electronics

MDPI

Article

Hyperspectral Image Classification Using Multi-Scale Lightweight Transformer

Quan Gu *, Hongkang Luan, Kaixuan Huang and Yubao Sun

Engineering Research Center of Digital Forensics, Ministry of Education, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211221026@nuist.edu.cn (H.L.); 20211249413@nuist.edu.cn (K.H.); sunyb@nuist.edu.cn (Y.S.)

* Correspondence: 20211249321@nuist.edu.cn

Abstract: The distinctive feature of hyperspectral images (HSIs) is their large number of spectral bands, which allows us to identify categories of ground objects by capturing discrepancies in spectral information. Convolutional neural networks (CNN) with attention modules effectively improve the classification accuracy of HSI. However, CNNs are not successful in capturing long-range spectral–spatial dependence. In recent years, Vision Transformer (VIT) has received widespread attention due to its excellent performance in acquiring long-range features. However, it requires calculating the pairwise correlation between token embeddings and has the complexity of the square of the number of tokens, which leads to an increase in the computational complexity of the network. In order to cope with this issue, this paper proposes a multi-scale spectral–spatial attention network with frequency-domain lightweight Transformer (MSA-LWFormer) for HSI classification. This method synergistically integrates CNN, attention mechanisms, and Transformer into the spectral–spatial feature extraction module and frequency-domain fused classification module. Specifically, the spectral–spatial feature extraction module employs a multi-scale 2D-CNN with multi-scale spectral attention (MS-SA) to extract the shallow spectral–spatial features and capture the long-range spectral dependence. In addition, The frequency-domain fused classification module designs a frequency-domain lightweight Transformer that employs the Fast Fourier Transform (FFT) to convert features from the spatial domain to the frequency domain, effectively extracting global information and significantly reducing the time complexity of the network. Experiments on three classic hyperspectral datasets show that MSA-LWFormer has excellent performance.

Keywords: hyperspectral image classification; multi-scale spectral attention; Transformer; long-range spectral dependence

Citation: Gu, Q.; Luan, H.; Huang, K.; Sun, Y. Hyperspectral Image Classification Using Multi-Scale Lightweight Transformer. *Electronics* **2024**, *13*, 949. https://doi.org/10.3390/electronics13050949

Academic Editor: Gemma Piella

Received: 28 January 2024
Revised: 24 February 2024
Accepted: 27 February 2024
Published: 29 February 2024

1. Introduction

The accelerated progress in satellite remote sensing technology has made HSI a compelling area of research [1]. Unlike traditional natural images, HSIs continuously record spectral responses in numerous narrow and uninterrupted bands, thereby achieving the ability to accurately identify and differentiate materials with small spectral changes [2]. HSI classification achieves the allocation of pixels within HSI data to distinct land cover types, facilitating the precise categorization and identification of surface features like farmland, forest, and water. This classification technique has extensive application in diverse fields, including agriculture [3], environmental monitoring [4], urban planning [5], and geological exploration [6].

HSI classification uses a variety of methods in the field of traditional machine learning. Some commonly used methods include random forest [7], the minimum distance classifier [8], support vector machine (SVM) [9], the K-nearest neighbor algorithm (KNN) [10], and the Bayesian classifier [11]. In addition, because of the high-dimensional nature of HSI, a variety of dimensionality reduction methods have also been widely used, including

principal component analysis (PCA) [12], isometric mapping (Isomap) [13], and local linear embedding (LLE) [14]. However, traditional methods often rely on manually extracting spatial and spectral features from HSI and struggle to capture complex non-linear relationships and high-order correlations in the data. This limitation results in a reduction in performance when dealing with complex datasets.

With the advancement of deep learning methods, its widespread application in various computer vision (CV) tasks has become increasingly evident in recent years [15,16]. These tasks include, but are not limited to, denoising [17,18], target detection [19,20], change detection [21,22], and classification [23–25]. Numerous research findings consistently demonstrate the superior performance of deep learning methods over traditional approaches in extracting high-level features. These features extracted by deep networks exhibit enhanced proficiency in capturing intricate and abstract information, contributing to a substantial improvement in the accuracy of HSI classification. Consequently, a range of deep learning methods have been proposed, encompassing recurrent neural networks (RNNs) [26], deep belief neural networks (DBNs) [27], stacked autoencoders (SAEs) [28], and more. However, these HSI classification methods often overlook spatial information, focusing solely on spectral details. This oversight may lead to challenges such as differentiating between spectral of the same substance and attributing the same spectrum to distinct substances.

To tackle the mentioned problems, various deep learning methods utilizing CNNs have been suggested to effectively extract spectral–spatial features in HSI classification. Yu et al. introduced a dedicated one-dimensional CNN (1D-CNN) to capture spectral correlations between spectral bands [29]. Gao et al. proposed a two-dimensional CNN (2D-CNN) to efficiently capture the spatial structure and texture information in images, thereby enhancing spatial feature extraction [30]. Meanwhile, Xu et al. introduced a three-dimensional CNN (3D-CNN) to capture both spectral and spatial characteristics in HSI, as well as their interactions with each other [31]. Roy et al. suggested the HybridSN network, a combination of a 2D-CNN and 3D-CNN, for the simultaneous extraction of spatial and spectral features [32]. Given the varied sizes of targets in HSI, researchers commonly explore feature extraction across multiple scales. He et al. introduced a multi-scale 3D-CNN capable of extracting spectral–spatial information from images at four different scales [33]. Hu et al. put forward a hybrid convolutional network that combined multi-scale 2D and 3D depthwise separable convolution [34]. Recognizing the great importance of information at different scales for specific tasks, many researchers have incorporated attention mechanisms into deep learning models. Mou et al. integrated an attention module into the spectrum, utilizing a gating mechanism to selectively emphasize frequency bands rich in information and adaptively recalibrate the spectral bands [35]. Cui et al. devised a network that combined multi-scale feature aggregation with a dual-channel spectral–spatial attention mechanism, aiming to adeptly capture local contextual information [36]. To enhance the extraction of long-range features, researchers have been progressively utilizing Transformers in the classification of HSI. Hong et al. introduced the SpectralFormer model, which incorporates a Transformer module that merges contextual information from neighboring frequency bands, capturing both local and spectral sequence information [37]. Sun proposed a method named the Spectral–Spatial Feature Tokenization Transformer (SSFTT), designed to capture both high-level semantic features and spectral–spatial features [38].

While the deep learning methods mentioned above have found extensive application in HSI classification, there are still some key obstacles, which can be summarized as follows.

(1) These approaches fail to effectively leverage the multi-scale features presenting in HSI and neglect to establish strong dependencies among spectral bands. Consequently, their capacity to distinguish long-range spectral disparities within HSI is limited.

(2) Existing Transformer-based HSI classification methods capture the contextual relationships among all input embeddings via the multi-head self-attention (MHSA) mechanism. However, it requires the correlation calculation of the square scale of the number of tokens, which results in an increase in computational complexity within the network.

In order to tackle the above difficulties, we propose a multi-scale spectral–spatial attention network with a frequency-domain lightweight Transformer. Specifically, we use a spectral–spatial feature extraction module to effectively extract the spectral–spatial features and capture the long-range spectral dependence of HSI. In addition, the frequency-domain lightweight Transformer applies the FFT to convert features from the spatial domain to the frequency domain, effectively extracting global information and significantly reducing the time complexity of the network. Our main contributions are as follows.

(1) MSA-LWFormer proposes a spectral–spatial feature extraction module aimed at extracting shallow spectral–spatial features and capturing long-range spectral dependencies. This module emphasizes cross-channel and multi-scale features by integrating the multi-scale 2D-CNN and MS-SA techniques. These designs enhance the model's ability to accurately capture and interpret complex spectral information.

(2) Applying the FFT to the query, key, and value matrices within a frequency-domain lightweight Transformer reduces the time complexity of the network. This transformation process serves to convert features from the spatial domain to the frequency domain, thereby enhancing the extraction of comprehensive global information and reducing the time complexity of the network.

(3) Our proposed network demonstrates good classification results across three classic HSI datasets, providing compelling evidence for the effectiveness of our approach.

The succeeding sections of this manuscript are organized as follows. Section 2 presents a comprehensive examination of the overall structure of the MSA-LWFormer and the design specifics of each sub-module. Section 3 presents the details of the experiments conducted and the corresponding experimental results. Section 4 is devoted to the discussion of the ablation experiments, time complexity analysis, and hyperparameter analysis of MSA-LWFormer. Lastly, Section 5 summarizes the study's conclusions and outlines prospects for future work.

2. Materials and Methods

Figure 1 depicts the overall framework of the MSA-LWFormer network. It consists of two parts: a spectral–spatial feature extraction module and a frequency-domain fused classification module.

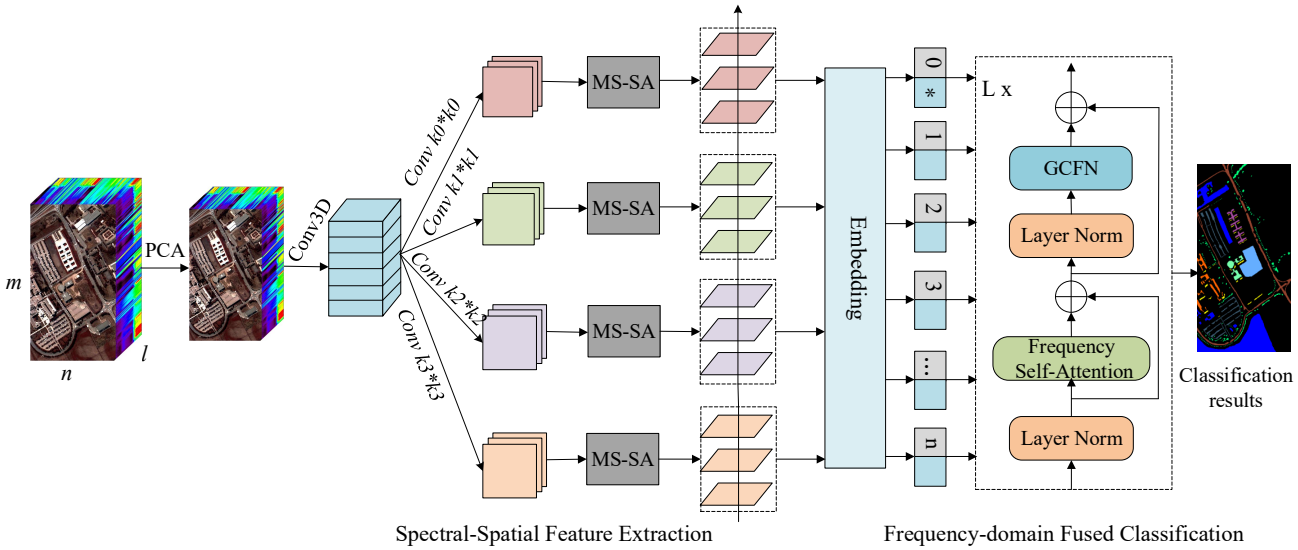


Figure 1. The overall architecture of the MSA-LWFormer network.

2.1. Spectral–Spatial Feature Extraction Module

The initial HSI is denoted as $I \in \mathbb{R}^{m \times n \times l}$, where m and n represent their height and width, respectively, and l represents the number of spectral bands. Despite the valuable spectral information carried by these l bands, they concurrently yield high-dimensional

data, leading to a substantial computational burden. Consequently, we utilize the PCA technique to reduce the spectral dimension, being beneficial in mitigating the computational complexity. This approach preserves the crucial elements of HSI while removing redundant and ineffective spectral bands. While preserving the spatial dimensions, the number of bands in HSI data is reduced from l to b, resulting in a transformed dataset denoted as $I_{pca} \in \mathbb{R}^{m \times n \times b}$.

Subsequently, a 3D convolution layer is used to convolve the HSI data across the entire spectral domain, aiming to extract multi-channel features. Within the 3D convolution layer, the output value v_{ij}^{xyz} of the jth feature cube in the ith layer at the (x, y, z) position is calculated as follows:

$$v_{i,j}^{x,y,z} = \Phi\left(\sum_{s=1}^{S} z \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} \sum_{d=0}^{D_i-1} \omega_{i,j,m}^{h,w,d} \cdot v_{(i-1),m}^{(x+h),(y+w),(z+d)} + b_{i,j}\right) \tag{1}$$

where Φ denotes the activation function, S is the total number of spectral bands, and s represents the sth channel among S channels. The parameters H_i, W_i, and D_i denote the dimensions of the 3D convolution kernel, specifically referring to its width, height, and band number, respectively. The weight parameter $\omega_{i,j,m}^{h,w,d}$ is linked to the position (h, w, d) and is intricately connected to the sth feature cube, while $b_{i,j}$ signifies the bias term.

As depicted in Figure 2, we propose an efficient MS-SA aimed at capturing long-range spectral dependencies. Given the significant spatial similarities and subtle spectral differences between different spectral groups, we employ multi-scale 2D convolution layers to extract spatial information from the input feature map. The input HSI composed of b spectral bands, is uniformly partitioned into four groups, effectively reducing the spectral dimension of the input tensor. This division facilitates the extraction of spectral features with spatial information at different scales. Each set of feature maps, denoted as $F_i \in \mathbb{R}^{m \times n \times c}$, shares a common channel dimension $c = \frac{b}{4}$, where $i = 1, \ldots, 4$. However, the augmentation of the convolution kernel size leads to a significant increase in the number of parameters. To address this challenge and efficiently process feature maps across different convolution kernel scales, we employ a group convolution method. Notably, this method is implemented without an increase in parameters, and the choice of an appropriate group size depends on the kernel size of each scale. The relationship between the multi-scale 2D convolution kernel size and the group size can be expressed as

$$n = 2^{\frac{1}{2}(m-1)} \tag{2}$$

where m represents the kernel size and n signifies the group size. F_i can be expressed as follows:

$$F_i = 2DConv(m_i \times m_i, n_i)(X) \quad i = 1, 2, 3, 4 \tag{3}$$

where the ith kernel size is defined as $m_i = 2(i+1)+1$, the ith group size is calculated as $n_i = 2^{\frac{1}{2}(m_i-1)}$, and $F_i \in \mathbb{R}^{m \times n \times c}$ represents feature maps corresponding to different scales. Moreover, to enhance the utilization of spectral similarities, an efficient spectral attention mechanism is employed to distinguish spectral features across various scales. We adopt the ECAWeight module to obtain spectral attention weight vectors across different scales from the derived multi-scale feature maps. As depicted in Figure 3, the ECAWeight module initiates the aggregation of global spatial features for each channel using Global Average Pooling (GAP). The calculation formula for GAP can be articulated as follows:

$$G_i = \frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} F_i(x, y) \quad i = 1, 2, 3, 4 \tag{4}$$

where $F_i(x, y)$ represents the values across all channels of F_i at the (x, y) position. The attention weight across all channels can be expressed as

$$Y_i = \sigma(C1D(G_i, k)) \quad i = 1, 2, 3, 4 \tag{5}$$

Here, the term $C1D$ denotes 1D convolution. The symbol σ signifies the Sigmoid activation function. The variable k denotes the number of neighbors surrounding a specific channel in G_i, and it can be determined by the channel dimension c. Consequently, for a given channel dimension c, k can be expressed as

$$k = \left| \frac{log_2(c)}{\gamma} + \frac{o}{\gamma} \right|_{odd} \tag{6}$$

The notation $|t|_{odd}$ represents the nearest odd number to t. In this study, we set γ to 2 and o to 1 in the experiments.

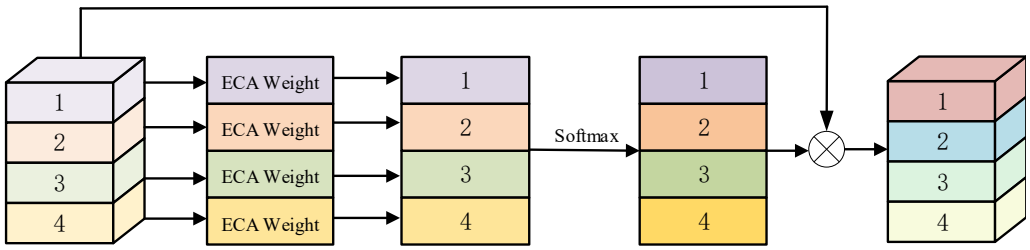


Figure 2. The architecture of the multi-scale spectral attention. Multi-scale spectral attention employs an effective channel attention mechanism and softmax operation to obtain feature maps at four scales at the spectral level, aiming to capture long-range spectral dependencies.

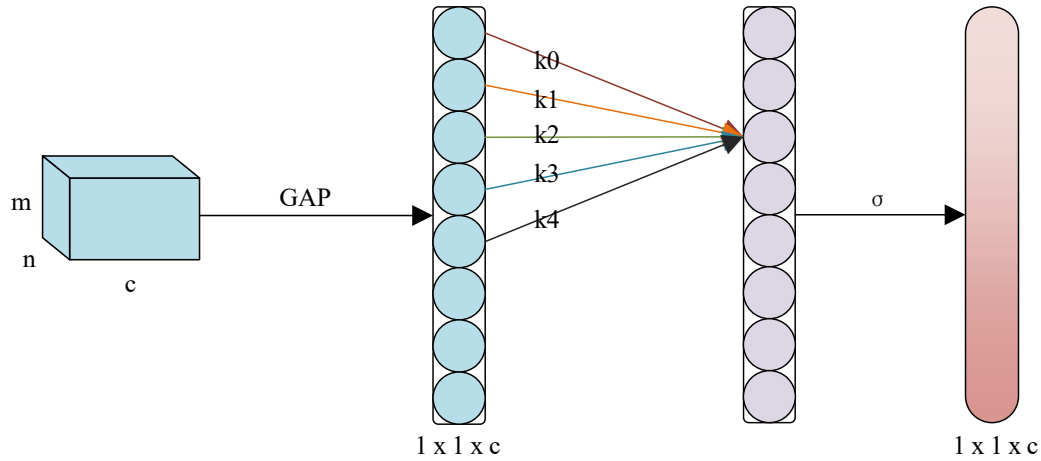


Figure 3. The architecture of the ECA weight. k_0 to k_4 denote the neighbors surrounding a certain channel in G_i, and its adaptive determination is contingent on the mapping of the channel dimension c, which is calculated according to Formula (7).

Therefore, the multi-scale 2D convolution layers effectively incorporate contextual information from each scale, thereby enhancing the pixel-level attention within the feature map. The introduction of long-range spectral attention is seamlessly achieved without altering the original spectral attention vector. This process culminates in the establishment of a comprehensive MS-SA vector. The attention weights for the multi-scale spectrum, derived from the softmax operation, are subsequently applied by element-wise multiplication with the feature map corresponding to the scale F_i, as formally articulated below:

$$X_i = F_i \odot softmax(Y_i) \quad i = 1, 2, 3, 4 \tag{7}$$

where Y_i signifies the attention value derived from F_i. The symbol $\odot$ denotes element-wise multiplication, and X_i denotes the resulting MS-SA feature map.

By employing the spectral–spatial feature extraction module, we investigate the reliance on extended spectral ranges in capturing spectral–spatial information across various scales. However, there are inherent limitations in comprehensively representing global features. Consequently, we introduce a frequency-domain lightweight Transformer. Its goal is to improve the extraction of global features, leading to a substantial enhancement in the representation of spectral–spatial information.

2.2. Frequency-Domain Fused Classification

The research conducted by [39] redefined the challenge of spatial feature extraction in high-resolution images by framing it as the extraction of spatial frequency-domain sequences within the domain of natural images. This innovative approach resulted in a noteworthy reduction in computational complexity. In the context of high-dimensional and large-scale HSI, the extraction of frequency-domain features emerges as a more effective strategy in enhancing the representation of global information. Through the utilization of the FFT, spatial-domain features are transformed into their frequency-domain counterparts, facilitating the efficient integration of global information during the self-attention calculation stage. This advancement not only improves the extraction of global features but also, in contrast to the original MHSA, significantly reduces the time complexity of the frequency-domain self-attention from $O(N^2)$ to $O(N \log^N)$.

As illustrated in Figure 1, the feature map obtained through the preceding MS-SA is partitioned based on the channel dimension. Each channel is treated as a patch, with each patch $p_1 \in \mathbb{R}^{w \times h \times 1}$ being flattened into a token $t_1 \in \mathbb{R}^{wh \times 1}$, where w represents the width, and h represents the height. Consequently, the feature map produced by MS-SA is flattened into a token sequence $T \in \mathbb{R}^{n \times c}$, where c is the number of tokens. To facilitate classification, we introduce an additional learnable classification token t_{class} (denoted as the number zero), positioned at the sequence's outset. To preserve positional information, a position embedding $T_{pe} \in \mathbb{R}^{n \times (c+1)}$ is introduced. The input of T_{in} this frequency-domain lightweight Transformer is formulated as

$$T_{in} = [t_{class}; T] + T_{pe} \tag{8}$$

Following this, the resulting token is fed into the frequency-domain lightweight Transformer. As illustrated in Figure 4a, this module contains frequency self-attention, a gated-conv feed-forward network (GCFN) layer, and layer normalization. Frequency self-attention employs three weight matrices (W_q, W_k, and W_v) for the input T_{in}. Through a linear transformation, each token is systematically mapped to three respective learnable parameter matrices, denoted as the query (F_q), key (F_k), and value (F_v). This process can be expressed as

$$\begin{aligned} F_q &= W_q T_{in} \\ F_k &= W_k T_{in} \\ F_v &= W_v T_{in} \end{aligned} \tag{9}$$

Next, apply the FFT to the features F_q and F_k to compute their correlation in the frequency domain. This process can be expressed as

$$A = FFT^{-1}\left(FFT(F_q) \odot \overline{FFT(F_k)}\right) \tag{10}$$

where $FFT^{-1}(\cdot)$ represents the inverse FFT, and $\overline{FFT(\cdot)}$ represents the conjugate of FFT. The notation $\odot$ represents element-wise multiplication. Finally, the feature representation, acquired through the utilization of frequency-domain self-attention, is articulated as follows:

$$Y_{att} = LN(softmax(A) \odot F_v) \tag{11}$$

The term LN denotes the linear layer within the network architecture. Figure 4b visually illustrates the specific structure of the frequency self-attention mechanism.

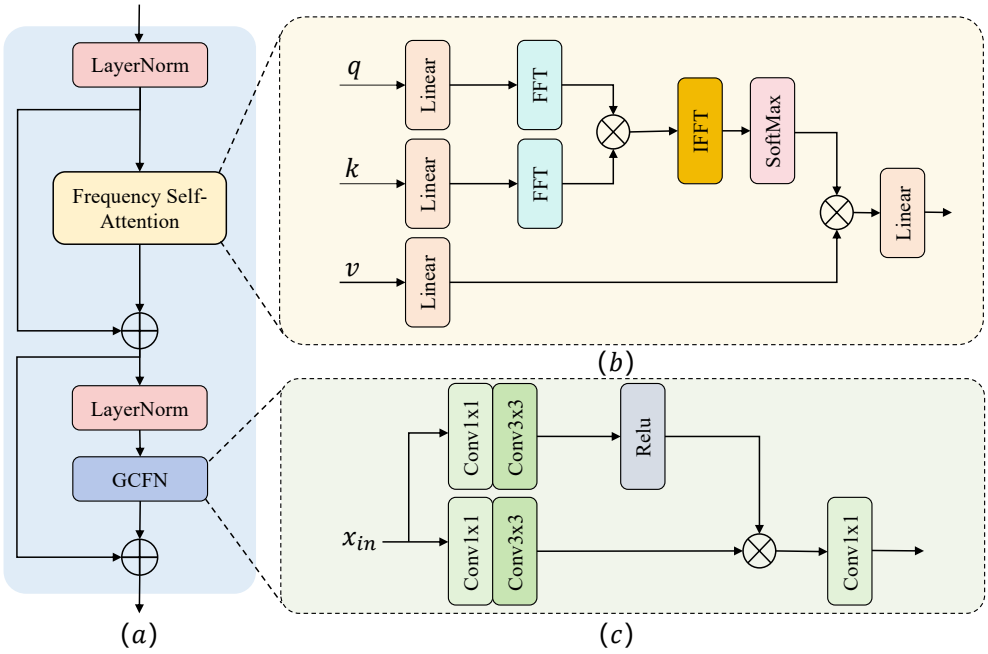


Figure 4. (**a**) The architecture of the frequency-domain lightweight Transformer. (**b**) Frequency self-attention. (**c**) Gated-conv feed-forward network.

In Figure 4c, the specific architecture of the GCFN layer is portrayed. GCFN conducts non-linear transformations and feature extraction on the attention values received from frequency self-attention, aiming to convert them into a more expressive feature representation. The GCFN layer employs a gate mechanism to obtain an element-wise dot product of the attention values from two parallel channels. Both channels utilize a 3×3 convolution to capture spatially adjacent pixel information, effectively extracting local structural details. The application of ReLU to one channel imparts non-linear characteristics to the GCFN layer. The expression for the GCFN layer's output is stated as follows:

$$GCFN(X) = Conv_{1\times1}(Conv_{3\times3}(Conv_{1\times1}(X))) \odot \Phi_R(Conv_{3\times3}(Conv_{1\times1}(X)))) \quad (12)$$

where Φ_R represents the ReLU non-linear activation function, and $Conv_{3\times3}$ and $Conv_{1\times1}$ represent the 3×3 convolution and 1×1 convolution, respectively. The details of the architecture of the frequency-domain lightweight Transformer are described in Table 1.

Table 1. The details of the architecture of the frequency-domain lightweight Transformer.

LayerName	Operations	Paramaters
Norm	Norm	dim = 64
Frequency Self-Attention	Linear	dim = 64, head = 8
	FFT	s = 64, dim = (−2, −1), norm = backward
	IFFT	s = 64, dim = (−2, −1), norm = backward
	softmax	dim = −1
	Linear	dim = 64
Norm	Norm	dim = 64
GCFN	2dConv	kernel size = 1, stride = 1, padding = 1
	2dConv	kernel size = 3, stride = 1, padding = 1
	2dConv	kernel size = 1, stride = 1, padding = 1

The frequency-domain lightweight Transformer ensures that the input size T_{in} matches the output size T_{out}. To perform classification, we have generated a dedicated learning vector for classification, denoted as t_{class}, which is input into the linear layer to obtain the final classification result. Using the softmax function, this linear layer computes the probabilities associated with specific categories based on the input. The assigned category for a sample corresponds to the label with the highest probability value.

2.3. Network Parameters Learning

MSA-LWFormer employs a multi-label cross-entropy loss function to constrain the learning of network parameters. This specific loss function calculates the cross-entropy loss independently for each category and sample and subsequently computes the average loss across all samples. It enables the assignment of multiple categories to each sample, considering both positive and negative categories. If we consider N samples, each categorized into K classes, the expression for the multi-label cross-entropy loss function is outlined as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\left(y_{i,j}\cdot\log(p_{i,j}) + (1-y_{i,j})\cdot\log(1-p_{i,j})\right) \tag{13}$$

where N denotes the number of samples, K represents the number of categories, $y_{i,j}$ signifies the actual label of sample i in the category j, and $p_{i,j}$ denotes the predicted probability by the model for the same sample and category. The terms $y_{i,j}\cdot\log(p_{i,j})$ and $(1-y_{i,j})\cdot\log(1-p_{i,j})$ correspond to the loss associated with treating the actual label as a positive and a negative category, respectively.

The comprehensive procedure of the MSA-LWFormer network is illustrated in Algorithm 1.

Algorithm 1 MSA-LWFormer Overall Network

Input: HSI data $I \in \mathbb{R}^{m\times n\times l}$, ground truth $Y \in \mathbb{R}^{m\times n}$; patch size s; PCA band number b; epoch number a; training sample rate μ%.
Output: Predicted labels for the testing phase.

```
Retrieve the I_pca following PCA transformation
Partition the sample patches within I_pca into distinct sets for training and testing, thereby creating datasets for training load and test load.
for i from 1 to a do
    Perform 3D convolution layer.
    Perform multi-scale 2D convolution layers.
    Perform MS-SA to obtain the multi-scale spectral attention feature map.
    Perform lightweight Transformer module.
end for
Input the feature to the linear layer.
Apply softmax function for label identification.
The trained model predicts labels for the test dataset.
return predicted labels
```

3. Results

3.1. Dataset Description

The Indian Pines (IP) HSI dataset, collected by the United States Department of Agriculture (USDA) through the AVIRIS sensor, is situated in the southwestern region of Indiana, USA. It has a spatial resolution of 20 × 20 m, covers 224 spectral bands, and has a total pixel count of 21,025. After eliminating noise, which involves removing 24 bands and 10,776 pixel samples, 200 bands and 10,249 pixel samples are retained for classification purposes. The detailed distribution of the samples across each category is presented in Table 2. The false color map image is depicted in Figure 5a, while the ground truth map is presented in Figure 5b.

The Pavia University (UP) HSI dataset, affiliated with the University of Pavia, forms a component of an environmental monitoring initiative for the city of Pavia, Italy, employing the Reflective Optics System Imaging Spectrometer(ROSIS-3) sensor jointly developed by Dornier Satellite Systems, the GKSS Research Center and the German Aerospace Center. Comprising 115 continuous spectral bands, the dataset retains 103 spectral bands for classification following the removal of 12 bands affected by noise, as detailed in Table 3. Encompassing 42,776 pixel samples distributed across nine distinct categories, the false color map and ground truth map of the dataset are depicted in Figure 6a,b, respectively.

Table 2. The land cover categories in the IP dataset, along with the total number of samples for each category, are specified, including counts for both training and testing samples.

No.	Land Cover Categories	Total	Training	Test
1	Alfalfa	46	5	41
2	Corn-notill	1428	143	1285
3	Corn-mintill	830	83	747
4	Corn	237	24	213
5	Grass-pasture	483	48	435
6	Grass-trees	730	73	657
7	Grass-pasture-mowed	28	3	25
8	Hay-windrowed	478	48	430
9	Oats	20	2	18
10	Soybean-notill	972	97	875
11	Soybean-mintill	2455	245	2210
12	Soybean-clean	593	59	534
13	Wheat	205	20	185
14	Woods	1265	126	1139
15	Buildings-grass-trees-drives	386	39	347
16	Stone-steel-towers	93	9	84

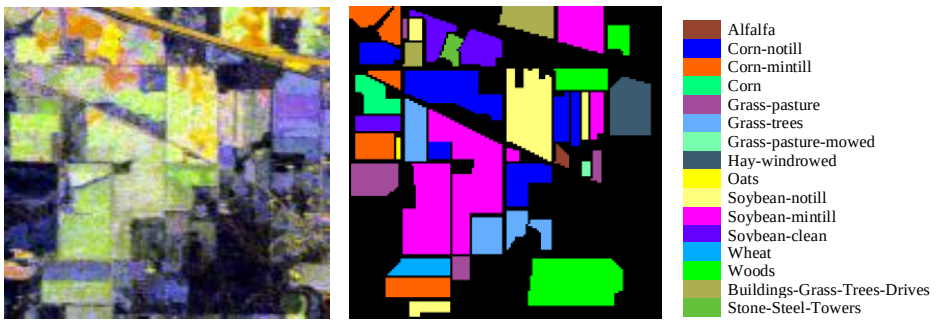
Table 3. The land cover categories in the UP dataset, along with the total number of samples for each category, are specified, including counts for both training and testing samples.

No.	Land Cover Classes	Total	Training	Test
1	Asphalt	6631	332	6299
2	Meadows	18,649	932	17,717
3	Gravel	2099	105	1994
4	Trees	3064	153	2911
5	Painted metal sheets	1345	67	1278
6	Bare soil	5029	251	4778
7	Bitumen	1330	67	1263
8	Self-blocking bricks	3682	184	3498
9	Shadows	947	47	900

Ultimately, the Salinas (SA) HSI dataset was obtained by NASA through the AVIRIS sensor, capturing imagery of the Salinas Valley region in Central California, USA. The dataset boasts an image resolution of 3.7 m. As outlined in Table 4, it comprises 54,129 samples categorized into 16 distinct land cover types. Following adjustments, the dataset is composed of 204 bands available for classification. Figure 7 visually illustrates the false color representation and ground truth representation of this dataset.

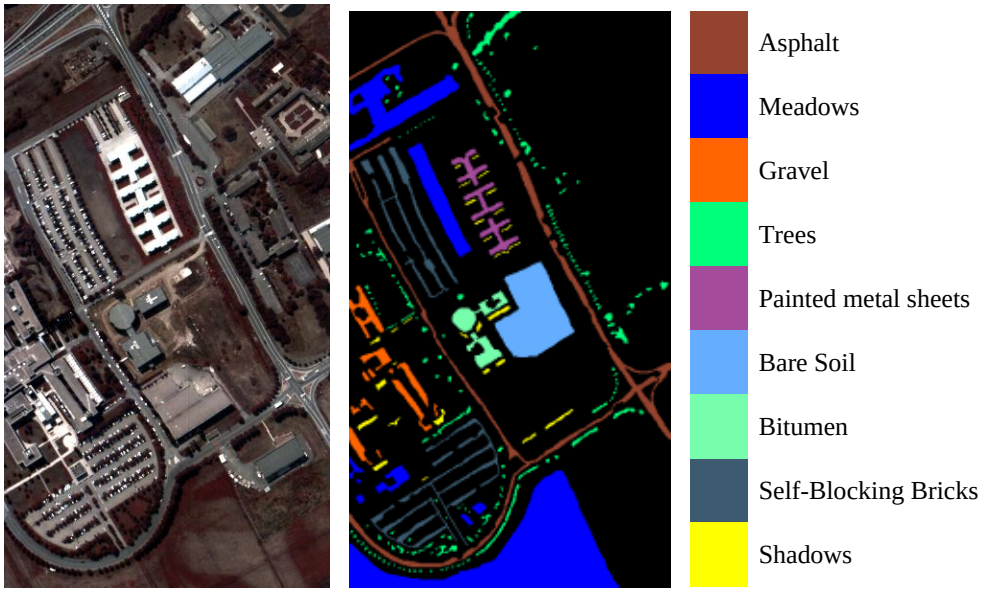
Table 4. The land cover categories in the SA dataset, along with the total number of samples for each category, are specified, including counts for both training and testing samples.

No.	Land Cover Classes	Total	Training	Test
1	Broccoli-green-weeds_1	2009	223	1786
2	Broccoli-green-weeds_2	3726	366	3360
3	Fallow	1976	187	1789
4	Fallow_rough_plow	1394	145	1249
5	Fallow_smooth	2678	272	2406
6	Stubble	3959	401	3558
7	Celery	3579	362	3217
8	Grapes_untrained	11,271	1138	10,133
9	Soil_vinyard_develop	6203	618	5585
10	Corn_senesced_green_weeds	3278	336	2942
11	Lettuce_romaine_4wk	1068	105	963
12	Lettuce_romaine_5wk	1927	201	1726
13	Lettuce_romaine_6wk	916	103	813
14	Lettuce_romaine_7wk	1070	104	966
15	Vinyard_untrained	7268	741	6527
16	Vinyard_vertical_trellis	1807	179	1628



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

Figure 5. IP dataset. (**a**) False color map. (**b**) Ground truth map.



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

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

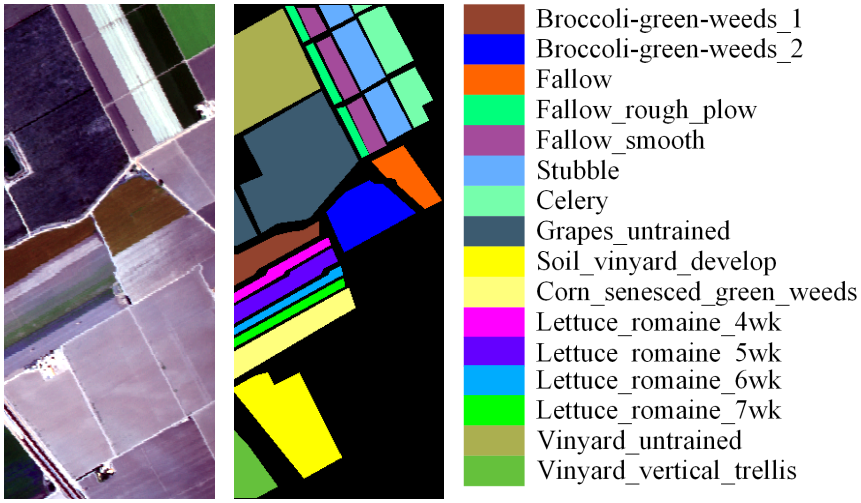


Figure 7. SA dataset. (**a**) False color map. (**b**) Ground truth map.

3.2. Experimental Configuration

All methods are tested on a server equipped with an NVIDIA GeForce RTX 3090 GPU. In terms of software, we use the Pycharm compiler on the Windows system, where the Python version is 3.7.12 and the Pytorch version is 1.10.1. The initial learning rate and training batch size are set to 0.001 and 64, respectively. The algorithm optimizer is the Adam optimizer. The training process for each dataset comprises 100 epochs.

3.3. Comparative Experiments

In this research, we utilize three classic HSI datasets for a comprehensive comparative analysis. The assessment of model performance encompasses four metrics: overall accuracy (OA), average accuracy (AA), kappa, and the accuracy of classification for individual land cover categories. We allocate 10% of each dataset for training purposes, reserving 90% for testing. Our proposed MSA-LWFormer is systematically compared with seven methods, including the traditional support vector machine (SVM) [9] and the deep learning methods 2D-CNN [30], 3D-CNN [31], HybridSN [32], SSFTT [38], and MDRDNet [34]. Tables 5–7 display the OA, AA, and kappa values for the IP, UP, and SA datasets. SVM depends on manually extracting spectral–spatial features from HSI and struggles to capture complex non-linear relationships. This limitation results in a reduction in performance when dealing with complex datasets. The 2D-CNN focuses on capturing the spatial structure and text information of HSI and ignores the extraction of spectral information. The 3D-CNN captures shallow spectral–spatial features but is at a disadvantage when processing complex information. HybridSN combines a 1D-CNN, 2D-CNN, and 3D-CNN to extract spectral–spatial features, but ignores the extraction of global features. SSFTT captures high-level semantic features but does not consider multi-scale feature extraction, which will cause insufficient feature extraction. MDRDNet considers the multi-scale spectral spatial features of HSI but does not consider the extraction of cross-channel features. MSA-LWFormer constructs the long-range spectral dependence of HSI, emphasizes the extraction of multi-scale and cross-channel features, and enhances the extraction of global information through the frequency-domain lightweight Transformer.

Table 5. Comparison of classification accuracy of different methods on the IP dataset (%).

NO.	SVM	2D-CNN	3D-CNN	HybridSN	SSFTT	MDRDNet	MSA-LWFormer
1	83.16	48.78	87.80	75.60	95.34	100	100
2	66.73	93.30	93.61	93.30	95.33	96.49	96.90
3	76.25	86.61	97.59	99.46	99.22	100	100
4	80.64	98.59	89.67	86.38	95.45	99.54	100
5	85.89	97.7	99.31	100	97.32	98.87	97.65
6	92.41	99.69	99.08	99.23	98.23	99.70	100
7	52.32	96	100	80	100	100	100
8	59.24	100	100	100	100	100	99.29
9	62.24	100	100	55.55	47.36	72.22	83.33
10	82.73	97.94	95.65	98.17	96.46	98.43	98.25
11	84.47	95.65	97.60	98.32	99.69	99.73	99.77
12	66.11	90.44	91.38	89.32	90.39	94.32	95.98
13	89.68	99.45	98.91	100	97.38	100	100
14	84.87	97.71	99.73	98.94	100	100	99.82
15	82.59	100	92.79	99.71	96.37	99.71	100
16	86.57	98.80	100	98.80	81.39	97.64	95.12
OA (%)	84.12	94.86	96.25	97.06	97.48	98.80	98.87
AA (%)	82.76	95.49	96.71	92.05	96.22	98.63	98.68
Kappa	0.8234	0.9379	0.9644	0.9664	0.9712	0.9729	0.9871

Table 6. Comparison of classification accuracy of different methods on the UP dataset (%).

NO.	SVM	2D-CNN	3D-CNN	HybridSN	SSFTT	MDRDNet	MSA-LWFormer
1	92.27	95.44	95.75	98.88	98.38	99.38	100
2	99.82	99.11	99.93	99.87	99.85	100	99.97
3	64.77	72.56	79.69	91.94	96.30	99.10	99.31
4	89.02	92.41	92.77	97.00	94.43	99.21	98.94
5	100	100	100	100	99.62	99.38	100
6	81.46	86.80	94.91	99.56	100	100	100
7	96.81	92.17	96.50	99.14	99.92	100	100
8	93.11	78.68	88.23	96.24	99.02	99.34	99.09
9	79.82	98.50	90.18	99.45	97.84	97.24	99.88
OA (%)	92.89	93.35	95.86	98.74	98.96	99.66	99.79
AA (%)	88.56	90.63	93.11	98.01	98.37	99.29	99.68
Kappa	0.9046	0.9113	0.9448	0.9834	0.9862	0.9956	0.9973

Table 7. Comparison of classification accuracy of different methods on the SA dataset (%).

NO.	SVM	2D-CNN	3D-CNN	HybridSN	SSFTT	MDRDNet	MSA-LWFormer
1	100	100	100	100	99.94	100	99.89
2	100	100	100	100	100	100	100
3	99.79	97.95	100	99.69	100	100	100
4	99.71	99.85	99.42	99.63	99.42	99.78	99.92
5	98.56	98.67	97.62	98.86	99.09	99.96	99.59
6	100	100	100	100	99.38	98.97	100
7	99.18	100	99.88	99.66	99.94	99.94	99.88
8	94.04	97.19	98.55	96.68	99.70	99.47	99.98
9	100	100	100	99.86	99.95	100	100
10	97.96	99.96	99.96	98.05	99.90	99.84	99.97
11	97.91	89.30	99.43	98.95	100	99.90	100
12	100	100	96.69	100	99.89	99.63	100

Table 7. *Cont.*

NO.	SVM	2D-CNN	3D-CNN	HybridSN	SSFTT	MDRDNet	MSA-LWFormer
13	75.74	96.69	67.25	96.03	95.25	99.00	100
14	96.03	85.55	97.63	99.05	99.43	99.81	100
15	85.24	94.03	90.13	91.02	98.40	99.04	100
16	99.10	94.13	98.15	99.44	100	99.77	100
OA(%)	95.95	97.72	97.44	97.74	99.50	99.62	99.96
AA(%)	96.45	97.08	96.54	98.56	99.39	99.69	99.95
Kappa	0.9549	0.9746	0.9714	0.9749	0.9945	0.9958	0.9995

3.3.1. Classification Results of IP

Table 5 offers a thorough summary of the performance demonstrated by the seven distinct algorithms when applied to the IP dataset. Of particular note is MSA-LWFormer, the novel algorithm introduced in this study, which demonstrates superior performance compared to its counterparts. Remarkably, MSA-LWFormer attains notable classification accuracy, achieving OA of 98.87% and AA of 98.68%. MSA-LWFormer surpasses its nearest competitor by 0.07% in OA and surpasses the best AA score by 0.05%.

These notable improvements in the classification metrics underscore the remarkable efficacy of MSA-LWFormer. A visual representation of the IP dataset's classification outcomes is illustrated in Figure 8.

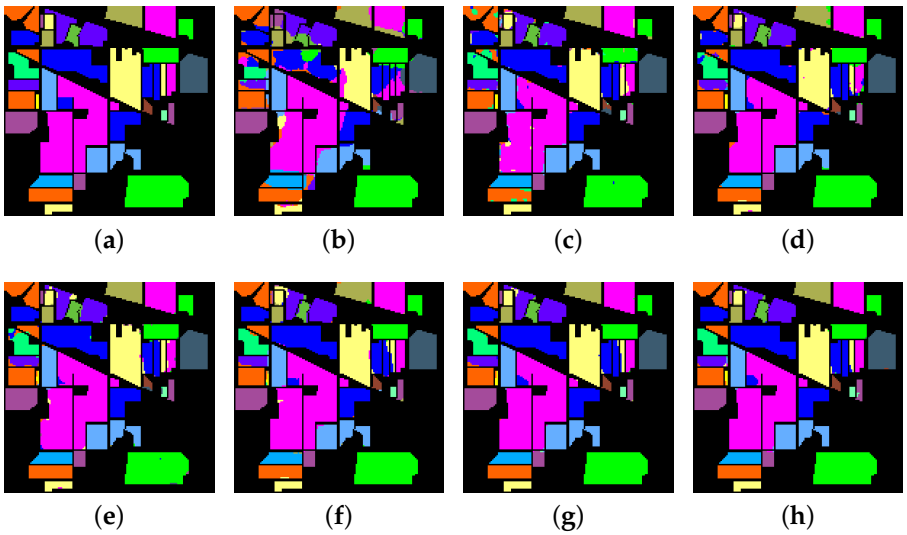


Figure 8. Classification maps for the IP dataset under seven methods. (**a**) Ground truth, (**b**) SVM, (**c**) 2D-CNN, (**d**) 3D-CNN, (**e**) HybridSN, (**f**) SSFTT, (**g**) MDRDNet, (**h**) MSA-LWFormer.

3.3.2. Classification Results of UP

In Table 6, the results of a comparative experiment are displayed, showcasing the application of the seven methods on the UP dataset. Consistently surpassing the alternative methods, MSA-LWFormer attains superior classification accuracy, achieving the highest scores in OA, AA, and kappa. It is noteworthy that MSA-LWFormer achieves OA of 99.66% and AA of 99.68%, surpassing the second-best performance by 0.7% and 0.39%, respectively.

Figure 9 shows the classification map of the seven methods on the UP dataset. Upon initial examination, the traditional SVM method demonstrates sensitivity to substantial speckle noise, resulting in OA and AA values of only 92.89% and 88.56%, respectively. The observed performance can be attributed to the limitations of SVM, specifically focusing on the spectral characteristics of individual pixels while neglecting the spatial dependencies among pixels.

In contrast, techniques based on deep learning proficiently harness both spectral and spatial features, leading to substantial enhancements in classification outcomes. Nonetheless, certain approaches like the 2D-CNN, 3D-CNN, and HybridSN encounter challenges related to misclassification, arising from the inadequate extraction of spectral–spatial features. In contrast, MSA-LWFormer exhibits precise classification outcomes, achieving optimal results in OA, AA, and kappa.

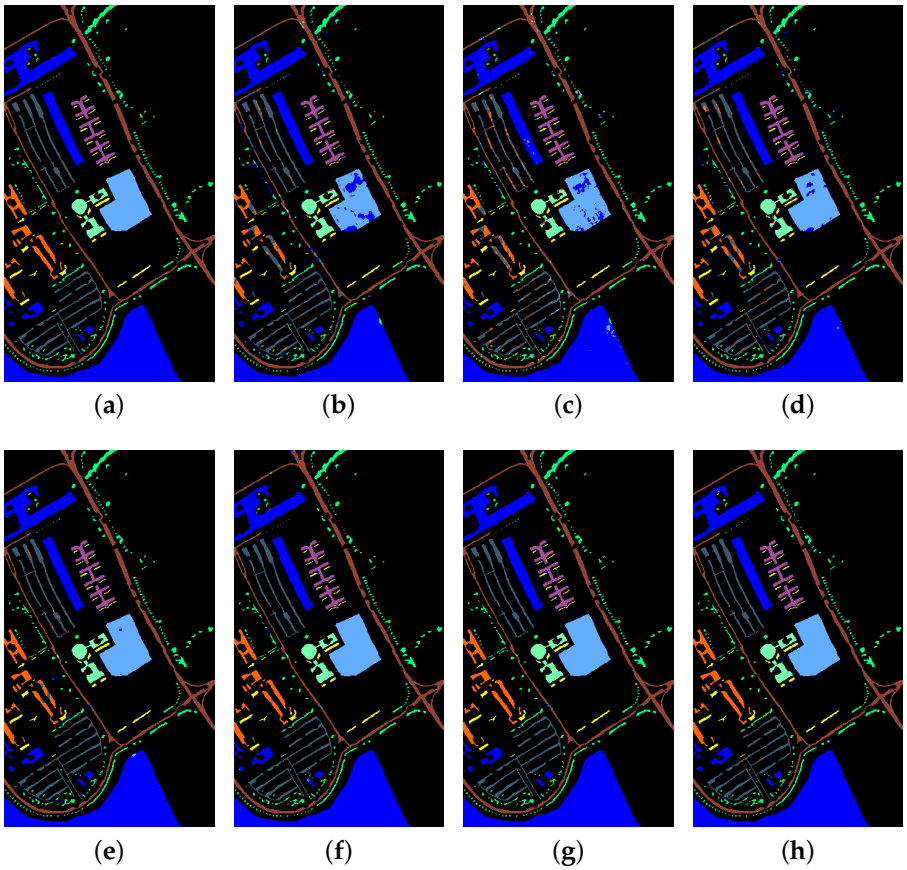


Figure 9. Classification maps for the UP dataset under seven methods. (**a**) Ground truth, (**b**) SVM, (**c**) 2D-CNN, (**d**) 3D-CNN, (**e**) HybridSN, (**f**) SSFTT, (**g**) MDRDNet, (**h**) MSA-LWFormer.

3.3.3. Classification Results of SA

The performance metrics for the SA dataset, as obtained by the seven different methods, are presented in Table 7. It is noteworthy that the SA dataset has a larger overall sample size compared to the IP and UP datasets, resulting in a larger number of training labels for each category. This inherent characteristic contributes to a comparatively higher level of classification accuracy.

MSA-LWFormer attains 100% classification accuracy across the 10 sample classes in the SA dataset, demonstrating outstanding results with accuracy rates of 99.96% for OA and 99.95% for AA. For improved visualization, Figure 10 illustrates the classification outcomes derived from the SA dataset, highlighting the evident excellence in MSA-LWFormer's performance.

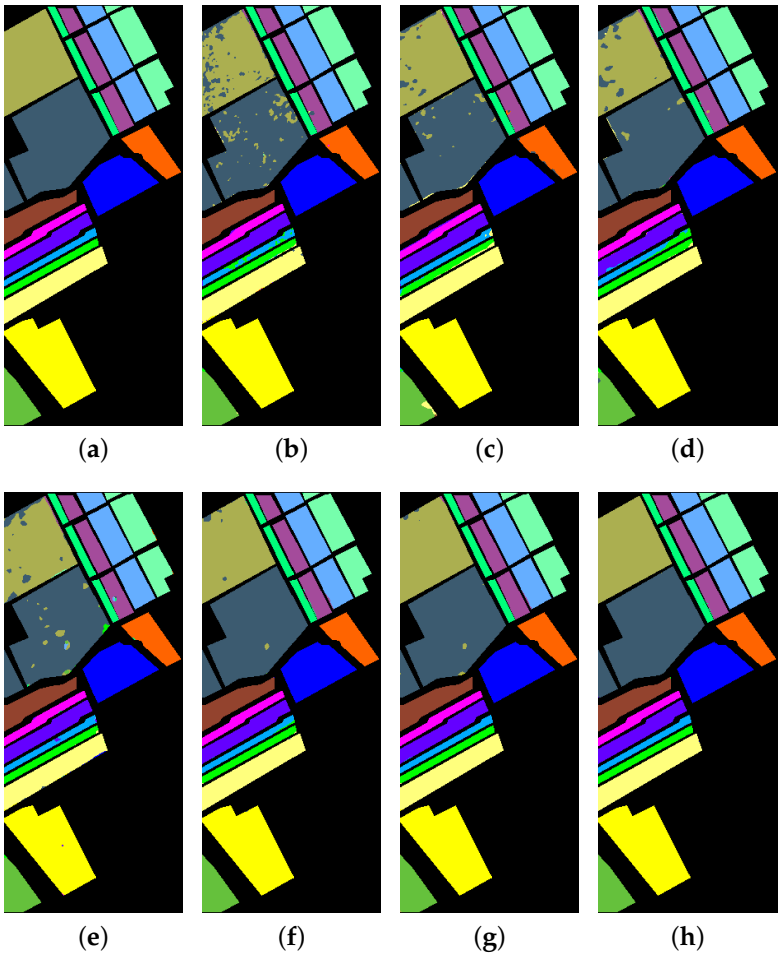


Figure 10. Classification maps for the SA dataset under seven methods. (**a**) Ground truth, (**b**) SVM, (**c**) 2D-CNN, (**d**) 3D-CNN, (**e**) HybridSN, (**f**) SSFTT, (**g**) MDRDNet, (**h**) MSA-LWFormer.

3.3.4. Experimental Results of Several Methods Using Varied Ratios of Training Samples

Figure 11 illustrates the classification accuracy of each approach across varying proportions of training samples. Labeled samples, constituting 1%, 3%, 5%, 7%, and 10% of the total samples, are chosen for training on the IP, UP, and SA datasets. The distinction in classification performance among different methods diminishes progressively with an increase in the number of training samples. Remarkably, our approach consistently demonstrates commendable performance, even when confronted with a limited sample size.

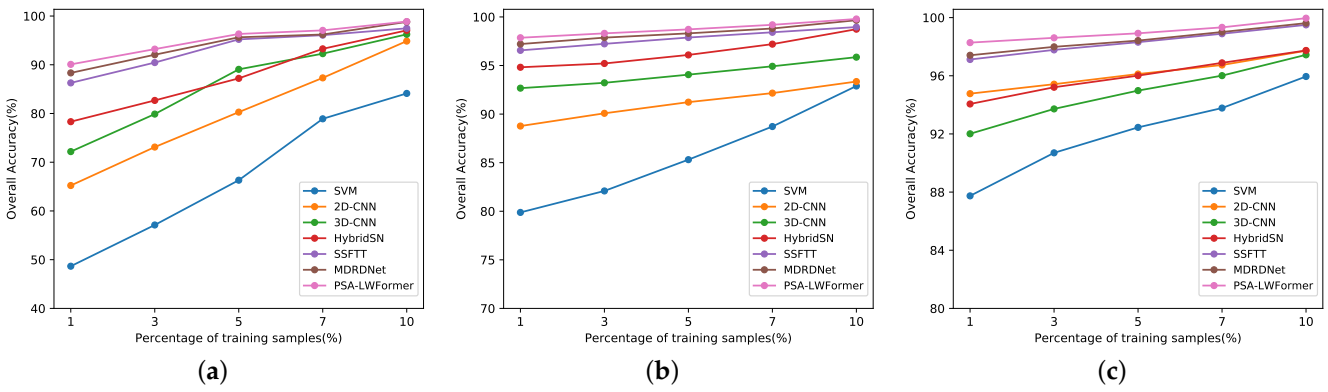


Figure 11. OA of different methods when using different ratios of training samples. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

4. Discussion

4.1. Ablation Experiment

4.1.1. Ablation Experiment Results of MSA-LWFormer on the UP Dataset

To thoroughly showcase the effectiveness of the proposed methodology, we performed ablation experiments on the UP dataset by testing different combinations of distinct components. Six specific combinations were systematically designed, and their impact on MSA-LWFormer was analyzed concerning the classification accuracy. The experimental outcomes are meticulously detailed and presented in Table 8. In particular, the overall model was deconstructed into four components, namely the 3D convolution layer (3D Conv), multi-scale 2D convolution layers (2D Conv), MS-SA, and a Transformer encoder based on frequency-domain attention (FDATE).

As presented in Table 8, the model without a multi-scale 2D convolution layer exhibited the lowest classification accuracy. In the absence of MS-SA, the model outperformed the second scenario. Specifically, in the first case, accuracy of 94.21% was achieved. In the fifth case, where two convolution layers were omitted to maintain a lightweight frequency attention Transformer and MS-SA, the classification accuracy reached 97.21%. While this performance is noteworthy, it is slightly inferior to that of our proposed method. The fifth case yielded classification accuracy of 90.88% for spectral–spatial features, obtained solely through two convolution layers. This underscores the significance of feature processing by the MSSA and FDATE modules in enhancing the performance. In conclusion, a thorough examination of the collective experimental results provides additional support for the effectiveness of our model.

Table 8. Ablation experiment results of MSA-LWFormer on the UP dataset.

Datasets	Case	Component				Indicator		
		3D Conv	2D Conv	MS-SA	FDATE	OA (%)	AA (%)	Kappa
UP	case1	✗	✓	✓	✓	92.88	90.77	0.9121
	case2	✓	✗	✓	✓	97.21	94.51	0.9538
	case3	✓	✓	✗	✓	94.21	92.72	0.9308
	case4	✓	✓	✓	✗	90.66	91.11	0.9021
	case5	✗	✗	✓	✓	88.68	89.72	0.8866
	case6	✓	✓	✓	✓	99.79	99.68	0.9973

4.1.2. Ablation Experiment Results for Transformer on the SA Dataset

We performed an additional ablation experiment on the Transformer encoder using a 10% sampling rate with the SA dataset. As illustrated in Table 9, the lightweight frequency-domain self-attention Transformer (Light-Transformer) achieves superior classification accuracy when compared to both the multi-head self-attention Transformer (MSA-Transformer) and the original softmax classifier. Additionally, in the testing phase, the Light-Transformer demonstrates the shortest running time. It is worth noting that optimal results were attained in terms of floating-point calculations/s (FLOPs) and parameter numbers. This outcome substantiates the claim that the Light-Transformer effectively mitigates the time complexity of the network.

Table 9. Ablation experiment results for Transformer on the SA dataset.

Methods	Accuracy Indicator			Efficiency Indicator		
	OA (%)	AA (%)	Kappa	Test Time (s)	Parameters (MB)	FLOPs (MB)
Softmax	98.91	99.07	0.9887	9.89	0.342773	1892.97
MSA-Transformer	99.72	99.63	0.9971	10.66	0.328125	1870.95
Light-Transformer	99.96	99.95	0.9995	9.48	0.310546	1841.97

4.2. Time Complexity and Parameter Count Analysis

In this section, we present a comparative analysis of the parameter quantities and test times for various innovative methods on the IP dataset. As depicted in Table 10, HybridSN incorporates multiple 3D convolutions, resulting in the highest parameter count and longest testing time. MDRDNet utilizes 3D depthwise separable convolutions, leading to a reduced time and parameter volume compared to HybridSN. SSFTT adopts a Transformer structure to alleviate the computational load. Our proposed MSA-LWFormer integrates the FFT into the Transformer structure, thereby achieving not only the shortest testing time but also the smallest parameter count.

Table 10. Test time (s) and parameters (MB) of several methods on the IP dataset.

Methods	Test Time (s)	Parameters (MB)
HybridSN	7.93	12.59
SSFTT	5.99	0.93
MDRDNet	7.12	1.92
MSA-LWFormer	5.81	0.86

4.3. Optimal Hyperparameters for MSA-LWFormer

To ascertain the optimal hyperparameters for the model, MSA-LWFormer underwent training on three publicly available datasets. Parameters including, but not limited to, the learning rate, batch size, training sample ratio, number of principal component analysis components, 3D convolution kernel size, and multi-scale 2D convolution kernel size were systematically varied and evaluated. The influence of these hyperparameters on the ultimate classification outcome was thoroughly examined.

4.3.1. Influence of Different Learning Rates

In the course of this investigation, the optimization of the model was executed utilizing the Adam optimizer, with a prescribed learning rate range comprising values of 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, and 0.0001. Figure 12 reveals that the optimal performance was achieved by the MSA-LWFormer architecture when the learning rate was configured at 0.001, particularly when applied to the three distinct hyperspectral datasets under consideration.

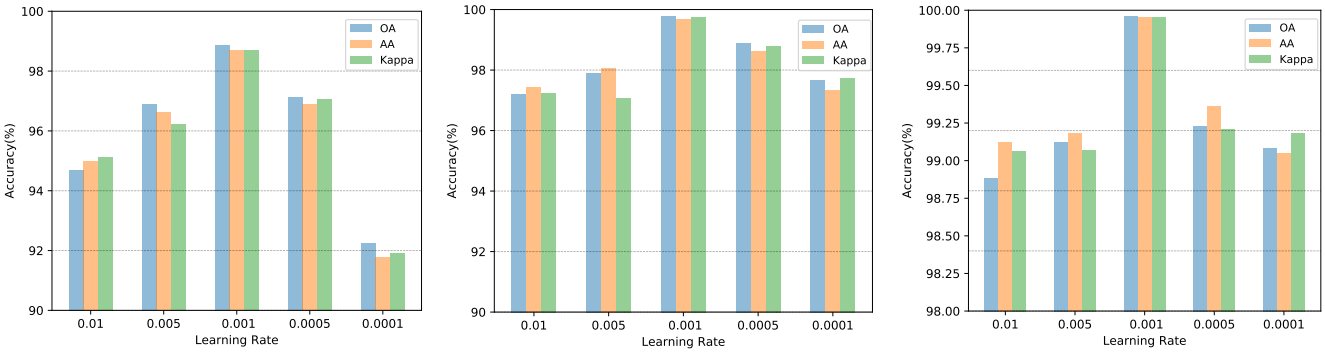


Figure 12. Influence of learning rate on classification accuracy. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

4.3.2. Influence of Different Patch Sizes

Figure 13 visually depicts the impact of the patch size on the classification accuracy across the three datasets. We examined patch sizes of 9, 11, 13, 15, and 17. The three subfigures reveal that the optimal classification results were consistently achieved at a patch size of 13 for all three datasets.

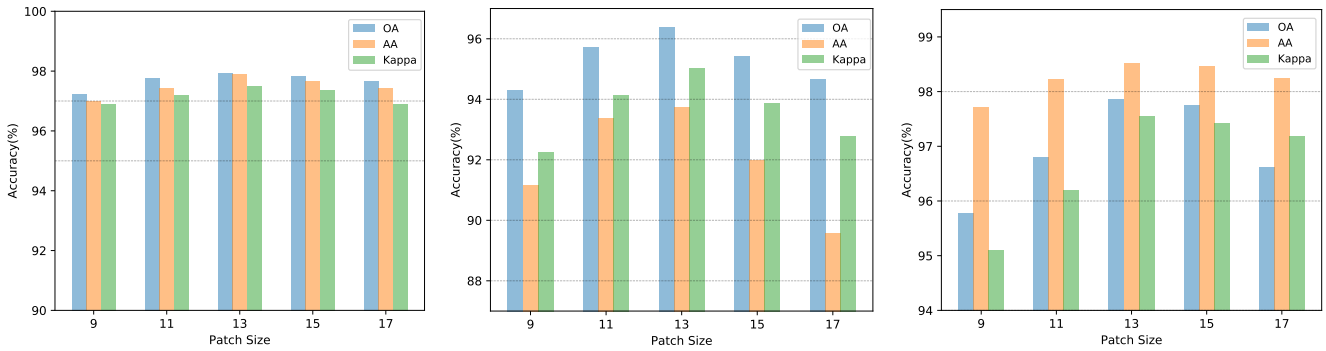


Figure 13. Influence of patch size on classification accuracy. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

4.3.3. Influence of the Number of Principal Components

The extensive continuous spectral range of HSI presents a classification challenge. In addressing this issue, MSA-LWFormer utilizes PCA on the initial HSI to reduce the dimensionality. We performed experiments using different numbers of principal components, namely 20, 25, 30, 35, and 40. As depicted in Figure 14, the optimal number of principal components for the IP, UP, and SA datasets was consistently determined to be 30.

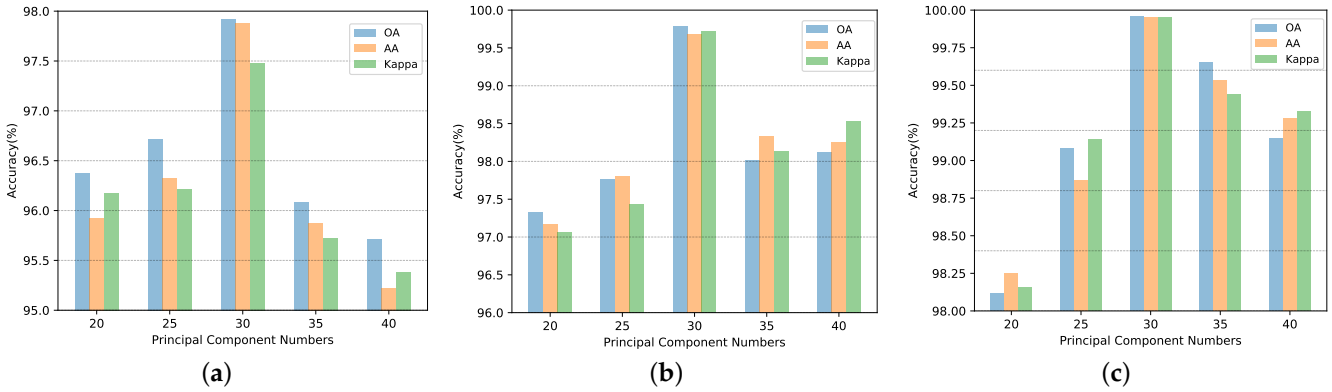


Figure 14. Influence of number of principal components on classification accuracy. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

4.3.4. Influence of 3D Convolution Kernel Number

The impact of varying the number of 3D convolution kernels on the OA, AA, and kappa is elucidated through the bar charts presented in Figure 15. Analyzing the IP dataset reveals a reduction in classification accuracy with an increasing number of 3D Convolution kernels, reaching its maximum at eight kernels. Similarly, for the UP dataset, the optimal configuration is attained with eight 3D kernels. Conversely, in the SA dataset, superior results are obtained when the quantity of 3D kernels is set to 24.

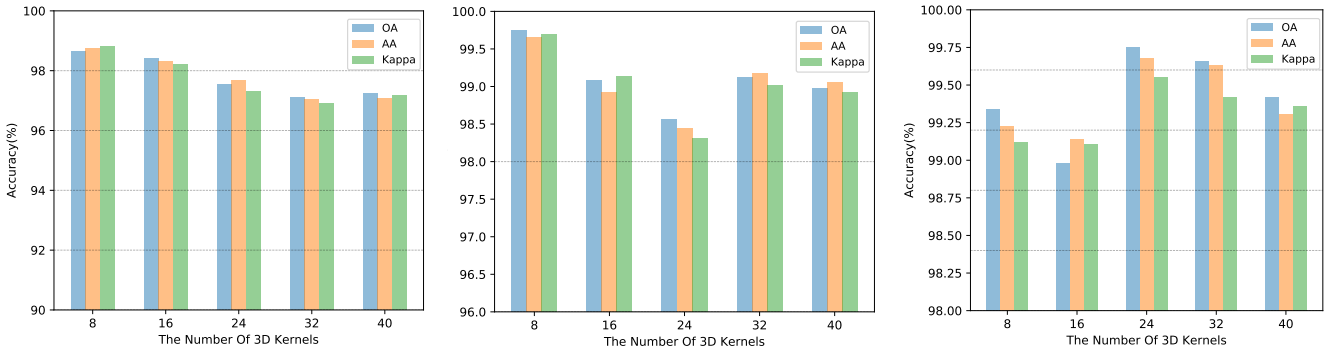


Figure 15. Influence of 3D convolution kernel number on classification accuracy. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

4.3.5. Influence of Multi-Scale 2D Convolution Kernel Size

Multi-scale 2D convolution employs four convolution kernels with varying sizes. The selection of the optimal kernel size plays a crucial role in influencing both the classification performance and computational complexity of the method. To examine the influence of different 2D convolution kernel sizes on the classification accuracy of MSA-LWFormer, we explored specific configurations, namely {{1, 3, 5, 7}, {2, 4, 6, 8}, {3, 5, 7, 9}, {4, 6, 8, 10}, and {5, 7, 9, 11}}. The corresponding results are presented in Figure 16.

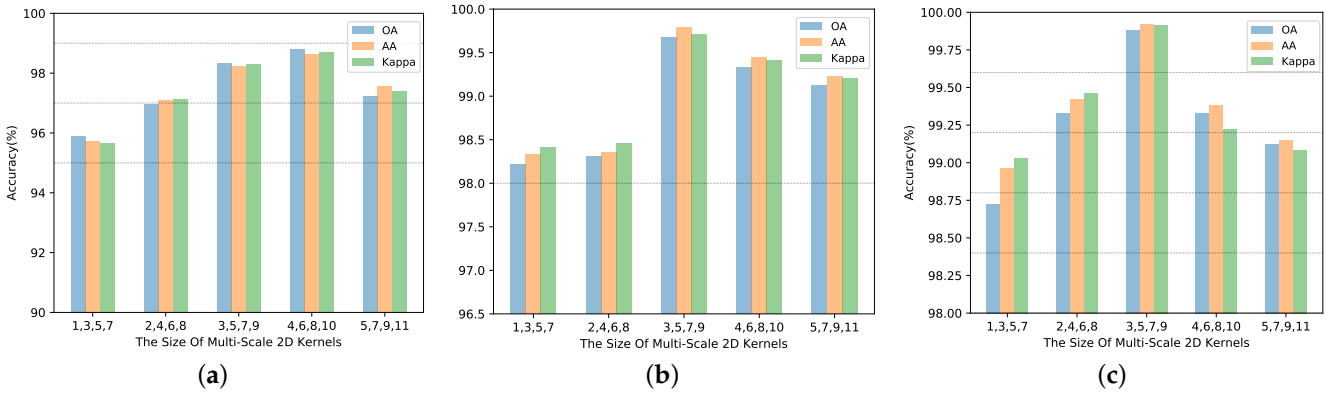


Figure 16. Influence of multi-scale 2D convolution kernel size on classification accuracy. (**a**) IP dataset. (**b**) UP dataset. (**c**) SA dataset.

The analysis of the IP dataset depicted in Figure 16 reveals that convolution kernel sizes of {4, 6, 8, 10} yield excellent accuracy across three standard metrics. In contrast, for the UP and SA datasets, the convolution kernel sizes {3, 5, 7, 9} exhibit superior performance. Consequently, we identify the optimal convolution kernel sizes for the three benchmark datasets as {4, 6, 8, 10}, {3, 5, 7, 9}, and {3, 5, 7, 9}, respectively.

5. Conclusions

This article presents a new approach known as MSA-LWFormer, designed to improve the efficacy of HSI classification. The technique incorporates a spectral–spatial feature extraction module, extracting shallow spectral–spatial features and establishing the long-range spectral dependence within HSI. The module contains 3D convolution layers, a multi-scale 2D-CNN, and MS-SA to integrate cross-channel and multi-scale features. In addition, a frequency-domain lightweight Transformer applies the FFT to convert features from the spatial domain to the frequency domain, effectively extracting global information and significantly reducing the time complexity of the network. Experiments on three classic HSI datasets demonstrate that MSA-LWFormer yields outstanding classification results.

While our proposed framework shows promise, there exists potential for further refinement. In future research endeavors, building upon MSA-LWFormer, we aim to integrate LIDAR data with HSI, thereby advancing the classification accuracy. Additionally, we intend to explore techniques such as transfer learning to bolster the network's accuracy when dealing with small sample sizes.

Author Contributions: Conceptualization: Q.G. and Y.S.; Data curation: Q.G. and H.L.; Methodology: Q.G.; Software: Q.G. and K.H.; Supervision: Y.S.; Writing—original draft: Q.G.; Writing—review and editing: Q.G. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grants U2001211 and 62276139.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the Managing Editor and anonymous reviewers for their insights and comments.

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

References

1. He, N.; Paoletti, M.E.; Haut, J.M.; Fang, L.; Li, S.; Plaza, A.; Plaza, J. Feature extraction with multiscale covariance maps for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 755–769. . [CrossRef]
2. Paoletti, M.; Haut, J.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [CrossRef]
3. Farmonov, N.; Amankulova, K.; Szatmári, J.; Sharifi, A.; Abbasi-Moghadam, D.; Nejad, S.M.M.; Mucsi, L. Crop type classification by DESIS hyperspectral imagery and machine learning algorithms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1576–1588. [CrossRef]
4. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]
5. Lu, B.; Dao, P.D.; Liu, J.; He, Y.; Shang, J. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens.* **2020**, *12*, 2659. [CrossRef]
6. Tan, Y.; Lu, L.; Bruzzone, L.; Guan, R.; Chang, Z.; Yang, C. Hyperspectral band selection for lithologic discrimination and geological mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 471–486. [CrossRef]
7. Zhang, Y.; Cao, G.; Li, X.; Wang, B.; Fu, P. Active semi-supervised random forest for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 2974. [CrossRef]
8. Liu, B.; Yu, X.; Zhang, P.; Yu, A.; Fu, Q.; Wei, X. Supervised deep feature extraction for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1909–1921. [CrossRef]
9. Wang, Y.; Duan, H. Classification of hyperspectral images by SVM using a composite kernel by employing spectral, spatial and hierarchical structure information. *Remote Sens.* **2018**, *10*, 441. [CrossRef]
10. Cariou, C.; Le Moan, S.; Chehdi, K. Improving K-nearest neighbor approaches for density-based pixel clustering in hyperspectral remote sensing images. *Remote Sens.* **2020**, *12*, 3745. [CrossRef]
11. Xu, L.; Li, J. Bayesian classification of hyperspectral imagery based on probabilistic sparse representation and Markov random field. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 823–827. [CrossRef]
12. Machidon, A.L.; Del Frate, F.; Picchiani, M.; Machidon, O.M.; Ogrutan, P.L. Geometrical approximated principal component analysis for hyperspectral image analysis. *Remote Sens.* **2020**, *12*, 1698. [CrossRef]
13. Li, W.; Zhang, L.; Zhang, L.; Du, B. GPU parallel implementation of isometric mapping for hyperspectral classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1532–1536. [CrossRef]
14. Fang, Y.; Li, H.; Ma, Y.; Liang, K.; Hu, Y.; Zhang, S.; Wang, H. Dimensionality reduction of hyperspectral images based on robust spatial information using locally linear embedding. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1712–1716. [CrossRef]
15. Yang, Y.; Wei, X.; Xu, R.; Wang, W.; Peng, L.; Wang, Y. Jointly beam stealing attackers detection and localization without training: An image processing viewpoint. *Front. Comput. Sci.* **2023**, *17*, 173704. [CrossRef]
16. Guo, K.; Wu, Z.; Wang, W.; Ren, S.; Zhou, X.; Gadekallu, T.R.; Luo, E.; Liu, C. GRTR: Gradient Rebalanced Traffic Sign Recognition for Autonomous Vehicles. *IEEE Trans. Autom. Sci. Eng.* **2023**. [CrossRef]
17. Shi, Q.; Tang, X.; Yang, T.; Liu, R.; Zhang, L. Hyperspectral image denoising using a 3-D attention denoising network. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 10348–10363. [CrossRef]

18. Wang, Q.; Wu, Z.; Jin, J.; Wang, T.; Shen, Y. Low rank constraint and spatial spectral total variation for hyperspectral image mixed denoising. *Signal Process.* **2018**, *142*, 11–26. [CrossRef]
19. Zhang, G.; Zhao, S.; Li, W.; Du, Q.; Ran, Q.; Tao, R. HTD-Net: A deep convolutional neural network for target detection in hyperspectral imagery. *Remote Sens.* **2020**, *12*, 1489. [CrossRef]
20. Zhang, Y.; Du, B.; Zhang, Y.; Zhang, L. Spatially adaptive sparse representation for target detection in hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1923–1927. [CrossRef]
21. Song, A.; Choi, J.; Han, Y.; Kim, Y. Change detection in hyperspectral images using recurrent 3D fully convolutional networks. *Remote Sens.* **2018**, *10*, 1827. [CrossRef]
22. Gong, M.; Jiang, F.; Qin, A.K.; Liu, T.; Zhan, T.; Lu, D.; Zheng, H.; Zhang, M. A spectral and spatial attention network for change detection in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
23. 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]
24. Gao, L.; Yao, D.; Li, Q.; Zhuang, L.; Zhang, B.; Bioucas-Dias, J.M. A new low-rank representation based hyperspectral image denoising method for mineral mapping. *Remote Sens.* **2017**, *9*, 1145. [CrossRef]
25. Yang, J.; Zhao, Y.Q.; Chan, J.C.W.; Xiao, L. A multi-scale wavelet 3D-CNN for hyperspectral image super-resolution. *Remote Sens.* **2019**, *11*, 1557. [CrossRef]
26. Wu, H.; Prasad, S. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sens.* **2017**, *9*, 298. [CrossRef]
27. Li, J.; Xi, B.; Li, Y.; Du, Q.; Wang, K. Hyperspectral classification based on texture feature enhancement and deep belief networks. *Remote Sens.* **2018**, *10*, 396. [CrossRef]
28. Liang, P.; Shi, W.; Zhang, X. Remote sensing image classification based on stacked denoising autoencoder. *Remote Sens.* **2017**, *10*, 16. [CrossRef]
29. Yu, S.; Jia, S.; Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **2017**, *219*, 88–98. [CrossRef]
30. Gao, H.; Lin, S.; Yang, Y.; Li, C.; Yang, M. Convolution neural network based on two-dimensional spectrum for hyperspectral image classification. *J. Sens.* **2018**, *2018*, 8602103. [CrossRef]
31. Xu, H.; Yao, W.; Cheng, L.; Li, B. Multiple spectral resolution 3D convolutional neural network for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 1248. [CrossRef]
32. 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]
33. He, M.; Li, B.; Chen, H. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), IEEE, Beijing, China, 17–20 September 2017; pp. 3904–3908. [CrossRef]
34. Hu, Y.; Tian, S.; Ge, J. Hybrid Convolutional Network Combining Multiscale 3D Depthwise Separable Convolution and CBAM Residual Dilated Convolution for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 4796. [CrossRef]
35. Mou, L.; Zhu, X.X. Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 110–122. [CrossRef]
36. Cui, B.; Wen, J.; Song, X.; He, J. MADANet: A Lightweight Hyperspectral Image Classification Network with Multiscale Feature Aggregation and a Dual Attention Mechanism. *Remote Sens.* **2023**, *15*, 5222. [CrossRef]
37. 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]
38. 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]
39. Kong, L.; Dong, J.; Ge, J.; Li, M.; Pan, J. Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5886–5895. [CrossRef]