%, and average accuracy was 97.31%, 99.02%, and 99.72%, for the Indian Pines, KSC, and XuZhou datasets, respectively.

Since the number of samples in the Indian Pines dataset is small and unevenly distributed, there is still room for improvement in the classification performance of the RSSAT model. In future work, we will study methods such as data expansion, loss constraints between features and HSI data, and transformer optimization to facilitate the classification performance of a small-sample HSI dataset.

Author Contributions: Conceptualization, A.W., K.Z., H.W., Y.I. and H.C.; methodology, A.W., K.Z., H.W. and Y.I.; software, K.Z.; validation K.Z.; writing—review and editing A.W., K.Z., H.W., Y.I. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Key Research and Development Plan Project of Heilongjiang (JD2023SJ19) and the Natural Science Foundation of Heilongjiang Province (LH2023F034).

Data Availability Statement: Indian Pines and KSC: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 20 May 2011); Xuzhou: https://ieee-dataport.org/documents/xuzhou-hyspex-dataset (accessed on 2 November 2018).

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

References

1. Plaza, A.; Benediktsson, J.A.; Boardman, J.W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A.; et al. Recent advances in techniques for hyperspectral image processing. *Remote Sens. Environ.* **2009**, *113*, 110–122. [CrossRef]
2. Landgrebe, D. Hyperspectral image data analysis. *IEEE Signal Process. Mag.* **2002**, *19*, 17–28. [CrossRef]
3. Yuen, P.; Richardson, M. An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. *Imaging Sci. J.* **2013**, *58*, 241–253. [CrossRef]
4. Yang, X.; Yu, Y. Estimating soil salinity under various moisture conditions: An experimental study. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2525–2533. [CrossRef]
5. Shimoni, M.; Haelterman, R.; Perneel, C. Hypersectral Imaging for Military and Security Applications: Combining Myriad Processing and Sensing Techniques. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 101–117. [CrossRef]
6. 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.* **2012**, *9*, 447–451. [CrossRef]
7. 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]
8. Blanzieri, E.; Melgani, F. Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1804–1811. [CrossRef]
9. Xu, M.; Zhao, Q.; Jia, S. Multiview Spatial–Spectral Active Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5512415. [CrossRef]
10. Liu, Z.; Tang, B.; He, X.; Qiu, Q.; Liu, F. Class-Specific Random Forest with Cross-Correlation Constraints for Spectral–Spatial Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 257–261. [CrossRef]
11. Friedl, M.; Brodley, C. Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* **1997**, *61*, 399–409. [CrossRef]
12. Kang, X.; Zhuo, B.; Duan, P. Dual-Path Network-Based Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 447–451. [CrossRef]
13. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep Learning-Based Classification of Hyperspectral Data. *IEEE J. Sel. Top. Appl. Earth Observ Remote Sens.* **2014**, *7*, 2094–2107. [CrossRef]
14. Li, Y.; Zhang, H.; Shen, Q. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* **2017**, *9*, 67. [CrossRef]
15. 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]
16. Praveen, B.; Menon, V. Study of Spatial–Spectral Feature Extraction Frameworks With 3-D Convolutional Neural Network for Robust Hyperspectral Imagery Classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 1717–1727. [CrossRef]
17. 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]
18. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 847–858. [CrossRef]
19. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [CrossRef]
20. Zhou, H.; Luo, F.; Zhuang, H.; Weng, X.; Gong, X.; Lin, Z. Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5508614. [CrossRef]
21. Guo, T.; Wang, R.; Luo, F.; Gong, X.; Zhang, L.; Gao, X. Dual-View Spectral and Global Spatial Feature Fusion Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5512913. [CrossRef]
22. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
24. Wang, W. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14408–14419.
25. Song, R.; Feng, F.; Cheng, W.; Mu, Z.; Wang, X. BS2T: Bottleneck Spatial–Spectral Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5532117. [CrossRef]
26. 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]
27. Zhang, X.; Su, Y.; Gao, L.; Bruzzone, L.; Gu, X.; Tian, Q. A Lightweight Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5517617. [CrossRef]
28. Liu, Z. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
29. Lin, H.; Cheng, X.; Wu, X.; Shen, D. CAT: Cross Attention in Vision Transformer. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6. [CrossRef]

30. Zhou, D.; Kang, B.; Jin, X. Deepvit: Towards deeper vision transformer. *arXiv* **2021**, arXiv:2103.11886. [CrossRef]
31. Graham, B. 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. 12239–12249. [CrossRef]
32. Ouyang, E.; Li, B.; Hu, W.; Zhang, G.; Zhao, L.; Wu, J. When multigranularity meets spatial–spectral attention: A hybrid transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4401118. [CrossRef]
33. Zu, B.; Li, Y.; Li, J.; He, Z.; Wang, H.; Wu, P. Cascaded convolution-based transformer with densely connected mechanism for spectral-spatial hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5513119. [CrossRef]
34. Liu, H.; Li, W.; Xia, X.G.; Zhang, M.; Gao, C.Z.; Tao, R. Central Attention Network for Hyperspectral Imagery Classification. *IEEE Trans. Neural Netw. Learn Syst.* **2023**, *34*, 8989–9003. [CrossRef]
35. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual Spectral-Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 449–462. [CrossRef]
36. Xu, F.; Zhang, G.; Song, C.; Wang, H.; Mei, S. Multiscale and Cross-Level Attention Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5501615. [CrossRef]
37. Waske, B.; van der Linden, S.; Benediktsson, J.A.; Rabe, A.; Hostert, P. Sensitivity of support vector machines to random featureselection in classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2021**, *48*, 2880–2889. [CrossRef]
38. He, K.; Zhang, M.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Li, Z. Hyperspectral Image Classification with Multi attention Fusion Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5503305. [CrossRef]
40. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.