remote sensing

MDPI

Article

MEA-EFFormer: Multiscale Efficient Attention with Enhanced Feature Transformer for Hyperspectral Image Classification

Qian Sun [1,2], Guangrui Zhao [2,3], Yu Fang [3], Chenrong Fang [4], Le Sun [2,3,5] and Xingying Li [6,*]

[1] School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; sunqian@nuist.edu.cn
[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China; cs_zhaogr@nuist.edu.cn (G.Z.); sunlecncom@nuist.edu.cn (L.S.)
[3] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; cs_yfang@nuist.edu.cn
[4] College of Intelligence and Computing, Tianjin University, Tianjin 300072, China; fangchenrong@tju.edu.cn
[5] Institute of Intelligent Network and Information System, Nanjing University of Information Science and Technology, Nanjing 210044, China
[6] Guangxi Forest Resources and Environment Monitoring Center, Nanning 530028, China
* Correspondence: caflxy@gmail.com

Abstract: Hyperspectral image classification (HSIC) has garnered increasing attention among researchers. While classical networks like convolution neural networks (CNNs) have achieved satisfactory results with the advent of deep learning, they are confined to processing local information. Vision transformers, despite being effective at establishing long-distance dependencies, face challenges in extracting high-representation features for high-dimensional images. In this paper, we present the multiscale efficient attention with enhanced feature transformer (MEA-EFFormer), which is designed for the efficient extraction of spectral–spatial features, leading to effective classification. MEA-EFFormer employs a multiscale efficient attention feature extraction module to initially extract 3D convolution features and applies effective channel attention to refine spectral information. Following this, 2D convolution features are extracted and integrated with local binary pattern (LBP) spatial information to augment their representation. Then, the processed features are fed into a spectral–spatial enhancement attention (SSEA) module that facilitates interactive enhancement of spectral–spatial information across the three dimensions. Finally, these features undergo classification through a transformer encoder. We evaluate MEA-EFFormer against several state-of-the-art methods on three datasets and demonstrate its outstanding HSIC performance.

Keywords: hyperspectral image (HSI) classification; multi-feature; channel attention mechanisms; transformer

Citation: Sun, Q.; Zhao, G.; Fang, Y.; Fang, C.; Sun, L.; Li, X.. MEA-EFFormer: Multiscale Efficient Attention with Enhanced Feature Transformer for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 1560. https://doi.org/10.3390/rs16091560

Academic Editors: Junjun Jiang, Bihan Wen, Kui Jiang, Leyuan Fang, Jiayi Ma and Gemine Vivone

Received: 19 March 2024
Revised: 22 April 2024
Accepted: 25 April 2024
Published: 27 April 2024

1. Introduction

Hyperspectral images (HSIs) are widely used in remote sensing (RS) due to their abundance of spatial and spectral information [1]. Compared with natural images, HSIs consist of numerous dense and narrow spectral bands [2], allowing for precise identification of land categories [3–6]. Consequently, HSIs have distinct advantages in various fields based on these characteristics, including ground material identification [7], precision agriculture [8,9] and scene understanding [10]. Among these applications, HSI classification emerges as a critical task.

In recent years, various HSI classification methods have been suggested, including support vector machine (SVM) [11,12], k-nearest neighbors (KNN) [13,14] and random forest (RF) [15]. These algorithms have achieved remarkable results by utilizing spectral information effectively. While SVM performs well in high-dimensional problems, it needs to select some indispensable parameters [16]. In [17], linear discriminant analysis (LDA)

was used for HSI classification. However, these classification methods still have much room for improvement due to their lack of spatial characteristics. Therefore, the extended morphological attribute profile (EMAP) took texture and morphological features into consideration [18]. Subsequently, there were also methods to capture texture and edge features in images based on Gabor filters [19,20]. Nevertheless, these methods fail to consider the correlation between spatial and spectral information. Thus, in [21–23], spectral and spatial information were jointly extracted for HSI classification. Nevertheless, these traditional methods extract shallow texture features and fail to reflect deep connections between spectral–spatial features. In this thesis, we develop a lower branch beyond the spectral–spatial features of the HSI itself to perform the processing of the local binary pattern (LBP) features under the first component band. LBP is an efficient approach for describing the local texture of an image and generates a binary encoding by comparing the differences in gray values between a pixel and its neighboring pixels. This operation can enhance the representation of spatial information by modeling spatial structures such as edges and corner points in the image. The upper and lower branches combine spectral and spatial information in order to more comprehensively and accurately describe the properties of HSI.

Compared with traditional machine learning classification technology, deep learning (DL) [24] has the characteristics of automatic learning and strong classification ability and is widely used in HSIC. It is important to note that the CNN model [25,26] stands as the most prevalent choice because of its proficiency in extracting features efficiently, its adaptability in handling high-dimensional data processing, its retention of spatial information, and its capability in managing large-scale data processing. Yang et al. [27] introduced four novel deep learning models, encompassing both two-dimensional CNN and three-dimensional CNN. Their research revealed that, while the 2D-CNN model excelled at exploiting spatial characteristics, it lacked consideration of spectral correlations. On the other hand, 3D-CNN models, despite having a higher number of network parameters compared to 2D models, effectively utilize spectral information alongside spatial features. To leverage both spatial and spectral information, Roy et al. [28] developed a hybrid spectral CNN (HybridSN) that integrates a spectral–spatial 3D-CNN with a spatial 2D-CNN. Unlike models solely reliant on 3D-CNNs, HybridSN incorporates elements of 2D-CNNs to extract a more abstract spatial representation, consequently streamlining the model's complexity. Zhu et al. [29] devised the deformable HSI classification network (DHCNet): a CNN-based method tailored for hyperspectral image classification. DHCNet integrates deformable convolutional sampling locations that dynamically conform in size and shape to accommodate the intricate spatial characteristics found in HSIs. This adaptive feature enables enhanced extraction of spatial features, leveraging complex structural information more efficiently. Jia et al. [30] proposed a lightweight convolutional neural network (LWCNN) for HSIC and designed a two-scale convolutional (DSC) module to process joint spatial–spectral information features [31]. By merging operations, the parameter sizes are greatly reduced. It has the advantage of efficiency and robustness when solving small-sample-set problems. Gong et al. [32] developed a multiscale convolutional and diversified metric CNN (DPP-DML-MS-CNN). Diversifying depth measurements rooted in multiscale features and determinantal point processes (DPPs) [33] has enhanced the characterization and classification abilities of HSI. However, general CNN for HSIC tends to focus overly on local information, and it is difficult to comprehensively capture the trends of spectral band curves. We thus use a hybrid 3D- and 2D-CNN as a feature extractor to refine representative local features and combine it deeply with a transformer, which is excellent at global modeling, to adequately understand the global spectral trend features.

Although a deep neural network provides an invaluable contribution to HSI classification, HSIs have a high data dimension and a large number of spectral channels, which greatly increases the number of parameters of the CNN model, requires more computing resources and is prone to overfitting problems. Recurrent neural networks (RNNs) [34,35] can utilize each band in the spectrum by applying the cyclic operator layer by layer, thus

obtaining fewer training parameters than convolutional neural networks (CNNs) and making the training and reasoning phases more efficient. By generating GANs [36–38], the discriminator training process continues to be effective through network confrontation and competition, which can alleviate the overfitting phenomenon in the training process. ResNet [39–41] mitigates the problem of disappearing gradients and avoids the loss of accuracy as the network deepens. In addition to the above methods, there are many classic deep learning methods such as autoencoders (AEs) [42,43], deep confidence networks (DBNs) [44,45], complete convolutional networks (FCNs) [46,47] and capsule networks (CapsNets) [48,49].

Transformers have significant advantages when processing sequential data and can establish global relationships, but they still encounter many challenges, such as limited spatial feature extraction capabilities or high computational costs [50]. A transformer is a neural network architecture based on a self-attention mechanism. Its emergence abandons the traditional RNN or CNN and allows the model to be trained in parallel and to have global information. The key feature of the transformer model is that it relies on multi-head self-attention mechanism (MHSAs) to capture dependencies between different elements in the input sequence, regardless of their position or distance in the sequence. With their powerful parallel computing capabilities, good scalability, ability to handle long-distance dependencies and advantages in processing long sequence data, transformers have broad application prospects in various fields and are still being continuously improved and expanded. Hong et al. [51] used transformers in HSI classification tasks for the first time. Later, Sun et al. [52] developed a spectral–spatial feature tokenization transformer (SSFTT) model to capture spectral–spatial features and high-level semantic features. Tu et al. [53] introduced an architecture named the local semantic feature aggregation transformer (LSFAT), which employs local semantic feature aggregation. This design enhances the capability of transformers to effectively capture long-term dependencies within multiscale features. Qiao et al. [54] recently developed a new type of hierarchical dual-frequency transformer network (DFTN) in which a frequency domain feature extraction (FDFE) block was proposed to capture high-frequency and low-frequency features separately, allowing the network to effectively utilize the input data. The multi-layer feature information in the system improves the modeling ability for complex relationships. Wang et al. [55] proposed a novel extended spectral spatial attention network (ESSAN) for HSI data classification when training samples are insufficient. For the whole network structure of a transformer, the information data for its global modeling all come from the token data generated by local patch transformation. In addition, some transformer networks based on the mask technique can effectively improve the classification performance in scenarios with insufficient samples [56,57]; such a technical improvement is also eye-catching. Our proposed SSEA module achieves further enhancement of the spectral–spatial features by computing the attention in three dimensions and also skillfully incorporates the LBP information. This operation achieves high-performance feature refinement for the subsequently generated token and also eliminates redundant information to a certain extent, providing accurate and representative global modeling information for the MHSA operation in the transformer.

In this paper, a novel multiscale efficient attention with enhanced feature transformer is presented for HSI classification. It mainly includes a multiscale efficient attention feature extraction module, a spectral–spatial enhancement attention module and a transformer encoder. The ingenious feature extraction method adequately exploits the abundant spatial and spectral information in HSI. The SSEA module enhances the interaction of spectral information with spatial features and LBP features from multiple perspectives. The transformer encoder fully integrates the key features through multi-head self-attention to optimize the feature representation. The main contributions of this paper are listed as follows:

(1) MEA-EFFormer is a multiscale efficient attentional feature extraction module that incorporates an efficient channel attention mechanism with multiscale convolution. It facilitates the mining of details in spectral–spatial information and solves the problem of fine-grained feature loss during single-scale sampling.

(2) MEA-EFFormer uses an SSEA module. Based on three directions, C-H, C-W and H-W, it captures the dependencies between spectral–spatial LBP information, refines the scale of the features and improves the perception of the attention mechanisms.
(3) The classification performance of MEA-EFFormer outperforms several classical and SOTA methods. Experiments on all three well-known datasets show that the proposed method has excellent classification performance.

2. Materials and Methods

Figure 1 depicts the overarching structure of the HSIC task incorporating the innovative MEA-EFFormer. This architecture is built upon three core components: a spectral–spatial multiscale efficient attention feature extraction module, a spectral–spatial enhancement attention module, and a transformer encoder module.

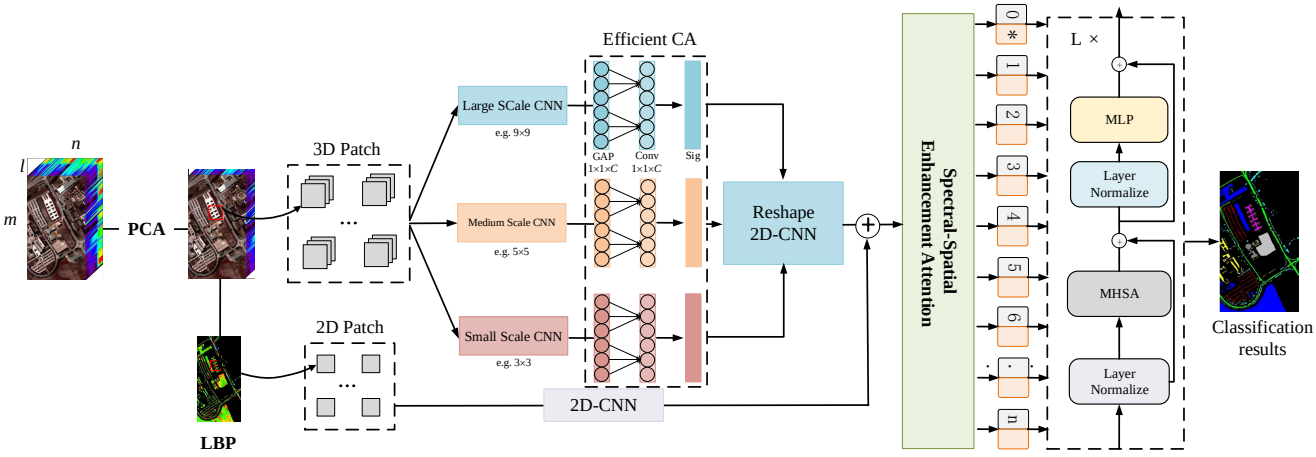


Figure 1. The architecture of the proposed MEA-EFFormer network. The network can be divided into three stages: data preprocessing, feature extraction and processing, and the transformer encoder. The data preprocessing stage includes principal component analysis (PCA) to extract the main bands from the raw HSI and local binary pattern (LBP) extraction. The feature extraction and processing stage is mainly a multiscale efficient attention feature extraction module and a spectral–spatial enhancement attention module. Finally, the obtained refined features are fed into the transformer encoder for classification operations.

2.1. Spectral–Spatial Multi-Feature Convolution Extraction

2.1.1. Multiscale Efficient Attention Feature Extraction Module

Original HSI data contain abundant spatial–spectral features that are manifested through specific spectrum reflectances and spatial morphologies. Preserving the relationship between spatial and spectral information is crucial for accurately distinguishing various land cover in classification tasks. Therefore, we employ a multiscale 3D-CNN that enables simultaneous extraction from multiple perspectives. The 3D-CNN utilizes three-dimensional kernels to perform convolution operations on the HSI along the spectral domain. However, sampling at a single scale may lead to the loss of fine-grained features. To address this issue, we utilize three distinct types of convolution kernels, each characterized by a unique scale size, to enable specialized extraction of relevant features. To prevent redundancy in high-dimensional data and mitigate the risk of overfitting, we conduct principal component analysis (PCA) dimensionality reduction on the raw HSI $\mathbf{I} \in \mathbf{R}^{m \times n \times l}$ to $\mathbf{I} \in \mathbf{R}^{m \times n \times b}$ prior to 3D convolution operations, where $m \times n$ represents the spatial size, l represents the number of bands in the original hyperspectral data, and b is the number of bands after PCA. This step aims to retain the most significant bands in order to optimize the model's performance and ensure the extracted features are more representative. Subsequently, we partition the HSI data into numerous 3D cubes $\mathbf{P} \in \mathbf{R}^{s \times s \times b}$ to facilitate

convolution operations. The multiscale propagation operation for the jth feature cube at the (x, y, z) position on the ith layer is expressed as:

$$\begin{aligned} \alpha_{ij}^{xyz} &= \theta\left(\sum_{m}\sum_{h=0}^{HL_{i-1}}\sum_{w=0}^{WL_{i-1}}\sum_{r=0}^{RL_{i-1}} w_{ijm}^{hwr}\alpha_{(i-1)m}^{(x+h)(y+w)(z+r)} + b_{ij}\right) \\ \beta_{ij}^{xyz} &= \theta\left(\sum_{m}\sum_{h=0}^{HM_{i-1}}\sum_{w=0}^{WM_{i-1}}\sum_{r=0}^{RM_{i-1}} w_{ijm}^{hwr}\beta_{(i-1)m}^{(x+h)(y+w)(z+r)} + b_{ij}\right) \\ \delta_{ij}^{xyz} &= \theta\left(\sum_{m}\sum_{h=0}^{HS_{i-1}}\sum_{w=0}^{WS_{i-1}}\sum_{r=0}^{RS_{i-1}} w_{ijm}^{hwr}\delta_{(i-1)m}^{(x+h)(y+w)(z+r)} + b_{ij}\right) \end{aligned} \tag{1}$$

Here, m denotes the feature cube associated with the jth feature cube. The three types of convolution kernel sizes are represented by $HL \times WL \times RL$, $HM \times WM \times RM$ and $HS \times WS \times RS$. The variables w and b correspond to the weight and bias parameters, respectively, while θ represents the activation function. The functions α_{ij}^{xyz}, β_{ij}^{xyz} and δ_{ij}^{xyz} are the extracted features at different scales. To mitigate the sensitive volatility of the spectral information due to scale changes, we input them into efficient channel attention (ECA) separately to realize the alignment fusion of the feature maps at different scales. The process of ECA is illustrated in Figure 2.

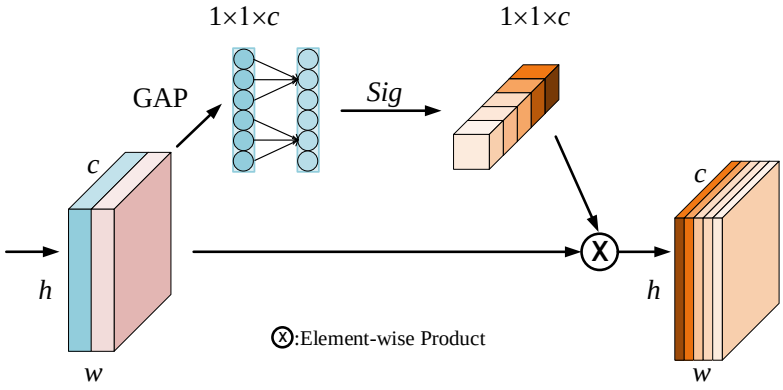


Figure 2. Illustration of ECA. It uses global average pooling and a one-dimensional convolution operation with an adaptive convolution kernel to compute the weights under each band, followed by an activation function to implement the mapping of the attention weights.

Efficient channel attention is a mechanism for enhanced modeling of neural networks by focusing on the importance of different channels within the feature map. It utilizes lightweight and highly efficient computation to break through the cost limitations of complex band computation for HSI data. For the spectral–spatial data α_{ij}^{xyz} output from one scale branch, we use the global average pool (GAP) to compress it into a feature map with a spatial dimension of 1×1. Subsequently, the one-dimensional convolution of the adaptive convolution kernel is utilized to compute the weights under each band, and finally, a *Sigmoid* function is introduced to map the weights between 0 and 1. These weights are then utilized to modulate the original feature map, highlighting information-rich bands while suppressing relatively irrelevant bands. The size of the adaptive convolution kernel is generically related to the number of spectral bands under the branch of the scale, as expressed by the formula:

$$k = \left|\frac{log_2(C)}{\gamma} + \frac{b}{\gamma}\right| \tag{2}$$

where C is the number of spectral bands characterized, and b and r are set to 1 and 2, respectively. The advantage of this module is that global dependencies between spectral bands are captured using a small computational cost, and the whole process can be represented as follows:

$$out = Sig(Conv(GAP(F), C_F) \times F), F = \alpha_{ij}^{xyz}, \beta_{ij}^{xyz}, \delta_{ij}^{xyz} \tag{3}$$

After performing ECA computation on the three scale branches, we cascade fuse them with $1 \times 1 \times 1$ unit convolution and extract 2DCNN features uniformly. This is done as follows:

$$v_{ij}^{xyz} = \text{cat}(\alpha_{ij}^{xyz}, \beta_{ij}^{xyz}, \delta_{ij}^{xyz}) \otimes \text{Filter}_{1\times1\times1} \tag{4}$$

$$v_{ij}^{xy} = \theta\left(\sum_{m}\sum_{h=0}^{H_{i-1}}\sum_{w=0}^{W_{i-1}} w_{ijm}^{hw} v_{(i-1)m}^{(x+h)(y+w)} + b_{ij}\right) \tag{5}$$

Here, m represents the feature map associated with the jth feature, while $H \times W$ denotes the dimensions of the 2D convolution kernel. The parameters w and b correspond to the weights and biases, respectively, and θ represents the activation function. The HSI features at this time have extensive detailed information and high expressive capability. To further enhance the surface texture information and capture the spatial fine-grained features, we next extract the LBP features from the original HSI image.

2.1.2. LBP Convolution Feature Processing

The local binary pattern (LBP) is an operator used to characterize local features of an image and has significant advantages such as grayscale invariance and rotational invariance. The LBP feature provides a description of the texture characteristics of the HSI surface, which utilizes spatially localized pixel grayscale differences to highlight landscape detail information. Each pixel in the image is compared as a center pixel with its domain pixel's gray value, and binary bits are set for the domain pixel depending on the result. Subsequently, this is converted to decimal to get the LBP code. The specific computational operations are as follows:

$$\begin{aligned} LBP_{P,R} &= \sum_{p=0}^{P-1} 2^p \cdot S_{\text{lbp}}(x_p, y_p) \\ S_{\text{lbp}}(x) &= \begin{cases} 1, & \text{if } I(x,y) \geq I_c \\ 0, & \text{if } I(x,y) < I_c \end{cases} \end{aligned} \tag{6}$$

where S_{lbp} is the binary bit of the neighboring pixel, $I(x,y)$ is the grayscale value of the neighboring pixel, and I_c is the grayscale value of the center pixel. P is the number of pixels in the neighborhood, R is the radius of the neighborhood, and (x_p, y_p) are the coordinates of the pixels in the neighborhood. In the following, we perform a 2D convolution operation on the LBP spatial features to obtain a feature map with the same dimensions as the spectral–spatial convolution in the previous section. The operation here is the same as Equation (5).

2.2. Spectral–Spatial Enhancement Attention Module

After extracting the LBP convolution features, we superimpose them onto the spectral–spatial branching features from the spectral dimension, i.e., $\mathbf{M} \in \mathbf{R}^{b+1\times h\times w}$, with h and w being the convolution feature map dimensions. Subsequently, the LBP convolution and spectral–spatial convolution were merged through a 2D convolution unit employing a 1×1 kernel size, resulting in a b-band convolution feature map. This fusion process effectively combines the complementary information from both convolutions, enriching the representational power of the feature map. However, the correlation between the original spectral features and the improved spatial features can significantly impact the final classification results. To address this, we propose a spectral–spatial enhancement attention (SSEA) module that facilitates the interactive enhancement of spatial and spectral information from two vertical directions and one horizontal direction. The specific process of SSEA is shown in Figure 3. SSEA establishes the dependency relationship between spectral dimension C and spatial dimensions H and W by rotational operation and calculates the attention weights in the three directions $C-H$, $C-W$ and $H-W$. Following this, the three types of correlation weights are combined to promote interaction and strengthen the representation

of spectral features, thereby ensuring a more comprehensive and nuanced understanding of the data.

In this stage, the features undergo 90-degree counterclockwise rotation along the H axis, resulting in dimensions of $W \times H \times C$. Following this transformation, both average and maximum pooling operations are performed in the W dimension. The average pooling operation calculates the mean value of image regions to smooth the data and suppress noise, thereby enhancing the stability and coupling of subsequent spectral–spatial information. In contrast, maximum pooling discards other information by retaining the maximum pixel value of each region in both the spatial and spectral dimensions. This operation helps to emphasize the salient features in the image. Finally, an activation function is applied to generate attention weights between the H and C dimensions; these weights are then used to modulate the convolution features. The second branch emphasizes the interaction between the spectral dimension C and spatial dimension W. In this phase, the convolution features undergo a 90-degree counterclockwise rotation along the W axis, resulting in dimensions of $H \times C \times W$. Attention weights for this branch are computed similarly to those in the first branch. The third branch is focused on the interaction between spatial dimensions H and W. It directly applies average and maximum pooling to the convolution features, enhancing both the original spatial information and the spatial information derived from LBP features. This approach further sharpens the ability to capture intricate ground category features. Once the three branches have been computed, their outputs are averaged and aggregated, fostering interaction between spectral and spatial dimensions and resulting in a richer and more comprehensive feature representation.

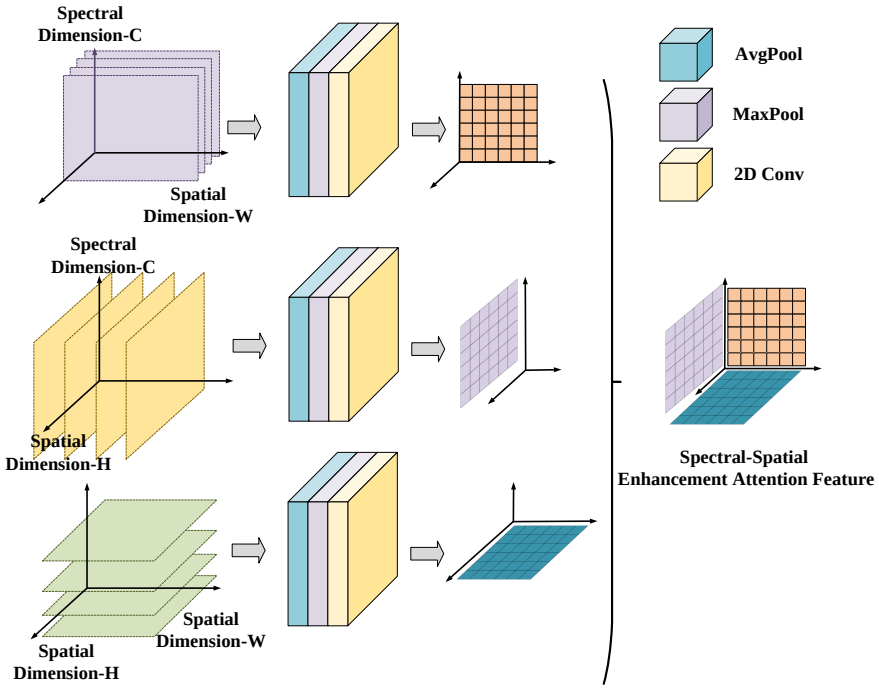


Figure 3. Illustration of SSEA. It consists of three branches that establish the dependencies between the spectral dimension C and the spatial dimensions H and W by means of rotational operations, and it computes the attention weights in each of the three directions.

2.3. Transformer Encoder Module

The transformer encoder block mainly consists of two residual structures, which are represented by the multi-head self-attention mechanism (MHSA) module and the multi-layer perceptron (MLP) module, as shown in Figure 4a. For the enhanced spectral–spatial features $\mathbf{N} \in \mathbf{R}^{c \times h \times w}$, which were extracted in the previous section, we perform an embedding flattening operation on their spatial data to obtain a 2D structure $\mathbf{N}' \in \mathbf{R}^{c \times h \times w}$ that is suitable as input for the transformer, where $d = h \times w$. As shown in Figure 4b, we

initialize three update matrices $\mathbf{W^Q}$, $\mathbf{W^K}$, $\mathbf{W^V}$ to map the input features $\mathbf{N'}$ to $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrices, respectively, which are composed of h head components:

$$\begin{aligned} \mathbf{Q} &= \{\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_i, \ldots, \mathbf{Q}_h\} \\ \mathbf{K} &= \{\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_i, \ldots, \mathbf{K}_h\} \\ \mathbf{V} &= \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_i, \ldots, \mathbf{V}_h\} \end{aligned} \tag{7}$$

where h is the number of heads, and $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbf{R}^{c\times(d/h)}$. Next, we use them to compute the attention scores:

$$\mathbf{SA}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d}}\right)\mathbf{V}_i \tag{8}$$

After obtaining the attention results, we combine the multiple results and perform a linear transformation to obtain the final global interaction result:

$$\mathbf{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\mathbf{SA}_1, \mathbf{SA}_2, \ldots, \mathbf{SA}_h)\mathbf{W} \tag{9}$$

where $\mathbf{W} \in \mathbf{R}^{d\times d}$ is a parameter matrix used to perform the linear transformation.

Next, the data proceed to the second residual structure MLP. it is composed of two linear layers. Following it, highly expressive and discriminative spectral–spatial features are fitted into multiple categories to realize the classification task for HSI.

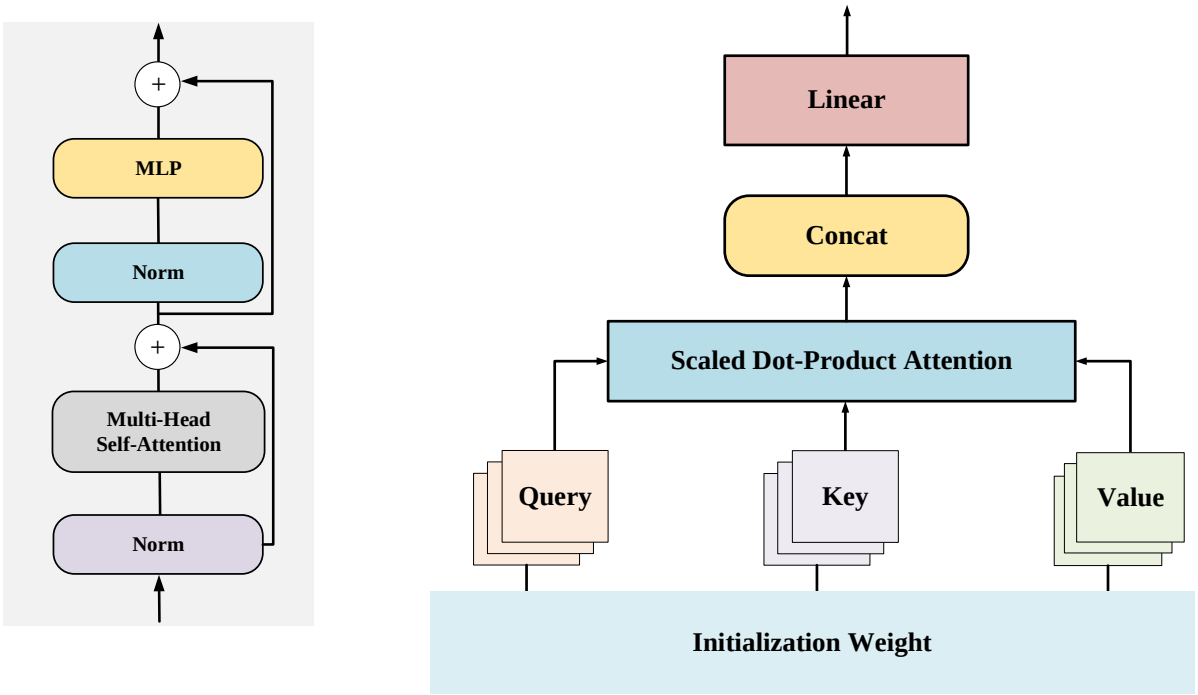


Figure 4. Graphical representation of the transformer encoder: (**a**) The general structure of encoder blocks. (**b**) Multi-head self-attention mechanism.

2.4. Algorithm Summarization for MEA-EFFormer

The overall process of the proposed MEA-EFFormer network is shown in Algorithm 1.

Algorithm 1 MEA-EFFormer network.

Require:
HSI data $\mathbf{I} \in \mathbf{R}^{m \times n \times l}$; ground truth $\mathbf{Y} \in \mathbf{R}^{m \times n}$; PCA bands number b; input patch size s; training sample rate μ%; attention head number h; epochs E; batch size = 128.
Ensure:
Predicted classification labels for the test dataset.
1: Extract the $\mathbf{I}_{pca}$ features from the HSI data and transform them into multiscale convolution features α_{ij}^{xyz}, β_{ij}^{xyz}, δ_{ij}^{xyz} by Equation (1).
2: Perform efficient channel attention operation on α_{ij}^{xyz}, β_{ij}^{xyz}, δ_{ij}^{xyz}, and stack to perform unit convolution and 2D convolution to get V_{ij}^{xy} by Equations (3)–(5).
3: Obtain the $\mathbf{I}_{lbp}$ features by Equation (6) from HSI data and transform them into the convolution features V_{lbp}^{xy} by Equation (5).
4: **for** $i = 1$ to E **do**
5: Perform SSEA operation on V_{ij}^{xy} and V_{lbp}^{xy} to achieve enhancement of spectral–spatial features.
6: Perform multi-head self-attention in the transformer encoder.
7: **end for**
8: Employ the attention outputs and feed them into a linear layer to determine the corresponding labels.
9: Employ the trained model on the test dataset to acquire predicted labels.

3. Experiment and Analysis

In this section, we employ three well-known HSI datasets—Indian Pines (IP), Salinas (SA) and Pavia University (PU)—to evaluate the effectiveness of the proposed method, and we use three metric indicators to give a quantitative assessment of the classification results.

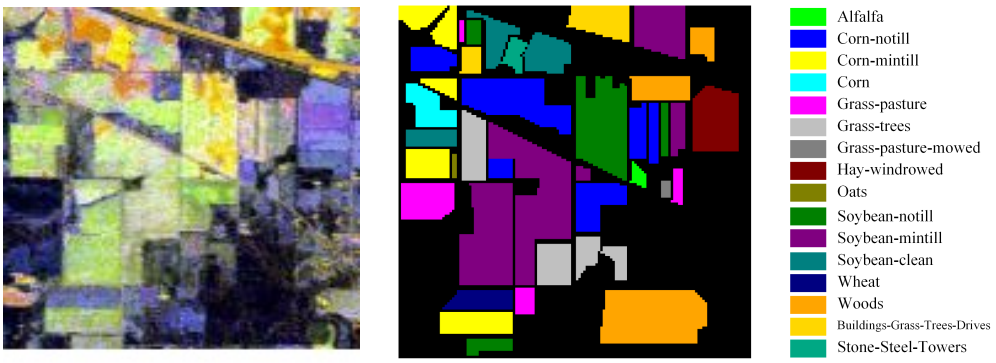
3.1. Data Description

3.1.1. Indian Pines

This dataset was acquired by the AVIRIS sensor over a test site in northwestern Indiana, USA. It has a spatial resolution of 20 m per pixel and covers an area of 145 × 145 pixels. After removing the water absorption bands, the dataset contains 200 spectral bands for analysis. The ground truth for Indian Pines identifies 16 distinct classes, including various crops, forests and other natural vegetation types. This dataset is frequently used to benchmark HSIC algorithms. Figure 5 illustrates the false-color image and labeling map of Indian Pines, and the specific division of the training and testing sets of the samples is shown in Table 1.

Table 1. Training and test samples in Indian Pines, Salinas and Pavia University datasets.

ID	Indian Pines			Salinas			Pavia University		
	Land Cover Class	Training	Test	Land Cover Class	Training	Test	Land Cover Class	Training	Test
C01	Alfalfa	3	43	Brocoli_green_weeds_1	11	1998	Asphalt	67	6564
C02	Corn-notill	72	1356	Brocoli_green_weeds_22	19	3707	Meadows	187	18,462
C03	Corn-mintill	42	788	Fallow	10	1966	Gravel	21	2078
C04	Corn	12	225	Fallow_rough_plow	7	1387	Trees	31	3033
C05	Grass-pasture	25	458	Fallow_smooth	14	2664	Painted metal sheets	14	1331
C06	Grass-tree	37	693	Stubble	20	3939	Bare Soil	51	4978
C07	Grass-pasture-mowed	2	26	Celery	18	3561	Bitumen	14	1316
C08	Hay-windrowed	24	454	Grapes_untrained	57	11,214	Self-Blocking Bricks	37	3645
C09	Oats	1	19	Soil_vinyard_develop	32	6171	Shadows	10	937
C10	Soybean-notill	49	923	Corn_senesced_green_weeds	17	3261			
C11	Soybean-mintill	123	2332	Lettuce_romaine _4wk	6	1062			
C12	Soybean-clean	30	563	Lettuce_romaine_5wk	10	1917			
C13	Wheat	11	194	Lettuce_romaine_6wk	5	911			
C14	Woods	64	1201	Lettuce_romaine_7wk	6	1064			
C15	Buildings-Grass-Trees	20	366	Vinyard_untrained	37	7231			
C16	Stone-Steel-Towers	5	88	Vinyard_vertical_trellis	10	1797			
	Total	513	9736	Total	271	53,858	Total	428	42,348

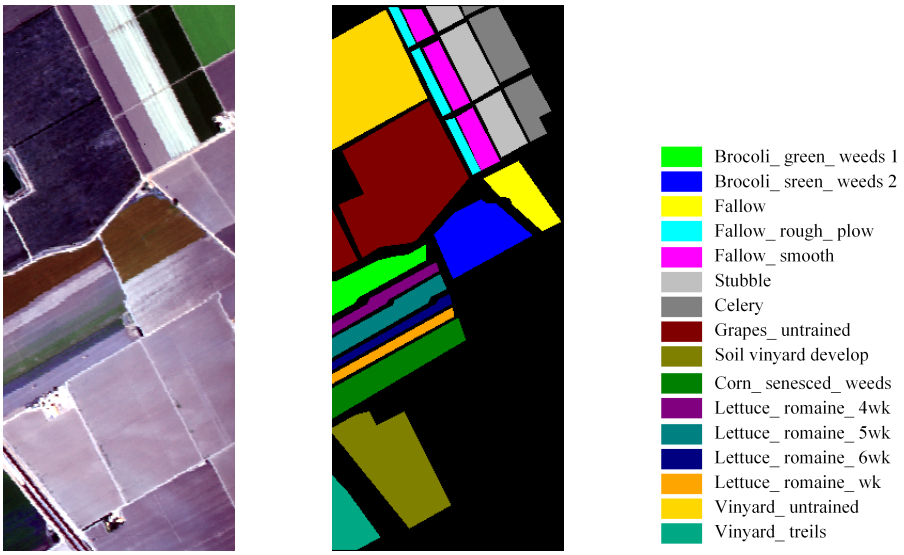


(**a**) (**b**)

Figure 5. Indian Pines dataset. (**a**) False-color map. (**b**) Ground-truth map.

3.1.2. Salinas

This dataset was captured by the AVIRIS sensor; the Salinas dataset focuses on the Salinas Valley in California. It offers a higher spatial resolution of 3.7 m per pixel with dimensions of 512 × 217 pixels. Similar to Indian Pines, the Salinas dataset typically uses 204 spectral bands after water absorption band removal. The ground truth comprises 16 classes representing agricultural fields, vineyards, and bare soil. Researchers often use this dataset to explore the challenges of classifying crops with finer spatial details. Figure 6 illustrates the false-color image and labeling map of Salinas, and the specific division of the training and testing sets of the samples is shown in Table 1.

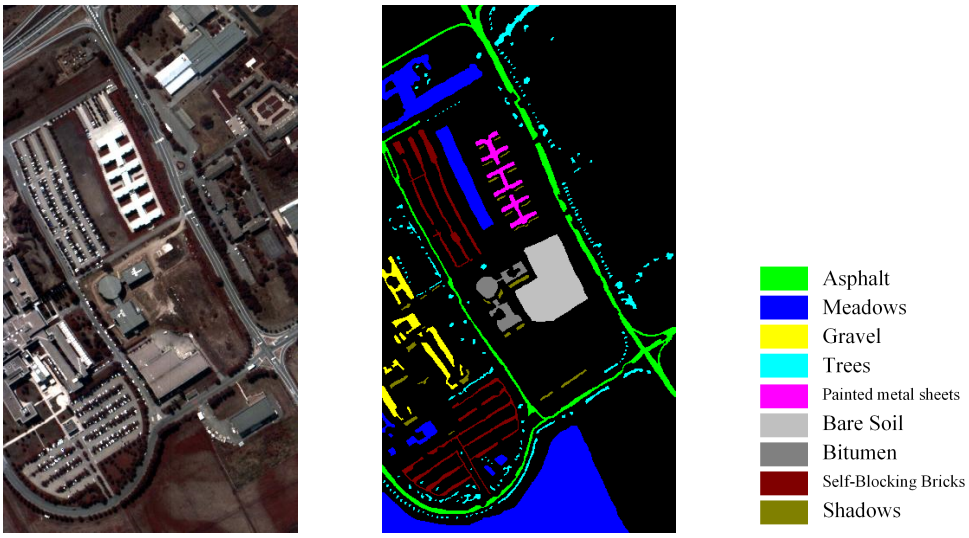


(**a**) (**b**)

Figure 6. Salinas dataset. (**a**) False-color map. (**b**) Ground-truth map.

3.1.3. Pavia University

The ROSIS sensor collected this dataset over an urban area in Pavia, Italy. It boasts a high spatial resolution of 1.3 m per pixel. The dataset contains 103 spectral bands and covers an image size of 610 × 340 pixels. Pavia University offers 9 land-cover classes focused on urban features. This dataset is commonly used to study the classification of urban environments and to address the challenge of working with noisy spectral bands. Figure 7 illustrates the false-color image and labeling map of Pavia University, and the specific division of the training and testing sets of the samples is shown in Table 1.



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

Figure 7. Pavia University dataset. (**a**) False-color map. (**b**) Ground-truth map.

3.2. Experimental Setting

3.2.1. Evaluation Criteria

In order to quantitatively evaluate the experimental results, three quantitative evaluation metrics were employed: overall accuracy (OA), average accuracy (AA) and kappa coefficient. First, OA measures the ratio between the number of correctly classified samples in a dataset and the total number of samples. OA provides an overall assessment of classification performance by indicating the model's ability to correctly classify samples. Second, AA calculates the average accuracy for each category in the dataset. It provides an evaluation of the model's performance on different categories and helps to determine whether the model performs well uniformly across all categories. Finally, the kappa coefficient measures the agreement between the predictions and the true classification while taking into account the stochastic agreement. A kappa value close to 1 indicates that there is strong agreement between prediction and true categorization beyond random consistency.

3.2.2. Environment Configuration

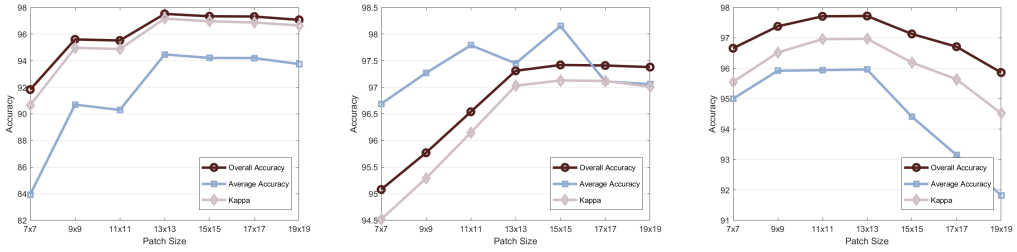
The proposed method was implemented using the PyTorch 2.2.0, while the traditional classical methods used for comparison were executed in the MATLAB R2018b environment. The computational setup included an Intel Xeon Silver 4314 CPU (Intel Corporation, Santa Clara, CA, USA) with 256 GB of RAM, along with an NVIDIA GeForce RTX 4090 GPU server (ASUS, Taipei, Taiwan) equipped with 24 GB of memory. In the comparison involving deep learning and transformer-based methods, parameters were configured as follows: the number of epochs was set to 100, and a batch size of 128 was employed.

3.2.3. Parameter Setting Adjustment

In this subsection, we analyze the impact of several important parameters on the classification results of the proposed network. These parameters are patch size, reduced spectral dimension, learning rate of the network, and the number of attention heads.

Figure 8 illustrates the impact of patch size on the classification metrics OA, AA and kappa. On all three datasets, the accuracy generally increases as the patch size increases. This is likely because larger patches capture more spectral information, which can help the model better distinguish between different classes. However, there is a point of diminishing returns at which increasing the patch size further does not improve accuracy. This is because larger patches may also include irrelevant information that can confuse the model. The specific patch size that yields the best accuracy varies depending on the dataset. For Indian Pines and Pavia University, the highest accuracies are achieved with a patch size of 13×13.

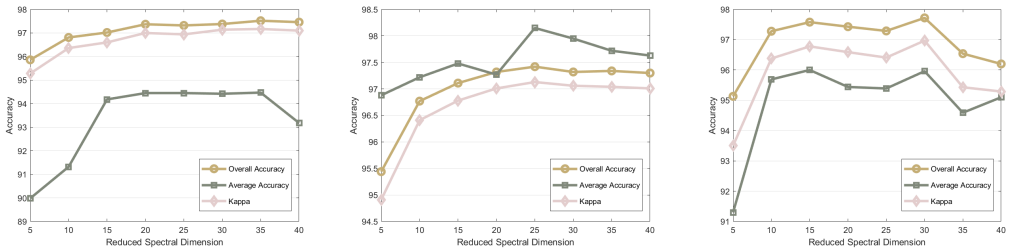
For Salinas, the best accuracy is achieved with a patch size of 15×15. Moreover, our proposed network achieves relatively good and stable performance over a wide range of patch sizes, e.g., $[11 \times 11, 19 \times 19]$. This demonstrates that our proposed network has certain robustness to the parameter of patch size.



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

Figure 8. Patch size as a function of OA, AA and kappa. (**a**) Indian Pines (IP). (**b**) Salinas (SA). (**c**) Pavia University (PU).

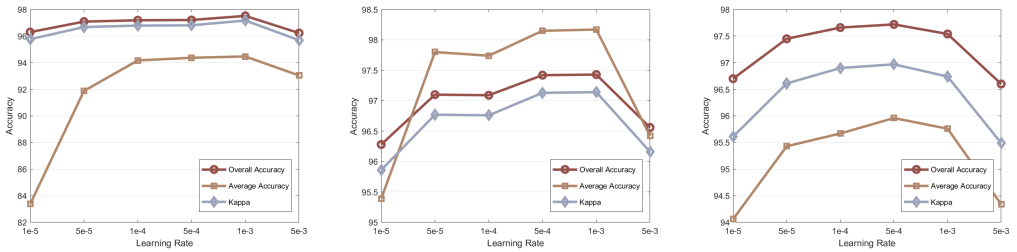
Figure 9 shows the classification results as a function of the reduced spectral dimensions. The reduced spectral dimensions indeed have a strong impact on the performance of the proposed network; however, the proposed network can achieve relatively stable results when the reduced dimensions lie in the range of [15, 35] for all three datasets. Specifically, it achieves the best performance when the reduced dimension is 20 for Indian Pines. For the Salinas and Pavia University datasets, the optimal values of the reduced dimensions are 25 and 30, respectively.



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

Figure 9. Effect of reducing spectral dimensionality on OA, AA and kappa coefficient: (**a**) Indian Pines (IP). (**b**) Salinas (SA). (**c**) Pavia University (PU).

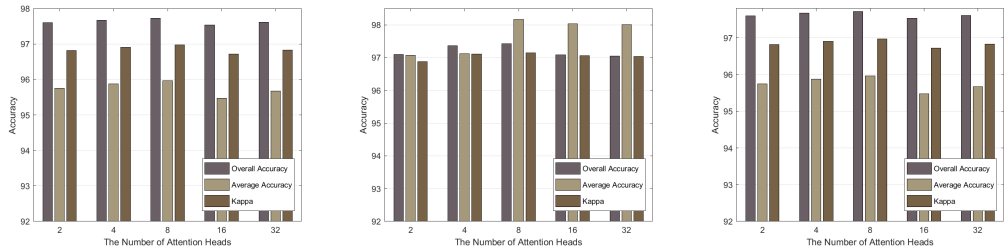
Figure 10 plots the impact of the learning rate on the classification accuracy. For three datasets, the learning rate of the proposed network indeed has a strong impact on the performance. It can be seen that on the three datasets, the OA, AA and kappa curves of the proposed network all show a trend of first increasing and then decreasing, indicating that these three indicators all have an optimal value. In addition, when the learning rate is in the range $[5 \times 10^{-5}, 1 \times 10^{-3}]$, the quantitative indicator values obtained by the proposed network are relatively stable, indicating that the network has certain robustness to the learning rate parameter. In the subsequent classification, we set the learning rate to 1×10^{-3}.



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

Figure 10. Effect of learning rate on OA, AA and kappa coefficient: (**a**) Indian Pines (IP). (**b**) Salinas (SA). (**c**) Pavia University (PU).

Figure 11 plots the number of attention heads as a function of the classification accuracy. The performance of the proposed network is quite stable when the number of attention heads lies in the range of [2, 32]. When the number is eight, the network achieves optimal performance.



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

Figure 11. Effect of the number of attention heads on the OA, AA and kappa coefficient. (**a**) Indian Pines (IP). (**b**) Salinas (SA). (**c**) Pavia University (PU).

3.3. Ablation Study

We performed an ablation experiment employing a 5% sample rate on the Indian Pines dataset. The framework was deconstructed into six discernible sections: principal component analysis (PCA), multiple scales (MS), efficient channel attention (ECA), LBP feature branch (LBP), spectral–spatial enhancement attention module (SSEA) and transformer encoder (TE). Subsequently, a comprehensive evaluation was conducted utilizing performance metrics including OA, AA and the kappa coefficient. The outcomes of these ablation experiments are meticulously tabulated in Table 2 for reference and analysis.

In Case 1, we eliminate the PCA component of the network and input all 200 bands of the Indian Pines dataset into the model. The amount of data computed is more than six times that of our proposed method. The large amount of redundant data being fed into the model computation also brings about a slight decrease in the accuracy metric.

In Case 2, we cancel the MS component of the network and employ a single-scale CNN to extract the HSI features. At this time, the accuracy metrics are all significantly decreased, especially the AA accuracy. This demonstrates the remarkable advantage of the MS component for exploiting the details of the imbalanced category samples.

In Case 3, we eliminate the ECA component of the network and do not compute the attention to the spectral dimension information. At this point, the degradation of AA accuracy is also obvious due to the direct and rough integration of the spectral information at multiple scales. This demonstrates that the ECA component can significantly mitigate the sensitivity of spectral information to scale transformations.

In Case 4, we directly cut the lower branch of the network and do not use the LBP features for spatial information enhancement. The AA accuracy decreases significantly in this case. This indicates that LBP features can provide effective spatial feature enhancement for samples with unbalanced distributions so as to provide the model's capture of the data as a whole.

In Case 5, we remove the SSEA module so that spectral–spatial information and LBP features are merely stacked and fed into the network. Each accuracy metric at this time is slightly decreased. This suggests that the fusion and de-redundancy operations within the SSEA module on the features can extract finer representations, which is conducive to the downstream recognition of the ground surface categories.

In Case 6, the transformer encoder is replaced with a deep residual convolutional network, and all accuracy metrics drop severely. This shows that with purely local features, it is difficult to carry out effective data modeling and the model lacks the overall consideration of global information. And it also further proves that the transformer is favorable for capturing trends within the global spectral curve and for calculating long-distance spatial information.

Table 2. Ablation experiment results (The optimal results are bolded).

Cases	Component						Indicators		
	PCA	MS	ECA	LBP	SSEA	TE	OA (%)	AA (%)	$k * 100$
1	×	✓	✓	✓	✓	✓	97.27	94.19	96.88
2	✓	×	✓	✓	✓	✓	96.92	93.07	96.48
3	✓	✓	×	✓	✓	✓	97.03	92.66	96.61
4	✓	✓	✓	×	✓	✓	97.18	92.04	96.78
5	✓	✓	✓	✓	×	✓	97.14	93.61	96.74
6	✓	✓	✓	✓	✓	×	95.31	90.97	92.24
7	✓	✓	✓	✓	✓	✓	**97.44**	**94.69**	**97.07**

In addition, we conducted additional ablation experiments for the three branches of the SSEA module to explore the impact of each branch on the final classification results. The results, as shown in Table 3, show that the computation of attention with the absence of any branch causes a decrease in the accuracy metric. This also proves that computation using each branch of our proposed SSEA module effectively strengthens the degree of coupling between spectral–spatial and LBP-HSI for refinement of feature representation.

Table 3. Ablation experimental results on spectral–spatial enhancement attention module (The optimal results are bolded).

Cases	Combination of Branches			Indicators		
	Spe-C Spa-W	Spe-C Spa-H	Spa-W Spa-H	OA (%)	AA (%)	$k * 100$
1	×	✓	✓	97.11	94.43	96.76
2	✓	×	✓	97.03	92.66	96.61
3	✓	✓	×	97.18	92.04	96.78
4	✓	✓	✓	**97.44**	**94.69**	**97.07**

3.4. Classification Results

In this subsection, we compare the proposed MEA-EFFormer network with state-of-the-art classifiers using quantitative and qualitative measures. These classifiers include random forest (RF) [58], support vector machine (SVM) [11], 1D-CNN [59], 2-DCNN [60], 3DCNN [61], HybridSN [28], GAHT [62], SpectralFormer [51], SSFTT [52] and GSC-ViT [63]. Tables 4–6, respectively, provide quantitative results of the compared algorithms on the Indian Pines, Salinas and Pavia University datasets. The parameter settings of the comparison methods were set according to the optimal settings of the reference source texts. To ensure the generality of the experimental results, we conducted ten separate rounds of each experiment and retained the means and variances. From the tables, it can be seen that traditional classification methods have a significant gap compared to deep learning methods. Among deep learning classifiers, methods based on transformers generally achieved better results; this is a benefit of the deep exploration of long-distance relationships between features. It is evident that our MEA-EFFormer method achieves the highest OA, AA and kappa values across all three datasets.

As well-known methods for HSIC in recent years, SSFTT and GSC-ViT effectively integrate spatial–spectral features from HSI and enhance feature discrimination through the integration of transformer networks. Compared with SSFTT on the Indian Pines dataset, MEA-EFFormer demonstrated a 0.47% increase in OA, a 1.26% increase in AA and a 0.53% increase in kappa while reducing bias by 0.11, 1.17, and 0.13, respectively. This comparison highlights the superior performance of MEA-EFFormer for enhancing classification accuracy and stability through the incorporation of multiscale information and an SSEA strategy for exploring discriminative features of land-cover objects.

Table 4. Classification accuracy of Indian Pines dataset using various methods (optimal results highlighted in bold).

No.	Traditional Classifiers		Deep-Learning-Based Classifiers								
	RF	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	GAHT	SpectralFormer	SSFTT	GSC-ViT	MEA-EFFormer
1	1.95 ± 1.83	3.41 ± 2.49	22.09 ± 6.6	30.0 ± 7.39	43.72 ± 11.76	32.79 ± 4.7	62.56 ± 11.32	27.67 ± 6.93	72.56 ± 17.18	40.7 ± 4.91	**87.67 ± 7.79**
2	57.73 ± 18.2	48.14 ± 4.75	90.16 ± 1.16	90.16 ± 2.06	92.19 ± 1.48	89.93 ± 2.52	94.26 ± 1.38	80.06 ± 1.83	94.67 ± 0.69	**95.76 ± 1.12**	95.19 ± 0.54
3	34.24 ± 10.38	37.99 ± 3.2	94.94 ± 2.19	94.65 ± 2.05	96.12 ± 1.51	91.04 ± 2.45	95.73 ± 2.88	91.65 ± 2.61	96.5 ± 1.14	98.64 ± 0.98	**98.81 ± 0.69**
4	4.98 ± 1.5	11.36 ± 1.93	82.41 ± 7.27	82.63 ± 5.2	78.66 ± 4.03	77.5 ± 6.6	90.94 ± 2.82	68.3 ± 4.45	92.9 ± 2.7	**94.91 ± 3.04**	94.73 ± 2.15
5	64.46 ± 5.99	61.2 ± 2.17	98.87 ± 1.45	98.67 ± 1.51	99.54 ± 0.6	98.85 ± 0.7	97.73 ± 3.26	97.45 ± 0.8	99.93 ± 0.1	98.82 ± 0.84	**99.98 ± 0.07**
6	55.76 ± 17.52	61.62 ± 12.91	98.67 ± 0.7	98.99 ± 0.42	99.74 ± 0.17	99.06 ± 0.68	97.75 ± 1.22	97.12 ± 0.56	98.87 ± 0.53	98.29 ± 0.58	**99.22 ± 0.48**
7	4.0 ± 5.06	1.6 ± 1.96	54.44 ± 23.49	59.26 ± 22.28	98.52 ± 1.81	55.19 ± 15.93	61.48 ± 31.91	12.59 ± 7.26	94.44 ± 9.83	88.89 ± 10.61	**97.04 ± 6.79**
8	54.98 ± 21.31	57.81 ± 12.68	99.96 ± 0.09	99.87 ± 0.18	**100.0 ± 0.0**	99.85 ± 0.4	99.4 ± 0.49	98.15 ± 2.3	99.6 ± 0.64	99.98 ± 0.07	99.45 ± 0.51
9	2.22 ± 4.44	1.11 ± 2.22	63.16 ± 12.23	61.58 ± 11.05	72.63 ± 16.94	64.74 ± 16.48	**86.32 ± 9.76**	26.32 ± 7.06	79.47 ± 10.38	80.0 ± 10.99	80.53 ± 4.74
10	50.79 ± 19.74	38.4 ± 5.3	94.45 ± 1.7	93.7 ± 1.77	91.84 ± 2.58	94.21 ± 1.77	92.18 ± 9.13	86.04 ± 1.25	96.46 ± 1.04	97.38 ± 1.12	**97.8 ± 0.52**
11	75.19 ± 14.49	68.72 ± 3.11	95.94 ± 1.88	95.72 ± 0.88	95.02 ± 1.65	95.46 ± 0.88	97.39 ± 0.59	91.89 ± 1.24	**98.76 ± 0.29**	98.06 ± 0.42	98.73 ± 0.32
12	23.11 ± 8.69	19.74 ± 3.6	83.57 ± 6.9	82.2 ± 4.48	81.76 ± 3.29	79.86 ± 4.15	90.53 ± 2.67	66.71 ± 5.55	91.47 ± 1.31	**92.86 ± 1.23**	90.39 ± 1.77
13	18.37 ± 7.58	45.0 ± 15.07	99.38 ± 0.5	99.38 ± 0.79	98.67 ± 0.7	98.56 ± 0.99	88.41 ± 8.34	98.21 ± 1.88	**100.0 ± 0.0**	97.18 ± 1.38	**100.0 ± 0.0**
14	84.84 ± 10.0	75.72 ± 6.73	99.56 ± 0.31	99.7 ± 0.28	98.5 ± 1.53	96.57 ± 1.7	98.8 ± 0.49	98.95 ± 0.65	99.17 ± 0.31	**99.85 ± 0.10**	99.26 ± 0.32
15	13.93 ± 4.63	27.75 ± 4.64	85.48 ± 5.82	83.95 ± 5.92	86.57 ± 2.96	81.61 ± 3.24	91.66 ± 3.52	90.0 ± 3.15	94.69 ± 3.8	90.84 ± 3.31	**95.86 ± 1.65**
16	1.69 ± 1.8	10.12 ± 3.01	92.07 ± 6.31	**92.76 ± 5.95**	88.28 ± 6.68	85.98 ± 11.75	63.91 ± 8.97	56.21 ± 5.33	85.29 ± 4.23	87.13 ± 3.55	80.46 ± 6.93
OA (%)	56.08 ± 8.24	52.67 ± 1.96	93.98 ± 0.84	93.78 ± 0.54	93.91 ± 0.68	92.66 ± 0.73	95.12 ± 1.28	88.57 ± 0.62	96.97 ± 0.28	97.01 ± 0.30	**97.44 ± 0.17**
AA (%)	34.26 ± 4.64	35.61 ± 2.24	84.70 ± 2.45	85.20 ± 2.19	88.86 ± 1.79	83.82 ± 1.77	88.07 ± 2.42	74.21 ± 0.89	93.43 ± 1.74	91.21 ± 1.50	**94.69 ± 0.57**
$k \times 100$	49.48 ± 9.09	46.02 ± 2.26	93.13 ± 0.96	92.90 ± 0.62	93.05 ± 0.77	91.62 ± 0.84	94.42 ± 1.48	86.92 ± 0.70	96.54 ± 0.32	96.59 ± 0.34	**97.07 ± 0.19**

Table 5. Classification accuracy of Salinas dataset using various methods (optimal results highlighted in bold).

No.	Traditional Classifiers		Deep-Learning-Based Classifiers								
	RF	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	GAHT	SpectralFormer	SSFTT	GSC-ViT	MEA-EFFormer
1	23.82 ± 18.27	56.82 ± 28.67	99.71 ± 0.37	99.88 ± 0.17	**99.99 ± 0.02**	99.78 ± 0.3	99.93 ± 0.16	99.38 ± 0.3	99.98 ± 0.03	99.32 ± 0.5	99.98 ± 0.03
2	50.95 ± 42.42	75.92 ± 19.78	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	99.85 ± 0.42	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**
3	14.88 ± 22.89	57.54 ± 5.5	99.99 ± 0.02	**100.0 ± 0.0**	99.99 ± 0.02	99.47 ± 0.45	**100.0 ± 0.0**	99.98 ± 0.05	99.98 ± 0.04	99.83 ± 0.21	**100.0 ± 0.0**
4	4.18 ± 6.32	34.52 ± 13.48	96.28 ± 1.44	94.26 ± 2.68	93.82 ± 2.74	94.86 ± 2.82	**98.48 ± 1.53**	94.58 ± 1.49	98.33 ± 0.99	92.54 ± 4.17	97.62 ± 0.99
5	6.73 ± 8.32	59.23 ± 7.03	**99.95 ± 0.11**	99.86 ± 0.21	99.21 ± 0.64	99.21 ± 0.47	99.65 ± 0.24	98.86 ± 0.47	99.33 ± 1.25	97.36 ± 0.76	99.92 ± 0.14
6	68.04 ± 34.3	74.01 ± 5.14	**100.0 ± 0.0**	99.76 ± 0.7	99.9 ± 0.24	99.94 ± 0.11	97.67 ± 0.71	99.95 ± 0.09	99.85 ± 0.14	99.87 ± 0.16	99.96 ± 0.07
7	91.08 ± 2.45	79.04 ± 13.67	**99.97 ± 0.06**	99.95 ± 0.06	99.92 ± 0.14	99.58 ± 0.37	99.87 ± 0.11	99.91 ± 0.09	99.16 ± 1.01	99.89 ± 0.12	99.1 ± 0.62
8	79.07 ± 16.99	67.52 ± 8.16	88.55 ± 1.04	88.02 ± 1.47	90.42 ± 1.79	86.82 ± 3.63	94.65 ± 0.8	84.85 ± 0.69	93.46 ± 0.94	89.99 ± 1.47	**95.09 ± 0.94**
9	77.78 ± 38.86	64.66 ± 15.31	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	99.83 ± 0.14	**100.0 ± 0.0**	100.0 ± 0.0	99.99 ± 0.02	99.92 ± 0.05	**100.0 ± 0.0**
10	33.34 ± 36.65	46.85 ± 2.77	98.25 ± 0.74	97.74 ± 0.92	98.79 ± 0.39	97.88 ± 1.78	**99.94 ± 0.07**	98.23 ± 0.48	98.73 ± 0.49	98.03 ± 0.46	98.36 ± 0.86
11	18.51 ± 33.65	10.01 ± 11.83	99.38 ± 0.47	99.13 ± 1.21	99.33 ± 0.64	98.43 ± 1.67	97.61 ± 1.17	98.51 ± 0.81	99.74 ± 0.26	97.06 ± 1.91	**99.75 ± 0.19**
12	5.41 ± 8.79	39.61 ± 13.54	99.81 ± 0.25	**99.84 ± 0.21**	99.61 ± 0.37	98.73 ± 1.07	99.66 ± 0.3	99.81 ± 0.32	99.48 ± 0.37	95.61 ± 4.25	99.39 ± 0.37
13	17.27 ± 34.49	18.33 ± 12.4	94.34 ± 10.6	95.63 ± 5.78	94.96 ± 4.1	92.13 ± 6.54	95.15 ± 2.85	**97.14 ± 1.94**	95.42 ± 3.07	74.4 ± 14.45	94.25 ± 2.64
14	15.62 ± 31.09	31.65 ± 13.26	**97.82 ± 2.18**	97.81 ± 2.28	97.06 ± 2.13	96.57 ± 4.11	86.67 ± 1.14	96.07 ± 1.65	95.86 ± 1.9	93.03 ± 3.55	95.39 ± 2.28
15	27.18 ± 19.38	56.11 ± 6.93	87.09 ± 1.01	87.33 ± 1.72	85.67 ± 4.22	86.27 ± 4.79	**92.68 ± 1.16**	85.17 ± 1.46	89.49 ± 1.18	82.6 ± 2.82	91.79 ± 1.15
16	21.14 ± 27.89	49.0 ± 13.96	99.28 ± 0.16	99.16 ± 0.11	99.09 ± 0.14	98.6 ± 1.2	**99.99 ± 0.02**	99.94 ± 0.05	99.49 ± 0.17	96.89 ± 1.03	99.72 ± 0.19
OA (%)	49.29 ± 9.65	59.91 ± 3.89	95.48 ± 0.33	95.32 ± 0.42	95.60 ± 0.42	94.70 ± 0.61	97.26 ± 0.21	94.35 ± 0.19	96.81 ± 0.22	94.19 ± 0.55	**97.42 ± 0.27**
AA (%)	34.69 ± 14.32	51.30 ± 4.10	97.53 ± 0.67	97.40 ± 0.44	97.36 ± 0.41	96.75 ± 0.53	97.62 ± 0.25	97.02 ± 0.17	98.02 ± 0.23	94.77 ± 1.12	**98.15 ± 0.23**
$k \times 100$	41.70 ± 10.89	55.34 ± 4.33	94.97 ± 0.37	94.79 ± 0.46	95.10 ± 0.48	94.10 ± 0.67	96.95 ± 0.24	93.72 ± 0.21	96.44 ± 0.24	93.53 ± 0.62	**97.13 ± 0.30**

Table 6. Classification accuracy of Pavia University using various methods (optimal results highlighted in bold).

No.	Traditional Classifiers		Deep-Learning-Based Classifiers								
	RF	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	GAHT	SpectralFormer	SSFTT	GSC-ViT	MEA-EFFormer
1	63.15 ± 27.63	40.51 ± 12.69	95.42 ± 2.52	96.09 ± 1.17	95.43 ± 1.24	94.54 ± 2.07	96.68 ± 2.36	93.78 ± 0.84	97.05 ± 1.21	89.17 ± 1.62	**97.37 ± 0.69**
2	93.89 ± 9.44	78.54 ± 4.51	99.41 ± 0.47	99.74 ± 0.15	99.52 ± 0.74	99.36 ± 0.47	99.82 ± 0.11	99.77 ± 0.16	99.89 ± 0.05	99.78 ± 0.18	**99.94 ± 0.03**
3	4.08 ± 2.83	21.1 ± 6.26	77.62 ± 6.96	82.52 ± 5.78	73.12 ± 4.41	76.41 ± 4.93	87.29 ± 4.38	72.25 ± 3.19	85.35 ± 6.65	76.03 ± 4.42	**88.67 ± 2.07**
4	19.31 ± 10.7	39.51 ± 4.94	91.32 ± 2.74	90.35 ± 1.98	87.72 ± 3.04	86.44 ± 3.64	82.98 ± 3.72	84.94 ± 2.36	**93.65 ± 1.04**	88.48 ± 1.51	93.33 ± 1.22
5	28.12 ± 26.35	48.56 ± 17.35	99.76 ± 0.33	99.68 ± 0.37	99.85 ± 0.15	99.81 ± 0.26	99.5 ± 0.35	98.64 ± 0.44	99.55 ± 0.42	**99.89 ± 0.15**	99.13 ± 0.48
6	28.18 ± 25.37	31.82 ± 4.73	93.84 ± 3.22	94.54 ± 1.17	97.19 ± 1.71	96.18 ± 1.86	98.84 ± 1.19	91.58 ± 1.54	99.5 ± 0.39	91.13 ± 2.68	**99.98 ± 0.04**
7	25.13 ± 19.44	15.78 ± 9.6	93.15 ± 5.25	93.12 ± 3.74	96.26 ± 3.08	94.31 ± 4.03	99.02 ± 1.15	93.55 ± 2.99	99.86 ± 0.26	88.56 ± 4.77	**99.96 ± 0.09**
8	19.26 ± 11.6	25.26 ± 1.97	85.58 ± 4.57	82.65 ± 8.29	89.18 ± 3.03	86.27 ± 4.78	**95.35 ± 2.63**	81.06 ± 1.04	92.3 ± 2.77	78.65 ± 2.58	92.85 ± 1.1
9	10.75 ± 12.87	27.65 ± 26.7	89.36 ± 6.71	88.77 ± 4.79	92.34 ± 6.66	89.1 ± 4.32	73.62 ± 4.18	71.51 ± 4.91	**94.4 ± 2.52**	92.78 ± 3.33	92.4 ± 3.08
OA (%)	59.18 ± 8.44	52.93 ± 1.59	94.89 ± 0.72	95.12 ± 0.55	95.33 ± 0.50	94.69 ± 0.57	96.40 ± 0.63	93.00 ± 0.33	97.46 ± 0.35	92.82 ± 0.34	**97.72 ± 0.24**
AA (%)	32.43 ± 9.57	36.53 ± 2.81	91.72 ± 1.62	91.94 ± 1.05	92.29 ± 1.34	91.38 ± 1.16	92.57 ± 1.03	87.45 ± 0.66	95.73 ± 0.86	89.38 ± 0.88	**95.96 ± 0.38**
$k \times 100$	42.45 ± 12.58	37.36 ± 2.00	93.19 ± 0.96	93.50 ± 0.73	93.79 ± 0.67	92.93 ± 0.76	95.21 ± 0.84	90.64 ± 0.45	96.62 ± 0.47	90.42 ± 0.46	**96.97 ± 0.32**

For the Salinas dataset, GAHT obtained the second-best classification accuracies in terms of OA, AA and kappa. This is a benefit of the group-aware hierarchical transformer

strategy in the network, which is good at classifying scenarios with objects that are relatively concentrated and uniform. However, MEA-EFFormer still achieves classification results that are similar to or even better than GAHT. Specifically, it improves OA by more than 0.16%, AA by more than 0.53% and kappa by more than 0.18%. This result further demonstrates the effectiveness of the proposed network for HSIC.

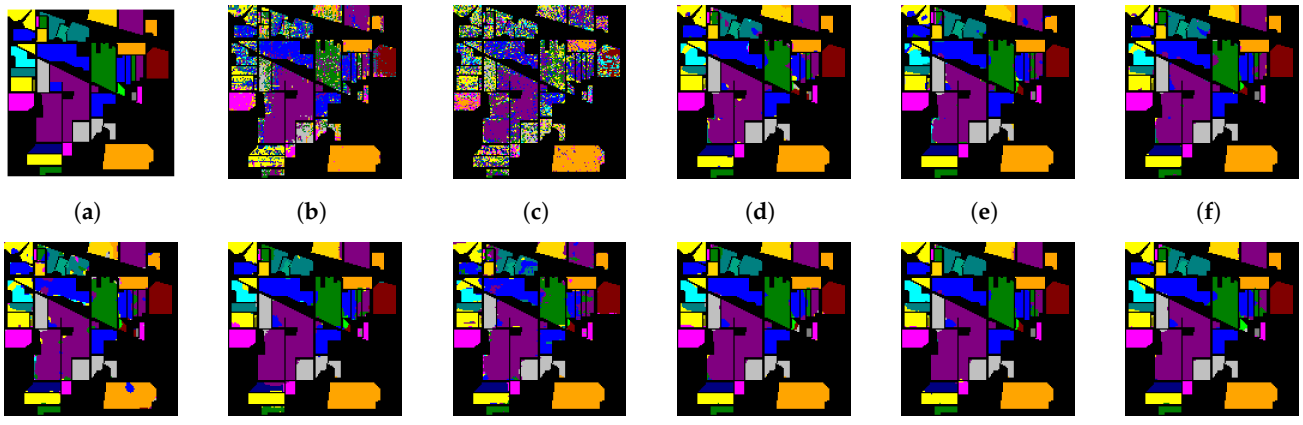
Finally, for the Pavia University dataset, which is known for its high spatial complexity, our method also achieves the best classification results, especially in terms of AA. This shows that MEA-EFFormer can effectively utilize the existing samples for global information even when dealing with scenarios with uneven sample distributions.

In a word, the proposed network has a significant advantage over the state-of-the-art transformer classifiers on these three well-known datasets and achieves the best results in terms of OA, AA and kappa.

3.5. Visual Evaluation

To qualitatively compare the performance of different algorithms, we illustrate the classification maps of different methods on the Indian Pines, Salinas and Pavia University datasets in Figures 12–14, respectively.

It can be clearly seen from the figures that the traditional methods exhibit numerous noisy points in the classification maps across the three datasets, indicating that their classification accuracies are relatively low. This is primarily attributed to the inherent limitations of traditional methods in terms of feature representation and exploration in high-dimensional data, which results in an inability to capture deep-level feature representations effectively. In contrast, the classification maps generated by deep neural network classifiers are generally smoother compared to those produced by traditional methods, aligning with the quantitative results. Particularly, transformer methods stand out for their superior performance: yielding classification maps that deliver satisfactory results both within categories and at their boundaries.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

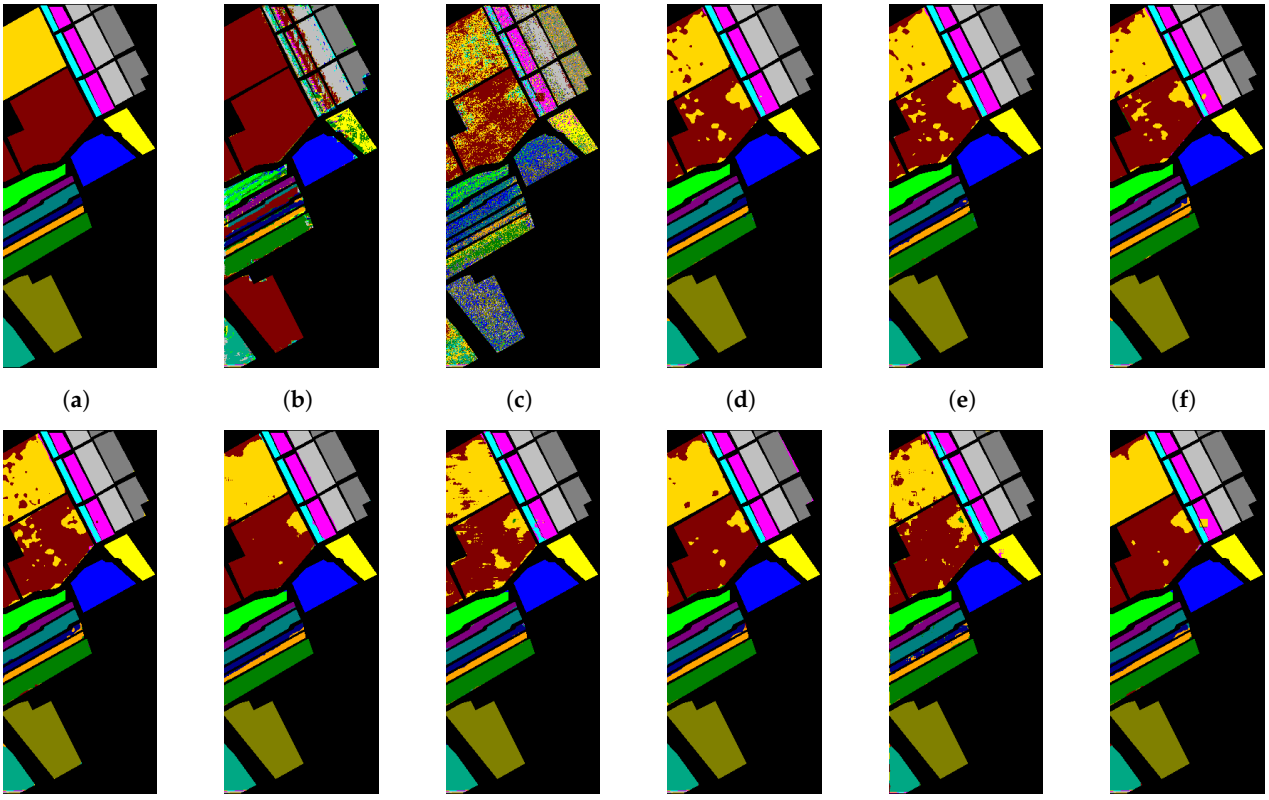
(**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 12. Maps depicting the classifications on Indian Pines dataset using various methods. (**a**) Ground truth. (**b**) RF. (**c**) SVM. (**d**) 1D-CNN. (**e**) 2D-CNN. (**f**) 3D-CNN. (**g**) HybridSN. (**h**) GAHT. (**i**) SpectralFormer. (**j**) SSFTT. (**k**) GSC-ViT. (**l**) MEA-EFFormer.

Notably, on the Indian Pines and Salinas datasets, the classification maps generated by the proposed MEA-EFFormer demonstrate enhanced classification performance, and the classification results of the boundary pixels are all relatively precise. At the same time, the smoothness and consistency of the classification maps are also appealing inside each category. For example, the "orange area" in Figure 13j is more accurate than for the other classification maps.

For the Pavia dataset, due to its high spatial resolution, the distribution of objects is more dispersed and the boundaries are more complex. The classification accuracy of most classifiers is lower, and there are many noise points in the classification map. However, the proposed

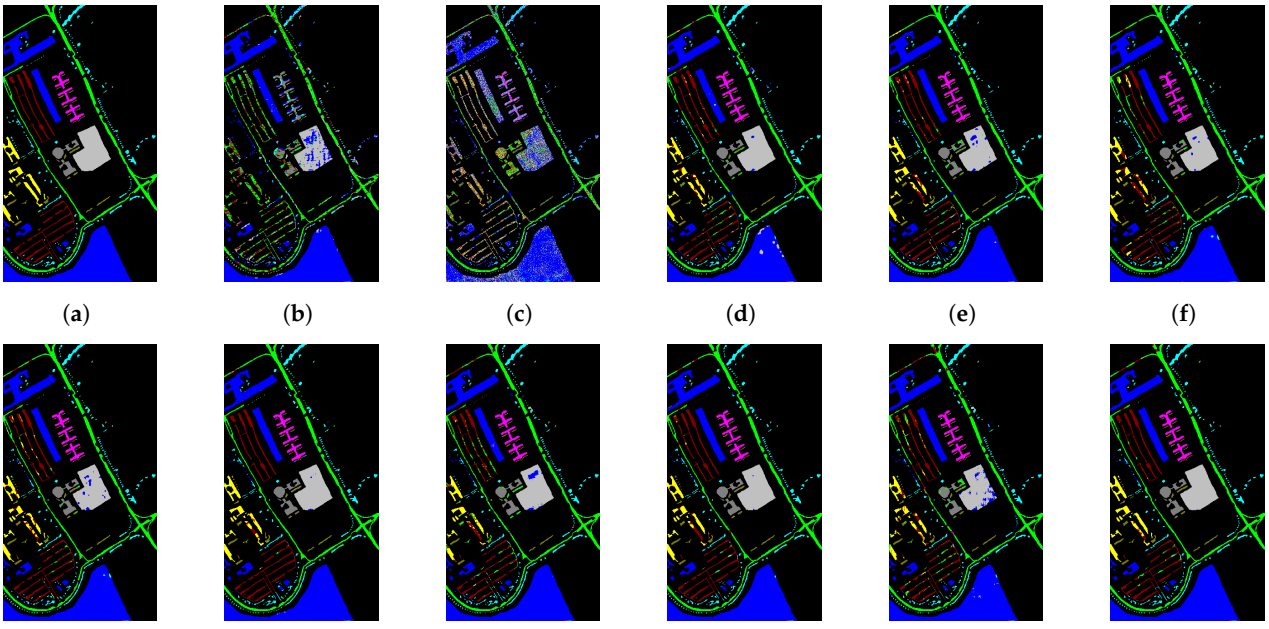
network MEA-EFFormer can still obtain a relatively satisfactory result. For example, the "gray area" in Figure 14l is significantly more accurate than that of other classification maps.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

(**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 13. Maps depicting the classification of Salinas dataset using various methods. (**a**) Ground truth. (**b**) RF. (**c**) SVM. (**d**) 1D-CNN. (**e**) 2D-CNN. (**f**) 3D-CNN. (**g**) HybridSN. (**h**) GAHT. (**i**) SpectralFormer. (**j**) SSFTT. (**k**) GSC-ViT. (**l**) MEA-EFFormer.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

(**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 14. Maps depicting the classification of Pavia University dataset using various methods. (**a**) Ground truth. (**b**) RF. (**c**) SVM. (**d**) 1D-CNN. (**e**) 2D-CNN. (**f**) 3D-CNN. (**g**) HybridSN. (**h**) GAHT. (**i**) SpectralFormer. (**j**) SSFTT. (**k**) GSC-ViT. (**l**) MEA-EFFormer.

3.6. Model Complexity and Efficiency Analysis

We analyzed the computational efficiency of several common deep-learning-based methods on the Pavia University dataset with a sampling rate of 1%. The results are shown in Table 7; our proposed method achieves a moderate advantage in terms of training time and parameter size while achieving the leading classification accuracy.

Table 7. Comparison of trainable parameters, testing times and accuracy of approaches based on deep learning on Pavia University dataset (the optimal results are bolded).

	Deep-Learning-Based Approaches						
	3D-CNN	**HybridSN**	**GAHT**	**SpectralFormer**	**SSFTT**	**GSC-ViT**	**MEA-EFFormer**
Testing Time (s)	**6.39**	7.29	13.69	18.52	7.24	10.85	8.54
Params. (K)	462.486 K	797.57	946.83	128.8	148.3	**77.90**	436.625
OA (%)	95.33 ± 0.50	94.69 ± 0.57	96.40 ± 0.63	93.00 ± 0.33	97.46 ± 0.35	92.82 ± 0.34	**97.72 ± 0.24**

For training time, we ranked second among the several classes of transformer-based methods that we compared. As for SSFTT, the simple fact is that it only performs convolutional extraction of a hybrid on raw HSI data, whereas MEA-EFFormer additionally uses LBP data for spatial information augmentation, which results in a slight increase in time. With the 3D-CNN and HybridSN methods, the runtime is faster since they only use convolutional networks.

For the model parameter sizes, our method is also preferred to most of the methods. The smaller parameter sizes of SpectralFormer and SSFTT are due to the fact that they are too simple, as they only have a single scale in the feature extraction stage, while our method adopts a multiscale strategy to fully exploit the spectral–spatial information in the HSI data. This is an important reason why MEA-EFFormer achieves higher accuracy. For GSC-ViT, the method itself is known for its light weight, and there is an obvious gap with the proposed method in terms of accuracy.

In summary, our proposed MEA-EFFormer can keep the computational efficiency as low as possible with small parameter sizes under the premise of leading accuracy, which again proves the superiority of our method.

4. Conclusions

In this paper, we propose a method called multiscale efficient attention and enhanced feature transformer (MEA-EFFormer) for hyperspectral image classification. We obtain further-refined spectral–spatial information through multiscale efficient attention feature extraction module. Then, we combine two-dimensional convolution features with local binary mode (LBP) spatial information, which effectively improves the representation of features. Then, we use the spectral–spatial enhancement attention module to make the feature enhanced. Finally, we classify these features through transformer encoders. Our experimental results on three HSI datasets show that the proposed method has excellent performance compared with other classification methods. Although our method achieves remarkable results in experiments, we also recognize that there is room for improvement. In the future, we will continue to work to improve our approach. We plan to introduce more efficient feature extraction and attention mechanisms as well as to optimize model structure and parameter settings to achieve better classification performance. At the same time, we also encourage other researchers to explore and innovate in this field and to jointly promote the development of HSIC technology.

Author Contributions: Conceptualization, Q.S. and L.S.; methodology, G.Z. and C.F.; software, Y.F.; validation, G.Z., Y.F. and X.L.; formal analysis, L.S.; investigation, Y.F.; data curation, G.Z. and C.F.; writing—original draft preparation, Q.S.; writing—review and editing, L.S. and X.L.; visualization, G.Z. and Y.F.; supervision, L.S. and X.L.; project administration, X.L.; funding acquisition, Q.S. and X.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, in part by the Startup Foundation for Introducing Talent of NUIST, grant number 2022r075, also in part by Guangxi Forestry Technology Promotion and Demonstration Application Project, grant number 2024GXLK19.

Data Availability Statement: Data are contained within 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.

References

1. Liu, S.; Liu, S.; Zhang, S.; Li, B.; Hu, W.; Zhang, Y.D. SSAU-Net: A Spectral–Spatial Attention-Based U-Net for Hyperspectral Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5542116. [CrossRef]
2. Sun, G.; Pan, Z.; Zhang, A.; Jia, X.; Ren, J.; Fu, H.; Yan, K. Large Kernel Spectral and Spatial Attention Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5519915. [CrossRef]
3. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. [CrossRef]
4. 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]
5. Sun, L.; Wang, X.; Zheng, Y.; Wu, Z.; Fu, L. Multiscale 3-D–2-D Mixed CNN and Lightweight Attention-Free Transformer for Hyperspectral and LiDAR Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 2100116. [CrossRef]
6. Sun, L.; Zhang, H.; Zheng, Y.; Wu, Z.; Ye, Z.; Zhao, H. MASSFormer: Memory-Augmented Spectral-Spatial Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
7. Yao, J.; Hong, D.; Xu, L.; Meng, D.; Chanussot, J.; Xu, Z. Sparsity-Enhanced Convolutional Decomposition: A Novel Tensor-Based Paradigm for Blind Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5505014. [CrossRef]
8. Dalponte, M.; Ørka, H.O.; Gobakken, T.; Gianelle, D.; Næsset, E. Tree Species Classification in Boreal Forests with Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2632–2645. [CrossRef]
9. Zhang, J.; Tao, D. Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet Things J.* **2021**, *8*, 7789–7817. [CrossRef]
10. Hang, R.; Liu, Q.; Li, Z. Spectral Super-Resolution Network Guided by Intrinsic Properties of Hyperspectral Imagery. *IEEE Trans. Image Process.* **2021**, *30*, 7256–7265. [CrossRef]
11. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
12. Fauvel, M.; Chanussot, J.; Benediktsson, J. Evaluation of Kernels for Multiclass Classification of Hyperspectral Remote Sensing Data. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 2, p. II. [CrossRef]
13. Tu, B.; Wang, J.; Kang, X.; Zhang, G.; Ou, X.; Guo, L. KNN-Based Representation of Superpixels for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4032–4047. [CrossRef]
14. Samaniego, L.; Bardossy, A.; Schulz, K. Supervised Classification of Remotely Sensed Imagery Using a Modified *k*-NN Technique. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2112–2125. [CrossRef]
15. Zhang, Y.; Cao, G.; Li, X.; Wang, B. Cascaded Random Forest for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1082–1094. [CrossRef]
16. 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]
17. Lobo, A. Image segmentation and discriminant analysis for the identification of land cover units in ecology. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 1136–1145. [CrossRef]
18. 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]
19. Liu, C.; Li, J.; He, L.; Plaza, A.; Li, S.; Li, B. Naive Gabor Networks for Hyperspectral Image Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 376–390. [CrossRef]
20. Jiang, C.; Su, J. Gabor Binary Layer in Convolutional Neural Networks. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3408–3412. [CrossRef]
21. Li, H.; Ye, Z.; Xiao, G. Hyperspectral Image Classification Using Spectral–Spatial Composite Kernels Discriminant Analysis. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2341–2350. [CrossRef]
22. Li, J.; Marpu, P.R.; Plaza, A.; Bioucas-Dias, J.M.; Benediktsson, J.A. Generalized Composite Kernel Framework for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4816–4829. [CrossRef]
23. Tang, Y.Y.; Lu, Y.; Yuan, H. Hyperspectral Image Classification Based on Three-Dimensional Scattering Wavelet Transform. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2467–2480. [CrossRef]

24. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
25. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised Deep Feature Extraction for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1349–1362. [CrossRef]
26. Cao, X.; Yao, J.; Xu, Z.; Meng, D. Hyperspectral Image Classification With Convolutional Neural Network and Active Learning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4604–4616. [CrossRef]
27. Yang, X.; Ye, Y.; Li, X.; Lau, R.Y.K.; Zhang, X.; Huang, X. Hyperspectral Image Classification With Deep Learning Models. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5408–5423. [CrossRef]
28. 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.* **2020**, *17*, 277–281. [CrossRef]
29. Zhu, J.; Fang, L.; Ghamisi, P. Deformable Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1254–1258. [CrossRef]
30. Jia, S.; Lin, Z.; Xu, M.; Huang, Q.; Zhou, J.; Jia, X.; Li, Q. A Lightweight Convolutional Neural Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4150–4163. [CrossRef]
31. Czaja, W.; Ehler, M. Schroedinger Eigenmaps for the Analysis of Biomedical Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1274–1280. [CrossRef]
32. Gong, Z.; Zhong, P.; Yu, Y.; Hu, W.; Li, S. A CNN With Multiscale Convolution and Diversified Metric for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3599–3618. [CrossRef]
33. Xie, P.; Salakhutdinov, R.; Mou, L.; Xing, E.P. Deep Determinantal Point Process for Large-Scale Multi-label Classification **2017**. pp. 473–482. [CrossRef]
34. Hang, R.; Liu, Q.; Hong, D.; Ghamisi, P. Cascaded Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5384–5394. [CrossRef]
35. Zhang, X.; Sun, Y.; Jiang, K.; Li, C.; Jiao, L.; Zhou, H. Spatial Sequential Recurrent Neural Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4141–4155. [CrossRef]
36. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063. [CrossRef]
37. Jia, S.; Wang, Z.; Li, Q.; Jia, X.; Xu, M. Multiattention Generative Adversarial Network for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624715. [CrossRef]
38. Neagoe, V.E.; Diaconescu, P. CNN Hyperspectral Image Classification Using Training Sample Augmentation with Generative Adversarial Networks. In Proceedings of the 2020 13th International Conference on Communications (COMM), Bucharest, Romania, 8–20 June 2020; pp. 515–519. [CrossRef]
39. He, K.; Zhang, X.; 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, 26 June–1 July 2016; pp. 770–778. [CrossRef]
40. Zhong, Z.; Li, J.; Ma, L.; Jiang, H.; Zhao, H. Deep residual networks for hyperspectral image classification. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1824–1827. [CrossRef]
41. Li, T.; Zhang, X.; Zhang, S.; Wang, L. Self-Supervised Learning With a Dual-Branch ResNet for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5512905. [CrossRef]
42. Zhang, X.; Liang, Y.; Li, C.; Huyan, N.; Jiao, L.; Zhou, H. Recursive Autoencoders-Based Unsupervised Feature Learning for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1928–1932. [CrossRef]
43. Koda, S.; Melgani, F.; Nishii, R. Unsupervised Spectral–Spatial Feature Extraction With Generalized Autoencoder for Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 469–473. [CrossRef]
44. Chen, Y.; Zhao, X.; Jia, X. Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2381–2392. [CrossRef]
45. Chen, C.; Ma, Y.; Ren, G. Hyperspectral Classification Using Deep Belief Networks Based on Conjugate Gradient Update and Pixel-Centric Spectral Block Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4060–4069. [CrossRef]
46. Sun, H.; Zheng, X.; Lu, X. A Supervised Segmentation Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2810–2825. [CrossRef]
47. Zheng, Z.; Zhong, Y.; Ma, A.; Zhang, L. FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5612–5626. [CrossRef]
48. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J.; Pla, F. Capsule Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2145–2160. [CrossRef]
49. Jiang, X.; Liu, W.; Zhang, Y.; Liu, J.; Li, S.; Lin, J. Spectral–Spatial Hyperspectral Image Classification Using Dual-Channel Capsule Networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1094–1098. [CrossRef]
50. Liang, L.; Zhang, Y.; Zhang, S.; Li, J.; Plaza, A.; Kang, X. Fast Hyperspectral Image Classification Combining Transformers and SimAM-Based CNNs. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5522219. [CrossRef]
51. 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.* **2022**, *60*, 5518615. [CrossRef]

52. 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]
53. Tu, B.; Liao, X.; Li, Q.; Peng, Y.; Plaza, A. Local Semantic Feature Aggregation-Based Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536115. [CrossRef]
54. Qiao, X.; Huang, W. A Dual Frequency Transformer Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 10344–10358. [CrossRef]
55. Wang, S.; Liu, Z.; Chen, Y.; Hou, C.; Liu, A.; Zhang, Z. Expansion Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6411–6427. [CrossRef]
56. Zhou, F.; Xu, C.; Yang, G.; Hang, R.; Liu, Q. Masked Spectral–Spatial Feature Prediction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4400913. [CrossRef]
57. Huang, L.; Chen, Y.; He, X. Spectral–Spatial Masked Transformer with Supervised and Contrastive Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5508718. [CrossRef]
58. Ham, J.; Chen, Y.; Crawford, M.M.; Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [CrossRef]
59. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4959–4962. [CrossRef]
60. 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]
61. 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.
62. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
63. Zhao, Z.; Xu, X.; Li, S.; Plaza, A. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5511817. [CrossRef]