remote sensing

MDPI

Article

Subpixel Multilevel Scale Feature Learning and Adaptive Attention Constraint Fusion for Hyperspectral Image Classification

Zixian Ge [1], Guo Cao [1,*], Youqiang Zhang [2,3], Hao Shi [1], Yanbo Liu [1], Ayesha Shafique [1] and Peng Fu [1]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; gezixian727@njust.edu.cn (Z.G.); hao1227@njust.edu.cn (H.S.); liuyanbo@njust.edu.cn (Y.L.); ayeshashafique@njust.edu.cn (A.S.); fupeng@njust.edu.cn (P.F.)
[2] Jiangsu Key Laboratory of Broadband Wireless Communication and Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; zhangyq@njupt.edu.cn
[3] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
* Correspondence: caoguo@njust.edu.cn

Abstract: Convolutional neural networks (CNNs) play an important role in hyperspectral image (HSI) classification due to their powerful feature extraction ability. Multiscale information is an important means of enhancing the feature representation ability. However, current HSI classification models based on deep learning only use fixed patches as the network input, which may not well reflect the complexity and richness of HSIs. While the existing methods achieve good classification performance for large-scale scenes, the classification of boundary locations and small-scale scenes is still challenging. In addition, dimensional dislocation often exists in the feature fusion process, and the up/downsampling operation for feature alignment may introduce extra noise or result in feature loss. Aiming at the above issues, this paper deeply explores multiscale features, proposes an adaptive attention constraint fusion module for different scale features, and designs a semantic feature enhancement module for high-dimensional features. First, HSI data of two different spatial scales are fed into the model. For the two inputs, we upsample them using bilinear interpolation to obtain their subpixel data. The proposed multiscale feature extraction module is intended to extract the features of the above four parts of the data. For the extracted features, the multiscale attention fusion module is used for feature fusion, and then, the fused features are fed into the high-level feature semantic enhancement module. Finally, based on the fully connected layer and softmax layer, the prediction results of the proposed model are obtained. Experimental results on four public HSI databases verify that the proposed method outperforms several state-of-the-art methods.

Keywords: HSI classification; convolutional neural network (CNN); multiscale features; subpixel; adaptive attention fusion; feature enhancement

Citation: Ge, Z.; Cao, G.; Zhang, Y.; Shi, H.; Liu, Y.; Shafique, A.; Fu, P. Subpixel Multilevel Scale Feature Learning and Adaptive Attention Constraint Fusion for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 3670. https://doi.org/10.3390/rs14153670

Academic Editor: Pietro Tizzani

Received: 10 June 2022
Accepted: 28 July 2022
Published: 31 July 2022

1. Introduction

Hyperspectral images (HSIs) obtained by an imaging spectrometer provide detailed spectral information for each pixel. Plentiful spectral signatures and spatial information make it possible for HSIs to detect objects that cannot be detected in ordinary images [1]. In recent years, the study of HSIs has become an important research direction in remote sensing due to their unique properties and massive information. Among them, HSI classification is a basic task that plays an important role in geological exploration [2,3], crop detection [4,5], national defense [6–8], and military fields [9] and is worthy of further study. The HSI classification task aims to use the prior information within the original data, such as partially labeled samples, to determine the object categories of other pixels in the image through learning [10,11].

Convolutional neural networks (CNNs), a part of the artificial intelligence research field, are currently one of the most popular neural networks. With their special structure with local connections and weight sharing, CNNs decrease model complexity and the

number of weights and cleverly achieve effect that other models cannot achieve with a small number of weights. It does not rely on a complex preprocessing process (extraction of artificial features, etc.) and can directly feed the original data into the constructed model. CNNs have demonstrated strong capabilities in several recent research fields, such as speech recognition [12,13], image recognition [14–16], and image segmentation [17,18]. Although these research fields are different, the feature extraction strategies can be summarized as follows: CNNs can automatically learn features from data and generalize the results to the same type of unknown data.

HSIs have the characteristics of high dual resolution, that is, high spectral resolution and high spatial resolution. Naturally, an HSI is thought to be a data cube. The effective utilization of spectral data and spatial neighborhood information of HSIs is the core to obtain accurate classification results. The 3D convolution kernel simultaneously extracts the hybrid spectral–spatial features of images, which are matched with the 3D features of HSIs. Therefore, 3D CNNs have become one of the important components of deep learning backbone networks in HSI classification [19]. Yu et al. [20] proposed an efficient CNN classification model, which combines data augmentation and larger drop rates in the dropout layers to effectively improve the classification performance of the model with limited training samples. Zhao et al. [21] proposed an HSI classification framework based on multiple convolutional layer fusion, which fuses image information extracted from different convolutional layers to enhance the feature representation of the model. Zheng et al. [22] constructed a feature extraction model by mixing 2D and 3D CNNs and used the covariance pooling technique to extract second-order information from the hybrid spectral–spatial feature map. Li et al. [23] proposed a data-driven joint spatial–spectral attention network that captures the interdependence of land cover and spectral bands by embedding an attention module in the network. Combined with improved triplet loss, Huang et al. [24] constructed a lightweight convolutional neural network model to solve the problems of intra-class diversity and inter-class similarity in HSIs. Gao et al. [25] presented an end-to-end hybrid dilated residual network. Based on residual connections, the spectral and spatial feature learning modules are constructed, respectively, and dilated convolution used in the spatial feature learning module can expand the receptive field and reduce the number of parameters. In [26], stacked autoencoders are used for dimensionality reduction of HSIs, and then 3D CNN and residual connection are combined to extract high-dimensional semantic features of the image. To explore the deep learning HSI classification model with minimal training samples, Huang et al. [27] combined an extended morphological profile, a Siamese CNN, and spectral–spatial feature fusion to construct the network model, which achieved good classification performance. Cai et al. [28] designed a densely connected convolutional extreme learning machine for HSI spectral–spatial classification. Gao et al. [29] proposed a spectral–spatial adaptive fusion network. This model can extract spectral and spatial features pertinently and fuse them adaptively. The above methods only considered the input information of a single scale, which may ignore some fine information and cannot guarantee the discriminability of extracted features.

For CNNs, based on a larger receptive field, the large-scale convolution kernel extracts more neighborhood information and spectral features. In addition, small-scale convolution is more sensitive to details and better at processing boundary information. The introduction of multiple scales can enhance the diversity of features and solve the limitation of a single scale [30–32]. Yang et al. [33] extracted features from multiple input data cubes of different scales according to the size difference of the objects to be classified. The coordinate attention cube composed of coordinate information embedding and coordinate attention generation is adopted to weight the high-dimensional features. Pu et al. [34] proposed a multilevel feature extraction model with an attention mechanism. The proposed attention module enhances the classification performance of the network by exciting or suppressing spectral and spatial features. Zhang et al. [35] designed a random multiscale convolutional network that uses a multiscale dimension reduction module to preprocess data and feeds the data into the random multiscale convolutional module. The model

considers the diversity of different regions of the HSI and extracts more abundant features. Gao et al. [36] proposed a multiscale residual network. In this network, depthwise separable convolution (DSC) is introduced, and mixed depth convolution (MDConv) is used to replace ordinary depth convolution in DSC. Roy et al. [37] proposed an adaptive spectral–spatial kernel residual network with an attention mechanism to learn a selective 3D convolution kernel for HSI classification. An efficient feature recalibration mechanism is used to better extract the nonlinear cross-channel correlation of feature mapping. Jia et al. [38] proposed a lightweight classification network that connects multiple dual-scale convolution modules and bichannel fusion modules to ensure the discrimination of extracted features. Xie et al. [39] designed a multiscale dense convolutional network. In this model, patches of multiple scales around pixels are extracted as the input of the network, and the features of shallow and deep layers are fully fused through skip connections for classification. Zhang et al. [40] designed a diverse region-based CNN model. This model utilizes inputs based on different regions to extract contextual interaction features for better classification performance. Lee et al. [41] designed a context depth CNN that better explores the local context interaction by jointly using the spectral–spatial relationship of neighborhood pixels. Cheng et al. [42] proposed a multiscale HSI classification model, which uses different spatial scales of image patches as the model inputs and extracts spectral and spatial features respectively by CNN and recurrent neural network (RNN). Sun et al. [43] grouped the spectral bands of hyperspectral data through correlation matrix and extracted the spectral and spatial features of each group of data, respectively. Wang et al. [44] designed an adaptive spectral–spatial multiscale feature extraction network. The network consists of spectral and spatial feature extraction subnetworks, and a convolutional long short-term memory (ConvLSTM) model is introduced to obtain context features. Multiscale inputs provide more options for the network model. Through learning, the network can assign weights to the input information of different spatial scales, which greatly reduces the lack of information caused by a too small input spatial scale and the interference information caused by a too large spatial scale. The introduction of the multiscale idea can solve the limitation of a single scale, the extracted features are richer, and the representation ability is stronger.

In recent years, the computer vision field has developed rapidly, and many excellent neural network models have been proposed by researchers. However, the current methods are still insufficient to study the information distribution within the data. The contributions of each element in the data or feature map to the final performance of the model are different, and the attention mechanism effectively solves the problem of information allocation [45]. In deep neural networks, attention mechanisms allocate more weight to the important parts and less weight to the unimportant parts. Visual attention is divided into several types, and its core idea is to explore correlations within data or features and then highlight more significant features, including channel attention [46], spatial attention [47], hybrid domain attention [48,49], nonlocal attention [50], and position attention [51]. Due to the high resolution of HSIs in spectral and spatial dimensions, the classification process is easily disturbed by noise and redundant information. Therefore, an attention mechanism is introduced into the classification model to improve the effectiveness of the data or features, thus enhancing the classification performance. Feng et al. [52] proposed an attention multibranch CNN model based on an adaptive regional search. This method adaptively searches different spatial scales according to the specific distribution of samples and feeds the features of different scales into different branches. In addition, the authors designed a branch attention mechanism to enhance the more discriminative network branch. In [53], a second-order pooling network combining an attention mechanism was proposed. In this method, the first-order discriminative operator is performed to extract the spectral–spatial information of the HSI, and a second-order discriminative operator is designed to model discriminative and representative features based on attention. To solve the insufficient information utilization problems of HSIs, Gao et al. [54] proposed a densely connected multiscale network with an attention mechanism. By introducing the attention module,

the method can further extract the fusion features of the channel and spectral and spatial dimensions, which makes an important contribution to improving the model performance. Combined with the spectral spatial attention mechanism, Guo et al. [55] designed a feature-grouped network. After spectral and spatial attention processing, the generated feature maps are divided into a series of groups along the spectral band direction, and each group extracts features through multiple spectral and spatial residual blocks. Yu et al. [56] designed a spatial–spectral dense CNN framework. The framework follows a compact connection mode to assemble spectral–spatial features, uses two independent dense CNN networks to extract sufficient information, and introduces a band attention module to enhance the feature representation. In [57], an attention-aided CNN model for HSIs was designed. In the method, the spectral and spatial features of the HSIs are separately extracted by two branches. Combined with the attention module, the network pays more attention to the discriminative channels or positions. In [58], a residual spectral–spatial attention network was designed. The input features of the network are first weighted by spectral and spatial attention. Then, further feature extraction is carried out through convolutional networks with residual connections. In general, the introduction of an attention mechanism brings additional performance improvement to the model. For data or feature maps, the attention mechanism can "take the essence and discard the dregs", and the extracted high-level semantic features are highly discriminative. However, notably, the attention mechanism is a weighted mode and an auxiliary functional module, and the model performance mainly depends on the information representation capability of the whole network architecture.

Most of the above deep learning-based HSI classification models contain feature fusion processes, such as using concatenation or addition for the fusion of feature maps at different scales or levels. The information of fused features is more abundant, which is helpful to the classification performance of HSIs. However, the premise of feature fusion using concatenation is that the feature maps should have the same shape in the fusion dimension, while addition requires that the shape of the feature maps be completely consistent. The existing methods mainly depend on the following approaches to fuse feature maps with different scales. In [33,39,56], the small-scale feature maps are upsampled to obtain the same scale as the large-scale feature maps for fusion. In general, the scale of the image changes after convolution. For the same input, the feature scale is reduced more after using a large convolution kernel. When small convolution kernels are used in [34], stride is introduced to ensure the scale consistency of feature maps after multiscale feature extraction. In addition, Lee et al. [41] and Feng et al. [52] used pooling to align feature maps. In the upsampling-based fusion methods, new elements are inserted between the pixels of the original feature to increase the scale of the feature map. New elements inserted in the upsampling process are generated based on existing elements, but these are still not actual information. Therefore, this operation introduces noise to the feature maps. For the downsampling, stride, and pooling methods, the feature map loses information, resulting in the restriction of subsequent classification tasks.

Inspired by the above methods, this paper proposes a subpixel multilevel scale feature learning and adaptive attention constraint fusion (SMS-AACF) to explore the HSI classification task. The SMS-AACF model is mainly composed of three parts, namely, a multilevel scale feature learning module, an adaptive attention constraint fusion module, and a high-level feature semantic enhancement module. There are three-level multiscale strategies in our proposed first module. After principal component analysis (PCA) dimensional reduction, image cubes with two different spatial scales are fed into the neural network. Then, the image cubes of the two scales are processed by interpolation upsampling to obtain their subpixel image cubes. For two original image cubes and their corresponding subpixel cubes, a total of four data cubes are fed into the multiscale feature extraction module. For the multiscale feature attention fusion module, we sampled the small-scale feature map to the same scale as the large-scale feature map and then fused the two feature maps through concatenation. After fusion, the attention weight of the fusion feature map in the spectral

and spatial dimensions is obtained through pooling, convolution, batch normalization, and the sigmoid function. After 1×1 convolution for dimension reduction, the two feature maps before fusion are weighted by an attention matrix, and then, the two feature maps are added point by point. Finally, we re-extracted the high-dimensional semantic features through different scale convolution kernels, which is an extension and enhancement of the original features.

The main contributions of this article are as follows.

(1) We propose a subpixel multilevel scale feature learning and adaptive attention constraint fusion method for HSI classification. The main advantages of our proposed method lie in its strong spectral–spatial feature learning ability, good discrimination of extracted features, and strong model generalization ability. Compared with the existing methods, the proposed method achieves good classification accuracy.
(2) The proposed method further mines the scale information and explores the effectiveness of multiscale information in HSI classification tasks from three levels. Firstly, different spatial scale inputs provide more choices for spatial scale learning of model. Secondly, the subpixel operation can greatly reduce the influence of different categories of pixels and significantly improve the classification performance of boundary locations and small-scale scenes. Finally, multiscale convolution can make the model adapt to different categories of input samples in different scenes.
(3) For the fusion of feature maps at different scales, we propose an adaptive attention constraint fusion method. This method solves problems such as feature loss and noise in the fusion process.
(4) A high-dimensional feature semantic enhancement module is designed, which can be easily inserted into a network model. Through further multiscale feature extraction to improve the semantic representation of the existing feature map, and the proposed method can obtain better classification results.

The remaining sections of this article are organized as follows. Section 2 introduces the motivation and the proposed method. Section 3 evaluates the effectiveness of the proposed method on real HSI datasets. Finally, Section 4 summarizes this article and suggests future work.

2. Motivation and Approach

2.1. Overall Architecture

In view of the problems in the existing deep learning-based HSI classification methods, this paper further explores the application of multiscale information and combines it with the attention mechanism, as shown in Figure 1. Because an HSI contains much noise and redundant information, it introduces much interference in the subsequent feature extraction. Therefore, we first process the original HSI through PCA. On the one hand, this reduces the interference of noise and redundant information. On the other hand, this compresses the spectral dimension of the data, thus reducing the amount of computation. Notably, PCA dimension reduction causes spectral information loss to some extent. After PCA processing, data are less affected by noise and redundant information, and the separability of data is improved. PCA dimension reduction makes it easier for the model to extract features with sufficient discriminative ability. Therefore, we believe that it is worth improving the performance at the expense of a small amount of spectral information.

In the proposed method, the image processed by PCA is divided into image cubes with the set window size and all spectral information. Based on the above image cubes L, the corresponding image cubes S with small spatial scale are generated. Therefore, based on the input image cubes S and L of two different scales, the branch structure of the model is naturally divided into upper and lower parts in Figure 1, where S represents the small-scale input and L represents the large-scale input. For image cube S, S_1 is the subpixel cube obtained after upsampling by interpolation. After S and S_1 are fed into the multiscale feature learning module (MSFL), the adaptive attention constraint fusion module (AACF) is used for feature fusion to obtain feature F_1. Similarly, for image cube L, feature map

F_2 is obtained through the same process as S. We introduce the selection of the scale of the feature maps mentioned above in detail in Section 3.3.4. For F_1 and F_2, AACF is used for fusion to obtain high-dimensional semantic feature F. Then, F is fed into the proposed high-dimensional feature semantic enhancement module (HFSE) for postprocessing; that is, the whole feature extraction process is completed. Finally, the feature map after HFSE treatment is classified through two fully connected layers and one softmax layer.

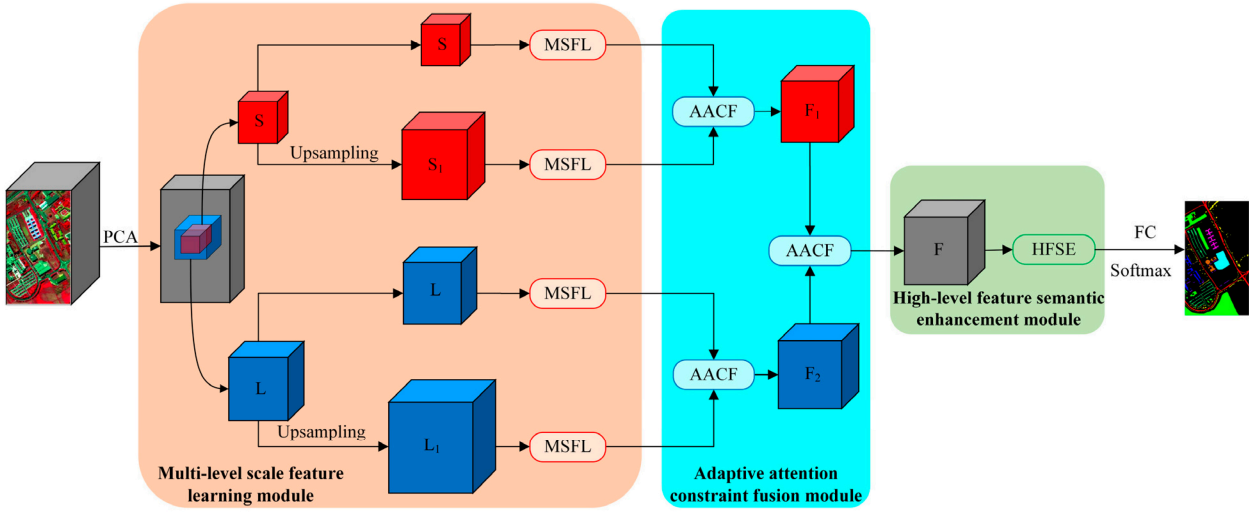

Figure 1. The structure of the proposed method. MSFL denotes the multiscale feature learning module, AACF denotes the adaptive attention constraint fusion module, and HFSE denotes the high-level feature semantic enhancement module. After PCA dimension reduction, two different scales of pixel neighborhood are selected to form S and L; S and L are learned by the upper and lower parts of the network, and then fused by AACF.

Next, we introduce the proposed network model in detail in three parts: subpixel multilevel scale feature learning, adaptive attention constraint fusion, and semantic feature enhancement.

2.2. Subpixel Multilevel Scale Feature Learning

2.2.1. Multiscale Inputs

In the classification of HSIs, the spatial neighborhood information of the center pixel of an image cube is an important basis for extracting discriminative features. However, HSIs have the "same spectral, different material" characteristics and "same material, different spectral" characteristics. An image cube with too large of a selected spatial scale introduces too much interference information. In contrast, if we choose an image cube with a spatial scale that is too small, the neighborhood information is insufficient, resulting in poor classification performance. Most HSI classification methods only use fixed-scale image cubes as input into deep learning models. For the single-scale feature input, the convolution kernel only extracts the image features at the current scale in the moving process but cannot extract the information outside the $m \times m$ neighborhood of the pixel in the HSI, where m is the spatial scale of the selected image cube. Therefore, model performance is sensitive to the selected domain size. In terms of scale selection, the idea of multiscale input is introduced in our proposed method as follows.

The proposed module is a double-input single-output model. Specifically, for each HSI pixel, we take the center pixel and its spatial neighborhood with size $m \times m$ as well as all spectral bands d of the pixel to represent the large-scale spectral–spatial information of the pixel. Based on the $m \times m \times d$ image cubes constructed above, we then select the spatial neighborhood of $n \times n$ based on the central pixel to form the small-scale image cubes of $n \times n \times d$, where $n < m$. Two inputs of different scales are fed into the network model in parallel.

2.2.2. Subpixel Operation

In HSI datasets, the image cubes formed by the center and neighborhood pixels basically belong to the same category. Therefore, large-scale convolutional kernels can extract richer neighborhood information, and small convolution kernels can extract the detailed information of the image. For boundary pixels or small target areas, large-scale convolution will introduce interference from other categories or background pixels, and the advantages of small-scale convolution are reflected at this moment. In addition, the smallest convolution (3×3) that can extract neighborhood information still introduces interference information to a boundary with a large difference or an area with very small pixels (consisting of only a dozen pixels). Based on this problem, subpixel processing of image cubes is proposed in this paper, as shown in Figure 2.

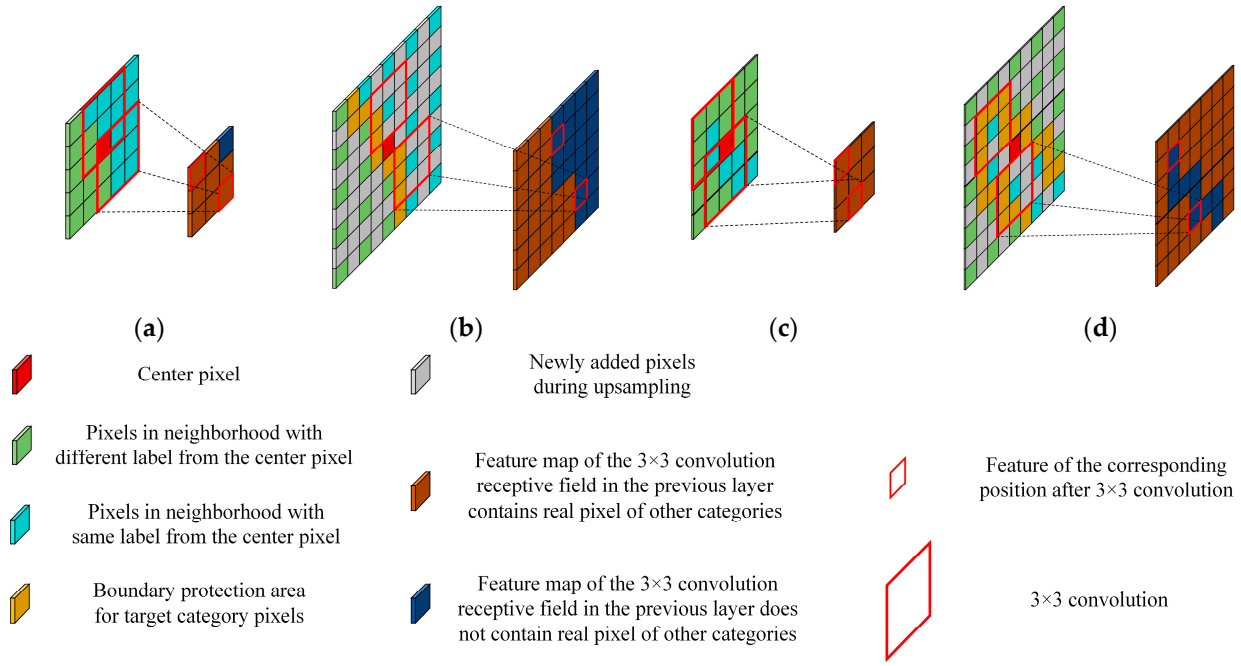

Figure 2. Subpixel operation for different datasets, (**a**) (128, 151) pixel in Salinas with 5×5 spatial scale; (**b**) subpixel operation on (**a**); (**c**) (963, 280) pixel in Pavia Center with 5×5 spatial scale; (**d**) subpixel operation on (**c**).

We select a pixel at the edge of the category and its 5×5 neighborhood information to represent a training sample in Salinas and Pavia Center datasets, respectively, as an example. In Figure 2a,c, the two pixels in red boxes of the same category label as the center pixel contain real sample information of other categories after 3×3 convolution, which will seriously affect the accuracy of subsequent feature extraction. After upsampling the input data (Figure 2b,d), the above two samples and their neighborhoods contain four parts of information: (1) neighborhood samples of the same category, (2) new samples generated from the same category of samples, (3) new samples generated from the same and different categories of samples, and (4) new samples generated from different categories of samples.

In the process of layer-by-layer convolution, the model can adjust parameters through back propagation to better fit the data characteristics. However, if there is a large amount of information different from the target category in the extracted feature maps, it is difficult to extract accurate features from the model. In Figure 2, for the above regions (1), (2), and (4), the model can distinguish these parts of information well in the learning process, regardless of whether upsampling is performed. For the adjacent boundary position of the two categories, the real information of other categories of pixels will be introduced in a large probability after convolution, resulting in serious interference with the extracted features. After upsampling, we can obtain a connected area ((3) of the above regions) composed of

new pixels generated by the target category and other category pixels, which is called the boundary protection area. This area is composed of pixels of the same category as the center pixel and pixels of other categories, which can separate the sample boundaries of different categories, such as the yellow area in Figure 2b,d. At this point, in the boundary target category sample of 3×3 convolution, the 3×3 neighborhood of the target category sample contains the same category of samples and the generated boundary protection area samples. When un-upsampled image cubes are convolved at corresponding positions, extracted features must contain other categories of samples. After upsampling, other categories of samples that 3×3 convolution should contain at this position are replaced by pixels of the boundary protection area. This area weakens the influence of other samples and increases the weight of samples of the same category. Compared with the samples at the corresponding position without upsampling, this method can greatly weaken the influence of other categories of samples. After the convolution operation, the characteristics of the corresponding position are closer to the target sample, which can provide sufficient correct information for the subsequent convolution layer. Therefore, for input image cubes of two different spatial scales, we use the bilinear interpolation upsampling method to obtain their corresponding subpixels. Directly upsampling the input image cube can reduce the pixel difference in the boundary and the interclass interference in a small target scene.

2.2.3. Multiscale Feature Extraction

Based on Sections 2.2.1 and 2.2.2, we adapt the multiscale method for feature extraction. For the two scales of input image cubes in Section 2.2.1, there are four scales of image cubes after the subpixel operation in Section 2.2.2. We carry out feature extraction of the above four parts of the data through the multiscale feature extraction module in Figure 3. In this module, a three-branch network is constructed by using different scales of convolution kernels, and different numbers of convolutional layers are chosen in each branch. Specifically, the first branch uses 16 of the $7 \times 7 \times 7$ convolution kernels. The second branch uses 16 of the $5 \times 5 \times 5$ convolution kernels and 32 of the $3 \times 3 \times 3$ convolution kernels. The third branch successively adopts 16, 32, and 64 convolution kernels with scales of $3 \times 3 \times 3$. Note that the first convolution in each branch of the module uses different scales of the convolution kernel. The convolution of different receptive fields can better extract different neighborhood information from the image. The three branches of this module use different levels of convolutional layers. Through concatenate fusion of extracted features of the three branches, the fusion features contain different levels of semantic information from shallow to deep, with richer information and stronger representation ability.

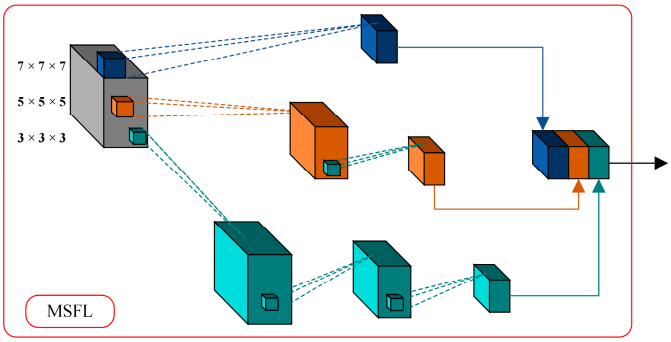

Figure 3. Multiscale feature extraction module.

2.3. Adaptive Attention Constraint Fusion

This paper designs an adaptive attention constraint fusion module in Figure 4 that enriches the semantic information of the fused features in a learnable way and enhances its discrimination. From another point of view, the proposed method modifies the upsampled small-scale feature map through a determined feature map (large-scale feature map) and

adjusts and constrains its internal feature distribution to reduce the noise caused by the upsampling process. The computation equation of the fusion module is as follows:

$$M = \sigma(E(\varphi(I, I')) \times E'(\varphi(I, I'))), \tag{1}$$

$$Out = I \times M + I' \times M, \tag{2}$$

where M is the extracted attention weight matrix; σ indicates the compressed 1×1 convolution of the weight matrix; E and E' represent the pooling process, 1×1 convolution, batch normalization, and the sigmoid activation function operation in the spectral and spatial weight generation branches; φ represents concatenation; and Out is the output of the fusion module. For the feature maps $I \in R^{d \times H \times W}$ and $I' \in R^{d \times H' \times W'}$ of two different spatial scales, assuming $H > H'$ and $W > W'$, we first process the small-scale feature maps through bilinear interpolation upsampling to align the spatial scales of the two feature maps and then fuse the two features by concatenation, where H and W indicate the spatial scale of the feature map and d is the spectral shape. The fused features are processed by pooling, convolution, batch normalization, and the sigmoid activation function. This is a two-branch structure: one branch learns the spatial attention weight, and the other branch learns the spectral attention weight. After multiplying the two weights, 1×1 convolution is used for channel compression. Finally, the weight matrix is weighted and then added to the two feature maps for adaptive multiscale feature attention fusion.

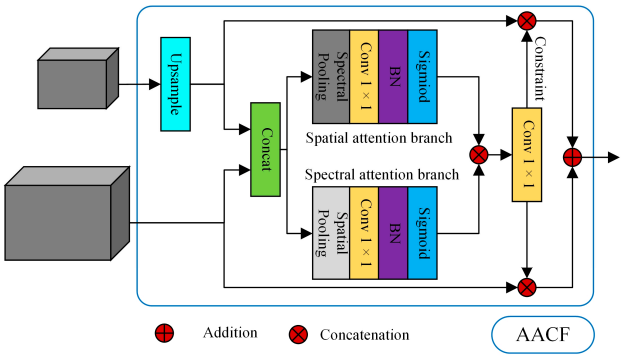

Figure 4. Adaptive attention fusion module.

2.4. High-Level Feature Semantic Enhancement Module

In the HSI classification task, the extracted high-dimensional semantic features are generally considered to have good discrimination ability based on the constructed network model. However, higher dimensional semantic features can be further mined to enhance their representational ability in some cases.

A postprocessing-based attention mechanism is an effective way to enhance high-dimensional semantic features. In this paper, the high-dimensional feature semantic enhancement module is proposed, which can be regarded as a postprocessing attention mechanism. Through this module, the extracted features have stronger discrimination, and the classification performance is further improved. The flowchart of the enhancement module is shown in Figure 5, and the computation equation of the enhancement module is as follows:

$$Out' = B_1 + B_2 \times F(B_3, B_4, B_5), \tag{3}$$

where Out' represents the output of the feature enhancement module; B_1, B_2, B_3, B_4, and B_5 are outputs of the five branches of the network from top to bottom; and F indicates the concatenation and convolution processes for the output of the bottom three branches. The proposed high-dimensional feature semantic enhancement module has a multibranch structure, including a backbone branch, a channel attention branch, and three information supplement branches for mining multiscale information. In the backbone branch, we only use one layer of 1×1 convolution to further extract the spectral information. For the

channel attention branch, feature maps are compressed to one dimension in the spatial scale through average pooling after 1×1 convolution and batch normalization; that is, a vector with the same channel scale is used as the input of the backbone branch. The vector represents the spectral distribution of the feature map. The three information supplement branches are first processed through a 1×1 convolution, and then $3 \times 3 \times 3$, $5 \times 5 \times 5$, and $7 \times 7 \times 7$ convolutions, and the corresponding batch normalization are used in each branch for multiscale feature extraction. After that, the sigmoid activation function is chosen to constrain the value of the three branches between 0 and 1. At this time, the features of the three branches have the feature representation weights of different scales of the feature map. Next, the information of the three branches is fused through concatenation, and then, the channel dimension is compressed to the same dimension as the feature maps in the first and second branches using a 1×1 convolution. Finally, the fused feature is multiplied by the feature map in the second branch and added to the feature map in the first branch.

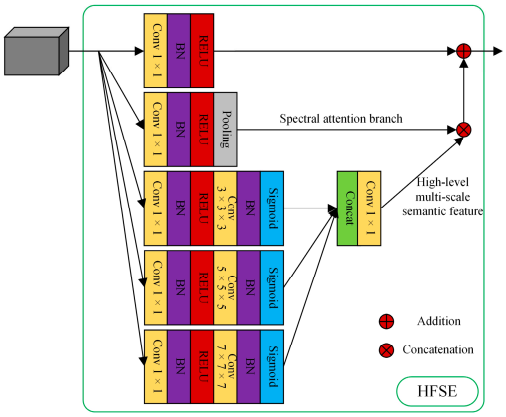

Figure 5. High-level feature semantic enhancement module.

3. Experimental Results

3.1. Dataset Description

Four real-world HSI datasets, Pavia University (PU) [59], Pavia Center (PC) [59], Salians (SA) [59], and Huston University (HU) [60], are considered to evaluate our proposed model. In the following, we describe the four datasets in detail. For PU, PC, and SA, 1% of the samples are selected for training, the other 1% for validation, and the remaining 98% for testing. For HU, 3% of the samples are selected for training, the other 3% for validation, and the remaining 94% for testing. The color of each category in the classification maps and the training validation and testing sample numbers are shown in Tables 1–4.

Table 1. Training validation and testing sample numbers in PU.

No.	Class Name	Train	Val	Test	Total
1	Asphalt	66	66	6499	6631
2	Meadows	186	186	18,277	18,649
3	Gravel	21	21	2057	2099
4	Trees	31	31	3002	3064
5	Painted metal sheets	13	13	1319	1345
6	Bare Soil	50	50	4929	5029
7	Bitumen	13	13	1304	1330
8	Self-Blocking Bricks	37	37	3608	3682
9	Shadows	9	9	929	947
Total		426	426	41,924	42,776

Table 2. Training validation and testing sample numbers in PC.

No.	Class Name	Train	Val	Test	Total
1	Water	659	659	64,653	65,971
2	Trees	76	76	7446	7598
3	Asphalt	31	31	3028	3090
4	Self-Blocking Bricks	27	27	2631	2685
5	Bitumen	66	66	6452	6584
6	Tiles	92	92	9064	9248
7	Shadows	73	73	7141	7287
8	Meadows	428	428	41,970	42,826
9	Bare Soil	29	29	2805	2863
Total		1481	1481	145,190	148,152

Table 3. Training validation and testing sample numbers in SA.

No.	Class Name	Train	Val	Test	Total
1	Brocoli_green_weeds_1	20	20	1969	2009
2	Brocoli_green_weeds_2	37	37	3652	3726
3	Fallow	20	20	1936	1976
4	Fallow_rough_plow	14	14	1366	1394
5	Fallow_smooth	27	27	2624	2678
6	Stubble	40	40	3879	3959
7	Celery	36	36	3507	3579
8	Grapes_untrained	113	113	11,045	11,271
9	Soil_vinyard_develop	62	62	6079	6203
10	Corn_senesced_green_weeds	33	33	3212	3278
11	Lettuce_romaine_4wk	11	11	1046	1068
12	Lettuce_romaine_5wk	19	19	1889	1927
13	Lettuce_romaine_6wk	9	9	898	916
14	Lettuce_romaine_7wk	11	11	1048	1070
15	Vinyard_untrained	73	73	7122	7268
16	Vinyard_vertical_trellis	18	18	1771	1807
Total		543	543	53,043	54,129

The PU dataset was collected in 2001 by the ROSIS sensor over Pavia University, Italy. The dataset comprises 610 × 340 pixels with a spatial resolution of 1.3 m. There are 103 bands in the wavelength range from 0.43 to 0.86 µm and nine distinguishable class labels.

The PC dataset was gathered in 1998 by an AVIRIS sensor in SA Valley, CA, USA. The dataset comprises 1096 × 715 pixels with a spatial resolution of 1.3 m. There are 102 bands in the wavelength range from 0.43 to 0.86 µm and nine distinguishable class labels.

The SA dataset was gathered in 1998 by AVIRIS in SA Valley, CA, USA. The dataset comprises 512 × 217 pixels with a spatial resolution of 3.7 m. There are 224 bands in the wavelength range from 0.4 to 2.5 µm and 16 distinguishable class labels.

The HU dataset was released in the 2013 IEEE GRSS Fusion Contest. The HU dataset is a more challenging task that was captured by the National Center for Airborne Laser Mapping (MCALM) over the HU campus. The dataset comprises 349 × 1905 pixels with

a spatial resolution of 2.5 m. There are 144 bands in the wavelength range from 0.36 to 1.05 μm and 15 complex land-cover classes.

Table 4. Training validation and testing sample numbers in HU.

No.	Class Name	Train	Val	Test	Total
1	Healthy grass	38	38	1175	1251
2	Stressed grass	38	38	1178	1254
3	Synthetic grass	21	21	655	697
4	Trees	37	37	1170	1244
5	Soil	37	37	1168	1242
6	Water	10	10	305	325
7	Residential	38	38	1192	1268
8	Commercial	37	37	1170	1244
9	Road	38	38	1176	1252
10	Highway	37	37	1153	1227
11	Railway	37	37	1161	1235
12	Parking Lot 1	37	37	1159	1233
13	Parking Lot 2	14	14	441	469
14	Tennis Court	13	13	402	428
15	Running Track	20	20	620	660
Total		452	452	14,125	15,029

3.2. Experimental Setup

To verify the superiority of our proposed model, some related methods are selected for comparison, including SVM, CDCNN, SSRN, FDSSC, HybridSN, DBDM, and MCNN. The details of the compared methods are described as follows:

(1) SVM: This is a method that relies only on spectral information and uses a SVM as a classifier.

(2) CDCNN: In this method, a 2D CNN at different scales is used to extract multiscale features, and then, high-dimensional semantic features are obtained by combining a 1×1 convolution and residual connection. In addition, image cubes with a size of $5 \times 5 \times L$ are selected as the model input. L represents the number of spectral bands of the image cube [41].

(3) SSRN: It is a classification network composed of spectral and spatial feature learning modules in parallel. Combined with residual connections, spectral features are extracted by a 1×1 convolution, and spatial features are extracted by a 3D convolution. Image cubes with a size of $7 \times 7 \times L$ are selected as the model input [32].

(4) FDSSC: It is a densely connected spectral–spatial feature extraction classification network, where the spectral and spatial features are separately extracted by a 1×1 convolution and 3D convolution. In this method, image cubes with a spatial scale of $9 \times 9 \times L$ are selected as model inputs [30].

(5) HybridSN: HybridSN is a single branch network that combines 3D convolution and 2D convolution, where $25 \times 25 \times L$ spatial scale image cubes are selected as model inputs [31].

(6) DBMA: This is a two-branch network model that extracts spectral and spatial features from two branches. Then, the features extracted from the two branches are weighted by channel attention and spatial attention. In this method, image cubes with a spatial scale of $7 \times 7 \times L$ are selected as model inputs [45].

(7) MCNN: This is an improved method of (5). Based on the backbone network of (5), the covariance pooling technique is used to extract the second-order feature information. In this method, 11 × 11 × L spatial scale image cubes are selected as model inputs [22].

The seven compared methods include traditional methods and deep learning methods in the feature learning level, single-scale and multiscale methods in the data or feature scale level, single branch and multibranch in the branch strategy, and single and hybrid spectral–spatial feature extraction in the feature extraction method. The above compared methods make a more comprehensive comparison of the proposed method to prove its effectiveness in multiple dimensions.

In our experiment, the learning rate and the batch size are set to 0.0005 and 32, respectively, during stochastic gradient decent training, and the training epoch is set as 80. In addition, we determine cross entropy as the loss function. After PCA dimensionality reduction, the number of bands in PU, PC, SA, and HU are reduced to 15, 15, 15, and 12, respectively. All the experiments are repeated 10 times, and the average of 1- experiments is given as a result. The experimental environments of the proposed method are all implemented in the Python operating language, using an Intel i7-9700 CPU and an NVIDIA GeForce RTX 2080ti GPU.

3.3. Classification Results

In this part, we provide quantitative and qualitative evaluations to verify the effectiveness of our proposed method and evaluate the method by the following classification metrics: overall accuracy (OA), average accuracy (AA), and kappa coefficient (KA).

3.3.1. Quantitative Analysis

Tables 5–8 present the quantitative results of the proposed method and comparison methods on the four datasets. Overall, the proposed method outperforms all of the compared methods in terms of OA, AA, and KA on the four datasets. For the seven comparison methods, the SVM and CDCNN obtain the weakest classification performance on the four datasets. Compared with those of the SVM and CDCNN, the performances of SSRN, HybridSN, DBMA, and MCNN are significantly improved but slightly lower than the performance of FDSSC.

Table 5. Classification results of different models on the PU dataset.

Class Name	SVM	CDCNN [41]	SSRN [32]	FDSSC [30]	Hybrid SN [31]	DBMA [45]	MCNN [22]	Proposed
Asphalt	84.65	90.21	98.90	**99.44**	95.76	96.50	97.75	98.46
Meadows	92.57	94.66	98.23	99.45	98.73	98.72	99.43	**100**
Gravel	74.94	64.95	98.93	99.52	85.03	**100**	93.83	97.55
Trees	70.53	97.24	**99.64**	97.61	97.83	97.85	88.84	90.80
Painted metal sheets	90.19	98.36	99.70	99.70	99.70	99.25	98.94	**100**
Bare Soil	66.41	93.11	98.62	98.50	**99.82**	99.15	95.50	99.42
Bitumen	78.87	96.88	94.25	**100**	89.09	96.97	93.33	99.85
Self-Blocking Bricks	83.84	88.98	84.91	80.08	88.47	83.23	90.94	**97.20**
Shadows	98.94	99.17	99.78	99.89	98.62	**100**	97.73	98.19
OA(%)	84.71	91.88	97.12	97.19	96.38	96.85	96.70	**98.63**
AA(%)	82.33	91.51	96.99	97.13	94.78	96.85	95.14	**97.94**
KA(%)	79.45	89.15	96.17	96.27	95.19	95.81	95.61	**98.18**

An HSI is an "image cube" that integrates not only spectral information, but also spatial features. The SVM classifier only considers the spectral information of pixels. Limited by the lack of neighborhood spatial information, the model cannot obtain sufficient discriminative features. In addition, classification based on spectral information alone is easily influenced by the "same materials, different objects" and "same objects, different materials" phenomena in the HSIs. Therefore, the classification performance of

the SVM is the weakest, and the OAs on PU, PC, SA, and HU are 84.71%, 97.04%, 86.89%, and 80.03%, respectively.

Table 6. Classification results of different models on the PC dataset.

Class Name	SVM	CDCNN [41]	SSRN [32]	FDSSC [30]	Hybrid SN [31]	DBMA [45]	MCNN [22]	Proposed
Water	99.84	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Trees	89.43	90.45	98.89	98.29	93.07	**99.93**	95.35	97.82
Asphalt	83.88	91.36	87.04	90.25	97.78	63.16	95.28	**99.74**
Self-Blocking Bricks	62.35	79.29	71.05	94.98	99.89	82.61	99.58	**99.96**
Bitumen	96.04	92.33	**99.83**	99.67	98.64	97.62	93.61	98.60
Tiles	92.41	96.47	99.33	98.40	99.32	98.16	97.02	**99.41**
Shadows	89.60	94.65	97.67	**100**	96.41	98.91	98.11	99.54
Meadows	99.44	99.81	99.97	**99.99**	99.75	99.94	99.64	99.72
Bare Soil	**99.97**	99.11	99.89	96.82	88.71	99.54	90.02	93.15
OA(%)	97.04	98.04	98.73	99.42	99.04	98.13	98.79	**99.55**
AA(%)	90.33	93.72	94.85	97.60	97.09	93.32	96.51	**98.66**
KA(%)	95.81	97.22	98.20	99.18	98.64	97.36	98.29	**99.36**

Table 7. Classification result of different methods on the SA dataset.

Class Name	SVM	CDCNN [41]	SSRN [32]	FDSSC [30]	Hybrid SN [31]	DBMA [45]	MCNN [22]	Proposed
Brocoli_green_weeds_1	99.10	64.11	99.95	**100**	98.42	**100**	**100**	**100**
Brocoli_green_weeds_2	97.61	99.92	99.97	**100**	99.97	99.92	**100**	**100**
Fallow	98.18	95.48	99.83	98.38	**100**	**100**	**100**	**100**
Fallow_rough_plow	97.20	94.00	96.37	94.45	95.50	97.47	99.78	**100**
Fallow_smooth	96.64	93.40	93.48	**99.92**	88.23	81.79	95.54	99.74
Stubble	98.41	99.01	**100**	99.97	**100**	**100**	99.97	99.97
Celery	99.16	99.07	**100**	99.91	99.69	**100**	**100**	**100**
Grapes_untrained	73.29	96.75	78.06	98.40	98.77	89.38	99.38	**100**
Soil_vinyard_develop	98.21	99.84	99.77	**100**	99.84	99.61	99.97	99.66
Corn_senesced_green_weeds	79.77	82.52	97.96	92.41	**99.63**	96.43	99.00	99.38
Lettuce_romaine_4wk	92.79	92.64	**100**	**100**	**100**	99.24	97.90	99.15
Lettuce_romaine_5wk	97.30	99.63	99.89	**100**	**100**	99.95	99.31	**100**
Lettuce_romaine_6wk	97.27	97.56	96.87	**100**	99.55	**100**	92.20	99.12
Lettuce_romaine_7wk	69.81	98.85	99.41	98.48	98.78	95.74	**99.62**	99.34
Vinyard_untrained	67.47	41.64	**98.82**	97.31	96.38	98.06	87.41	97.82
Vinyard_vertical_trellis	94.19	98.87	**100**	99.09	**100**	99.88	99.60	**100**
OA(%)	86.89	76.84	93.43	98.52	98.33	95.86	97.67	**99.57**
AA(%)	91.03	90.83	97.52	98.65	98.42	97.34	98.11	**99.64**
KA(%)	85.37	74.63	92.65	98.36	98.15	95.38	97.40	**99.52**

At present, most deep learning-based HSI classification methods combine the spectral and spatial information of images and achieve significant performance improvement. Although CDCNN considers both spectral and spatial information of images, its classification model is relatively simple. In addition to using different scale convolutions for multiscale feature extraction at the front end of the model, the subsequent model is composed of a 1×1 convolution. Only relying on a parallel single-layer multiscale feature extraction strategy to mine the global spatial information of the image is insufficient for HSIs with considerable noise and redundant information. Compared with SVM, the classification performances of CDCNN on PU, PC, and HU are improved by 7.17%, 1%, and 4.16%, respectively, in terms of the OA. Because different categories of ground objects in SA are concentrated, most pixel neighborhoods are consistent with the central pixel category. Therefore, exploiting spatial information is an effective way to classify the dataset. However, there are a large number of 1×1 convolutions in the CDCNN network, which leads

to the poor classification performance of SA. Among them, the OA and KA are both lower than those of the SVM, and only the AA is close to that of the SVM.

Table 8. Classification result of different methods on the HU dataset.

Class Name	SVM	CDCNN [41]	SSRN [32]	FDSSC [30]	Hybrid SN [31]	DBMA [45]	MCNN [22]	Proposed
Healthy grass	85.61	85.94	95.15	93.16	98.43	98.11	96.43	**98.93**
Stressed grass	93.30	93.43	98.43	98.72	98.03	98.98	97.28	**100**
Synthetic grass	97.70	97.95	99.54	**100**	99.41	**100**	99.85	99.85
Trees	98.79	96.65	95.94	**100**	90.72	97.88	92.65	98.76
Soil	98.15	93.55	92.07	99.83	**100**	93.20	**100**	**100**
Water	82.15	99.22	**100**	**100**	90.48	**100**	95.08	98.41
Residential	82.65	89.52	91.79	90.43	82.85	92.92	94.88	**97.64**
Commercial	52.65	96.73	93.38	**96.95**	72.91	95.35	92.39	91.63
Road	77.64	60.40	84.42	92.34	87.08	94.12	87.52	**96.13**
Highway	86.72	75.64	97.34	99.14	**99.66**	89.09	97.66	99.41
Railway	66.00	72.28	87.20	98.98	95.49	96.15	98.11	**99.58**
Parking Lot 1	44.77	78.80	92.75	**97.48**	82.27	86.57	97.15	95.57
Parking Lot 2	55.86	91.57	83.20	74.56	81.32	88.31	88.66	**95.82**
Tennis Court	98.60	86.88	99.25	**100**	**100**	92.68	**100**	**100**
Running Track	97.27	94.01	97.47	98.40	**100**	95.21	**100**	**100**
OA(%)	80.30	84.46	93.22	96.10	91.50	94.32	95.72	**97.94**
AA(%)	81.19	87.50	93.86	96.00	91.91	94.57	95.84	**97.77**
KA(%)	78.72	83.20	92.67	95.78	90.81	93.86	95.37	**98.08**

The feature extraction abilities of SSRN, FDSSC, and HybridSN are based on the spectral and spatial information of the HSIs. Both SSRN and FDSSC extract the spectral information by a 1×1 convolution and then extract the spatial information through a 3D CNN; that is, the spectral and spatial feature extraction modules are constructed in series. HybridSN extracts spectral–spatial features by a hybrid 2D-3D CNN. The classification performances of the three methods on PU, PC, and SA are superior to those of the SVM and CDCNN. Because the HU dataset is complex, there are many categories of objects, and the distribution of features is scattered, making it difficult to classify them. HybridSN is limited by its relatively simple model, and its prediction ability is insufficient for the object categories in HU.

The proposal of an attention mechanism can highlight the important information in the feature map and weaken the less useful part, which is a way to enhance model performance. DBMA constructs spectral and spatial attention mechanisms in two network branches of the model. Although the MCNN does not directly use the attention mechanism, the channelwise shift and channelwise weighting constructed by the MCNN move and weight the data processed by PCA in the channel dimension, so the MCNN can change the information distribution. Therefore, DBMA and MCNN obtained OAs greater than 96.7%, 98.1%, 95.8%, and 94.3% on the four datasets. Among them, for the HU dataset, which is difficult to classify, the classification performances of DBMA and MCNN are significantly improved compared with those of the SVM, CNCNN, SSRN, and HybridSN, and are only slightly lower than the classification performance of FDSSC. Note that the classification performance of FDSSC on multiple datasets is superior to that of other compared methods (including those using attention methods), but this does not mean that the attention mechanism has no advantage. The addition of an attention mechanism can enhance the model performance, but it cannot play a leading role. An attention mechanism is a kind of auxiliary functional module. The feature extraction ability of the model is the main factor that determines the classification performance.

Based on multiscale input, the subpixel strategy, and multiscale convolution feature extraction, the proposed method fully mines the multiscale information of the HSI, which improves its classification ability for some small objects. The constructed adaptive attention

constraint fusion module and high-dimensional feature semantic enhancement module also play a significant role in model performance. The classification results obtained on all experimental datasets prove the effectiveness of the proposed method. For the HU dataset with many small objects, the proposed method shows significant performance improvement. Compared with the other seven methods with the best classification accuracies on each dataset, our method achieves OAs that are improved by 1.44%, 0.13%, 1.05%, and 1.84%, respectively.

3.3.2. Qualitative Analysis

To reflect the classification performance of each method more directly, Figures 6–9 show the classification results of the best trained models on the four datasets, along with the ground-truth maps and their false color images. As seen in Figures 6–9, the SVM and CDCNN only depend on the spectral information or do not fully use the spatial information, resulting in serious salt-and-pepper noise. SSRN, FDSSC, HybridSN, DBMA, and MCNN extract sufficient spectral and spatial information of HSIs, so smooth classification maps are obtained. The proposed method not only deeply mines the spectral and spatial information of HSIs, but also reduces noise interference of different scale features in the fusion process and enhances the high-dimensional semantic features. In addition, the combination of original image cubes and subpixel cubes improves the classification performance of the model for small target scenes (HU and some targets in PU and PC). Overall, the proposed method shows a good classification effect on the four datasets and delivers the most accurate and smooth classification maps.

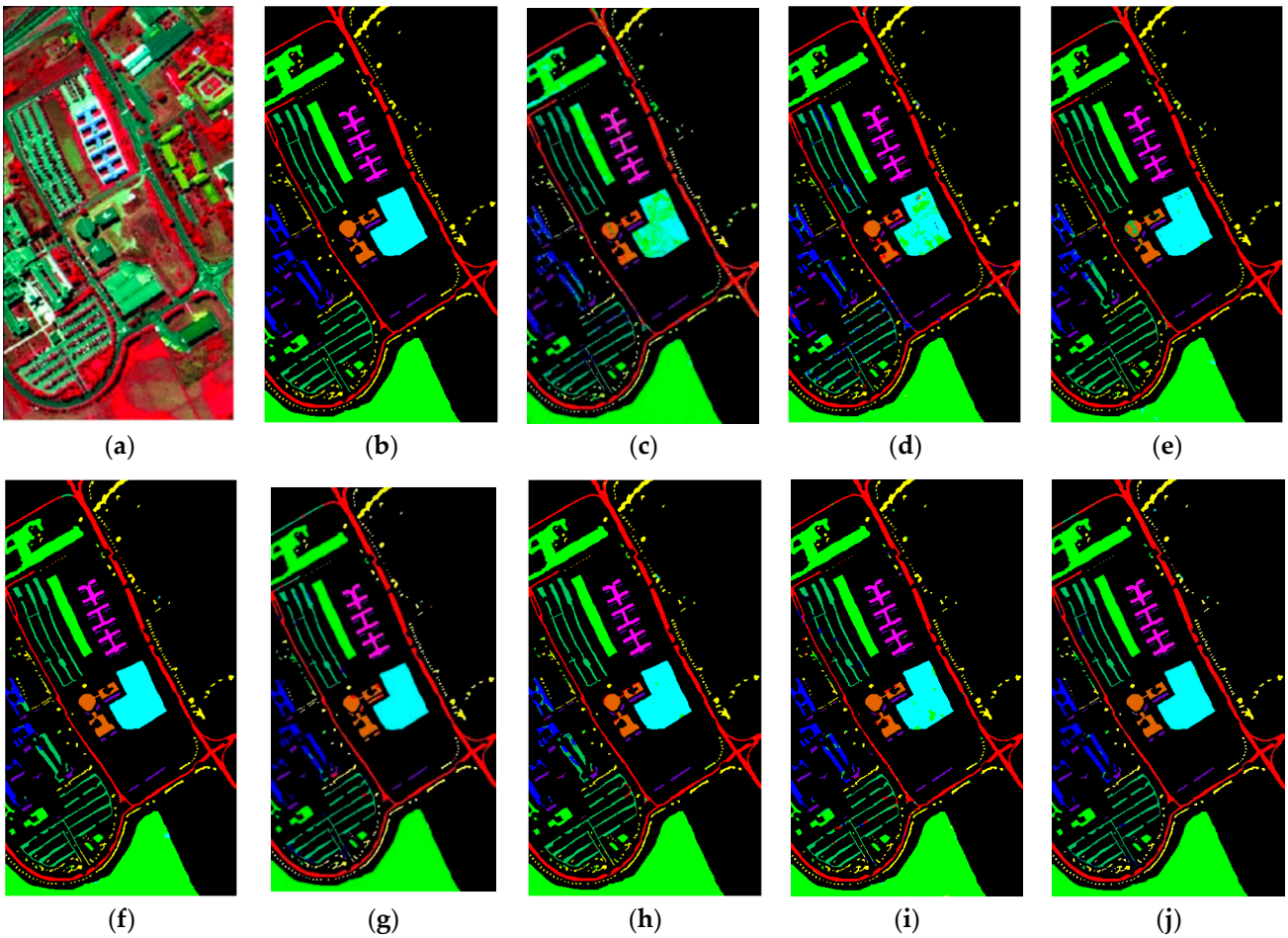

Figure 6. The classification map of Pavia University. (**a**) False-color image; (**b**) ground truth; (**c**) SVM; (**d**) CDCNN; (**e**) SSRN; (**f**) FDSSC; (**g**)HybirdSN; (**h**) DBMA; (**i**) MCNN; (**j**) the proposed method.

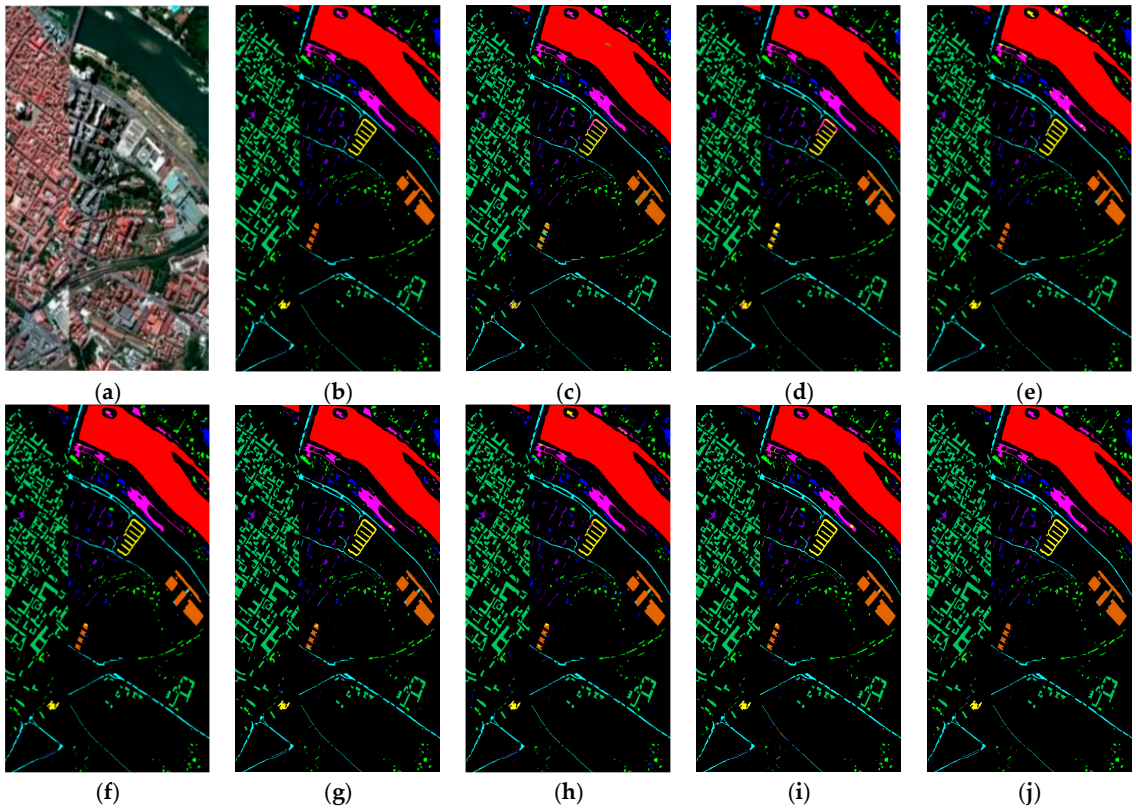

Figure 7. The classification map of Pavia Center. (**a**) False-color image; (**b**) ground truth; (**c**) SVM; (**d**) CDCNN; (**e**) SSRN; (**f**) FDSSC; (**g**) HybirdSN; (**h**) DBMA; (**i**) MCNN; (**j**) the proposed method.

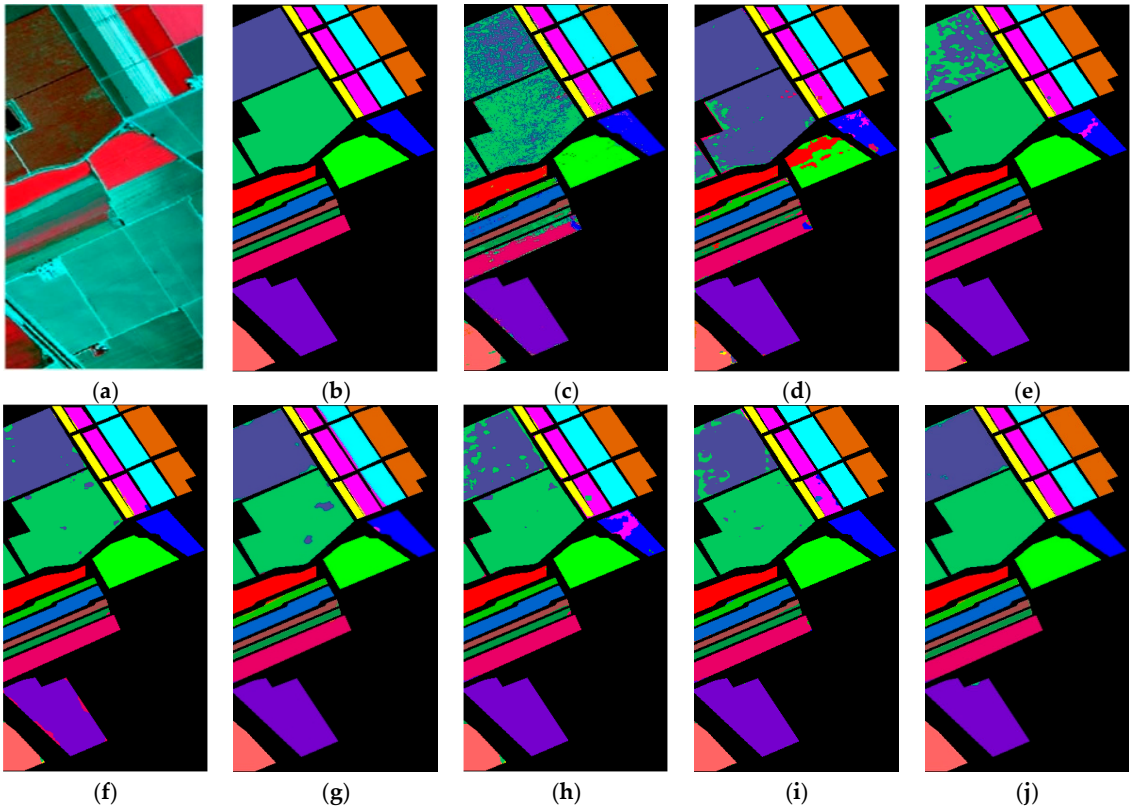

Figure 8. The classification map of Salinas. (**a**) False-color image; (**b**) ground truth; (**c**) SVM; (**d**) CDCNN; (**e**) SSRN; (**f**) FDSSC; (**g**) HybirdSN; (**h**) DBMA; (**i**) MCNN; (**j**) the proposed method.

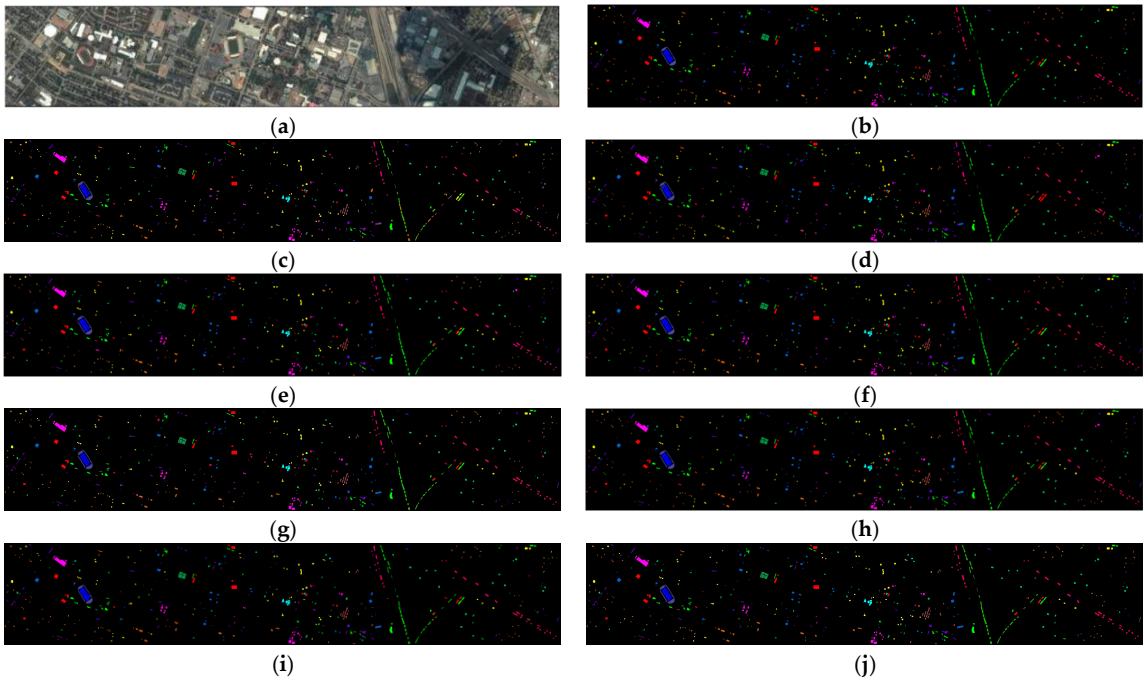

Figure 9. The classification map of Huston University. (**a**) False-color image; (**b**) ground truth; (**c**) SVM; (**d**) CDCNN; (**e**) SSRN; (**f**) FDSSC; (**g**) HybirdSN; (**h**) DBMA; (**i**) MCNN; (**j**) the proposed method.

3.3.3. Comparison Analysis of Using Different Percentages of Training Samples

A deep neural network has powerful feature extraction ability, but it relies on a large number of samples to train the network. The classification accuracy of the network model is also greatly related to the number of training samples. To test the robustness and generalization of the proposed method, 1%, 3%, 5%, 8%, 10%, and 15% of samples are randomly selected to train the proposed model. The experimental results are given in Figure 10. Because the classification performances of the SVM and CDCNN are obviously lower than those of other methods, there is no comparison between these two methods in this part.

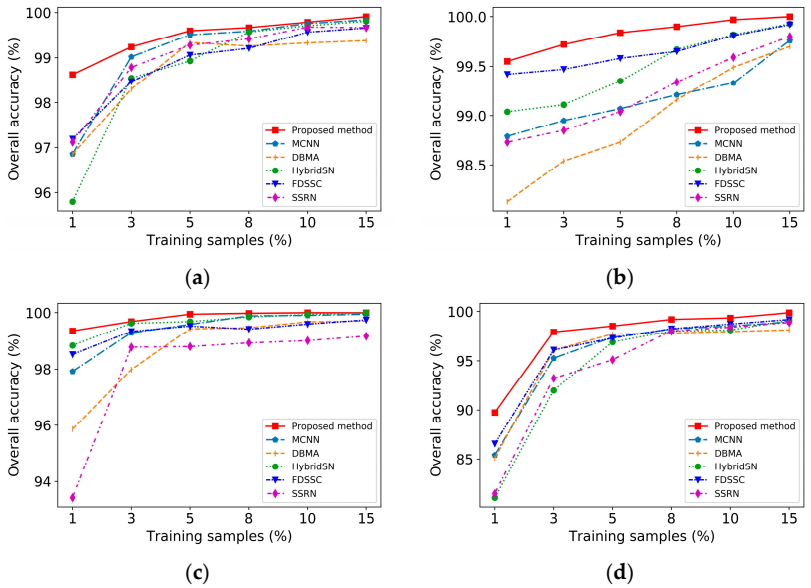

Figure 10. OA results of the six methods with different numbers of training samples. (**a**) PU, (**b**) PC, (**c**) SA, and (**d**) HU.

For a deep learning model, the more labeled samples that are used for training, the stronger the feature representation ability of the model and the better the generalization ability. Therefore, the classification accuracy of all methods in different datasets shows a positive correlation with the number of training samples. The proposed method is more thorough for scale feature mining, and the features extracted from the model can distinguish between different categories of land cover objects. Therefore, the proposed method shows the optimal classification performance under all test conditions.

3.3.4. Classification Performance for Different Spatial Sizes

The spatial scale of the input feature map is a key factor that determines the model performance. A spatial scale that is too large includes a lot of neighborhood information. Domain information may contain many other categories of land cover objects, which causes great interference with the feature extraction process. In contrast, if the selected spatial scale is too small, the model lacks considerable spatial information in feature extraction. The information loss leads to the weak ability of the model to discriminate between the extracted features, meaning that the model cannot accurately classify different categories of objects. According to the study of the works in HSI classification field, we find that the input scales of HSIs are mostly distributed from 7×7 to 25×25. Therefore, based on the large spatial scale input and small spatial scale input proposed in the manuscript, 15×15 is chosen as the partition point of the two scales. The spatial scales less than 15×15 are defined as small spatial scales, and the spatial scales greater than or equal to 15×15 are defined as large spatial scales. To select the best spatial scale and explore the model performance with scale variation, we test a total of 16 combinations of the 7×7, 9×9, 11×11, and 13×13 small-scale information and 15×15, 17×17, 19×19, and 21×21 large-scale information. The experimental results are shown in Figure 11. Among them, the upsampling of both small-scale input and large-scale input in the model is 2S-1, where S represents the spatial scale of the input patch.

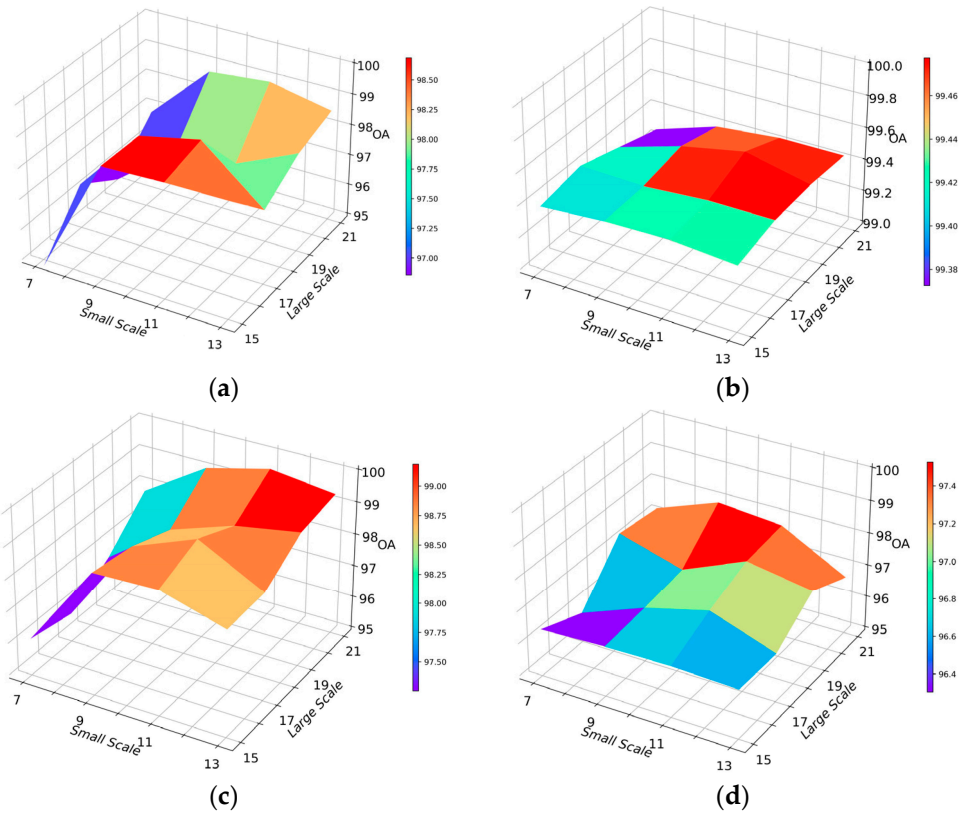

Figure 11. Analysis of the effect of different combinations of spatial scales. (**a**) PU, (**b**) PC, (**c**) SA, and (**d**) HU.

With the increase in the two input scales, the model performance shows an overall trend of increasing first and then decreasing, which is consistent with the previous theoretical expectations. When the input scale of the model gradually increases, the neighborhood information contained in the input information increases gradually. Under this condition, the spatial information becomes more abundant, and the classification performance of the model shows an upward trend. When the scale information increases to a certain extent, further increasing the spatial scale makes the input image cube contain a large amount of information from other categories, which weakens the ability of the model to discriminate between extracted features and eventually results in a decline in classification performance. According to the optimal experimental performance, the small-scale input and large-scale input of the model are 21 × 21 and 11 × 11 on PU, 17 × 17 and 11 × 11 on PC, 21 × 21 and 11 × 11 on SA, and 19 × 19 and 9 × 9 on HU, respectively. Please note that the above analysis is only limited to the proposed method.

3.4. Ablation Study

3.4.1. Effectiveness Analysis of the Multiscale Input Strategy

This paper explores the importance of scale information in HSI classification from three aspects: multiscale input, input data subpixel, and multiscale convolution feature extraction. In this part, the multiscale input ablation experiment is performed to prove the effectiveness of the proposed method. For the proposed multilevel scale feature extraction module, we fix the scale of the input image cube and only use a single scale image cube as the model input for a comparative experiment. The experimental results are shown in Figure 12, where S_scale means small scale alone; that is, only the upper parts of the multi-level scale feature learning module and the adaptive attention constraint fusion module are retained in Figure 5. L_scale represents large scale alone; that is, only the lower parts of the above two modules are retained. Finally, M_scale indicates that the entire network structure is executed.

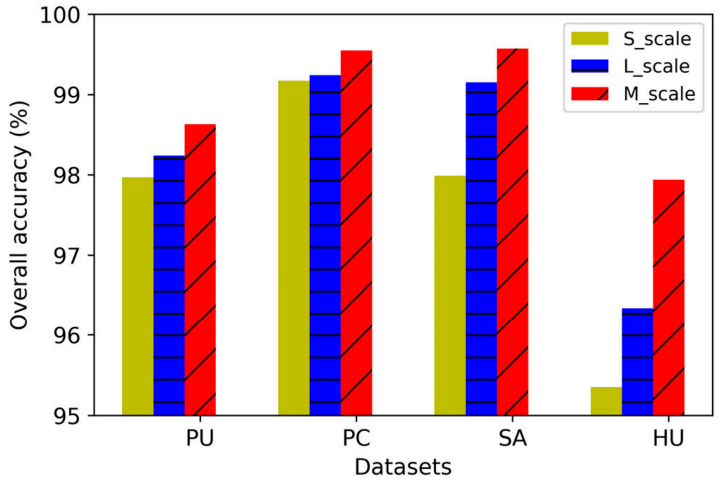

Figure 12. Effectiveness of multiscale inputs.

In all datasets, the OA of the model using only large-scale input is better than that using only small-scale input. The combined use of large-scale and small-scale image cubes as input further improves the classification accuracy. In the experiment of the three input methods, the model performance gap is most obvious on HU. Because the HU is captured over the HU campus, there are many categories of pixels, and the distribution of objects in HU is relatively scattered. Meanwhile, the same category of pixels in HU do not show a large area aggregation, the classification complexity is high. Therefore, when extracting the feature of HU, the spatial neighborhood information is easy to be mixed with other categories of pixels, which greatly increases the difficulty of classification. Through learning, the model can allocate the weight of parameters to different scales of information

and extract highly discriminative semantic features from rich neighborhood information. In summary, using the input of neighborhood information at different scales can improve the classification performance.

3.4.2. Effectiveness Analysis of Subpixels

In this paper, the subpixel operation of the input information is performed to improve the classification accuracy of the model for small objects in complex environments. We compare large-scale input subpixel processing, small-scale input subpixel processing, and no subpixel processing. Figure 13 shows the experiment studying the influence of subpixel operations on model performance, where WO_S indicates that neither of the two input image cubes of the model is subjected to subpixel processing, WI_SS indicates that only small-scale input information is subjected to subpixel processing, WI_LS indicates that only large-scale input is subjected to subpixel processing, and WI_SLS indicates that both inputs are subjected to subpixel processing.

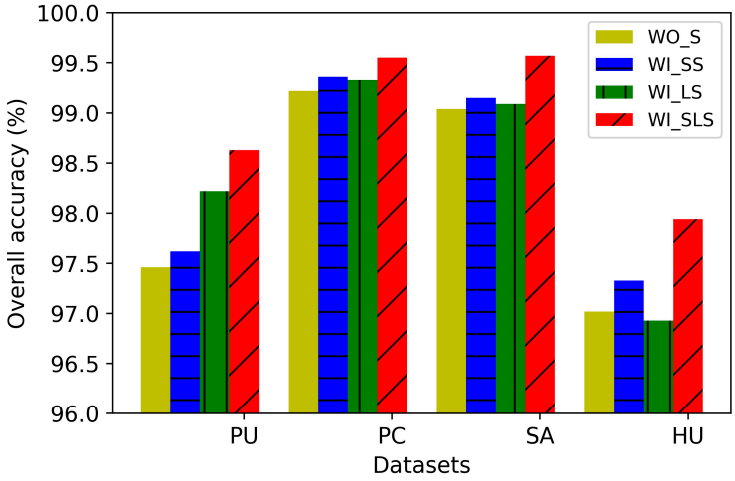

Figure 13. Effectiveness analysis of subpixel operations.

When the two inputs are processed with subpixel operations, the model performance is significantly improved compared with the classification performance without subpixel processing. The OAs on the four datasets increased by 0.77%, 0.17%, 0.43%, and 0.92%. HU has a slightly different phenomenon from the other three datasets; that is, after subjecting the large-scale input to only a subpixel operation, the model classification performance on this dataset is even weaker than that without a subpixel operation. We believe that because the same category of objects in HU does not show a large distribution state, the large-scale input introduces more neighborhood information that is different from the center pixel to a certain extent. After subpixel operations, the amount of interference information increases. Therefore, the extracted features inaccurately represent different categories of objects, and the classification performance is relatively weakened. In general, performing subpixel operations on inputs at different scales at the same time is effective for improving the performance of the model.

3.4.3. Effectiveness Analysis of the Adaptive Attention Constraint Fusion Mechanism

The proposed adaptive attention constraint fusion method aims to solve the noise and feature loss problem caused by upsampling, downsampling, and convolution in the process of scale unaligned feature fusion. This section conducts performance tests for the fusion module. We compare the proposed method with (1) using the large-scale feature map to align with the small-scale feature map (L_D in Figure 14) and (2) aligning the samples on the small-scale feature map with the large-scale feature map (S_U in Figure 14); the results are shown in Figure 14.

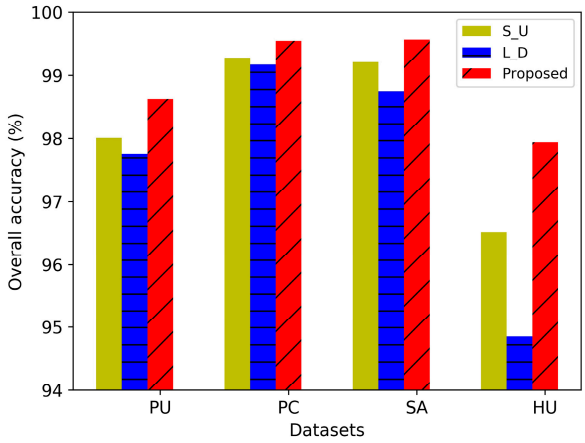

Figure 14. Effective analysis of fusion method.

As shown in Figure 14, the proposed attention fusion method can obviously improve the model performance. For the fusion of different scale feature maps, the method of upsampling small-scale feature maps to align with large-scale feature maps obtains better classification accuracy than downsampling large-scale feature maps to align with small-scale feature maps. Compared with part of the noise caused by upsampling, high-dimensional semantic features are more sensitive to the loss of features in the downsampling process. Compared with (1), the classification accuracies of the proposed attention fusion module improved by 0.77%, 0.24%, 0.62%, and 3.49% on the four datasets. Compared with (2), 0.59%, 0.15%, 0.26%, and 1.12% performance improvements are obtained. This module can be easily applied to deep neural networks where different spatial scale feature maps need to be fused.

3.4.4. Analysis of the Effectiveness of the Semantic Feature Enhancement Mechanism

To explore the ability of high-dimensional semantic features to represent data, a semantic feature enhancement module is constructed in this paper. This module improves the feature discrimination ability and generalization ability by enhancing feature representation, which leads to better classification performance. Figure 15 shows the performance comparison of the models with and without feature enhancement, where WO_E is the method without the enhancement module and WI_E is the method with the enhancement module.

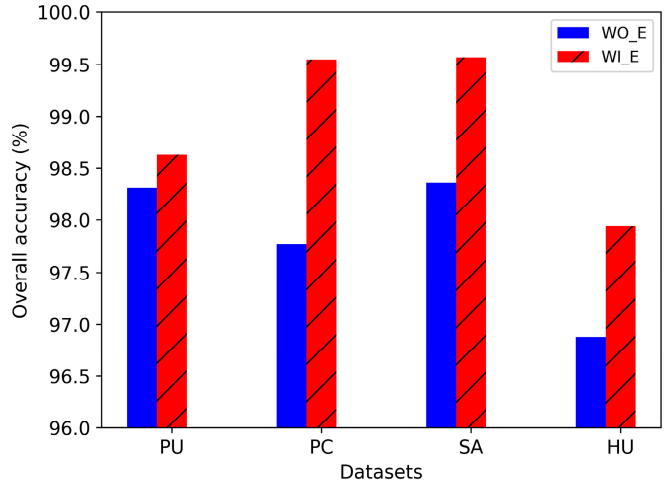

Figure 15. Effectiveness of feature enhancement module.

As shown in Figure 15, the performance of the model improves significantly after using the semantic feature enhancement module. The classification accuracy was improved by 0.32%, 1.78%, 1.21%, and 1.06% on the four datasets, with an average of 1.09%. This shows

that this module further enriches the features and enhances the ability to discriminate between features.

3.4.5. Training and Testing Times of Compared Methods Based on Attention Mechanism

A good HSI classification method needs not only high classification accuracy, but also good computational efficiency. We compare DBMA and MCNN, which use the attention mechanism, as shown in Table 9:

Table 9. Comparison of training and testing times of different attention-based methods on four datasets.

		PU	PC	SA	HU
DBMA	Train (s)	104.93	66.29	260.34	77.98
	Test (s)	36.99	43.88	55.85	5.88
MCNN	Train (s)	12.97	26.60	13.25	10.26
	Test (s)	2.56	8.55	3.25	1.32
Proposed	Train (s)	92.81	72.23	161.27	66.49
	Test (s)	36.84	31.71	40.06	6.33

As shown in Table 9, MCNN achieves the best computational efficiency among the three comparative methods. The computational efficiency of the proposed method is better than that of DBMA as a whole, and only slightly weaker than that of DBMA in the training time of PC and the testing time of HU.

MCNN has a simple structure, few network layers, and a low number of training parameters, and it achieves better computational efficiency. However, its simple model cannot fully mine image features, leading to its lower classification accuracy. Although the training and testing times of our proposed method is slower than that of MCNN, it is better than DBMA, and the classification accuracy is significantly higher than that of DBMA and MCNN.

4. Conclusions

This paper focuses on the classification of complex objects in small areas and proposes three main components: a sub-pixel multilevel scale feature learning module (SMSFL), an adaptive attention constraint fusion module (AACF), and a high-level feature semantic enhancement module (HFSE). SMSFL collects multiscale input, subpixel, and multiscale convolutions for richer feature extraction. AACF aims to reduce the noise and information loss problems introduced in the scaling process for multiscale feature fusion. HFSE is designed to enhance the feature representation and semantic information of high-level semantic features. Experimental results show that our approach obtains state-of-the-art performance on four HSI datasets.

In our proposed multiscale input module, different scale information needs to be set manually. In addition, although the proposed method is better than the comparison method in OA, AA, and KA of the four datasets, the classification accuracy of bare soil categories in PU is slightly lower. In the future, we will use adaptive scale selection to obtain the most suitable input scale combination for the model in order to achieve better classification performance.

Author Contributions: Conceptualization, Z.G.; methodology, Z.G.; validation, Z.G.; formal analysis, Z.G.; writing—original draft preparation, Z.G.; writing—review and editing, G.C., Y.Z., H.S., Y.L. and A.S.; funding acquisition, G.C., P.F. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61801222, in part by the Nature Science Foundation of Jiangsu Province under Grant BK20191284, in part by the Start Foundation of Nanjing University of Posts and Telecommunications

(NUPTSF) under Grant NY220157, and in part by Natural Science Research Project of Colleges and Universities of Jiangsu Province under Grant 22KJB510037.

Data Availability Statement: Publicly available datasets are analyzed in this study, which can be found here: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 27 February 2022) and https://hyperspectral.ee.uh.edu/?page_id=459 (accessed on 27 February 2022).

Acknowledgments: The authors would like to thank the editors and reviewers for their insightful suggestions.

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

References

1. Ge, Z.; Cao, G.; Zhang, Y.; Li, X.; Shi, H.; Fu, P. Adaptive Hash Attention and Lower Triangular Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5509119. [CrossRef]
2. Van der Meer, F.D.; Van der Werff, H.M.; Van Ruitenbeek, F.J.; Hecker, C.A.; Bakker, W.H.; Noomen, M.F.; Van der Meijde, M.; Carranza, E.J.; Boudewijn de Smeth, J.; Woldai, T. Multi-and hyperspectral geologic remote sensing: A review. *Int. J. Appl. Earth Observ. Geoinf.* **2022**, *14*, 112–128. [CrossRef]
3. Acosta, I.C.; Khodadadzadeh, M.; Tusa, L.; Ghamisi, P.; Gloaguen, R. A machine learning framework for drill-core mineral mapping using hyperspectral and high-resolution mineralogical data fusion. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 4829–4842. [CrossRef]
4. Gerhards, M.; Schlerf, M.; Mallick, K.; Udelhoven, T. Challenges and future perspectives of multi-/Hyperspectral thermal infrared remote sensing for crop water-stress detection: A review. *Remote Sens.* **2019**, *11*, 1240. [CrossRef]
5. Zhang, J.; Huang, Y.; Reddy, K.N. Assessing crop damage from dicamba on non-dicamba-tolerant soybean by hyperspectral imaging through machine learning. *Pest Manag. Sci.* **2019**, *75*, 3260–3272. [CrossRef]
6. Bioucas-Dias, J.M.; Plaza, A.; Capms-valls, G.; Scheunders, P.; Nasrebidi, N.; Chanussot, G. Hyperspectral remote sensing data analysis and future challenges. *IEEE Trans. Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
7. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
8. Kang, X.; Zhang, X.; Li, S.; Li, K.; Li, J.; Benediktsson, J.A. Hyperspectral anomaly detection with attribute and edge-preserving filters. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5600–5611. [CrossRef]
9. Tao, R.; Zhao, X.; Li, W.; Li, H.; Du, Q. Hyperspectral anomaly detection by fractional Fourier entropy. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 4920–4929. [CrossRef]
10. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
11. Shao, Y.; Lan, J.; Niu, B. Dual-channel networks with optimal-band selection strategy for arbitrary cropped hyperspectral images classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 550805. [CrossRef]
12. Lin, Y.; Guo, D.; Zhang, J.; Chen, Z.; Yang, B. A Unified Framework for Multilingual Speech Recognition in Air Traffic Control Systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3608–3620. [CrossRef] [PubMed]
13. Liu, Z.; Wu, Z.; Li, T.; Shen, C. GMM and CNN hybrid method for short utterance speaker recognition. *IEEE Trans. Ind. Inform.* **2018**, *14*, 244–3252. [CrossRef]
14. Lee, S.; Kim, H.; Lieu, Q.X.; Lee, J. CNN-based image recognition for topology optimization. *Knowl.-Based Syst.* **2020**, *198*, 105887. [CrossRef]
15. Li, J.; Jin, K.; Zhou, D.; Kubota, N.; Ju, Z. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **2020**, *411*, 340–350. [CrossRef]
16. You, H.; Tian, S.; Yu, L.; Lv, Y. Pixel-level remote sensing image recognition based on bidirectional word vectors. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 1281–1293. [CrossRef]
17. Kim, W.; Kanezaki, A.; Tanaka, M. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Trans. Image Process.* **2020**, *29*, 8055–8068. [CrossRef]
18. Sultana, F.; Sufian, A.; Dutta, P. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowl.-Based Syst.* **2020**, *201*, 106062. [CrossRef]
19. Xu, H.; Yao, W.; Cheng, L.; Li, B. Multiple Spectral Resolution 3D Convolutional Neural Network for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 1248. [CrossRef]
20. Yu, S.; Jia, S.; Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **2017**, *219*, 88–98. [CrossRef]
21. Zhao, G.; Liu, G.; Fang, L.; Tu, B.; Ghamisi, P. Multiple convolutional layers fusion framework for hyperspectral image classification. *Neurocomputing* **2019**, *339*, 149–160. [CrossRef]
22. Zheng, J.; Feng, Y.; Bai, C.; Zhang, J. Hyperspectral image classification using mixed convolutions and covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 522–534. [CrossRef]

23. Li, L.; Yin, J.; Jia, X.; Li, S.; Han, B. Joint Spatial–Spectral Attention Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1816–1820. [CrossRef]
24. Huang, K.; Ren, C.; Liu, H. Hyperspectral image classification via discriminative convolutional neural network with an improved triplet loss. *Pattern Recognit.* **2021**, *112*, 107744. [CrossRef]
25. Gao, F.; Guo, W. Deep hybrid dilated residual networks for hyperspectral image classification. *Neurocomputing* **2020**, *384*, 170–181. [CrossRef]
26. Zhao, J.; Hu, L.; Dong, Y.; Huang, L.; Weng, S.; Zhang, D. A combination method of stacked autoencoder and 3D deep residual network for hyperspectral image classification. *Int. J. Appl. Earth Observ. Geoinf.* **2021**, *102*, 102459. [CrossRef]
27. Huang, L.; Chen, Y. Dual-path siamese CNN for hyperspectral image classification with limited training samples. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 518–522. [CrossRef]
28. Cai, Y.; Zhang, Z.; Yan, Q.; Zhang, D.; Banu, M. Densely connected convolutional extreme learning machine for hyperspectral image classification. *Neurocomputing* **2021**, *434*, 21–32. [CrossRef]
29. Gao, H.; Chen, Z.; Xu, F. Adaptive spectral-spatial feature fusion network for hyperspectral image classification using limited training samples. *Int. J. Appl. Earth Observ. Geoinf.* **2022**, *107*, 102687. [CrossRef]
30. Wang, W.; Dou, S.; Jiang, Z.; Sun, L. A fast dense spectral–spatial convolution network framework for hyperspectral images classification. *Remote Sens.* **2018**, *10*, 1068. [CrossRef]
31. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaduri, B.B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 277–281. [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 Trans. Geosci. Remote Sens.* **2018**, *56*, 847–858. [CrossRef]
33. Yang, L.; Zhang, F.; Wang, P.; Li, X.; Meng, Z. Multi-scale spatial-spectral fusion based on multi-input fusion calculation and coordinate attention for hyperspectral image classification. *Pattern Recognit.* **2022**, *122*, 108348. [CrossRef]
34. Pu, C.; Huang, H.; Yang, L. An attention-driven convolutional neural network-based multi-level spectral–spatial feature learning for hyperspectral image classification. *Expert Syst. Appl.* **2021**, *185*, 115663. [CrossRef]
35. Zhang, T.; Wang, J.; Zhang, E.; Yu, K.; Zhang, Y.; Peng, J. RMCNet: Random Multiscale Convolutional Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1826–1830. [CrossRef]
36. Gao, H.; Yang, Y.; Li, C.; Gao, L.; Zhang, B. Multiscale residual network with mixed depthwise convolution for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3396–3408. [CrossRef]
37. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-based adaptive spectral-spatial kernel resnet for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [CrossRef]
38. Jia, S.; Lin, Z.; Xu, M.; Huang, Q.; Zhou, J.; Jia, X.; Li, Q. A lightweight convolutional neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4150–4163. [CrossRef]
39. Xie, J.; He, N.; Fang, L.; Chamisi, P. Multiscale densely-connected fusion networks for hyperspectral images classification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 246–259. [CrossRef]
40. Zhang, M.; Li, W.; Du, Q. Diverse region-based CNN for hyperspectral image classification. *IEEE Trans. Image Process.* **2018**, *27*, 2623–2634. [CrossRef]
41. Lee, H.; Kwon, H. Going Deeper With Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
42. Cheng, S.; Chi-Man, P. Multi-scale hierarchical recurrent neural networks for hyperspectral image classification. *Neurocomputing* **2018**, *294*, 82–93. [CrossRef]
43. Sun, G.; Zhang, X.; Jia, X.; Ren, J.; Zhang, A.; Yao, Y.; Zhao, H. Deep Fusion of Localized Spectral Features and Multi-scale Spatial Features for Effective Classifification of Hyperspectral Images. *Int. J. Appl. Earth Observ. Geoinf.* **2020**, *91*, 102157. [CrossRef]
44. Wang, D.; Du, B.; Zhang, L.; Xu, Y. Adaptive Spectral–Spatial Multiscale Contextual Feature Extraction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 2461–2477. [CrossRef]
45. 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]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
47. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inform. Process. Syst. Netw. (NIPS)* **2015**, *28*, 2017–2025.
48. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164. [CrossRef]
49. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [CrossRef]
50. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]

51. Huang, Z.; Wang, X.; Huang, L.; Shi, H.; Liu, W.; Huang, T. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 603–612. [CrossRef]
52. Feng, J.; Wu, X.; Shang, R.; Sui, C.; Li, J.; Jiao, L. Zhang, X. Attention multibranch convolutional neural network for hyperspectral image classification based on adaptive region search. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5054–5070. [CrossRef]
53. Xue, Z.; Zhang, M.; Liu, Y.; Du, P. Attention-Based Second-Order Pooling Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9600–9615. [CrossRef]
54. Gao, H.; Miao, Y.; Cao, X.; Li, C. Densely Connected Multiscale Attention Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 2563–2576. [CrossRef]
55. Guo, W.; Ye, H.; Cao, F. Feature-Grouped Network With Spectral-Spatial Connected Attention for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5500413. [CrossRef]
56. Yu, C.; Han, R.; Song, M.; Liu, C.; Chang, C.I. Feedback Attention-Based Dense CNN for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5501916. [CrossRef]
57. Hang, R.; Li, Z.; Liu, Q.; Chamisi, P.; Bhattacharyya, S.S. Hyperspectral image classification with attention-aided CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2281–2293. [CrossRef]
58. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 449–462. [CrossRef]
59. Grupo de Inteligencia Computacional (GIC). Available online: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 27 February 2022).
60. 2013 IEEE GRSS Data Fusion Contest–Fusion of Hyperspectral and LiDAR Data. Available online: https://hyperspectral.ee.uh.edu/?page_id=459 (accessed on 27 February 2022).