remote sensing

MDPI

Article

Spectral–Spatial Feature Extraction for Hyperspectral Image Classification Using Enhanced Transformer with Large-Kernel Attention

Wen Lu [1], Xinyu Wang [2], Le Sun [2,3] and Yuhui Zheng [1,*]

[1] The College of Computer, Qinghai Normal University, Xining 810000, China; luwen@qhnu.edu.cn
[2] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; cs_xywang@nuist.edu.cn (X.W.)
[3] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: zheng_yuhui@nuist.edu.cn

Abstract: In the hyperspectral image (HSI) classification task, every HSI pixel is labeled as a specific land cover category. Although convolutional neural network (CNN)-based HSI classification methods have made significant progress in enhancing classification performance in recent years, they still have limitations in acquiring deep semantic features and face the challenges of escalating computational costs with increasing network depth. In contrast, the Transformer framework excels in expressing high-level semantic features. This study introduces a novel classification network by extracting spectral–spatial features with an enhanced Transformer with Large-Kernel Attention (ETLKA). Specifically, it utilizes distinct branches of three-dimensional and two-dimensional convolutional layers to extract more diverse shallow spectral–spatial features. Additionally, a Large-Kernel Attention mechanism is incorporated and applied before the Transformer encoder to enhance feature extraction, augment comprehension of input data, reduce the impact of redundant information, and enhance the model's robustness. Subsequently, the obtained features are input to the Transformer encoder module for feature representation and learning. Finally, a linear layer is employed to identify the first learnable token for sample label acquisition. Empirical validation confirms the outstanding classification performance of ETLKA, surpassing several advanced techniques currently in use. This research provides a robust and academically rigorous solution for HSI classification tasks, promising significant contributions in practical applications.

Keywords: CNN; Transformer; spectral–spatial feature; HSI

check for updates

Citation: Lu, W.; Wang, X.; Sun, L.; Zheng, Y. Spectral–Spatial Feature Extraction for Hyperspectral Image Classification Using Enhanced Transformer with Large-Kernel Attention. *Remote Sens.* **2024**, *16*, 67. https://doi.org/10.3390/rs16010067

Academic Editor: Pedro Melo-Pinto

Received: 7 November 2023
Revised: 14 December 2023
Accepted: 19 December 2023
Published: 23 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Classification is one of the most important tasks [1,2] in HSI processing, which provides the basis for many subsequent applications, such as urban planning, military target recognition, and geological prospecting. In addition, HSI classification is also a prerequisite for many subsequent processing tasks of HSIs, such as semantic segmentation [3,4], content understanding [5,6], target recognition [7,8], and anomaly detection [9,10].

One of the key points of classification is feature extraction. For decades, many conventional feature extraction methods have been proposed for HSI classification. For instance, the Extended Multiple Attribute Profile (EMAP), as a popular method for spectral–spatial feature extraction, is widely used in HSI processing. This method achieves the purpose of selecting the best features by connecting multiple morphological attribute filters without considering the pixel problem [11–13]. Later, Kwan et al. [14,15] used EMAP to enhance image bands. Zhang et al. [16] proposed a new classification framework based on the gravitational optimized multilayer perception classifier and EMAP, combined with Sentinel-2 Multispectral images (MSI), to draw complex coastal wetland maps. Huang et al. [17]

proposed to use EMAP to explore spatial features and remove suspicious abnormal pixels to achieve the effect of image background purification.

In addition to EMAP, there is a series of other notable techniques, such as Support Vector Machine [18] and Discriminant Analysis [19,20]. For example, Baassou et al. [21] proposed a novel method using Support Vector Machine (SVM) with Spatial Pixel Association (SPA) features to enhance SVM's classification performance by extracting regional texture information from hyperspectral data. Guo et al. [22] addressed the particular demands of HSI classification for SVM by introducing a spectral-weighted kernel. They selected a specific set of weights by optimizing the estimation of generalization error or appraising the practicality of each band. Melgani et al. [23] comprehensively evaluated the potential of SVM in HSI classification through a combination of theoretical exploration and experimental analysis, providing a comprehensive perspective for a deeper understanding of its performance. Kang et al. [24] suggested a novel PCA-based Edge-Preserving Features (PCA–EPFs) method, processing HSIs by constructing standard Edge-Preserving Features (EPFs), followed by dimension reduction utilizing Principal Component Analysis (PCA), and ultimately employing SVM for classification. Wang et al. [25] introduced a supervised approach utilizing PCA Network (PCANet) and Gaussian-Weighted Support Vector Machine (Gaussian-SVM), obtaining HSI classification results through threshold decision. Villa et al. [26] presented a method using Independent Component (IC) Discriminant Analysis (ICDA). They utilized ICDA to choose the transformation matrix that maximizes the independence of components and applied the Bayesian rule for the final classification. Bandos et al. [27] introduced an efficient version of Regularized Linear Discriminant Analysis (RLDA) for HSI classification, addressing the challenges when the ratio between the number of spectral features and the number of training samples is large. These traditional methods perform well in small-sample classification problems. However, as training datasets become complex and scale increases, they may encounter performance bottlenecks due to limitations such as linear assumptions, computational complexity, dimension constraints, and specific assumptions about data distribution. Recently, there has been widespread adoption of deep learning methods in HSI classification, addressing limitations posed by traditional approaches.

The rapid advancement of deep learning technology has significantly influenced various domains, notably making substantial contributions to the field of image processing [28,29]. In the domain of remote sensing data classification, the application of deep learning methods has garnered considerable attention for its efficacy in analyzing fragmented data with improved efficiency and precision. Multiple approaches have been proposed for the classification of HSIs by leveraging deep models [30–32]. Hu et al. [33] devised a method that employs a 1D CNN comprising five convolutional layers. This method utilizes spectral dimension information as input, accurately extracting spectral features. However, this network inadequately considers the importance of spatial information. To overcome this limitation, Zhao and Du [34] designed a 2D CNN model that, through dimensionality reduction of spectral information, extracts valuable spatial features from the data. Nevertheless, both of these methods analyze data from a singular feature dimension. Yang et al. [35] employed a dual-branch structure, utilizing both one-dimensional and two-dimensional CNNs to extract spectral and spatial features simultaneously. Subsequently, Chen et al. [36] introduced three-dimensional CNNs from the natural image domain to address HSI classification problems. To extract deep spectral–spatial features, Roy et al. [37] used a concatenation of three-dimensional and two-dimensional CNNs to extract features from HSI.This method not only comprehensively extracts spatial–spectral feature information but also improves classification accuracy while reducing computational complexity.

Given the widespread acceptance of residual networks, He et al. [38] introduced residual networks (ResNets) for HSI classification. This approach ensures more comprehensive feature information extraction, allowing the model to minimize information loss at each convolutional layer to address the challenge of gradient vanishing. Zhong et al. [39] intro-

duced a spatial–spectral residual network (SSRN) that supplements information on the previous layer's features with the next layer's features to achieve enhanced classification performance. In [40], adding a residual network to the model to increase the network's depth and feature map dimensions has successfully extracted feature information that traditional convolutional filters may overlook. Dense convolutional network structures, such as Cubic-CNN [41] and lightweight heterogeneous kernel convolutions [42], are also capable of achieving effective feature extraction for HSI and yielding satisfactory classification results.

All the aforementioned methods have employed strategies based on CNN backbones and their variants, effectively enhancing the classification performance of hyperspectral images (HSI). However, challenges remain in the face of decreasing classification performance due to limited training samples and increasing network depth. Additionally, these methods face the significant challenge of feature redundancy.

In recent years, Vision Transformer (ViT) [43] has found widespread application in various computer vision tasks [44–46], serving as an extension of the Transformer [47] architecture into the visual domain. While traditional CNNs excel in various visual tasks such as classification, detection, and segmentation, HSIs typically comprise hundreds of contiguous spectral bands, posing challenges for CNNs in effectively capturing the global dependencies within spectral information. In contrast, ViT leverages the same self-attention mechanism as the original Transformer, enabling it to establish relationships between different positions within an image, effectively capturing global information. This capability has propelled ViT to excel in HSI classification tasks, even surpassing traditional CNNs in certain scenarios [48].

Hong et al. [49] reexamined the HSI classification problem from a sequential perspective and introduced a novel Transformer-based backbone network known as SpectralFormer. SpectralFormer incorporates two simple yet effective modules, grouped spectral embedding (GSE) and cross-layer adaptive fusion (CAF). These modules are designed to facilitate the learning of local detailed spectral representations and to transmit memory-like components from shallow layers to deeper layers. Sun et al. [50] presented a novel model called SSFTT, designed to convert shallow-level features into deep semantic tokens. This model effectively captures spectral–spatial joint features through a combination of convolutional layers and Transformers. In the study by Xue et al. [51], they introduced a local Transformer for HSI classification. Within this Transformer, there is a Spatial Partitioned Recurrent Local Transformer Network (SPRLT-Net). SPRLT-Net not only acquires global contextual information but also its dynamic attention weights can adaptively accommodate spatial variations among different pixels in HSI. Huang et al. [52] presented the 3D SwinT (3DSwinT) model, tailored to accommodate the 3D characteristics of HSI and capture the abundant spatial–spectral information within HSI. Additionally, they introduced a novel hierarchical contrastive learning method based on 3DSwinT (3DSwinT-HCL), which effectively harnesses multi-scale semantic representations of images. Fang et al. [53] introduced an approach called MAR-LWFormer, which utilizes a multi-attention joint mechanism and a lightweight Transformer to achieve multi-channel feature representation. The design of MAR-LWFormer aims to effectively leverage the multispectral and multi-scale spectral–spatial information within HSI data, significantly enhancing classification accuracy, particularly under extremely low sampling rates. Xu et al. [54] introduced a method called spatial–spectral 1DSwin (SS1DSwin) Transformer, which comprises two critical components, the grouped Feature Tokenization module (GFTM) and the 1DSwin Transformer with a cross-block normalization connection module (TCNCM). The design of the SS1DSwin Transformer investigates local and hierarchical spatial–spectral relationships from two distinct perspectives. Zhang et al. [55] proposed a novel and efficient lightweight spectral–spatial Transformer (ELS2T). This method employs a global multi-scale attention module (GMAM) to emphasize feature distinctiveness and proposes an adaptive feature fusion module (AFFM) for the effective integration of spectral and spatial features.

Currently, HSI classification is one of the hot research areas in HSI processing [56,57]. Researchers have made significant progress in unsupervised learning [58], autoencoders [59], latent representation learning [60], adversarial representation learning [61], and other fields. They have opened up new directions for handling HSI classification tasks. However, unsupervised feature learning and adversarial learning have not achieved satisfactory results in extracting spectral–spatial features from HSI. In our proposed ETLKA model, we design a novel architecture that consists of dual-branch shallow feature extraction, an innovative attention mechanism, and an efficient Transformer framework.

This paper introduces an innovative network that enhances the understanding of image data within the Transformer model and improves its robustness. We have enhanced the module for extracting Spectral–Spatial Shallow Features and integrated it with a Transformer architecture featuring a Large-Kernel Attention module to thoroughly extract spatial and spectral information from HSIs. During the phase of extracting Spectral–Spatial Shallow Features, We adopted a dual-branch structure, with the first branch extracting spectral–spatial features from HSI and the second branch extracting spatial features. To further enhance the quality of extracted features, strengthen the Transformer's comprehension of image data, alleviate the computational complexity of attention mechanisms, and effectively mitigate the impact of redundant information, we designed a Large-Kernel Attention Module preceding the Transformer encoder. This enhancement aims to reduce redundancy and elevate the overall performance of the model.

The main contributions of the ETLKA method can be condensed into the following three points:

- In order to more comprehensively extract spatial and spectral feature information from HSIs, a high-performance network has been designed that combines a dual-branch CNN with a Transformer framework equipped with a Large-Kernel Attention mechanism. This further enhances the classification performance of the CNN–Transformer combined network;
- In the shallow feature extraction module, we designed a dual-branch network that uses 3D convolutional layers to extract spectral features and 2D convolutional layers to extract spatial features. These two discriminative features are then processed by a Gaussian-weighted Tokenizer module to effectively fuse them and generate higher-level semantic tokens;
- From utilizing a CNN network for shallow feature extraction to effectively capturing global contextual information within the image using the Transformer framework, our proposed ETLKA allows for comprehensive learning of spatial–spectral features within HSI, significantly enhancing joint classification accuracy. Experimental validation on three classic public datasets has demonstrated the effectiveness of the proposed network framework.

2. Materials and Methods

Figure 1 illustrates the overall framework of the ETLKA model proposed for HSI classification. It is primarily divided into four main components: Feature Extraction via Dual-Branch CNNs, HSI Feature Tokenization, Large-Kernel Attention (LKA), and Transformer Encoder (TE) modules.

2.1. Feature Extraction via Dual-Branch CNNs

We represent the obtained HSI data using a 3D tensor $\boldsymbol{P} \in \mathbb{R}^{c \times v \times e}$, where $c \times v$ represents the spatial dimensions of the HSI data and e represents the number of spectral dimensions in the HSI. Each pixel in the HSI contains e spectral dimensions, and we typically use one-hot encoded class vectors to represent the labels for this feature, denoted as $L = (l_1, l_2, \ldots, l_D) \in \mathbb{R}^{1 \times 1 \times D}$, where D is the number of land cover categories present in the current region. As HSI data often have a high number of spectral dimensions, we preprocess the HSI using PCA to significantly reduce the computational burden of the model. PCA reduces the number of spectral bands from e to b, while keeping the

spatial dimensions as $c \times v$. The HSI data after dimension reduction are represented as $\boldsymbol{P}_{\text{pca}} \in \mathbb{R}^{c \times v \times b}$, where b represents the number of spectral dimensions obtained through the PCA operation.

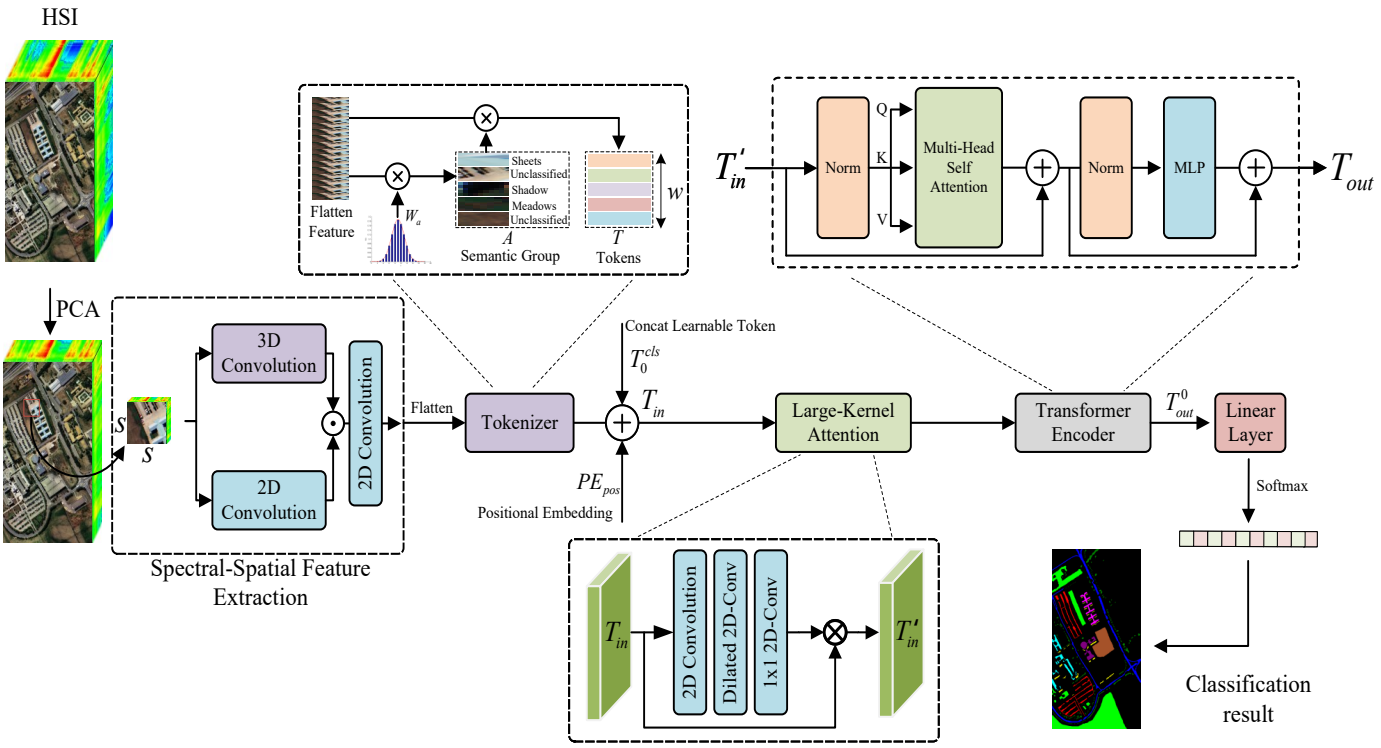


Figure 1. The overall framework of the proposed ETLKA network for the HSI classification.

After obtaining the 3D blocks $\boldsymbol{B} \in \mathbb{R}^{s \times s \times b}$ from the preprocessed HSI data $\boldsymbol{P}_{\text{pca}}$, these extracted 3D blocks serve as inputs to the entire model. Here, $s \times s$ represents the spatial dimensions when extracting 3D blocks, and b represents the spectral dimensions of the 3D block. The center coordinates in the spatial dimensions of each block obtained from the HSI are denoted as (y_m, y_n), where m and n range from $0 \leq m < c$ and $0 \leq n < v$. The true labels for each 3D block are determined by the class of the element located at the central coordinates.

When extracting 3D blocks from the edge pixels, due to the lack of some edge pixels, padding is applied to the original HSI data with a width of $(s-1)/2$. The number of generated 3D blocks equals the number of spatial pixels contained in the HSI ($c \times v$). After removing all 3D blocks with the background class (label 0), the remaining 3D blocks are divided into training and testing sets for model training and evaluation.

Following data preprocessing, we proceed to extract shallow spectral–spatial features from each acquired 3D sample block using a dual-branch convolutional layer. In this module, the 3D convolutional layer theoretically consists of 8 3D convolution kernels. The training samples pass through the first branch of this module's 3D convolutional layer, generating 8 3D feature cubes, which contain rich spectral–spatial features obtained from the HSI.

At the same time, we convert the 3D HSI cube data into 2D data in order to feed it into a 2D convolutional layer. In the 2D convolution layer, the size of the convolutional kernel is set to 64@3 $\times$ 3. Set the size of the padding to (1×1). Next, we convert the 3D feature information obtained from the 3D convolution branch into 2D data, in order to fuse it with the 2D feature information obtained from the 2D convolution branch through concatenation. Finally, we use a 2D convolutional layer with 64@3 $\times$ 3 2D convolutional

kernels to further extract spatial features from the obtained 2D feature information. The entire module is as follows:

$$\boldsymbol{X}_{out} = (Conv3D(\boldsymbol{X}_{in}, k_1, p_1) \odot Conv2D(\boldsymbol{X}_{in}, k_2, p_2)) \oplus Conv2D(\boldsymbol{X}'_{in}, k_3, p_3) \quad (1)$$

where $\boldsymbol{X}_{in}$ represents the HSI cube data, $\boldsymbol{X}'_{in}$ represents the new feature information obtained by concatenating two-dimensional feature information from two branches, k represents the size of the convolution kernel, k_1 is $(3 \times 3 \times 3)$, k_2 and k_3 are (3×3), p represents the fill size, p_1 is $(0 \times 1 \times 1)$, and p_2 and p_3 are (1×1).

2.2. HSI Feature Tokenization

The training samples extract rich shallow spectral–spatial features using the dual-branch CNNs that we designed. However, there is still deeper feature information to be explored. To address this issue, we redefine the obtained shallow spectral–spatial features as semantic tokens. We represent the flattened feature map from the input (obtained by flattening the 2D feature map) as $\boldsymbol{F} \in \mathbb{R}^{uv \times z}$, where uv represents the height and width of the 2D feature map, and z represents the number of channels. The final output obtained by this module is represented as $\boldsymbol{T} \in \mathbb{R}^{t \times z}$, where t is the token number. To obtain feature tokens $\boldsymbol{T}$ from the feature map $\boldsymbol{F}$, one can use the following equation:

$$\boldsymbol{T} = \underbrace{softmax(\boldsymbol{F}\boldsymbol{W}_a)}_{Q}{}^{T} \boldsymbol{F} \quad (2)$$

In the formula, $\boldsymbol{F}\boldsymbol{W}_a$ represents element-wise multiplication with dimensions 1×1, and $\boldsymbol{W}_a \in \mathbb{R}^{z \times t}$ represents a weight matrix. We initialize this weight matrix using a Gaussian distribution. At this stage, the newly generated semantic groups are represented by $\boldsymbol{Q} \in \mathbb{R}^{uv \times t}$. Next, $\boldsymbol{Q}$ is transposed. We apply the softmax function (softmax$(\cdot)$) to the transposed result to emphasize important semantic components. Finally, the multiplication of $\boldsymbol{Q}$ and $\boldsymbol{F}$ generates the module's final output semantic tokens $\boldsymbol{T}$. The specific process is visualized in Figure 2.

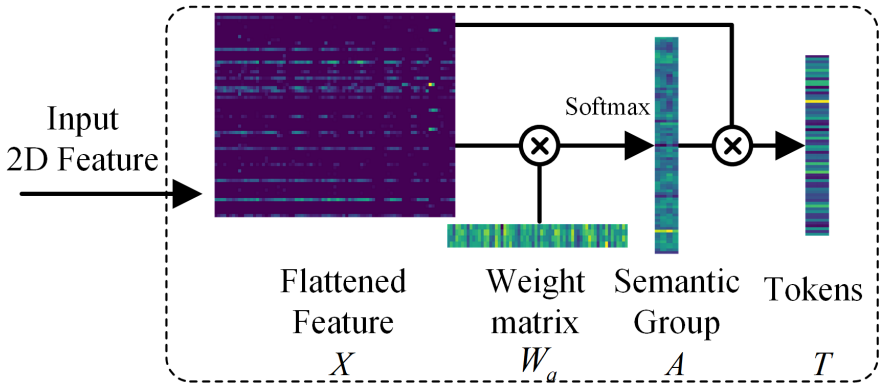


Figure 2. The visualization process of the HSI Feature Tokenization [50].

2.3. Large-Kernel Attention

The tokens output from the previous module can be represented as $[T_1, T_2, \ldots, T_k]$. To make our model more suitable for our classification task, we include a learnable classification token T_0^{cls} in the first position of these tokens. In order to not lose the positional information inherent in the image, we embed positional information PE_{pos} into each semantic token. For $\left[T_0^{cls}, T_1, \ldots, T_k\right]$ and PE_{pos}, the input T_{in} for the Large-Kernel Attention can be represented by the following equation:

$$\boldsymbol{T}_{in} = \left[T_0^{cls}, T_1, \ldots, T_k\right] + PE_{pos} \quad (3)$$

In the realm of academic research focused on HSI processing, Large-Kernel Attention offers several valuable contributions to subsequent Transformer modules. First and foremost, it elevates the quality of feature extraction, enhancing the Transformer's comprehension of image data and thereby advancing overall model performance. Significantly, Large-Kernel Attention serves to alleviate the computational complexity of attention mechanisms, particularly evident when dealing with extensive image datasets, as it effectively reduces the number of positional elements that necessitate consideration, ultimately enhancing computational efficiency. Furthermore, it effectively mitigates the impact of redundant information, thereby fortifying the robustness of the Transformer model. This is especially pertinent within the context of image data processing, given the inherent wealth of information typically contained therein. Lastly, Large-Kernel Attention plays a pivotal role in augmenting computational efficiency, facilitating accelerated model training convergence, and curtailing memory and computational requisites. These advantages collectively position Large-Kernel Attention as an invaluable component within the Transformer architecture for scholarly endeavors in HSI processing.

As depicted in Figure 1, this module primarily consists of a 2D convolution layer with 4@3 × 3 kernels, a dilated 2D convolution layer, and a 2D convolution layer with 4@1 × 1 kernels. The process can be expressed by the following equations:

$$\boldsymbol{T}'_{in} = (Conv2D(DilatedConv2D((Conv2D(\boldsymbol{T}_{in}, k_1, p_1), k_2, p_2, d_1)), k_3))\boldsymbol{T}_{in} \tag{4}$$

where k represents the size of the convolution kernel, k_1 and k_2 are (3×3), k_3 is (1×1), p represents the fill size, p_1 is (1×1), p_2 is (3×3), d represents the size of dilation, and d_1 is 3.

2.4. Transformer Encoder Module

After the Large-Kernel Attention, we obtain enhanced quality features, which are then fed into the TE module. This module uses self-attention mechanisms to handle relationships between tokens, capturing both global and local features. As can be observed from Figure 1, this module primarily consists of four components, including two Layer Normalization (LN) layers, a Multilayer Perceptron (MLP) layer, and a Multi-Head Self-Attention (MHSA) block. To facilitate deep neural network training and optimization, alleviate gradient vanishing problems, and enhance performance, residual connections are added after the MHSA block and MLP layer.

The TE module includes two normalization layers placed before MHSA and MLP, which help alleviate gradient explosion, reduce vanishing gradient problems, and accelerate training. The core of TE is the MHSA block. MHSA integrates the Self-Attention (SA) mechanism, with its essential components typically named $\boldsymbol{Q}$(Queries), $\boldsymbol{K}$(Keys), and $\boldsymbol{V}$(Values). These three matrices are learned during the model's training process to adapt to the classification task and HSI data. The SA mechanism computes attention scores using $\boldsymbol{Q}$ and $\boldsymbol{K}$, and the weights of these scores are determined using the softmax function. Subsequently, the scores are multiplied by $\boldsymbol{V}$ to obtain the output of SA, as shown in Figure 3b. These descriptions can be expressed by the following equations:

$$SA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{5}$$

where $\boldsymbol{K}^T$ is the transpose of K and d_K is the dimension of $\boldsymbol{K}$.

Compared to SA, MHSA uses multiple sets of weight matrices, allowing it to map multiple sets of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. Following the same operations as described earlier, it computes attention values for each set to calculate Multi-Head Attention values. The attention results from each head are then concatenated together. The final step involves concatenating each attention value and multiplying it with the weight matrix $\boldsymbol{W} \in \mathbb{R}^{n \times w_K \times w_d}$, where n represents the number of attention heads and w_d represents the number of tokens. The computation equation for MHSA can be expressed as follows:

$$MHSA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(SA_1, SA_2, \ldots, SA_h)\boldsymbol{W} \quad (6)$$

The MLP component consists of two fully connected layers. After passing through the TE module, T'_{in} is transformed into T_{out}. We extract the classification token vector $\hat{T}_0^{cls}$ embedded in Equation (3) for the classification task. Next, we use a linear layer with an output dimension equal to the number of land cover classes in the HSI data. Finally, we use the softmax function to assign the label of the input sample to the category represented by the class with the highest probability in the final vector.

The complete procedure of the ETLKA method, as proposed, is outlined in Algorithm 1.

Algorithm 1 Enhanced Transformer with Large-Kernel Attention Model

Input: Input HSI data $\boldsymbol{P} \in \mathbb{R}^{c \times v \times e}$ and ground truth labels $\boldsymbol{Y} \in \mathbb{R}^{c \times v}$; Set the spectral dimension after PCA preprocessing to $b = 30$; Extract patch size $s = 13$; and specify the training sample rate as μ.

Output: Predicted labels for the test dataset.

1: Configure batch size = 64, use the Adam optimizer with a learning rate of 5×10^{-4}, and set the number of iterations to $\epsilon = 150$.
2: Obtain PCA-transformed HSI, denoted as $\boldsymbol{P_{pca}}$, from which generate patches for all samples and split them into training and test sets according to the training sample rate.
3: Create training and test data loaders.
4: **for** $i = 1$ to ϵ **do**
5: Generate shallow features using the spectral–spatial shallow feature extraction module.
6: Flatten the extracted 2D shallow spectral–spatial feature maps to obtain a 1D feature vector.
7: Execute tokenization transformation using feature vectors and initialized weights to produce semantic tokens.
8: The first position of the semantic token sequence combines with a learnable class token, and positional embeddings are applied to these semantic tokens.
9: Perform Large-Kernel Attention and TE module
10: Input the learnable class tokens into a linear layer and use the softmax function to obtain classification probabilities.
11: **end for**
12: Use the trained model with the test dataset to obtain predicted labels.

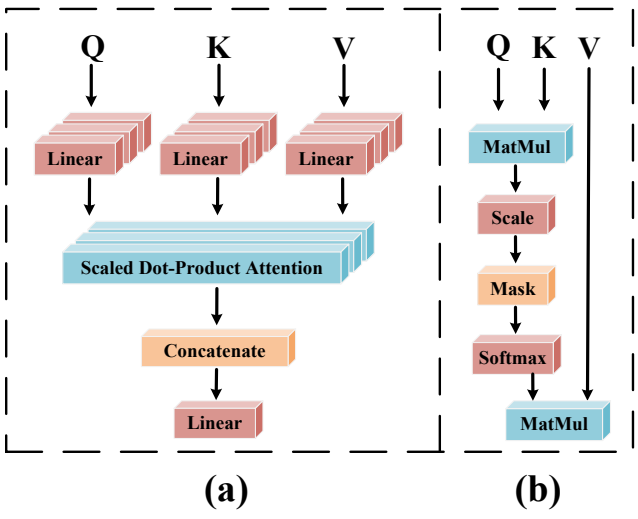


Figure 3. Visual representations of (**a**) Multi-Head Self-Attention Module and (**b**) self-attention in the Transformer architecture.

3. Results

3.1. Data Description

The proposed method was tested on three classical HSI datasets, including the Indian Pines dataset, the Pavia University dataset, and the Houston2013 dataset.

Indian Pines dataset: This dataset was acquired in northwestern Indiana, USA, in 1992, using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. The HSI consists of 224 bands with a wavelength range from 0.4 to 2.5 micrometers. The image comprises 145 × 145 pixels with a spatial resolution of 20 m and includes 16 land cover categories. In our experiments, a total of 200 bands were chosen (excluding water-absorbing and noisy bands). Pseudo-color images and ground-truth maps are illustrated in Figure 4a and b, respectively.

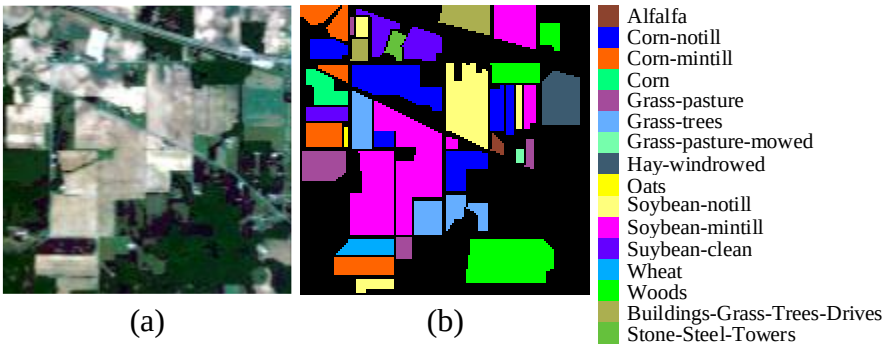


Figure 4. Visualization of the India Pines (IP) dataset. (**a**) Pseudo-color image for the dataset. (**b**) Ground-truth map for the dataset.

Pavia University dataset: This dataset was captured in 2001 over the University of Pavia in the northern part of Italy using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. The dataset is an HSI comprising 115 bands with a wavelength ranging from 0.43 to 0.86 micrometers. The image has dimensions of 610 × 340 pixels with a spatial resolution of 1.3 m. It includes 9 land cover categories. In the experiments, 103 bands were selected for analysis, and 12 noisy bands were removed. Pseudo-color images and ground-truth classification maps are displayed in Figure 5a,b.

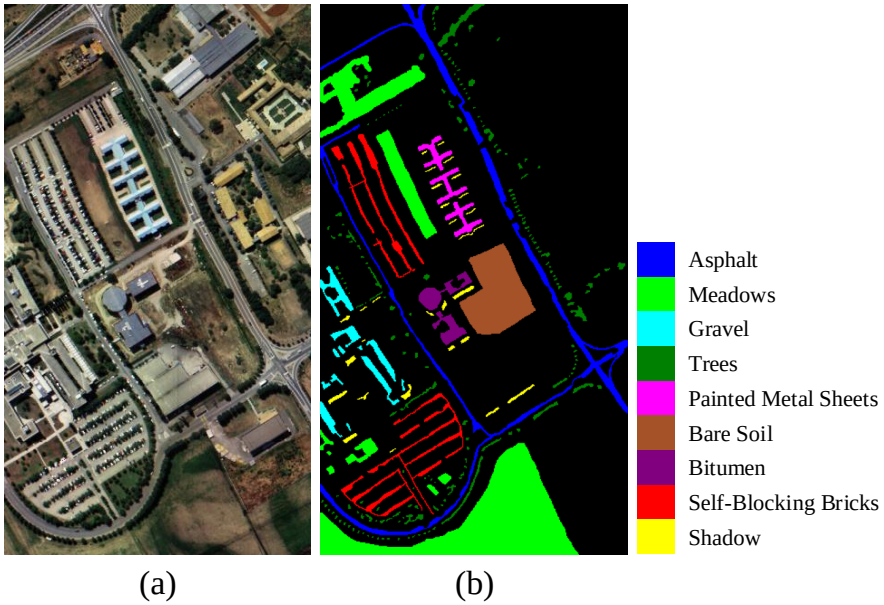


Figure 5. Visualization of the Pavia University (PU) dataset. (**a**) Pseudo-color image for the dataset. (**b**) Ground-truth map for the dataset.

Houston2013 dataset: This dataset was provided by two organizations, one being the Hyperspectral Imaging Group at the University of Houston, and the other being the National Center for Airborne Laser Mapping (NCALM), funded by the National Science Foundation (NSF). The dataset includes 15 land cover categories. It is a hyperspectral image consisting of 144 bands with a wavelength range from 0.38 to 1.05 micrometers. The dataset comprises 349 × 1905 pixels with a spatial resolution of 2.5 m. Pseudo-color images and ground-truth classification maps are shown in Figure 6a,b.

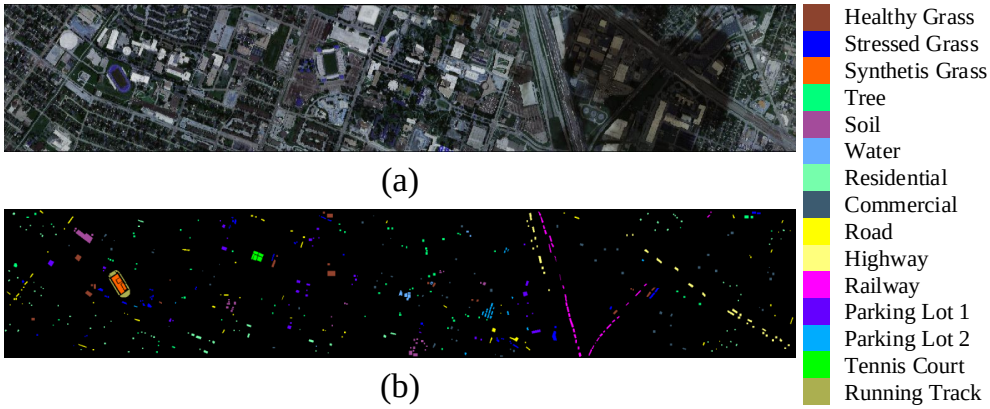


Figure 6. Visualization of the Houston2013 dataset. (**a**) Pseudo-color image for the dataset. (**b**) Ground-truth map for the dataset.

The land cover category names, number of training samples, and number of test samples regarding these three datasets are listed in Table 1. Each dataset takes 3% of the samples as the training set and the rest as the test set.

Table 1. Training and test sample numbers in the India Pines dataset, Pavia University dataset, and Houston2013 dataset.

NO.	India Pines			Pavia University			Houston2013		
	Class	**Training**	**Test**	**Class**	**Training**	**Test**	**Class**	**Training**	**Test**
#1	Alfalfa	1	45	Asphalt	199	6432	Healthy Grass	125	1126
#2	Corn-notill	43	1385	Meadows	559	18,090	Stressed Grass	125	1129
#3	Corn-mintill	25	805	Gravel	63	2036	Synthetis Grass	70	627
#4	Corn	7	230	Trees	92	2972	Tree	124	1120
#5	Grass-pasture	14	469	Metal Sheets	40	1305	Soil	124	1118
#6	Grass-tree	22	708	Bare soil	151	4878	Water	33	292
#7	Grass-pasture-mowed	1	27	Bitumen	40	1290	Residential	127	1141
#8	Hay-windrowed	14	464	Bricks	111	3571	Commercial	124	1120
#9	Oats	1	19	Shadows	28	919	Road	125	1127
#10	Soybean-notill	29	943				Highway	123	1104
#11	Soybean-mintill	73	2382				Railway	123	1112
#12	Soybean-clean	18	575				Parking Lot 1	123	1110
#13	wheat	6	199				Parking Lot 2	47	422
#14	woods	38	1227				Tennis Court	43	385
#15	Buildings-Grass-Trees	12	374				Running Track	66	594
#16	Stone-Steel-Towers	3	90						
	Total	307	9942	Total	1283	14,493	Total	1502	13,527

3.2. Parameter Analysis

We analyzed several crucial hyperparameters that have a significant impact on both the classification performance and the training process. These parameters encompass the patch size and the batch size.

(1) Patch Size: To empirically study the effect of patch size, we kept a fixed set of values for the remaining hyperparameters. The patch size was systematically chosen from a predetermined set of candidate values, namely, {9, 11, 13, 15, 17}. Figure 7 shows the effect of different patch sizes on the OA. An optimal OA was achieved in three datasets when the patch size was set to 13.

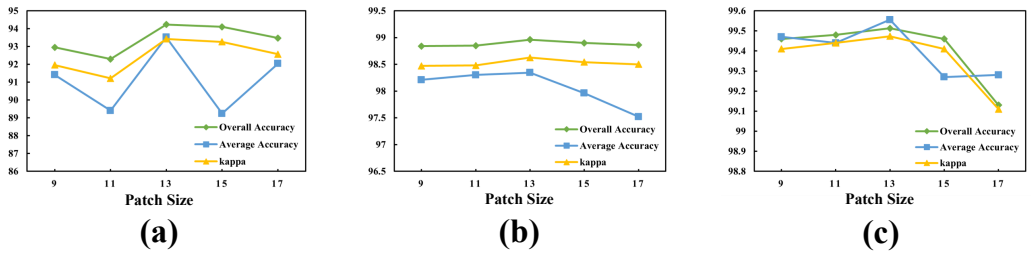


Figure 7. The impact of different patch sizes on Overall Accuracy, Average Accuracy, and kappa coefficient. (**a**) India Pines dataset, (**b**) Pavia University dataset, (**c**) Houston2013 dataset.

(2) Batch Size: In Figure 8, the impact of batch size on OA of the three datasets is presented. The batch size was chosen from a set of candidates {16, 32, 64, 128, 256}. Clearly, the classification metrics we have presented achieve the best results when the batch size is set to 64.

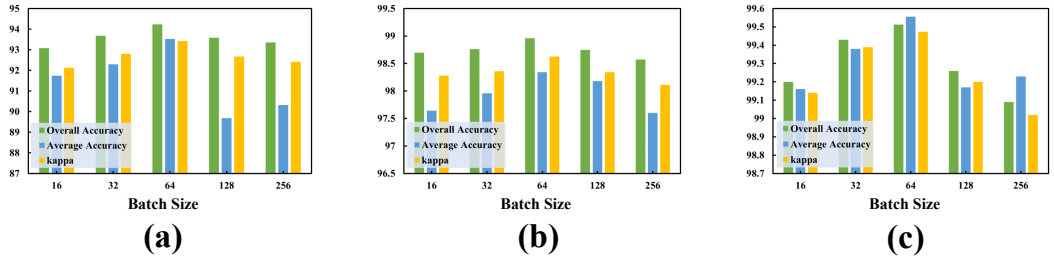


Figure 8. The impact of different batch sizes on Overall Accuracy, Average Accuracy, and kappa coefficient. (**a**) India Pines dataset, (**b**) Pavia University dataset, (**c**) Houston2013 dataset.

3.3. Classification Results and Analysis

At this stage, we conducted comparative experiments with several advanced traditional and deep learning methods to validate the effectiveness and superiority of the proposed ETLKA model. They are SVM [21], EMAP [17], 1D-CNN [33], 2D-CNN [34], 3D-CNN [36], HybridSN [37], and SSFTT [50]. For the listed comparative methods, the network parameters and training strategies in the original paper remain unchanged in the experiments. The number of training and testing samples for the above methods is as shown in Table 1. To ensure a fair comparison, samples are randomly selected.

(1) Quantitative Results and Analysis: In Tables 2–4, we present the superiority of the proposed ETLKA method over each of the comparison methods on the Indian Pines, Pavia University, and Houston2013 datasets in four aspects: Overall Accuracy (OA), Average Accuracy (AA), kappa coefficient (κ), and class-wise accuracy. OA, AA and κ can be obtained from the following equations:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$AA = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i} \tag{8}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

where TP is truly positive, TN is truly negative, FP is false positive, FN is false negative, C is the number of classes, and TP_i and FP_i are the true positives and false positives for class i, respectively. Additionally, p_o is observed accuracy and p_e is random accuracy. In classification problems, p_o can be calculated through OA, while p_e can be calculated through the marginal probabilities of the classes. The best results for each metric in the tables are highlighted in bold. It is evident that our proposed ETLKA model outperforms the seven compared methods. Taking the India Pines dataset as an example, our proposed approach demonstrated superior classification performance for categories 'Alfalfa', 'Corn-min-till', 'Corn', 'Grass-tree', 'Grass-pasture-mowed', 'Oats', 'Soybean-clean', and 'woods'. This owes to the innovative feature extraction method, efficient attention strategy, and the incorporation of the Transformer architecture in our proposed framework. In Figure 9, we visualized the performance comparison of different methods on different datasets. It is evident from the visualization that the proposed ETLKA model exhibits the best performance.

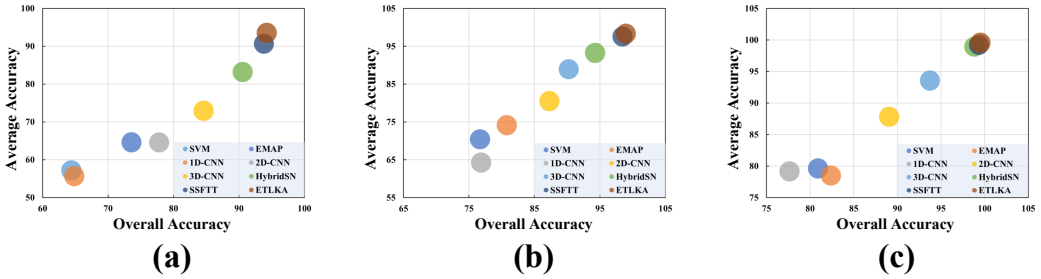


Figure 9. The overall visualization presents the performance of different methods on various datasets. (**a**) India Pines dataset, (**b**) Pavia University dataset, (**c**) Houston2013 dataset.

(2) Visual Evaluation and Analysis: Our proposed method, as illustrated in Figures 10–12, generated classification maps for three datasets. By comparing these maps in terms of spatial characteristics, edge contours, and the presence of noise, it is clear that our method's visual maps closely resemble the actual ground-truth maps. In contrast, traditional classifiers generally performed less effectively in comparison to deep learning classifiers, with their classification maps displaying a higher degree of misclassification and noticeable noise.

For 1D-CNN, 2D-CNN, 3D-CNN, and HybridSN methods, while misclassification was reduced, noise outliers were still present. Our method notably mitigated noise-related issues in the classification maps and accurately represented the shapes of the classified regions. In the case of Transformer-based methods, such as SSFTT, the consideration of global information interaction further improved classification accuracy and led to clearer delineation of classification map boundaries. However, our method's classification maps showed the fewest instances of misclassification and noise, affirming its superior visual classification performance at the same sampling rate. Specifically, in the Indian Pines dataset, various methods performed relatively poorly in classifying categories 'Alfalfa,' 'Corn,' and 'Oats.' Our method achieved highly accurate results. In the Pavia University dataset, two traditional methods exhibited pronounced misclassification in the 'Bitumen' category region, while other methods displayed more noise in the 'Bare soil' category. Our method had the fewest classification errors in the 'Bare soil' category region. For the Houston2013 dataset, the classification results of Transformer-based methods significantly outperformed traditional methods and other deep learning methods.

Table 2. Different methods exhibit various classification metrics on the Indian Pines dataset. The optimal results are bolded.

NO.	SVM [21]	EMAP [17]	1D-CNN [33]	2D-CNN [34]	3D-CNN [36]	HybridSN [37]	SSFTT [50]	Proposed
1	22.22	17.07	11.68	17.46	46.13	34.09	48.00	**78.67**
2	62.31	70.55	81.87	68.46	75.84	91.96	**92.74**	91.89
3	49.68	70.01	19.41	66.54	63.50	79.59	86.55	**93.90**
4	17.82	66.19	20.39	40.18	44.37	64.00	88.96	**98.35**
5	44.77	74.25	77.13	66.67	**96.51**	94.12	88.32	84.14
6	86.58	61.34	94.04	97.54	98.86	90.62	98.80	**99.00**
7	88.89	60.00	46.67	67.04	70.37	85.19	94.81	**97.04**
8	95.04	82.32	97.29	98.37	**99.45**	97.80	98.15	97.16
9	10.53	20.45	11.76	15.26	26.32	68.42	88.42	**97.37**
10	47.08	71.89	32.92	84.06	85.36	84.72	**95.26**	88.12
11	71.99	80.99	83.76	85.71	93.48	**96.78**	95.54	96.56
12	30.98	63.11	48.61	48.67	64.65	78.15	87.43	**90.19**
13	94.97	93.46	93.10	98.97	98.97	95.38	**99.85**	96.68
14	92.75	87.09	91.63	89.09	96.92	**98.92**	97.47	98.22
15	15.77	64.84	42.07	55.04	70.57	85.83	93.50	**94.97**
16	93.33	69.05	72.15	71.26	74.71	85.23	**96.44**	94.11
OA (%)	64.39	73.58	64.83	77.81	84.59	90.57	93.81	**94.23**
AA (%)	57.17	64.61	55.66	64.57	72.95	83.18	90.64	**93.52**
$\kappa \times 100$	59.07	69.87	62.74	74.56	82.26	89.17	92.94	**93.43**

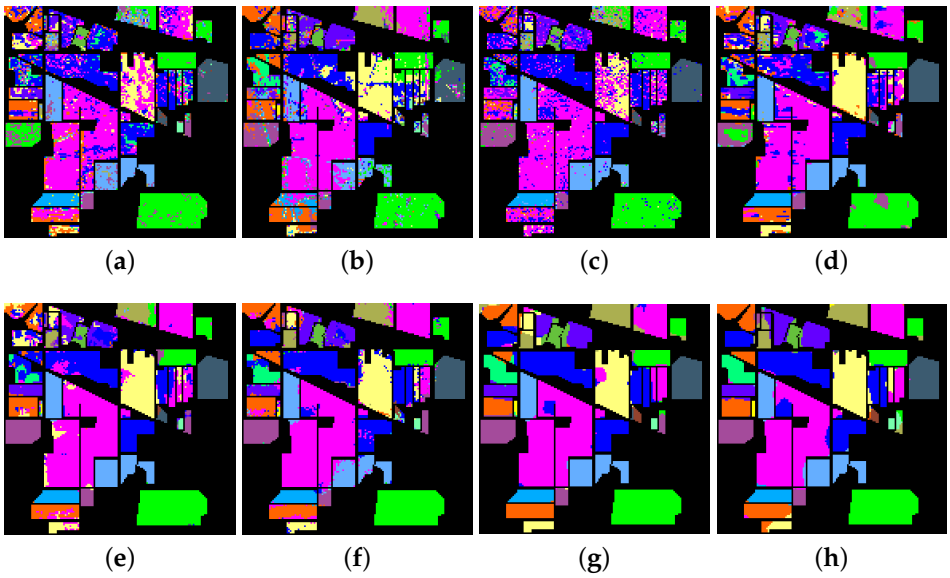


Figure 10. Classification maps of Indian Pines dataset. (**a**) SVM, (**b**) EMAP, (**c**) 1D-CNN, (**d**) 2D-CNN, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) SSFTT, (**h**) Proposed.

In summary, visual comparisons across multiple classifiers confirm the outstanding classification performance of our proposed method.

3.4. Inference Speed Analysis

As shown in Table 5, with the epoch set to 150 and the batch size set to 64, we made corresponding records of the network's training time and testing time. The experimental results indicated that the network exhibited fast inference speed.

Table 3. Different methods exhibit various classification metrics on the Pavia University dataset. The optimal results are bolded.

NO.	SVM [21]	EMAP [17]	1D-CNN [33]	2D-CNN [34]	3D-CNN [36]	HybridSN [37]	SSFTT [50]	Proposed
1	83.05	87.50	79.27	88.64	92.95	87.84	97.39	**99.26**
2	87.36	93.35	89.33	97.75	97.85	99.87	**99.89**	99.78
3	42.59	55.97	47.99	42.97	82.45	85.08	94.47	**96.67**
4	81.47	85.36	79.10	92.98	95.37	91.38	94.86	**97.84**
5	95.17	97.47	98.92	99.32	95.63	91.50	98.67	**99.74**
6	40.15	48.54	76.42	76.14	79.59	93.93	99.77	**99.88**
7	14.92	28.39	20.70	63.78	76.44	89.38	99.38	**99.53**
8	71.23	85.13	74.19	72.13	88.08	90.33	**96.25**	95.21
9	87.54	96.84	71.47	98.93	**99.78**	96.12	97.48	97.17
OA (%)	76.68	80.76	76.85	87.28	90.23	94.26	98.44	**98.96**
AA (%)	70.43	74.15	64.27	80.52	88.95	93.27	97.57	**98.34**
$\kappa \times 100$	74.87	79.11	69.15	82.84	89.72	93.73	97.94	**98.62**

Table 4. Different methods exhibit various classification metrics on the Houston2013 dataset. The optimal results are bolded.

NO.	SVM [21]	EMAP [17]	1D-CNN [33]	2D-CNN [34]	3D-CNN [36]	HybridSN [37]	SSFTT [50]	Proposed
1	95.65	87.85	86.20	94.02	93.75	98.85	**99.66**	99.42
2	97.52	92.53	95.13	96.30	94.91	99.73	99.14	**99.81**
3	99.84	99.84	**100.00**	89.73	95.54	99.84	99.68	99.79
4	93.66	92.63	95.43	98.31	94.91	96.07	**99.34**	98.40
5	98.30	97.26	98.98	97.88	**100.00**	**100.00**	99.99	**100.00**
6	84.25	82.16	95.15	73.46	89.86	**100.00**	98.15	99.73
7	82.65	80.47	76.60	89.63	91.84	97.63	99.11	**99.56**
8	56.61	70.51	58.12	80.96	80.36	97.95	98.61	**99.26**
9	73.91	72.69	63.16	70.98	93.58	98.67	98.74	**99.26**
10	83.51	86.78	55.57	84.65	94.28	99.00	**99.99**	99.84
11	68.97	64.59	70.16	90.96	92.64	99.28	99.82	**99.94**
12	60.72	59.18	54.74	91.46	94.37	99.46	**99.60**	99.19
13	25.69	45.29	41.93	85.65	90.97	98.10	96.49	**99.43**
14	93.25	96.48	97.05	73.71	94.58	**100.00**	**100.00**	**100.00**
15	99.66	98.45	99.04	99.52	99.65	99.49	**100.00**	99.70
OA (%)	80.92	82.39	77.64	89.05	93.73	98.80	99.34	**99.51**
AA (%)	79.61	78.49	79.15	87.82	93.56	98.93	99.22	**99.56**
$\kappa \times 100$	79.33	80.64	75.81	88.15	93.63	98.71	99.28	**99.47**

3.5. Ablation Analysis

To comprehensively demonstrate the effectiveness of our proposed method, we conducted ablation experiments on the Indian Pines dataset, analyzing the impact of different components on the overall model. As shown in Table 6, we considered four combinations and evaluated their impact on OA, AA, and κ.

Specifically, the complete model was divided into four components, including the Spectral–Spatial Feature Extraction module (SSFE) comprising a dual-branch structure of 3D and 2D convolutional layers, the Tokenizer, LKA, and the TE Module. Notably, when we replaced the Spectral–Spatial Shallow Feature Extraction module with a single 2D convolution layer, the overall accuracy decreased by 1.42%. In the second combination experiment, removing TE resulted in a 2.64% decrease in overall accuracy, with a significant decline in average accuracy. The third combination experiment involved transitioning from parallel 3D–2D convolution layers to a serial arrangement, while removing the LKA module led to a 0.42% decrease in overall accuracy and a substantial 2.88% reduction in average accuracy.

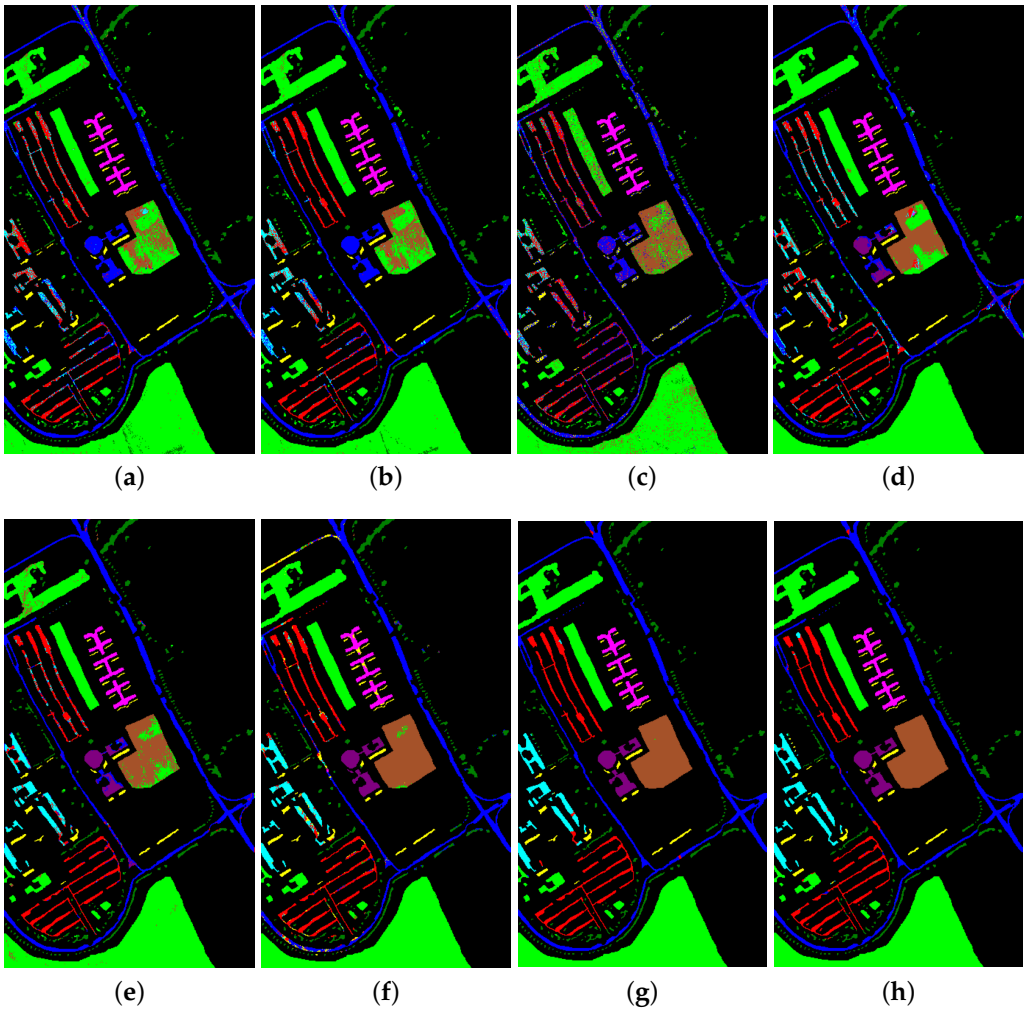


Figure 11. Classification maps of Pavia University dataset. (**a**) SVM, (**b**) EMAP, (**c**) 1D-CNN, (**d**) 2D-CNN, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) SSFTT, (**h**) Proposed.

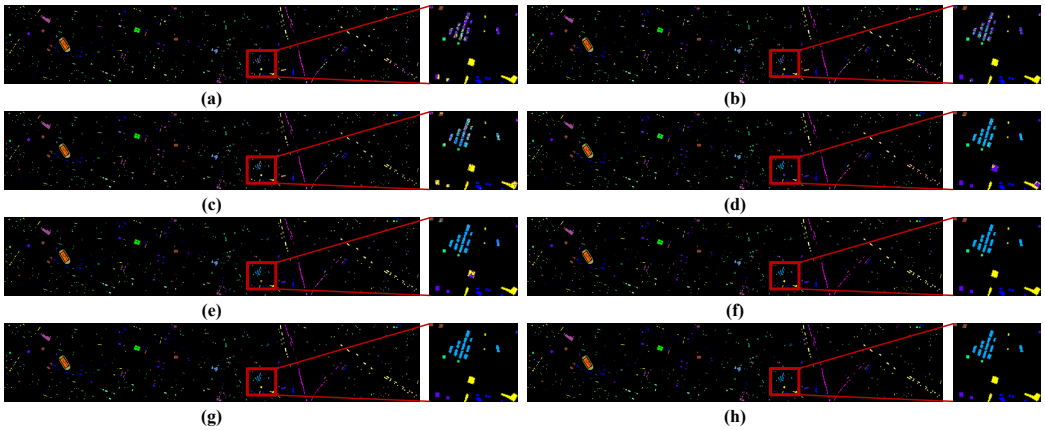


Figure 12. Classification maps of Houston2013 dataset. (**a**) SVM, (**b**) EMAP, (**c**) 1D-CNN, (**d**) 2D-CNN, (**e**) 3D-CNN, (**f**) HybridSN, (**g**) SSFTT, (**h**) Proposed.

In the final experiment, only retaining the Spectral–Spatial Shallow Feature Extraction module caused an average accuracy decrease of 4.21%. This underscores the positive impact of the Transformer architecture on performance enhancement.

In summary, a comprehensive analysis of these combination experiment results further substantiates the effectiveness of our model.

Table 5. Inference time analysis on different datasets.

Dataset	India Pines		Pavia University		Houston2013	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Time (s)	8.09	71.85	26.92	276.65	32.01	100.31

Table 6. Ablation studies on different components for Indian Pines dataset (accuracy in %). The optimal results are bolded.

Cases	Components				Trainable	Indicators			
	SSFE	**Tokenizer**	**LKA**	**TE**	**Parameters (MB)**	**OA (%)**	**AA (%)**	$\kappa \times 100$	***p*** **(*t*-Test)**
1	Only 2D	√	√	√	0.14	92.81	91.91	91.81	0.4834
2	√	√	√	×	0.71	91.59	85.10	90.41	0.4780
3	3D-Conv+2D-Conv	√	×	√	0.56	93.81	90.64	92.94	0.4846
4	√	×	×	×	0.70	90.02	86.12	89.56	0.4727
5	√	√	√	√	0.77	**94.23**	**93.52**	**93.43**	0.4926

4. Conclusions

This paper introduces a network that utilizes a Transformer architecture with a Large-Kernel Attention Module to deeply extract spatial and spectral information from HSIs, significantly improving classification accuracy. In the spectral–spatial shallow feature extraction stage, we employed a dual-branch structure. The first branch primarily utilized 3D convolution to extract spectral–spatial features from the HSI data, while the second branch used 2D convolution to extract spatial features from the HSI data. Before the Transformer Encoder, we developed a Large-Kernel Attention Module, with the objective of improving the quality of feature extraction. This enhancement leads to a better understanding of image data by the Transformer, resulting in an overall improvement in the model's performance. Additionally, it effectively mitigates the impact of redundant information, strengthening the robustness of the Transformer model. Experimental evaluations were conducted on three hyperspectral imaging datasets, comparing this approach with existing classification methods. The results confirm the method's effectiveness and superiority. Future research can focus on the design of an end-to-end lightweight Transformer architecture and leverage a multi-scale feature extraction network to extract diverse feature types, enabling deep exploration of rich feature information in hyperspectral images for further accuracy improvement.

Author Contributions: Conceptualization, W.L. and L.S.; methodology, W.L. and X.W.; software, X.W.; validation, L.S.; investigation, W.L. and X.W.; writing—original draft preparation, X.W.; writing—review and editing, L.S. and Y.Z.; visualization, X.W.; supervision, Y.Z. and L.S.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62076137

Data Availability Statement: The data presented in this study are available in the article

Acknowledgments: The authors thank the anonymous reviewers and the editors for their insightful comments and helpful suggestions that helped improve the quality of our manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

AA	Average Accuracy
EMAP	Extended Multiple Attribute Profile
HSI	Hyperspectral image
LKA	Large-Kernal Attention
LN	Normalization layers

κ	Kappa coefficient
MHSA	Multi-Head Self-Attention
OA	Overall Accuracy
PCA	Principal Component Analysis
SVM	Support Vector Machine
TE	Transformer Encoder

References

1. Li, J.; Zheng, K.; Liu, W.; Li, Z.; Yu, H.; Ni, L. Model-Guided Coarse-to-Fine Fusion Network for Unsupervised Hyperspectral Image Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
2. Li, J.; Zheng, K.; Li, Z.; Gao, L.; Jia, X. X-Shaped Interactive Autoencoders with Cross-Modality Mutual Learning for Unsupervised Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
3. Sun, L.; Cheng, S.; Zheng, Y.; Wu, Z.; Zhang, J. SPANet: Successive pooling attention network for semantic segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4045–4057. [CrossRef]
4. García, J.L.; Paoletti, M.E.; Jiménez, L.I.; Haut, J.M.; Plaza, A. Efficient semantic segmentation of hyperspectral images using adaptable rectangular convolution. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
5. Ben-Ahmed, O.; Urruty, T.; Richard, N.; Fernandez-Maloigne, C. Toward content-based hyperspectral remote sensing image retrieval (CB-HRSIR): A preliminary study based on spectral sensitivity functions. *Remote Sens.* **2019**, *11*, 600. [CrossRef]
6. Sun, L.; Wang, Q.; Chen, Y.; Zheng, Y.; Wu, Z.; Fu, L.; Jeon, B. CRNet: Channel-enhanced Remodeling-based Network for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5618314. [CrossRef]
7. Fu, L.; Zhang, D.; Ye, Q. Recurrent thrifty attention network for remote sensing scene recognition. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 8257–8268. [CrossRef]
8. Ren, H.; Du, Q.; Wang, J.; Chang, C.I.; Jensen, J.O.; Jensen, J.L. Automatic target recognition for hyperspectral imagery using high-order statistics. *IEEE Trans. Aerosp. Electron. Syst.* **2006**, *42*, 1372–1385. [CrossRef]
9. Matteoli, S.; Diani, M.; Corsini, G. A tutorial overview of anomaly detection in hyperspectral images. *IEEE Aerosp. Electron. Syst. Mag.* **2010**, *25*, 5–28. [CrossRef]
10. Li, L.; Li, W.; Qu, Y.; Zhao, C.; Tao, R.; Du, Q. Prior-based tensor approximation for anomaly detection in hyperspectral imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 1037–1050. [CrossRef]
11. Pedergnana, M.; Marpu, P.R.; Mura, M.D.; Benediktsson, J.A.; Bruzzone, L. A Novel Technique for Optimal Feature Selection in Attribute Profiles Based on Genetic Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3514–3528. [CrossRef]
12. Song, B.; Li, J.; Dalla Mura, M.; Li, P.; Plaza, A.; Bioucas-Dias, J.M.; Benediktsson, J.A.; Chanussot, J. Remotely Sensed Image Classification Using Sparse Representations of Morphological Attribute Profiles. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 5122–5136. [CrossRef]
13. Xia, J.; Dalla Mura, M.; Chanussot, J.; Du, P.; He, X. Random Subspace Ensembles for Hyperspectral Image Classification with Extended Morphological Attribute Profiles. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4768–4786. [CrossRef]
14. Kwan, C.; Gribben, D.; Ayhan, B.; Bernabe, S.; Plaza, A.; Selva, M. Improving Land Cover Classification Using Extended Multi-Attribute Profiles (EMAP) Enhanced Color, Near Infrared, and LiDAR Data. *Remote Sens.* **2020**, *12*, 1392. [CrossRef]
15. Kwan, C.; Ayhan, B.; Budavari, B.; Lu, Y.; Perez, D.; Li, J.; Bernabe, S.; Plaza, A. Deep Learning for Land Cover Classification Using Only a Few Bands. *Remote Sens.* **2020**, *12*, 2000. [CrossRef]
16. Zhang, A.; Sun, G.; Ma, P.; Jia, X.; Ren, J.; Huang, H.; Zhang, X. Coastal Wetland Mapping with Sentinel-2 MSI Imagery Based on Gravitational Optimized Multilayer Perceptron and Morphological Attribute Profiles. *Remote Sens.* **2019**, *11*, 952. [CrossRef]
17. Huang, J.; Liu, K.; Xu, M.; Perc, M.; Li, X. Background Purification Framework With Extended Morphological Attribute Profile for Hyperspectral Anomaly Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8113–8124. [CrossRef]
18. Ye, Q.; Huang, P.; Zhang, Z.; Zheng, Y.; Fu, L.; Yang, W. Multiview learning with robust double-sided twin SVM. *IEEE Trans. Cyber.* **2021**, *52*, 12745–12758. [CrossRef]
19. Fu, L.; Li, Z.; Ye, Q.; Yin, H.; Liu, Q.; Chen, X.; Fan, X.; Yang, W.; Yang, G. Learning Robust Discriminant Subspace Based on Joint $L_{2,p}$-and $L_{2,s}$-Norm Distance Metrics. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 130–144. [CrossRef]
20. Ye, Q.; Li, Z.; Fu, L.; Zhang, Z.; Yang, W.; Yang, G. Nonpeaked discriminant analysis for data representation. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3818–3832. [CrossRef]
21. Baassou, B.; He, M.; Mei, S. An accurate SVM-based classification approach for hyperspectral image classification. In Proceedings of the 2013 21st International Conference on Geoinformatics, Kaifeng, China, 20–22 June 2013; pp. 1–7.
22. Guo, B.; Gunn, S.R.; Damper, R.I.; Nelson, J.D. Customizing kernel functions for SVM-based hyperspectral image classification. *IEEE Trans. Image Proc.* **2008**, *17*, 622–629. [CrossRef] [PubMed]
23. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
24. Kang, X.; Xiang, X.; Li, S.; Benediktsson, J.A. PCA-based edge-preserving features for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7140–7151. [CrossRef]

25. Wang, F.; Zhang, R.; Wu, Q. Hyperspectral image classification based on PCA network. In Proceedings of the 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Los Angeles, CA, USA, 21–24 August 2016; pp. 1–4.
26. Villa, A.; Benediktsson, J.A.; Chanussot, J.; Jutten, C. Hyperspectral image classification with independent component discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4865–4876. [CrossRef]
27. 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]
28. Su, Y.; Gao, L.; Jiang, M.; Plaza, A.; Sun, X.; Zhang, B. NSCKL: Normalized Spectral Clustering With Kernel-Based Learning for Semisupervised Hyperspectral Image Classification. *IEEE Trans. Cybern.* **2023**, *53*, 6649–6662. [CrossRef]
29. Lee, H.; Kwon, H. Going Deeper With Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef]
30. Sun, L.; Fang, Y.; Chen, Y.; Huang, W.; Wu, Z.; Jeon, B. Multi-structure KELM with attention fusion strategy for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
31. Gao, H.; Yang, Y.; Li, C.; Gao, L.; Zhang, B. Multiscale residual network with mixed depthwise convolution for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3396–3408. [CrossRef]
32. Gao, H.; Chen, Z.; Xu, F. Adaptive spectral-spatial feature fusion network for hyperspectral image classification using limited training samples. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102687. [CrossRef]
33. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 1–12. [CrossRef]
34. Zhao, W.; Du, S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4544–4554. [CrossRef]
35. Yang, J.; Zhao, Y.Q.; Chan, J.C.W. Learning and transferring deep joint spectral–spatial features for hyperspectral classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4729–4742. [CrossRef]
36. 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]
37. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281. [CrossRef]
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
39. 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.* **2017**, *56*, 847–858. [CrossRef]
40. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.J.; Pla, F. Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 740–754. [CrossRef]
41. Wang, J.; Song, X.; Sun, L.; Huang, W.; Wang, J. A novel cubic convolutional neural network for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4133–4148. [CrossRef]
42. Roy, S.K.; Hong, D.; Kar, P.; Wu, X.; Liu, X.; Zhao, D. Lightweight heterogeneous kernel convolution for hyperspectral image classification with noisy labels. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
44. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10012–10022.
46. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7262–7272.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst. (NIPS)* **2017**, *30*, 1–11.
48. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral Image Transformer Classification Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
49. 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]
50. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
51. 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]
52. Huang, X.; Dong, M.; Li, J.; Guo, X. A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]

53. Fang, Y.; Ye, Q.; Sun, L.; Zheng, Y.; Wu, Z. Multi-Attention Joint Convolution Feature Representation with Lightweight Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14.
54. Xu, Y.; Xie, Y.; Li, B.; Xie, C.; Zhang, Y.; Wang, A.; Zhu, L. Spatial-Spectral 1DSwin Transformer with Group-wise Feature Tokenization for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16.
55. Zhang, S.; Zhang, J.; Wang, X.; Wang, J.; Wu, Z. ELS2T: Efficient Lightweight Spectral–Spatial Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
56. Zhang, K.; Zhu, D.; Min, X.; Zhai, G. Implicit Neural Representation Learning for Hyperspectral Image Super-Resolution. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6. [CrossRef]
57. Dong, W.; Wang, H.; Wu, F.; Shi, G.; Li, X. Deep Spatial–Spectral Representation Learning for Hyperspectral Image Denoising. *IEEE Trans. Comput. Imaging* **2019**, *5*, 635–648. [CrossRef]
58. Tulczyjew, L.; Kawulok, M.; Nalepa, J. Unsupervised Feature Learning Using Recurrent Neural Nets for Segmenting Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 2142–2146. [CrossRef]
59. Nalepa, J.; Myller, M.; Imai, Y.; Honda, K.I.; Takeda, T.; Antoniak, M. Unsupervised Segmentation of Hyperspectral Images Using 3-D Convolutional Autoencoders. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1948–1952. [CrossRef]
60. Sellami, A.; Tabbone, S. Deep neural networks-based relevant latent representation learning for hyperspectral image classification. *Pattern Recognit.* **2022**, *121*, 108224. [CrossRef]
61. Zhang, S.; Zhang, X.; Li, T.; Meng, H.; Cao, X.; Wang, L. Adversarial Representation Learning for Hyperspectral Image Classification with Small-Sized Labeled Set. *Remote Sens.* **2022**, *14*, 2612. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.