

Article

Attention-Embedded Triple-Fusion Branch CNN for Hyperspectral Image Classification

Erlei Zhang , Jiayi Zhang, Jiaxin Bai, Jiarong Bian, Shaoyi Fang, Tao Zhan  and Mingchen Feng * 

School of Information Engineering, Northwest A&F University, Xi'an 712100, China;
erlei.zhang@nwafu.edu.cn (E.Z.); jiayiz@nwafu.edu.cn (J.Z.); bai123@nwafu.edu.cn (J.B.);
jiarong98@nwafu.edu.cn (J.B.); shaoyi.fang@nwafu.edu.cn (S.F.); omegazhant@gmail.com (T.Z.)

* Correspondence: mingchen@nwafu.edu.cn

Abstract: Hyperspectral imaging (HSI) is widely used in various fields owing to its rich spectral information. Nonetheless, the high dimensionality of HSI and the limited number of labeled data remain significant obstacles to HSI classification technology. To alleviate the above problems, we propose an attention-embedded triple-branch fusion convolutional neural network (AETF-Net) for an HSI classification. The network consists of a spectral attention branch, a spatial attention branch, and a multi-attention fusion branch (MAFB). The spectral branch introduces the cross-channel attention to alleviate the band redundancy problem in high dimensions, while the spatial branch preserves the location information of features and eliminates interfering image elements by a bi-directional spatial attention module. These pre-extracted spectral and spatial attention features are then embedded into a novel MAFB with large kernel decomposition technique. The proposed AETF-Net achieves multi-attention features reuse and extracts more representative and discriminative features. Experimental results on three well-known datasets demonstrate the superiority of the method AETF-Net.

Keywords: hyperspectral image classification; attention mechanism; feature fusion; deep learning



Citation: Zhang, E.; Zhang, J.; Bai, J.; Bian, J.; Fang, S.; Zhan, T.; Feng, M. Attention-Embedded Triple-Fusion Branch CNN for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 2150. <https://doi.org/10.3390/rs15082150>

Academic Editor: Javier Marcello

Received: 9 March 2023

Revised: 3 April 2023

Accepted: 14 April 2023

Published: 19 April 2023



Copyright: © 2023 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

Hyperspectral remote sensing can obtain the intrinsic characteristics and change patterns of objects by recording the electromagnetic wave characteristics without direct contact, making it a cutting-edge remote sensing technology [1]. Hyperspectral imaging (HSI) can record the spatial information under each waveband and the spectral information under the same position. Therefore, it has excellent application prospects in many fields, such as agriculture and forestry [2–5], ocean [6], disaster [7], mineral exploration [8,9], and urban construction [10–12]. HSI classification assigns category labels to each pixel based on sample features, which is increasingly becoming a key technology in hyperspectral remote sensing.

In the first two decades of the evolution of HSI classification, there were many machine learning algorithms based on hand-crafted features from the perspective of learning spectral and spatial features, for instance, spectral angle map [13], support vector machine [14], sparse representation [15], manifold learning [16], Markov Random Fields [17], Morphological Profiles [18], Random Forests [19], etc. However, due to the significant variability among different objects, classification algorithms based on manual feature extraction face challenges in fitting an optimal set of features for different objects and require greater robustness and discriminability..

Recently, studies on HSI classification have heavily focused on deep learning (DL) technology, since it could adaptively extract features from the input data in a hierarchical manner [20–22]. This allowed DL to learn data features in both spectral and spatial dimensions without requiring prior statistical knowledge of the input data. Chen et al. [23] first introduced DL to the HSI classification by applying deep Stacked Auto-Encoder (SAE). Similarly, in [24], the feasibility of using deep belief network (DBN) for HSI classification was

investigated. However, implementing SAE and DBN could potentially lead to decreased performance, as they use complex structures to modify the input data [25]. Researchers later discovered that Convolutional Neural Networks (CNNs) [26] could effectively extract multi-level features from large samples, thus eliminating the need for complicated feature extraction techniques. Hu et al. [27] first applied a one-dimensional CNN (1D-CNN) to HSI classification and obtained greater classification accuracy than many conventional machine learning techniques. Nevertheless, 1D-CNN has limited ability to capture spatial relationships between features in the input data. In contrast, two-dimensional CNN (2D-CNN) [28] learns how pixels in an image are related, allowing it to capture complex spatial patterns that are important for accurate image classification. However, it may struggle to capture the spectral relationships between features in the input data, as it considers the different spectral bands only as separate channels of the image. To incorporate the advantages of both 1D-CNN and 2D-CNN, researchers have attempted various methods. Yu et al. [29] utilized a 1D-CNN to extract spectral features and a 2D-CNN to extract spatial-spectral features, resulting in highly accurate classification. Conversely, the three-dimensional CNN (3D-CNN) [30] was proposed to operate on 3D HSI data and was capable of learning both spatial and spectral relationships between features in the input data, compensating for the weaknesses of 2D-CNNs. Nowadays, CNNs have gained significant attention and popularity among scholars [31], as evidenced by recent studies. Zhong et al. [32] proposed a spectral-spatial residual network (SSRN) that combines 3D-CNN for extracting discriminative features. Li et al. [33] developed a two-branch dual attention network (DBDA) that integrates spectral and spatial attention mechanisms for refining extracted feature maps. Yan et al. [34] designed a dual-branch network structure to relieve the issue of insufficient samples in HSI classification by incorporating transfer learning. Through this novel network structure, both [33] and [34] investigated how multimodel features can be used to improve HSI task performance. Although CNNs are well adapted to the high-dimensional and complex features of HSIs, high computational complexity arises, and its classification accuracy can suffer as a result of samples with insufficient data annotation. Furthermore, CNNs may require more refined feature extractors for specific tasks, and CNN models are prone to problems such as overfitting in small samples.

In supervised learning, sufficient labeled samples are required to provide a foundation for the classification algorithm [35]. However, labeling the samples pixel by pixel is time consuming and costly. Thus, the limited number of labeled samples and high-dimensional data can lead to the generation of the Hughes phenomenon [36], a type of model overfitting caused by insufficient training data, which affects classification accuracy heavily. Zhang et al. [37] proposed a lightweight 3D network based on transfer learning to address the sample-limited problem. Sellami et al. [38] proposed a semi-supervised network with adaptive band selection to reduce the dimensional redundancy and alleviate the Huges phenomenon. Although deeper networks can extract richer features to achieve high classification accuracy, a problem arises when the number of training samples is vastly smaller than the data dimensionality, leading to the explosive growth of parameters and vanishing gradients during the training process. Li et al. [39] designed a depth-wise separable Res-Net framework, which permitted separating spectral and spatial information in HSI and reduced network size to avoid overfitting issues. CNNs have shown remarkable performance in HSI classification tasks. Researchers have proposed various techniques, including transfer learning, adaptive band selection, and depth-wise separable networks, to improve the classification accuracy and robustness of the HSI small-sample classification model. However, convolution operations tend to assign equal weights to all pixels or bands in an image, despite the fact that some pixels and bands may be more beneficial for classification than others, or may even interfere with classification.

Currently, the introduction of an attention mechanism provides a solution to the aforementioned issue [40–43]. The attention mechanism draws inspiration from the visual focus region of the human brain, which aids the network in concentrating on significant regions while ignoring irrelevant ones and performing adaptive weight fitting on features. This

enhances the efficiency of feature extraction in models and reduces the need for unnecessary computation and data preprocessing, thereby making it a promising approach for HSI classification. Yu et al. [44] proposed a spatial-spectral dense CNN framework based on a feedback attention mechanism to extract high-level semantic features. Roy et al. [45] proposed an end-to-end trained adaptive spectral-spatial kernel improved residual network (A2S2K) with an attention-based mechanism to capture discriminative features for HSI classification. Li et al. [46] proposed a multi-attention fusion network (MAFN) that employs spatial and spectral attention mechanisms, respectively, to mitigate the effects of band redundancy and interfering pixels. Xue et al. [47] proposed the attention-based second-order pooling network (A-SPN) for modeling distinct and representative features by training the model with adaptive attention weights and second-order statistics. The attention mechanism learns more effective feature information but can lead to overfitting when the sample size is limited. Additionally, the high dimensional data of hyperspectral data carry a large amount of redundant information. The traditional single-attention mechanism needs to locate adequate information quickly and accurately, resulting in the need for a deeper network.

We propose an attention-embedded triple-branch fusion convolutional neural network (AETF-Net) for HSI classification to address the aforementioned issue. As is shown in Figure 1, the network comprises a spectral attention branch, a spatial attention branch, and a multi-attention fusion branch (MAFB). The spectral attention branch and spatial attention branch, respectively, address the issues of feature redundancy and correlation between spectral and spatial dimensions. We design a global band attention module (GBAM) in the spectral branch with a novel SMLP to extract more discriminative band features. In the spatial branch, we reference a bi-directional spatial attention module (BSAM) to extract spatial feature information in both horizontal and vertical directions. To incorporate the extracted spectral and spatial features and reduce the computational cost, we introduce the large kernel decomposition technique in the MAFB, which replaces large kernel convolution operation with some small kernel depth convolution and deep dilated convolution. In the proposed AETF-Net, multiple kinds of attention are used and fused to provide a reference basis for the relative importance of bands and pixels for 3D convolution with different weight values. Consequently, the proposed AETF-Net ensures efficient feature extraction while avoiding the gradient disappearance and feature dissipation issues caused by deep neural networks. In conclusion, the main contributions of this paper are as follows.

1. A novel multi-attention-based module is introduced that incorporates spatial attention, spectral attention, and joint spatial-spectral attention. The proposed approach embeds spatial and spectral feature information into each level of the joint spatial-spectral feature extraction module via cascading to compensate for the feature loss issue of the deep neural network.
2. An improved spectral feature extraction mechanism is designed to generate more accurate band features and weighting information. Moreover, we introduce an innovative weight fusion strategy for feature enhancement to prevent data loss during feature fusion and preserve the relative size relationship between weights.
3. The proposed method AETF-Net has been validated on three public datasets (i.e., IN, UP, and KSC) and has shown significantly better classification results. Particularly, at small sample rates, our method outperforms both traditional and advanced methods. The effectiveness of the method is verified.

The rest of this paper is arranged as follows. Section 2 elaborates on the proposed AETF-Net. Section 3 describes the detailed datasets and analyzes the experimental results. Section 4 provides a comprehensive discussion of the differences between the proposed method and the comparative algorithms. Section 5 summarizes the core of the whole paper and provides some suggestions for further research.

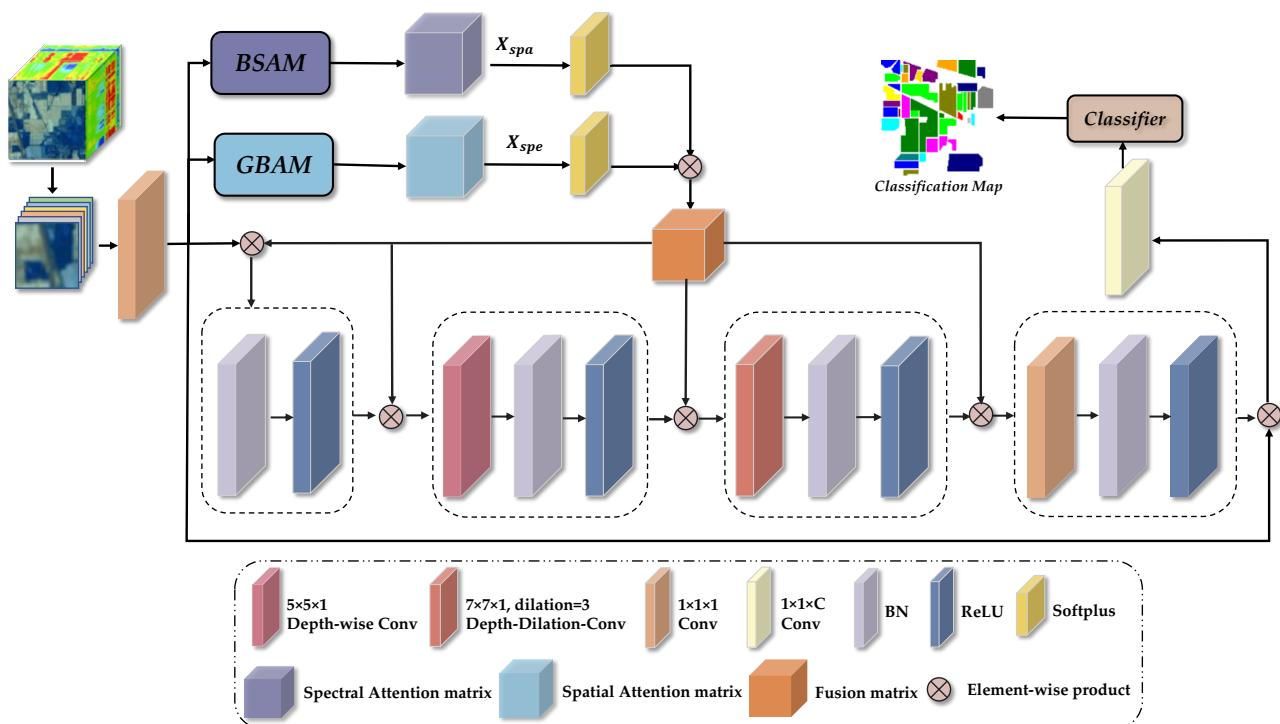


Figure 1. The overall architecture of the proposed AETF-Net model.

2. Materials and Methods

As is shown in Figure 1, the proposed AETF-Net framework is composed of three primary submodules: (1) spectral attention module, using 1D-CNN to extract attention features and eliminate global band redundancy; (2) spatial attention module, using 2D-CNN to extract attention features from both spatial horizontal and vertical directions to capture more discriminative and detailed edge features; (3) spectral-spatial fusion module, aiming at fusing joint spatial-spectral features by spectral and spatial attentional weights to improve 3D convolution feature extraction efficiency.

2.1. Spectral Attention Module

HSI typically has a large number of spectral bands, while not all of them are useful for classification. Thus, significant spectral bands should be highlighted for feature extraction. Inspired by the channel attention mechanism [48], we regard spectral bands as the channels and develop a new band attention module (GBAM) for spectral feature learning. The structure of the proposed GBAM is shown in Figure 2.

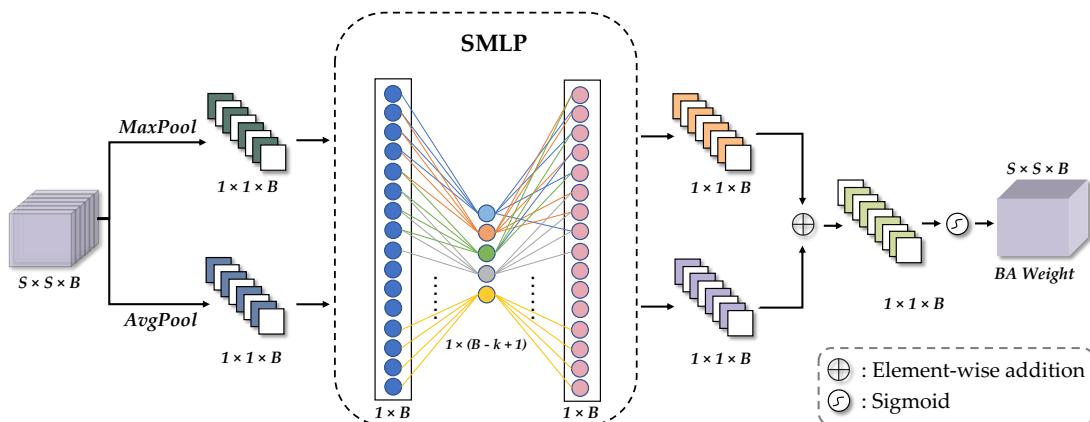


Figure 2. The proposed GBAM structure.

Specifically, the original hyperspectral data cube is first dealt with using a 3D convolution operation to extract low-order spectral features from the HSI, and the output feature map is defined as:

$$\mathbf{X}' = \mathbf{X} * \mathbf{w} + \mathbf{b} \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{S \times S \times B}$ denotes the 3D input HSI patch, S denotes the size of input patch and B denotes the number of bands, \mathbf{w} and \mathbf{b} denote the weights and biases of the network, and $*$ denotes the 3D convolution operation. The output feature map \mathbf{X}' is then squeezed in spatial dimension with maximum pooling and average pooling:

$$\begin{cases} \mathbf{X}_{avg} = \frac{1}{S \times S} \sum_{i=1}^S \sum_{j=1}^S \mathbf{X}'_{i,j} \\ \mathbf{X}_{max} = \max_{i \leq S, j \leq S} \mathbf{X}'_{i,j} \end{cases} \quad (2)$$

where $\mathbf{X}'_{i,j} \in \mathbb{R}^{S \times S \times B}$ is the element in the input patch at pixel (i, j) , $\mathbf{X}_{avg} \in \mathbb{R}^{1 \times 1 \times B}$ represents the output of the average pooling operation, and $\mathbf{X}_{max} \in \mathbb{R}^{1 \times 1 \times B}$ represents the output of the max pooling operation.

They are subsequently delivered into a new shared selective multilayer perceptron (SMLP). The typical MLP consists of an input layer, a simple hidden layer, and an output layer. The hidden layer is commonly designed to reduce the parameters by a squeezing operation, which can lead to the loss of band information in our spectral band. Thus, we propose a new SMLP in which the hidden layer is refined to model the long-range dependencies of all bands by considering k local neighborhoods. Based on the best experimental results, we set the value of k to 9. The SMLP output vector $\mathbf{L} \in \mathbb{R}^{B-k+1}$ is composed of \mathbf{L}_i , $i = (\lceil \frac{k}{2} \rceil, \dots, B - \lceil \frac{k}{2} \rceil)$, which is:

$$\mathbf{L}_i = \sum_{j=1}^k \mathbf{y}_i^j \mathbf{w}_i, \quad \mathbf{y}_i^j \in \Omega_i^k \quad (3)$$

where $\lceil \cdot \rceil$ denotes the ceiling function, which rounds a given number up to the nearest integer, Ω_i^k denotes the set of k spectral bands adjacent to the i th element of the average pooling vector or max pooling vector, and \mathbf{w}_i is the shared parameters of each \mathbf{y}_i^j . Next, the deconvolution operation is applied to the feature vector \mathbf{S} to generate a vector of the same size as the input, facilitating subsequent processing. To enhance the robustness and generalization ability of the deconvolution operation, the activate function ReLU and batch normalization are introduced.

After \mathbf{X}_{max} and \mathbf{X}_{avg} pass through the SMLP module, the element-wise addition operation and the sigmoid nonlinear activation function yield the band attention weight matrix $f(\mathbf{X}) \in \mathbb{R}^{S \times S \times B}$. The band attention module can be expressed as:

$$f(\mathbf{X}) = sigmoid(\mathbf{L}(\mathbf{X}_{max}) + \mathbf{L}(\mathbf{X}_{avg})) \quad (4)$$

where *sigmoid* is the nonlinear activation function.

2.2. Bi-Directional Spatial Attention Module

As far as we know, spatial information is helpful for HSI classification because the neighboring pixels are likely to belong to the same class. Furthermore, the spatial feature from multiple neighboring pixels can suppress noise interference and redundant information. In this paper, we develop a bi-directional spatial attention module (BSAM) to obtain the abstract spatial representation for HSI classification. Instead of using 2D pooling operation in spatial feature extraction, which may lead to a loss of location information [49], BSAM separates spatial attention into two parallel 1D feature encoding processes. Two separate attention feature maps are independently embedded with orientation-specific

information. Each of them captures the long-range dependencies of the input feature map along one of the spatial directions, while preserving the location information in the other direction in the generated attention map. The structure of improved BSAM is shown in Figure 3.

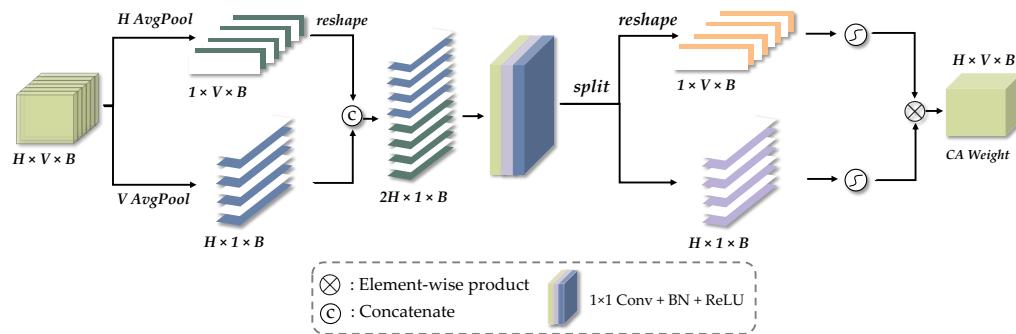


Figure 3. The proposed BSAM structure.

BSAM first performs two independent global average pooling operations with kernels $(H, 1)$ and $(1, V)$ in two spatial dimensions (the horizontal and vertical axes) on each channel to encode attention maps. H denotes the size of the pooling kernel in the horizontal direction, and V denotes the size in the vertical direction. It is noteworthy that the values of H and V are equivalent to the size of the original image patch $S \times S$, while in this section, different symbols are expressed for the purpose of direction distinction. The global average poolings are calculated by:

$$\begin{cases} Z_H^h = \frac{1}{V} \sum_{0 \leq i < V} x(h, i), h = [1, \dots, H] \\ Z_V^v = \frac{1}{H} \sum_{0 \leq j < H} x(j, v), v = [1, \dots, V] \end{cases} \quad (5)$$

where $x(h, i)$ is the input feature pixel at height h , $x(j, v)$ is the input feature pixel at width v , and $Z_H \in \mathbb{R}^{h \times 1 \times B}$ and $Z_V \in \mathbb{R}^{1 \times vs. \times B}$ are the h th horizontal average pooling result and the v th vertical average pooling result, respectively.

The global pooling operation in both directions generates paired direction-aware feature maps. Those feature maps not only capture directional information over a wide spatial range but also preserve location information. They can help the deep neural network locate target locations of interest. The obtained direction-aware feature maps Z_H and reshaped feature map $\tilde{Z}_V \in \mathbb{R}^{v \times 1 \times B}$ are then applied to feature fusion, along with channel compression and feature nonlinear restructuring, to yield the feature map $U \in \mathbb{R}^{(h+v) \times 1 \times B}$:

$$U = \text{sigmoid}(\text{Conv}(\text{Cat}[Z_H, \tilde{Z}_V])) \quad (6)$$

where Conv denotes a 1×1 convolution layer and Cat denotes the concatenate operation. Then, the feature map is separated again into two tensor matrices, U_H and U_V , along the spatial horizontal and vertical directions, where U_V needs to be reshaped back to its original shape. Then, $U_H \in \mathbb{R}^{h \times 1 \times B}$ and $U_V \in \mathbb{R}^{1 \times vs. \times B}$ are delivered into the convolution layer with kernel 1×1 and the sigmoid nonlinear activation function, respectively, making the shape of the output tensor matrix the same as the input data patch X .

Finally, feature fusion is performed by element-wise multiplication to acquire the spatial attention weight matrix $g(X) \in \mathbb{R}^{S \times S \times B}$:

$$g(X) = \text{sigmoid}(\text{Conv}(U_H)) \otimes \text{sigmoid}(\text{Conv}(U_V)) \quad (7)$$

2.3. Multi-Attention Spectral-Spatial Feature Fusion Branch

The multi-attention fusion branch (MAFB) is designed mainly following the structure of 3D-CNN. The attention weight matrices $f(\mathbf{X})$ and $g(\mathbf{X})$ learned from 1D-GBAM and 2D-BSAM branches are fused in the MAFB's convolution operations. In comparison, there are some different points from typical 3D-CNN.

MAFB is developed based on a lightweight CNN (LCNN) [50], which consists of a combination of deep convolution DW-Conv, deep dilated convolution DW-D-Conv, and 1×1 convolution procedures with small kernels. In this structure, the improved MAFB not only captures relative long-distance features for the large-scale visual field, but also relieves the requirements of many training samples and massive computational resources due to the large kernel convolutional operations.

MAFB designs a new multi-attention fusion strategy. The attention weight matrices $f(\mathbf{X})$ and $g(\mathbf{X})$ learned from 1D-GBAM and 2D-BSAM branches help the network be more attentive to the channels and locations contributing to the feature classification task. However, there are some issues with fusing two attention and convolutional operations by the simple multiplication strategy. As is shown in Figure 4a, the weight values of $f(\mathbf{X})$ and $g(\mathbf{X})$ are in $[0, 1]$, which leads to a smaller value after they multiply. This may exacerbate the feature intensity to decay and lose the most critical information when using the simple multiplication strategy. Thus, we introduce a softplus-post multiplication in MAFB, as is shown in Figure 4b. Before multiplying the two weight matrices by the input map \mathbf{X} , the softplus activation function performs feature enhancement and linear activation on the weight matrices, respectively. By doing so, the weights between different features can be scaled up to avoid feature dissipation while the relative sizes are guaranteed to be constant.

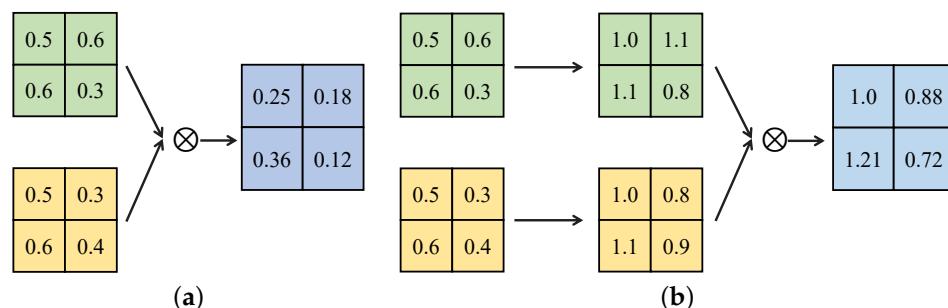


Figure 4. (a) Comparison of direct multiplication and (b) softplus-post multiplication.

Compared with the traditional ReLU activation function, the softplus activation function is closer to the activation model of brain neurons and solves the Dead ReLU problem. The equation of the softplus activation function is shown below:

$$G_{ij} = \log(1 + \exp(m_{ij})) \quad i, j = [1, \dots, S] \quad (8)$$

where m_{ij} denotes the element in the i th row and j th column of the $f(\mathbf{X})$ or $g(\mathbf{X})$ weight matrix and G_{ij} is the output of the activate operation softplus. The $G_M \in \mathbb{R}^{S \times S \times B}$ is acquired by element-wise multiplication of the two weight matrices:

$$G_M = \text{softplus}(f(\mathbf{X})) \otimes \text{softplus}(g(\mathbf{X})) \quad (9)$$

The fused feature weights matrix G_M is multiplied by the original input feature map \mathbf{X} as the input of the depth convolution DW-Conv. Then, the output feature map of DW-Conv multiplied by the attention weight matrix G_M is to be used as the input of the next layer of the depth-void convolution DW-D-Conv. Similarly, the output feature map of DW-D-Conv multiplied by the attention weight matrix G_M is to be used as the input of the 1×1 convolution. Finally, the output of the convolution operation is multiplied by the original input feature map \mathbf{X} to obtain the final attention map of the joint spatial-spectral feature extraction module. The overall fusion equation can be expressed as:

$$F = \text{Conv}((C_{DW-D}((C_{DW}(G_M \otimes X)) \otimes G_M)) \otimes G_M) \quad (10)$$

where X denotes the input feature map, C_{DW} denotes deep convolution operation, C_{DW-D} denotes deep dilated convolution operation, and $F \in \mathbb{R}^{S \times S \times B}$ denotes the feature map, which finally feeds into the classifier.

3. Results

3.1. Dataset Description

The datasets used in this paper are Indian pines (IP), University of Pavia (UP), and Kennedy Space Center (KSC). The sample numbers and corresponding colors of the three datasets are in Tables 1–3.

Table 1. The training and testing sample numbers and colors of the IP dataset.

No.	Class Name	Train/Validate	Test	Total	Color
1	Alfalfa	2	42	46	
2	Corn-notill	14	1400	1428	
3	Corn-mintill	8	814	830	
4	Corn	2	233	237	
5	Grass-pasture	4	475	483	
6	Grass-trees	7	716	730	
7	Grass-pasture-mowed	2	24	28	
8	Hay-windrowed	4	470	478	
9	Oats	2	16	20	
10	Soybean-notill	9	954	972	
11	Soybean-mintill	24	2407	2455	
12	Soybean-clean	5	583	593	
13	Wheat	2	201	205	
14	Woods	12	1241	1265	
15	Buildings-Grass-Trees-Drives	3	380	386	
16	Stone-Steel-Towers	2	89	93	
Total		102	10,045	10,249	

Table 2. The training and testing sample numbers and colors of the UP dataset.

No.	Class Name	Train/Validate	Test	Total	Color
1	Asphalt	66	6499	6631	
2	Meadows	186	18,277	18,649	
3	Gravel	20	2059	2099	
4	Trees	30	3004	3064	
5	Painted metal sheets	13	1319	1345	
6	Bare Soil	50	4929	5029	
7	Bitumen	13	1304	1330	
8	Self-Blocking Bricks	36	3610	3682	
9	Shadows	9	929	947	
Total		423	41,930	42,776	

Table 3. The training and testing sample numbers and colors of the KSC dataset.

No.	Class Name	Train/Validate	Test	Total	Color
1	Scrub	7	747	761	
2	Willow swamp	2	239	243	
3	Cabbage palm hammock	2	252	256	
4	Cabbage palm/oak hammock	2	248	252	
5	Slash pine	2	157	161	
6	Oak/broadleaf hammock	2	225	229	
7	Hardwood swamp	2	101	105	
8	Graminoid marsh	4	423	431	
9	Spartina marsh	5	510	520	
10	Cattail marsh	4	396	404	
11	Salt marsh	4	411	419	
12	Mudd flats	5	493	503	
13	Water	9	909	927	
Total		50	5111	5211	

The IP dataset is a widely used hyperspectral remote sensing image dataset, which contains a scene captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor at the Indian Pines test site in northwestern Indiana. The scene comprises two-thirds agricultural land and one-third forest or other natural perennial plants. Its data size is 145×145 , with a spatial resolution of 20 meter/pixel (m/p) and a wavelength range from 0.4 to $2.5 \mu\text{m}$ containing 224 bands, with 200 remaining after removing the overlying absorption region, and 16 species.

The ROSIS sensor, an optical reflection system imaging spectrometer for urban areas, captured the UP dataset in 2003 at the University of Pavia, Northern Italy. It possesses a spatial resolution of 1.3 m/p, an image size of 610×340 , 103 bands within the wavelength range from 0.43 to $0.86 \mu\text{m}$, and 9 classes. Compared to the IP dataset, the UP dataset has fewer bands while still having a high dimensionality and complex classification task.

The KSC dataset is a hyperspectral remote sensing image dataset collected and released by the National Aeronautics and Space Administration (NASA), which is collected at the Kennedy Space Center by an AVIRIS sensor in March 1996. It has a spatial resolution of 1.8 m/p, 512×614 pixels, 224 bands from 0.4 to $2.5 \mu\text{m}$, with 176 bands after removing absorbance and noise bands, and covers 13 different ground cover types. The KSC dataset has the same number of bands as the IP dataset while its spatial resolution is lower, thus requiring higher algorithmic requirements.

3.2. Experimental Setup

To demonstrate the efficiency of the proposed method, we conducted a series of classification experiments on three well-known hyperspectral datasets. These included CNN-based methods can be divided into two categories, traditional CNN-based methods (2D-CNN, 3D-CNN, Res-Net, and SSRN) and CNN-based methods with attention mechanism (DBDA, A2S2K, MAFN, and A-SPN). All comparison methods have the same parameter settings as in their corresponding references. The performance of classification will be evaluated using three metrics: overall accuracy (OA), average accuracy (AA), and the statistical kappa coefficient (Kappa) for the results. All methods were repeated ten times independently, after which the average value and standard deviation were taken to guarantee the generalizability of the experimental results.

In our experiments, three datasets were each divided into a 1% training set, a 1% validation set, and a 98% test set. During the training phase, we continuously adjusted certain hyperparameters of the model, such as the size of the convolution kernel, patch size, and learning rate, based on the training results obtained through experimentation. The model was trained using the Adam optimizer and cross-entropy loss function. In the validation phase, 1% of the samples were randomly selected as a validation set to select

the best model. Performance metrics were calculated on the validation set to select the best-performing model as the final model. During the final testing phase, the remaining 98% of the test set was used to test the best model and obtain the test results. The experiments were set up to take two samples for any class with a sample number less than 2 in the 1% training samples case.

By employing an early stopping strategy during the training phase, we found that our model converges in terms of loss and accuracy stabilizes around 200 epochs. Thus, we ultimately set 200 epochs to train the model. The batch size was set to 64 and the Adam optimizer was used in the proposed method. The learning rate was initialized at 0.01 and then adjusted using the cosine annealing algorithm. The k value in the SMLP structure of GBAM was set to 9 based on the optimal experimental results. All experiments were finished on the software environment PyTorch and a computer with a process of Inter(R) Xeon(R) Platinum 8124M CPU @ 3.00 GHz, 64G RAM, and an NVIDIA GeForce RTX 3090 graphics card.

3.2.1. The Effect of the Number of Training Samples

To further analyze the effect of the number of training samples on the proposed AETF-Net, we split the three datasets into a training set, a validation set, and a test set with varying proportions. The size of the validation set is always consistent with that of the train set, while the remaining portion constitutes the test set. The remaining hyperparameters were set to be consistent with the above. For the IP dataset, the number of training samples varies from 1%, 3%, 5%, 10%, and 20% of the dataset samples. For the UP and KSC, the number of training samples varies from 1%, 3%, 5%, 7%, and 10% of the dataset samples, respectively. The validation sets in the above three datasets are divided from the remaining data into data samples the same size as the divided training set, while the remaining part is employed as the test set.

Figure 5 shows the classification results of the proposed with the different numbers of training samples. The vertical axis represents OA, and the horizontal axis represents the training set ratio. For three datasets, the values of OA increase with the number of training samples increase until a stable case. For the IP dataset, the value of OA stabilizes when the training set size is between 3% and 5%. It improves dramatically after 5% and reaches stable when the ratio of training size is 10%. The data distribution of the UP and KSC is not as heterogeneous as that of the IP dataset. Therefore, after the training set size reaches 4%, OA becomes stable and can reach nearly 100% accuracy, especially after 1% for the UP dataset, which has a sufficient number of samples.

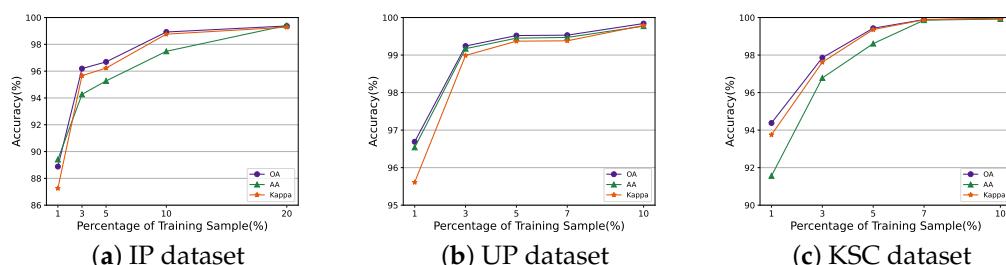


Figure 5. The effect of the number of training samples.

3.2.2. Effectiveness of the k Value in SMLP Structure

A series of experiments were conducted to verify the effectiveness of the improved SMLP structure in the GBAM module by setting various values of hyperparameter k . The remaining hyperparameter settings of the experiments were consistent with those described above. Firstly, we conducted experiments on the original MLP structure, followed by experiments on the improved SMLP structure with different values of hyperparameter k (3, 5, 7, 9, 11, 13, 15). To ensure fairness, all the experiments were conducted independently 10 times, and the final average results were compared. As shown in Table 4, when using the original MLP in the channel attention module, the classification accuracy OA was 2.29%

lower than that of the improved SMLP structure ($k = 9$), indicating that the improved SMLP structure could utilize the inter-band correlation information during the sliding window step of the convolutional kernel to extract more useful features than the original MLP structure.

Table 4. Performance of the SMLP structure with different k value on the IP dataset

IP (1%)	MLP	SMLP (k Value)						
		3	5	7	9	11	13	15
OA	0.8729	0.8556	0.8738	0.8811	0.8958	0.8920	0.8778	0.8781
AA	0.8379	0.8371	0.8739	0.8768	0.8791	0.8701	0.8461	0.8473
Kappa	0.8452	0.8341	0.8555	0.8641	0.8817	0.8767	0.8601	0.8604

However, the performance decreased significantly when the k value was set to 3 or 5, even lower than that of the original MLP structure, because a small k value cannot capture all the local features, leading to feature loss. With the increase of k value, the sliding window of convolution can capture more inter-band correlation information and local features. However, when the k value exceeds 11, the classification accuracy starts to decline due to the high overlap of the windows, which leads to overfitting, and the same local information is extracted multiple times. Therefore, based on the best experimental results, we set the k value to 9, increase the number of convolutional kernels to extract various features of the data, and set the stride small enough to retain more local features. Combined with deconvolution, we can reduce the dimensionality while retaining the important features of the input data, eliminate the influence of the one-dimensional convolution layer, and restore the data to its original dimensionality.

3.2.3. The Effect of Patch Size

The patch size of the training network has an essential effect on classification performance. Typically, the larger the patch, the more spatial information it contains, leading to a better classification performance of the classification. However, a larger patch causes massive parameters and exacerbates the limited sample learning issue.

In this section, we design several experiments based on the dataset partitioning method and parameter settings described above to analyze the effect of the patch size on the proposed method. Figure 6 shows the classification results of the proposed method with different patch sizes. The OA has achieved 80% when the patch size is 5×5 for the IP dataset. As the value of patch size increases, the OA gradually improves and plateaus until the OA tends to 90% when the sample block size exceeds 13×13 . For the KSC dataset, the OA decreases when the patch size increases to a certain sample block size. The reason is that the objects in KSC are small and in dispersed distribution, so the large patch contains multiple classes, which provide negative information for classifying the center pixel in the patch. Thus, considering the computational cost and HSI scene, we set the patch size to 11×11 in our experiments.

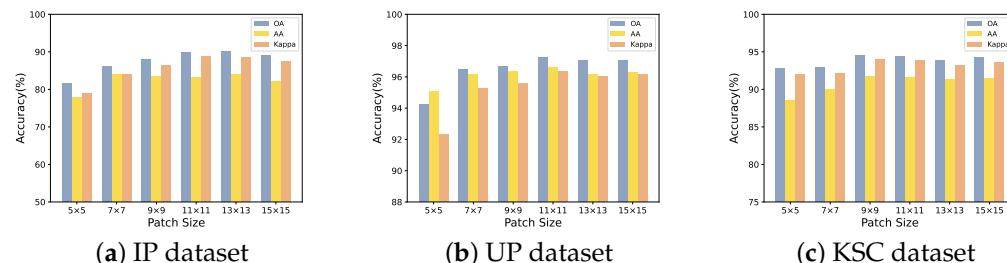


Figure 6. The classification results of the proposed AETF-Net with different patch sizes.

3.3. Result and Analysis

Experimental results on the IP dataset: As is shown in Table 5 and Figure 7, the proposed AETF-Net method obtains the highest accuracy among all the methods with 89.58%, 87.91%, and 88.17%, and achieves the most detailed and smooth classification maps. The 3D-CNN has better feature extraction capability than 2D-CNN because it can incorporate both spectral and spatial information. However, in the case of insufficient training samples, the overfitting caused by the conflict between the high dimension of processing data and the insufficient number of samples makes the accuracy lower than the 2D-CNN with 5.47% OA. Res-Net has the worst classification result with 36.06% OA because it has many layers of the network, which is very redundant, and the adequate depth is inadequate. A2S2K adds the attention mechanism to the residual structure to weigh the valuable features, which obtained a 0.27% improvement in OA compared to SSRN, which illustrates the effectiveness of the attention mechanism.

Table 5. The classification results (%) of all compared methods on the IP dataset.

Class	2D-CNN	3D-CNN	Res-Net	SSRN	A2S2K	MAFN	DBDA	A-SPN	AETF-Net
1	1.56 ± 1.02	18.44 ± 6.13		54.53 ± 50.19	55.93 ± 22.66	77.81 ± 25.78	79.60 ± 27.86	90.00 ± 15.18	54.52 ± 9.96
2	40.77 ± 2.93	45.79 ± 7.34		40.04 ± 16.46	74.40 ± 7.40	76.92 ± 6.63	74.17 ± 8.95	79.33 ± 7.79	62.86 ± 6.08
3	24.77 ± 2.97	25.49 ± 8.13		42.65 ± 23.13	77.60 ± 9.86	77.94 ± 6.60	55.09 ± 10.03	80.78 ± 11.38	49.05 ± 10.71
4	0.26 ± 0.04	9.83 ± 4.75		62.66 ± 40.38	74.03 ± 12.81	80.60 ± 8.92	48.62 ± 18.67	82.94 ± 8.13	24.43 ± 8.10
5	57.87 ± 3.62	46.03 ± 15.22		65.94 ± 31.15	87.63 ± 12.55	95.92 ± 5.81	87.47 ± 7.92	98.71 ± 1.95	78.08 ± 6.53
6	97.60 ± 0.42	89.44 ± 6.98		46.11 ± 28.57	92.34 ± 4.58	91.85 ± 05.66	87.51 ± 13.69	90.69 ± 6.38	78.08 ± 3.18
7	00.00 ± 0.00	00.00 ± 0.00		58.40 ± 30.86	59.04 ± 21.82	58.54 ± 19.17	52.58 ± 30.00	33.31 ± 16.24	100.00 ± 0.00
8	99.87 ± 0.22	93.33 ± 3.65		80.61 ± 20.02	96.90 ± 4.39	99.57 ± 0.41	92.33 ± 10.09	99.29 ± 1.54	81.88 ± 10.61
9	00.00 ± 0.00	00.00 ± 0.00		11.64 ± 23.12	31.75 ± 16.78	33.72 ± 08.84	48.42 ± 27.89	63.84 ± 18.03	85.00 ± 20.19
10	35.06 ± 1.99	44.59 ± 11.17		52.04 ± 31.21	73.35 ± 13.11	84.78 ± 05.44	72.77 ± 14.16	78.25 ± 11.82	55.07 ± 6.02
11	80.36 ± 2.73	60.67 ± 11.00		37.55 ± 5.85	78.34 ± 5.61	69.92 ± 05.61	83.71 ± 5.58	79.11 ± 8.29	92.80 ± 3.61
12	20.87 ± 4.07	20.81 ± 6.84		42.60 ± 21.51	76.38 ± 8.98	82.96 ± 14.12	59.53 ± 13.41	81.33 ± 19.32	39.30 ± 7.79
13	82.76 ± 5.03	64.78 ± 14.12		70.15 ± 30.02	90.85 ± 4.00	91.84 ± 04.22	78.31 ± 16.10	91.57 ± 7.15	98.72 ± 0.89
14	99.38 ± 0.17	90.39 ± 5.99		67.80 ± 18.49	92.24 ± 4.65	90.24 ± 4.65	96.89 ± 3.08	92.97 ± 7.18	99.67 ± 0.49
15	15.18 ± 2.97	13.39 ± 6.86		42.51 ± 35.54	76.51 ± 15.15	81.60 ± 10.79	66.10 ± 15.07	87.68 ± 12.96	36.74 ± 11.49
16	9.78 ± 4.24	10.55 ± 8.26		77.51 ± 38.93	78.14 ± 5.96	76.33 ± 12.19	87.55 ± 15.26	75.54 ± 11.83	98.57 ± 1.71
OA	60.40 ± 5.01	54.66 ± 2.18	36.06 ± 6.01	78.93 ± 3.25	79.20 ± 1.94	75.80 ± 2.67	82.19 ± 4.27	74.68 ± 1.31	89.58 ± 0.36
AA	41.63 ± 6.43	39.60 ± 2.24		53.30 ± 12.61	75.60 ± 4.03	78.04 ± 3.60	73.05 ± 5.59	80.93 ± 4.81	74.21 ± 1.60
Kappa	53.31 ± 5.66	47.93 ± 2.35	23.59 ± 6.49	75.89 ± 3.42	75.90 ± 2.30	72.44 ± 3.02	79.54 ± 4.99	70.30 ± 4.99	88.17 ± 0.49

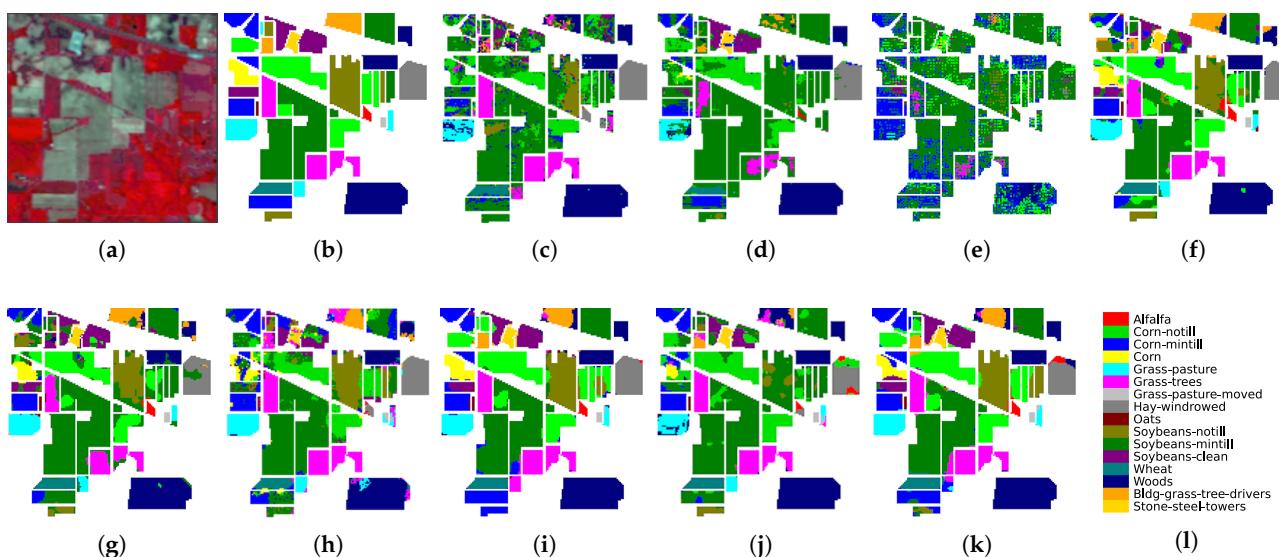


Figure 7. Classification maps of different methods on the IP dataset. **(a)** False-color; **(b)** Ground truth map; **(c)** 2D-CNN; **(d)** 3D-CNN; **(e)** Res-Net; **(f)** SSRN; **(g)** A2S2K; **(h)** MAFN; **(i)** DBDA; **(j)** A-SPN; **(k)** AETF-Net; **(l)** Color bar.

MAFN, DBDA, and A-SPN also introduce the attention mechanism and obtain higher accuracy than the method without introducing the attention mechanism. Among them, DBDA captures many spatial and spectral features using a two-branch and densely connected network and obtains the highest accuracy among the compared methods with 82.19% OA. However, DBDA has not used the attention mechanism to locate the region of interest at the very beginning, so the evaluation indices of our method are 7.39%, 6.46%, and 8.63% higher than those of DBDA. The unbalanced distribution of samples in the IP dataset results in very few training samples for some classes after dividing 1%. A-SPN performs better for small sample classes, where the accuracies for classes 7 and 9 (i.e., Grass-pasture-mowed and Oats) were higher than the proposed method. However, the performance on classes 3, 4, 12, and 15 (i.e., Corn-mintill, Grass-pasture, Soy-bean-clean, and Buildings-Grass-Trees-Drives) is significantly lower than that of the proposed method because these classes are at the edge of the image and have a large number of neighboring species, making it difficult to classify them with blurred boundaries correctly. As a result, the overall OA, AA, and Kappa are 14.90%, 13.18%, and 17.87% lower, respectively, than the proposed method. To further evaluate the classification performance from a visual perspective, the ground-truth map and the classification results of eight comparison methods are shown in Figure 7. 2D-CNN, 3D-CNN, and Res-Net obtained considerable noise within and at the class boundary. The noise point within the classification maps of SSRN, A2S2K, MAFN, and A-SPN are fewer, while the misclassification rates are higher than DBDA. By comparison, the classification map of our proposed methods has minor noise points and misclassified pixels on the boundary between classes and is closest to the ground-truth map.

Experimental results on the UP dataset: Table 6 and Figure 8 show the numerical results and visual results of UP dataset comparison experiments. It could be seen that the OA of the proposed AETF-Net method was improved compared with those attention-based methods A2S2K, MAFN, DBDA, and A-SPN for 1.76%, 0.82%, 0.86%, and 2.66%, respectively. Due to the relatively balanced distribution of each class in the dataset, 2D-CNN, 3D-CNN, and Res-Net obtained relatively higher classification accuracy. MAFN and DBDA all outperformed SSRN, A2S2K, and A-SPN. Compared with the similar multi-attention fusion method DBDA, our method has higher accuracy with 0.86% OA, 1.09% AA, and 0.45% Kappa. Because the MAFN lacks information interaction and feature transfer during the extraction of spectral and band attributes, resulting in one-sided extracted results, which demonstrates the effectiveness of our proposed feature fusion strategy. In addition, MAFN achieved the second-best classification results throughout a multi-scale

multi-attention feature extraction framework. Our method can reduce the network depth while extracting sufficient feature information, which avoids the overfitting problem caused by limited samples. Our method has absolute advantages. The classification map of our method performed better on the UP dataset. In 2D-CNN, 3D-CNN, Res-Net, SSRN, and A-SPN, class 2 and class 6 have considerable noise, while the noise points were significantly reduced in other methods because of the multi-attention structure used in MAFN, DABA, and the proposed method. This demonstrates the effectiveness of the multi-attention strategy. Overall, the produced classification map of our method has more precise edges of features and was closest to the ground-truth map.

Table 6. The classification results (%) of all compared methods on the UP dataset.

Class	2D-CNN	3D-CNN	Res-Net	SSRN	A2S2K	MAFN	DBDA	A-SPN	AETF-Net		
1	94.68 ± 1.14	88.15 ± 3.47		67.75 ± 10.04	89.58 ± 3.43	92.83 ± 3.33	96.99 ± 0.63	95.90 ± 6.15	97.15 ± 1.05	94.95 ± 2.11	
2	97.41 ± 0.33		87.63 ± 10.02	83.75 ± 7.10	97.49 ± 1.31	97.60 ± 1.51	99.46 ± 0.47	99.17 ± 1.08	99.36 ± 0.44	99.62 ± 0.12	
3	58.24 ± 2.61	76.79 ± 9.91		72.24 ± 12.16	80.21 ± 14.30	84.82 ± 6.10	95.72 ± 2.80	96.23 ± 3.07	74.21 ± 6.31	96.54 ± 1.46	
4	85.19 ± 1.01	88.80 ± 4.98	98.16 ± 1.82	98.57 ± 1.76		99.40 ± 0.49	98.19 ± 0.72	97.02 ± 1.05	92.89 ± 1.51	98.49 ± 0.10	
5		100.00 ± 0.00	98.05 ± 1.51	98.48 ± 2.17	99.25 ± 0.77	99.55 ± 0.64	99.33 ± 0.77	98.95 ± 1.81		100.00 ± 0.00	99.39 ± 0.38
6	70.90 ± 1.39		69.08 ± 16.21	93.17 ± 5.03	96.13 ± 2.96	98.43 ± 1.36	98.42 ± 0.32	98.74 ± 1.19	86.76 ± 5.43	98.78 ± 0.33	
7	48.72 ± 4.60		69.68 ± 16.77	76.18 ± 16.83	94.05 ± 8.84	97.15 ± 2.45	94.05 ± 6.46	97.82 ± 4.32	85.53 ± 7.90	98.96 ± 1.00	
8	74.00 ± 2.48		83.10 ± 17.54	74.55 ± 9.03	88.65 ± 3.79	86.84 ± 4.31	93.40 ± 3.72	90.12 ± 3.82	91.08 ± 5.33	86.94 ± 4.25	
9	93.79 ± 2.57	96.72 ± 2.02		89.03 ± 15.01	96.62 ± 3.74	98.31 ± 0.92	95.39 ± 1.91	98.42 ± 1.51	94.85 ± 2.65	97.09 ± 1.91	
OA	87.54 ± 3.58	84.66 ± 4.01	79.73 ± 4.19	94.12 ± 1.81		95.51 ± 1.05	96.45 ± 0.64	96.41 ± 1.86	94.61 ± 0.83	97.27 ± 0.49	
AA	80.33 ± 8.51	84.22 ± 3.03	83.70 ± 3.23	93.39 ± 2.63		94.99 ± 0.96	95.66 ± 0.98	96.15 ± 1.11	91.31 ± 1.43	96.75 ± 0.87	
Kappa	83.21 ± 4.83	79.96 ± 4.71	71.84 ± 6.24	92.16 ± 2.43		94.02 ± 1.41	95.94 ± 0.85	95.56 ± 2.50	92.80 ± 1.31	96.39 ± 0.65	

Experimental results on the KSC dataset: The KSC dataset has only 50 training samples at the 1% data division method, as shown in Table 7, and the proposed method still achieved the best classification accuracy with 96.48% OA, 95.00% AA, and 96.08% Kappa, and the clearest classification results were obtained for some hard distinguish categories like class 4, 6, 8, and 9. Regarding the classification accuracy of each of the thirteen classes of features, eight classes achieve the highest accuracy. Classes 10 and 13 achieved the best precision. Although the number of KSC dataset training samples is the smallest, it obtains better classification accuracy. Because the dataset is relatively balanced, the feature distribution is dispersed, and the inter-class differences are less influential. However, due to the limited samples, the classification accuracy of 2D-CNN, 3D-CNN, and Res-Net still needs to be improved. Although the MAFN method performed well on the IN and UP datasets, it needs to catch up on the KSC dataset due to the minimal and balanced number of samples in each class. It also indicates that the MAFN method is unsuitable for small sample classification. In addition, A2S2K has the best classification accuracy among all the compared methods due to the attention mechanism employed at the beginning of the framework to extract valuable characteristics. As shown in Figure 9, the proposed method had a smoother visual image compared with other methods and the classification map was closest to the ground-truth map.

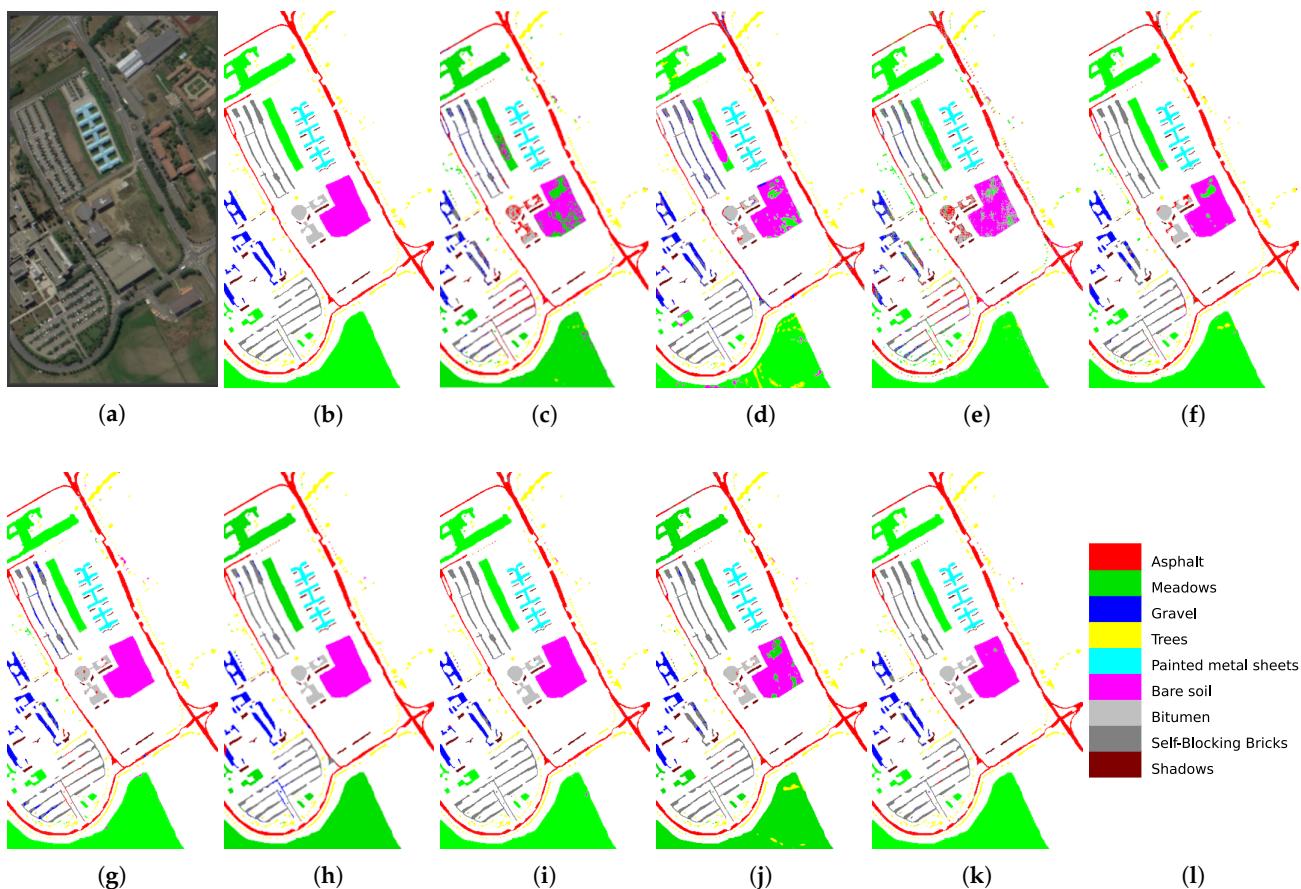


Figure 8. Classification maps of different methods on the UP dataset. (a) False-color; (b) Ground truth map; (c) 2D-CNN; (d) 3D-CNN; (e) Res-Net; (f) SSRN; (g) A2S2K; (h) MAFN; (i) DBDA; (j) A-SPN; (k) AETF-Net; (l) Color bar.

Furthermore, when viewed in the context of the proposed method, the standard deviation of the results of ten runs for almost every class and OA, AA, and Kappa is lower than that of the other methods. It can be demonstrated that the proposed method produces less variation and more stable results for small samples of different datasets, implying that the method is more robust and can be adapted for a broader range of hyperspectral datasets.

3.4. Ablation Study

To further validate the contribution of the GBAM, BSAM, and MAFB in the proposed framework to the final classification results, ablation experiments were conducted while maintaining the original experimental setup.

The effectiveness of the three branches is examined: (1) GBAM: only employ the GBAM to extract spectral feature extraction and the classifier; (2) BSAM: only employ the BSAM to extract spatial feature extraction and the classifier; (3) LCNN: LCNN network of MAFB without fusing the GBAM and BSAM; (4) LCNN + GBAM: LCNN network of MAFB with fusing the GBAM and without BSAM; (5) LCNN + BSAM: LCNN network of MAFB with fusing the BSAM and without GBAM.

From the results in Table 8, we can see that the performance of the GBAM and BSAM could be better than the other methods because the classification method based on single spectral or spatial feature extraction is significantly inferior to the methods based on spectral-spatial feature fusion. The LCNN overperforms GBAM and BSAM by about 3.86–10% on OA because it utilizes spectral-spatial feature combination by the 3D convolutional operation.

Table 7. The classification results (%) of all compared methods on the KSC dataset.

Class	2D-CNN	3D-CNN	Res-Net	SSRN	A2S2K	MAFN	DBDA	A-SPN	AETF-Net
1	87.58 ± 4.09	76.94 ± 18.98	74.17 ± 16.44	97.22 ± 4.19	95.69 ± 3.16	92.21 ± 5.79	98.11 ± 2.23	95.70 ± 3.91	99.94 ± 0.16
		34.02 ± 17.58	76.31 ± 21.21	88.00 ± 18.13	97.50 ± 4.20	78.03 ± 19.35	94.05 ± 8.15	70.62 ± 10.76	96.29 ± 4.91
2	2.49 ± 2.73	33.04 ± 15.09	50.57 ± 22.08	77.29 ± 12.88	78.61 ± 13.25	58.90 ± 10.74	75.81 ± 13.66	95.45 ± 5.49	77.33 ± 10.23
		25.24 ± 13.64	57.08 ± 26.11	83.62 ± 12.76	91.46 ± 7.17	57.87 ± 12.14	67.41 ± 25.84	45.28 ± 13.19	93.37 ± 9.02
3	36.48 ± 3.87	15.09	22.08	12.88	78.61 ± 13.25	58.90 ± 10.74	75.81 ± 13.66	95.45 ± 5.49	77.33 ± 10.23
		20.25 ± 5.65	15.67 ± 10.91	37.33 ± 27.51	78.59 ± 14.04	87.31 ± 11.89	84.16 ± 9.91	63.63 ± 25.86	89.24 ± 9.57
4	22.69 ± 1.82	13.64	26.11	12.76	91.46 ± 7.17	12.14	25.84	13.19	87.42 ± 15.21
		15.67 ± 10.91	37.33 ± 27.51	78.59 ± 14.04	87.31 ± 11.89	84.16 ± 9.91	63.63 ± 25.86	89.24 ± 9.57	87.42 ± 15.21
5	20.25 ± 5.65	10.91	27.51	14.04	88.46 ± 9.04	79.63 ± 17.63	81.84 ± 11.88	86.46 ± 9.52	98.24 ± 2.10
		20.66 ± 4.71	11.78 ± 8.08	60.67 ± 32.81	88.46 ± 9.04	79.63 ± 17.63	81.84 ± 11.88	86.46 ± 9.52	98.24 ± 2.10
6	37.30 ± 8.96	20.29 ± 13.40	82.37 ± 14.88	72.37 ± 18.47	74.63 ± 14.66	59.58 ± 18.63	56.97 ± 16.10	100.00 ± 0.00	88.81 ± 17.33
		63.40 ± 10.01	24.09 ± 9.68	51.56 ± 19.49	88.67 ± 8.23	95.33 ± 5.81	71.99 ± 12.56	73.75 ± 28.61	91.71 ± 10.54
7	73.98 ± 3.98	74.16 ± 14.03	58.88 ± 16.72	94.83 ± 9.10	99.04 ± 0.98	78.70 ± 10.00	79.48 ± 13.38	88.25 ± 6.73	99.63 ± 0.50
		15.20 ± 6.42	25.91 ± 9.71	97.13 ± 2.92	99.22 ± 1.58	99.19 ± 1.36	76.07 ± 8.07	92.13 ± 9.84	99.93 ± 0.11
8	93.86 ± 2.82	79.81 ± 6.88	96.26 ± 5.82	99.48 ± 0.53	99.69 ± 0.49	95.16 ± 6.55	96.14 ± 4.89	95.15 ± 2.71	99.69 ± 0.77
		36.43 ± 9.61	48.15 ± 17.14	78.64 ± 20.55	96.12 ± 2.83	95.63 ± 5.35	88.36 ± 6.93	85.71 ± 13.67	99.20 ± 0.89
9	81.79 ± 6.89	96.33 ± 4.19	67.87 ± 22.84	99.80 ± 0.32	96.60 ± 10.12	98.06 ± 4.91	99.94 ± 0.13	100.00 ± 0.00	100.00 ± 0.00
		44.70 ± 1.34	51.45 ± 6.75	56.84 ± 7.88	91.07 ± 2.70	93.02 ± 3.91	79.28 ± 0.04	84.84 ± 7.67	90.95 ± 1.76
OA	57.50 ± 1.39	56.69 ± 6.02	62.06 ± 6.73	91.98 ± 2.43	93.75 ± 3.50	81.43 ± 0.03	86.40 ± 6.88	91.86 ± 1.58	96.48 ± 1.10
AA	46.70 ± 1.99	43.49 ± 5.77	68.37 ± 5.00	89.24 ± 3.09	92.24 ± 2.62	78.36 ± 0.04	81.92 ± 7.76	89.00 ± 1.95	95.00 ± 1.61
Kappa	44.70 ± 1.34	51.45 ± 6.75	56.84 ± 7.88	91.07 ± 2.70	93.02 ± 3.91	79.28 ± 0.04	84.84 ± 7.67	90.95 ± 1.76	96.08 ± 1.23

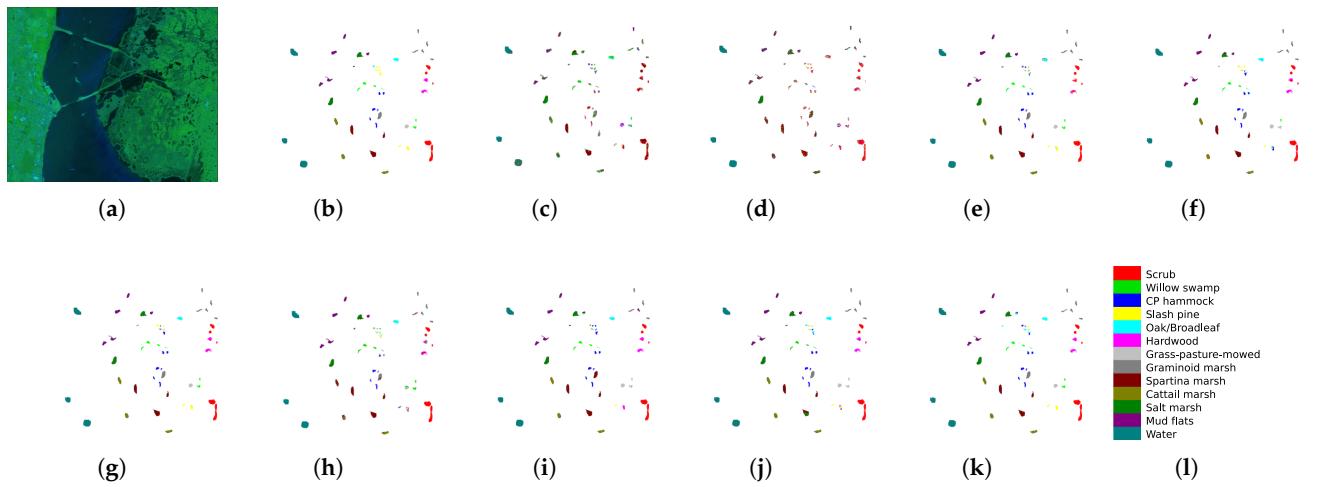
**Figure 9.** Classification maps of different methods on the KSC dataset. (a) False-color; (b) Ground truth map; (c) 2D-CNN; (d) 3D-CNN; (e) Res-Net; (f) SSRN; (g) A2S2K; (h) MAFN; (i) DBDA; (j) A-SPN; (k) AETF-Net; (l) Color bar.

Table 8. The best ablation study results on the IP dataset.

Methods	IP (1%)		
	OA	AA	Kappa
GBAM	0.7682	0.7173	0.7344
BSAM	0.8299	0.8096	0.8050
LCNN	0.8685	0.8161	0.8497
LCNN + GBAM	0.8723	0.8239	0.8534
LCNN + BSAM	0.8857	0.8571	0.8686
AETF-Net	0.8958	0.8791	0.8817

Additionally, the OA of the “LCNN + GBAM” and “LCNN + BSAM” increased by 0.38% to 1.72% compared with the OA of the “LCNN” method. It proved the effectiveness of the attention in GBAM and BSAM for classification. Especially the BSAM has obvious help in improving the AA by about 4.1%. It demonstrated that obtaining spatial features between feature mappings or long-range dependencies via the attention mechanism can significantly enhance the performance of the HSI classification model.

Lastly, the best classification results can be obtained when the spatial context information and the band dependencies are added concurrently to each stage of the MAFB for spatial-spectral joint attention feature extraction. It demonstrates the effectiveness of the proposed multiple attention fusion mechanism.

3.5. Analysis of the Multi-Attention Fusion Strategy

The fusion strategy is essential for multi-attention fusion, which significantly affects the classification method’s performance. In this section, we designed six multi-attention fusion strategies following the AETF-Net framework and did some experiments to analyze and discuss the effect on classification performance.

Six multi-attention fusion strategies are shown in Figure 10. They can mainly be split into two groups: attention weight fusion (Figure 10a) and attention feature map fusion (Figure 10b). The outputs of each attention module in attention weight fusion strategies are the combination of the weights, while the outputs of each attention module in attention feature maps fusion strategies are the combination of the weights and input maps. Especially the six multi-attention fusion strategies are designed as follows:

- (1) Figure 10a(1): the attention weight matrices produced by the GBA and BSA modules are element-wise multiplied and then multiplied with the original feature maps.
- (2) Figure 10a(2): the attention weight matrices produced by the GBA and BSA modules are element-wise added and then multiplied with the original feature maps.
- (3) Figure 10b(1): the feature maps produced by the GBAM and BSAM modules are element-wise added and then added with the original feature maps.
- (4) Figure 10b(2): the feature maps produced by the GBAM and BSAM modules are element-wise added and then multiplied with the original feature maps.
- (5) Figure 10b(3): the feature maps produced by the GBAM and BSAM modules are element-wise multiplied and then added to the original feature maps.
- (6) Figure 10b(4): the feature maps produced by the GBAM and BSAM modules are element-wise multiplied and then multiplied with the original feature maps.

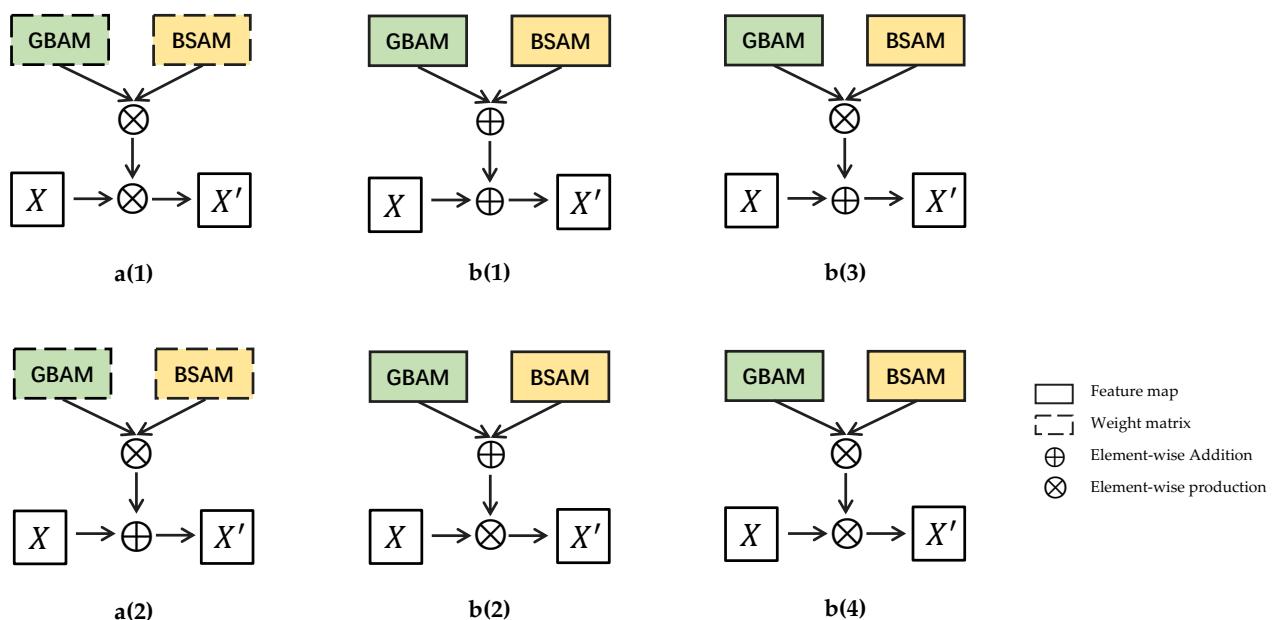


Figure 10. The illustration of multi-attention fusion strategy. (a) Attention weight fusion strategies, (b) attention feature maps fusion strategies.

Table 9 shows the classification results of the proposed AETF-Net with six different multi-attention fusion strategies. Compared with the two groups, the attention weight fusion strategy has a slight advantage over the attention feature map fusion strategy from the classification. The reasons are that both multi-attention fusion strategy groups have utilized effective characteristics of the attention mechanism for spectral-spatial feature learning, while the attention feature maps fusion strategies cost more computing resources to generate the feature map and lead to information redundancy.

In the attention weight fusion strategies, the multiplication strategy retains the relative size relationship between different feature mappings better than the addition strategy. Because it retains the variability between features and is superior to the addition strategy, the classification performance of a model can be improved by strengthening the compelling features after eliminating redundant ones. Thus, the proposed AETF-Net adopts the attention weight fusion strategy with multiplication (Figure 10a(1)).

Table 9. The effect of multi-attention fusion strategy on the IP dataset.

		IP (1%)		
		Methods	OA	AA
Weights	(GBA × BSA) × Input	0.8958	0.8791	0.8817
	(GBA + BSA) × Input	0.8971	0.8402	0.8782
Maps	(GBAM + BSAM) + Input	0.8927	0.8241	0.8779
	(GBAM + BSAM) × Input	0.8924	0.8219	0.8777
	(GBAM × BSA) + Input	0.8853	0.8204	0.8810
	(GBAM × BSAM) × Input	0.8812	0.7963	0.8647

3.6. Running Time Analysis

We computed the training and testing times of different methods using randomly selected samples. As shown in Figure 11, the proposed approach significantly improves training time compared to traditional DL methods. This is primarily attributed to the fact that traditional DL methods incorporate multiple convolutional and pooling layers in their network architecture to extract feature information, which leads to a large number of parameters when processing high-dimensional hyperspectral data. However, our proposed method did not show the least training and testing time compared to DL methods based on attention mechanisms, suggesting that further improvements are needed in our method. According to our model framework, it is possible that the increased computational cost of our method is due to the need for multiple information fusion processes in the backbone network. Nevertheless, our proposed method can fully extract and fuse the spatial and spectral features of HSI and the increase in computational cost is justifiable given the significant improvement in classification accuracy. The method A-SPN, which has the shortest processing time, may be attributed to its abandonment of the hierarchical structure composed of traditional convolution and pooling layers, resulting in a significant reduction in the computational cost of parameters. This is a direction worth exploring in our future work.

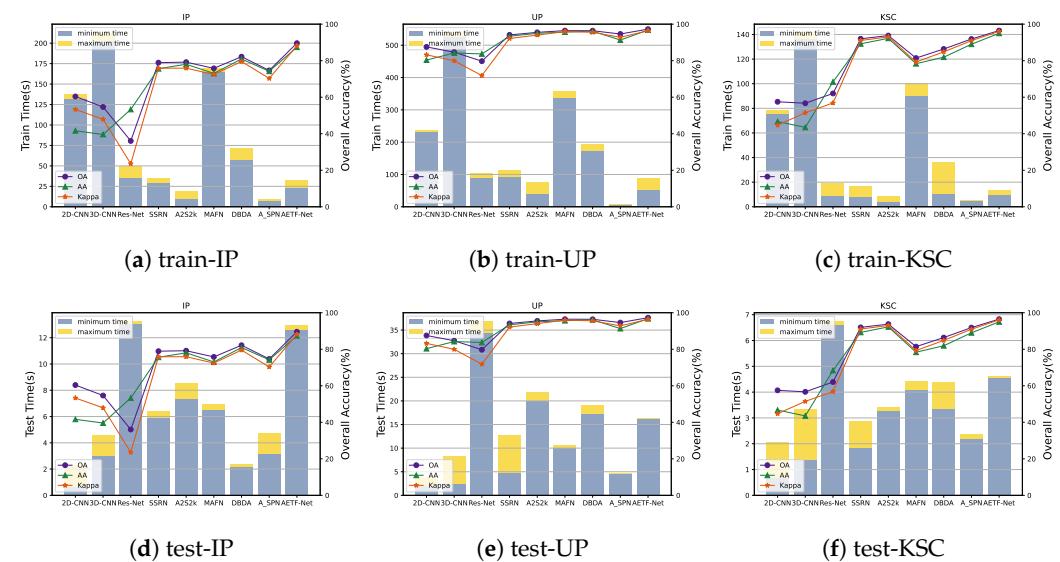


Figure 11. Comparison of computation time and overall accuracy of different methods.

4. Discussion

The proposed AETF-Net method has shown remarkable performance in terms of accuracy and classification map quality on three publicly available datasets, surpassing existing state-of-the-art methods.

Firstly, one of the key factors affecting the accuracy of deep learning-based image classification is the number of training samples. The 3D-CNN outperforms the 2D-CNN in feature extraction capabilities. However, overfitting can be a challenge when the number of training samples is insufficient. Additionally, Res-Net's redundant layers lead to worse classification results. Attention mechanisms, as seen in A2S2K, MAFN, DBDA, and A-SPN methods, have been demonstrated to improve accuracy, especially for small sample classes.

Additionally, despite having a limited number of training samples, the AETF-Net method achieved the best classification accuracy due to the dataset's balanced feature distribution and dispersed inter-class differences. The study emphasizes the limitations of existing methods for small sample classification and highlights the importance of attention mechanisms in achieving high accuracy. Furthermore, the results demonstrate the potential of AETF-Net to improve image classification tasks and its robustness for a broader range of datasets.

Furthermore, the study's findings also suggest that AETF-Net has the potential to overcome challenges associated with unbalanced sample distributions and misclassification at class boundaries by minimizing noise and improving classification accuracy. This has significant implications for the development of more reliable and accurate image classification in practical applications.

In conclusion, the results of this study have important implications for the development of deep learning-based image classification methods. The study emphasizes the importance of continued research in this area to improve accuracy and overcome the challenges associated with small sample classification, unbalanced sample distributions, and misclassification at class boundaries.

5. Conclusions

In this paper, we propose a novel HSI classification algorithm named AETF-Net to implement high-accuracy classification under a small sample rate. The model is divided into two sections, the spatial and spectral attention branch, and the spatial-spectral joint attention fusion branch. The first section of the spatial attention module models pixel-distant dependencies from two directions in space while preserving pixel position information, increasing the effectiveness and richness of spatial information. The band attention module establishes inter-band dependencies with adaptive convolution kernels to locate the band of interest. The second section of the spatial-spectral joint attention fusion branch extracts spatial-spectral joint features with three-stage 3D convolution. It embeds spatial and spectral attention features extracted in the first section before each convolution stage, and thus, enhancing the expressiveness and discriminative power of the spatial-spectral joint features extracted by 3D convolution. With a series of comparison and ablation experiments, the proposed AETF-Net achieved outstanding performance on limited training samples from three well-known HSI datasets.

The effectiveness of the multiple attention mechanism in dealing with small sample scales has been initially verified; however, the overfitting problem still exists at tiny sample rates. Further work will be to combine the attention mechanism and multi-scale to enhance the accuracy of HSI classification with tiny sample rates.

Author Contributions: E.Z.; Conceptualization, Writing—review and editing, Supervision. J.Z.; Methodology, Software, Formal analysis, Writing—original draft. J.B. (Jiaxin Bai); Conceptualization, Writing—review and editing. J.B. (Jiarong Bian); Investigation. S.F.; Resources. T.Z.; Supervision, Data curation. M.F.; Visualization, Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 62006188, 62103311), the Natural Science Basic Research Program of Shaanxi Province under Grant (2021JQ-195), the Qin Chuangyuan high-level innovation and entrepreneurship talent program of Shaanxi (2021QCYRC4-50), and the Chinese Universities Scientific Fund (No. 2452022341).

Data Availability Statement: Not applicable.

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

References

1. 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 Observ. Remote Sens.* **2022**, *15*, 968–999. [[CrossRef](#)]
2. Kang, K.K.K.; Hoekstra, M.; Foroutan, M.; Chegoonian, A.M.; Zolfaghari, K.; Duguay, C.R. Operating Procedures and Calibration of a Hyperspectral Sensor Onboard a Remotely Piloted Aircraft System For Water and Agriculture Monitoring. In Proceedings of the IGARSS, Yokohama, Japan, 28 July–2 August 2019; pp. 9200–9203. [[CrossRef](#)]
3. Lanthier, Y.; Bannari, A.; Haboudane, D.; Miller, J.R.; Tremblay, N. Hyperspectral Data Segmentation and Classification in Precision Agriculture: A Multi-Scale Analysis. In Proceedings of the IGARSS, Boston, MA, USA, 7–11 July 2008; Volume 2, pp. 585–588. [[CrossRef](#)]
4. Ang, K.L.M.; Seng, J.K.P. Big Data and Machine Learning With Hyperspectral Information in Agriculture. *IEEE Access* **2021**, *9*, 36699–36718. [[CrossRef](#)]

5. 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. [[CrossRef](#)] [[PubMed](#)]
6. Torrecilla, E.; Piera, J.; Aymerich, I.F.; Pons, S.; Ross, O.N.; Vilaseca, M. Hyperspectral Remote Sensing of Phytoplankton Assemblages in the Ocean: Effects of the Vertical Distribution. In Proceedings of the WHISPERS, Reykjavik, Iceland, 14–16 June 2010; pp. 1–4. [[CrossRef](#)]
7. Kruse, F.A.; Clasen, C.C.; Kim, A.M.; Carlisle, S.C. Effects of Spatial and Spectral Resolution on Remote Sensing for Disaster Response. In Proceedings of the IGARSS, Munich, Germany, 22–27 July 2012; pp. 7086–7089. [[CrossRef](#)]
8. Contreras, C.; Khodadadzadeh, M.; Tusa, L.; Loidolt, C.; Tolosana-Delgado, R.; Gloaguen, R. Geochemical and Hyperspectral Data Fusion for Drill-Core Mineral Mapping. In Proceedings of the WHISPERS, Amsterdam, The Netherlands, 24–26 September 2019; pp. 1–4. [[CrossRef](#)]
9. Murphy, R.J.; Schneider, S.; Monteiro, S.T. Consistency of Measurements of Wavelength Position From Hyperspectral Imagery: Use of the Ferric Iron Crystal Field Absorption at ~900 nm as an Indicator of Mineralogy. *IEEE Geosci. Remote Sens.* **2014**, *52*, 2843–2857. [[CrossRef](#)]
10. Ghandehari, M.; Aghamohamadnia, M.; Dobler, G.; Karpf, A.; Cavalcante, C.; Buckland, K.; Qian, J.; Koonin, S. Ground based Hyperspectral Imaging of Urban Emissions. In Proceedings of the WHISPERS, Los Angeles, CA, USA, 21–24 August 2016; pp. 1–3. [[CrossRef](#)]
11. Hsieh, T.H.; Kiang, J.F. Comparison of CNN Algorithms on Hyperspectral Image Classification in Agricultural Lands. *Sensors* **2020**, *20*, 1734. [[CrossRef](#)]
12. Ghamisi, P.; Dalla, M.M.; Benediktsson, J.A. A Survey on Spectral–Spatial Classification Techniques based on Attribute Profiles. *IEEE Geosci. Remote Sens.* **2015**, *53*, 2335–2353. [[CrossRef](#)]
13. Rashmi, S.; Swapna, A.; Venkat, S. Spectral Angle Mapper Algorithm for Remote Sensing Image Classification. *IJSET* **2014**, *1*. [[CrossRef](#)]
14. Tarabalka, Y.; Fauvel, M.; Chanussot, J.; Benediktsson, J.A. SVM- and MRF-Based Method for Accurate Classification of Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 736–740. [[CrossRef](#)]
15. Zhang, E.; Zhang, X.; Liu, H.; Jiao, L. Fast Multifeature Joint Sparse Representation for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1397–1401. [[CrossRef](#)]
16. Huang, H.; Chen, M.L.; Duan, Y.L.; Shi, G.Y. Hyperspectral Image Classification using Spatial-Spectral Manifold Reconstruction. *Opt. Precis. Eng.* **2018**, *26*, 1827–1836. [[CrossRef](#)]
17. Ghamisi, P.; Benediktsson, J.A.; Ulfarsson, M.O. The Spectral-Spatial Classification of Hyperspectral Images based on Hidden Markov Random Field and its Expectation-Maximization. In Proceedings of the IGARSS, Melbourne, VIC, Australia, 21–26 July 2013; pp. 1107–1110. [[CrossRef](#)]
18. Kumar, B.; Dikshit, O. Hyperspectral Image Classification based on Morphological Profiles and Decision Fusion. *Int. J. Remote Sens.* **2017**, *38*, 5830–5854. [[CrossRef](#)]
19. Ham, J.; Chen, Y.; Crawford, M.; Ghosh, J. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. *IEEE Geosci. Remote Sens.* **2005**, *43*, 492–501. [[CrossRef](#)]
20. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
21. Alipourfard, T.; Arefi, H.; Mahmoudi, S. A Novel Deep Learning Framework by Combination of Subspace-Based Feature Extraction and Convolutional Neural Networks for Hyperspectral Images Classification. In Proceedings of the IGARSS, Valencia, Spain, 22–27 July 2018; pp. 4780–4783. [[CrossRef](#)]
22. Rissati, J.V.; Molina, P.C.; Anjos, C.S. Hyperspectral Image Classification Using Random Forest and Deep Learning Algorithms. In Proceedings of the IEEE LAGIRS, Santiago, Chile, 22–26 March 2020; p. 132. [[CrossRef](#)]
23. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep Learning-Based Classification of Hyperspectral Data. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2014**, *7*, 2094–2107. [[CrossRef](#)]
24. Chen, Y.; Zhao, X.; Jia, X. Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2015**, *8*, 2381–2392. [[CrossRef](#)]
25. Li, B.; Wang, Q.W.; Liang, J.H.; Zhu, E.Z.; Zhou, R.Q. SquonNet: Deep Sequencer Convolutional Network for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 983. [[CrossRef](#)]
26. Zhang, H.; Li, Y.; Zhang, Y.; Shen, Q. Spectral-Spatial Classification of Hyperspectral Imagery using a Dual-channel Convolutional Neural Network. *Remote Sens. Lett.* **2017**, *8*, 438–447. [[CrossRef](#)]
27. Wei, H.; Yangyu, H.; Li, W.; Fan, Z.; Hengchao, L. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [[CrossRef](#)]
28. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep Supervised Learning for Hyperspectral Data Classification through Convolutional Neural Networks. In Proceedings of the IGARSS, Milan, Italy, 26–31 July 2015; pp. 4959–4962. [[CrossRef](#)]
29. Zhang, Y.; Huynh, C.P.; Ngan, K.N. Feature Fusion with Predictive Weighting for Spectral Image Classification and Segmentation. *IEEE Geosci. Remote Sens.* **2019**, *57*, 6792–6807. [[CrossRef](#)]
30. Li, Y.; Zhang, H.; Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67. [[CrossRef](#)]

31. Ge, H.; Wang, L.; Liu, M.; Zhu, Y.; Zhao, X.; Pan, H.; Liu, Y. Two-Branch Convolutional Neural Network with Polarized Full Attention for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 848. [[CrossRef](#)]
32. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Geosci. Remote Sens.* **2018**, *56*, 847–858. [[CrossRef](#)]
33. Li, R.; Zheng, S.; Duan, C.; Yang, Y.; Wang, X. Classification of Hyperspectral Image Based on Double-Branch Dual-Attention Mechanism Network. *Remote Sens.* **2020**, *12*, 582. [[CrossRef](#)]
34. Yan, H.; Zhang, E.; Wang, J.; Leng, C.; Peng, J. MTFFN: Multimodal Transfer Feature Fusion Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
35. Yue, J.; Fang, L.; Rahmani, H.; Ghamisi, P. Self-Supervised Learning with Adaptive Distillation for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens.* **2022**, *60*, 1–13. [[CrossRef](#)]
36. Hughes, G. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Trans. Inf. Theory* **1968**, *14*, 55–63. [[CrossRef](#)]
37. Zhang, H.; Li, Y.; Jiang, Y.; Wang, P.; Shen, Q.; Shen, C. Hyperspectral Classification Based on Lightweight 3-D-CNN with Transfer Learning. *IEEE Geosci. Remote Sens.* **2019**, *57*, 5813–5828. [[CrossRef](#)]
38. Sellami, A.; Farah, M.; Riadh Farah, I.; Solaiman, B. Hyperspectral Imagery Classification based on Semi-Supervised 3-D Deep Neural Network and Adaptive Band Selection. *Expert Syst. Appl.* **2019**, *129*, 246–259. [[CrossRef](#)]
39. Li, T.; Zhang, X.; Zhang, S.; Wang, L. Self-Supervised Learning with a Dual-Branch ResNet for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
40. Yang, K.; Sun, H.; Zou, C.; Lu, X. Cross-Attention Spectral–Spatial Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens.* **2022**, *60*, 1–14. [[CrossRef](#)]
41. Xiang, J.; Wei, C.; Wang, M.; Teng, L. End-to-End Multilevel Hybrid Attention Framework for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
42. Huang, H.; Luo, L.; Pu, C. Self-Supervised Convolutional Neural Network via Spectral Attention Module for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
43. Tu, B.; He, W.; He, W.; Ou, X.; Plaza, A. Hyperspectral Classification via Global-Local Hierarchical Weighting Fusion Network. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 184–200. [[CrossRef](#)]
44. Yu, C.; Han, R.; Song, M.; Liu, C.; Chang, C.I. Feedback Attention-Based Dense CNN for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens.* **2022**, *60*, 1–16. [[CrossRef](#)]
45. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [[CrossRef](#)]
46. Li, Z.; Zhao, X.; Xu, Y.; Li, W.; Zhai, L.; Fang, Z.; Shi, X. Hyperspectral Image Classification with Multiattention Fusion Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
47. Xue, Z.; Zhang, M.; Liu, Y.; Du, P. Attention-Based Second-Order Pooling Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens.* **2021**, *59*, 9600–9615. [[CrossRef](#)]
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 3–19. [[CrossRef](#)]
49. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the CVPR, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 13713–13722. [[CrossRef](#)]
50. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual Attention Network. *arXiv* **2022**, arXiv:2202.09741.

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.