remote sensing

MDPI

Article

Hyperspectral Image Classification Based on Multi-Scale Convolutional Features and Multi-Attention Mechanisms

Qian Sun [1], Guangrui Zhao [1], Xinyuan Xia [2,*], Yu Xie [2], Chenrong Fang [3], Le Sun [4,5], Zebin Wu [2] and Chengsheng Pan [1]

[1] School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; sunqian@nuist.edu.cn (Q.S.); cs_zhaogr@nuist.edu.cn (G.Z.); pancs@sohu.com (C.P.)
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; xavier-xie@njust.edu.cn (Y.X.); wuzb@njust.edu.cn (Z.W.)
[3] College of Intelligence and Computing, Tianjin University, Tianjin 300000, China; fangchenrong@tju.edu.cn
[4] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; sunlecncom@nuist.edu.cn
[5] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: xinyuan@njust.edu.cn

Abstract: Convolutional neural network (CNN)-based and Transformer-based methods for hyperspectral image (HSI) classification have rapidly advanced due to their unique characterization capabilities. However, the fixed kernel sizes in convolutional layers limit the comprehensive utilization of multi-scale features in HSI land cover analysis, while the Transformer's multi-head self-attention (MHSA) mechanism faces challenges in effectively encoding feature information across various dimensions. To tackle this issue, this article introduces an HSI classification method, based on multi-scale convolutional features and multi-attention mechanisms (i.e., MSCF-MAM). Firstly, the model employs a multi-scale convolutional module to capture features across different scales in HSIs. Secondly, to enhance the integration of local and global channel features and establish long-range dependencies, a feature enhancement module based on pyramid squeeze attention (PSA) is employed. Lastly, the model leverages a classical Transformer Encoder (TE) and linear layers to encode and classify the transformed spatial–spectral features. The proposed method is evaluated on three publicly available datasets—Salina Valley (SV), WHU-Hi-HanChuan (HC), and WHU-Hi-HongHu (HH). Extensive experimental results have demonstrated that the MSCF-MAM method outperforms several representative methods in terms of classification performance.

Keywords: hyperspectral image classification; convolutional neural network; Transformer; attention mechanism; multi-scale features

check for updates

Citation: Sun, Q.; Zhao, G.; Xia, X.; Xie, Y.; Fang, C.; Sun, L.; Wu, Z.; Pan, C. Hyperspectral Image Classification Based on Multi-Scale Convolutional Features and Multi-Attention Mechanisms. *Remote Sens.* **2024**, *16*, 2185. https://doi.org/10.3390/rs16122185

Academic Editors: Xudong Kang, Pedram Ghamisi and Mingmin Chi

Received: 8 May 2024
Revised: 31 May 2024
Accepted: 11 June 2024
Published: 16 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of spectral imaging technology, we can now acquire numerous high-quality hyperspectral images (HSIs) [1,2]. These images offer abundant spectral details through dense narrow-band spectral imaging and find extensive utility in various domains like vegetation surveys [3], precision agriculture [4,5], atmospheric assessments [6], geological explorations [7], food safety [8], medical diagnostics [9], and military surveillance [10]. As the utilization of HSI applications proliferates, the swift evolution of HSI processing and analysis methods including unmixing [11,12], denoising [13,14], pan-sharpening [15], super-resolution [16], target detection [17,18], and classification [19,20] has been observed. Within the realm of these HSI data processing techniques, the significance of HSI land cover classification emerges distinctly. Nonetheless, the task of HSI classification poses considerable challenges owing to the intricate high-dimensional nature of HSI and the nuanced spectral distinctions among features [21].

Traditional HSI classification methods typically depend on manual feature extraction and conventional machine learning (ML) algorithms. Many traditional methods initially reduce the dimensionality of high-dimensional spectral data using methods like principal component analysis (PCA) [22] and linear discriminant analysis [23] to preserve key spectral features. Then, the obtained spectral feature is used to assign pixels to different classes using traditional ML classifiers, such as support vector machines (SVMs) [24], random forests [25], and k-nearest neighbors [26], etc. In addition to using only spectral features, traditional classification methods often incorporate some spatial features to more fully capture these features. Some methods based on mathematical morphology are used to characterize the shape features, such as morphological profile (MP) [27], extended MP (EMP) [28], and extended multi-attribute profile [29]. Some local spatial statistical features based on the neighborhood around a pixel help to capture the local spatial structure of features, such as local binary pattern (LBP) [30] and Gabor wavelet transform [31]. These traditional spatial features are usually used in combination with spectral features to form a fused feature set, which is then classified by applying traditional ML classifiers. Although these traditional methods have achieved some success in certain scenarios, they typically have a limited ability to model the nonlinear relationships and complex spatial structures of spectral information.

The emergence of deep learning (DL) methods has ushered in superior performance and enhanced feature learning flexibility for HSI classification tasks, surpassing traditional machine learning (ML) classification techniques [32]. The initial DL methods in HSI classification tried to use stacked autoencoders [33] and deep belief networks [34]. However, such methods still regard the pixels as 1D vectors as inputs to the network, which cannot effectively retain spatial information. Subsequently, a series of prominent DL backbone models have been introduced to the HSI classification task. Generative adversarial networks (GANs) [35,36] based on generative models can be used to generate virtual samples with similar spectral features for HSI data enhancement. In the HSI classification task, the availability of labeled sample data is typically limited. Augmenting the dataset using GANs has been proven to be beneficial in enhancing the model's performance and generalization capabilities. Capsule networks [37] based on dynamic routing mechanisms can learn the feature representation of hierarchical structures through the capsule layer, which can effectively deal with the complex hierarchical structure of HSIs. As a feed-forward neural network, recurrent neural networks [38,39] can effectively model the correlation information between long-range spectral bands. Graph convolutional networks (GCNs) [40,41] have demonstrated the capability to effectively capture spatial relationships among pixels in HSIs. This is achieved by treating pixels as nodes and dynamically constructing graph representations based on their interrelations. This is very helpful in processing spatial structure and texture information in HSIs. Convolutional neural networks (CNNs) are extensively employed in image processing for their capability to capture spatial structural information and have emerged as the predominant approach in HSI classification [42]. Hu et al. [43] first designed a 1D CNN structure to achieve HSI classification. When processing HSIs, the method needs to be converted into a 1D feature sequence, which loses some spatial information. Zhao and Du [44] constructed a 2D CNN structure based on image spatial dimension, which mainly consists of five 2D convolutional layers, for the spectral–spatial classification of several principal component bands of HSIs. Zhao et al. [45] proposed a 2D CNN with spectral attention based on a compact band weighting module for capturing spectral information in HSIs. To extract correlation information between spatial and spectral dimensions in HSIs, Hamida et al. [46] proposed a 3D CNN to jointly process spatial and spectral features from multiple dimensions of the data. Apart from a single CNN structure, Roy et al. [47] designed a hybrid 3D and 2D CNN (HybridSN) model, which effectively extracts complementary spatial–spectral features and spectral features for HSI classification. Song et al. [48] introduced a deep feature fusion network (DFFN) based on residual learning, which effectively constructs special relationships between multilevel features in deep networks. While methods based on CNN have proved to be successful in

HSI classification, they are not without limitations. Firstly, the feature receptive fields of CNN-based methods are limited to a square region with convolutional kernel size, which makes it difficult for CNN-based methods to model global information and long-range dependencies in images. Secondly, when the HSI contains land cover objects with complex shapes and different scales, it is difficult for the CNN-based method to effectively extract multi-scale features.

Recently, a deep learning model named Transformer, utilizing a multi-head self-attention (MHSA) architecture, was originally introduced for natural language processing tasks [49]. Leveraging the self-attention mechanism, the Transformer model excels at capturing long-range dependencies within sequences, enabling focused feature extraction across various scales and positions in the input sequence. It is also due to this advantage that Transformers are widely used in image processing [50]. CNNs usually use a fixed-size convolutional kernel when processing images, while Transformers can establish global correlations over the entire image, which helps one better understand the overall structure of the image [51]. Similarly, many Transformer-based methods are becoming increasingly used in HSI classification tasks. He et al. [52] first proposed a spatial–spectral Transformer model for HSI classification. This model utilizes a CNN-based backbone to extract spatial features and a densely connected Transformer to capture relationships between neighboring spectra. Hong et al. [53] developed the spectralformer, a spectral-based model that extracts spectral representation information from grouped neighboring bands and incorporates a cross-layer Transformer encoder module. Subsequently, Sun et al. [54] introduced the spectral–spatial feature tokenization Transformer (SSFTT) model to capture spectral–spatial features and extract high-level semantic features. Mei et al. [55] proposed a local–global attention-based group-aware hierarchical Transformer (GAHT) for HSI classification. Combining graph convolutional networks and Transformer networks, Yang et al. [56] designed an end-to-end classification method for joint spatial–spectral feature extraction. With the help of the superpixel segmentation algorithm, Feng et al. [57] proposed a center attention Transformer with a stratified spatial–spectral token generated by superpixel sampling to further improve the classification performance. Zou et al. [58] proposed a local-enhanced spectral–spatial Transformer method. By converting the HSI into an adaptive spectral–spatial token representation, this method enhanced the discriminative ability of these tokens and improved the classification performance by capturing both local and global information. Peng et al. [59] designed a cross-attention spatial–spectral Transformer to achieve effective HSI classification by fusing features from both spatial and spectral branches. With the help of spatial and spectral morphological convolutions, Roy et al. [60] proposed a spectral–spatial morphological attention Transformer (MorghAT). Ouyang et al. [61] designed a hybrid Transformer encoder with spatial–spectral serial attention to effectively improve the classification accuracy. Fang et al. [62] fused traditional and deep features by constructing a multi-attention Transformer with multi-level features, which effectively improved the classification performance. These methods make full use of the long sequence modeling capability of the Transformer to tap the correlation of spatial- and spectral-based feature sequences, but seldom take into account the spatial shape scale variations of the image.

As a result, this paper proposes the multi-scale convolutional feature and multi-attention mechanism (MSCF-MAM) classification method, which utilizes CNN and Transformer to address these issues and enhance existing HSI classification techniques. The network architecture includes a multi-scale convolutional feature extraction module, a feature enhancement module leveraging pyramid squeeze attention (PSA), and a Transformer encoder (TE) module. The multi-scale convolutional feature extraction module aims to capture diverse scale information present in HSI patches by employing three 3D convolutional blocks with varying kernel sizes to effectively represent different object shapes and scale variations. The feature enhancement module consists mainly of the PSA. The PSA mainly consists of multi-scale group convolutional layers and multiple squeeze and excitation (SE) attention. This component aims to bolster the multi-scale feature representation and facilitate the establishment of long-range feature dependencies through

diverse attention mechanisms. The TE module, built on top of the multi-head self-attention (MHSA) mechanism, further refines the internal feature correlations and generates the ***cls*** token for classification. The classification is performed using a linear layer to obtain the ultimate category label. The primary contributions of this paper are detailed below.

- This method employs multi-scale convolutional kernels to capture features at different scales in HSIs. By using the multi-scale convolutional kernels, the model can adapt and capture detail and structural information on these different scales, helping to retain more information and thus improving the ability to identify complex land cover environments.
- Multiple attention mechanisms based on the pyramid squeeze attention and multi-head self-attention mechanism are introduced to enhance the modeling capabilities of the perception and utilization of critical information in HSIs. By using multiple attention mechanisms, the redundant or secondary information present in HSIs is filtered to improve the modeling efficiency and comprehensively capture the image features.
- The systematic combination of the multi-scale CNNs and multiple attention mechanisms can fully and efficiently exploit the spectral and spatial features in HSIs, thereby significantly improving the classification performance. The experiments conducted on three public datasets demonstrate the superior performance of the proposed method.

The structure of this paper is as follows. Section 2 provides a comprehensive description of the MSCF-MAM method. Section 3 presents the three experimental datasets, the design of the experimental parameters, extensive experimental results, and further analyses of the proposed method. Some related conclusions and possible future improvements are summarized in Section 4.

2. Materials and Methods

The overall framework of the proposed MSCF-MAM method for HSI classification is illustrated in Figure 1. It is divided into four main modules—HSI data pre-processing, multi-scale convolutional features extraction, feature enhancement-based pyramid squeeze attention, and Transformer encoder.

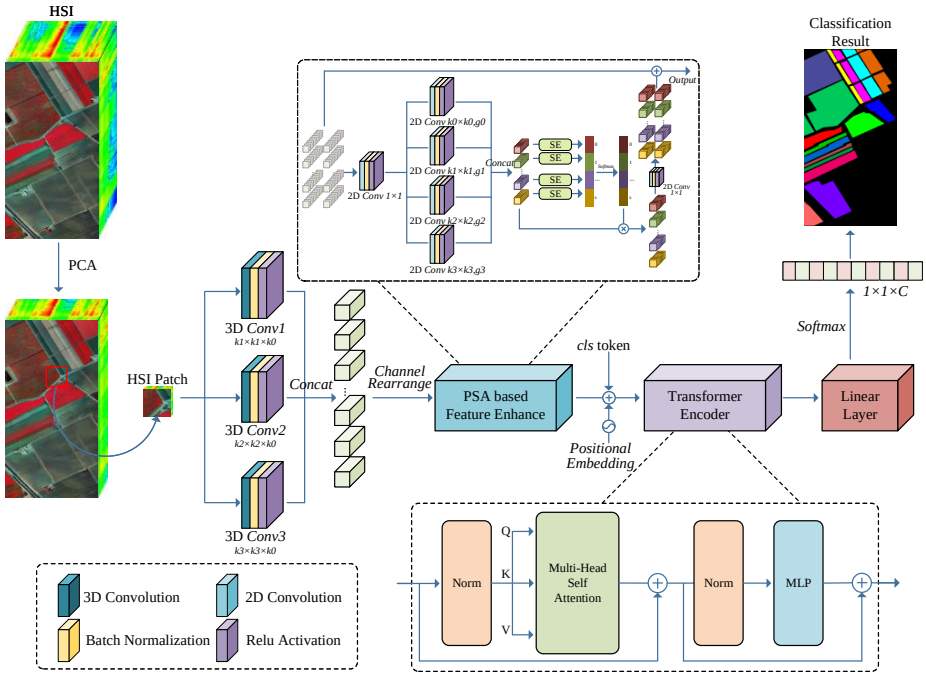


Figure 1. The overall framework of the proposed MSCF-MAM method.

2.1. HSI Data Pre-Processing

We represent the obtained HSI as $\boldsymbol{X} \in \mathbb{R}^{m \times n \times d}$, where m is the image spatial height, n is the image spatial width, and d represents the image spectral band number. While the HSI captures spectral data from multiple bands and offers detailed land cover information, the complexity of high-dimensional data presents processing challenges. Hence, preprocessing is necessary to augment the retention of valuable image data and alleviate computational load. PCA, a conventional data preprocessing technique, is adept at reducing the dimensionality of high-dimensional HSI data. It extracts the first b principal components of $\boldsymbol{X}$ while preserving the spatial dimensions. The dimension-reduced HSI is represented as $\boldsymbol{X_{pca}} \in \mathbb{R}^{m \times n \times b}$, where b corresponds to the number of spectral dimensions resulting from the PCA transformation.

Then, for multi-scale convolutional feature extraction, 3D patches are extracted from $\boldsymbol{X_{pca}}$. Each 3D neighbouring patch $\boldsymbol{P} \in \mathbb{R}^{s \times s \times b}$ is generated from $\boldsymbol{X_{pca}}$, where $s \times s$ is the size of the spatial window. The center coordinates in each patch's spatial dimensions are set to (x_i, x_j), with $0 \leq i < m$ and $0 \leq j < n$. The category of the central pixel determines the true category of each patch. Moreover, when the 3D patches around the edge pixels are extracted, a padding operation of width $(s-1)/2$ is performed on these pixels. Hence, the total number of 3D patches generated from $\boldsymbol{X_{pca}}$ is $m \times n$. Following the removal of patches labeled as zero, the remaining samples are partitioned into training and test sets.

2.2. Multi-Scale Convolutional Feature Extraction

After data pre-processing, the multi-scale convolutional features of the 3D patches were extracted using a set of multi-scale 3D convolution blocks. As shown in Figure 1, three 3D convolutions blocks, namely, $3Dconv1$, $3Dconv2$, $3Dconv3$, with kernel sizes $k_1 \times k_1 \times k_0$, $k_2 \times k_2 \times k_0$, $k_3 \times k_3 \times k_0$ operate on the 3D image patch simultaneously. The primary objective of this design is to undertake multi-scale feature extraction of the image characteristics in the spatial domain. Given a 3D patch $\boldsymbol{P}$ and a 3D convolution kernel $\boldsymbol{K}$, the 3D convolution operation can be expressed as

$$G(x,y,z) = \sum_{\alpha=1}^{k_i} \sum_{\beta=1}^{k_i} \sum_{\gamma=1}^{k_0} \boldsymbol{P}(x+\alpha, y+\beta, z+\gamma) \cdot \boldsymbol{K}(\alpha,\beta,\gamma) \tag{1}$$

$G(x,y,z)$ is an element of the output feature map. $\boldsymbol{P}(x+\alpha, y+\beta, z+\gamma)$ denotes the input 3D patch at position $(x+\alpha, y+\beta, z+\gamma)$. $\boldsymbol{K}(\alpha,\beta,\gamma)$ is the weight of the convolution kernel at position (α,β,γ). k_i represents the width and height size of the convolution kernel, $i \in [1,2,3]$.

Next, based on the size of the 3D convolution kernel, a feature padding operation is performed after each 3D convolution to ensure that the feature maps obtained after the three convolution operations are maintained at the same size. Next, we fused the multi-scale features using a simple concatenation operation. The multi-scale 3D convolutional feature output G_{concat} using a concatenation operation can then be expressed as

$$G_{concat} = \text{Concat}(G_1, G_2, G_3, \text{axis} = \alpha) \tag{2}$$

concat denotes the concatenation operation, and axis $= \alpha$ specifies the dimension in which the splicing operation takes place. We set the concatenation operation on the channel dimension of the feature maps and concatenate the feature maps obtained from different convolutional blocks directly. This operation efficiently integrates features of varying scales while retaining all output information from the convolution blocks, thereby enhancing the model's capability to represent features of diverse dimensions.

2.3. Feature Enhancement Based on Pyramid Squeeze Attention

The shallow multi-scale convolutional features obtained from multi-scale 3D convolutional operations may not sufficiently represent rich feature information. Utilizing a

feature enhancement module based on PSA enables the effective acquisition of a more comprehensive multi-scale feature representation, dynamically adjusting the attention weights of each channel.

As illustrated in Figure 1, the feature enhancement module mainly comprises the PSA module, which involves three steps: Firstly, leveraging a multi-scale pyramid structure to acquire spatial information of various scales on each channel feature map. Secondly, feeding the multi-scale feature map into the SE module to establish the attention mechanism of the channel within the multi-scale feature map. Lastly, recalibrating the attention weights of the channel within the multi-scale feature map using the softmax algorithm and multiplying them with the feature maps from the initial step to generate multi-scale spatial–spectral feature representations.

Initially, a 2D point-by-point convolution is used to reduce the number of channels by mapping the acquired multi-scale features through channel mapping. This operation effectively reduces parameters, enhances nonlinear expressive capabilities, and preserves spatial information. Subsequently, the multi-scale pyramid structure facilitates the extraction of diverse scale features. Features are extracted concurrently through multiple branches, each utilizing convolutional kernels of varying sizes to capture features with different perceptual fields. The input channel dimension for each branch is denoted as D. The dimension of the feature channels of the output feature maps F_i for each branch at different scales is $D^{'} = \frac{D}{S}, i = 0, 1, 2, \cdots, S-1$; D should be divisible by S. However, as the size of the convolution kernel increases, the corresponding number of parameters also increases. Therefore, the group convolution method is used to control the corresponding computational cost and ensure the timeliness of feature processing. The group size g dynamically adjusts in response to the multi-scale convolution kernel size. The correlation between them can be expressed as follows:

$$g = 2^{\frac{k-1}{2}} \tag{3}$$

where k is the kernel size, and g is the group number. Moreover, the padding operations are used to guarantee that each branch produces output feature maps with uniform dimensions. The output feature maps from each branch F_i are illustrated by the following equation:

$$F_i = \mathrm{Conv}(k_i \times k_i, g_i)(G_{concat}), i = 0, 1, 2, \cdots, S-1 \tag{4}$$

where the convolution kernel size is denoted as $k_i \times k_i$, with $k_i \times k_i = (2i+3) \times (2i+3)$.

The SE block is used to extract channel attention weights for multi-scale feature maps. The recalibration of attention weights for feature maps at various scales using the softmax operation facilitates the integration of local and global information. Following this, the feature vectors and attention weights are combined with the feature maps at the corresponding scales to construct a comprehensive feature representation. The specific formula is as follows:

$$T_{concat} = \mathrm{Concat}(F_i \odot \mathrm{Softmax}(\mathrm{SE}(F_i))), i = 0, 1, 2, \cdots, S-1 \tag{5}$$

where T_{concat} is the output feature map with multi-scale channel attention weights, and $\odot$ represents the channel-wise multiplication operation.

As shown in Figure 2, an SE block comprises two steps: squeeze and excitation, respectively. It is calculated as follows. The input feature map $G \in \mathbb{R}^{h \times w \times c}$ is defined, where h, w, and c represent the height, width, and number of input channels, respectively. Subsequently, a global average pooling operation is conducted on the input feature map to compute the average feature values across each channel, resulting in a scalar value. The global average pooling operator is computed as follows:

$$z = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} x(i, j) \tag{6}$$

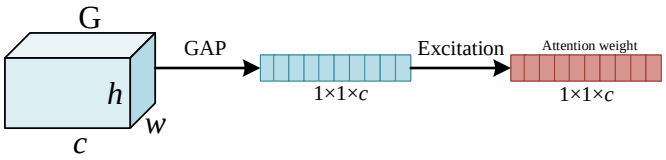


Figure 2. Illustration of the SE block.

The resulting z feature tensor is input into a fully connected layer (usually containing a ReLU activation function) to learn the importance weights for each channel. The output of the fully connected layer is then passed through a Sigmoid function that limits the value of the weights learned on each channel to between 0 and 1, which is used to adjust the feature response of each channel. The output y in the SE block can be written as

$$y = \sigma(w_2\delta(w_1(z))) \tag{7}$$

where δ represents the ReLU activation function, and w_1 and w_2 represent the weight matrices of the fully connected layer. The symbol σ represents the Sigmoid activation function. Employing the activation function enables the assignment of weights to post-channel interactions, thereby enhancing the efficiency of information extraction.

Finally, the obtained multi-scale features are channel mapped by 2D point-by-point convolution to increase the number of channels and the input feature map is summed with the enhanced features to obtain the final output features.

2.4. Transformer Encoder

The Transformer encoder (TE) module is primarily utilized for capturing profound long-range correlations among image features. It comprises a multi-head self-attention (MHSA) block (Figure 3), a Multilayer Perceptron (MLP) block, and two normalization layers. Residual connections are incorporated before both the MHSA block and the MLP block.

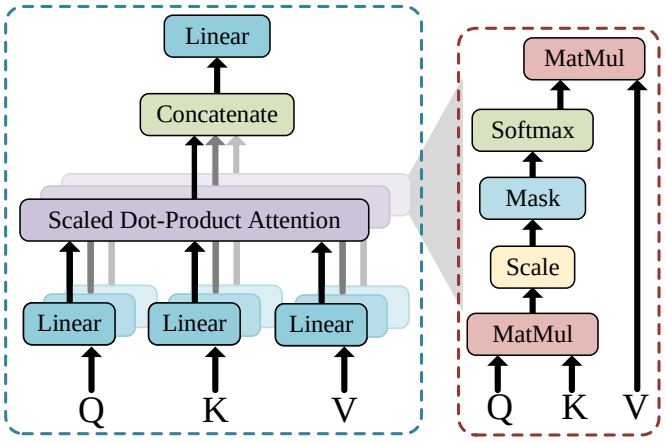


Figure 3. Illustration of the MHSA block.

The TE is recognized for its central MHSA block. This block employs the self-attention (SA) mechanism to effectively capture feature correlations. To enable learning of diverse semantics, three trainable weight matrices, $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$, are pre-defined. These matrices linearly map the feature tokens into three-dimensional invariant representations, comprising query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. The attention scores are calculated using all $\boldsymbol{Q}$ and $\boldsymbol{K}$, and their weights are determined through the softmax function. In conclusion, the self-attention mechanism is elucidated as follows:

$$\mathrm{SA} = \mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V} \tag{8}$$

where softmax is the activation function used to calculate weights, and d_K is the key's dimension.

The MHSA block utilizes various sets of weight matrices to map the query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$, calculating the multi-head attention values using the same operational procedure. Following this, the outputs from each attention head are concatenated. This procedure is represented by the equation below:

$$\mathrm{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}(\mathrm{SA}_1, \mathrm{SA}_2, \ldots, \mathrm{SA}_h)\boldsymbol{W} \tag{9}$$

where h is the head number, $\boldsymbol{W}$ is a linear transformation matrix, $\boldsymbol{W} \in \mathbb{R}^{h \times d_K \times d_w}$, where $d_w = w$ (number of feature channels).

Subsequently, the weight matrix generated by the MHSA is passed through the MLP block, which comprises two fully connected layers. An interposed Gaussian Error Linear Unit (GELU) nonlinear activation function exists between these layers. Following the MLP block, a normalization layer is applied, aiding in mitigating the gradient explosion, addressing gradient vanishing, and accelerating the training process.

After undergoing the aforementioned steps, the final feature tokens are derived, with the initial embedding learned classification token *cls* serving as a classification vector. This vector is then input into a linear classifier based on the softmax function to acquire the ultimate category information. The comprehensive procedure of the proposed method is outlined in Algorithm 1.

Algorithm 1 The proposed MSCF-MAM method.

Input: HSI data $\boldsymbol{X} \in \mathbb{R}^{m \times n \times d}$, ground truth $\boldsymbol{Y} \in \mathbb{R}^{m \times n}$, batch size = 64, the PCA bands $b = 30$, patch size $s = 13$, epoch number $e = 100$, learning rate = $1e-3$, and the training sample rate $\mu\% = 0.1\%$.

Output: Predicted labels of the test dataset.

1: Perform PCA transformation to obtain $\boldsymbol{X_{pca}}$.
2: Create sample patches from $\boldsymbol{X_{pca}}$, then partition them into training and testing sets.
3: **for** $i = 1$ to e **do**
4: Perform multi-scale 3D convolution block to obtain multi-scale convolutional features.
5: Flatten multi-scale 3D convolutional features into 2D feature maps.
6: Perform feature enhancement using PSA through Equations (4) and (5).
7: Concatenate the learnable tokens to create feature tokens and embed position information into the tokens.
8: Perform Transformer encoder module using Equations (8) and (9).
9: Input the first *cls* token into the linear layer.
10: Use the softmax function to identify the labels.
11: **end for**
12: Utilize the test dataset along with the trained model to obtain predicted labels.

3. Experiments and Results

Three publicly available HSI datasets are employed in the experiments to evaluate the efficacy of the proposed MSCF-MAM method. Firstly, the specifics of the three datasets employed in the experiments are outlined. Subsequently, a detailed description of the experimental configuration parameters was provided. The quantitative and visual results of the classification experiment are then exhibited to demonstrate the effectiveness and advantages of the proposed method. Finally, parameter and ablation analysis experiments are carried out to assess the influence of various parameters and modules on the model performance, accompanied by additional comparative time analysis to showcase the efficiency of the proposed method.

3.1. Dataset Description

In our experiments, three publicly available HSI datasets from different imaging platforms are used to validate the designed method's effectiveness. Detailed descriptions of the three datasets are given below.

(1) The Salinas Valley (SV) dataset was collected over Salinas Valley, California, utilizing the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor. The dataset comprises 224 spectral bands spanning from 0.4 to 2.5 µm, with a spatial resolution of 3.7 m. For the experiment, we discarded 20 water-absorbing bands and chose to use the 204 band for the experiment. The area covered consists of 512 × 217 pixels and comprises 16 categories of land cover. Figure 4 shows the pseudo-color image, ground truth map, and the land cover categories' names and corresponding colors, respectively.

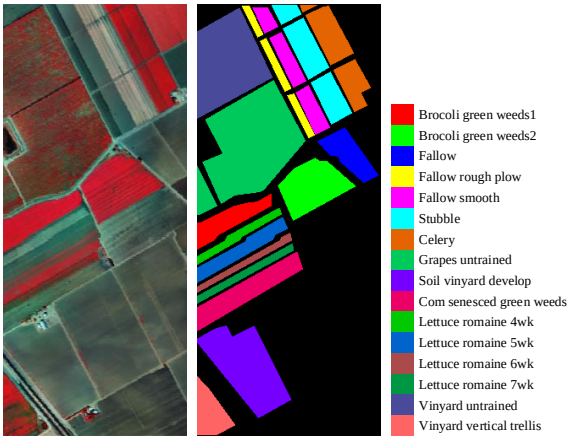


Figure 4. SV dataset. (**a**) Pseudo-color image. (**b**) Ground truth map.

(2) The WHU-Hi-HanChuan (HC) dataset was acquired in Hanchuan, Hubei province, China, with a 17 mm focal length Headwall Nano-Hyperspec imaging sensor equipped on a Leica Aibot X6 UAV V1 platform [63]. The dataset comprises 274 spectral bands spanning from 0.4 to 1 µm. The image spatial dimensions are 1217 × 303 pixels, having a spatial resolution of 0.109 m and encompassing a total of 16 land cover categories. Figure 5 provides the pseudo-color image and ground truth map alongside the names of land cover categories and their corresponding colors.

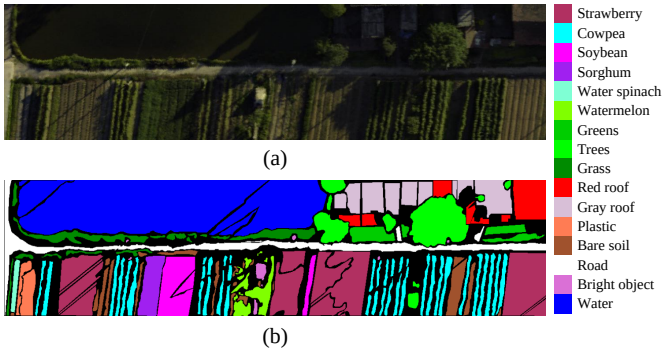


Figure 5. HC dataset. (**a**) Pseudo-color image. (**b**) Ground truth map.

(3) The WHU-Hi-HongHu (HH) dataset was obtained in Honghu City, Hubei Province, China. The dataset was captured using a DJI Matrice 600 Pro UAV platform equipped with a 17 mm focal length Headwall Nano-Hyperspec imaging sensor [64]. It contains 270 bands ranging from 0.4 to 1 µm, and the spatial resolution of the UAV-borne HSI dataset is 0.043 m. The size of this dataset is 940 × 475 pixels, and 22 land cover categories are

covered. Figure 6 displays the pseudo-color image, ground truth map, and the names of land cover categories along with their respective colors.

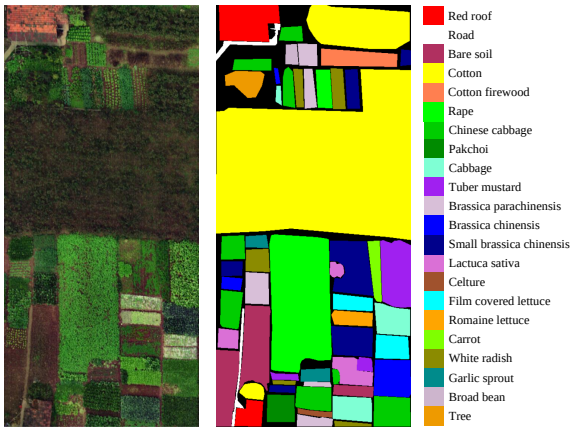


Figure 6. HH dataset. (**a**) Pseudo-color image. (**b**) Ground truth map.

Table 1 lists the land cover categories' names along with the corresponding counts of training and test samples for the three datasets. In each dataset, 0.1% of the total samples were randomly allocated to the training set, while the remainder formed the test set.

Table 1. Training and test sample numbers of each category for the SV, HC, and HH datasets.

NO.	SV			HC			HH		
	Class Name	**Training**	**Test**	**Class Name**	**Training**	**Test**	**Class Name**	**Training**	**Test**
C01	Brocoli green weeds1	2	2007	Strawberry	45	44,690	Red roof	14	14,027
C02	Brocoli green weeds2	4	3722	Cowpea	23	22,730	Road	4	3508
C03	Fallow	2	1974	Soybean	10	10,277	Bare soil	22	21,799
C04	Fallow rough plow	1	1393	Sorghum	5	5348	Cotton	163	163,122
C05	Fallow smooth	3	2675	Water spinach	1	1199	Cotton firewood	6	6212
C06	Stubble	4	3955	Watermelon	5	4528	Rape	45	44,512
C07	Celery	4	3575	Greens	6	5897	Chinese cabbage	24	24,079
C08	Grapes untrained	11	11,260	Trees	18	17,960	Pakchoi	4	4050
C09	Soil vinyard develop	6	6197	Grass	9	9460	Cabbage	11	10,808
C10	Com senesced green weeds	3	3275	Red roof	10	10,506	Tuber mustard	12	12,382
C11	Lettuce romaine 4 wk	1	1067	Gray roof	17	16,894	Brassica parachinensis	11	11,004
C12	Lettuce romaine 5 wk	2	1925	Plastic	4	3675	Brassica chinensis	9	8945
C13	Lettuce romaine 6 wk	1	915	Bare soil	9	9107	Small brassica chinensis	22	22,485
C14	Lettuce romaine 7 wk	1	1069	Road	19	18,541	Lactuca sativa	7	7349
C15	Vinyard untrained	7	7261	Bright object	1	1135	Celture	1	1001
C16	Vinyard vertical trellis	2	1805	Water	75	75,326	Film covered lettuce	7	7255
C17							Romaine lettuce	3	3007
C18							Carrot	3	3214
C19							White radish	9	8703
C20							Garlic sprout	4	3482
C21							Broad bean	1	1327
C22							Tree	4	4036
-	Total	54	54,075	Total	257	257,273	Total	386	386,307

3.2. Experimental Setting

(1) Evaluation Indicators: Four representative quantitative evaluation indicators, including overall accuracy (OA), average accuracy (AA), kappa coefficient (κ), and accuracy per class, were chosen to evaluate and compare the classification performance of the proposed

method with other comparative methods. Higher values for each indicator denote a more precise classification.

(2) Configuration: The evaluation experiments were performed on the PyTorch platform using a server configured with an Inter Xeon Silver 4314 2.40-GHz CPU, 192 GB RAM, and an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). The Adam optimizer was used to optimize the network. For the training phase, we set the batch size to 64 and the number of training epochs to 100.

3.3. Quantitative and Visual Classification Results

To evaluate the classification performance of the MSCF-MAM method, we compared it with several representative methods. These included the traditional SVM [24] method, CNN-based methods, namely, 1D CNN [43], 3D CNN [46], DFFN [48], and HybridSN [47], and the Transformer-based methods MorghAT [60], GAHT [55], and SSFTT [54]. To be fair, we conducted experiments on all methods using the same settings on all three datasets. We calculated the average values for the 10 sets of experiments and highlighted the optimal results in bold.

(1) Quantitative Analysis: The detailed classification results for the three HSI datasets are presented in Tables 2–4. Overall, the performance of the traditional SVM classifiers is considered average. In addition, the classification accuracy achieved by the earlier CNN-based method is notably higher than that of the traditional method, underscoring the superiority of deep learning approaches. Furthermore, the recently introduced Transformer-based method has shown even more promising results, surpassing the performance of the CNN-based method. For instance, in the SV dataset, despite using only 0.1% of the sample size with some categories having only one sample, our method achieved over 99% classification accuracy in certain categories. This success underscores the effectiveness and robustness of our model, particularly in scenarios with limited samples. In the HC and HH datasets, our method also demonstrates commendable classification accuracy, despite challenges such as sample imbalance and lower accuracy for sparsely distributed categories. In comparison to other existing methods, our results stand out, showcasing the efficacy of our method even in challenging classification scenarios.

Table 2. Classification results obtained by different methods with 0.1% training samples per class for the SV dataset (optimal results are highlighted in bold).

NO.	SVM [24]	1-D CNN [43]	3-D CNN [46]	DFFN [48]	HybirdSN [47]	MorghAT [60]	GAHT [55]	SSFTT [54]	MSCF-MAM
C01	65.25 ± 5.10	62.77 ± 24.44	97.70 ± 1.96	97.12 ± 1.06	**98.23 ± 1.65**	80.72 ± 3.70	99.05 ± 1.26	96.47 ± 1.81	97.82 ± 1.60
C02	86.24 ± 4.27	98.52 ± 3.26	93.18 ± 8.58	90.78 ± 3.00	99.41 ± 0.51	93.23 ± 6.35	99.80 ± 0.27	**100.00 ± 0.00**	**100.00 ± 0.00**
C03	53.70 ± 8.23	54.26 ± 5.98	32.69 ± 5.48	51.00 ± 6.26	83.32 ± 8.59	83.42 ± 6.97	53.45 ± 0.50	99.52 ± 0.46	**99.87 ± 0.14**
C04	84.11 ± 4.07	86.86 ± 9.25	**97.96 ± 1.29**	96.72 ± 2.79	51.54 ± 4.60	93.36 ± 6.18	97.60 ± 1.86	75.78 ± 12.93	85.83 ± 9.70
C05	80.09 ± 6.14	87.48 ± 7.85	**96.95 ± 0.34**	85.44 ± 2.79	94.12 ± 8.15	86.25 ± 9.87	96.10 ± 1.58	93.41 ± 3.85	93.44 ± 2.71
C06	97.09 ± 0.86	99.40 ± 0.22	98.93 ± 0.75	84.32 ± 6.57	94.55 ± 3.14	97.88 ± 1.97	99.90 ± 0.17	**99.99 ± 0.02**	98.68 ± 1.39
C07	93.68 ± 1.47	98.89 ± 0.27	99.69 ± 0.23	81.40 ± 4.67	**100.00 ± 0.00**	96.26 ± 0.66	98.40 ± 0.46	99.87 ± 0.24	99.99 ± 0.02
C08	50.80 ± 7.64	55.63 ± 4.29	77.16 ± 12.58	70.75 ± 9.13	86.61 ± 3.67	77.16 ± 7.95	83.31 ± 3.28	85.04 ± 0.48	**88.25 ± 2.46**
C09	97.32 ± 1.23	97.04 ± 2.45	99.95 ± 0.05	95.39 ± 2.66	98.94 ± 1.51	98.08 ± 1.70	99.87 ± 0.23	**100.00 ± 0.00**	99.92 ± 0.13
C10	77.59 ± 6.59	81.57 ± 5.86	29.79 ± 10.26	81.15 ± 1.08	94.12 ± 3.67	85.10 ± 3.48	88.87 ± 2.37	**96.42 ± 0.69**	95.70 ± 1.23
C11	56.10 ± 7.28	57.16 ± 28.62	60.63 ± 4.33	38.43 ± 12.02	46.86 ± 5.77	37.52 ± 7.22	92.02 ± 3.99	98.57 ± 3.03	**99.19 ± 1.75**
C12	92.36 ± 2.96	97.24 ± 2.54	94.91 ± 5.60	95.42 ± 7.93	60.92 ± 6.32	61.05 ± 10.69	**99.98 ± 0.03**	96.32 ± 3.08	99.03 ± 1.12
C13	46.78 ± 4.15	41.33 ± 19.65	50.10 ± 1.86	**94.90 ± 4.10**	83.86 ± 14.04	90.04 ± 1.53	94.27 ± 5.28	87.28 ± 11.22	86.95 ± 8.88
C14	75.97 ± 2.71	81.09 ± 7.66	82.73 ± 6.32	**85.68 ± 2.11**	80.51 ± 6.17	82.30 ± 8.48	66.56 ± 3.61	67.72 ± 16.67	78.99 ± 9.00
C15	49.86 ± 10.52	59.19 ± 5.15	43.84 ± 10.63	71.83 ± 2.15	72.32 ± 16.36	**84.43 ± 1.55**	66.56 ± 3.61	50.14 ± 7.52	61.02 ± 8.02
C16	75.35 ± 3.64	72.82 ± 3.64	47.22 ± 7.33	57.40 ± 2.45	90.66 ± 0.23	47.73 ± 7.41	80.72 ± 2.01	**98.91 ± 0.73**	98.55 ± 0.40
OA(%)	69.52 ± 3.45	75.91 ± 2.00	74.61 ± 1.81	77.70 ± 2.77	85.40 ± 2.53	82.16 ± 0.96	87.94 ± 0.39	87.81 ± 1.05	**90.44 ± 1.86**
AA(%)	72.89 ± 4.04	75.94 ± 4.34	71.87 ± 3.00	77.98 ± 4.58	79.83 ± 8.19	79.24 ± 1.65	89.37 ± 1.74	90.33 ± 1.21	**92.70 ± 1.99**
$\kappa \times 100$	67.15 ± 3.02	73.21 ± 2.22	73.28 ± 2.60	75.15 ± 3.20	83.72 ± 2.85	80.11 ± 1.02	86.54 ± 0.45	86.39 ± 1.18	**89.33 ± 2.08**

Table 3. Classification results obtained by different methods with 0.1% training samples per class for the HC dataset (optimal results are highlighted in bold).

NO.	SVM [24]	1-D CNN [43]	3-D CNN [46]	DFFN [48]	HybirdSN [47]	MorghAT [60]	GAHT [55]	SSFTT [54]	MSCF-MAM
C01	92.73 ± 0.57	77.58 ± 8.49	82.08 ± 7.40	78.05 ± 10.67	**97.69 ± 1.77**	94.05 ± 2.23	90.97 ± 3.01	94.29 ± 2.28	95.79 ± 3.32
C02	53.70 ± 9.23	58.08 ± 4.92	65.86 ± 4.83	50.08 ± 2.08	82.56 ± 3.46	80.37 ± 1.19	85.41 ± 4.28	86.77 ± 6.65	**92.69 ± 3.47**
C03	60.10 ± 5.16	64.28 ± 3.21	58.76 ± 2.39	55.78 ± 7.58	67.54 ± 2.00	67.67 ± 8.22	69.79 ± 4.26	**79.79 ± 5.03**	72.83 ± 4.49
C04	46.54 ± 14.01	40.75 ± 11.11	41.12 ± 14.21	49.89 ± 6.54	**99.08 ± 0.75**	72.63 ± 7.86	91.30 ± 6.91	92.53 ± 6.76	94.49 ± 2.74
C05	15.67 ± 9.06	14.65 ± 8.69	15.48 ± 3.38	22.89 ± 5.01	21.18 ± 8.69	21.07 ± 4.42	17.29 ± 5.31	**26.56 ± 11.93**	23.86 ± 2.33
C06	13.21 ± 4.08	12.74 ± 3.96	16.71 ± 2.71	10.72 ± 3.63	33.83 ± 2.67	31.99 ± 3.29	31.56 ± 4.50	47.91 ± 8.81	**50.68 ± 2.10**
C07	65.77 ± 5.27	61.49 ± 4.68	53.29 ± 3.44	61.87 ± 7.68	64.53 ± 0.24	80.43 ± 4.35	67.46 ± 3.89	**82.99 ± 3.76**	73.93 ± 4.59
C08	45.72 ± 5.25	46.82 ± 3.02	55.69 ± 2.29	72.24 ± 4.98	74.08 ± 2.28	74.35 ± 3.06	72.42 ± 1.28	67.40 ± 9.73	**75.96 ± 7.06**
C09	16.86 ± 12.37	16.97 ± 15.19	17.71 ± 4.29	17.36 ± 4.93	43.63 ± 1.85	25.42 ± 1.96	33.04 ± 7.92	60.17 ± 5.40	**63.79 ± 2.66**
C10	26.21 ± 6.29	25.98 ± 4.13	29.71 ± 4.61	33.31 ± 5.73	82.08 ± 4.16	53.95 ± 9.60	87.78 ± 10.65	91.21 ± 5.11	**95.80 ± 1.68**
C11	25.17 ± 6.32	41.70 ± 11.05	48.94 ± 5.44	44.38 ± 12.19	85.30 ± 4.86	90.55 ± 10.41	**95.59 ± 0.75**	94.78 ± 3.49	89.07 ± 3.86
C12	12.10 ± 4.47	10.09 ± 5.27	12.64 ± 3.27	15.26 ± 3.32	31.84 ± 0.97	38.37 ± 9.05	42.62 ± 7.76	45.91 ± 1.92	**52.22 ± 6.36**
C13	13.51 ± 6.78	14.62 ± 5.84	14.82 ± 3.27	20.00 ± 5.85	38.54 ± 4.03	51.03 ± 11.41	**51.60 ± 7.08**	42.51 ± 6.42	46.51 ± 3.14
C14	74.04 ± 4.21	69.38 ± 3.94	71.58 ± 2.98	71.60 ± 2.84	77.04 ± 5.11	72.19 ± 11.77	**88.72 ± 2.11**	86.51 ± 5.71	88.71 ± 4.82
C15	11.19 ± 15.11	12.62 ± 5.24	38.01 ± 6.75	**57.17 ± 8.79**	44.11 ± 1.98	24.94 ± 9.03	10.29 ± 4.54	29.91 ± 7.96	37.94 ± 3.67
C16	98.06 ± 0.64	92.53 ± 7.22	99.54 ± 0.70	98.26 ± 1.00	**99.58 ± 0.26**	98.64 ± 0.06	99.04 ± 0.42	98.51 ± 1.10	99.40 ± 0.49
OA(%)	60.27 ± 3.28	57.41 ± 3.01	65.06 ± 1.19	62.86 ± 3.95	83.06 ± 1.04	77.09 ± 1.15	83.19 ± 1.71	85.80 ± 1.40	**86.94 ± 0.47**
AA(%)	40.37 ± 9.54	39.42 ± 3.53	53.51 ± 5.54	48.04 ± 4.21	62.04 ± 1.13	52.91 ± 2.41	60.95 ± 1.92	67.67 ± 2.12	**68.04 ± 1.40**
$\kappa \times 100$	53.93 ± 6.11	49.49 ± 3.44	58.05 ± 3.09	55.48 ± 5.33	80.06 ± 1.23	72.98 ± 1.50	80.29 ± 2.00	83.35 ± 1.61	**84.65 ± 0.57**

Table 4. Classification results obtained by different methods with 0.1% training samples per class for the HH dataset (optimal results are highlighted in bold).

NO.	SVM [24]	1-D CNN [43]	3-D CNN [46]	DFFN [48]	HybirdSN [47]	MorghAT [60]	GAHT [55]	SSFTT [54]	MSCF-MAM
C01	74.93 ± 2.42	55.11 ± 10.49	79.64 ± 2.65	82.58 ± 11.98	91.35 ± 3.18	92.35 ± 3.53	95.81 ± 1.09	86.95 ± 5.70	**96.93 ± 0.59**
C02	33.81 ± 11.23	18.65 ± 16.10	56.64 ± 6.45	33.67 ± 5.66	36.95 ± 5.34	45.44 ± 9.67	54.84 ± 9.60	59.96 ± 10.47	**61.92 ± 8.05**
C03	93.24 ± 2.16	91.11 ± 2.49	92.71 ± 1.56	**96.32 ± 0.77**	87.95 ± 2.31	92.99 ± 0.74	89.16 ± 3.94	91.50 ± 3.56	93.43 ± 3.95
C04	98.61 ± 1.01	96.44 ± 0.66	99.28 ± 0.14	**99.72 ± 0.11**	98.87 ± 0.89	98.32 ± 0.67	99.33 ± 0.32	99.02 ± 0.32	98.42 ± 1.77
C05	25.67 ± 10.04	19.43 ± 6.83	25.48 ± 4.38	22.76 ± 4.76	42.22 ± 5.85	51.07 ± 14.42	46.16 ± 12.27	**62.80 ± 4.48**	56.84 ± 6.85
C06	87.75 ± 0.92	82.44 ± 6.19	88.11 ± 1.35	88.31 ± 0.95	92.65 ± 0.58	**96.11 ± 1.32**	91.57 ± 2.91	93.13 ± 4.02	93.98 ± 1.02
C07	75.46 ± 3.27	53.65 ± 13.73	80.44 ± 3.05	66.96 ± 12.94	70.37 ± 2.92	76.99 ± 4.68	78.14 ± 2.29	85.52 ± 4.82	**92.98 ± 2.77**
C08	35.16 ± 9.25	29.91 ± 4.96	30.80 ± 10.21	30.06 ± 9.05	36.29 ± 5.15	37.32 ± 4.67	**45.25 ± 3.97**	34.02 ± 3.24	34.32 ± 4.99
C09	83.21 ± 4.16	62.99 ± 17.45	86.64 ± 1.88	84.63 ± 4.37	90.43 ± 2.64	94.45 ± 0.75	**98.99 ± 0.50**	91.90 ± 3.58	96.12 ± 1.45
C10	10.31 ± 8.29	9.79 ± 4.18	49.51 ± 9.50	44.77 ± 4.46	50.07 ± 7.87	70.02 ± 7.49	79.07 ± 7.18	84.89 ± 3.18	**87.84 ± 1.72**
C11	14.47 ± 14.32	16.53 ± 15.19	17.38 ± 3.94	16.00 ± 12.52	50.62 ± 10.83	51.03 ± 8.04	46.48 ± 10.84	66.00 ± 3.88	**68.03 ± 7.70**
C12	50.48 ± 8.47	26.08 ± 10.44	52.64 ± 4.64	47.23 ± 3.45	58.78 ± 10.41	55.59 ± 7.25	**65.06 ± 2.24**	58.39 ± 7.43	58.95 ± 4.28
C13	56.19 ± 6.28	46.49 ± 5.88	68.71 ± 0.44	65.89 ± 6.13	**78.81 ± 2.93**	72.39 ± 6.31	75.36 ± 4.71	70.43 ± 10.58	72.40 ± 5.30
C14	36.89 ± 6.15	31.37 ± 11.33	30.91 ± 13.94	44.36 ± 12.92	60.83 ± 5.28	66.45 ± 7.12	**78.94 ± 4.31**	72.61 ± 5.27	73.45 ± 9.92
C15	11.23 ± 10.04	15.11 ± 9.13	12.17 ± 8.29	10.54 ± 11.05	45.50 ± 3.68	41.93 ± 11.61	41.17 ± 3.61	50.07 ± 4.09	**56.85 ± 3.00**
C16	83.89 ± 3.64	85.19 ± 11.56	90.33 ± 1.50	89.04 ± 3.08	80.50 ± 9.70	88.30 ± 7.18	**95.96 ± 1.65**	88.74 ± 4.59	90.59 ± 2.03
C17	23.58 ± 11.03	12.95 ± 8.63	16.60 ± 2.94	16.40 ± 4.03	34.94 ± 6.54	40.44 ± 3.90	40.16 ± 2.27	40.15 ± 4.22	**48.09 ± 10.79**
C18	16.54 ± 10.24	10.34 ± 5.59	10.27 ± 10.47	17.40 ± 6.45	34.60 ± 10.29	46.56 ± 7.33	36.91 ± 1.62	52.33 ± 7.34	**65.50 ± 4.93**
C19	45.18 ± 6.35	23.84 ± 7.03	49.69 ± 10.32	15.11 ± 9.84	70.33 ± 3.54	74.41 ± 11.44	73.09 ± 10.91	**81.00 ± 8.41**	79.61 ± 5.31
C20	14.71 ± 8.66	12.40 ± 13.70	28.04 ± 14.29	22.89 ± 15.53	28.44 ± 4.75	75.33 ± 4.46	**81.01 ± 4.29**	43.28 ± 3.29	50.27 ± 2.27
C21	13.90 ± 13.45	10.10 ± 8.17	12.64 ± 4.57	17.29 ± 12.14	19.82 ± 5.89	21.71 ± 2.55	19.86 ± 10.87	24.24 ± 3.56	**30.30 ± 2.65**
C22	21.54 ± 12.36	15.19 ± 13.78	24.64 ± 8.36	25.21 ± 12.47	56.55 ± 5.18	51.84 ± 8.15	69.87 ± 10.53	**73.47 ± 3.48**	72.88 ± 3.00
OA(%)	73.52 ± 5.28	69.52 ± 2.92	79.09 ± 1.19	77.38 ± 1.73	82.44 ± 0.76	85.09 ± 0.78	86.34 ± 0.62	86.74 ± 1.00	**87.87 ± 0.60**
AA(%)	45.52 ± 9.04	42.77 ± 6.82	54.91 ± 3.42	51.90 ± 3.54	55.77 ± 2.47	59.14 ± 1.60	62.16 ± 2.16	64.29 ± 1.54	**65.99 ± 0.48**
$\kappa \times 100$	67.90 ± 5.02	60.42 ± 4.80	73.05 ± 2.28	75.15 ± 3.20	77.68 ± 0.95	81.07 ± 0.96	82.66 ± 0.79	83.05 ± 1.44	**84.55 ± 0.80**

(2) Visual Evaluation: The classification result maps of the method presented in this paper and several comparison methods on the three datasets are illustrated in Figures 7–9. The comparison of the spatial and edge clarity in these classification result maps reveals that the method's classifications align closely with the ground truth maps. Nonetheless, the utilization of only 0.1% of the training samples results in some level of noise in the classification result maps obtained. Traditional classifiers typically exhibit a lower effectiveness than deep learning classifiers, evident from their higher misclassification rates and prevalent noise in classification maps. For conventional SVM classifiers and several CNN-based methods, the classification outcomes are comparatively inferior, displaying extensive noise distribution areas and an inadequate representation of feature class boundaries. In contrast, Transformer-based techniques like GAHT and SSFTT produce cleaner classification result maps that distinctly delineate the boundaries of various feature classes by considering the

global correlation information of features. The proposed method in this paper demonstrates minimal misclassification among diverse classification results, maintaining superior edge information and affirming its optimal classification performance at a 0.1% sampling rate.

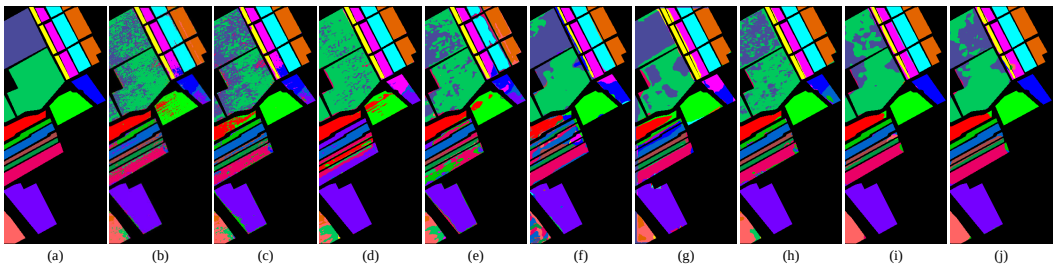


Figure 7. Classification maps of different methods for the SV dataset with 0.1% training samples per class. (**a**) Ground truth map. (**b**) SVM [24]. (**c**) 1-D CNN [43]. (**d**) 3-D CNN [46]. (**e**) DFFN [48]. (**f**) HybridSN [47]. (**g**) MorghAT [60]. (**h**) GAHT [55]. (**i**) SSFTT [54]. (**j**) MSCF-MAM.

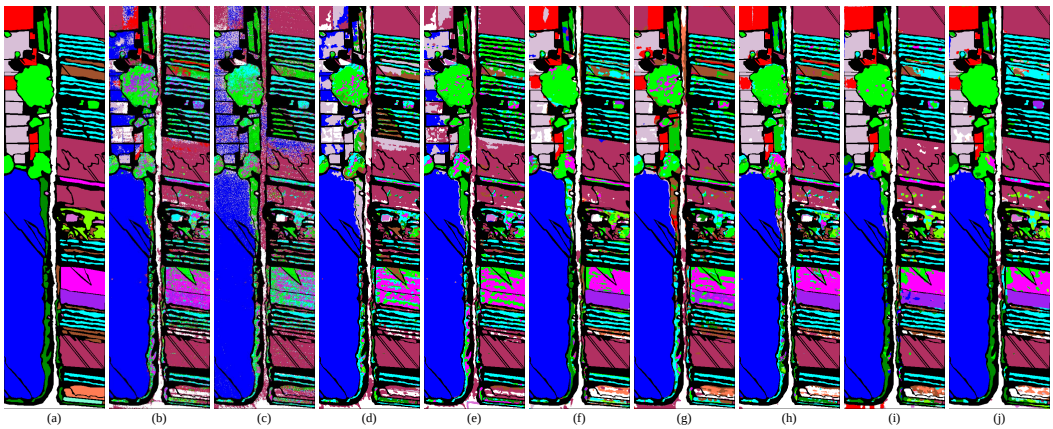


Figure 8. Classification maps obtained by different methods with 0.1% training samples per class for the HC dataset. (**a**) Ground truth map. (**b**) SVM [24]. (**c**) 1-D CNN [43]. (**d**) 3-D CNN [46]. (**e**) DFFN [48]. (**f**) HybridSN [47]. (**g**) MorghAT [60]. (**h**) GAHT [55]. (**i**) SSFTT [54]. (**j**) MSCF-MAM.

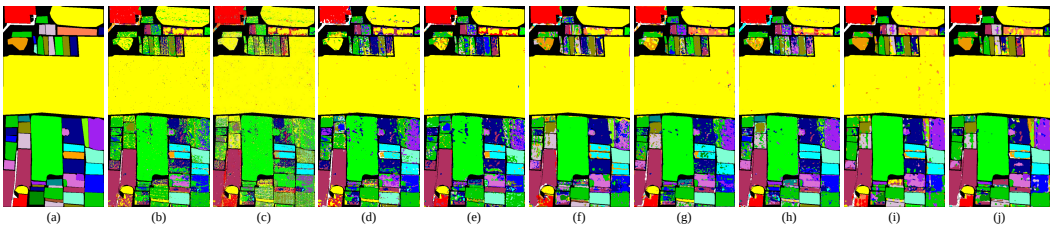


Figure 9. Classification maps of different methods for the HH dataset with 0.1% training samples per class. (**a**) Ground truth map. (**b**) SVM [24]. (**c**) 1-D CNN [43]. (**d**) 3-D CNN [46]. (**e**) DFFN [48]. (**f**) HybridSN [47]. (**g**) MorghAT [60]. (**h**) GAHT [55]. (**i**) SSFTT [54]. (**j**) MSCF-MAM.

3.4. Parameter Analysis

We have analyzed several crucial hyper-parameters that significantly influence the classification accuracy, including the patch size (s) and the number of PCA bands (b).

(1) Patch Size: We select the patch size (s) from the candidate set $\{9, 11, 13, 15, 17\}$ for evaluation. The impact of various patch sizes (s) on OA, AA, and κ is depicted in Figure 10. The optimal classification accuracy is observed when the patch size is set to either 11 or 13 in the three datasets.

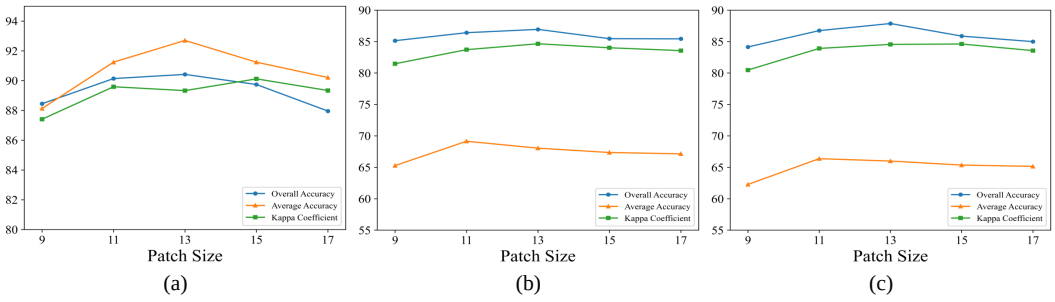


Figure 10. The impact of different patch sizes on OA, AA, and κ. (**a**) SV dataset. (**b**) HC dataset. (**c**) HH dataset.

(2) PCA Band: To assess the impact of PCA band numbers (b), we choose the parameter b from the set of candidates $\{5, 10, 15, 20, 25, 30\}$. Figure 11 illustrates how the OA, AA, and κ of the proposed method change with varying PCA band numbers across the three datasets. The three indicators show an initial increase followed by stabilization as b increases. Taking into account the computational cost and classification performance, we set b to 30.

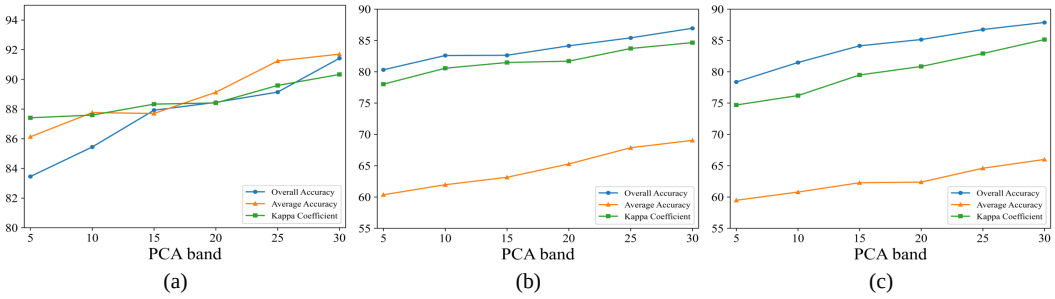


Figure 11. The impact of different PCA bands on OA, AA, and κ. (**a**) SV dataset. (**b**) HC dataset. (**c**) HH dataset.

3.5. Ablation Analysis

To thoroughly validate the proposed method, we performed ablation experiments on the SV dataset to analyze the influence of various modules on the network's overall performance. Four module combinations were assessed using the OA, AA, and k values, as presented in Table 5. The model comprises three modules: the multi-scale feature extraction module with three 3D convolutional blocks of varying scales, the feature enhancement module utilizing the PSA structure, and the TE module. The results from Table 5 show that employing a single 3D convolutional block for spatial–spectral feature extraction reduced the OA and AA by 2.70% and 4.57% compared to the full model results. Conversely, utilizing two 3D convolutional blocks of different scales in parallel led to a 1.3% improvement in the OA and a 2.30% increase in the AA. This underscores the benefit of extracting multi-scale convolutional features for enhancing the model's classification performance. The third combined experiment focused on evaluating the impact of the PSA-based feature enhancement module on classification performance. Removing the PSA structure caused a 3.66% decrease in the OA and a substantial 5.49% decline in the AA, highlighting the crucial role of the feature enhancement module in the model. Furthermore, eliminating the TE module in the final experiment resulted in a 1.21% OA decrease, affirming the affirmative impact of the TE structure on performance improvement. Overall, a detailed analysis of the results of these ablation experiments further confirms the efficacy of this paper's method.

Table 5. Ablation analysis of the proposed method with 0.1% training samples per class for the SV dataset (optimal results are highlighted in bold).

Cases	Component					Indicators		
	3D Conv1	*3D Conv2*	*3D Conv3*	PSA	TE	OA (%)	AA (%)	$\kappa \times 100$
1	√	×	×	√	√	87.74	88.13	87.58
2	√	√	×	√	√	89.12	90.43	88.71
3	√	√	√	×	√	86.78	87.21	86.58
4	√	√	√	√	×	89.23	90.33	88.24
5	√	√	√	√	√	**90.44**	**92.70**	**89.33**

3.6. Comparison of Computational Efficiency

Table 6 presents a comparative analysis of the computational efficiency for the 1D CNN [43], 3D CNN [46], DFFN [48], HybridSN [47], MorghAT [60], GAHT [55], SSFTT [54], and our proposed MSCF-MAM method. It is evident that, with the exceptions of the 1D CNN and SSFTT, our proposed method demonstrates a relatively higher speed compared to the other methods. Notably, our method incurs significantly lower time costs when contrasted with DFFN, MorghAT, and GAHT. These findings underscore the effectiveness of our method in reducing the computational time and enhancing the classification efficiency. At the same time, there are several comparative methods of FLOP values and the number of parameters can also be analyzed to obtain, due to the design of the feature enhancement module, the addition of a multi-group convolution operation. This paper's method, in terms of the computational complexity, is better than the SSFTT, and at the same time, the number of parameters is also increased by a factor of two, but still can maintain a better performance.

Table 6. Comparison of computational efficiency between the proposed method and the comparison methods on the three datasets (optimal results are highlighted in bold, our method's results are underlined).

Methods	SV		HC		HH		FLOPs (M)	Params (M)
	Training (s)	Test (s)	Training (s)	Test (s)	Training (s)	Test (s)		
1-D CNN [43]	4.17	5.72	**3.45**	**7.68**	**4.84**	**11.69**	**277.46**	**0.12**
3-D CNN [46]	8.53	4.97	40.49	21.50	58.37	25.76	1728.97	0.16
DFFN [48]	15.90	10.01	74.63	44.05	124.21	54.86	1328.13	0.42
HybridSN [47]	6.65	9.18	10.08	19.41	13.57	38.82	3252.56	4.84
MorghAT [60]	26.33	23.38	95.46	37.63	136.51	50.34	2725.21	0.20
GAHT [55]	17.02	10.05	77.98	39.25	110.22	47.58	3046.75	0.97
SSFTT [54]	**1.93**	**3.36**	8.65	19.32	5.64	15.75	781.38	0.16
MSCF-MAM	3.26	4.98	11.64	21.94	12.06	26.20	2551.71	0.59

4. Conclusions

In this paper, an HSI classification method using multi-scale convolutional features and a multi-attention mechanism is proposed, which aims to comprehensively capture multi-scale spatial–spectral features by fusing multi-scale information and introducing a multi-attention mechanism and can process HSI data more efficiently to significantly improve the classification accuracy in the case of limited data samples. Multiple CNNs operating at diverse scales are leveraged for parallel feature extraction from the HSI data. Additionally, a feature enhancement module based on the PSA structure is incorporated to significantly enrich the feature representation. By introducing this feature enhancement module, the input of image features into the Transformer encoder structure contributes to a deeper understanding of the HSI data, thereby improving the overall performance of the model. This approach promotes a more comprehensive understanding of the HSI data, which in turn improves the accuracy and generalization of the classification model. The remarkable classification performance of the proposed method is confirmed by extensive experiments on three HSI datasets. Prospective research directions may involve

the refinement of lightweight Transformer network architecture and the amalgamation of graph convolutional networks to collectively extract spatial and spectral features, aiming to deepen the exploration of abundant feature details in HSIs and elevate the classification efficacy and precision.

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

Funding: This study was supported by the National Natural Science Foundation of China under Grant 62076137, and Grant 61931004; in part by the Startup Foundation for Introducing Talent of NUIST, Grant number 2022r075; in part by the Natural Science Foundation of Jiangsu Province under Grant BK 20211539; and in part by the Jiangsu Innovation & Entrepreneurship Group Talents Plan.

Data Availability Statement: The datasets presented in this paper can be obtained through https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 4 March 2024) and http://rsidea.whu.edu.cn/resource_WHUHi_sharing.htm (accessed on 4 March 2024).

Acknowledgments: The authors would like to thank the National Natural Science Foundation of China and the Natural Science Foundation of Jiangsu Province for supporting our work and thank the anonymous reviewers and the editors for their insightful comments and helpful suggestions for improving the quality of our manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

HSI	Hyperspectral image
CNN	Convolutional neural network
MHSA	Multi-head self-attention
PSA	Pyramid squeeze attention
ML	Machine learning
PCA	Principal component analysis
SVM	Support vector machine
MP	Morphological profile
DL	Deep learning
TE	Transformer encoder
GCN	Graph convolutional network
GAN	Generative adversarial network
DFFN	Deep feature fusion network
SE	Squeeze and excitation
κ	Kappa coefficient
OA	Overall accuracy
AA	Average accuracy

References

1. Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
2. Wu, Z.; Sun, J.; Zhang, Y.; Wei, Z.; Chanussot, J. Recent Developments in Parallel and Distributed Computing for Remotely Sensed Big Data Processing. *Proc. IEEE* **2021**, *109*, 1282–1305. [CrossRef]
3. 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]
4. Gevaert, C.M.; Suomalainen, J.; Tang, J.; Kooistra, L. Generation of Spectral–Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3140–3146. [CrossRef]
5. Khanal, S.; Kc, K.; Fulton, J.P.; Shearer, S.; Ozkan, E. Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. *Remote Sens.* **2020**, *12*, 3783. [CrossRef]

6. Soppa, M.A.; Silva, B.; Steinmetz, F.; Keith, D.; Scheffler, D.; Bohn, N.; Bracher, A. Assessment of Polymer Atmospheric Correction Algorithm for Hyperspectral Remote Sensing Imagery over Coastal Waters. *Sensors* **2021**, *21*, 4125. [CrossRef]
7. Shirmard, H.; Farahbakhsh, E.; Müller, R.D.; Chandra, R. A Review of Machine Learning in Processing Remote Sensing Data for Mineral Exploration. *Remote Sens. Environ.* **2022**, *268*, 112750. [CrossRef]
8. Virtriana, R.; Riqqi, A.; Anggraini, T.S.; Fauzan, K.N.; Ihsan, K.T.N.; Mustika, F.C.; Suwardhi, D.; Harto, A.B.; Sakti, A.D.; Deliar, A.; et al. Development of Spatial Model for Food Security Prediction Using Remote Sensing Data in West Java, Indonesia. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 284. [CrossRef]
9. Dremin, V.; Marcinkevics, Z.; Zherebtsov, E.; Popov, A.; Grabovskis, A.; Kronberga, H.; Geldnere, K.; Doronin, A.; Meglinski, I.; Bykov, A. Skin Complications of Diabetes Mellitus Revealed by Polarized Hyperspectral Imaging and Machine Learning. *IEEE Trans. Med. Imaging* **2021**, *40*, 1207–1216. [CrossRef]
10. Shimoni, M.; Haelterman, R.; Perneel, C. Hypersectral Imaging for Military and Security Applications: Combining Myriad Processing and Sensing Techniques. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 101–117. [CrossRef]
11. Sun, L.; Chen, Y.; Li, B. SISLU-Net: Spatial Information-Assisted Spectral Information Learning Unmixing Network for Hyperspectral Images. *Remote Sens.* **2023**, *15*, 817. [CrossRef]
12. Sun, L.; Wu, F.; He, C.; Zhan, T.; Liu, W.; Zhang, D. Weighted Collaborative Sparse and L1/2 Low-Rank Regularizations with Superpixel Segmentation for Hyperspectral Unmixing. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 5500405. [CrossRef]
13. Sun, L.; Cao, Q.; Chen, Y.; Zheng, Y.; Wu, Z. Mixed Noise Removal for Hyperspectral Images Based on Global Tensor Low-Rankness and Nonlocal SVD-Aided Group Sparsity. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5506617. [CrossRef]
14. Sun, L.; He, C.; Zheng, Y.; Wu, Z.; Jeon, B. Tensor Cascaded-Rank Minimization in Subspace: A Unified Regime for Hyperspectral Image Low-Level Vision. *IEEE Trans. Image Process.* **2022**, *32*, 100–115. [CrossRef] [PubMed]
15. Diao, W.; Zhang, F.; Sun, J.; Xing, Y.; Zhang, K.; Bruzzone, L. ZeRGAN: Zero-Reference GAN for Fusion of Multispectral and Panchromatic Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 8195–8209. [CrossRef] [PubMed]
16. Sun, L.; Cheng, Q.; Chen, Z. Hyperspectral Image Super-Resolution Method Based on Spectral Smoothing Prior and Tensor Tubal Row-Sparse Representation. *Remote Sens.* **2022**, *14*, 2142. [CrossRef]
17. Nasrabadi, N.M. Hyperspectral Target Detection: An Overview of Current and Future Challenges. *IEEE Signal Process. Mag.* **2013**, *31*, 34–44. [CrossRef]
18. 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]
19. Zhao, G.; Ye, Q.; Sun, L.; Wu, Z.; Pan, C.; Jeon, B. Joint Classification of Hyperspectral and Lidar Data Using a Hierarchical CNN and Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5500716. [CrossRef]
20. 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*, 5539217. [CrossRef]
21. Liu, S.; Marinelli, D.; Bruzzone, L.; Bovolo, F. A Review of Change detection in Multitemporal Hyperspectral Images: Current Techniques, Applications, and Challenges. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 140–158. [CrossRef]
22. Licciardi, G.; Marpu, P.R.; Chanussot, J.; Benediktsson, J.A. Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data based on the Extended Morphological Profiles. *IEEE Geosci. Remote Sens. Lett.* **2011**, *9*, 447–451. [CrossRef]
23. 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]
24. Melgani, F.; Bruzzone, L. Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
25. 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]
26. Ma, L.; Crawford, M.M.; Tian, J. Local Manifold Learning-based *K*-Nearest-Neighbor for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4099–4109. [CrossRef]
27. Fauvel, M.; Benediktsson, J.A.; Chanussot, J.; Sveinsson, J.R. Spectral and Spatial Classification of Hyperspectral Data Using SVMs and Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3804–3814. [CrossRef]
28. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of Hyperspectral Data from Urban Areas Based on Extended Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
29. Dalla Mura, M.; Villa, A.; Benediktsson, J.A.; Chanussot, J.; Bruzzone, L. Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 542–546. [CrossRef]
30. Li, W.; Chen, C.; Su, H.; Du, Q. Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [CrossRef]
31. Jia, S.; Shen, L.; Li, Q. Gabor Feature-based Collaborative Representation for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1118–1129. [CrossRef]
32. Fang, S.; Li, X.; Tian, S.; Chen, W.; Zhang, E. Multi-Level Feature Extraction Networks for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 590. [CrossRef]
33. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep Learning-based Classification of Hyperspectral Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2094–2107. [CrossRef]

34. 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]
35. 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]
36. Zhan, Y.; Hu, D.; Wang, Y.; Yu, X. Semisupervised Hyperspectral Image Classification Based on Generative Adversarial Networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *15*, 212–216. [CrossRef]
37. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J.; Pla, F. Capsule Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2145–2160. [CrossRef]
38. 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]
39. Hao, S.; Wang, W.; Salzmann, M. Geometry-Aware Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2448–2460. [CrossRef]
40. Wan, S.; Gong, C.; Zhong, P.; Du, B.; Zhang, L.; Yang, J. Multiscale Dynamic Graph Convolutional Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3162–3177. [CrossRef]
41. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5966–5978. [CrossRef]
42. Wang, Z.; Cao, B.; Liu, J. Hyperspectral Image Classification via Spatial Shuffle-Based Convolutional Neural Network. *Remote Sens.* **2023**, *15*, 3960. [CrossRef]
43. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*. [CrossRef]
44. 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]
45. Zhao, L.; Yi, J.; Li, X.; Hu, W.; Wu, J.; Zhang, G. Compact Band Weighting Module Based on Attention-Driven for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9540–9552. [CrossRef]
46. Hamida, A.B.; Benoit, A.; Lambert, P.; Amar, C.B. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4420–4434. [CrossRef]
47. 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]
48. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral Image Classification with Deep Feature Fusion Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3173–3184. [CrossRef]
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 7 May 2024).
50. 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*, 5516415. [CrossRef]
51. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
52. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
53. 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*, 5518615. [CrossRef]
54. 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]
55. 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]
56. Yang, A.; Li, M.; Ding, Y.; Hong, D.; Lv, Y.; He, Y. GTFN: GCN and transformer fusion with spatial-spectral features for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 6600115. [CrossRef]
57. Feng, J.; Wang, Q.; Zhang, G.; Jia, X.; Yin, J. CAT: Center Attention Transformer with Stratified Spatial-Spectral Token for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5615415. [CrossRef]
58. Zou, J.; He, W.; Zhang, H. Lessformer: Local-enhanced spectral-spatial transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5535416. [CrossRef]
59. Peng, Y.; Zhang, Y.; Tu, B.; Li, Q.; Li, W. Spatial–spectral transformer with cross-attention for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5537415. [CrossRef]
60. Roy, S.K.; Deria, A.; Shah, C.; Haut, J.M.; Du, Q.; Plaza, A. Spectral–Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5503615. [CrossRef]
61. Ouyang, E.; Li, B.; Hu, W.; Zhang, G.; Zhao, L.; Wu, J. When Multigranularity Meets Spatial–Spectral Attention: A Hybrid Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5524916. [CrossRef]

62. Fang, Y.; Ye, Q.; Sun, L.; Zheng, Y.; Wu, Z. Multi-Attention Joint Convolution Feature Representation with Lightweight Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5513814. [CrossRef]
63. Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; Zhang, L. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sens. Environ.* **2020**, *250*, 112012. [CrossRef]
64. Zhong, Y.; Wang, X.; Xu, Y.; Wang, S.; Jia, T.; Hu, X.; Zhao, J.; Wei, L.; Zhang, L. Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications. *IEEE Geosci. Remote Sens. Mag.* **2018**, *6*, 46–62. [CrossRef]

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