applied sciences

MDPI

Article

DCTransformer: A Channel Attention Combined Discrete Cosine Transform to Extract Spatial–Spectral Feature for Hyperspectral Image Classification

Yuanyuan Dang, Xianhe Zhang, Hongwei Zhao and Bing Liu *

College of Computer Science and Engineering, Changchun University of Technology, Changchun 130000, China; dangyuanyuan@ccut.edu.cn (Y.D.); 2202203040@stu.ccut.edu.cn (X.Z.); zhaohw@jlu.edu.cn (H.Z.)
* Correspondence: liubing@ccut.edu.cn

Abstract: Hyperspectral image (HSI) classification tasks have been adopted in huge applications of remote sensing recently. With the rise of deep learning development, it becomes crucial to investigate how to exploit spatial–spectral features. The traditional approach is to stack models that can encode spatial–spectral features, coupling sufficient information as much as possible, before the classification model. However, this sequential stacking tends to cause information redundancy. In this paper, a novel network utilizing the channel attention combined discrete cosine transform (DCTransformer) to extract spatial–spectral features has been proposed to address this issue. It consists of a detail spatial feature extractor (DFE) with CNN blocks and a base spectral feature extractor (BFE) utilizing the channel attention mechanism (CAM) with a discrete cosine transform (DCT). Firstly, the DFE can extract detailed context information using a series of layers of a CNN. Further, the BFE captures spectral features using channel attention and stores the wider frequency information by utilizing the DCT. Ultimately, the dynamic fusion mechanism has been adopted to fuse the detail and base features. Comprehensive experiments show that the DCTransformer achieves a state-of-the-art (SOTA) performance in the HSI classification task, compared to other methods on four datasets, the University of Houston (UH), Indian Pines (IP), MUUFL, and Trento datasets. On the UH dataset, the DCTransformer achieves an OA of 94.40%, AA of 94.89%, and kappa of 93.92.

Keywords: hyperspectral image (HSI) classification; discrete cosine transform (DCT); channel attention mechanism

Citation: Dang, Y.; Zhang, X.; Zhao, H.; Liu, B. DCTransformer: A Channel Attention Combined Discrete Cosine Transform to Extract Spatial–Spectral Feature for Hyperspectral Image Classification. *Appl. Sci.* **2024**, *14*, 1701. https://doi.org/10.3390/app14051701

Academic Editors: Junseop Lee and Atsushi Mase

Received: 11 December 2023
Revised: 21 January 2024
Accepted: 2 February 2024
Published: 20 February 2024

1. Introduction

Hyperspectral images (HSIs) are extensively used in remote sensing applications, such as agriculture [1], environmental monitoring [2,3], and urban planning [4]. A wider range of wavelengths are contained in HSIs compared with traditional images, such as the visible spectrum, as well as near-infrared and short-wave infrared bands. HSIs play a crucial role in the remote sensing (RS) classification task. Refs. [5,6] consider the reason HSIs are fit for the classification task, because of the valuable spectral information in the continuous hundreds of spectral bands of HSIs. Ref. [7] demonstrates that spectral–spatial feature extraction is the key to improving the classification performance.

Algorithms of HSI classification have been introduced, such as the Random Forest [8] and Support Vector Machines (SVMs) [9]. However, these methods still have a weak capacity for spectral–spatial information. With the development of deep learning, CNNs have dominated HSI classification. Ref. [10] suggested the utilization of a 2D-CNN, which can flexibly encode spatial features. Ref. [11] introduced a methodology that combines morphological attributes with 2D-CNN feature extraction methods for HSI classification. The 2D-CNN methods can successfully capture the spatial features, but the 2D-CNN methods cannot extract the spectral features, which mainly contain the category information. It causes an inadequate performance. To tackle this problem, ref. [12] proposed the use of

3D-CNNs, and [13] proposed a Spectral Hierarchical Network (SHN) that utilizes sequential layers of Conv3D and HetConv2D to extract robust and discriminative features from hyperspectral images (HSIs) in the backbone stage. However, the 3D-CNN approaches cannot extract the long-range information due to the restriction of the convolution kernel size, where the long-range information is crucial for understanding complex spatial patterns and contextual relationships in HSI data. Since the vision transformer (ViT) [14] has been proposed and attracted considerable attention in RS image classification, the various variations of vision transformers have been widely studied and adopted for HSI classification [15,16]. Recently, researchers have presented a spectral learning model named SpectralFormer [17] to derive spectral adjacency dependencies in HSIs. It utilizes a transformer encoder module, which can effectively present spectral features. Nonetheless, it does not capture dense semantic details or make optimal use of the local spatial information. The channel attention mechanism has been demonstrated to help highlight important spectral bands. Ref. [18] introduces the "Squeeze-and-Excitation" (SE) block, which remodels the channel-wise responses using channel attention. Ref. [19] proposes the three-branch network, and the SeNet is deployed in the first branch to increase the classification performance successfully. Ref. [20] introduces a bi-branch attention-assisted network, which utilizes the SE block to obtain the attention weights from multi-scale feature maps. However, the traditional SE block has been demonstrated to only utilize the lowest-frequency information of spectral features due to the global average pooling operation.

In summary, existing methods either fail to extract global spatial and spectral information or only utilize the low-frequency information of spectral features. Furthermore, previous methods directly fuse spatial and spectral features, neglecting the need for feature filtering to obtain the necessary information. To address these issues, we propose the DCTransformer.

Our DCTransformer is a two-branch model, consisting of a base feature extractor (BFE) and a detail feature extractor (DFE). The BFE branch extracts global spectral features, addressing the lack of global spectral information. The DFE branch specifically focuses on extracting local spatial features. Additionally, we propose a simple feature dynamic fusion module (DFM), introducing two learnable parameters for the dynamic fusion of spectral and spatial features. By combining these features, the spectral and spatial data required for the ultimate prediction can be captured. A CNN-based network, incorporating dilation convolution (DC), serves as the lightweight component of the detail feature extractor, capturing local spatial information. Depthwise Separable Convolutions (DSCs) [21] are introduced to reduce computational costs and parameters. The base feature extractor adopts a novel SeNet-based channel attention mechanism, utilizing the discrete cosine transform (DCT) [22,23] instead of traditional average pooling operations. We introduce the DCT block to obtain the frequency domain representation of different frequency components of feature maps, using average frequency responses to remodel channel-wise responses through a simple multi-layer perception (MLP).

Extensive experiments demonstrate that the DCTransformer achieves a state-of-the-art (SOTA) performance in the Houston, Indian Pines, Trento, and MUUFL datasets. The effectiveness of the DCTransformer has also been proved with numerous ablation studies.

In summary, our main contributions are as follows:

- We present a specialized two-branch network architecture named the DCTransformer. Within this network, spectral and spatial features are extracted flexibly and effectively by their respective branches. Owing to this advantage, the network facilitates a subsequent dynamic focus module (DFM) that adaptively learns spatial and spectral features with varying emphases to address different land cover characteristics. These characteristics may prioritize either spatial visual features or the categorical information conveyed by spectral features.
- We introduce the DFE module, which effectively extracts spatial features while minimally increasing the total number of parameters.

- We introduce the BFE module, which is composed of the channel attention mechanism combined with the DCT method, to effectively extract spectral features and to extract the base spectral features, including global frequent information.
- We provide a DFM that fuses the spectral and spatial features self-adaptively.

This paper's remainder is structured thusly. Section 2 describes the materials and methods, the 3D-CNN, the DCT, and the attention mechanism. ResNet will be introduced in the related work subsection, and the DCTransformer model detail will be elaborated. Section 3 describes the experiment results compared to other methods. Section 4 and Section 5 present the discussion and conclusions.

2. Methods

2.1. Related Work

2.1.1. Backbone Based on CNNs

For the HSI classification task, ref. [24] proposed the utilization of a 1D-CNN architecture. Surveying the hyperspectral feature extraction, this architecture is composed of five layers: the input layer, convolutional layer, maxpooling layer, fully connected layer, and output layer. Ref. [25] surveys the hyperspectral feature extraction methods utilizing 2D-CNN and demonstrates the importance of deep feature extraction techniques. The 3D-CNN's capacity to extract spatial–spectral data is remarkable, as it captures the 3D feature embedding of the three-dimensional input data directly. The 3D-CNN has been utilized by [13]. The sequential layers and stacking encode spatial–spectral features and effectively reduce the dimension. This paper proposed the integration of the HetConv2D block following the Conv3D layer. This innovative block comprises two parallel Conv2D layers. One of these Conv2D layers has a variant DW convolution. The HetConv2D block effectively extracts multi-scale information by utilizing two kernels of different sizes.

2.1.2. Discrete Cosine Transform

The discrete cosine transform (DCT) [26] technique has been employed for image tasks. Ref. [12] investigates the convolution theorem of the DCT and proposes a faster spectral convolution method for CNNs. DCT convolution kernels are employed to transform the input feature map into the frequency domain, allowing them to execute the novel convolution operation in the spectral domain, inspired by simple spatial convolution. This technique is then used to reshape high-resolution images in the frequency domain, instead of resizing them in the spatial domain, for super-resolution and other reconstruction purposes. The DCT (discrete cosine transform) convolutional kernel applies the principles of the DCT to perform convolutional operations on image data. Unlike traditional convolutional kernels that operate in the spatial domain, DCT convolutional kernels operate in the frequency domain, focusing on capturing frequency components of the input image. With the DCT convolution, the characters of HSI data can be fully used and lead to a more efficient representation. Ref. [27] regards channel attention as a compression problem and introduces the DCT in channel attention and compress more information with multiple frequency components of the 2D-DCT.

2.1.3. Attention Mechanism

Hu et al. [18] introduced the "Squeeze-and-Excitation" block, a noteworthy architectural unit, which is extensively used in the classification of hyperspectral images through the channel attention mechanism. The channel attention mechanism is extensively employed in HSIs. This SE block effectively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Ref. [18] redistributes feature weights through three operations. Firstly, the Squeeze process compresses features across the spatial dimension, using GAP to transform each two-dimensional feature channel into a single value that has a global receptive field. Further, the Exception operation generates weights for each feature channel through parameters, where parameters are learned to explicitly model the correlation between feature channels. Finally, the Reweight oper-

ation uses multiplication to weigh each channel of the previous features and keep the dimension the same. Ref. [15] introduces a new visual attention-driven technique for the HSI classification.

The transformer for HSI classification is introduced by SpectralFormer [17]; the author first adopts purely transformers, which include two main modules (GSE and CAF) that improve the fine-grained spectral difference capture ability. SpectralFormer is suitable for both pixel-wise hyperspectral image classification and spatial–spectral classification. Ref. [28] proposed HSI-BERT, which consists of a bidirectional encoder representation used to obtain a global receptive field. Ref. [29] proposed the SSFTT method, which combines spectral–spatial feature extraction, Gaussian-weighted feature tokenization, and transformer encoding to efficiently capture spectral–spatial and high-level semantic features. Ref. [30] introduced a Multimodal Fusion Transformer (MFT) network with multihead cross patch attention (mCrossPA) for hyperspectral image classification, leveraging complementary data sources for improved generalization.

2.2. Proposed Method

This paper's overall structure shall be elucidated in this section. Subsequent sections will discuss the details of the DCTransformer model and the dynamic fusion mechanism (DFM). In summary, the introduction will be divided into four parts:

1. The overall architecture;
2. The DFE model;
3. The BFE model;
4. The DFM model;
5. The dataset and evaluation metrics.

2.2.1. Overall Network

In recent years, many studies have provided many extraordinary models for HSI classification. For example, the sequential layers of Conv3D and HetConv2D have been introduced to extract robust and discriminative features from HSIs, and the classification transformer, which has a shared cross-attention mechanism, has been adopted to effectively improve the classification performer with the class token (*CLS*). Based on the above, we propose our DCTransformer model to further consummate this task. The architecture of the DCTransformer and the data flow are depicted in their entirety.

Firstly, the HSIs have been split into 11×11 patches, represented as $X \in \mathbb{R}^{11\times11\times C}$, where C means the number of spectral bands. The patches were fed the Conc3D and HetConv2D backbone as the input. Then, we obtained the feature maps, denoted as $X^{'} \in \mathbb{R}^{11\times11\times C^{'}}$. In the feature extraction stage, we adopted the DFE and BDE modules to capture the spatial and spectral features and keep the dimensions consistent as $X^{'}$, respectively. Further, the spatial and spectral features have been used to input the DFM to obtain the spatial–spectral features. Finally, the output of the fusion module was fed to the classification performer. Figure 1 provides an outline of the proposed DCTransformer in the HSI classification task.

2.2.2. Detail Feature Extractor

The DC block has mainly consisted of depthwise convolution (DWC) of the kernel size 3×3 and Point-wise Convolution (PWC) of the kernel size 1×1, to reduce the computation cost and the number of parameters compared to the traditional convolution operation. Rather than the traditional convolution operation, dilated convolution (DC) has been employed to acquire a greater receptive field. The DC block can be expressed as follows:

$$X_D = X + PWC(GELU(PWC(BatchNorm(DWC(X))))) \quad (1)$$

where X is the input of the HSIs. $DWC_r(\cdot)$ is the depthwise dilated convolution with the dilation rate r. $BatchNorm$ is a standard batch normalization layer. $GELU$ is an activation

operation. $PWC(\cdot)$ denotes a PWC operation with kernel size 1×1, which is equivalent to a linear layer.

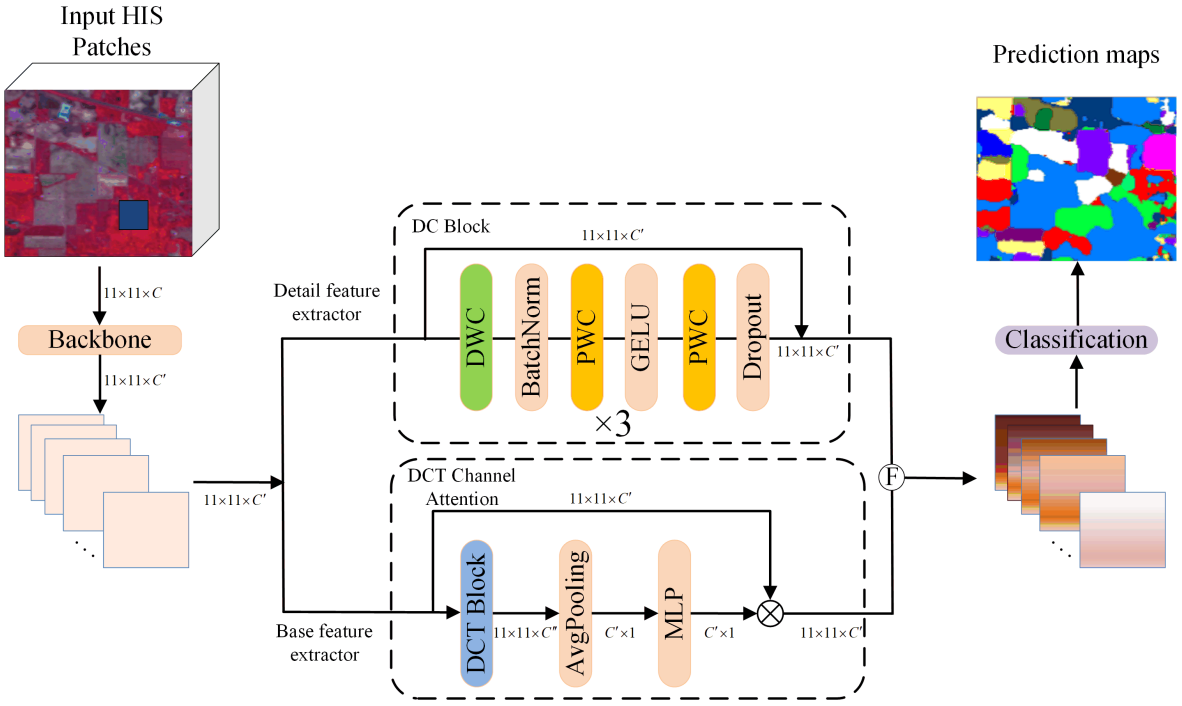

Figure 1. Overview of DCTransformer network for the HSI classification task.

Three DC blocks with a similar structure have been introduced consecutively, with the only adjustment of the dilation rate in the DWC stage. In the first block, the dilation rate $r = 1$ was utilized, corresponding to a standard non-dilated convolution. We continuously increased the value of r until it reached its upper limit value, which is constrained by the patch size. Figure 2 illustrates this stage.

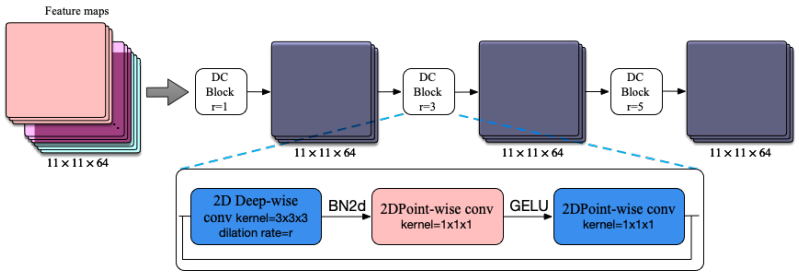

Figure 2. The structure of the proposed DFE module.

2.2.3. Base Feature Extractor

A sequence of convolutional layers and pooling operations is contained within the SENet (Squeeze-and-Excitation Network). In the input stage, the global average pooling (GAP) layer calculates the average value of all the pixel values within each feature map, thereby generating a singular value that encapsulates the global semantic information. Noting the importance, GAP only captures the low-frequency information of the input features, potentially overlooking valuable high-frequency details. Subsequent paragraphs then explain the principles of the discrete cosine transform (DCT) and demonstrate that GAP can be seen as a particular instance of the DCT.

For an input $X \in \mathbb{R}^{H \times W \times C}$ where H is the height, W is the width, and C is the number of spectral channels of X, the 2D-DCT frequency spectrum $g \in \mathbb{R}^{H \times W}$ is defined as follows:

$$g_{h,w} = \sum_{p=0}^{H-1} \sum_{j=0}^{W-1} x_{p,q} \cos\left(\frac{\pi h}{H}\left(p + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(q + \frac{1}{2}\right)\right) \tag{2}$$

where $h \in \{0, 1, \ldots, H-1\}$ and $w \in \{0, 1, \ldots, W-1\}$, which control the frequency of the cosine functions. Suppose h and w are 0; we can demonstrate that the GAP is an extreme stage that has the lowest-frequency component, where H and W can be regarded as the normalizing constants:

$$\begin{aligned} g_{0,0} &= \sum_{p=0}^{H-1} \sum_{q=0}^{W-1} x_{p,q} \cos\left(\frac{0}{H}\left(p + \frac{1}{2}\right)\right) \cos\left(\frac{0}{W}\left(q + \frac{1}{2}\right)\right) \\ &= \sum_{p=0}^{H-1} \sum_{q=0}^{W-1} x_{p,q} \\ &= \mathrm{GAP}(x) \cdot H \cdot W. \end{aligned} \tag{3}$$

In this paper, we introduced the DCT block to replace the traditional GAP, because the high-frequency information between spectral features is also crucial for the classification task. We constructed a certain number of frequency components that are denoted as J; each frequency component is designed by a special set of $[h, w]$. To apply the optimal number of J, we perform convolutional operations on the input to reduce the spatial size to 8×8 ($H \times W$). Following a simple idea, for an input $X \in \mathbb{R}^{H \times W \times C}$, we design $h \in \{1, 2, \ldots, H-1\}$ and $w \in \{1, 2, 3, \ldots, W-1\}$, and those frequency components can represented as follows:

$$J_i = \cos\left(\frac{\pi h_i}{H}\left(p + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w_i}{W}\left(q + \frac{1}{2}\right)\right) \tag{4}$$

where $i \in \{1, \ldots, N\}$ and N is the maximum of the H and W and where p and q indicate the position of each pixel. In other words, the $J \in \mathbb{R}^{N \times H \times W}$ can be regarded as a special filter, and each frequency component $J_i \in \mathbb{R}^{H \times W}$ can be written as a two-dimensional DCT matrix A, such as the following formula:

$$J_{i_{p,q}} = \begin{bmatrix} C_h(0)C_w(0) & C_h(0)C_w(1) & \cdots & C_h(0)C_w(q) \\ C_h(1)C_w(0) & C_h(1)C_w(1) & \cdots & C_h(1)C_w(q) \\ \vdots & \vdots & \ddots & \vdots \\ C_h(p)C_w(0) & C_h(p)C_w(1) & \cdots & C_h(p)C_w(q) \end{bmatrix} \tag{5}$$

With Equation (5), we observe that the DCT operation places high-frequency information in the bottom-right corner of the spatial domain, while low-frequency information is placed in the top-left corner. $C_h(\cdot) = \cos\left(\frac{\pi h_i}{H}\left(\cdot + \frac{1}{2}\right)\right)$, and $C_w(\cdot) = \cos\left(\frac{\pi w_i}{W}\left(\cdot + \frac{1}{2}\right)\right)$. For each channel of the input HSIs, calculate it using Equation (6), until all frequency components are traversed. The result $\hat{X}_i \in \mathbb{R}^{C \times H \times W}$ is concatenated as massive feature maps $\hat{X}' \in \mathbb{R}^{N \times C \times H \times W}$ and summed on its spatial dimension, which can be represented as follows:

$$\hat{X}_i = X \odot J_i \tag{6}$$

$$\hat{X}' = Concat(X_1, X_2, \ldots, X_i) \tag{7}$$

$$A_B = AvgPooling(\hat{X} = \sum_{p=0}^{H-1} \sum_{q=0}^{W-1} \hat{X}') \tag{8}$$

$$X_B = X \otimes A_B \tag{9}$$

where $\hat{X} \in \mathbb{R}^{N \times C \times 1}$, $\odot$ is the Hadamard product, and the $Concat(\cdot)$ is the concatenation operation. In the ultimate stage of the DCT block, to preserve the profuse and entire frequency information, the average strategy is adopted to the first dimension on the extracted feature map, where $A_B \in \mathbb{R}^{C \times 1}$ means the average frequency response for each channel of the attention maps and $AvgPooling$ is the average operation. In the ultimate stage, the attention maps will be fed into an MLP, which increases non-linear relationships between features and leads to a complete dimension transformation. Finally, we could specify the attention maps to achieve spectral attention calculation using Equation (9). Figure 3 illustrates the proposed base feature extractor (BFE) module.

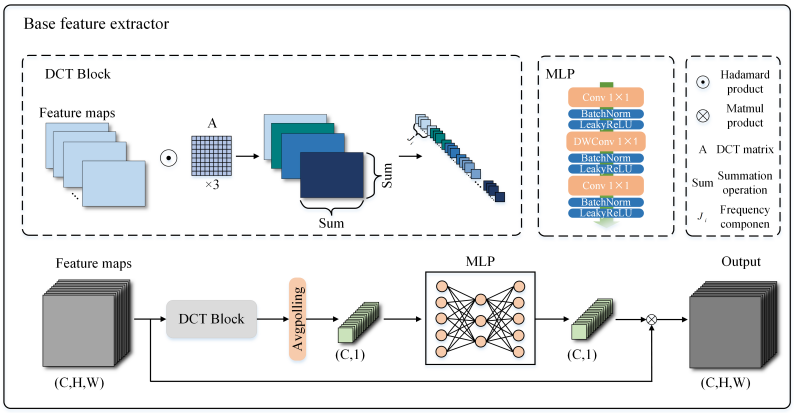

Figure 3. The structure of the proposed BFE module.

2.2.4. Dynamic Feature Fusion Mechanism

After the X_D and X_B have been claimed, they can be fed to the dynamic fusion module. In this stage, two learnable parameters have been introduced, denoted as α and β, respectively. Further, the multiplication operation has been applied between α and X_D and β and X_B, respectively, and their results have finally been added. As the training progresses, the α and β will stabilize at an appropriate weight, which is beneficial for the network to choose the ratio of two types of features that have been fused. The above procedure can be represented as follows:

$$X = \alpha X_D + \beta X_B \tag{10}$$

2.3. Dataset and Evaluation Metrics

2.3.1. Dataset Description

The performance of the classification has been evaluated by comparative experiments on four datasets, namely the University of Houston (UH), Indian Pines (IP), MUUFL, and Trento datasets.

The University of Houston (UH) conducted research published by the IEEE Geoscience and Remote Sensing Society. The dataset used by the UH was gathered in 2013 and involved the deployment of the Compact Airborne Spectrographic Imager (CASI). Each image in the dataset comprises 340 × 1905 pixels, encompassing 144 distinct spectral bands, covering a wavelength range spanning from 0.38 to 1.05 µm. The spatial resolution is set at 2.5 m per pixel (MPP). The ground truth for the images comprises 15 unique classes that correspond to various land cover types. The dataset is divided into testing and partitioning sets for each of the 15 classes. For a more comprehensive breakdown of the separate sequences and testing samples, generated by the methodology outlined in the work of [17], for each land cover type, see Table 1.

Table 1. Land cover classes of the UH dataset, with the standard training and testing sets for each class.

Class No.	Class Name	Training	Testing
1	Healthy Grass	198	1053
2	Stressed Grass	190	1064
3	Synthetic Grass	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot 1	192	1041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	184	473
	Total	2832	12,197

In 1992, the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) was employed to acquire the Indian Pines (IP) dataset, with a GSD of 20 sensors. The data collection took place over Northwestern Indiana, USA. A 10 m spectral resolution of the hyperspectral (HS) images, with 145 × 145 pixels, encompasses 220 spectral bands, extending from 400–2500 nm. After eliminating 20 bands that are affected by noise and water absorption, the dataset still contains 200 spectral bands, particularly bands 1–103, 109–149, and 164–219. Within this scene, 16 primary land cover categories are examined. Table 2 should be consulted to ascertain the class names and the number of samples allocated to training and testing in the classification task. The spatial distribution of training and testing sets is also detailed in the work by [17].

Table 2. Land cover classes of the IP dataset, with the standard training and testing sets for each class.

Class No.	Class Name	Training	Testing
1	Corn Notill	50	1384
2	Corn Mintill	50	784
3	Corn	50	184
4	Grass Pasture	50	447
5	Grass Trees	50	697
6	Hay Windrowed	50	439
7	Soybean Notill	50	918
8	Soybean Mintill	50	2418
9	Soybean Clean	50	564
10	Wheat	50	162
11	Woods	50	1244
12	Buildings, Grass, Trees, Drives	50	330
13	Stone and Steel Towers	50	45
14	Alfalfa	15	39
15	Grass, Pasture, Towers	15	11
16	Oats	15	5
	Total	695	9671

In November 2010, the University of Southern Mississippi Gulf Park (MUUFL) in Long Beach, MS, USA, was the site of the acquisition of data using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. The dataset consists of 325 × 220 pixels and contains 72 spectral bands. Additionally, LiDAR data are available, providing elevation

information from two rasters. The first and last eight bands were removed due to noise, resulting in a total of 64 bands. The ground truth data for the MUUFL dataset includes 53,687 pixels, which are categorized into 11 different classes representing urban land cover. For training, 5% of the samples from each of the 11 classes were randomly selected [13], as depicted in Table 3.

Table 3. Land cover classes of the MUUFL dataset, with the standard training and testing sets for each class.

Class No.	Class Name	Training	Testing
1	Trees	1162	22,084
2	Grass Groundsurface	344	6538
3	Road Materials	334	6353
4	Buildings' Shadow	112	2121
5	Sidewalk	69	1316
6	Cloth Panels	13	256
7	Grass—Pure	214	4056
8	Dirt and Sand	91	1735
9	Water	23	443
10	Buildings	312	5928
11	Yellow Curb	9	174
	Total	2683	28,920

The Trento dataset was collected from rural regions located in the southern part of Trento, Italy, using the AISA Eagle sensor. In parallel, corresponding LiDAR data were acquired using the Optech ALTM 3100EA sensor (Optech Incorporated, Toronto, ON, Canada). The hyperspectral (HS) images are composed of 63 spectral channels, covering wavelengths from 0.42 to 0.99 µm. Additionally, the LiDAR data encompass two elevation rasters. The HS image dataset consists of 600 × 166 pixels, representing six distinct vegetation land cover classes. It boasts a spatial resolution of 1 m per pixel (MPP) and a spectral resolution of 9.2 nm. To facilitate the training and testing procedures, the samples have been partitioned into six exclusive sets, each for training and testing purposes, as detailed in the work by [13]. Table 4 furnishes a summary of the class-wise distribution of samples in the Trento dataset.

Table 4. Land cover classes of the Trento dataset, with the standard training and testing sets for each class.

Class No.	Class Name	Training	Testing
1	Buildings	125	2778
2	Woods	154	8969
3	Roads	122	3052
4	Apples	129	3905
5	Ground	105	374
6	Vineyard	184	10,317
	Total	819	29,395

2.3.2. Evaluation Metrics

Quantitative evaluation metrics are used to evaluate the effectiveness of the suggested technique, as well as other methods for comparison. These metrics include the following:

- Overall accuracy (OA): A comprehensive assessment of the classification;
- Average accuracy (AA): The average accuracy provides a balanced representation of the model's performance by calculating the mean classification accuracy across all land cover categories;

- Kappa coefficient (κ): The kappa coefficient assesses the agreement between the anticipated classifications and the ground truth while considering any agreement that might happen by chance.

3. Results

This section delves into the particulars of experiments, including the setting of the experiment, the dataset utilized, and other specifics. The following introduction will be divided into four aspects in the corresponding subsections:

1. Experiment setting;
2. Comparative experiments;
3. Ablation experiments.

3.1. Experiment Setting

We conducted all experiments using the PyTorch framework on the NVIDIA Tesla P40 GPU. We trained our model with a batch size of 64. The AdamW [31] optimizer was deployed for training all networks rapidly. On the UH, MUUFL, and IP datasets, the learning rate was 5×10^{-4}, using a cosine learning rate schedule [32] with a gamma of 0.9 and steps of size 50, and on the Trento dataset, the learning rate was 1×10^{-4}. The training epoch was set to 500. During the experiment, patches with a size of $11 \times 11 \times B$ were taken from the HSIs and used as input to the model. The experiments were conducted on sets of training and testing samples that are spectrally and spatially disjoint [33]. This ensures that there is no overlap or interaction between the respective samples.

3.2. Comparative Experiments

Experiments conducted on the DCTransformer model proposed by us have been conducted, and the results have been compared to those of traditional methods, such as DeepHyperX [34]. They have also been compared with the state-of-the-art (***SOTA***) transformer-based techniques, such as the vision transformer (ViT) [14], SSFTT [29], SpectralFormer [17], MorphFormer [13], HybridSN [35], and double-branch multi-attention mechanism network (DBMA) [36]. During the training of DeepHyperX, we set the learning rate to 1×10^{-3} and the epoch to 500. A learning rate of 5×10^{-4} and epoch of 600 were set for the SpectralFormer's training. We trained the ViT module with a learning rate of 5×10^{-4} and an epoch of 1000. For the SSFTT, the learning rate was 1×10^{-3}, and 100 epochs were used. The results of the classification performances are in an upcoming table, with the best results being emphasized in bold.

Table 5 shows the classification performance on the UH dataset. The DCTransformer model we proposed shows an OA of 94.40%, an AA of 94.89%, and kappa coefficients (*kappa*) of 93.92%, which achieves the best performance compared with other methods. Notably, the SSFTT and MorphFormer remain the preeminent methods in the individual categories, yet the model we suggested has a clear superiority in precision overall. The traditional ViT only extracts the global spatial features and ignores the spectral features, which fails to exploit the characteristics of HSIs. Therefore, it has the worst performance. The 3D-CNN-based methods, such as DeepHyperX, can extract spatial–spectral features and therefore achieve a better performance than SpectralFormer, which can effectively extract the spectral information but has a weak capacity to extract spatial features. The SSFTT adopts the Se Block and 2D-CNN network to capture the spatial and spectral features, yet it can only obtain the lowest-frequency information. The training sample size of the Houston dataset is relatively small, and complex models such as DBMA may overfit the training data, resulting in its OA (92.20%) being inferior to the model constituting the simpler HybridSN, with an OA of 92.27%, while our proposed model uses channel attention with the DCT to fully extract spectral features in the spectral branch and a 2D-CNN-based module in the spatial branch to extract and integrate spatial features; then, we combined spatial and spectral features to complete the final classification. The classification accuracy achieved by our DCTransformer model is approximately 5.2% higher than that of

the SSFTT model. Particularly, the performance for class 15 in the UH HSI DATASET is exceptional, because, compared with other methods, our network uses the DFE, and the DFE can extract the fine spatial features and inter-class information because of gaining a larger receptive field via a gradual extensive dilation rate. Correspondingly, the BFE module has been equipped to extract spectral features with global frequency information and less information loss. Figure 4 shows the visualization of classification maps on the UH dataset and prominently displays some visual qualitative comparisons for class 15. This outcome underscores the effectiveness of the proposed BFE (global spectral feature extraction) module in capturing global spectral features. The classification result of our chosen approach is more refined and has a more distinct edge.

Table 5. Classification performance (IN%) on the **UH** HSI dataset.

Class No.	DeepHyperX	ViT	SpectralFormer	SSFTT	MorphFormer	HybridSN	DBMA	DCTransformer
1	88.98	83.00	82.15	**82.53**	81.67	80.15	81.86	82.43
2	94.27	98.50	98.78	100.00	100.00	100.00	95.77	**100.00**
3	76.43	85.74	94.65	96.24	97.82	94.85	**99.20**	99.00
4	**99.91**	99.53	91.73	99.72	96.69	96.68	98.29	95.92
5	80.69	66.57	87.32	80.88	95.87	99.24	**99.71**	90.78
6	99.34	95.93	96.59	**100.00**	97.63	**100.00**	**100.00**	99.72
7	95.80	72.03	90.91	92.31	95.80	93.56	95.42	**95.80**
8	88.07	61.05	68.07	91.23	81.75	88.03	79.58	**91.23**
9	87.41	90.21	90.30	**95.90**	93.28	90.65	88.66	95.06
10	70.30	83.11	79.89	95.92	97.53	88.52	89.67	**98.58**
11	77.30	63.72	73.69	74.45	81.95	90.03	96.29	**91.93**
12	83.38	68.56	76.02	84.51	89.71	**92.02**	92.12	90.75
13	79.34	89.38	51.54	80.79	93.63	82.80	88.77	**94.40**
14	83.81	93.12	93.52	100.00	99.60	95.54	98.78	**100.00**
15	73.57	85.62	77.17	82.24	87.53	**100.00**	99.36	97.67
OA	85.35	83.17	83.78	89.80	92.74	92.27	92.20	**94.40**
AA	85.24	82.74	84.01	90.45	92.71	92.80	93.16	**94.89**
κ ($\times$100)	84.09	81.74	82.39	88.93	92.11	91.43	91.53	**93.92**
F1 score	85.37	83.21	83.72	89.69	92.64	92.08	92.19	**94.41**
precision	85.86	84.49	84.89	90.25	92.97	92.44	92.76	**94.83**

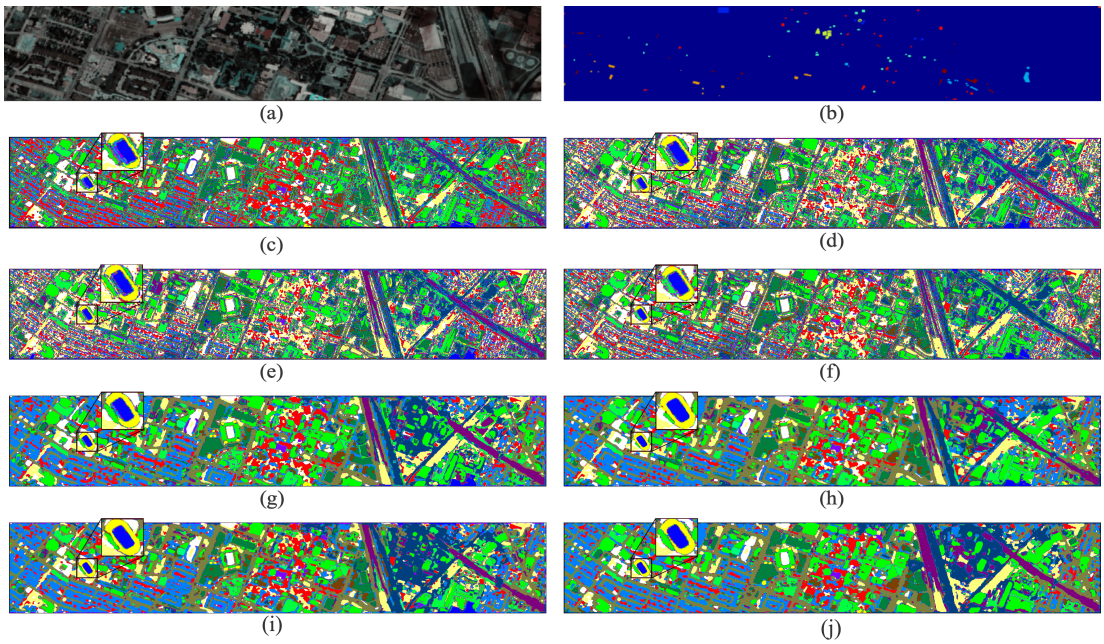

Figure 4. Classification maps for the UH dataset with emphasis on class 15 (Running Track): (**a**) pseudocolor image; (**b**) ground truth; (**c**) DeepHyperX; (**d**) ViT; (**e**) SpectralFormer; (**f**) SSFTT; (**g**) MorphFormer; (**h**) HybridSN; (**i**) DBMA; (**j**) DCTransformer.

Table 6 presents the classification performance metrics for the IP dataset, providing an assessment of its generalization capabilities. The DCTransformer model, as proposed,

exhibits a remarkable overall accuracy (OA) of 96.73%, an average accuracy (AA) of 97.80%, and a kappa coefficient (κ) of 96.25%. Our proposed approach achieves a better performance by invoking superior attention mechanisms. Moreover, Figure 5 visualizes the classification maps generated for the IP dataset.

Table 6. Classification performance (IN%) on the **IP** HSI dataset.

Class No.	DeepHyperX	ViT	SpectralFormer	SSFTT	MorphFormer	HybridSN	DBMA	DCTransformer
1	67.27	52.10	47.18	91.26	92.56	94.07	94.71	**94.72**
2	78.44	34.82	70.66	99.620	98.31	98.08	98.08	**98.60**
3	83.15	75.54	77.17	100.00	100.00	100.00	100.00	**100.00**
4	88.59	90.16	80.76	**97.54**	96.86	98.43	97.09	95.75
5	90.10	72.45	76.47	**99.57**	99.43	99.28	99.56	99.43
6	97.95	94.99	92.94	99.21	99.54	100.00	99.31	**99.54**
7	66.99	65.14	69.39	94.01	87.91	92.81	90.30	**94.44**
8	61.29	57.07	76.26	97.48	94.09	96.73	96.15	**97.48**
9	57.45	36.88	66.49	**96.10**	93.97	90.60	91.31	86.17
10	98.77	92.59	98.15	100.00	100.00	100.00	100.00	**100.00**
11	92.60	78.94	91.40	97.03	98.39	98.63	97.58	**98.64**
12	94.24	74.24	78.79	100.00	100.00	99.39	100.00	**100.00**
13	93.33	97.78	93.33	97.28	100.00	100.00	99.13	**100.00**
14	89.74	61.54	92.31	100.00	89.74	94.87	96.43	**100.00**
15	99.98	100.00	100.00	97.11	100.00	100.00	100.00	**100.00**
16	100.00	100.00	100.00	89.16	100.00	100.00	99.97	**100.00**
OA	75.38	63.13	74.00	96.69	95.37	95.61	96.15	**96.73**
AA	85.00	74.02	81.96	95.10	96.93	96.68	97.12	**97.80**
κ ($\times$100)	72.02	58.46	70.21	93.36	94.69	95.10	95.63	**96.25**
F1 score	75.51	63.82	73.98	96.76	96.06	95.51	95.75	**96.76**
precision	76.71	66.73	75.23	96.85	96.25	95.58	95.74	**96.97**

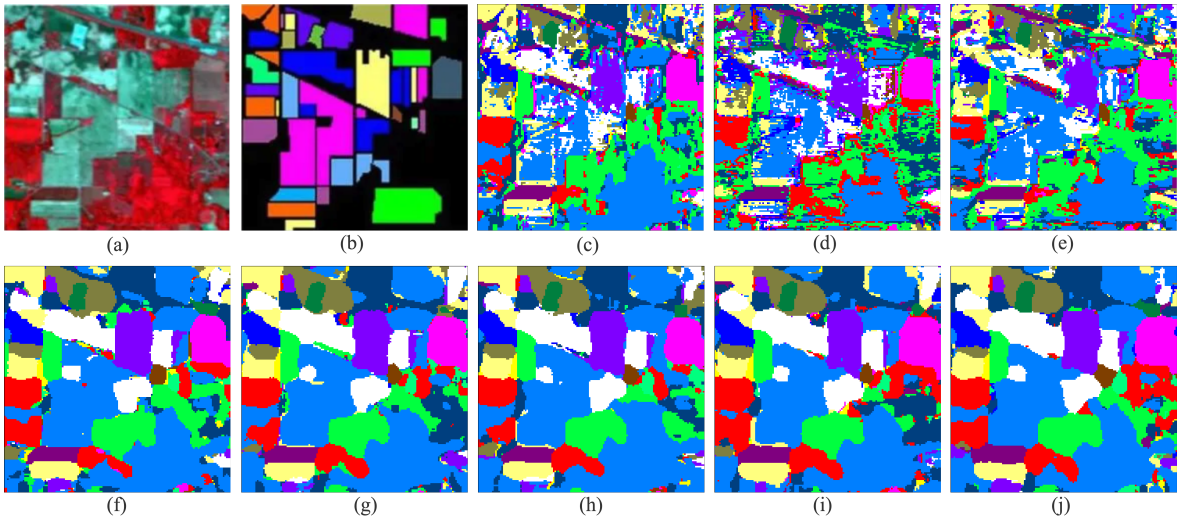

Figure 5. Classification maps for the IP dataset: (**a**) pseudocolor image; (**b**) ground truth; (**c**) DeepHyperX; (**d**) ViT; (**e**) SpectralFormer; (**f**) SSFTT; (**g**) MorphFormer; (**h**) HybridSN; (**i**) DBMA; (**j**) DCTransformer.

In Table 7, the Trento dataset's results are demonstrated for the DCTransformer model we proposed, which displays an OA of 97.45%, an AA of 95.07%, and kappa coefficients of 96.58%, respectively, which is approximately 0.6%, 1.8%, and 0.9% greater than the SSFTT. Figure 6 shows the 2D graphical plots depicting the features extracted with t-SNE on the Trento dataset. Through analysis of the t-SNE, the category points produced by our methods are more concentrated, which means our DCTransformer has a better performance in classification. Figure 7 shows the visualization with the confusion matrix for the Trento dataset. Our method has a higher True Positive Rate (TPR) compared with other methods.

Table 7. Classification performance (IN%) on the **Trento** HSI dataset.

Class No.	DeepHyperX	ViT	SpectralFormer	SSFTT	MorphFormer	HybridSN	DBAM	DCTransformer
1	95.62	82.94	92.37	95.72	98.23	97.02	95.05	**99.44**
2	72.28	95.43	89.96	86.25	88.12	89.52	90.35	**92.76**
3	66.31	59.36	42.25	70.05	86.36	**97.86**	95.98	92.78
4	99.40	83.39	94.25	99.89	98.37	97.25	98.71	**99.13**
5	99.33	96.77	97.38	99.56	99.91	99.91	99.20	**99.94**
6	81.42	54.23	57.80	**89.09**	89.02	85.02	81.55	86.37
OA	94.02	85.83	90.25	96.43	96.80	96.17	95.80	**97.45**
AA	85.72	78.69	79.00	90.09	93.33	94.43	93.47	**95.07**
κ (×100)	91.97	81.17	86.98	95.20	95.72	94.87	94.37	**96.58**
F1 score	93.96	85.91	90.19	96.39	96.50	96.19	95.84	**97.35**
precision	93.97	88.07	91.16	96.40	96.51	96.27	96.02	**97.38**

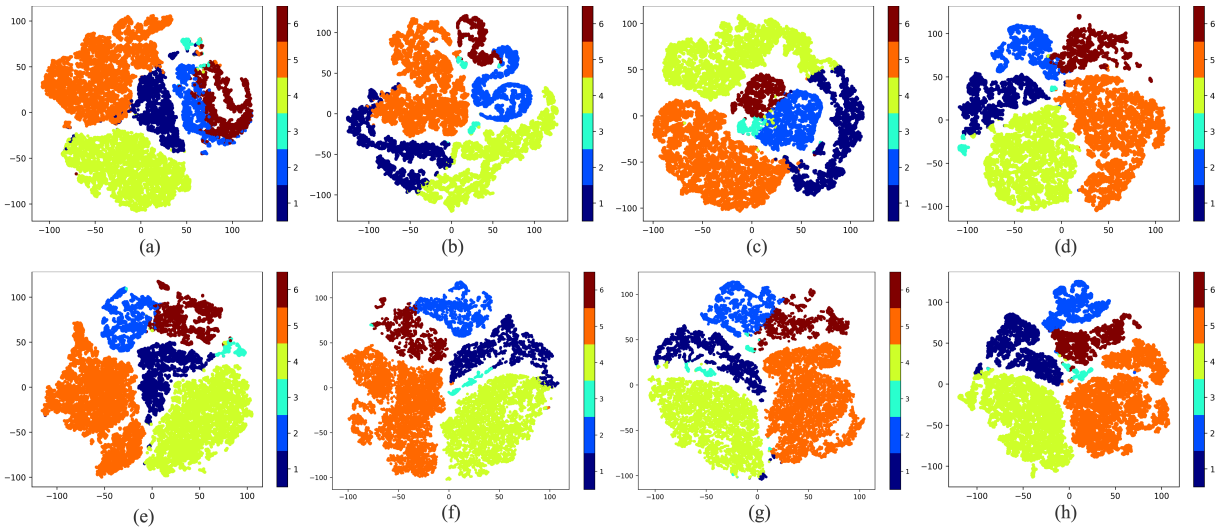

Figure 6. Visualization with t-SNE for the Trento dataset: (**a**) DeepHyperX; (**b**) ViT; (**c**) SpectralFormer; (**d**) SSFTT; (**e**) HybridSN; (**f**) DBMA; (**g**) MorphFormer; (**h**) DCTransformer.

In Table 8, The MUUFL dataset reveals the results of our proposed DCTransformer model, which displays an OA of 94.46%, an AA of 82.86%, and kappa coefficients of 92.65%. The SSFTT, however, exhibits a remarkable performance, with an accuracy of 97.50% in class 8 and parameters of 0.22 MB and GFLOPs of 0.81. In comparison to our model, the SSFTT's performance has increased by 65% and 21%, as Table 9 demonstrates. Figure 8 presents 2D graphical representations illustrating the features extracted via t-SNE on the MUUFL dataset. In Figure 9, a visualization is provided, depicting the confusion matrix for the MUUFL dataset.

Table 8. Classification performance (IN%) on the **MUUFL** HSI dataset.

Class No.	DeepHyperX	ViT	SpectralFormer	SSFTT	MorphFormer	HybridSN	DBAM	DCTransformer
1	96.27	94.98	97.06	97.24	**97.96**	97.45	97.46	97.91
2	78.53	72.26	71.45	**93.64**	89.97	89.57	85.87	92.33
3	82.55	77.79	78.33	89.98	90.72	90.65	90.34	**91.85**
4	84.11	73.49	82.77	**96.20**	92.16	91.00	94.52	94.70
5	91.36	89.53	91.00	95.14	**95.40**	93.93	92.08	94.19
6	72.23	36.12	65.91	83.07	84.65	83.29	86.45	**90.52**
7	97.03	68.32	83.40	88.21	90.43	86.61	91.51	**92.88**
8	91.84	79.47	91.50	**97.50**	96.66	97.16	96.86	97.12
9	55.24	42.86	62.84	48.86	61.93	**67.32**	63.37	64.23
10	20.69	0.00	0.00	5.17	20.69	13.79	24.13	**24.71**
11	65.62	13.67	42.58	73.05	73.44	72.65	**76.95**	73.05
OA	89.50	84.06	76.35	93.57	93.98	93.50	93.22	**94.46**
AA	75.05	58.95	76.89	78.91	81.27	80.31	81.78	**82.86**
κ (×100)	86.09	78.82	91.61	91.50	92.03	91.39	91.01	**92.65**
F1 score	89.35	83.60	88.07	93.32	93.87	93.43	93.16	**94.36**
precision	89.35	83.66	87.93	93.31	93.81	93.39	93.17	**94.31**

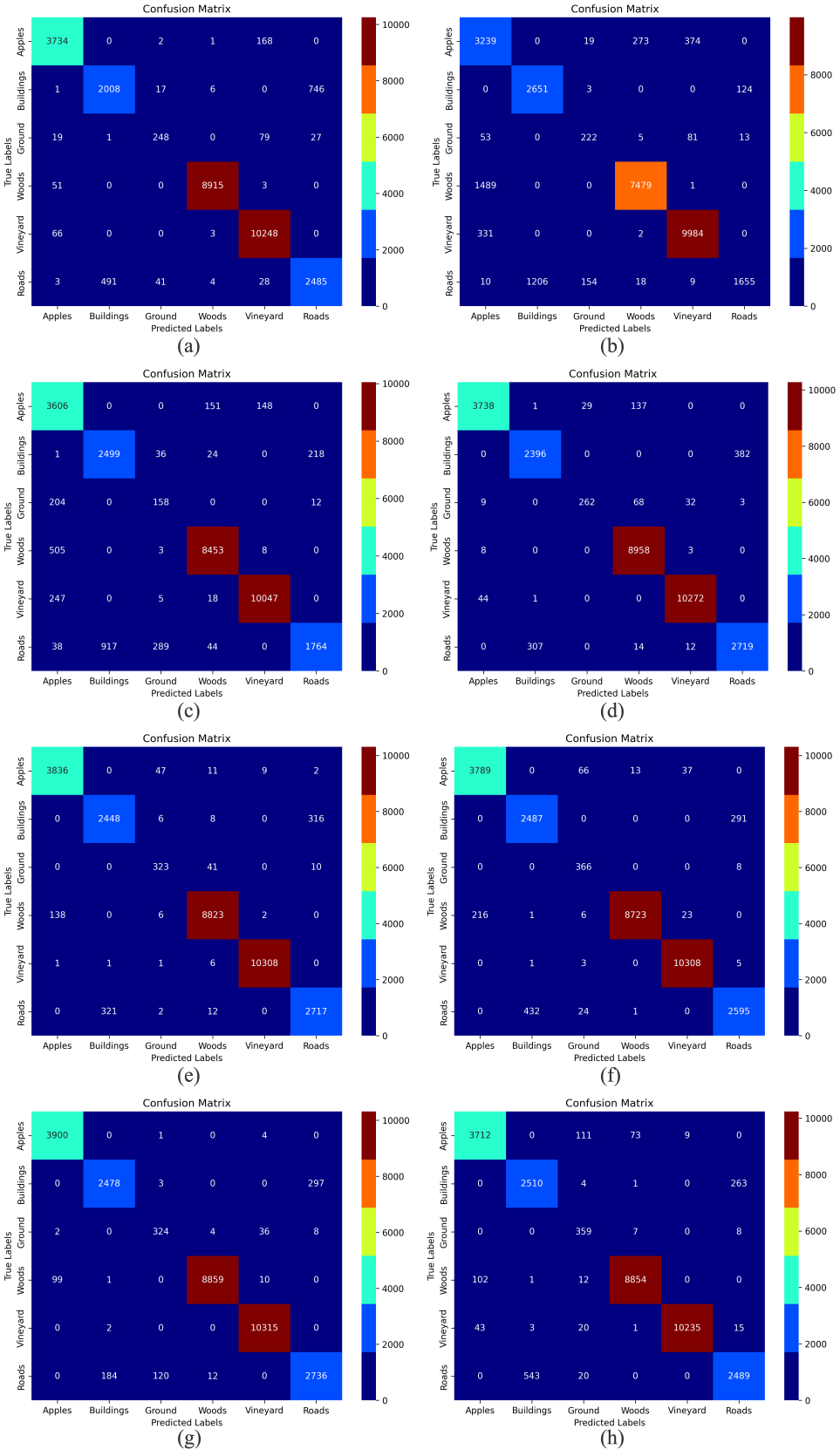

Figure 7. Visualization with confusion matrix for the Trento dataset: (**a**) DeepHyperX (93.97%); (**b**) ViT (88.07%); (**c**) SpectralFormer (91.16%); (**d**) SSFTT (96.40%); (**e**) MorphFormer (96.51%); (**f**) HybridSN; (**g**) DBMA; (**h**) DCTransformer (97.38%).

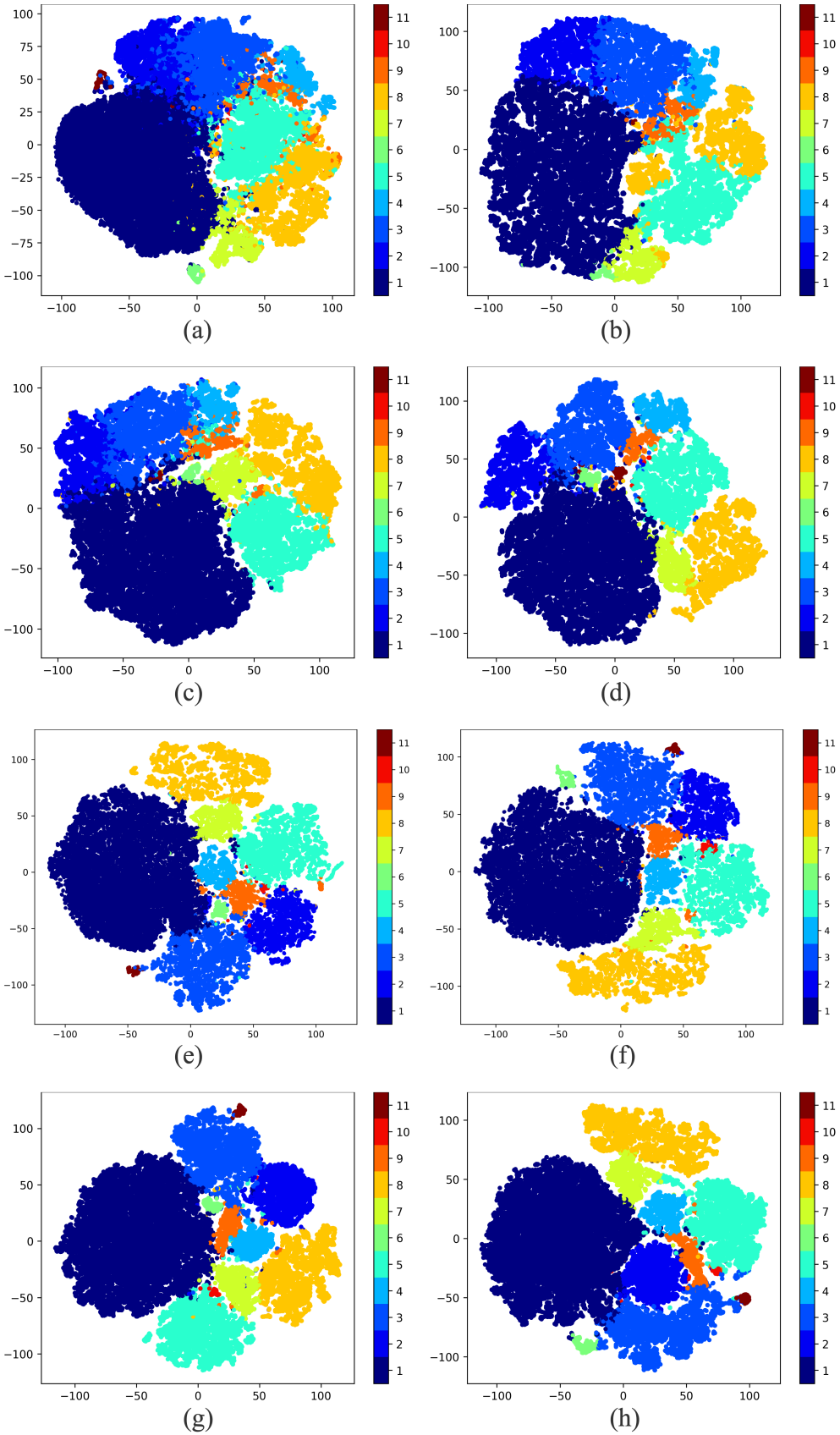

Figure 8. Visualization with t-SNE for the MUUFL dataset: (**a**) DeepHyperX; (**b**) ViT; (**c**) SpectralFormer; (**d**) SSFTT; (**e**) MorphFormer; (**f**) HybridSN; (**g**) DBMA; (**h**) DCTransformer.

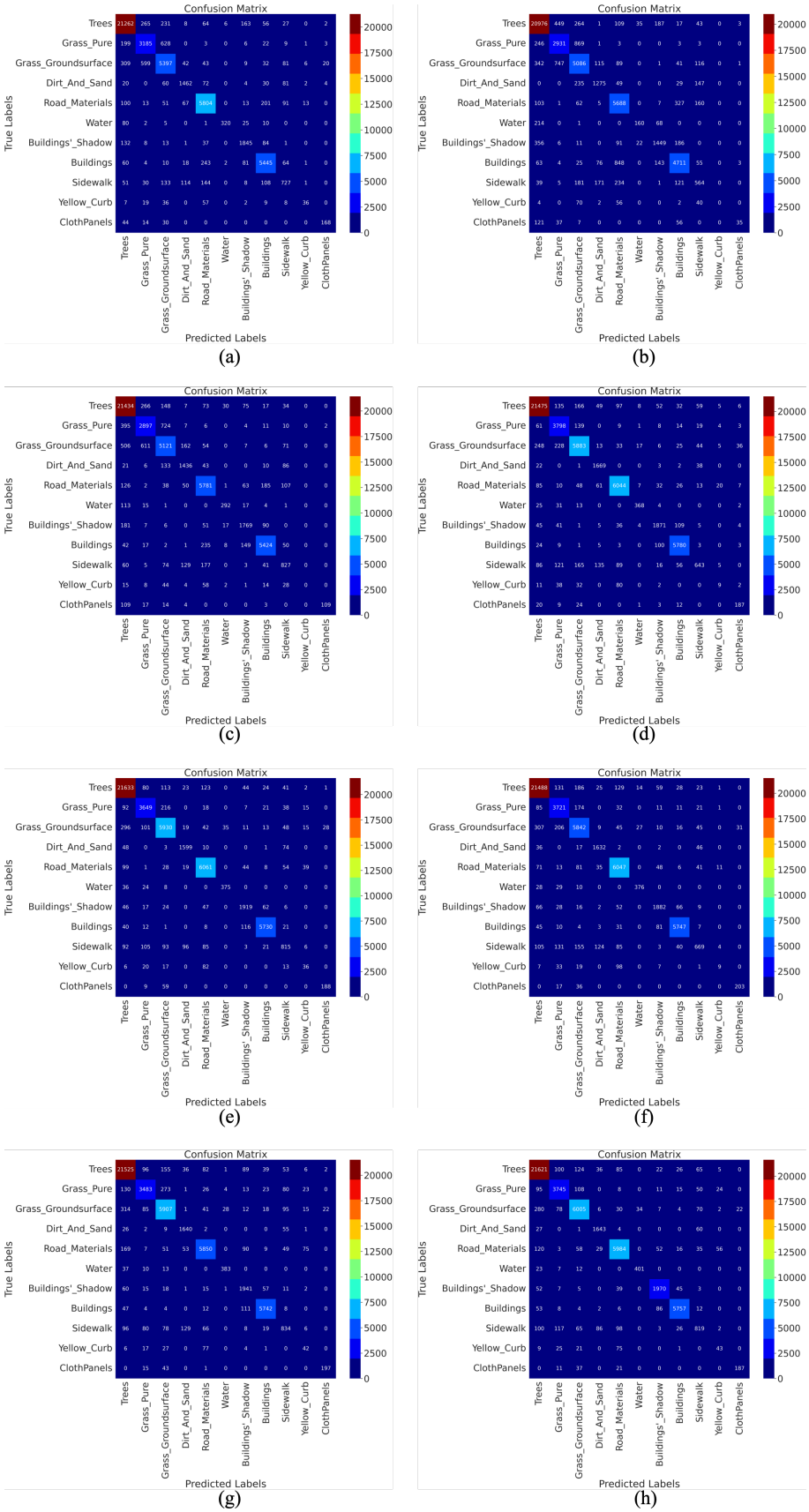

Figure 9. Visualization with confusion matrix for the MUUFL dataset: (**a**) DeepHyperX (89.35%); (**b**) ViT (83.66%); (**c**) SpectralFormer (87.93%); (**d**) SSFTT (93.31%); (**e**) MorphFormer (93.81%); (**f**) HybridSN (93.39%); (**g**) DBMA (93.17%); (**h**) DCTransformer (94.31%).

Table 9 exhibits the parameters and GFLOPs of the different methods compared. Through analysis of the parameters, the ViT embraces the least results yet has the worst performance on the AA of all datasets. The SSFTT module has the highest parameters but has not obtained the best classification result. As we know, MorphFormer achieves the ***SOTA*** performance of the HSI classification and has parameters of 0.22 MB. In comparison, our model has parameters of 0.17 MB. Through analysis of the GFLOPs, DeepHyperX has the worst result on the GFLOPs of 2.65, due to the 3D CNN involving convolution operations along the depth dimension. Our approach attains a relatively low level of GFLOPs.

Table 9. Parameters and FLOPS of all methods on the MUUFL dataset.

Method	Params (MB)	GFLOPs
DeepHyperX	0.20	2.65
ViT	0.08	0.41
SpectralFormer	0.03	0.51
SSFTT	0.28	0.98
MorphFormer	0.22	0.72
HydridSN	0.16	1.53
DBMA	0.23	0.97
DCTransformer	0.17	0.81

3.3. Ablation Experiments

By employing the MUUFL dataset, we conducted ablation experiments to evaluate the potency of the two modules we had introduced. Table 10 displays the experimental results for different cases. The DFE and BFE models we proposed have been removed in the network of DBFFS, and this structure is used as the baseline. The version without the DFE model is denoted as w/o DFE and maintains the same remaining architecture and conditions. Similarly, we designed the experiment by discarding the BFE model, which is named w/o BFE. By analyzing the above results of ablation experiments, the effectiveness of the DFE and the BFE is significantly reflected in the AA metrics. With only a minimal parameter increase, the classification performance leaps greatly. It is worth noting that the DBFFS suffers from the adverse effect of overfitting, although we have cleverly designed the dual-branch structure and fully utilized lightweight technology, which is also present in universal models. When the special models have been removed, the overfitting will be mitigated, and the model will be trained better, supposing the epoch and learning rate are set consistently. It makes the experiment results increase in a rational range and affects the significance of the results of ablation experiments, such as in OA and kappa metrics. This does not mean those models have a leak effectiveness and are unnecessary; when comparing the accuracy of the baseline and DBFSS, the results slump in three metrics we adopted due to capturing few spatial–spectral features. To conclude, further research should concentrate on optimizing the structure and augmenting the effectiveness of feature extraction modules.

Table 10. Ablation experiment on the MUUFL dataset.

Cases	OA (%)	AA (%)	Kappa (%)	Params (MB)
baseline	93.83	80.39	91.83	0.12
w/o DFE	94.26	81.75	92.40	0.165
w/o BFE	94.04	80.54	92.10	0.13
DCTransformer	94.46	82.86	92.65	0.17

The different window sizes are investigated thoroughly in this paper. Table 11 exhibits the impact of the window size of the input HSI data for the classification task. We can deduce from the table that as the window size expands, both the computational complexity and the training time increase. Additionally, the average accuracy initially ascends and subsequently declines. The reason is that, with more neighboring pixels, the DFE with a CNN may work better due to encoding more context information. Still, the BFE with the DCT has a constant number of frequency components, which hardly adapt to the window size increases. After considering achieving a good trade-off between the three evaluation

metrics and the computational volume, we ultimately chose a window size of 11 × 11 for all datasets. Figure 10 shows the OA, AA, and kappa for the different window size cases. When a patch size of 13 × 13 is set, our model achieves the best performance, yet it brings an additional and undeniable increase in GFLOPs.

Table 11. Different window sizes affect the performance on the Indian Pines dataset.

Window Size	OA (%)	AA (%)	Kappa (%)	GFLOPs
11	96.73	97.80	96.25	2.38
13	96.87	98.18	96.41	3.25
15	95.77	97.31	95.15	4.27
17	95.81	97.76	95.20	5.44
19	96.00	97.63	95.41	6.75

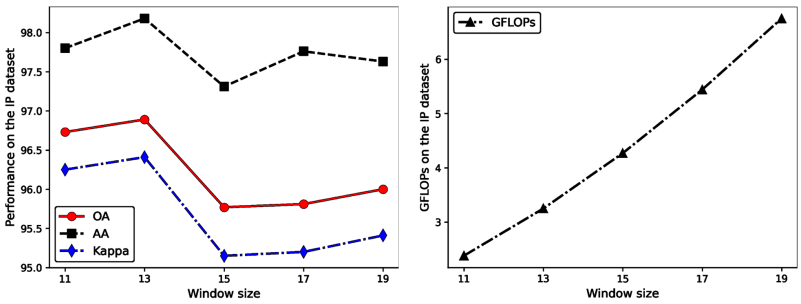

Figure 10. (**a**) Description of the effect of the different window sizes on the OA, AA, and kappa on the IP dataset. (**b**) Description of the effect of the different window sizes on the GFLOPs on the IP dataset.

4. Experiments and Discussion

The experimental results demonstrate that the DCTransformer technique is notably more effective than other classification techniques. The ViT failed to consider the spatial–spectral features, resulting in poor precision across all indicators. The SpectralFormer has the worst result in OA and AA, meaning the model performs poorly in special categories due to the deficiency of local detail spatial information, but it has higher kappa metrics, which may indicate it performs well in considering consistency and randomness in classification. The SSFTT and MorphFormer performed well, in contrast. Furthermore, the classification performance was assessed using various window sizes. The final window size of 11 × 11 was chosen by considering the OA, AA, and kappa. Table 9 shows our method's parameters on the Houston datasets and Table 11 shows different window sizes' impacts on the results. Our method demonstrates that highly frequent information is important to the HSI classification and proposes an effective model to capture the spatial–spectral features to achieve an ***SOTA*** performance.

4.1. Impact of the BFE and DFE Modules

To verify that the BFE and DFE have an enhancing effect on the classification performance, we conducted the following ablation experiments: removing both the DFE and BFE modules (baseline); removing the DFE module but keeping the BFE module (w/o DFE); removing the BFE module but keeping the DFE module (w/o BFE); and keeping all modules (DCTransformer). The results of the comparison experiments are shown in Table 10.

4.2. Impact of Different Window Sizes on Image Pathes

To find the best window size for a patched split, we designed the following comparative experiments: we gradually expanded the size of image patches from 11 × 11 to 19 × 19. Due to the salience of the central pixel position, we exclusively employed patch sizes with odd dimensions. The experimental results indicate that, considering a comprehensive

trade-off between the inference speed and classification performance, a patch size of 13×13 demonstrates an optimal performance, as shown in Table 11.

5. Conclusions

In this paper, we proposed a model, the DCTransformer, to classify HSIs. The DCTransformer includes a base feature extractor and detail feature extractor, to capture the local features, restore high-frequency information, and analyze global information in HSIs, separately. The BFE employs the DCT method to distinguish high-frequency and low-frequency information and weighs them in a cross-channel way using channel attention. The BDE module employs convolution networks to capture local information, whilst the DFE stack utilizes blocks with dilation convolution to extract detailed features. A dynamic focus mechanism further refines the resulting embedding features from both modules. Classification is achieved using the transformer. Comparative experimental results on four popular HSIs demonstrate that our framework outperforms current state-of-the-art (SOTA) algorithms, especially in the case of few samples, and also improves in terms of efficiency due to the abandonment of 3D convolution in favor of 2D convolution, where the parameters of the model are greatly reduced. The results of experiments and analyses on four datasets substantiate the method's efficacy in improving the classification performance while concurrently reducing the count of parameters. Yet, the HSI classification still encounters overfitting. Further, we strive to discover a more efficacious technique and concise architecture to amplify the efficacy and exactness of categorization.

Author Contributions: Conceptualization, Y.D. and X.Z.; Methodology, X.Z.; Software, B.L.; Validation, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

Abbreviations

The following abbreviations are used in this manuscript:

HSIs	Hyperspectral Images
CNN	Convolution Neural Network
CAM	Channel Attention Mechanism
DCT	Discrete Cosine Transform
SVMs	Support Vector Machines
ViT	Vision Transformer
SA	Self-Attention
DC	Dilation Convolution
DSC	Depthwise Separable Convolution
CLS	Class Token
GAP	Global Average Pooling
PU	Pavia University
CASI	Compact Airborne Spectrographic Imager
MPP	Meters Per Pixel
GSD	Group Sampling Distance
IP	Indian Pines
UH	University of Houston
POSIS	Reflective Optics System Imaging Spectrometer
OA	Overall Accuracy
AA	Average Accurary
κ	Kappa Coefficient

References

1. Fan, J.; Zhou, N.; Peng, J.; Gao, L. Hierarchical learning of tree classifiers for large-scale plant species identification. *IEEE Trans. Image Process.* **2015**, *24*, 4172–4184.
2. Sabbah, S.; Rusch, P.; Gerhard, J.H.; Stöckling, C.; Eichmann, J.; Harig, R. Remote sensing of gases by hyperspectral imaging: Results of measurements in the Hamburg port area. In Proceedings of the Electro-Optical Remote Sensing, Photonic Technologies, and Applications V. *SPIE* **2011**, *8186*, 261–269.
3. 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*, 3140–31460 [CrossRef]
4. Lu, B.; He, Y.; Dao, P.D. Comparing the performance of multispectral and hyperspectral images for estimating vegetation properties. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1784–1797. [CrossRef]
5. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [CrossRef]
6. Gao, L.; Hong, D.; Yao, J.; Zhang, B.; Gamba, P.; Chanussot, J. Spectral superresolution of multispectral imagery with joint sparse and low-rank learning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2269–2280. [CrossRef]
7. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.; Wu, X.; Yao, J.; Khan, A.M.; Mazzara, M.; Distefano, S.; Chanussot, J. Hyperspectral image classification—Traditional to deep models: A survey for future prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 968–999 [CrossRef]
8. 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]
9. Seifi Majdar, R.; Ghassemian, H. A probabilistic SVM approach for hyperspectral image classification using spectral and texture features. *Int. J. Remote Sens.* **2017**, *38*, 4265–4284. [CrossRef]
10. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; IEEE: Piscataway Township, NJ, USA, 2015; pp. 4959–4962.
11. Aptoula, E.; Ozdemir, M.C.; Yanikoglu, B. Deep learning with attribute profiles for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1970–1974. [CrossRef]
12. Xu, Y.; Nakayama, H. Dct-based fast spectral convolution for deep convolutional neural networks. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway Township, NJ, USA, 2021; pp. 1–8.
13. 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*, 1–15. [CrossRef]
14. 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 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Haut, J.M.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Li, J. Visual attention-driven hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8065–8080. [CrossRef]
16. Hu, X.; Li, T.; Zhou, T.; Liu, Y.; Peng, Y. Contrastive learning based on transformer for hyperspectral image classification. *Appl. Sci.* **2021**, *11*, 8670. [CrossRef]
17. 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]
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Asker, M.E. Hyperspectral image classification method based on squeeze-and-excitation networks, depthwise separable convolution and multibranch feature fusion. *Earth Sci. Inform.* **2023**, *16*, 1427–1448. [CrossRef]
20. Huang, W.; Zhao, Z.; Sun, L.; Ju, M. Dual-branch attention-assisted CNN for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 6158. [CrossRef]
21. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
22. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *100*, 90–93. [CrossRef]
23. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.K.; Ren, F. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1740–1744.
24. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
25. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88. [CrossRef]
26. Khayam, S.A. The discrete cosine transform (DCT): Theory and application. *Mich. State Univ.* **2003**, *114*, 31.
27. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 783–792.

28. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [CrossRef]
29. 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]
30. Roy, S.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal fusion transformer for remote sensing image classification. *arXiv* **2022**, arXiv:2203.16952.
31. Dubey, S.R.; Chakraborty, S.; Roy, S.K.; Mukherjee, S.; Singh, S.K.; Chaudhuri, B.B. diffGrad: An optimization method for convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 4500–4511. [CrossRef]
32. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
33. Hendrix, E.M.; Paoletti, M.; Haut, J.M. On Training Set Selection in Spatial Deep Learning. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 327–339.
34. Zhang, M.; Wang, Z.; Wang, X.; Gong, M.; Wu, Y.; Li, H. Features kept generative adversarial network data augmentation strategy for hyperspectral image classification. *Pattern Recognit.* **2023**, *142*, 109701. [CrossRef]
35. 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]
36. Ma, W.; Yang, Q.; Wu, Y.; Zhao, W.; Zhang, X. Double-branch multi-attention mechanism network for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 1307. [CrossRef]