remote sensing

MDPI

Article

CSA-MSO3DCNN: Multiscale Octave 3D CNN with Channel and Spatial Attention for Hyperspectral Image Classification

Qin Xu, Yong Xiao, Dongyue Wang and Bin Luo *

Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China; xuqin@ahu.edu.cn (Q.X.); xy@stu.ahu.edu.cn (Y.X.); e18201065@stu.ahu.edu.cn (D.W.)

* Correspondence: luobin@ahu.edu.cn

Received: 20 November 2019; Accepted: 31 December 2019; Published: 5 January 2020

Abstract: 3D convolutional neural networks (CNNs) have been demonstrated to be a powerful tool in hyperspectral images (HSIs) classification. However, using the conventional 3D CNNs to extract the spectral–spatial feature for HSIs results in too many parameters as HSIs have plenty of spatial redundancy. To address this issue, in this paper, we first design multiscale convolution to extract the contextual feature of different scales for HSIs and then propose to employ the octave 3D CNN which factorizes the mixed feature maps by their frequency to replace the normal 3D CNN in order to reduce the spatial redundancy and enlarge the receptive field. To further explore the discriminative features, a channel attention module and a spatial attention module are adopted to optimize the feature maps and improve the classification performance. The experiments on four hyperspectral image data sets demonstrate that the proposed method outperforms other state-of-the-art deep learning methods.

Keywords: hypersctral image classification; octave convolution; feature extraction; channel and spatial attention

1. Introduction

Hyperspectral images (HSIs) are obtained by a series of hyperspectral imaging sensors and composed of hundreds of successive spectral bands. Because the wavelength interval between every two neighboring bands is quite small (usually 10 nm), HSIs generally have a very high spectral resolution [1]. Analysis of HSIs has been widely used in a large variety of fields, including materials analysis, precision agriculture, environmental monitoring and surveillance [2–4]. Among the hyperspectral community, the HSIs classification is most vibrant filed of research which is to assign a unique class to each pixel in the image [5]. However, due to the excessively redundant spectral band information and limited training samples, it also poses a great challenge to the classification of HSIs [6].

Early attempts for HSIs classification including the radial basis functions (RBFs) and K-nearest neighbor (kNN) methods are all pixel-wise and focus on the spectral signatures of hyperspectral data. But besides the spectral aspect, the spatial dependency which indicates the adjacent pixels likely belong to the same category is another useful information in the hyperspectral data. According to this, in aim to characterize the relationship between the samples, several spatial methods such as sparse representation and graph-based methods were proposed [7–10]. However, they used the class label to construct the manifold structure by both labeled and unlabeled data for classification which didn't incorporate the spectral feature. Consequently, a promising way is to combine the spectral and spatial information for classification that can enhance the performance of classification

due to taking full advantage of HSIs contained [11–13]. In [13], to cleanse the label noise the authors proposed a random label propagation algorithm (RLPA) which is guided by the spectral−spatial constraint-based knowledge. The RLPA algorithm constitutes two steps: spectral–spatial probability transform matrix (SSPTM) generation and random label propagation. However, the feature extraction of all the above-mentioned methods is hand-crafted which is not enough for the large intra-class difference or subtle inter-class difference.

Recently, deep learning (DL) has witnessed a great surge of interest that minds the deep feature and extracts the high-level feature of big data automatically in computer vision and big data field. Among the various DL models, the deep convolutional neural network (CNN) has become an efficient and popular tool for image recognition [14–18]. For the HSIs classification task, a series of CNN methods have been exploited [19–22]. In [19], a CNN architecture containing five layers was developed to classify hyperspectral images directly in spectral domain. In [20], a CNN network with a multi-layer perceptron was proposed to encode the pixels' spectral and spatial information to accomplish the classification task. In [21], Li et al. proposed a pixel-pair method to increase the training samples and used deep CNN to learn the pixel-pair features which are expected to have more discriminative power. However, these methods require a large amount of data as a training set and the over-fitting occurs easily that greatly decreases the classification accuracy when the labeled hyperspectral data is limited. In order to reduce the probability to over-fitting, a CNN based on diverse regions was proposed [22]. In this paper, on the one hand, the authors proposed a data enhancement method to obtain more training data. They cropped and filled the pixel patch according to diverse regions, and then flipped the original samples and added tiny Gaussian noise to the obtained training samples. On the other hand, inspired by the [18], they designed the network structure as a residual network, which has been proved to be effective to prevent the network from over-fitting.

To directly extract the spectral–spatial feature, some methods based on 3D CNN were proposed. Chen et al. [23] proposed a novel DL framework of 3D CNN to extract the spectral–spatial features effectively. In [24], Li et al. proposed a lightweight network based on 3D CNN, which required fewer parameters and performed better compared with 2D CNN method. Inspired by this, He et al. [25] provided a method based on 3D CNN with multi-scale convolution kernel, which could extract multiple sets of features by using convolution kernel with different sizes to enlarge receptive field. In [26], the authors not only effectively integrated spectral–spatial information through the use of 3D CNN but also employed residual network structure to alleviate the declining-accuracy phenomenon and facilitated the backpropagation of gradients. Inspired by [26], Wang et al. proposed a fast dense spectral–spatial convolution network for HSIs classification [27]. They used 3D convolution kernel of different sizes to extract more recognizable features and utilized dense network structure to prevent the proposed framework from over-fitting, which has been reported to be more effective than residual network structure [28]. In [29], a novel HSIs classification method was proposed, which combined the adaptive dimensionality reduction and semi-supervised 3D CNN. It can overcome the problem of high dimensionality curse of HSIs and limited training samples.

With the sustaining development of DL technology, some auxiliary technologies have been emerged. It is remarkable that attention mechanism as a representative has played an important role in many fields and interested numerous researchers [30–35]. The attention mechanism is based on the fact that humans focus attention selectively on parts of the visual space to acquire information when and where it is needed to. This is of significant interest for analyzing remotely sensed hyperspectral images. In [36], the author incorporated attention mechanisms to a ResNet to better characterize the spectral–spatial information contained in the data. In [37], a novel DL framework based on dense connectivity with spectral-wise attention for HSI classification was proposed. In this framework the dense connectivity is employed to prevent the network from over-fitting and a new spectral-wise attention is used to refine the features maps.

Although the DL methods for HSIs classification have achieved excellent results, which are better than the traditional approaches, there are still some problems.

- There is a lot of spatial redundancy in the hyperspectral data processing which takes up much memory space. Especially when 3D CNN is adopted to learn the feature, it will include numerous parameters which is disadvantage to the classification performance, compared to the 2D CNN and 1D CNN.
- Although the methods of combining DL method and attention mechanism have achieved successes for HSIs classification, to my best knowledge, there is not much research on the spatial attention for HSIs classification which does play an important role for HSIs classification.

In this paper, we propose a novel multiscale octave 3D CNN with channel and spatial attention (CSA-MSO3DCNN) for hyperspectral image classification. In our method, as 3D CNN can mine information hidden in the hyperspectral data more effectively whereas 1D CNN and 2D CNN can not, 3D CNN serves as the foundation of the entire architecture to directly extract the spectral–spatial features. In order to extract the spectral–spatial features of different scales, we design 3D CNN convolution kernels of different sizes. Due to 3D CNN has a lot of parameters and redundancy, we propose to use octave 3D CNN to replace the standard 3D CNN to decompose the features into high frequency and low frequency and reduce the spatial redundancy. Before feeding into the full connection layer, the channel and spatial attention mechanism modules are added to refine the feature maps. Through a series of optimization design, our method can extract higher and more recognizable features compared with the standard methods based on deep learning, like 2D CNN, 3D CNN, etc. Finally, the contributions of this paper can be summarized as follows:

1. The proposed network takes full advantages of octave 3D CNN with different kernels to capture diverse features and reduce the spatial redundancy simultaneously. Given the same input and structure, our proposed method works more effectively than the method based on normal 3D CNN.
2. A new attention mechanism with two attention modules is employed to refine the feature maps, which selects the discriminative features from the spectral and spatial views. This boosts the performance of our proposed network which further captures the similarity of adjacent pixels and the correlation of various spectral bands.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the CNN and the attention mechanism in DL. The detailed design of CSA-MSO3DCNN method is given in Section 3. In Section 4, we present and discuss the experimental results, including ablation experiments. Finally, Section 5 summarizes this paper.

2. Related Works

2.1. Convolutional Neural Network

Convolutional neural network (CNN) is a hierarchical structure composed of a deep stack of convolutional layers. It is because of this structure that CNN has a good capability of extracting the features of the visual data such as images and videos, which is very helpful for the subsequent operations. The mechanism of CNN is that it is based on the receptive fields and follows the behavior of neurons in the primary visual cortex of a biological brain [36]. In order to promote the efficiency of CNN, some improved convolutions, such as group convolution [14], separable convolution [38], depthwise convolution [39], and dilated convolution [40], have been proposed, which are mainly distinguished by the different ways of convolution.

At present, convolutional neural network has three different forms of convolution kernels that are 1D ($s_1 \times n$), 2D ($s_1 \times s_2 \times n$) and 3D ($s_1 \times s_2 \times s_3 \times n$). They have the same principle, specifically, they have the same element calculation and all adopt the back propagation algorithm to modify the parameters and train the network. For HSIs classification, the difference between the three forms is that they characterize different forms of feature, specifically, 1D CNN explores the spectral feature, 2D CNN explores the spatial feature, 3D CNN explores the spatial and spectral feature.

Due to the HSIs are originally 3-D and high dimensional, 3D CNN is more suitable for feature extraction and also used in our proposed network. To build a network for HSIs classification, only 3D convolution kernel is not enough. The activation function and some regularization measures are also needed. Therefore, for the input or intermediate feature map, the processing through a layer of the network can be described by the following formula,

$$x_l = H(F(x_{l-1}) + b) \tag{1}$$

where x_{l-1}, x_l and b are the input, output and corresponding bias respectively, $F(.)$ is the convolution operation and $H(.)$ is a subsequent processing function which can be batch normalization (BN) and rectified linear units (RELU). By stacking more and more layers, and adding the pooling layer and fully connected layer, a trainable network is established. In our proposed network, the same construction is employed.

2.2. Attention Mechanisms

As early as around the year 2000, studies have shown that attention mechanisms play an important role in human visual perception [41,42]. Subsequent to these, attention mechanisms have penetrated into various tasks in the field of information recognition, such as machine translation [43], object recognition [30], pose estimation [44], saliency detection [45]. In [43], the author proposed an architecture based on convolutional neural networks which used gated linear units to ease gradient propagation and equipped each decoder layer with a separate attention module. In [30], a recurrent neural network (RNN) model which is capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution was proposed.

In recent years, several attention mechanism networks of great significance have been developed. Hu et al. [35] proposed a squeeze-and-excitation network based on channel attention. It employed squeeze and excitation operations, which are composed of pooling and fully connecting, to assign different weights to different channels of the feature map to achieve the purpose of re-calibrating the feature map. Furthermore, it is a plug-and-play lightweight model that can be easily combined with classic DL models such as residual models and can be conveniently applied to various applications. To investigate attention from more aspects, in [46], Park et al. proposed an effective attention module called bottleneck attention module (BAM) from channel and spatial pathways separately which can be embedded in any feed-forward convolutional neural networks. Similar to this, to boost representation power of CNNs, Woo et al. [47] proposed an attention module along channel and spatial dimensions separately and plug it at every convolutional block, whereas Park et al. placed BAM module at every bottleneck of the network.

Inspired by [47], we propose a novel deep learning method combined with channel and spatial attention, which not only decreases the noise along the spectral bands but also explores the correlation between them. In the next section, our proposed method will be discussed in detail.

3. CSA-MSO3DCNN for Hyperspectral Images Classification

In this section, we first introduce the octave CNN and attention module, and then present the proposed network architecture, which is a novel multiscale octave 3D CNN with channel and spatial attention for HSIs classification (CSA-MSO3DCNN).

3.1. Octave Convolution

The main feature of the CNN is that the parameters of the network and the demand for memory will increase dramatically as the number of convolution layers increases, especially for the 3D CNN used in the field of HSIs classification. Recent proposed octave convolution [48] decomposes the feature map produced by CNN into high and low spatial frequency, which are updated separately and

exchanged with each other and finally merged together, to reduce the spatial redundancy and enlarge the receptive field.

Based on this, in this paper we propose to leverage three layers octave 3D convolution. Let $X_j = \{x_1, x_2, \ldots, x_i, \ldots, x_{n_{channels}} | x_i \in \mathbb{R}^{n_{bands} \times l_1 \times l_2}\}$ denote the $n_{channels}$ feature maps in the jth layer, where $l_1 \times l_2$ denotes the spatial dimensions and n_{bands} is the number of spectral bands. As shown in Figure 1, the first layer of the octave 3D convolution is to decompose the input feature maps into high frequency group X_1^H and low frequency group X_1^L along the channel dimension by a super-parameter α, which is the ratio of the low frequency group to the total. The channels of X_1^H and X_1^L are calculated as $(1-\alpha) \times c$ and $\alpha \times c$ respectively. Given the input feature maps X_1, the output of the first layer of 3D octave convolution is:

$$X_1^H = F(X_1) \tag{2}$$

$$X_1^L = F(Avg\_pool(X_1)) \tag{3}$$

where $F(.)$ is normal 3D convolution operation and Avg_pool is average pooling operation. In the middle layer, the high frequency group and low frequency group perform intra-feature update and inter-feature communication. The output of the middle layer is computed specifically as:

$$X_2^H = F(X_1^H) + Up(F(X_1^L)) \tag{4}$$

$$X_2^L = F(X_1^L) + F(Avg\_pool(X_1^H)) \tag{5}$$

where $Up(.)$ is up-sampling operation. Aiming at HSIs classification, in the last layer, the high and low frequency group are processed to obtain the same shape and then add up to decrease the feature redundancy, as illustrated in Figure 1. The output of the last layer Y is:

$$Y = F(Avg\_pool(X_2^H)) + F(X_2^L) \tag{6}$$

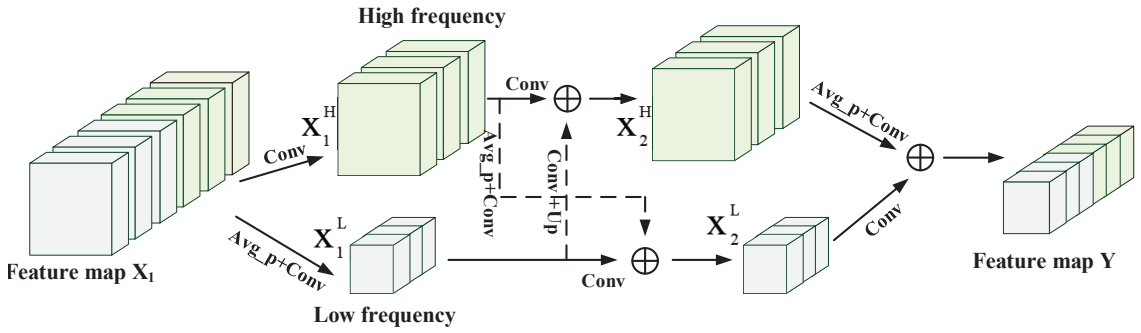


Figure 1. Three layers octave 3D convolution.

Therefore, in the octave 3D convolution, the group of spatial resolution of low frequency is reduced by sharing information between neighboring regions. It is believed that the octave 3D convolution has two distinct advantages, reducing spatial redundancy and enlarging the receptive field. Accordingly, we propose to use the octave 3D convolution to replace the traditional 3D convolution to improve the HSIs classification.

3.2. Channel and Spatial Attention

To boost the representation ability of our network, with consideration of the abundant spectral and spatial information that HSIs have, we propose to employ the channel attention mechanism which attempts to assign different significance to the channels of feature maps and the spatial attention mechanism which is in aim to find which portions are more important in a feature map. We adopt the convolutional block attention module proposed by Woo et al. that is general and end-to-end trainable

along with the basic CNNs [47]. The structures of the channel and spatial attention module are shown in Figure 2.

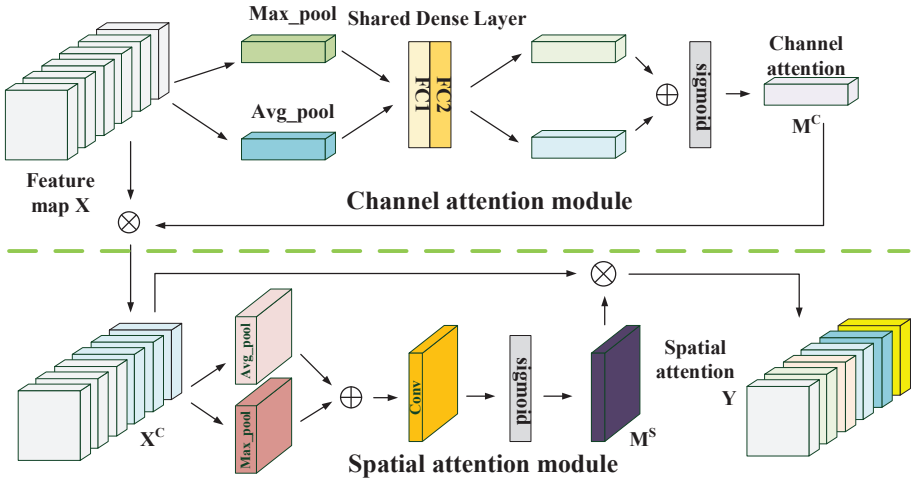


Figure 2. The detailed structure of channel and spatial attention module. (Top) Channel-wise attention module. (Bottom) Spatial-wise attention module.

Channel-wise attention is an attention mechanism which emphasizes reducing channel redundancy and building a channel attention map through capturing the inter-channel relationship of features [47]. As exhibited in the top of Figure 2, given an intermediate layer of feature maps $X = \{x_1, x_2, \ldots, x_i, \ldots, x_{n_{channels}} | x_i \in \mathbb{R}^{n_{bands} \times l_1 \times l_2}\}$, to squeeze and aggregate the feature average-pooling and max-pooling are performed simultaneously to generate two different feature maps: max-pooled features X^{max} and average-pooled features X^{avg}. Then, X^{max} is fed into a shared network which is composed of two dense layers to train. With the learned weight the X^{avg} is also fed into the shared network. As a result, the channel attention $M^C \in \mathbb{R}^{n_{channels} \times 1 \times 1 \times 1}$ is obtained. In addition, we adopt a reduction ratio r to reduce parameters [35] and the hidden activation size is set to $n_{channels}/r \times 1 \times 1 \times 1$. In summary, the channel attention is calculated as:

$$M^C = \sigma(FC(Max\_pool(X)) + FC(Avg\_pool(X))) \tag{7}$$

where σ denotes the sigmoid function and Max_pool is max pooling operation.

In order to further explore where to focus on in a channel of feature map, the spatial-wise attention mechanism is adopted which can be seen as a supplement to the channel-wise attention. As illustrated in the bottom of the Figure 2, the spatial attention module is connected behind the channel-wise attention module. The input of the spatial attention module X^C is the channel refined feature maps,

$$X^C = X \otimes M^C \tag{8}$$

where $\otimes$ denotes element-wise multiplication. For taking full advantage of channel information, global average-pooling and max-pooling operations are both applied to generate 3D feature maps X^{Cmax} and X^{Cavg}. Then these are concatenated and convolved by a standard convolution layer to generate the 3D spatial attention map. The spatial-wise attention is computed as:

$$M^S = \sigma(F^{3\times3\times3}([Max\_pool(X^C); Avg\_pool(X^C)])) \tag{9}$$

where $F^{3\times3\times3}$ denotes a normal 3D convolution with kernel size of $3 \times 3 \times 3$. At this point, the output feature map Y of the two attention module is

$$Y = X^C \otimes M^S \tag{10}$$

The obtained Y is the optimized feature map of X through the two sequential attention modules and can improve the classification performance.

3.3. Proposed Network Architecture

In this section, we design our CSA-MSO3DCNN architecture as follows. First, for the high-dimensional HSIs data, we apply principal component analysis (PCA) to reduce the dimension before the formal network training, which can decrease the parameters and keep the essential information as illustrated in Figure 3. Then, a pixel and surrounding pixels are selected from the processed HSIs data to form a 3D patch which has the shape of $s \times s \times d$, where $s \times s$ is the spatial dimension and d is the spectral bands. This is to explore the relationship between the neighbouring pixels, which has a large probability of belonging to the same category.

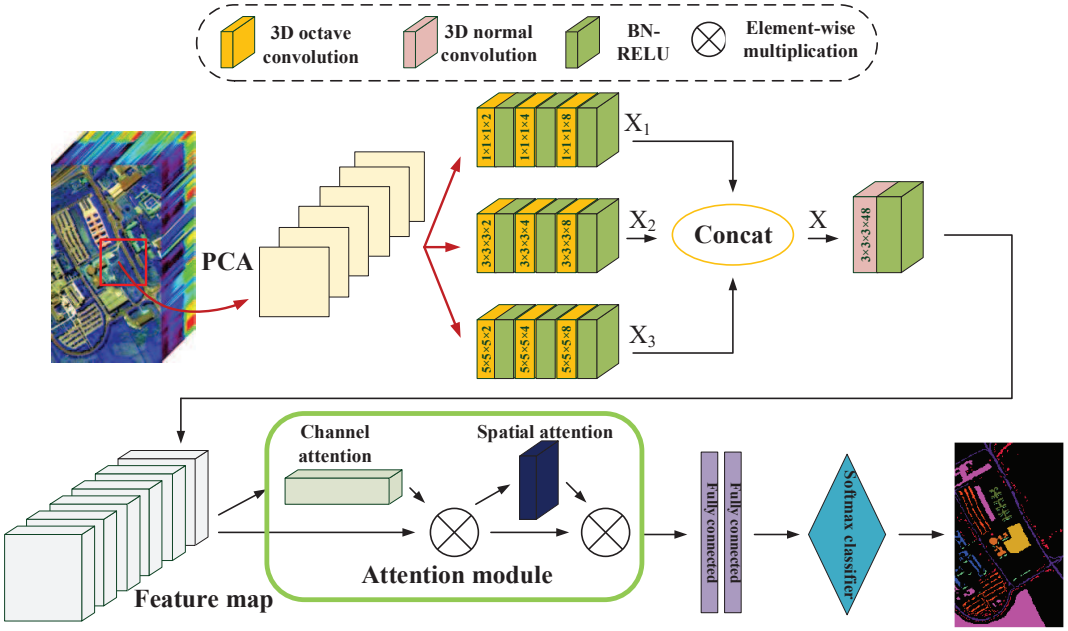


Figure 3. The overall flowchart of the proposed method. (Top) Using principal component analysis (PCA) and octave 3D CNN to extract features. (Bottom) Utilizing the channel and spatial attention module to refine features and finally to classify.

Secondly, the 3D patch is fed to three network branches, each of which is composed of three 3D octave convolution layers to extract multi-scale features. In the three network branches, each octave 3D convolution layer is followed by batch normalization-RELU(BN-RELU). We denote the outputs of the three branches as X_1, X_2, X_3. It is worth noting that the three branches differ in the size of the octave 3D convolution kernel, where the convolution kernel sizes of the three branch are $1 \times 1 \times 1$, $3 \times 3 \times 3$ and $5 \times 5 \times 5$ respectively. In each branch, each octave 3D convolution layer is designed with a different number of convolution kernels, as shown in Figure 3.

Thirdly, for the convenience of concatenation, we keep the size of the original data and the feature map consistent. Therefore, we concatenate the outputs of the three branches to get X:

$$X = G(X_1, X_2, X_3) \tag{11}$$

where $G(.)$ is concatenation operation. Then a normal 3D convolutional layer is used to abstract the feature map.

Fourthly, we employ an attention module with channel-wise attention and spatial-wise attention (see in Section 3.2) to refine the obtained feature maps, so that the feature maps can become more discriminative. Finally, two fully connected layers (FC) and a softmax classifier are used as a classifier.

Reasonably, we also use 'dropout' technology in the fully connected layer, which can effectively suppress over-fitting without adding a large number of parameters. In addition, the categorical cross entropy is adopted as a loss function:

$$E = -\sum_{k} t_k \log y_k \tag{12}$$

where t_k is the correct label and the y_k is the output of the network. In order to continuously reduce the loss and update network parameters, the Adam method is adopted.

In summary, a novel multi-scale octave 3D CNN based on channel and spatial attention for HSIs classification has been proposed. It is obvious that our approach can greatly reduce spatial redundancy and enlarge the receptive field, which are beneficial for improving the classification performance.

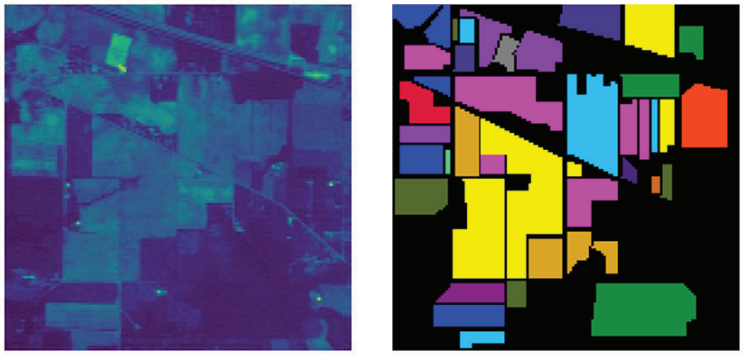
4. Experimental Results and Analysis

In this section, we evaluated the performance of our CSA-MSO3DCNN on four public HSI data sets for HSIs classification. Four popular indicators, class accuracy, overall accuracy (OA), average accuracy (AA), kappa coefficient (κ) are used to measure the pros and cons of our approach and the compared five state-of-the-art methods. All experiments are implemented with an NVIDIA 1060 GPU and a Titan graphics card server, Tensorflow-gpu and Keras with Python 3.6.

4.1. Experimental Data

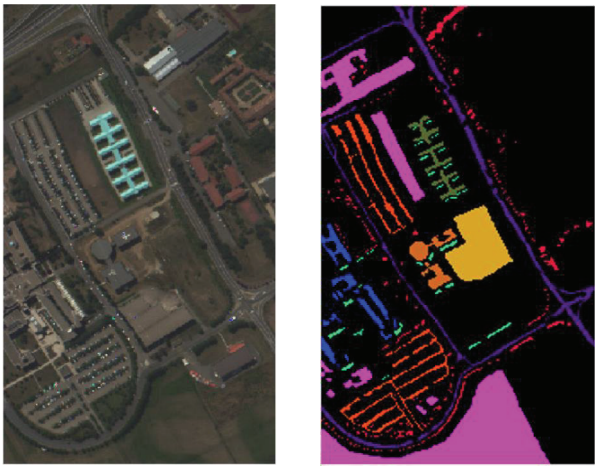
The experiments were conducted on four standard HSIs data sets, including two popular data sets and two contest data sets, that are, Indian Pines, University of Pavia, grss_dfc_2013 [49] and grss_dfc_2014 [50].

- **Indian Pines** Indian Pines is a very popular hyperspectral data set which has 16 different classes. It was obtained by airborne visible/infrared imaging spectrometer (AVIRIS) which contains 200 spectral bands after removing the noisy bands. The data set has a spatial dimension of 145 × 145 with 10,249 labeled pixels and covers the wavelengths between 0.4 to 2.5 µm with 20 m spatial resolution. Figure 4a,b are a false color image and the corresponding ground truth map.
- **University of Pavia** is over an urban area surrounding University of Pavia, Italy. It is collected by the reflective optics system imaging spectrometer (ROSIS) and has been widely used in HSIs classification. The data set has a spatial dimension of 610 × 340 and a spatial resolution of 1.3 m per pixel. It has 115 spectral bands ranging from 0.43 to 0.86 µm with 12 noisy bands. In experiments, the 12 noisy bands are removed. The false color and reference ground truth image are shown in Figure 5a,b, respectively.
- **Grss_dfc_2013** is a public HSI data set, which was released in the 2013 IEEE GRSS Data Fusion Contest, collected by NSF-funded Center for Airborne Laser Map- ping (NCALM), and acquired over the University of Houston campus and the surrounding area in 23 June 2012. It has a spatial dimension of 349 × 1905 with 2.5 m spatial resolution, in the range of 380 nm to 1050 nm, and has 144 spectral bands. In Figure 6a,b, a false color composite image and the ground truth map are displayed.
- **Grss_dfc_2014** is a coarser-resolution long-wave infrared (LWIR, thermal infrared) hyperspectral data set, which is more challenging and employed in 2014 IEEE GRSS Data Fusion Contest. It was acquired by an 84-channel imager that covered the wavelengths between 7.8 to 11.5 µm with approximately 1-m spatial resolution. The size of this data set is 795 × 564 pixels with 22532 labeled pixels and is classed into seven classes. Figure 7a,b give a false color image of Grss_dfc_2014 and the ground truth map.



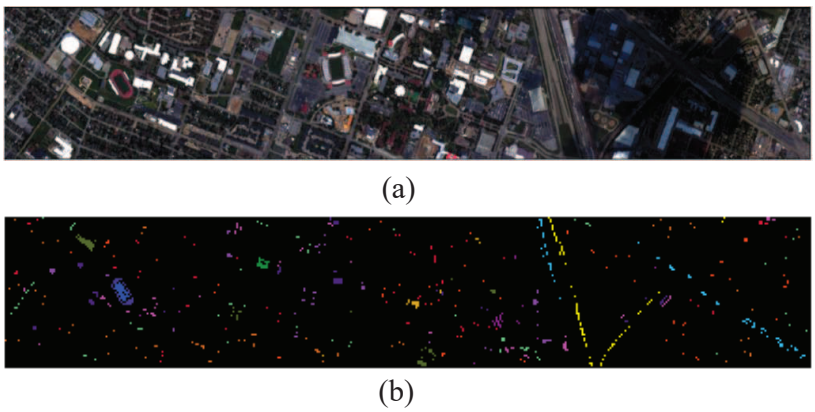
(a) (b)

Figure 4. The Indian Pines data set. (**a**) A false color map. (**b**) The ground truth map.



(a) (b)

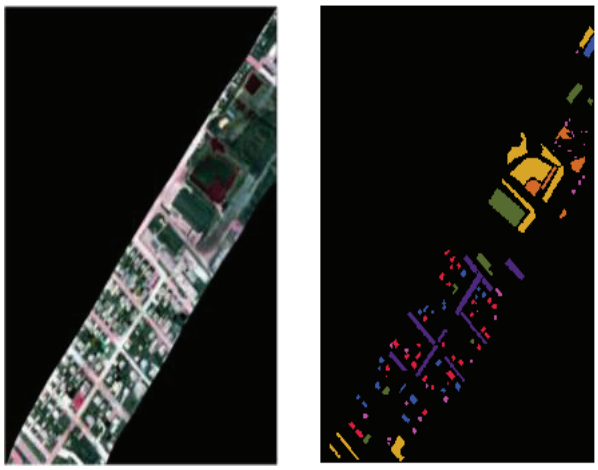
Figure 5. The University of Pavia data set. (**a**) A false color map. (**b**) The ground truth map.



(a)

(b)

Figure 6. The Grss_dfc_2013 data set. (**a**) A false color map. (**b**) The ground truth map.



(a) (b)

Figure 7. The Grss_dfc_2014 data set. (**a**) A false color map. (**b**) The ground truth map.

Tables 1–4 show the distribution of training and testing samples of the four experimental data sets. For the Indian Pines data set, as shown in Table 1, because the number of samples in different categories varies widely, for each class we randomly selected a half as the training samples if the number of samples was less than 600, and took 300 as the training set if the number of samples was more than 600 and the rest were set as the testing samples. For the University of Pavia and Grss_dfc_2014 data sets, 200 training samples were randomly selected from each class. The rest of the samples were taken for testing, as shown in Tables 2 and 4. It can be seen that several classes had a large number of test samples, such as 'meadows' and 'asphalt' in the University of Pavia and 'vegetation' in the Grss_dfc_2014, which increased the difficulty of classification. For the Grss_dfc_2013 data set, 200 training samples were randomly selected from each class, except the 'water' class. Because of the limited samples in class 'water', we randomly selected 162 samples as the training set. The rest samples were defined for testing, as shown in Table 3.

Table 1. The distribution of training and testing samples of the Indian Pines data set.

Label	Class Name	Train	Test
1	Alfalfa	23	23
2	Corn-notill	300	1128
3	Corn-min	300	530
4	Corn	118	119
5	Grass/Pasture	241	242
6	Grass/Trees	300	430
7	Grass/Pasture-mowed	14	14
8	Hay-windrowed	239	239
9	Oats	10	10
10	Soybeans-notill	300	672
11	Soybeans-min	300	2155
12	Soybeans-clean	296	297
13	Wheat	102	103
14	Woods	300	965
15	Building-Grass-Trees-Drives	193	193
16	Stone-steel Towers	46	47
-	Total	3082	7176

Table 2. The distribution of training and testing samples of the University of Pavia data set.

Label	Class Name	Train	Test
1	Asphalt	200	6431
2	Meadows	200	18,449
3	Gravel	200	1899
4	Trees	200	2864
5	Sheets	200	1145
6	Baresoil	200	4829
7	Bitumen	200	1130
8	Bricks	200	3482
9	Shadows	200	747
-	Total	1800	40,976

Table 3. The distribution of training and testing samples of the Grss_dfc_2013 data set.

Label	Class Name	Train	Test
1	Healthy grass	200	1051
2	Stressed grass	200	1054
3	Synthetic grass	200	497
4	Trees	200	1044
5	Soil	200	1042
6	Water	162	163
7	Residential	200	1068
8	Commercial	200	1044
9	Road	200	1052
10	Highway	200	1027
11	Railway	200	1035
12	Parking Lot 1	200	1033
13	Parking Lot 2	200	269
14	Tennis Court	200	228
15	Running Track	200	460
-	Total	2962	12,067

Table 4. The distribution of training and testing samples of the Grss_dfc_2014 data set.

Label	Class Name	Train	Test
1	Road	200	4243
2	Trees	200	893
3	Red roof	200	1654
4	Grey roof	200	1926
5	Concrete roof	200	3688
6	Vegetation	200	7157
7	Bare soil	200	1571
-	Total	1400	21,132

4.2. Experimental Setup

In all the experiments, the size of the 3D cube was set to $22 \times 22 \times 20$, where the '20' is the spectral bands after PCA dimension reduction. The processed data contained no less than 99.96% information of the original data. Because the last octave 3D convolution layer contains a half-size operation (Avg_pooling), the shape of the cube was reduced to $11 \times 11 \times 10$, which omitted the padding operation and reduced the noise in this step. In addition, the size of all 3D convolution operation is depicted in Figure 3, such as $3 \times 3 \times 3$ with padding 'SAME'. The setting of attention module is referenced to Figure 2.

The parameter α was set to 0.25 which was used to reduce the spatial redundancy. The learning rate was set to $1e-4$ to make sure the convergence speed. We used training steps to represent the number of iterations of the network, and each iteration was the parameter update of the whole network. If the number of training steps was too large, it would lead to over fitting. In all the experiments, the number of training steps was set to 1000 and the training batch was set to 128. The number of hidden layers and the regularization are shown in Figures 2 and 3. Each branch had three hidden layers, and each convolution layer was followed by a batch normalization for regularization.

4.3. Experimental Results and Discussion

To evaluate the effectiveness of the proposed method, we compared our method with five state-of-the-art methods. For fair comparison, for each data set the size of the training set adopted is same for all the methods. The compared methods are as follows:

- CNN [20]: A method exploits CNN to encode the spectral–spatial information of pixels and a MLP to conduct the classification task.
- M3DCNN [25]: A multiscale 3D CNN method for HSIs classification, which different branches have different sizes of 3D convolution kernel.
- SRN: [26] A spectral–spatial 3D deep learning network with residual structure, which effectively mitigates over-fitting.
- MSDN-SA: [37] A dense 3D CNN framework with spectral-wise attention mechanism.
- MSO3DCNN: Our proposed method without attention module.

In the experimental setup, we randomly chose the training and testing samples for the classification task. Because each random selection produced a different classification result, for each data set and each class we ran the experiment for 10 times to obtain the accuracy. We finally computed the average accuracy and the standard deviation for each class and compute the overall accuracy (OA), average accuracy (AA) and kappa coefficient (κ) for each data set.

4.3.1. Results for Indian Pines Data Set

The comparison results of the classification accuracy for the Indian Pines data set are presented in Table 5. From the Table 5, it can be seen that the proposed CSA-MSO3DCNN obtained the best OA, AA, and κ, which are 99.68%, 99.45%, and 99.62% respectively. Compared with the CNN method, OA, AA, and κ of our method have improved much more, where the AA has increased by about 8%. The classification results of SSRN and MSO3DCNN methods show that the spatial–spectral feature obtained by the octave 3D CNN was better than the feature obtained by the normal 3D CNN. Furthermore, the proposed method was better than MSO3CNN method, which proves the positive effect of the channel and spatial attention module.

To show the visual classification results, the classification maps and normalized confusion matrices are shown in Figure 8 and Figure 9 respectively. From Figure 8, the classification map produced by CSA-MSO3DCNN method is closest to the ground truth map, which means that CSA-MSO3DCNN method obtains the best classification result. The diagonal values of the matrix in Figure 9 also prove this, where the row represents prediction value and the column represents actual value.

4.3.2. Results for The University of Pavia Data Set

The comparison results of the classification accuracy and classification maps for the University of Pavia data set are reported in Table 6 and Figure 10. From Table 6 we can see that the proposed CSA-MSO3DCNN achieved the best OA, AA, and κ, which are 99.76%, 99.66% and 99.67% respectively. The classification results of each class obtained by the CSA-MSO3DCNN were much better than others such as the first three categories 'Asphalt', 'Meadows' and 'Gravel'. Compared with the MSO3DCNN method (without attention module), the OA, AA, and κ of CSA-MSO3DCNN all improved, which indicates the effectiveness of the attention module. The results also demonstrate the superiority of

octave 3D CNN which replaces the normal 3D CNN in the CSA-MSO3DCNN method. Furthermore, from the results we can conclude that methods based on 3D CNN were significantly better than the methods based on CNN. Although the MSDN-SA method provided the second best results, from the standard deviation, our proposed CSA-MSO3DCNN performed more stably than MSDN-SA. For the visual results, from Figure 10, the classification maps obtained by the proposed method were closest to the ground truth map. For example, the center yellow area, which represents the class 'Baresoil', in our feature map, was all yellow without other color, which means there were no misclassified pixels. The normalized confusion matrices of classification results are depicted in Figure 11. From Figure 11a, there are several colors of diamonds in the confusion matrix obtained by the CNN method, which means some pixels were misclassified into other categories. From Figure 11a–f, it can be seen that the classification accuracy of each class obtained by the CSA-MSO3DCNN was higher than others.

Table 5. Classification accuracy for the Indian Pines data set.

Class	CNN	M3DCNN	SSRN	MSDN-SA	MSO3DCNN	CSA-MSO3DCNN
1	50.72 ± 10.55	98.55 ± 2.51	95.65 ± 4.35	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
2	87.85 ± 6.24	98.32 ± 1.19	99.02 ± 0.54	98.38 ± 0.64	99.56 ± 0.15	98.97 ± 0.80
3	94.53 ± 3.71	99.87 ± 0.11	99.75 ± 0.22	100.00 ± 0.00	100.00 ± 0.00	99.87 ± 0.11
4	96.08 ± 5.40	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
5	99.69 ± 2.71	100.00 ± 0.00	99.17 ± 0.41	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
6	99.67 ± 0.36	99.92 ± 0.13	100.00 ± 0.00	99.92 ± 0.13	100.00 ± 0.00	99.92 ± 0.13
7	30.95 ± 27.56	97.62 ± 4.12	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
8	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
9	70.00 ± 36.10	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
10	90.28 ± 2.67	99.65 ± 0.48	99.50 ± 0.23	99.80 ± 0.23	99.90 ± 0.17	99.85 ± 0.15
11	91.13 ± 5.35	98.36 ± 0.72	98.55 ± 0.26	98.70 ± 0.52	98.79 ± 0.31	99.14 ± 0.53
12	96.41 ± 3.60	99.21 ± 0.85	99.33 ± 0.34	98.88 ± 1.36	99.66 ± 0.34	99.66 ± 0.33
13	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
14	97.55 ± 1.20	99.79 ± 0.18	99.83 ± 0.16	99.55 ± 0.47	99.97 ± 0.06	100.00 ± 0.00
15	97.41 ± 2.88	99.83 ± 0.30	100.00 ± 0.00	99.83 ± 0.30	100.00 ± 0.00	99.82 ± 0.30
16	100.00 ± 00.00	99.29 ± 1.23	99.29 ± 1.23	99.29 ± 1.23	98.58 ± 2.46	100.00 ± 0.00
OA(%)	93.03 ± 1.68	99.12 ± 0.24	99.32 ± 0.06	99.21 ± 0.09	99.53 ± 0.07	**99.68** ± 0.06
AA(%)	91.75 ± 1.62	99.06 ± 0.09	99.24 ± 0.11	98.72 ± 0.55	99.41 ± 0.23	**99.45** ± 0.54
$\kappa \times 100$	98.95 ± 0.28	99.01 ± 0.31	99.27 ± 0.07	99.07 ± 0.10	99.44 ± 0.08	**99.62** ± 0.07

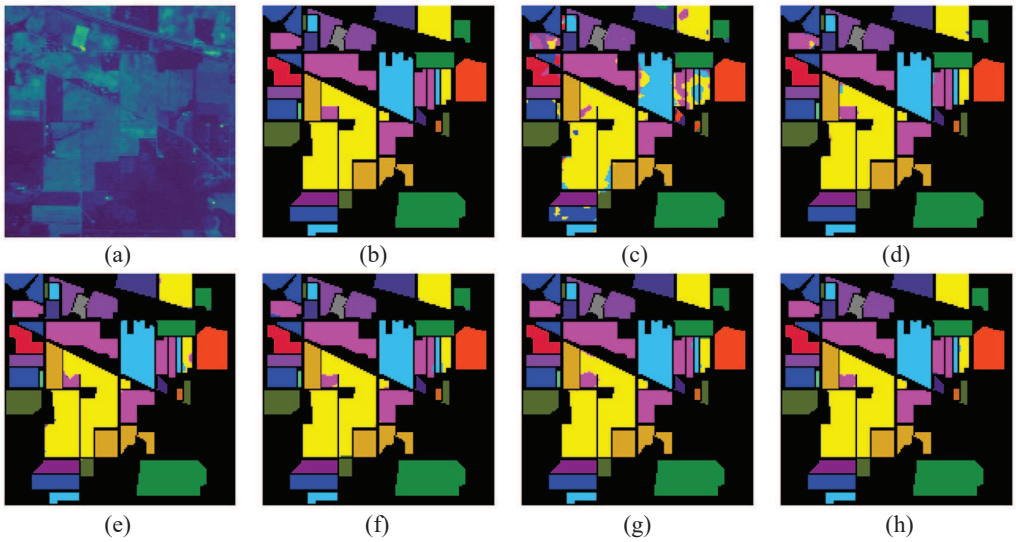


Figure 8. Classification maps provided for the Indian Pines data set by different methods. (**a**) A false color map. (**b**) The ground truth map. (**c**) CNN (93.03%). (**d**) M3DCNN (99.12%). (**e**) SSRN (99.32%). (**f**) MSDN-SA (99.21%). (**g**) MSO3DCNN (99.53%). (**h**) CSA-MSO3DCNN (99.68%).

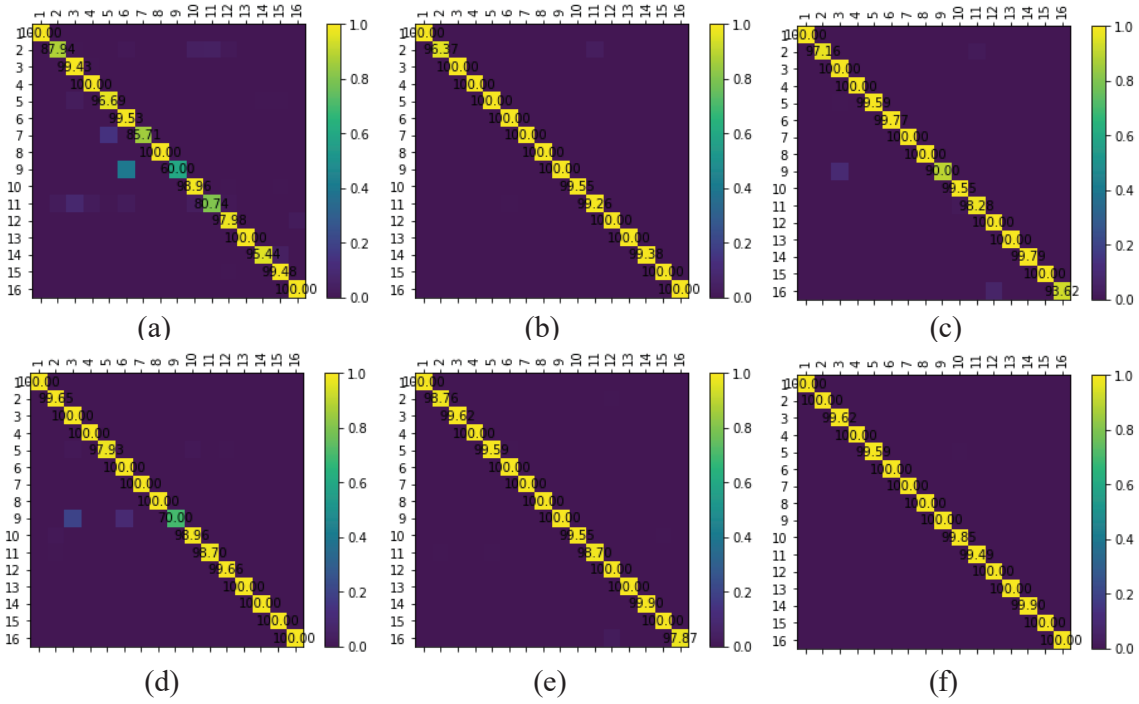


Figure 9. Normalized confusion matrices of classification results for the Indian Pines data set. (**a**) CNN. (**b**) M3DCNN. (**c**) SSRN. (**d**) MSDN-SA. (**e**) MSO3DCNN. (**f**) CSA-MSO3DCNN.

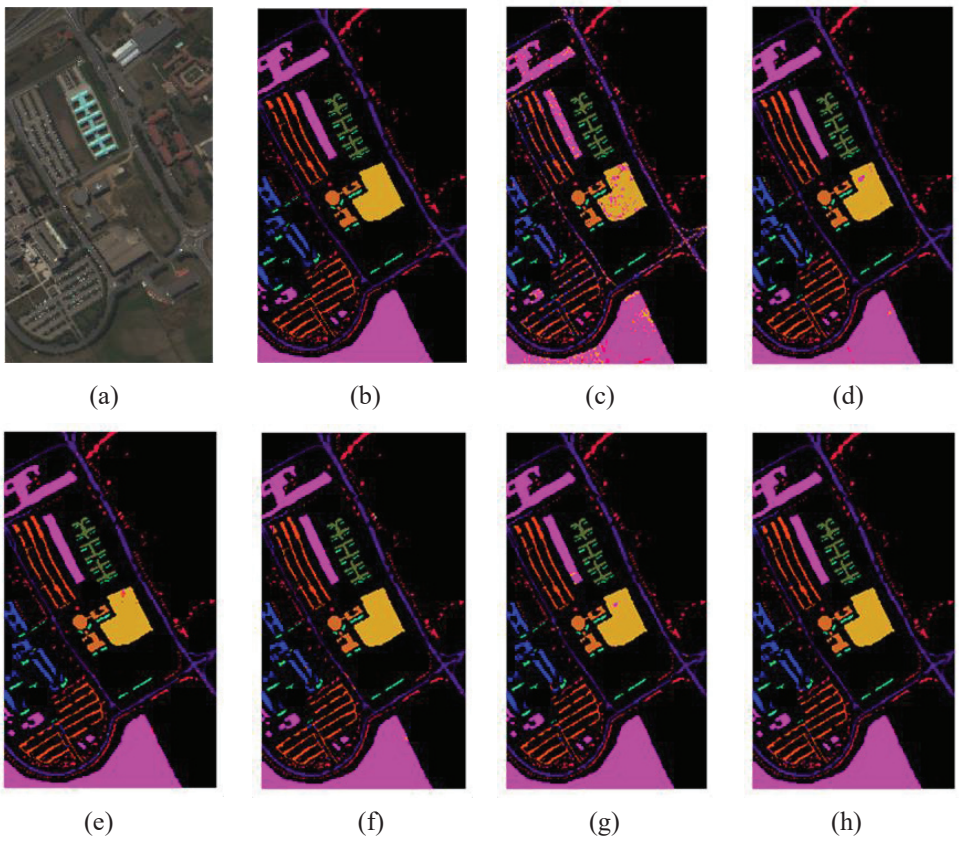


Figure 10. Classification maps provided for the University of Pavia data set by different methods. (**a**) A false color map. (**b**) The ground truth map. (**c**) CNN (85.89%). (**d**) M3DCNN (99.03%). (**e**) SSRN (99.19%). (**f**) MSDN-SA (99.31%). (**g**) MSO3DCNN (99.54%). (**h**) CSA-MSO3DCNN (99.76%).

Table 6. Classification accuracy for the University of Pavia data set.

Class	CNN	M3DCNN	SSRN	MSDN-SA	MSO3DCNN	CSA-MSO3DCNN
1	76.85 ± 0.32	99.15 ± 0.41	98.91 ± 0.40	99.15 ± 0.13	99.48 ± 0.34	99.89 ± 0.06
2	86.52 ± 0.19	99.35 ± 0.16	99.60 ± 0.11	99.52 ± 0.14	99.69 ± 0.11	99.83 ± 0.11
3	87.41 ± 0.34	96.84 ± 0.57	96.96 ± 0.89	97.96 ± 0.36	99.53 ± 0.15	99.77 ± 0.11
4	95.43 ± 0.45	99.56 ± 0.31	99.37 ± 0.01	98.77 ± 0.41	98.65 ± 0.18	98.78 ± 0.27
5	99.83 ± 0.21	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.91 ± 0.12	99.94 ± 0.08
6	89.09 ± 0.45	98.94 ± 0.75	99.30 ± 0.54	99.83 ± 0.12	100.00 ± 0.00	99.98 ± 0.03
7	96.64 ± 0.12	100.00 ± 0.00	99.94 ± 0.08	99.91 ± 0.07	99.76 ± 0.03	99.85 ± 0.21
8	75.10 ± 0.23	97.16 ± 0.08	97.84 ± 0.99	98.37 ± 1.01	98.62 ± 0.55	99.11 ± 0.68
9	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.0 ± 0.00	99.69 ± 0.23
OA(%)	85.89 ± 0.09	99.03 ± 0.01	99.19 ± 0.13	99.31 ± 0.06	99.54 ± 0.05	**99.76** ± 0.02
AA(%)	82.33 ± 0.34	98.51 ± 0.22	98.66 ± 0.11	98.90 ± 0.08	99.32 ± 0.12	**99.66** ± 0.01
$\kappa \times 100$	81.52 ± 0.12	98.70 ± 0.20	99.02 ± 0.05	99.08 ± 0.06	99.38 ± 0.08	**99.67** ± 0.02

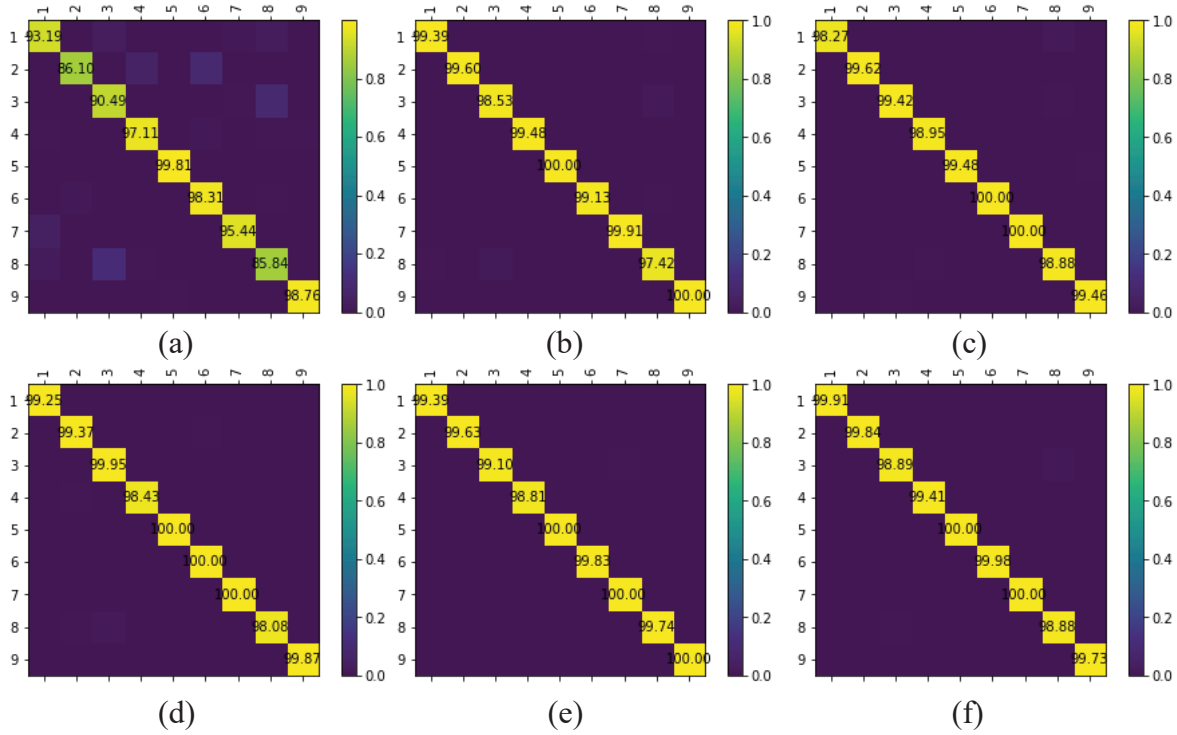


Figure 11. Normalized confusion matrices of classification results for the University of Pavia data set. (**a**) CNN. (**b**) M3DCNN. (**c**) SSRN. (**d**) MSDN-SA. (**e**) MSO3DCNN. (**f**) CSA-MSO3DCNN.

4.3.3. Results for the Grss_dfc_2013 Data Set

The classification accuracy, the classification maps and the normalized confusion matrices of classification results of all methods for Grss_dfc_2013 data set are listed in Table 7, in Figures 12 and 13 respectively. It can be seen that the proposed CSA-MSO3DCNN method achieved the best OA, AA, and κ, which were 99.69%, 99.72%, and 99.66% respectively from the Table 7. The classification result of each class was 99.09% at least, and our obtained lowest accuracies were all higher than lowest accuracies obtained by other methods.

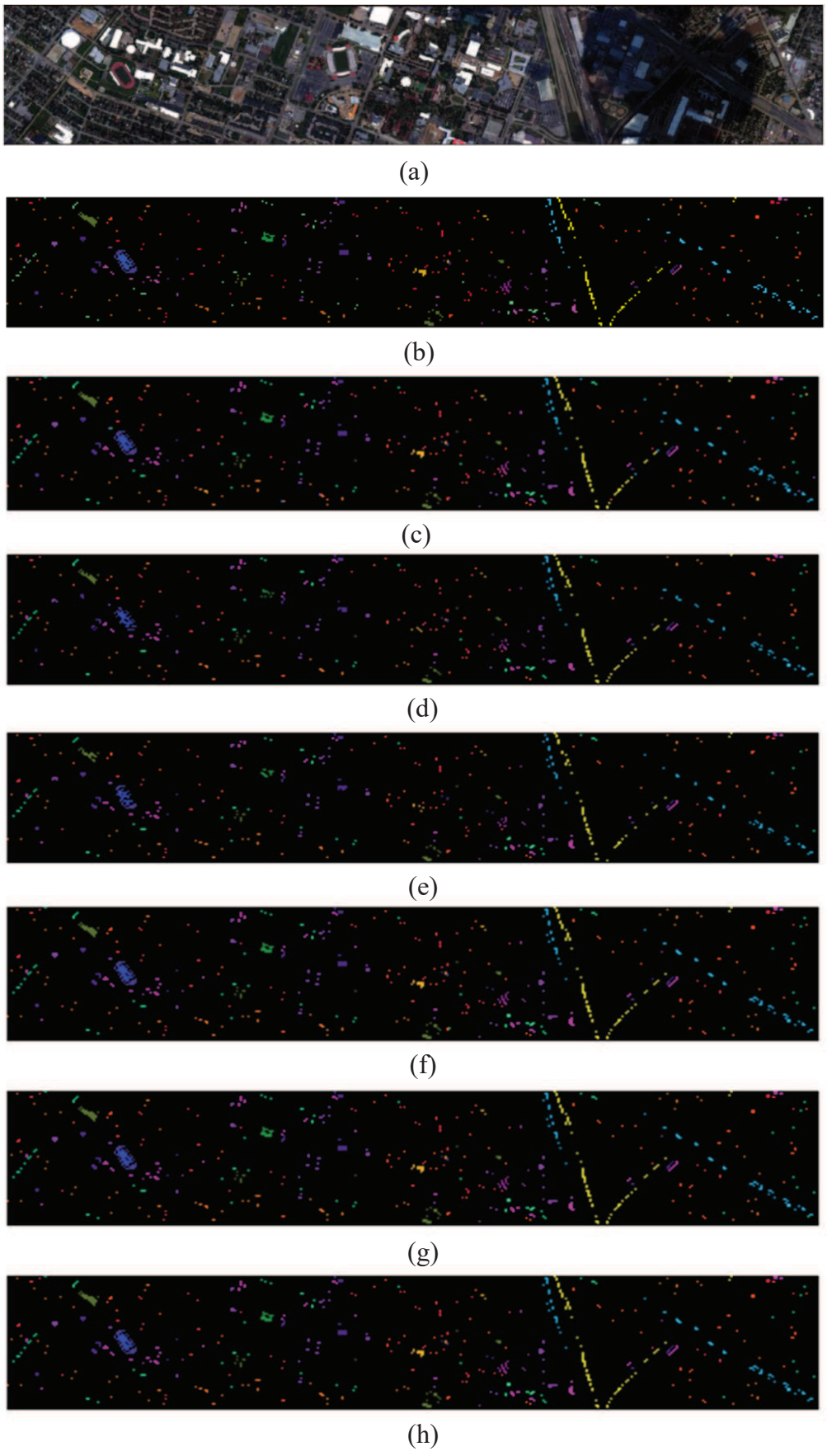


Figure 12. Classification maps provided for the Grss_dfc_2013 data set by different methods. (**a**) A false color map (**b**) The ground truth map (**c**) CNN (94.59%). (**d**) M3DCNN (99.10%). (**e**) SSRN (99.32%). (**f**) MSDN-SA (99.45%). (**g**) MSO3DCNN (99.37%). (**h**) CSA-MSO3DCNN (99.69%).

Table 7. Classification accuracy for the Grss_dfc_2013 data set.

Class	CNN	M3DCNN	SSRN	MSDN-SA	MSO3DCNN	CSA-MSO3DCNN
1	98.48 ± 0.31	99.21 ± 0.15	99.78 ± 0.09	99.18 ± 0.38	99.65 ± 0.43	99.75 ± 0.36
2	96.49 ± 0.07	99.49 ± 0.19	99.78 ± 0.12	99.68 ± 0.32	99.81 ± 0.20	99.84 ± 0.16
3	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.93 ± 0.09	99.55 ± 0.63	100.00 ± 0.00
4	98.95 ± .26	99.97 ± 0.05	99.36 ± 0.05	99.87 ± 0.05	100.00 ± 0.00	99.81 ± 0.16
5	99.14 ± 0.05	100.00 ± 0.00	99.93 ± 0.09	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
6	96.93 ± 0.29	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	98.69 ± 0.67	100.00 ± 0.00
7	89.42 ± 0.63	98.41 ± 0.99	98.56 ± 0.19	98.50 ± 0.61	98.37 ± 0.54	99.09 ± 0.54
8	92.53 ± 0.49	98.31 ± 0.93	97.83 ± 0.82	99.01 ± 0.67	98.20 ± 0.53	99.55 ± 0.44
9	89.64 ± 0.14	97.21 ± 0.94	98.07 ± 0.55	98.95 ± 0.08	99.45 ± 0.32	99.11 ± 0.60
10	97.08 ± 0.35	98.96 ± 1.45	99.48 ± 0.36	99.97 ± 0.05	99.58 ± 0.59	99.81 ± 0.28
11	91.21 ± 0.25	99.48 ± 0.43	99.48 ± 0.39	99.84 ± 0.16	99.42 ± 0.47	99.68 ± 0.39
12	90.61 ± 0.42	98.48 ± 0.54	99.32 ± 0.36	98.68 ± 0.30	99.42 ± 0.23	99.77 ± 0.32
13	82.53 ± 0.07	99.88 ± 0.18	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
14	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
15	98.04 ± 0.04	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
OA(%)	94.59 ± 0.06	99.10 ± 0.28	99.32 ± 0.06	99.45 ± 0.03	99.37 ± 0.05	**99.69** ± 0.05
AA(%)	93.91 ± 0.21	99.09 ± 0.33	99.24 ± 0.11	99.34 ± 0.15	99.27 ± 0.16	**99.72** ± 0.05
$\kappa \times 100$	94.13 ± 0.05	99.01 ± 0.31	99.27 ± 0.07	99.40 ± 0.03	99.31 ± 0.05	**99.66** ± 0.06

From Table 7, comparing the results of the MSDN-SA method and the CSA-MSO3DCNN method, the OA, AA, and κ obtained by the CSA-MSO3DCNN method were all improved, which indicates the effectiveness of octave 3D CNN. Comparing with MSO3DCNN method, the proposed CSA-MSO3DCNN method gets better classification accuracy. It reveals that the feature maps selected by the channel and spatial attention module are more discriminative and efficient. From Figure 12, the classification map obtained by the proposed CSA-MSO3DCNN method is closest to the ground truth map. From the Figure 13, the normalized confusion matrices also prove this.

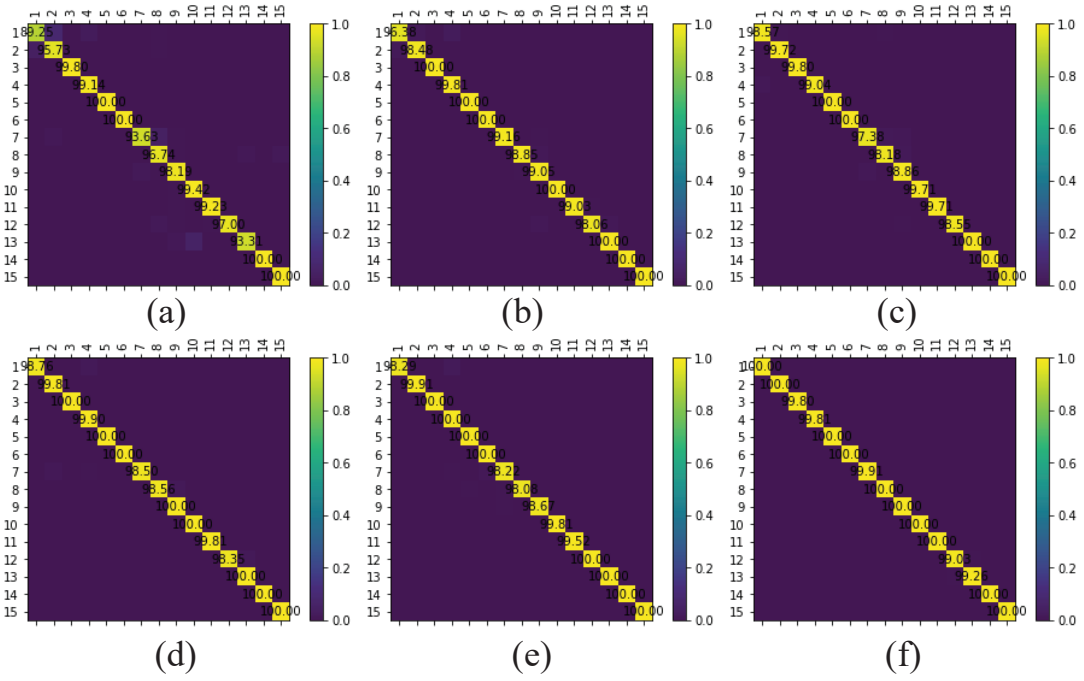


Figure 13. Normalized confusion matrices of classification results for the Grss_dfc_2013 data. (**a**) CNN. (**b**) M3DCNN. (**c**) SSRN. (**d**) MSDN-SA. (**e**) MSO3DCNN. (**f**) CSA-MSO3DCNN.

4.3.4. Results for The Grss_dfc_2014 Data Set

This data set is a challenging data set due to its low resolution. The comparison results of the classification accuracy and classification maps for The Grss_dfc_2014 data set are shown in Table 8 and Figure 14.

Table 8. Classification accuracy for the Grss_dfc_2014 data set.

Class	CNN	M3DCNN	SSRN	MSDN-SA	MSO3DCNN	CSA-MSO3DCNN
1	99.18 ± 0.67	99.29 ± 0.53	99.68 ± 0.21	99.59 ± 0.51	99.52 ± 0.28	99.81 ± 0.23
2	38.19 ± 2.40	92.53 ± 2.61	87.72 ± 3.82	95.78 ± 1.87	96.57 ± 1.58	97.50 ± 0.05
3	78.05 ± 3.51	90.27 ± 1.02	86.28 ± 2.72	92.22 ± 2.91	95.75 ± 1.32	95.83 ± 1.54
4	10.54 ± 4.20	88.65 ± 1.67	89.22 ± 2.92	91.87 ± 2.46	95.41 ± 1.89	96.99 ± 0.29
5	46.64 ± 1.31	92.81 ± 2.15	94.38 ± 2.02	96.07 ± 1.09	97.46 ± 0.75	97.94 ± 0.42
6	77.73 ± 0.89	84.74 ± 2.06	84.29 ± 1.50	88.27 ± 1.51	93.71 ± 0.29	97.24 ± 0.28
7	50.32 ± 2.35	95.86 ± 1.44	93.87 ± 2.65	96.77 ± 1.48	99.43 ± 0.29	99.26 ± 0.42
OA(%)	66.78 ± 2.11	90.45 ± 0.89	90.60 ± 1.56	93.49 ± 0.45	96.39 ± 0.41	**97.96** ± 0.20
AA(%)	62.45 ± 1.45	86.25 ± 1.33	85.46 ± 2.16	89.17 ± 0.67	93.62 ± 0.91	**96.19** ± 0.07
$\kappa \times 100$	58.18 ± 1.23	88.84 ± 1.32	88.33 ± 1.91	91.89 ± 0.56	95.48 ± 0.51	**97.37** ± 0.20

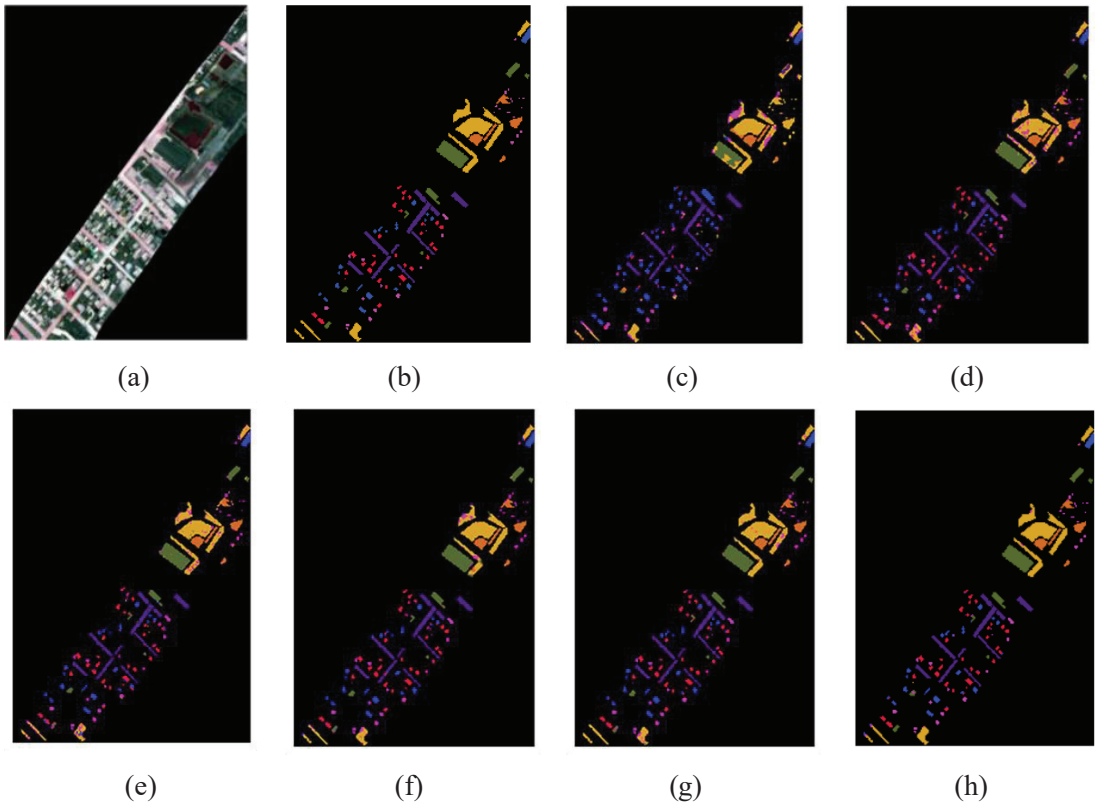


Figure 14. Classification maps provided for the Grss_dfc_2014 data set by different methods. (**a**) A false color map. (**b**) The ground truth map. (**c**) CNN (66.78%). (**d**) M3DCNN (90.45%). (**e**) SSRN (90.60%). (**f**) MSDN-SA (93.49%). (**g**) MSO3DCNN (96.39%). (**h**) CSA-MSO3DCNN (97.96%).

From Table 8, the best OA, AA and κ were obtained by our proposed CSA-MSO3DCNN method, which were 97.96%, 96.19% and 97.37% respectively. Compared with the most competitive MSDN-SA method, the proposed CSA-MSO3DCNN method made a great improvement, especially for the indicator of κ, CSA-MSO3DCNN increased by 7%. For further analysis of two results obtained by MSO3DCNN and CSA-MSO3DCNN, it could be found that the CSA-MSO3DCNN method was more

outstanding. The ablation experiment demonstrated that the feature map was optimized by the channel and spatial attention module, which can be comprehended as feature selection process. Moreover, it is obvious that the MSO3DCNN method also outperformed the other methods. It can be inferred that this data set was sensitive to spatial redundancy and the methods based on octave 3D CNN could obtain better result.

For the visual results, from Figure 14, the classification maps obtained by the CSA-MSO3DCNN method is closest to the ground truth map. For example, the 'vegetation' class, at the top right of the classification map, is all yellow, which is same to the ground truth map, and the classification maps obtained by other methods have wrong colors to some extent. The corresponding normalized confusion matrices of classification results are reported in Figure 15. It could be seen that the confusion matrix obtained by the CSA-MSO3DCNN method diagonal color is closest to yellow, which shows the best classification accuracy. The experimental results of this data set also demonstrate the superiority of our approach.

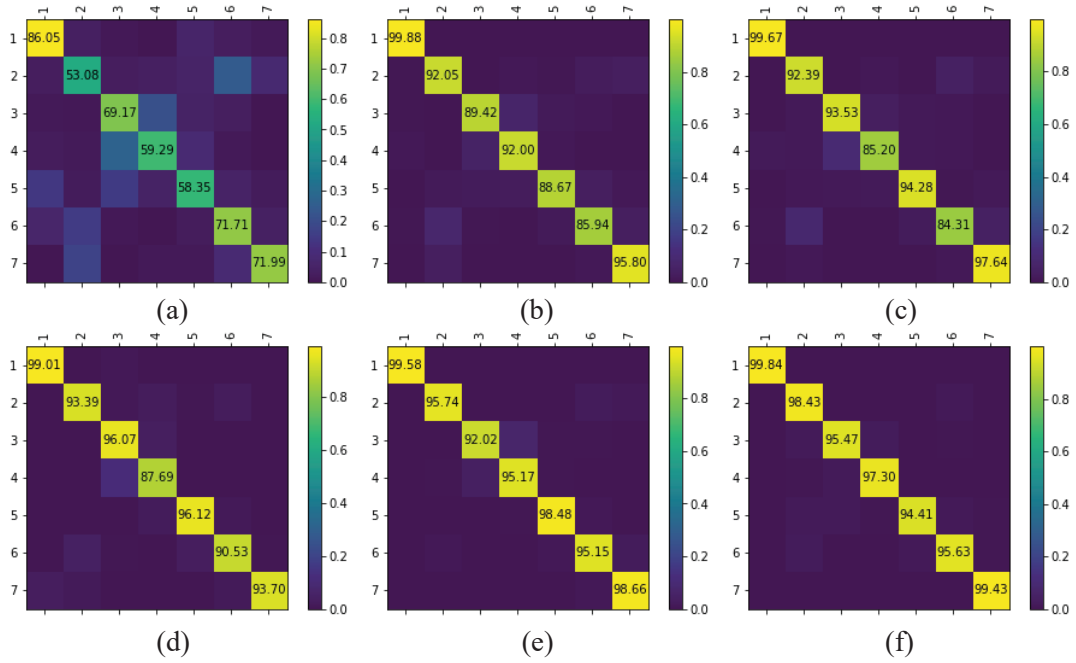


Figure 15. Normalized confusion matrices of classification results for the Grss_dfc_2014 data. (**a**) CNN. (**b**) M3DCNN. (**c**) SSRN. (**d**) MSDN-SA. (**e**) MSO3DCNN. (**f**) CSA-MSO3DCNN.

Overall, our approach excels on all the four data sets compared to other competitive methods, which indicates the robustness and stability of the CSA-MSO3DCNN method. It is worth noting that the experimental results show that spatial information has a greater influence on the results. Therefore, in the next sub-subsection, the effects of several parameters on the experiment are discussed.

4.3.5. The Effects of Parameters and Number of Training Samples

In the deep learning framework, the parameters play a significant role in the experiments. There are three main parameters in our method, which are α, spatial size and dropout. As the number of the training samples affects the quality of the ultimate model [51], the effects of the number of training samples on the ultimate model are also analyzed.

1. In our method, α characterizes the ratio between high frequency and low frequency, which decides the balance of spatial information and spatial redundancy. Thus we test a series of different α values to evaluate and get the OA results which are listed in Table 9. The test experimental results reveal that the best results are obtained for four data sets when $\alpha = 0.25$.

Table 9. Overall accuracy (OA) of the proposed method with different α.

α	Indian Pines	University of Pavia	Grss_dfc_2013	Grss_dfc_2014
0.1	97.76	98.38	99.22	96.67
0.2	98.91	99.19	99.56	97.12
0.25	**99.68**	**99.76**	**99.69**	**97.96**
0.3	98.86	98.15	99.45	97.53

2. To figure out the influence of the size of the 3D patch $s \times s \times d$, different spatial sizes $s \times s$ are conducted on the four data sets where d is set to 20. The OA results are provided in Table 10. The experimental results show that too large or too small spatial size is not recommended which means excessive noise or too little spatial information is included. It is not beneficial for the classification.

Table 10. OA of the proposed method with different spatial sizes.

Spatial Size	Indian Pines	University of Pavia	Grss_dfc_2013	Grss_dfc_2014
14×14	97.87	98.33	98.22	93.47
18×18	98.63	99.42	99.46	96.42
22×22	**99.68**	**99.76**	**99.69**	**97.96**
26×26	99.10	**99.76**	99.45	94.23
30×30	97.54	98.35	99.23	90.57

3. In the fully connected layer, the drop out is generally employed to overcome over-fitting. The effects of various drop out are depicted in Table 11. It could be observed that 0.5 was a suitable value for all four data sets, which can suppress over-fitting and train model in a balanced way.

Table 11. OA of the proposed method with different drop out.

Drop Out	Indian Pines	University of Pavia	Grss_dfc_2013	Grss_dfc_2014
0.2	92.41	90.13	91.43	87.47
0.4	98.96	99.26	99.00	97.38
0.5	**99.68**	**99.76**	**99.69**	**97.96**
0.6	98.63	98.85	99.26	96.56
0.8	93.33	91.47	90.35	92.31

In order to explore the effects of the number of training samples on the ultimate model, we have implemented more experiments by using different percentage quantity of the whole samples as the training set for all the six methods and all the data sets. For the Indian Pines data set the ratio of the training set to sample is selected from 5% to 20%, for the University of Pavia data set and Grss_dfc_2014 data set the ratio of the training set to sample is selected from 1% to 6%, and for the Grss_dfc_2013 data set the ratio of the training set to sample is selected from 10% to 20%. The obtained OA results for the four data sets are shown in Figure 16. It can be seen that, for the six methods, with the increase of the training data, the OA also increases, and our proposed method can always provide a better OA compared with the state-of-the-art methods.

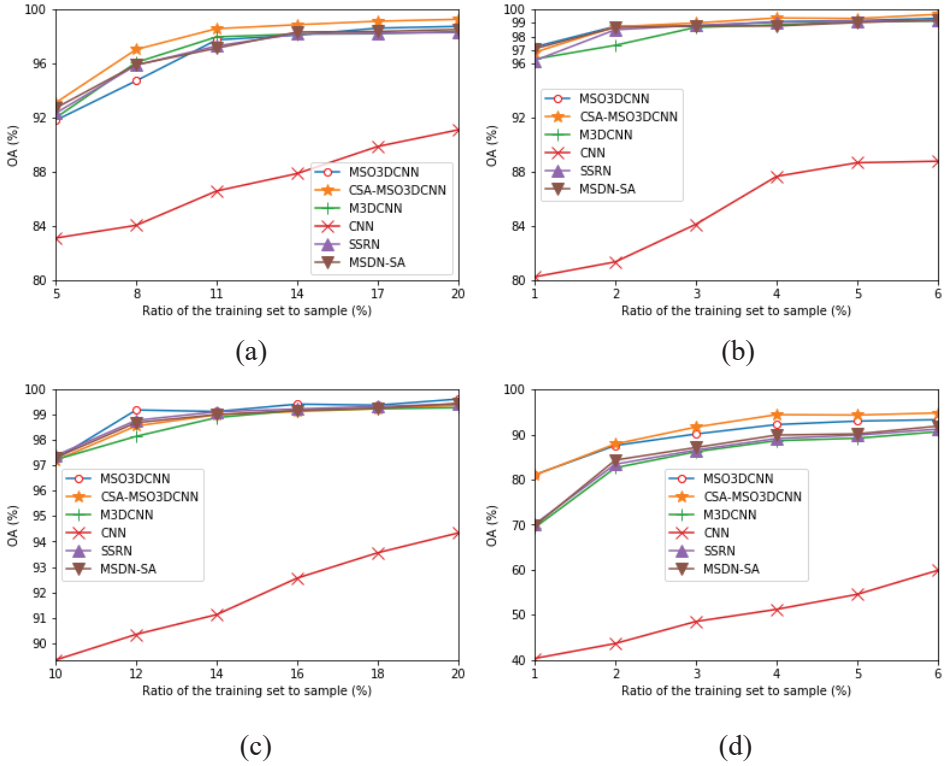


Figure 16. The OA of different training set sizes for four data sets. (**a**) Indian Pines data set. (**b**) University of Pavia data set. (**c**) Grss_dfc_2013 data set. (**d**) Grss_dfc_2014 data set.

5. Conclusions

In this paper, we have proposed a new framework based on DL for HSIs classification. Although the method based on DL has achieved good results in HSIs classification, the automatically extracted features are still rough and contain a lot of noise. Therefore, we investigate to reduce the noise of features and select more appropriate features by octave 3D CNN and attention mechanism operations.

The multi-scale octave 3D convolution is designed to decrease the spatial redundancy and expand the receptive field which are proven to be important for extracting appropriate features. Then, three different group feature maps are cascaded into one. In addition, a channel attention module and a spatial attention module are employed to refine the feature maps, which not only assign different weights to the feature maps along the channel dimension but also along the spatial dimension. The refined feature maps have been demonstrated to be beneficial for improving the classification performance. The results of ablation experiments have shown the efficiency of the attention modules. The experimental results on four public HSIs data sets have demonstrated that the proposed CSA-MSO3DCNN outperforms the state-of-the-art methods. Accordingly, it can be concluded that our method is more suitable for HSIs classification.

Because of the limit labeled HSIs pixel samples and the difficult of labeling the HSIs pixel samples, as future work, we intend to explore the HSI classification methods combined with the data enhancement techniques and semi-supervised HSIs classification in order to overcome the problem of the limit labeled HSIs pixel samples.

Author Contributions: All the authors made significant contributions to this work. Y.X. and D.W. designed and performed the experiments; Q.X. and Y.X. analyzed the results; Q.X. and Y.X. wrote the paper; B.L. and J.L. acquired the funding support. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 61502003, Grant 61860206004, Grant 61671018 and Grant 61976003, by the Key Research Project of Humanities and Social Sciences in Colleges and Universities of Anhui Province under Grant SK2019A0013.

Acknowledgments: The authors would like to thank Telops Inc. (Québec, Canada) for acquiring and providing the data used in this study, the IEEE GRSS Image Analysis and Data Fusion Technical Committee and Michal Shimoni (Signal and Image Centre, Royal Military Academy, Belgium) for organizing the 2014 Data Fusion Contest, the Centre de Recherche Public Gabriel Lippmann (CRPGL, Luxembourg) and Martin Schlerf (CRPGL) for their contribution of the Hyper-Cam LWIR sensor, and Michaela De Martino (University of Genoa, Italy) for her contribution to data preparation.

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

References

1. Wang, Q.; He, X.; Li, X. Locality and structure regularized low rank representation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 911–923. [CrossRef]
2. Yuan, Y.; Feng, Y.; Lu, X. Projection-Based NMF for Hyperspectral Unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2632–2643. [CrossRef]
3. Li, W.; Du, Q.; Zhang, B. Combined sparse and collaborative representation for hyperspectral target detection. *Pattern Recognit.* **2015**, *48*, 3904–3916. [CrossRef]
4. Pan, B.; Shi, Z.; Xu, X. MugNet: Deep learning for hyperspectral image classification using limited samples. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 108–119. [CrossRef]
5. Zhou, S.; Xue, Z.; Du, P. Semisupervised Stacked Autoencoder With Cotraining for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3813–3826. [CrossRef]
6. Ghamisi, P.; Plaza, J.; Chen, Y.; Li, J.; Plaza, A.J. Advanced Spectral Classifiers for Hyperspectral Images: A review. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–32. [CrossRef]
7. Camps-Valls, G.; Marsheva, T.V.B.; Zhou, D. Semi-Supervised Graph-Based Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3044–3054. [CrossRef]
8. Gu, Y.; Feng, K. L1-graph semisupervised learning for hyperspectral image classification. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 1401–1404.
9. Ma, L.; Crawford, M.M.; Yang, X.; Guo, Y. Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2832–2844. [CrossRef]
10. Shao, Y.; Sang, N.; Gao, C.; Ma, L. Probabilistic class structure regularized sparse representation graph for semi-supervised hyperspectral image classification. *Pattern Recognit.* **2017**, *63*, 102–114. [CrossRef]
11. Roscher, R.; Waske, B.; Forstner, W. Incremental import vector machines for classifying hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3463–3473. [CrossRef]
12. Li, J.; Marpu, P.R.; Plaza, A.; Bioucas-Dias, J.M.; Benediktsson, J.A. Generalized composite kernel framework for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4816–4829. [CrossRef]
13. Jiang, J.; Ma, J.; Wang, Z.; Chen, C.; Liu, X. Hyperspectral image classification in the presence of noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 851–865. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
17. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]

20. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4959–4962.
21. Li, W.; Wu, G.; Zhang, F.; Du, Q. Hyperspectral image classification using deep pixel-pair features. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 844–853. [CrossRef]
22. Zhang, M.; Li, W.; Du, Q. Diverse region-based CNN for hyperspectral image classification. *IEEE Trans. Image Process.* **2018**, *27*, 2623–2634. [CrossRef] [PubMed]
23. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
24. Li, Y.; Zhang, H.; Shen, Q. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* **2017**, *9*, 67. [CrossRef]
25. He, M.; Li, B.; Chen, H. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3904–3908.
26. 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.* **2017**, *56*, 847–858. [CrossRef]
27. 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]
28. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
29. Sellami, A.; Farah, M.; Farah, I.R.; 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]
30. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2204–2212.
31. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv* **2014**, arXiv:1412.7755.
32. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
33. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
34. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
36. Haut, J.M.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Li, J. Visual Attention-Driven Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8065–8080. [CrossRef]
37. Fang, B.; Li, Y.; Zhang, H.; Chan, J.C.W. Hyperspectral Images Classification Based on Dense Convolutional Networks with Spectral-Wise Attention Mechanism. *Remote Sens.* **2019**, *11*, 159. [CrossRef]
38. Mamalet, F.; Garcia, C. Simplifying convnets for fast learning. In Proceedings of the International Conference on Artificial Neural Networks, Lausanne, Switzerland, 11–14 September 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 58–65.
39. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
40. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
41. Rensink, R.A. The dynamic representation of scenes. *Vis. Cogn.* **2000**, *7*, 17–42. [CrossRef]
42. Corbetta, M.; Shulman, G.L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **2002**, *3*, 201–215. [CrossRef]

43. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional Sequence to Sequence Learning. In Proceedings of the 34th International Conference on Machine Learning, JMLR.org, ICML'17, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1243–1252.
44. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
45. Zhang, X.; Wang, T.; Qi, J.; Lu, H.; Wang, G. Progressive attention guided recurrent network for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 714–722.
46. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
47. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
48. Chen, Y.; Fang, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. *arXiv* **2019**, arXiv:1904.05049.
49. 2013 IEEE GRSS Data Fusion Contest. Available online: http://www.grss-ieee.org/community/technical-committees/data-fusion/ (accessed on 23 June 2012).
50. 2014 IEEE GRSS Data Fusion Contest. Available online: http://www.grss-ieee.org/community/technical-committees/data-fusion/ (accessed on 27 January 2014).
51. Ahmad, M.; Khan, A.; Khan, A.M.; Mazzara, M.; Distefano, S.; Sohaib, A.; Nibouche, O. Spatial Prior Fuzziness Pool-Based Interactive Classification of Hyperspectral Images. *Remote Sens.* **2019**, *11*, 1136. [CrossRef]