applied sciences

MDPI

Article

Hyperspectral Image Classification Based on Hybrid Depth-Wise Separable Convolution and Dual-Branch Feature Fusion Network

Hualin Dai, Yingli Yue and Qi Liu *

School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China; daihualin@tcu.edu.cn (H.D.); yyl2635@126.com (Y.Y.)
* Correspondence: liuqicj@126.com

Abstract: Recently, advancements in convolutional neural networks (CNNs) have significantly contributed to the advancement of hyperspectral image (HSI) classification. However, the problem of limited training samples is the primary obstacle to obtaining further improvements in HSI classification. The traditional methods relying solely on 2D-CNN for feature extraction underutilize the inter-band correlations of HSI, while the methods based on 3D-CNN alone for feature extraction lead to an increase in training parameters. To solve the above problems, we propose an HSI classification network based on hybrid depth-wise separable convolution and dual-branch feature fusion (HDCDF). The dual-branch structure is designed in HDCDF to extract simultaneously integrated spectral–spatial features and obtain complementary features via feature fusion. The proposed modules of 2D depth-wise separable convolution attention (2D-DCAttention) block and hybrid residual blocks are applied to the dual branch, respectively, further extracting more representative and comprehensive features. Instead of full 3D convolutions, HDCDF uses hybrid 2D–3D depth-wise separable convolutions, offering computational efficiency. Experiments are conducted on three benchmark HSI datasets: Indian Pines, University of Pavia, and Salinas Valley. The experimental results show that the proposed method showcases superior performance when the training samples are extremely limited, outpacing the state-of-the-art method by an average of 2.03% in the overall accuracy of three datasets, which shows that HDCDF has a certain potential in HSI classification.

Keywords: hyperspectral image classification; hybrid residual unit; depth-wise separable convolution; spatial–spectral features

check for updates

Academic Editor: Atsushi Mase

Received: 3 January 2025
Revised: 25 January 2025
Accepted: 26 January 2025
Published: 29 January 2025

Citation: Dai, H.; Yue, Y.; Liu, Q. Hyperspectral Image Classification Based on Hybrid Depth-Wise Separable Convolution and Dual-Branch Feature Fusion Network. *Appl. Sci.* **2025**, *15*, 1394. https://doi.org/10.3390/app15031394

Copyright: © 2025 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

Remote sensing techniques are becoming increasingly mature. Among them, hyperspectral remote sensing is an advanced technology combining imaging technology and spectral technology, based on numerous narrow and contiguous spectral bands [1,2]. In a hyperspectral image (HSI), each pixel contains spectral information across different bands and spatial information. Regarding the exploitation of distinguishing subtle spectral differences, HSI has a lot of application prospects in cross-scene coastal wetlands [3], precision agriculture [4], and soil carbon content prediction [5].

In recent years, the research on HSI classification has played a significant role in fully utilizing the value of remote sensing images and has drawn broad attention in the field of remote sensing [6]. Compared to traditional methods depending on the handcrafted feature [7–11], deep learning (DL) has occupied a dominant position in computer vision [12],

due to its robust feature representation capabilities and prominent performance. The methods based on DL include recursive autoencoder (RAE) [13], convolutional neural network (CNN) [14–17], recurrent neural network (RNN) [18,19], and generative adversarial network (GAN) [20], which have achieved exciting performance.

Due to the reliance on spectral features in previous methods, the spatial features were often neglected, resulting in the performance of those models remaining unsatisfactory. Then, more and more classification frameworks based on spectral–spatial features have been developed, which incorporate spatial characteristics of ground objects into pixel-wise classifiers. For instance, a model called HybridSN for HSI classification is explored in ref. [21], which constructs a 2D-CNN and a 3D-CNN to extract more abstract-level feature representation after principle component analysis (PCA) [22] operation on original images, improving the classification effects. The proposed pyramid-based hierarchical spatial–spectral transformer (PyFormer) [23] enhances processing efficiency by organizing input data into hierarchical pyramid segments at different abstraction levels. Each level uses a dedicated transformer encoder to effectively capture local and global contexts. The final input representation integrates outputs from various levels. The pyramid excels at capturing spatial features and local patterns, while the transformer models spatial–spectral correlations and long-range dependencies. One popular structure is decoupling the classification task for two branches to extract the spectral features and spatial features, respectively. Ma et al. [24] present an HSI classification method using a double-branch multi-attention mechanism network (DBMA). The network designs two branches that separately extract spectral features and spatial features, reducing interference between them. Each branch employs a specific attention mechanism suited to its characteristics, ensuring the extraction of more discriminative features. The features from the double branch are then finally fused for classification. Another widely used approach is cropping HSI data into local cubes in a certain size as inputting data and then transferring them to the CNN module for acquiring the joint features [25]. For example, in [26], spatial–spectral information is acquired by feeding data local cubes into a network that combines a 3D CNN with a feature pyramid structure. Additionally, this network employs multiscale fusion to enhance sensitivity to spatial–spectral information.

In recent years, many well-recognized networks based on transformers [27] have been widely and successfully applied to HSI classification since Vision Transformer (ViT) [28] was proposed in 2020. For example, Hong et al. [29] proposed a novel backbone network SpectralFormer which is able to learn spectrally local sequence information from neighboring bands of HSI, thereby generating group-wise spectral embeddings. To prevent information loss during layer propagation, it uses cross-layer skip connections to adaptively fuse residuals across layers. Extended Vision Transformer (ExViT) [30] is a new framework based on multi-modal transformation that extends traditional ViT with minimal changes. Unlike traditional linear projections, ExViT is built on a lightweight separable backbone, extracting features of image patches through local and global perception in both spatial and channel dimensions. To enhance information exchange between multi-modalities, an attention fusion module is designed, which improves the model's discriminative ability by integrating attention at different stages. In ref. [31], considering that existing CNN-based models cannot sufficiently utilize the sequence attributes of spectral features, Yang et al. embedded the convolution process into the changed structure of the transformer. MFT provides multi-head cross-patch attention for land cover classification tasks, which also utilizes other complementary information to achieve better generalization. Although the above methods based on transformers can obtain global features to increase accuracy effectively, these methods come with notable limitations, including high demands on memory and computational resources.

Among numerous DL models, the CNN has demonstrated its prominent advantage of spectral–spatial characteristic extraction. A CNN has three main specialties: (1) the local receptive field, (2) the sharing of weights, and (3) the multiple convolution kernel. However, a CNN also has some problems that can be further optimized; particularly in the pooling layer, a significant amount of valuable information can be lost and the connection between the global and the local features may be overlooked. In CNN-based methods, the traditional method of using a 1D CNN [32] alone can capture spectral features from each pixel of HSI, but spatial information is not adopted. An HSI classification approach based on the structure of only utilizing a 2D CNN is called the spatial-based classification method, which extracts spatial features following a similar approach to traditional RGB images, which brings an inevitable weakness caused by ignoring the unified spectral–spatial attributes of the hypercube [33,34]. Based on this, the appearance of the 3D CNN successfully captures the integrated spectral–spatial features across various spectral bands. For example, in [35], Fang et al. propose a 3D dense convolutional network with a spectral-wise attention mechanism (MSDN-SA), which simultaneously captures the spectral–spatial features at multiple scales via 3D dilated convolutions. Despite experimental evidence from HSI public datasets showing that utilizing 3D CNNs enhances HSI classification performance, the increased computational complexity of deep 3D CNNs results in suboptimal outcomes for classes with analogous textures over multiple spectral bands. Therefore, to alleviate the mentioned issue, gradually integrating 2D and 3D CNNs with diverse kernel sizes has been paid more attention recently, which contributes to capturing more abstract joint spectral–spatial features, ultimately making the network strong enough. Kanthi et al. leveraged the profound feature extraction capabilities of 3D CNNs for HSI classification and achieved better performance [36]. In ref. [37], ESFNet provided a 3D convolutional neural network based on the fusion of different spectral strides (SSFCNN), which can handle weighted spectral data to facilitate the efficient learning of spectral characteristics. The model optimizes the preservation of spectral continuity, thereby enhancing classification accuracy.

The attention mechanism is also incorporated to enhance the feature matrices, optimizing the differentiation of the extracted features. In [38], the key model is the spectral–spatial domain attention network (SSDA), designed for few-shot HSI classification. SSDA comprises three modules: the spectral–spatial module for learning robust features, the domain attention mechanism to amplify relevant features while mitigating the influence of irrelevant ones, and the multifaceted loss module, implemented to tackle domain adaptation challenges. Ref. [39] presents a novel network that synergistically combines residual modules with attention mechanisms. Separately selecting features from each band by weight-adaptive learning and enhancing the importance of the center pixel are executed in both networks. Furthermore, considering the problem that available training samples for training DL models are generally very limited, Wang et al. [40] propose a depth-wise separable fully convolutional residual network (DFRes) combined with convolutional conditional random fields and region growing to tackle the question of relying on limited training samples, where residual learning and depth-wise separable convolution effectively addresses the challenges of gradient vanishing, overtraining, and high memory requirements.

Consequently, augmenting the capacity for feature extraction and bolstering the classification accuracy of limited training samples have emerged as pressing issues in the realm of HSI classification, particularly within the framework of deep learning networks. In this paper, inspired by the hybrid 2D–3D CNN and feature fusion [41], we construct a hybrid depth-wise separable convolution and dual-branch feature fusion network (HDCDF) for classification. The proposed architecture contains two branches based on 2D–3D depth-wise separable convolutions, combined with an attention mechanism and hybrid residual blocks separately, which both capture joint spectral–spatial features. By fusing the extracted

features of the two branches, we obtain further features and send them to the classifier to determine the final classification result. In addition, we evaluate the efficacy of the proposed HDCDF with a reduced number of training samples. The primary contributions of this study are as follows:

1. We propose an innovative network architecture that leverages 2D–3D depth-wise separable convolution along with dual-branch feature fusion for HSI classification;
2. The proposed model incorporates 2D depth-wise separable convolution attention (2D-DCAttention) and hybrid residual blocks, which effectively improve the classification performance;
3. The classification performance of the HDCDF network is analyzed using limited training data, achieving state-of-the-art results across three public datasets.

2. Related Work

This section provides a concise overview of the fundamental modules included in HDCDF, containing depth-wise separable convolution, a batch-normalized 3D-CNN and Mish function, an attention mechanism, and a residual unit.

2.1. Depth-Wise Separable Convolution and Standard Convolution

Depth-wise separable convolution (DSC) [40] is composed of depth-wise convolution (DWC) and pointwise convolution (PWC). Among them, DWC operates independently on each channel of the input image, effectively extracting spatial characteristics along each dimension, with one convolution kernel responsible for one channel, combining all output feature maps to generate the final result. This operation greatly reduces the number of parameters compared with standard convolution. PWC is actually 1×1 standard convolution, which has two functions including the freedom to change the number of output channels and the fusion of feature map channels after DWC. The operation of PWC has a less computationally intensive nature, which makes depth-wise separable convolution possess significant advantages in lightweight models and environments with limited computing resources.

Figure 1 is a contrast diagram of standard convolution and depth-wise separable convolution. Compared with standard convolution, DSC can be decomposed into two steps. We assume the input image size is $D_I \times D_I \times k$, where D_I is the height and width of the input image, and k represents the number of channels of the image, ignoring other convolution parameters. Given a kernel of size $D_K \times D_K$ and a number of kernels (n), we can obtain an output feature map of $D_O \times D_O \times k$ (D_O is the final output image size). Respectively, the computational complexity of the standard convolution and DSC is as follows:

$$P_1 = D_O \times D_O \times D_K \times D_K \times k \times n \tag{1}$$

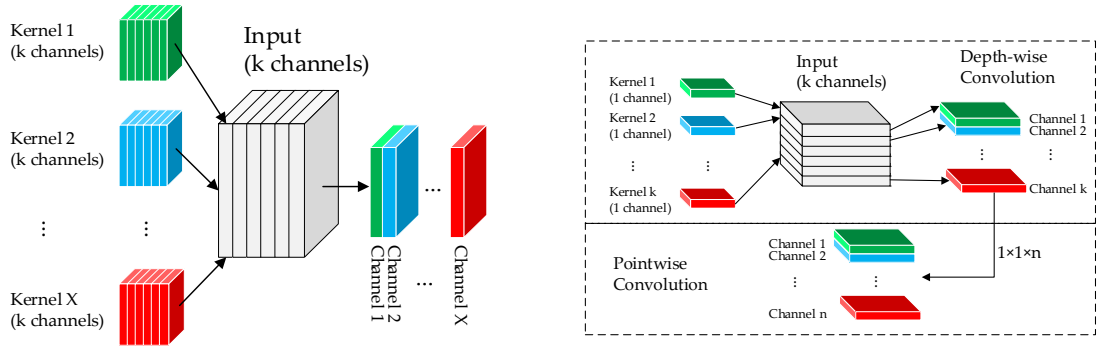
$$P_2 = D_O \times D_O \times D_K \times D_K \times k + k \times n \times D_O \times D_O \tag{2}$$

From Equation (2), the first part is the calculational complexity of depth-wise convolution, and the second part refers to pointwise convolution. Relative to standard convolution, the ratio of the computational complexity of DSC to standard convolution is shown below:

$$R = \frac{P_2}{P_1} = \frac{1}{n} + \frac{1}{{D_K}^2} \tag{3}$$

It is evident that the parameters and computations involved in depth-wise separable convolution are merely a fraction, specifically R times, of those in standard convolution, significantly enhancing model training speed and reducing computational cost. There-

fore, instead of full 3D convolutions, we used mixed 2D and 3D depth-wise separable convolutions, offering computational efficiency.

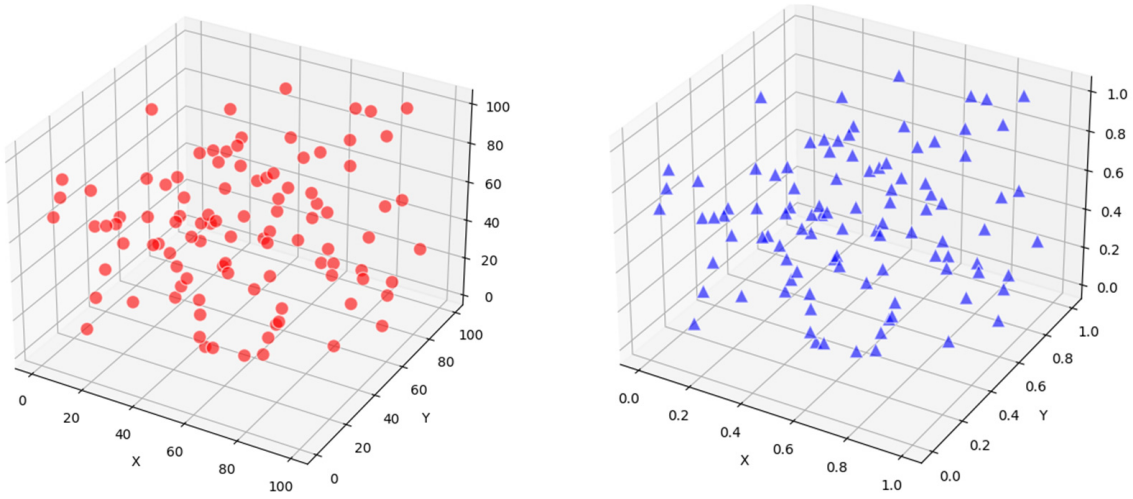


(**a**) (**b**)

Figure 1. A comparison graph of standard convolution and depth-wise separable convolution. (**a**) The structure of standard convolution; (**b**) the structure of depth-wise separable convolution.

2.2. Batch-Normalized 3D-CNN and Mish Activation

In deep learning, the normalization layer is a very important part of the neural network that contributes to stabilizing the network during training, accelerates convergence, and improves the generalization ability of the model. Among them, normalization operations on cube data can adjust the voxel value to a fixed range, commonly 0 to 1, as shown in Figure 2, the cube data scatter plot.



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

Figure 2. Data scatter plot of original data and normalized data. The red circles represent random numbers and blue triangles are corresponding normalized data. (**a**) Original randomly generated data; (**b**) normalized data.

Since HSI contains a large amount of spatial and spectral information, we add a BN layer in each 3D convolutional layer to improve model stability and speed up training convergence. There are many normalization methods, such as maximum and minimum value normalization, mean normalization, and standardization. This paper adopts a standardized approach; the results can be expressed as Equation (4), where μ represents the mean of the sample data and σ is the standard deviation of the sample data. After

standardization, the optimal parameters can be obtained more easily, thus achieving the effect of accelerating convergence.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{4}$$

The activation function is an important component of the neural network model, which is designed to help the network learn complex patterns in the data and introduce nonlinearity into the neural network architecture. A suitable activation function can speed up the back-propagation and enhance network convergence. We used a non-monotonic function, Mish, as opposed to the conventional ones: $ReLU(x) = max(0, x)$. The equation for Mish and its differential coefficient is

$$Mish(x) = x \times tanh(softplus(x)) = x \times tanh(ln1 + e^x) \tag{5}$$

$$f'(x) = \frac{e^x \omega}{\delta^2} \tag{6}$$

where x is the input of the activation, $\omega = 4(x+1) + 4e^x + e^{3x} + e^x(4x+6)$, and $\delta = 2e^x + e^{2x} + 2$.

As depicted in Figure 3, the comparison between Mish and ReLU reveals that Mish is unbounded above and bounded below, with a range approximately from -0.31 to infinity, with similar characteristics to ReLU, while the ReLU function is a piecewise linear operator that disregards all negative inputs. Conversely, Mish preserves non-positive inputs as a non-monotonic function, contributing to the stabilization of the gradient flow within the network. If the input is non-positive, neurons using the ReLU activation function cannot be updated like Mish, even though inputs might contain valued information. In addition, the Mish function is a smooth function with enhanced generalization capability to elevate the outcome quality, accelerating the counter-propagation and convergence processes.

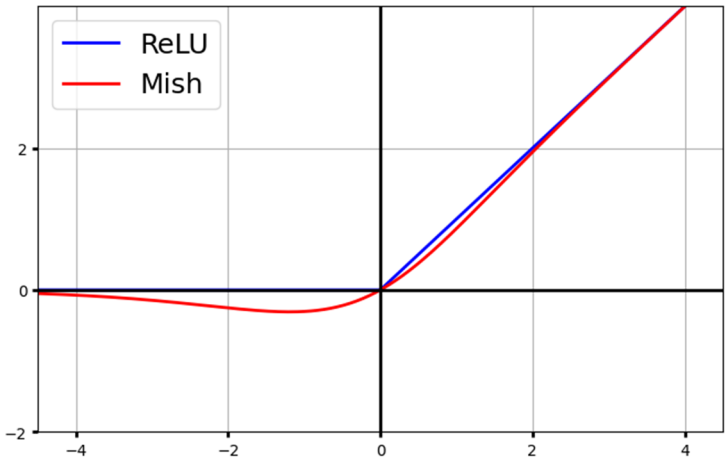


Figure 3. The comparison between Mish and ReLU.

2.3. Attention Mechanism

One disadvantage of convolution is that all spatial–spectral pixels undergoing convolution have equal weights in both spatial and spectral domains. However, in fact, different spectral bands and spatial pixels have different contributions to the extraction of spectral–spatial information. The proposed attention mechanism emphasizes the most influential information, inhibits the influence of useless information on the result, and thus makes the classification result more superior. Therefore, the attention mechanism is a powerful technique in processing hyperspectral remote sensing image classification tasks. A spatial attention mechanism is added to the channel attention mechanism, which extracts the

spatial position correlation of the target image, and this joint spatial–spectral attention mechanism enables the network to extract more abstract and decisive features. The convolutional block attention module (CBAM) [42] is a representative module in the joint spatial–spectral attention mechanism.

Thus, we propose 2D depth-wise separable convolution attention (2D-DCAttention) by combining CBAM and DSC, as shown in Figure 4. Specifically, the input is first extracted by 2D-DWC for spatial dimension, then further processed by BN and the Mish activation function. The obtained features are sent into CBAM to strengthen features in the spatial dimension and spectral dimension, and the enhanced features are processed by PWC to enhance the features between each channel.

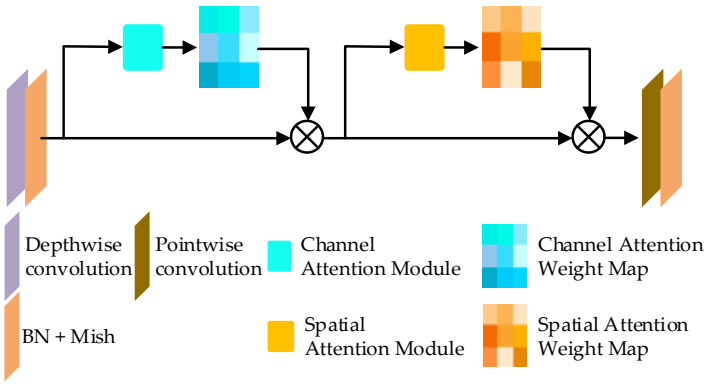


Figure 4. The map of 2D-DCAttention.

The channel attention module and spatial attention module are described in detail as follows.

2.3.1. Channel Attention Module

The structure of the channel attention module (CAM) is shown in Figure 5, whose core purpose is to highlight those feature channels that are more important to the task. We assume an intermediate feature map $F \in \mathbb{R}^{B\times W\times H\times C}$ as input, and then global average pooling and global max pooling are applied to F to generate two feature vectors, $f_i \in \mathbb{R}^{B\times 1\times 1\times C}$ and $g_i \in \mathbb{R}^{B\times 1\times 1\times C}$, representing the global maximum feature and average feature of each channel. This step aims to concentrate the information on the feature map of the channel so as to better capture it. Using these two features can greatly improve the representation capability of the network.

$$f_i = AvgPool(F) \tag{7}$$

$$g_i = MaxPool(F) \tag{8}$$

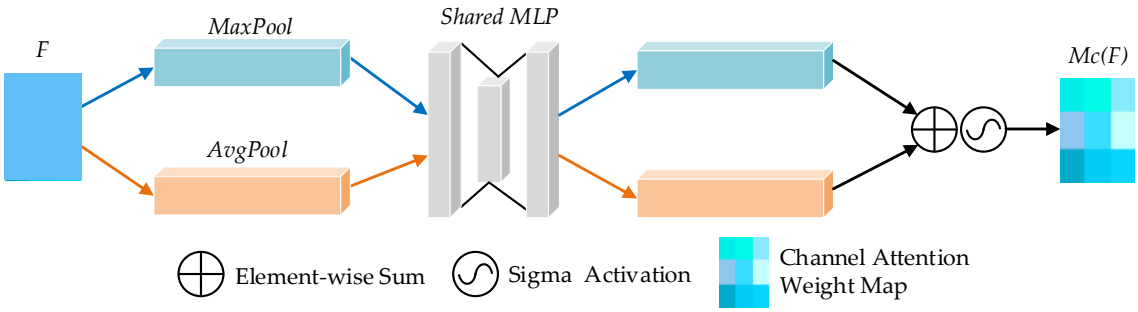


Figure 5. The structure of the channel attention module.

f_i and g_i pass through a weighted shared multilayer perceptron (MLP). The shared MLP is composed of two convolution functions and a ReLU activation function, which firstly reduces the dimension and then raises the dimension. Because the weights in MLP are shared, the weight matrix of inputs used is the same, and this operation can reduce the parameters.

$$S_i = MLP(f_i) = W_1(W_0(f_i)) \tag{9}$$

$$T_i = MLP(g_i) = W_1(W_0(g_i)) \tag{10}$$

where $W_0 \in \mathbb{R}^{C \times C/r}$ and $W_1 \in \mathbb{R}^{C \times C/r}$; r is the reduction ratio in the hidden layer. After applying the shared MLP to each feature, the element-wise sum is used to merge the output feature vectors S_i and T_i. Then, the output feature vector is generated by the sigmoid function to generate the weight vector $M_c(F)$.

$$M_c(F) = \sigma(S_i + T_i) \tag{11}$$

2.3.2. Spatial Attention Module

The goal of the spatial attention module (SAM) is to spatially emphasize more critical areas, as shown in Figure 6. Assuming that the feature map obtained after processing by CAM is $F_c \in \mathbb{R}^{B \times W \times H \times C}$, through average pooling and maximum pooling operations on the channel of F_c, the channel dimension (C) can be compressed to 1 to integrate the global channel information into the single-channel feature maps $x_i \in \mathbb{R}^{B \times 1 \times \mathrm{H} \times W}$ and $y_i \in \mathbb{R}^{B \times 1 \times \mathrm{H} \times W}$. At the same time, for each sample (B), the model will average the characteristics of the sample on the channel, realizing the compression and consolidation of the global channel information. These operations help to reduce the number of parameters and the computational complexity while preserving important channel characteristics.

$$x_i = AvgPool(F_c) \tag{12}$$

$$y_i = MaxPool(F_c) \tag{13}$$

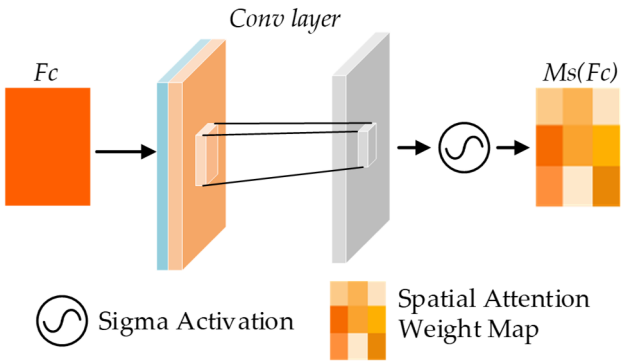


Figure 6. The structure of the spatial attention module.

We concatenate two matrices together on the channel and, through a 7×7 convolutional kernel $f^{7\times7}$, the number of channels is changed to 1 again so that all the obtained feature information is distributed on one channel. Then, the weight vector $M_s(F_c)$ is generated through the sigmoid function on the output of the convolutional layer.

$$M_s(F_c) = \sigma(f^{7\times7}(Concat(x_i + y_i))) \tag{14}$$

2.4. Residual Unit

The classification accuracy of HSI has been substantially improved based on CNNs. However, as the depth of the network exceeds a certain extent, the phenomenon of gradient

vanishing occurs, which directly leads to a deterioration in the network's performance. However, the skipping connection of residual blocks can retain the original image information, which is conducive to the back-propagation of the gradient, mitigating the phenomenon of accuracy decline and contributing to preventing network degradation. Therefore, some studies have built HSI classification networks through residual module connection and tried to solve the gradient vanishing problem of deep networks by means of stacking residual modules to achieve better classification results. The basic residual unit is shown Figure 7 and its expression as follows:

$$X_{i+1} = F(X_i, W_i) + H(X_i) \tag{15}$$

where X_i and X_{i+1} represent the inputs and outputs of the i-th residual unit. F is the residual function and H refers to the method of shortcut connection; $H(X_i) = X_i$ represents identity.

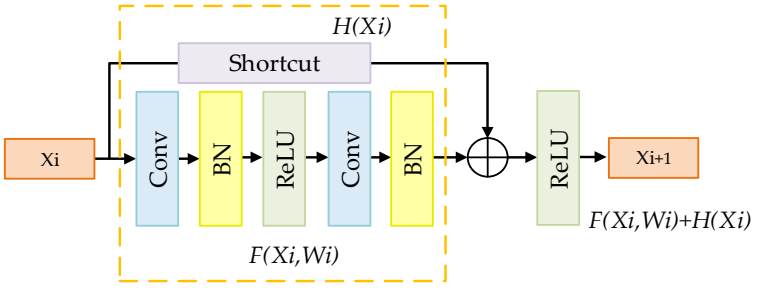


Figure 7. The structure of the basic residual unit.

Although some structures proposed by the above block have achieved relatively good results, the accumulation of basic modules leads to the explosive growth of parameters. In order to ensure that the features can be extracted further and the network load can be reduced simultaneously, we propose a hybrid residual module consisting of the residual block based on 2D-DSC and the residual block based on 3D-DSC, as shown in Figure 8.

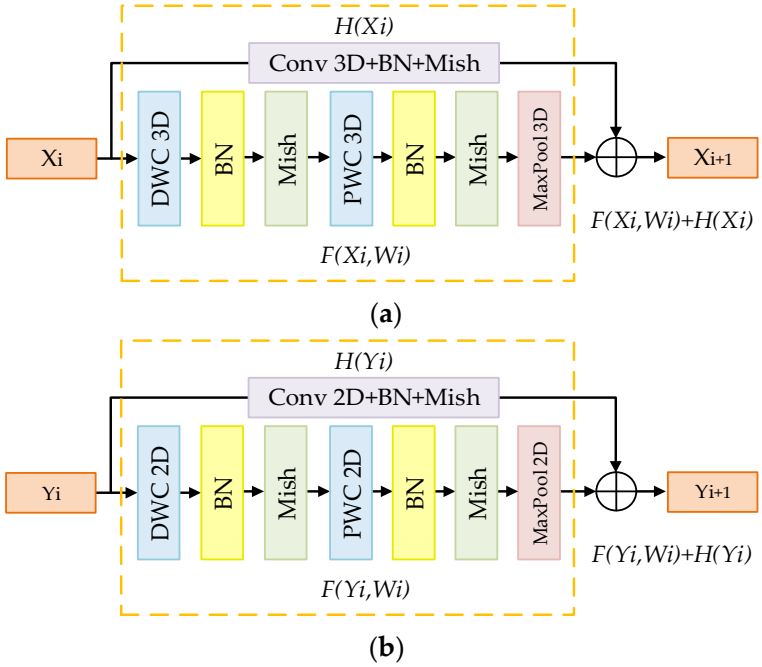


Figure 8. The structure of the hybrid residual module. (**a**) The residual block based on 3D-DSC; (**b**) the residual block based on 2D-DSC.

For the hybrid residual module, different residual blocks are used for the input of data of different dimensions. Specifically, the 3D-DSC residual block and 2D-DSC residual block are, respectively, used for 3D data input and 2D data input.

In contrast to the basic residual unit, our hybrid residual module omits the final rectified linear unit (ReLU). In the shortcut branch, we can express $H(X_i) = Mish(BN(Conv(X_i)))$. In the lower branch, we introduce depth-wise convolution (DWC) and pointwise convolution (PWC) to replace standard convolution, which alleviates the problem of parameter surge when residual blocks are superimposed. Finally, the characteristics of the lower branch are obtained through the maximum pooling layer. Specifically, the order of execution of the layers can be described as follows: DWC $\rightarrow$ BN $\rightarrow$ Mish $\rightarrow$ PWC $\rightarrow$ BN $\rightarrow$ Mish $\rightarrow$ MaxPool. As the network gradually goes deeper, the parameters do not increase significantly, so we can apply the lightweight residual module to our model, HDCDF.

3. Methodology

Figure 9 shows the architecture of the procedure of the HDCDF framework, which contains three parts. The purpose of data preprocessing is to increase the number of training samples and prepare for HDCDF network input.

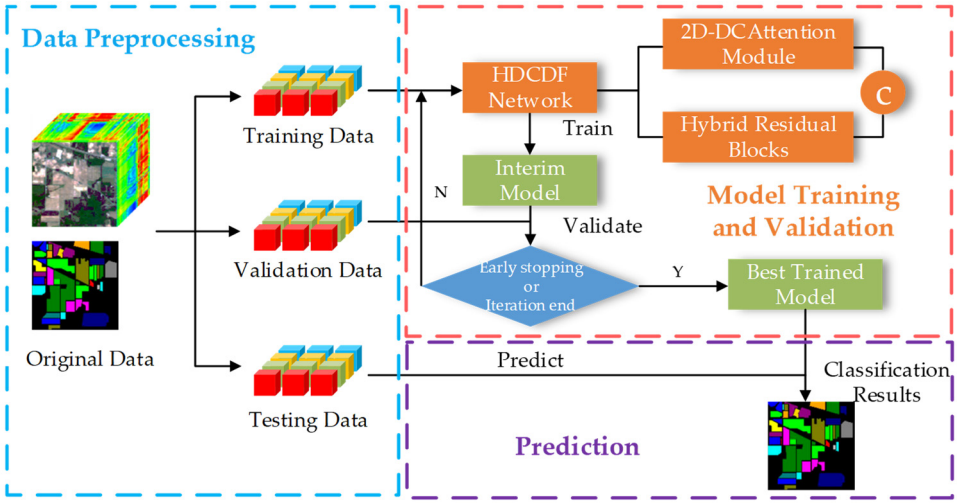


Figure 9. Procedure of hyperspectral image classification based on HDCDF.

In the model training and validation part, the training data and the validation data are used to update the network parameters and monitor the performance of models with different parameters, respectively, to obtain the best-trained model. Due to the early stopping strategy, the best-trained model can be obtained and effectively alleviate the overfitting problem. Finally, the test data are used to verify the performance of the best-trained model.

3.1. Data Preprocessing for Dataset Generation

We record the original spectral–spatial hyperspectral data cube as $X \in \mathbb{R}^{W \times H \times D}$, where X represents the original HSI data, W is the width, H is the height, and D is the number of spectral bands. Each pixel in HSI has a pixel sequence $Y = (y_1, y_2, \cdots, y_c) \in \mathbb{R}^{1 \times 1 \times C}$, where C donates the land cover categories. Due to the huge computational burden caused by the high dimensionality and existing redundant data in HSI, we apply the PCA strategy to the original HSI along spectral bands for spectral feature enhancement and to reduce redundant data. The original data $X \in \mathbb{R}^{W \times H \times D}$ are modified to $X \in \mathbb{R}^{W \times H \times B}$, where B refers to the number of spectral bands after PCA. The PCA not only reduces the number of spectral bands from D to B but also preserves the spatial dimension information,

which is of great significance for pixel classification. The data cube is divided into a number of small 3D neighboring patch sets, $Z \in R^{P \times P \times B}$, from X; the final label depends on the label of the center pixel. We set the missing adjacent pixels as zero when the core pixel is located at the edge of X so that all patches can cover the $P \times P$ size of the space and all spectral bands of HSI. Then, the patch set is randomly split into a training set Z_{train}, testing set Z_{test}, and validation set Z_{val}, and their respective labels are divided into Y_{train}, Y_{test}, and Y_{val}.

Considering the fact that training deep networks requires numerous samples while available training samples are limited, we adopt the date enhancement strategy to increase the number of training samples and extract more sufficient features, which can effectively avoid overfitting and the problem of sample imbalance, improving the generalization ability of the model HDCDF. With equal probability, patches are rotated 45 degrees, 90 degrees, or 135 degrees; vertically flipped; horizontally flipped; or kept the same. Then, the total number of training patches can be multiplied n times, compared with Z_{train}, obtaining better network parameters. We represent the enhanced training samples by Z_{en_train} and, accordingly, their corresponding new label set is Y_{train_en}.

3.2. Proposed HDCDF Structure

In this paper, in order to overcome the question mentioned in the Introduction, we propose a hybrid depth-wise separable CNN and dual-branch feature fusion model (HDCDF) for hyperspectral image classification. The whole structure of the HDCDF network is displayed in Figure 10, and the different dimensional figures of the same color in the legend represent operations on data of different dimensions.

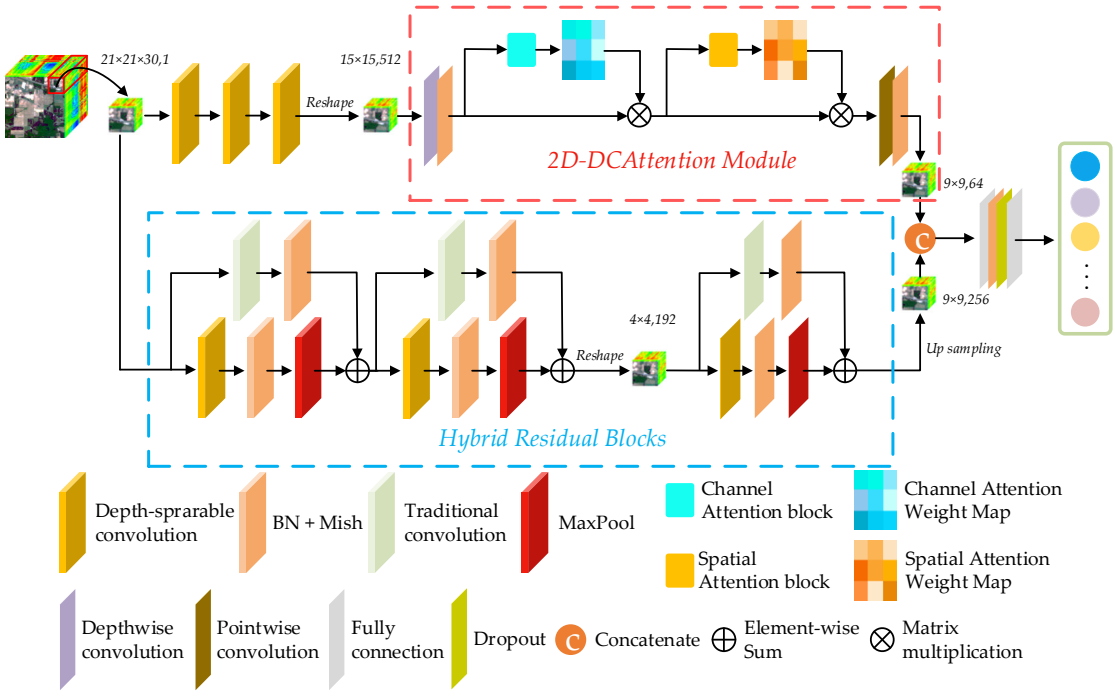


Figure 10. Structure of the proposed hybrid depth-wise separable CNN and dual-branch feature fusion (HDCDF).

In general, HDCDF is an end-to-end parallel network structure that extracts features from the training patches Z_{train} simultaneously and then fuses the features extracted from the two branches to obtain the joint feature F_f, which is sent into the classifier to obtain the classification map.

Specifically, the upper-branch network stream firstly extracts the abundant spectral–spatial features in the input image using continuous 3D depth-wise separable convolution (3D-DC). After reshaping the extracted features, the 2D-DCAttention module is used to further enhance the features of the spatial dimension and spectral dimension, while retaining key features and suppressing noise and irrelevant information. Finally, we obtain the feature F_u extracted from the upper branch. The mixed residual module is adopted in the lower-branch network flow. First, Z_{train} is processed through two 3D residual blocks to obtain a more comprehensive image feature and then through the 2D residual block after the reshape operation on features of the previous layer. Finally, the characteristics F_d of the lower branch are obtained by up-sampling. Hybrid residual blocks reduce the complexity of the lower branches by jumping connections, while retaining the relatively original image features. In this way, we obtain complementary feature F_f by fusing the feature F_d with more original image information and F_u with joint spectral–spatial features, which is passed into the classifier to select the best model.

4. Experimental Results

4.1. Dataset Description

To evaluate the classification performance of the HDCDF network, we used three representative HSI datasets in our experiments: Indian Pines (IP), University of Pavia (UP), and Salinas Valley (SV). The false-color composite, ground-truth maps, and details of each category are shown in Figures 11–13. Further details regarding these datasets are presented in the following.

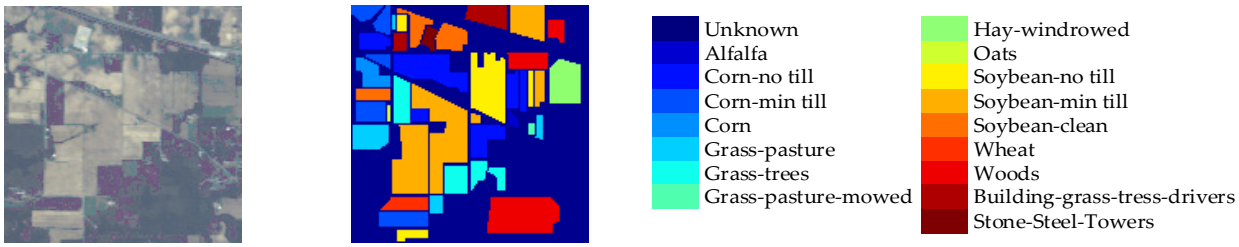


Figure 11. IP dataset. (**a**) False-color composite; (**b**) ground-truth map; (**c**) land cover categories in the IP dataset.

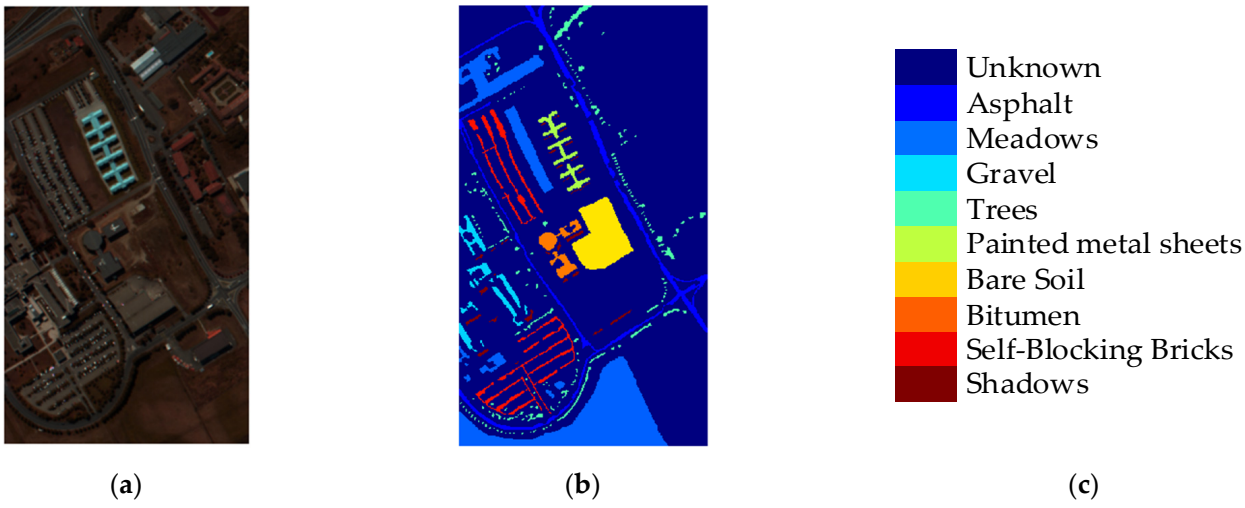


Figure 12. UP dataset. (**a**) False-color composite; (**b**) ground-truth map; (**c**) land cover categories in the UP dataset.

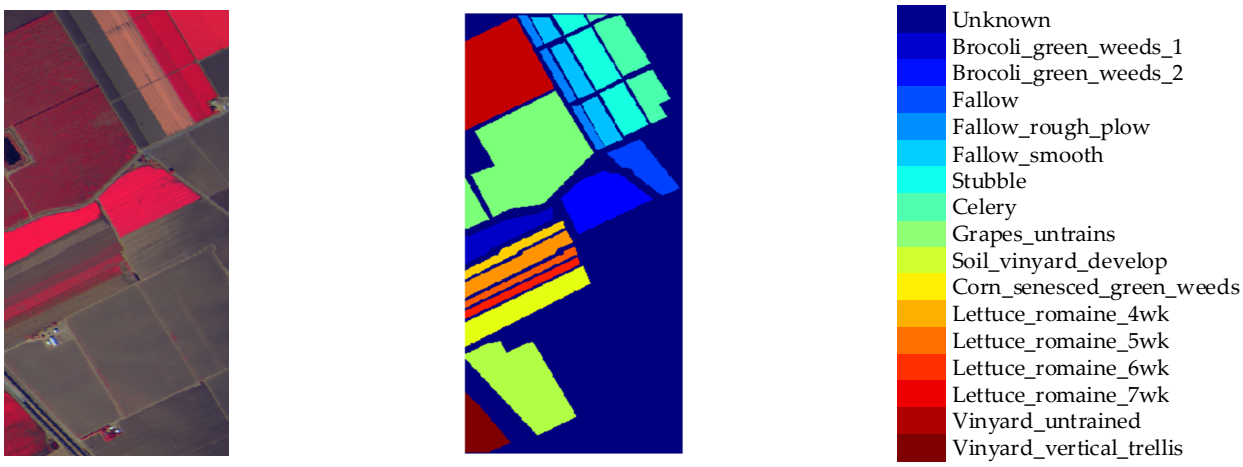


Figure 13. SV dataset. (**a**) False-color composite; (**b**) ground-truth map; (**c**) land cover categories in the SV dataset.

Indian Pines (IP) dataset: The hyperspectral image was acquired by the AVIRIS sensor (manufactured by NASA's Jet Propulsion Laboratory (JPL) in Pasadena, CA, USA) over the Indian Pines test site in north-western Indiana in 1992, consisting of 16 land cover classes, mainly including corn, soybean, and oats. This dataset consists of 145 × 145 pixels and 220 spectral bands in the wavelength scope of 0.4~2.5 μm. We removed 20 water absorption and low-signal-to-noise-ratio bands and kept the remaining 200 bands for our experiments.

University of Pavia (UP) dataset: The UP dataset was captured by the ROSIS-03 sensor (developed by the German Aerospace Center (DLR) in Oberpfaffenhofen, Germany) in Northern Italy in 2001, with a size of 610 × 340 and a high spatial resolution of 1.3 m per pixel. A total of 103 spectral bands were obtained after removing some low-signal-to-noise-ratio bands in the spectral coverage ranging from 0.43 to 0.86 μm. In addition, the existing ground truths are divided into nine classes, such as meadows, trees, bitumen, and so on.

Salinas Valley (SV) dataset: Collected by the AVIRIS sensor in 1998 over the Salinas Valley in California, this dataset features a high resolution of 3.7 m per pixel with a size of 512 × 217. Since 20 of the bands cannot be reflected by water, we selected the remaining 204 bands with a wavelength scope of 0.4 μm to 2.5 μm to conduct the experiments. The available ground truth for the SV contains 16 categories.

Deep learning structures are data-driven, meaning they need numerous labeled training samples to fully train a network for perfect classification results, but more data mean higher computation complexity and more time consumption. Therefore, we performed experiments to verify that the proposed HDCDF can maintain excellent classification performance in the case of very limited training samples. In Tables 1–3, we can see the details of each dataset used in our experiments. During the experiment, we set the size of the training sample and the validation sample at the minimum level. Since each class of UP and SV samples is more sufficient than IP, only 0.5% training samples and 0.5% validation samples are selected, respectively.

Table 1. The details of the Indian Pines (IP) dataset with 5% training samples.

No	Class	Total	Training	Val	Test
1	Alfalfa	46	2	2	42
2	Corn—no till	1428	71	71	1286
3	Corn—min till	830	42	42	746

Table 1. *Cont.*

No	Class	Total	Training	Val	Test
4	Corn	237	12	12	213
5	Grass—pasture	483	24	24	435
6	Grass—trees	730	37	37	656
7	Grass—pasture (mowed)	28	1	1	26
8	Hay—windrowed	478	24	24	430
9	Oats	20	1	1	18
10	Soybean—no till	972	49	49	874
11	Soybean—min till	2455	123	123	2209
12	Soybean—clean	593	30	30	533
13	Wheat	205	10	10	185
14	Woods	1265	63	63	1139
15	Building/grass/trees/drivers	386	19	19	348
16	Stone/Steel/Towers	93	5	5	83
	Total	10,249	513	513	9223

Table 2. The details of the University of Pavia (UP) dataset with 0.5% training samples.

No	Class	Total	Training	Val	Test
1	Asphalt	6631	33	33	6565
2	Meadows	18,649	93	93	18,463
3	Gravel	2099	10	10	2079
4	Trees	3064	15	15	3034
5	Painted metal sheets	1345	7	7	1331
6	Bare Soil	5029	25	25	4979
7	Bitumen	1330	7	7	1316
8	Self-Blocking Bricks	3682	18	18	3646
9	Shadows	947	5	5	937
	Total	42,776	213	213	42,350

Table 3. The details of the Salinas Valley (SV) dataset with 0.5% training samples.

No	Class	Total	Training	Val	Test
1	Brocoli_green_weeds_1	2009	10	10	1989
2	Brocoli_green_weeds_2	3726	19	19	3688
3	Fallow	1976	10	10	1956
4	Fallow_rough_plow	1394	7	7	1380
5	Fallow_smooth	2678	13	13	2652
6	Stubble	3959	20	20	3919
7	Celery	3597	18	18	3561
8	Grapes_untrains	11,271	56	56	11,159
9	Soil_vinyard_develop	6203	31	31	6141
10	Corn_senesced_green_weeds	3278	16	16	3246
11	Lettuce_romaine_4wk	1068	5	5	1058
12	Lettuce_romaine_5wk	1927	10	10	1907
13	Lettuce_romaine_6wk	916	5	5	906
14	Lettuce_romaine_7wk	1070	5	5	1060
15	Vinyard_untrained	7268	36	36	7196
16	Vinyard_vertical_trellis	1807	9	9	1789
	Total	54,147	270	270	53,607

4.2. Experimental and Optimal Parameter Settings

For the sake of validating the effectiveness and generalization of the HDCDF network, three widely available datasets are used and the experiments are performed using a computer with an NVIDIA GeForce GTX 1660 graphics card (manufactured by NVIDIA Corporation, Santa Clara, CA, USA) Super 6 GB, an Intel Core i7-8700K CPU (Intel, Santa Clara, CA, USA), and a memory of 16 GB. The software framework was PyTorch 1.12.0, CUDA 11.3, and the programming language was Python 3.9.18. For image cube display, we adopted the ENVI 5.6 software to process datasets. We used Adam Optimizer to optimize the network parameters with a batch size of 32. The number of epochs in the experiment was set to 50 and the learning rate was 0.001. In addition, we also adopted the early stopping strategy and set patience to 10, which means that if the validation loss is not improved after 10 consecutive epochs, early stopping will be triggered. For experimental evaluation, three widely used quantitative metrics are utilized to assess the HDCDF performance, as well as a confusion matrix ($C_{n\times n}$), a specific table layout used to visualize the performance of supervised learning algorithms. In the confusion matrix, the prediction labels and the actual labels are represented, respectively, by each column and row.

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix} \tag{16}$$

(1) Overall accuracy (OA) is the proportion of correctly classified samples.

$$p_o = \frac{\sum_{i=1}^{n} c_{ii}}{\sum_{i}^{n} \sum_{j}^{n} c_{ij}} \tag{17}$$

(2) Average accuracy (AA) is the average of all class accuracies.

$$p_a = \frac{1}{n} \times \sum_{i=1}^{n} \frac{C_{ii}}{C_{ij}} \tag{18}$$

(3) The Kappa coefficient (Kappa) is a robust measure of the degree of agreement between the ground-truth map and classification map.

$$p_e = \sum_{i=1}^{n} \left(\sum_{i}^{n} C_{ij} \times \sum_{j}^{n} C_{ij} \right) \tag{19}$$

$$k = \frac{p_o - p_e}{1 - p_e} \tag{20}$$

where p_o represents OA, p_a refers to AA, and p_e computes the measures of agreement between the true and predicted classes. $p_o - p_e$ describes the difference among OA values obtained by chance and actual accuracy; $1 - p_e$ portrays the maximum value for the difference. Usually, the Kappa falls between 0 and 1; the closer Kappa is to 1, the better and more stable the model classification effect is.

It is vital to find the optimal parameters to run the model. During dataset generation, the HSI cube is divided into overlapping small 3D patches of size $Z \in R^{P\times P\times B}$, where $P \times P$ refers to the spatial input size and B represents the number of spectral bands after PCA processing. This kind of processing method, instead of processing the entire HSI cube, can reduce the number of parameters and, at the same time, significantly reduces the required number of parameters. To obtain the best-fitting spatial size and amount of spectral bands, we conducted several experiments with different sizes and numbers of

bands to select the most suitable value. As a result, we concluded that the OAs begin to decline when the spatial size exceeds 21 × 21 and the value of *B* is over 30. Figure 14 shows the HDCDF classification performance with different spatial sizes and amounts of spectral bands.

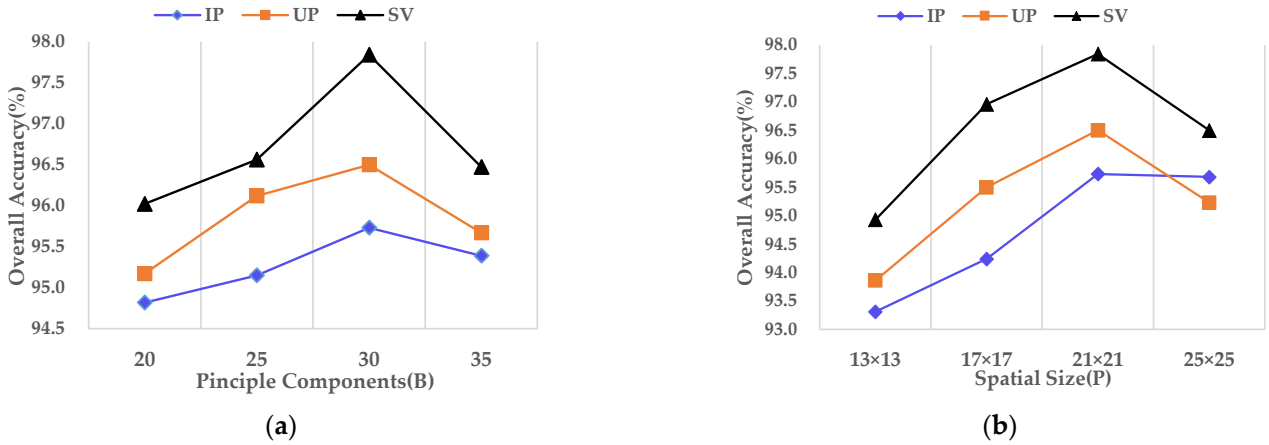


Figure 14. Results of parameter analysis. (**a**) The Effect of B on OAs on three datasets; (**b**) the effect of P on OAs on three datasets.

To assess the performance of our proposed structure, we also select some DL-related classification methods with better effects than comparative methods, including 3DCNN, the double-branch multi-attention mechanism network (DBMA) [24], SpectralFormer [29], the extended Vision Transformer (ExViT) [30], and the lightweight spectral–spatial neural architecture with a multi-attention feature extraction network (LSSMA) [43]. Specifically, SpectralFormer and ExViT are novel networks developed based on the transformer. SpectralFormer focuses on spectral characteristics in classification, and the author designed two key modules and integrated them into the framework of ViT, improving the ability to capture details of subtle spectral differences. ExViT is a new multi-modal deep learning framework that extends traditional ViT with minimal modification. Full 3D convolutions are used in 3DCNN to process both spatial and spectral information, but it requires millions of parameters. Inspired by ResNet, LSSMA uses convolutional kernels of different sizes to perform residual joins and feature fusion, reducing the number of trainable parameters and effectively preventing overfitting. Under the influence of an attention mechanism, DBMA includes two branches to extract spectral features and spatial features, respectively, which can reduce the interference between the two types of features.

4.3. Classification Maps and Categorized Results

In this section, to evaluate the classification performance of the proposed HDCDF, we present the detailed results of different methods under three datasets. Tables 4–6 and Figures 14–17 show the detailed accuracy and visual results of the HDCDF method compared with other methods. It can be seen from the experimental results that, although the model does not achieve optimal results in all categories, HDCDF performs the best in three indicators, OA, AA, and Kappa, with values of 95.73%, 96.51%, and 95.22% on the IP dataset, 96.5%, 97.33%, and 95.26% on the UP dataset, and 97.84%, 98.74%, and 95.46% on the SV dataset, showing a better classification ability on those categories and that other models do not perform well generally, compared with other networks.

To be specific, it is obvious that the overall classification effect of the proposed model on the UP and SV datasets is better than that on the IP dataset. Although we only selected 0.5% training samples for these two datasets, there are still enough training samples for each category of them compared with the IP dataset. As a result, SpectralFormer and ExViT,

based on the transformer design, are sufficiently trained in that the final classification results are higher than other networks and closer to HDCDF. However, the number of training samples on the IP dataset is limited, so the SpectralFormer and ExViT networks have lower classification results (OAs: SpectralFormer: 91.13%; ExViT: 94.59%) than HDCDF, which also shows that HDCDF can still maintain excellent classification performance in the case of insufficient samples. This may be due to the double-branch feature fusion structure of HDCDF, which helps to extract more abundant features, and thus the training results are better than other models. From the IP dataset, the classification effect of the 3DCNN model is not ideal as the model only relies on a 3D convolution operation; the extracted features are not as sufficient as DBMA with the introduction of an attention mechanism and LSSMA with the introduction of a residual module. In addition, we can see that the phenomenon of salt-and-pepper noise is severe in the 3DCNN network classification figure. In contrast, the HDCDF classification far outperforms the SpectralFormer and ExViT networks. The SpectralFormer network pays more attention to the role of spectral features in classification due to the large number samples in the IP dataset, so the correlation between space and spectral features should be fully utilized to improve classification accuracy. HDCDF includes convolutions of different sizes and dimensions for residual connection and feature fusion, which can extract deeper features and reduce the number of training parameters.

In the UP dataset, we find the sixth category in the middle of the picture; the 3DCNN and DBMA models have many pixels identified as the second category. This may be due to the similarity of the spectral features of the two categories, and the model cannot accurately identify the extracted features. The network proposed in this paper achieves almost 100% in this category. This result also proves the superiority of HDCDF in extracting features, which makes the problem of large spatial differences in spectral features of HSI effectively solved.

Table 4. Classification results of the IP dataset with 5% training samples and the best accuracy each class is in bold.

No	Class	3DCNN	DBMA	SpectralFormer	ExViT	LSSMA	Proposed
1	Alfalfa	83.93	95.75	93.41	97.64	94.99	**100**
2	Corn—no till	87.68	**96.62**	93.37	78.12	93.74	95.06
3	Corn—min till	89.04	94.26	90.25	94.72	83.52	**97.72**
4	Corn	82.33	92.98	89.04	90.50	91.66	**95.36**
5	Grass—pasture	82.36	93.65	**99.59**	97.59	68.74	95.16
6	Grass—trees	96.42	97.69	81.97	89.83	89.47	**100**
7	Grass—pasture (mowed)	84.28	66.84	99.34	98.54	95.76	**100**
8	Hay—windrowed	89.62	94.69	64.28	83.51	80.14	**99.81**
9	Oats	78.85	89.57	89.02	**93.61**	91.75	91.89
10	Soybean—no till	91.55	88.36	90.93	**98.38**	96.33	93.22
11	Soybean—min till	92.59	98.75	93.69	96.48	93.97	**100**
12	Soybean—clean	88.78	83.19	98.62	**100**	99.21	90.45
13	Wheat	84.02	95.65	95.34	98.04	84.74	**98.20**
14	Woods	93.61	95.57	90.47	87.74	85.52	**98.82**
15	Building/grass/trees/drivers	84.47	93.45	91.27	93.39	**94.42**	90.85
16	Stone/Steel/Towers	93.55	92.27	90.12	91.86	87.15	**99.75**
OA (%)		89.95	92.55	91.13	94.59	94.28	**95.73**
AA (%)		87.69	91.83	90.57	93.27	89.49	**96.51**
Kappa (%)		88.04	91.48	90.81	93.04	92.89	**95.22**

Table 5. Classification results of the UP dataset with 0.5% training samples and the best accuracy each class is in bold.

No	Class	3DCNN	DBMA	SpectralFormer	ExViT	LSSMA	Proposed
1	Asphalt	90.24	95.48	95.75	**98.24**	97.88	97.27
2	Meadows	96.48	**98.99**	98.92	98.86	98.16	98.91

Table 5. *Cont.*

No	Class	3DCNN	DBMA	SpectralFormer	ExViT	LSSMA	Proposed
3	Gravel	92.11	88.34	96.27	94.92	93.20	**96.36**
4	Trees	94.13	96.63	96.55	96.84	**97.01**	96.68
5	Painted metal sheets	91.04	96.71	96.65	96.85	98.42	**99.16**
6	Bare Soil	79.83	74.93	96.89	98.16	96.55	**99.19**
7	Bitumen	88.95	95.46	98.04	96.93	98.00	**98.66**
8	Self-Blocking Bricks	86.54	**92.61**	92.49	87.45	86.03	91.03
9	Shadows	97.15	97.35	97.14	97.16	97.23	**98.72**
OA (%)		90.81	93.27	96.83	96.25	95.41	**96.50**
AA (%)		90.72	92.94	96.52	96.16	95.83	**97.33**
Kappa (%)		89.64	93.04	93.48	90.11	92.89	**95.26**

Table 6. Classification results of the SV dataset with 0.5% training samples and the best accuracy each class is in bold.

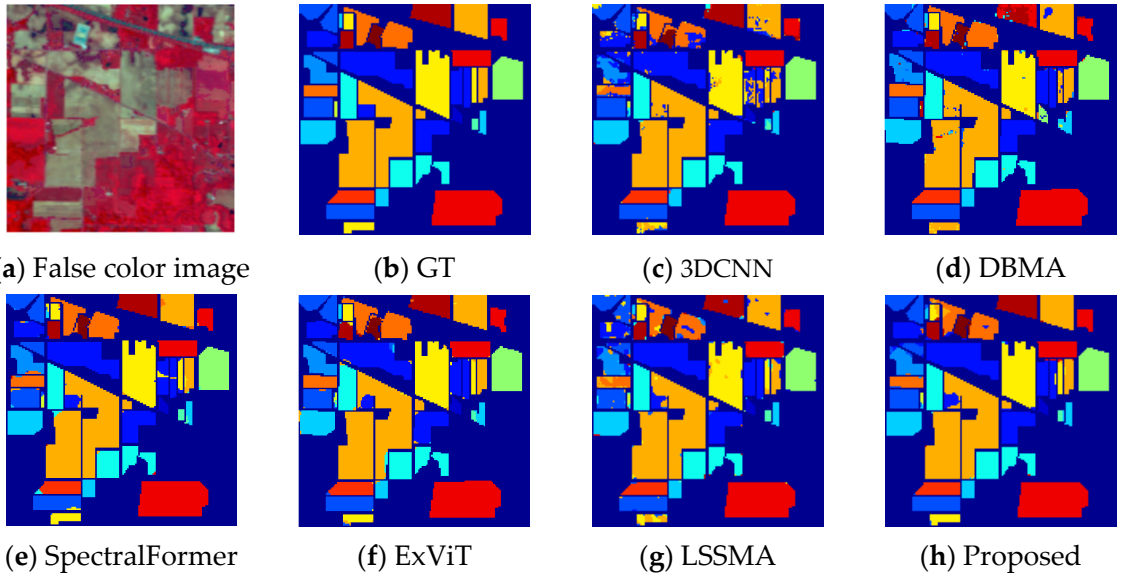
No	Class	3DCNN	DBMA	SpectralFormer	ExViT	LSSMA	Proposed
1	Brocoli_green_weeds_1	99.87	99.87	99.91	**100**	99.98	**100**
2	Brocoli_green_weeds_2	98.9	99.47	**100**	**100**	97.81	**100**
3	Fallow	91.29	98.9	98.29	89.93	87.44	**98.92**
4	Fallow_rough_plow	97.32	97.41	96.96	96.99	96.87	**98.74**
5	Fallow_smooth	93.64	94.13	**100**	**100**	89.66	98.95
6	Stubble	97.91	98.15	98.94	**99.63**	98.72	99.29
7	Celery	99.67	99.71	99.75	**100**	99.96	99.41
8	Grapes_untrains	74.62	89.64	93.69	92.13	88.2	**99.64**
9	Soil_vinyard_develop	98.02	98.21	98.54	97.85	96.97	**99.70**
10	Corn_senesced_green_weeds	90.01	94.27	92.38	91.05	92.14	**95.50**
11	Lettuce_romaine_4wk	90.26	97.74	89.9	93.13	95.91	**97.77**
12	Lettuce_romaine_5wk	95.24	98.93	95.67	95.4	98.93	**99.11**
13	Lettuce_romaine_6wk	95.61	96.62	94.55	97.7	**100**	97.75
14	Lettuce_romaine_7wk	89.22	94.42	96.64	96.65	97.87	**98.68**
15	Vinyard_untrained	74.96	90.29	**96.55**	91.72	80.07	96.42
16	Vinyard_vertical_trellis	98.82	99.67	99.71	99.74	91.14	**100**
OA (%)		91.05	94.59	92.55	92.41	94.28	**97.84**
AA (%)		92.84	96.71	96.97	96.37	94.48	**98.74**
Kappa (%)		88.04	93.04	91.48	91.20	92.89	**95.46**

At the same time, from the classification diagram, we can see that the network proposed in this paper handles the boundary well. Taking class 16, located in the lower left corner of Figure 17, which only has nine training samples in the SV dataset, as an example, the method proposed performs the best among these models and finally obtains a remarkable result of 100%, while the performance of other networks is not accurate, especially LSSMA. It can be seen from Figure 17 that except for HDCDF and LSSMA with poor recognition effects, there are a few cases of incorrect classification on the boundary of other models.

In general, the accuracy of the proposed HDCDF network on most of the three datasets exceeds 93% for each class; it has strong feature extraction capability and is superior to other models in boundary processing.

The computational efficiency of the HDCDF model is apparent in terms of parameter training and testing times on the IP dataset using 5% training samples in Table 7. A network with good performance should balance accuracy and efficiency properly. The above experiments have proven that our proposed method can obtain excellent performance with limited data. From Table 7, on account of 3DCNN containing more parameters (1481.6 K) to be trained, other methods take less time than 3DCNN. HDCDF takes less time and memory compared to models based on transformer. Although LSSMA possesses fewer parameters

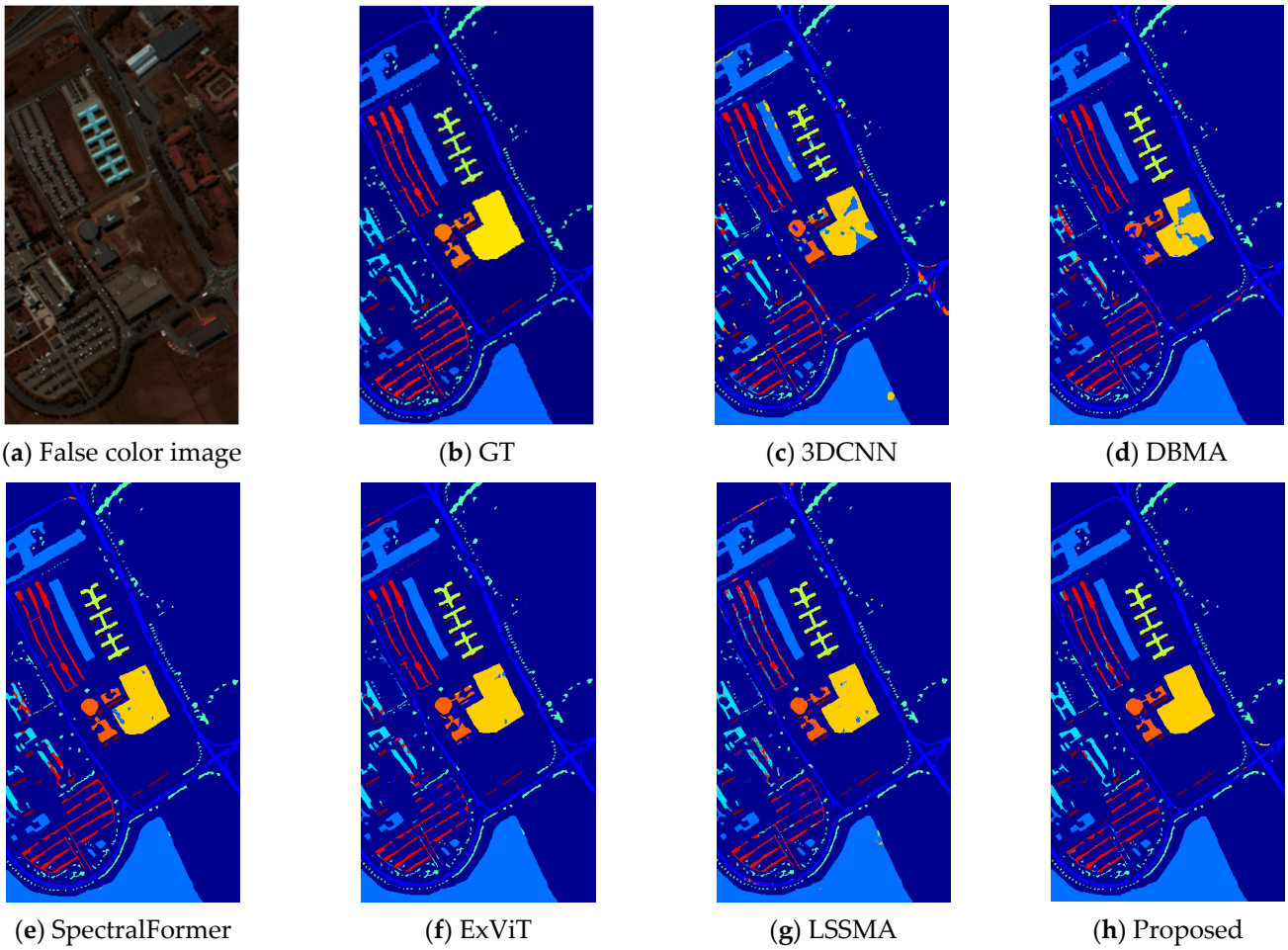
(8.7 K) than the proposed model (22.5 K), HDCDF has a shorter training time and a higher accuracy of nearly 1.5%.



(**a**) False color image (**b**) GT (**c**) 3DCNN (**d**) DBMA

(**e**) SpectralFormer (**f**) ExViT (**g**) LSSMA (**h**) Proposed

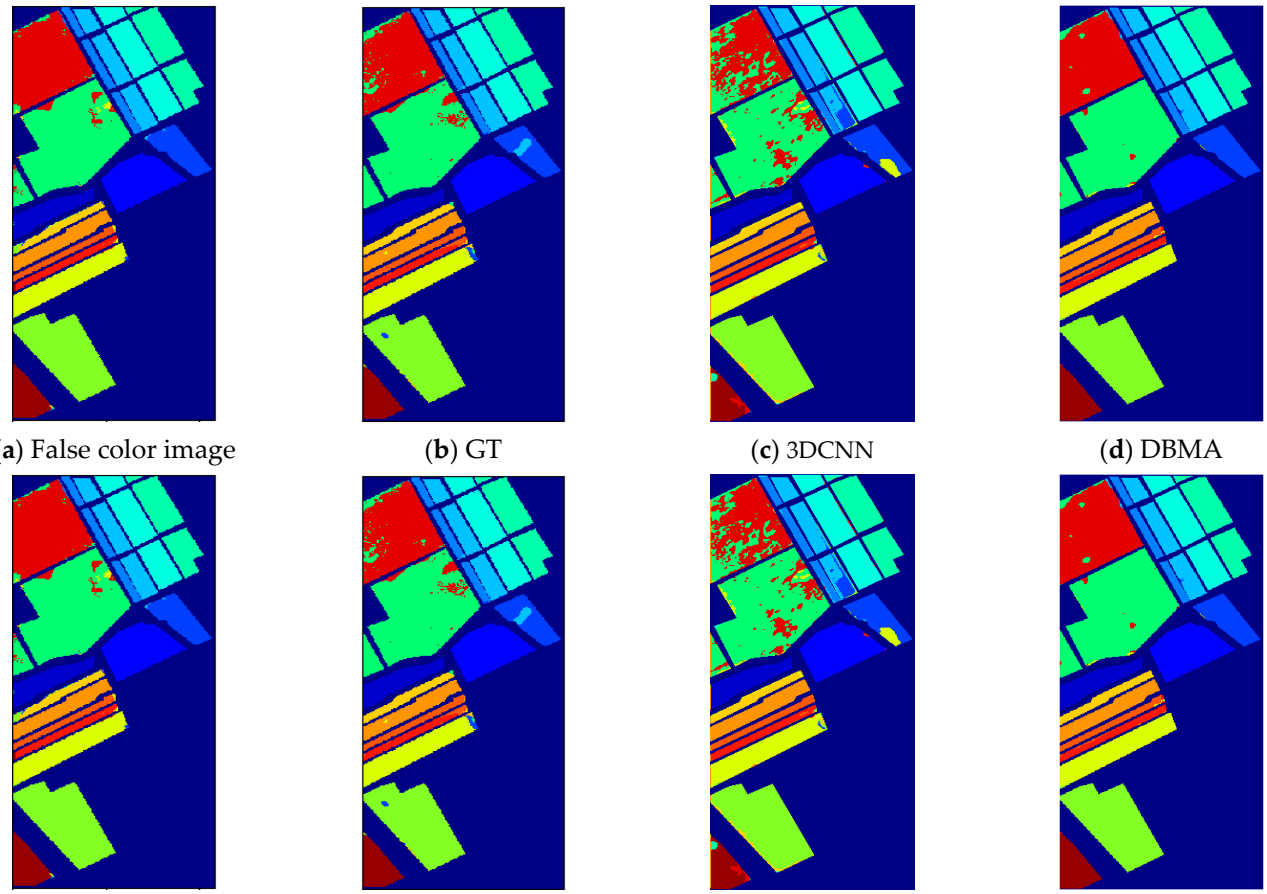
Figure 15. Classification maps of the different models for the IP dataset using 5% training samples: (**a**) false-color image; (**b**) ground-truth map; (**c**–**h**) maps with different models.



(**a**) False color image (**b**) GT (**c**) 3DCNN (**d**) DBMA

(**e**) SpectralFormer (**f**) ExViT (**g**) LSSMA (**h**) Proposed

Figure 16. Classification maps of the different models for the UP dataset using 0.5% training samples: (**a**) false-color image; (**b**) ground-truth map; (**c**–**h**) maps with different models.



(**a**) False color image (**b**) GT (**c**) 3DCNN (**d**) DBMA

(**e**) SpectralFormer (**f**) ExViT (**g**) LSSMA (**h**) Proposed

Figure 17. Classification maps of the different models for the SV dataset using 0.5% training samples: (**a**) false-color image; (**b**) ground-truth map; (**c**–**h**) maps with different models.

Table 7. Complexity analysis of different models and our method on the IP dataset using 5% training samples in 16 classes.

Methods	Training (s)	Testing (s)	Parameter (K)
3DCNN	96.69	8.49	1481.6
DBMA	82.60	11.17	609.0
SpectralFormer	73.06	9.85	351.2
ExViT	71.64	9.18	346.4
LSSMA	61.72	10.41	8.7
Proposed	70.11	9.35	22.5

5. Discussion

In this section, different experiments are conducted to further explore HDCDF's performance. We discuss the assessment of the dual-branch structure, the ablation experiments to assess the performance of the hybrid DSC, and the results of the different activation functions.

5.1. Assessment of Dual-Branch Structure

Our proposed HDCDF is an end-to-end dual-branch structure, which combines two joint feature extraction-independent branches, which are the upper-branch network based on 2D-DCAttention and the lower-branch network based on hybrid residual modules. Theoretically speaking, the information obtained by the single-branch network is very limited, while the features extracted by the two-branch network with different methods are

usually richer and more comprehensive than those extracted by the single-branch network, so the two-branch network can obtain better results in the classification task. To verify the conjecture, we split the HDCDF network into an attention branch and residual branches and conducted experiments on three datasets. As can be seen from Table 8, the performance of our proposed two-branch-structure HDCDF on the three datasets is superior to each single branch of HDCDF; the OA is at least 1.2% higher than that of a single branch. Compared with the residual branch, the attention branch obtains higher classification accuracy through three layers of 3D depth-wise separable convolution that further strengthens the features in the spatial dimension and spectral dimension. Compared with the attention branch, the residual module retains the original classification features and also achieves relatively good classification results.

Table 8. Classification results of each single branch and HDCDF.

Methods	IP	UP	SV
Attention Branch	93.42	95.71	96.03
Residual Branch	81.26	88.68	90.19
HDCDF	95.73	96.50	97.84

In short, the two branches in the double-branch structure can complement each other to obtain more abundant features. In order to obtain a better classification effect, it is necessary to choose the double-branch structure over the single-branch structure.

5.2. Accuracy Comparison of the Proposed Model with 2D-DC Versus 3D-DC

To validate the effectiveness of integrating 2D-DC and 3D-DC in the HDCDF network, we develop two supplementary models for comparative analysis. The first model exclusively uses 3D-DC by replacing the 2D-DC in two branches of the proposed model, while maintaining the output sizes of the feature-extracted branches as in the original model. The second model employs 2D-DC exclusively, following a similar design. The comparative OA results, illustrated in Figure 18, demonstrate that the proposed hybrid depth-wise separable convolution structure consistently achieves superior classification performance across varying numbers of training samples in three datasets. Generally, the classification accuracies of models using 2D-DC or 3D-DC alone are comparable to the original model, with the 2D-DC-only model showing slightly lower performance. This reduction in performance is attributable to 2D-DC's inability to capture the inter-spectral band information of HSI, directly resulting in lower classification accuracy.

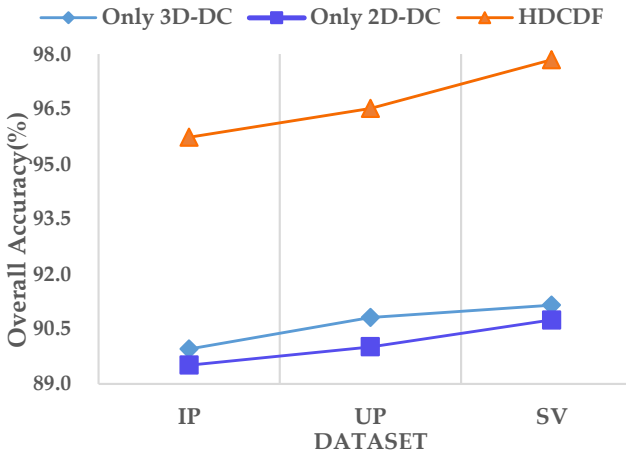


Figure 18. The classification performance of the proposed method employing hybrid depth-wise separable convolution, that using only 3D-DC, and that using only 2D-DC.

Overall, HDCDF derives more distinctive features by means of the hybrid integration of 2D-DC and 3D-DC, substantially enhancing the final classification performance.

5.3. Effectiveness of the Mish Activation Function

In Section 2.2, we explain why we chose Mish as the activation function instead of ReLU. Here, we will test the two functions separately to further verify the much better performance of HDCDF based on Mish. As you can see from Figure 19, HDCDF based on Mish surpasses HDCDF based on ReLU. Specifically, there was an increase of 2.27%, 1.97%, and 2.84% on the IP, UP, and SV datasets, respectively. Thus, the Mish activation function is more suitable than ReLU to deal with the nonlinear relationship between the spectral and spatial dimensions of hyperspectral images.

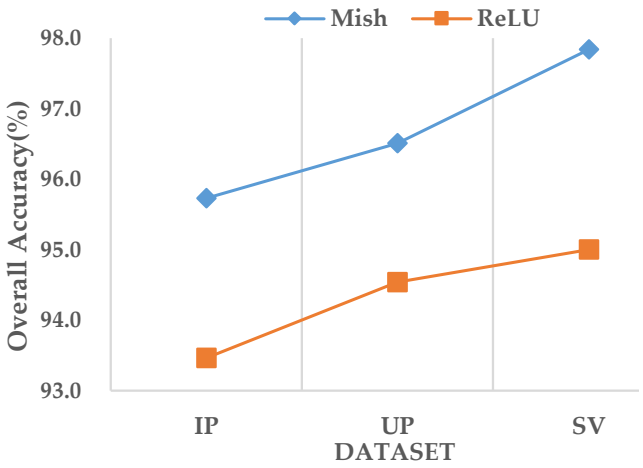


Figure 19. Results of different activation functions.

6. Conclusions

To address the challenges associated with the limited labeled training samples and high computational complexity for HSI classification, this study proposes an end-to-end dual-branch network structure, HDCDF. Instead of full 3D convolutions, we use mixed 2D and 3D depth-wise separable convolutions, offering computational efficiency. The input of the HDCDF framework is achieved after a series of data preprocessing tasks, including the PCA strategy to reduce dimensionality and rotation with equal probability to increase the number of trainable samples. Based on hybrid depth-wise separable convolutions, 2D-DCAttention module, and hybrid residual blocks, we design two branches to simultaneously capture joint spectral–spatial features and introduce some strategies to prevent overfitting involving early stopping and a dropout layer. To accelerate the counter-propagation and convergence, Mish is introduced as the activation function, taking full advantage of the spectral–spatial correlation of HSI. Experimental results demonstrate the validity and generalization of the HDCDF. In detail, the overall accuracy of the three widely used datasets is as high as 95.73%, 96.5%, and 97.84%, respectively, in cases with limited training samples.

A prospective avenue for our research involves extending the framework's application to forthcoming hyperspectral datasets, beyond just the existing public datasets. The main focus is still on creating models that can make the most out of a small amount of training data and maintain lower computational costs.

Author Contributions: Conceptualization, H.D. and Q.L.; methodology, Y.Y. and H.D.; software, Y.Y.; validation, H.D.; resources, H.D. and Q.L.; data curation, Y.Y.; writing—original draft, Y.Y.; writing—review and editing, Y.Y. and H.D.; visualization, Y.Y.; supervision, H.D. and Q.L.; funding acquisition, H.D and Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tianjin Science and Technology Planning Project, grant number 23YDTPJC00350.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The IP, UP, and SV datasets can be obtained from www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (accessed on 23 December 2024).

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

References

1. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.F.; 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.* **2022**, *15*, 968–999. [CrossRef]
2. Kumar, B.; Dikshit, O.; Gupta, A.; Singh, M.K. Feature extraction for hyperspectral image classification: A review. *Int. J. Remote Sens.* **2020**, *41*, 6248–6287. [CrossRef]
3. Xin, Z.; Li, Z.; Xu, M.; Wang, L.; Ren, G.; Wang, J.; Hu, Y. Feature disentanglement based domain adaptation network for cross-scene coastal wetland hyperspectral image classification. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103850. [CrossRef]
4. Ram, B.G.; Oduor, P.; Igathinathane, C.; Howatt, K.; Sun, X. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Comput. Electron. Agric.* **2024**, *222*, 109037. [CrossRef]
5. Li, X.; Li, Z.; Qiu, H.; Chen, G.; Fan, P. Soil carbon content prediction using multi-source data feature fusion of deep learning based on spectral and hyperspectral images. *Chemosphere* **2023**, *336*, 139161. [CrossRef]
6. Kumar, D.; Kumar, D. Hyperspectral image classification using deep learning models: A review. *J. Phys. Conf. Ser.* **2021**, *129*, 012087. [CrossRef]
7. Okwuashi, O.; Ndehedehe, C.E. Deep support vector machine for hyperspectral image classification. *Pattern Recognit.* **2020**, *103*, 107298. [CrossRef]
8. Jin, S.; Zhang, F.; Zheng, Y.; Zhou, L.; Zuo, X.; Zhang, Z.; Zhao, W.; Zhang, W.; Pan, X. CSKNN: Cost-sensitive K-Nearest Neighbor using hyperspectral imaging for identification of wheat varieties. *Comput. Electr. Eng.* **2023**, *111*, 108896. [CrossRef]
9. Tong, F.; Zhang, Y. Spectral–spatial and cascaded multilayer random forests for tree species classification in airborne hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4411711. [CrossRef]
10. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 809–823. [CrossRef]
11. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098. [CrossRef]
12. Gao, H.; Zhang, Y.; Chen, Z.; Li, C. A multiscale dual-branch feature fusion and attention network for hyperspectral images classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8180–8192. [CrossRef]
13. Zhang, X.; Liang, Y.; Li, C.; Huyan, N.; Jiao, L.; Zhou, H. Recursive autoencoders-based unsupervised feature learning for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1928–1932. [CrossRef]
14. Esmaeili, M.; Abbasi-Moghadam, D.; Sharifi, A.; Tariq, A.; Li, Q. Hyperspectral image band selection based on CNN embedded GA (CNNeGA). *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1927–1950. [CrossRef]
15. Chen, S.-Y.; Hsu, K.-H.; Kuo, T.-H. Hyperspectral Target Detection-Based 2D-3D Parallel Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 9451–9469. [CrossRef]
16. Zhao, R.; Zhang, C.; Xue, D. A multi-scale multi-channel CNN introducing a channel-spatial attention mechanism hyperspectral remote sensing image classification method. *Eur. J. Remote Sens.* **2024**, *57*, 2353290. [CrossRef]
17. Zhang, M.; Liu, L.; Jin, Y.; Lei, Z.; Wang, Z.; Jiao, L. Tree-shaped multiobjective evolutionary CNN for hyperspectral image classification. *Appl. Soft Comput.* **2024**, *152*, 111176. [CrossRef]
18. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
19. Dong, Y.; Gao, D.; Li, Y.; Shi, G.; Liu, D. Degradation estimation recurrent neural network with local and non-local priors for compressive spectral imaging. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5520115. [CrossRef]
20. Yu, Z.; Cui, W. Robust hyperspectral image classification using generative adversarial networks. *Inf. Sci.* **2024**, *666*, 120452. [CrossRef]
21. 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 Trans. Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281. [CrossRef]
22. Rodarmel, C.; Shan, J. Principal component analysis for hyperspectral image classification. *Surv. Land Inf. Sci.* **2002**, *62*, 115–122.

23. Ahmad, M.; Butt, M.H.F.; Mazzara, M.; Distefano, S.; Khan, A.M.; Altuwaijri, H.A. Pyramid hierarchical spatial-spectral transformer for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 17681–17689. [CrossRef]
24. 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]
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 Trans. Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88. [CrossRef]
26. Zhou, J.; Zeng, S.; Gao, G.; Chen, Y.; Tang, Y. A novel spatial-spectral pyramid network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5519314. [CrossRef]
27. Zhao, Z.; Xu, X.; Li, S.; Plaza, A. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5511817. [CrossRef]
28. Alexey, D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. 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]
30. Yao, J.; Zhang, B.; Li, C.; Hong, D.; Chanussot, J. Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5514415. [CrossRef]
31. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [CrossRef]
32. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sensors* **2015**, *2015*, 258619. [CrossRef]
33. Ge, Z.; Cao, G.; Li, X.; Fu, P. Hyperspectral image classification method based on 2D–3D CNN and multibranch feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5776–5788. [CrossRef]
34. Yu, C.; Han, R.; Song, M.; Liu, C.; Chang, C.-I. A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2485–2501. [CrossRef]
35. Fang, B.; Li, Y.; Zhang, H.; Chan, J.C.-W. Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism. *Remote Sens.* **2019**, *11*, 159. [CrossRef]
36. Kanthi, M.; Sarma, T.H.; Bindu, C.S. A 3D-deep CNN based feature extraction and hyperspectral image classification. In Proceedings of the 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), Ahmedabad, India, 1–4 December 2020; pp. 229–232.
37. Zhou, J.; Zeng, S.; Xiao, Z.; Zhou, J.; Li, H.; Kang, Z. An enhanced spectral fusion 3D CNN model for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 5334. [CrossRef]
38. Zhang, Z.; Gao, D.; Liu, D.; Shi, G. Spectral-spatial domain attention network for hyperspectral image few-shot classification. *Remote Sens.* **2024**, *16*, 592. [CrossRef]
39. Chhapariya, K.; Buddhiraju, K.M.; Kumar, A. A Deep Spectral-Spatial Residual Attention Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15393–15406. [CrossRef]
40. Wang, Y.; Li, K.; Xu, L.; Wei, Q.; Wang, F.; Chen, Y. A depthwise separable fully convolutional ResNet with ConvCRF for semisupervised hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4621–4632. [CrossRef]
41. Imani, M.; Ghassemian, H. An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. *Inf. Fusion* **2020**, *59*, 59–83. [CrossRef]
42. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
43. Ding, S.; Ruan, X.; Yang, J.; Sun, J.; Li, S.; Hu, J. LSSMA: Lightweight Spectral-Spatial Neural Architecture with Multi-Attention Feature Extraction for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6394–6413. [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.