remote sensing

MDPI

Article

Attention-Guided Multispectral and Panchromatic Image Classification

Cheng Shi [1], Yenan Dang [1], Li Fang [2,*], Zhiyong Lv [1] and Huifang Shen [2]

[1] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; C_Shi@xaut.edu.cn (C.S.); 2201221089@stu.xaut.edu.cn (Y.D.); zhiyongLyu@xaut.edu.cn (Z.L.)
[2] The Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou 362000, China; shenhf@fjirsm.ac.cn
* Correspondence: fangli@fjirsm.ac.cn

Abstract: Multi-sensor image can provide supplementary information, usually leading to better performance in classification tasks. However, the general deep neural network-based multi-sensor classification method learns each sensor image separately, followed by a stacked concentrate for feature fusion. This way requires a large time cost for network training, and insufficient feature fusion may cause. Considering efficient multi-sensor feature extraction and fusion with a lightweight network, this paper proposes an attention-guided classification method (AGCNet), especially for multispectral (MS) and panchromatic (PAN) image classification. In the proposed method, a share-split network (SSNet) including a shared branch and multiple split branches performs feature extraction for each sensor image, where the shared branch learns basis features of MS and PAN images with fewer learn-able parameters, and the split branch extracts the privileged features of each sensor image via multiple task-specific attention units. Furthermore, a selective classification network (SCNet) with a selective kernel unit is used for adaptive feature fusion. The proposed AGCNet can be trained by an end-to-end fashion without manual intervention. The experimental results are reported on four MS and PAN datasets, and compared with state-of-the-art methods. The classification maps and accuracies show the superiority of the proposed AGCNet model.

Keywords: multi-sensor classification; attention mechanism; deep neural network; multispectral and panchromatic image

Citation: Shi, C.; Dang, Y.; Fang, L.; Lv, Z.; Shen, H. Attention-Guided Multispectral and Panchromatic Image Classification. *Remote Sens.* **2021**, *13*, 4823. https://doi.org/10.3390/rs13234823

Academic Editor: Xinghua Li

Received: 13 October 2021
Accepted: 8 November 2021
Published: 27 November 2021

1. Introduction

The rapid development of aerospace technology has generated a large number of remote sensing images from a variety of sensors [1–4], and the research interests in multi-sensor image classification is also increasing, especially for multispectral (MS) and panchromatic (PAN) images. The MS and PAN images are usually captured using the optical satellites, and have different characteristics. Generally, the MS image consists of four spectral bands, and the PAN image has only one band. However, the PAN image has higher spatial resolution than that of MS image. For taking full use of the complementary spectral and spatial information, the processing methods of MS and PAN images are usually classified into two models: fusion-based classification model and classification-based fusion model. The fusion-based classification model is to pan-sharpen the MS image for improving its spatial resolution, followed by a classification process on the pan-sharpened MS image. The classification-based fusion model is to capture the features of MS and PAN images respectively and then combine these features for classification. The fusion-based classification model pays more attention to obtain effective fusion images [5], while the classification-based fusion model focuses more on the effective classification [6]. To avoid the influence of the fusion effect on the classification results, the classification-based fusion model is adopted in this study for MS and PAN image classification.

Depth perception has proved to be effective in remote sensing image classification [7,8]. The common deep learning practice of classification-based fusion model is shown in Figure 1. In the classification-based fusion model, a feature extraction model for MS image and another one for PAN image are trained separately by minimizing a loss function, and the prediction result could be obtained by classifying the simply fused higher-level features [6,9]. The classification-based fusion model may be preferable to train joint features. However, two limitations could be analyzed: (1) two feature extraction models with two independent networks are usually trained with a higher time cost; (2) the simple feature fusion method does not consider the importance level of each model for classification. For solving these two limitations, a low-complexity multi-sensor feature extraction method and an adaptive feature fusion method are studied in this paper.

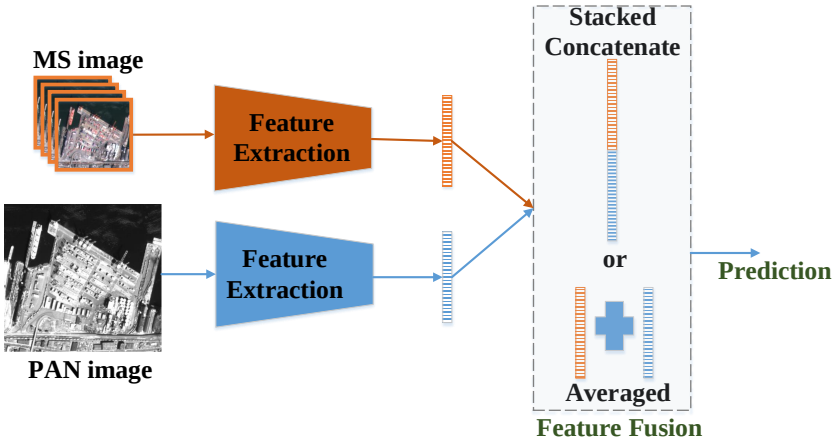

Figure 1. General framework of MS and PAN image classification based on a classification-based fusion model. (Quickbird satellite with 2.44-m MS image and 0.61-m PAN image).

An attention-guided classification network (AGCNet) is proposed for MS and PAN image classification to tackle these two limitations of the former classification-based fusion model. The AGCNet mainly consists of two networks: a share-split network (SSNet) and a selective classification network (SCNet). The network architecture is shown in Figure 2a. The whole network is an end-to-end form that can be simply trained.

Training the multi-sensor image usually requires more learn-able parameters. Unsurprisingly, the simplest way of reducing the training cost is to reduce the number of learn-able parameters [10–12]. A reasonable way of reducing the learn-able parameters in multi-sensor classification network is to construct a shared branch, where the parameters are shared for MS and PAN images. Meanwhile, inspired by the squeeze-and-excitation network (SENet) [13], multiple split branches with task-specific attention units are designed to capture the specific features of MS and PAN images respectively. The task-specific attention units can adaptively re-weight the share-channel feature for selecting emphasis information and suppressing the less useful ones. Although the learn-able parameters of task-specific attention units are privileged for MS and PAN images, the training cost is slightly increased.

The classification performance also depends on the effective fusion of privileged features of MS and PAN images. The contributions of MS and PAN images to classification result are imbalanced, so a weighted fusion method is more effective than the general stacked fusion method (shown in Figure 1). Building upon the idea of selective kernel network (SKNet) [14], this study uses an attention-based selective kernel unit to generate an adaptive selection weight. The privileged features of MS and PAN images are adaptively weighted for classification. This selection operator is also computationally lightweight.

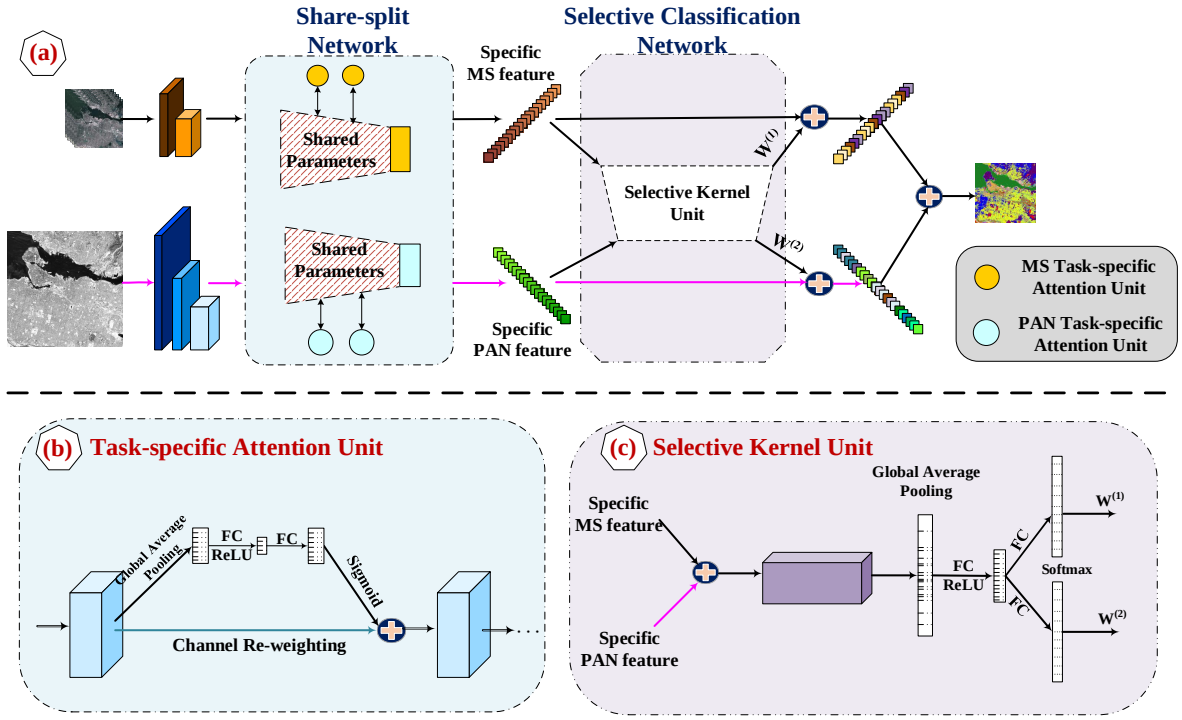

Figure 2. (**a**) The architecture of the proposed attention-guided classification network. In SSNet, the shared branch is designed for extracting basis features of the MS and PAN image, and multiple split branches with MS task-specific attention units and PAN task-specific attention units are designed to capture the specific features of MS and PAN images, respectively. In SCNet, the specific features of MS and PAN images are fused with a selective kernel unit. (**b**) Task-specific attention unit. (**c**) Selective kernel unit.

The contributions of this study are listed as follows.

(1) A novel multi-sensor feature extraction approach is proposed to learn a SSNet, which consists of a shared branch and multiple split branches with task-specific attention units. The shared branch is designed for learning basis features and reducing the learnable parameters, and the task-specific attention units are constructed to learn the specific features of MS and PAN images.

(2) The privileged features of MS and PAN images are combined by an attention-based selective kernel unit. The selective kernel unit can generate adaptive global weights with fewer additional learn-able parameters.

(3) In the experiments, four groups of experimental images are used for multi-sensor classification. Compared with the general multi-sensor classification models, the performance of the proposed attention-guided network is improved with fewer training and testing costs.

The rest of the paper is organized as follows. The literature reviews are surveyed in Section 2. The details of the proposed method are described in Section 3. Section 4 presents the experimental results and analyses. The conclusions and future work of this study are discussed in Section 5.

2. Literature Reviews

For designing the multi-sensor feature extraction approach, Section 2.1 summarizes the overview of the multi-sensor remote sensing image classification with deep learning techniques, and advantages and limitations of these techniques are discussed. Furthermore, Section 2.2 presents a brief introduction to the attention mechanisms, and the key ideas of each attention mechanism are illustrated for designing the attention-based selective operator in this study.

2.1. Multi-Sensor Remote Sensing Image Classification with Deep Learning

In recent studies, deep learning-based remote sensing classification techniques have achieved promising results [15,16]. Several typical deep learning networks like stacked auto-encoders (SAE) [17], convolutional auto-encoders (CAE) [18], deep belief networks (DBN) [19], convolutional neural network (CNN) [20,21], and recurrent neural network (RNN) [22] have been adopted for remote sensing image classification. To further improve the classification accuracy, multi-scale feature learning techniques [23–25] and generative adversarial network (GAN) [26–29] have received widespread attention. These methods are dedicated to single-sensor image classification tasks. Actually, remote sensing image classification combined with multi-sensor image is possible to further improve the classification accuracy.

For instance, ref. [30] proposed an advanced multi-sensor remote sensing classification method for urban land use. A fusion-FCN (Fusion-fully convolutional network) was proposed to well maintain the boundary information and reduce the spatial loss in the classification map. The fusion-FCN received three-sensor images as inputs and was trained separately; a stacked concatenate layer was adopted for feature fusion and a softmax classifier was used for classification.

In another studies, a hyperspectral and multispectral-based fusion classification method was proposed to include a compressive measurement model for extracting the features of each sensor image [31]; the feature fusion problem was defined to estimate the new features that could better capture the useful information from multi-sensor compressive measurements. Furthermore, ref. [32] extracted the features of each sensor-image via a compressive measure technology, and the acquired features are stacked classified with a support vector machine (SVM).

Another related studies are [6,33]. In [6], a superpixel-based multiple local CNN model was proposed to recognize MS and PAN images; MS images were adopted to obtain an initial classification, and PAN images were used to modify the detailed errors. However, the acquisition of an initial classification map requires training six local regions separately, which is time-consuming. In [33], a multi-instance network was proposed to improve MS and PAN image classification; one instance was used for extracting the spectral feature of MS image, and the other instance was used for extracting the spatial features of PAN image. The extracted features from these two instances were stacked concatenated for fusion and classification.

The difference of the proposed AGCNet from [6] is that this study tries to learn an effective feature representation from the MS and PAN images directly, without post-processing. In addition, compared with above mentioned methods, the proposed attention-guided classification architecture consists of a computationally lightweight multi-sensor feature extraction network and an adaptive feature fusion network. This mechanism can easily be extended to the general multi-feature classification framework [34,35].

2.2. Attention Mechanisms

Recently, the visual attention mechanisms have been proposed to improve the network performance [36]. On the one hand, the attention mechanism is introduced to the spatial dimension, such as integrating multi-scale spatial information or spatial dependencies into network [37–40]. On the other hand, some studies focus to capture the relationship between channels, and propose channel-wise attention mechanisms, such as squeeze-and-excitation (SE) block [13,41] and SKNet [14]. In particular, the SE block can learn the global information to selectively important features and can be freely inserted into any network, and therefore, some studies extended SE block to remote sensing applications. In [42], the authors incorporated the spatial attention and channel attention to residual network for scene classification. In [43], the channel-based DenseNet was proposed for remote sensing image scene classification. In [44], channel-wise attention block was embedded into a dual-level semantic concept network for multi-label remote sensing image annotation. In [45], a multi-scale visual attention network was proposed for object detection in remote sensing

image; this is the first time that attention has been introduced into the encoder-decoder model for object detection. In [46], an enhanced attention module was presented for remote sensing scene classification; a global average pooling and a global max pooling were used to aggregate the global spatial feature, and a multilayer.

Perception was designed to learn the channel attention map. In [47], an task-specific attention domain adaptation method was proposed for Satellite-to-Aerial scene. In [48], a spectral-spatial squeeze-and-excitation residual bag-of-feature network was proposed for hyperspectral image (HSI) classification, of which two residual SE blocks were used to extract the spectral and spatial features, respectively. Another recent related work is [49], a spatial attention module and a spectral attention module were designed to strengthen the spatial features of PAN image and the spectral features of MS image, respectively; furthermore, a dual-branch attention fusion network was proposed for multiresolution remote sensing image classification. In above-mentioned studies, the introduction of the attention model is to further improve the performance of network. In contract to these studies, this paper focuses on reducing the number of learn-able parameters and better balancing the time cost and the classification effect.

3. Learning to Attention-Guided Classification Network

In this section, the details of the proposed AGCNet are presented in two subsections. In Section 3.1, a lightweight SSNet is proposed to extract the deep-level features of MS and PAN images, respectively; compared to the general multi-sensor classification model, the learn-able parameters in SSNet could be significantly reduced. In Section 3.2, a SCNet is provided for adaptive feature fusion, in which the weight calculation fully considers the global information of the features; compared with the simply fusion strategy, the increase of learn-able parameters in SCNet is lightweight. The architecture of AGCNet is shown in Figure 2a. The stages are presented in the following subsections.

3.1. Share-Split Network for Multi-Sensor Feature Extraction

In the general multi-sensor image classification framework, the feature extraction model of each sensor image (i.e., MS and PAN images) is trained separately, resulting in an increase in learn-able parameters. The goal of this study is to construct a shared branch to reduce the learn-able parameters and design multiple split branches with task-specific attention units for extracting the specific features of MS and PAN images.

Initial feature extraction of MS and PAN images. The shared branch requires the input feature size of the two sensors to be the same. Due to the differences in the spatial and spectral resolution of PAN and MS images, an initial feature extraction process is necessary.

PAN and MS images are denoted as $f_{PAN}^{(1)}$ and $f_{MS}^{(1)}$. Since the size of PAN image is four times larger than the MS image, three feature extraction layers with convolution filtering and max-pooling are used to reduce the feature size of PAN image, and two feature extraction layers are applied on the MS image for feature extraction.

The features of PAN and MS images in the l-th layer are written as $f_{PAN}^{(l)}$ and $f_{MS}^{(l)}$ (shown in Equations (1) and (2)).

$$f_{PAN}^{(l)} = g_{pooling}(g_{Relu}(f_{PAN}^{(l-1)} * W_{PAN}^{(l-1)})), \ if \ 2 \leqslant l \leqslant 4, \tag{1}$$

$$f_{MS}^{(l)} = \begin{cases} g_{Relu}(f_{MS}^{(2)} * W_{MS}^{(2)}), & if \ l = 3, \\ g_{pooling}(g_{Relu}(f_{MS}^{(3)} * W_{MS}^{(3)})), & if \ l = 4. \end{cases} \tag{2}$$

Here * represents convolution, $W_{PAN}^{(l-1)}$ and $W_{MS}^{(l-1)}$ are the 2D spatial filters, and $g_{pooling}$ is the ReLU activation function. The features of $f_{PAN}^{(4)}$ and $f_{MS}^{(4)}$ have the same spatial and channel sizes.

Shared branch for shared feature learning. This study designs a shared branch to reduce the number of learn-able parameters. The shared branch consists of several

convolution filtering and max-pooling operators. If the layer l is larger than 5, shared filters W in Equations (3) and (4) will be adopted for shared feature extraction.

$$f_{PAN}^{(l)} = g_{pooling}(g_{Relu}(f_{PAN}^{(l-1)} * W_{PAN}^{(l-1)})), \; if \; 5 \leqslant l \leqslant l_n, \tag{3}$$

$$f_{MS}^{(l)} = g_{pooling}(g_{Relu}(f_{MS}^{(l-1)} * W_{MS}^{(l-1)})), if \; 5 \leqslant l \leqslant l_n, \tag{4}$$

where l_n is the total layer number. In the shared branch, all of the MS and PAN features are used to train the same network. Therefore, the number of training samples is increased by half, but the number of learn-able parameters is reduced by half. The shared branch can learn basis features more effectively. However, the MS and PAN images also have some different characteristics that cannot be represented well by basis features. Therefore, task-specific attention units are designed to capture the privileged information of MS and PAN images.

Task-specific attention unit for privileged feature learning. To achieve the specific feature extraction, each convolution layer of the shared branch is followed by a MS task-specific attention unit and a PAN task-specific attention unit. The parameters of MS and PAN task-specific attention units are privileged and trained in accordance with the MS and PAN images, respectively. By performing a task-specific re-weighting operator on shared convolution features, the MS and PAN task-specific attention units can better learn the different complementary features.

The structure of the task-specific attention unit is shown in Figure 2b. For capturing the relationship between channels, a global average pooling operator is applied on each channel of the convolution feature to obtain a global statistic. The convolution feature is denoted as $f \in \Re^{H \times W \times c}$ and adopted into Equation (5) for extracting the global feature of channel.

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_c(i, j). \tag{5}$$

The notations in Equation (5) are listed as follows. H and W are the spatial size of feature f; f_c is the convolution feature of channel C; $z \in \Re^{1 \times 1 \times C}$ is the global feature; Z_c is the global feature of channel c; C is the channel number. For capturing the channel dependencies, two fully connected layers are used for feature combination. Equation (6) is designed to learn the task-specific weight vector.

$$w = f_{Sigmoid}(W_{Full}^{(2)}(f_{Relu}(W_{Full}^{(1)} z))), \tag{6}$$

where W_{Full} is the fully connected parameter vector, and $w \in \Re^{1 \times 1 \times C}$ is the task-specific weight vector. The task-specific weight vector w and the feature f are adopted to obtain a re-weighted feature $\tilde{f} \in \Re^{H \times W \times C}$ by Equation (7) based on the element-wise multiplication.

$$\tilde{f} = w \otimes f, \tag{7}$$

where $\otimes$ represents the channel-wise multiplication. The weight vector w can select the emphasis information and suppress the fewer useful ones. The re-weighted feature $\tilde{f}$ and the input feature have the same spatial and channel sizes.

3.2. Selective Classification Network for Adaptive Feature Fusion

In Section 3.1, specific features of MS and PAN images are obtained by a SSNet. In this subsection, an adaptive SCNet is applied on these two specific features for feature fusion and classification. The importance levels of different sensor features are usually not considered by the simply fusion strategies (e.g., stacked concatenated operation [50–53] and averaged operation [9]). For further considering the channel-dependencies and sensor-importance, an attention-based selective kernel unit is designed for multi-sensor feature fusion, which is shown in Figure 2c.

Selective kernel unit for multi-sensor features fusion. The specific features of MS and PAN images obtained by Section 3.1 are recorded as $f_{MS} \in \Re^{H\times W\times C}$ and $f_{PAN} \in \Re^{H\times W\times C}$, where C is the channel number of the feature f_{MS} and f_{PAN}. These two features are first integrated via an element-wise summation by Equation (8).

$$f_{add} = f_{MS} + f_{PAN}. \tag{8}$$

Equation (5) on the feature f_{add} is applied to obtain a global statistic feature $Z \in \Re^{1\times 1\times C}$. For capturing the channel-dependencies, a compact fully-connected layer is used to obtain a dimensional-reduced feature $s \in \Re^{1\times 1\times d}(d < C)$ by Equation (9).

$$s = f_{Relu}(W_c z), \tag{9}$$

where $W_c \in \Re^{d\times C}$ is the fully connected parameter vector. Guided by the compact feature s, the fully connected operator is used to extract channel-wise attention information of each sensor and adopted into Equations (10) and (11).

$$\tilde{s}^{(1)} = W_{Full}^{(1)} s, \tag{10}$$

$$\tilde{s}^{(2)} = W_{Full}^{(2)} s, \tag{11}$$

where $W_{Full}^{(1)} \in \Re^{C\times d}$ and $W_{Full}^{(2)} \in \Re^{C\times d}$ are the fully connected parameters; $\tilde{s}^{(1)} \in \Re^{1\times 1\times C}$ and $\tilde{s}^{(2)} \in \Re^{1\times 1\times C}$ denote the channel-wise attention feature of each sensor. A softmax operator on the channel-wise is applied to obtain the final adaptive fusion weight by Equations (12) and (13).

$$w^{(1)} = \frac{e^{\tilde{s}^{(1)}}}{e^{\tilde{s}^{(1)}} + e^{\tilde{s}^{(2)}}}, \tag{12}$$

$$w^{(2)} = \frac{e^{\tilde{s}^{(2)}}}{e^{\tilde{s}^{(1)}} + e^{\tilde{s}^{(2)}}}. \tag{13}$$

The softmax result makes that the sum of $w^{(1)}$ and $w^{(2)}$ equals to l. The specific features f_{MS} and f_{PAN} are re-weighted by the fusion weights $w^{(1)}$ and $w^{(2)}$ to obtain the fused feature by Equation (14).

$$F = w^{(1)} \cdot f_{MS} + w^{(2)} \cdot f_{PAN}. \tag{14}$$

Classification. Finally, a softmax classifier [44] is used for classification. For training the network, the loss function is designed as Equation (15).

$$\zeta = -\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i log(y_i) + (1-\tilde{y}_i) log(1-y_i)] + \alpha \sum_{j=1}^{N} W_j^2, \tag{15}$$

where $\tilde{y}_i$ and y_i are the i-th predicted-label and true-label; m is the mini-batch-size; N is the number of learn-able parameters; α is a free parameter. In the experiments of this study, α is selected as 10^{-5}. In Equation (15), the first term is the cross-entropy loss, and the second term is the L2 regularization to prevent overfitting. The proposed AGCNet is trained end-to-end by using mini-batch stochastic gradient descent.

4. Experiments and Discussions

4.1. Datasets

The performance of the proposed method is evaluated on four datasets, which were obtained by two different satellites. In the following paragraphs, the details of these four datasets are illustrated.

Level 1B datasets and Level 1C dataset were obtained by DEIMOS-2 satellites in Vancouver, Canada, on 31 March 2015 and 30 May 2015, which were provided by 2016 IEEE GRSS Data Fusion Contest [54]. Each dataset contains an MS image with 4-m spatial resolution and a PAN image with 1-m spatial resolution. For Level 1B dataset, the size of MS image is 3249 × 2928 with four spectral bands. The size of PAN image is 12,996 × 11,712 with one band. The Level 1B dataset contains 11 available categories. For Level 1C dataset, the sizes of MS and PAN images are 1311 × 873 and 5244 × 3492 respectively. The Level 1C dataset contains 8 available categories. Figure 3a,b show the Level 1B and Level 1C datasets and their ground-truth maps.

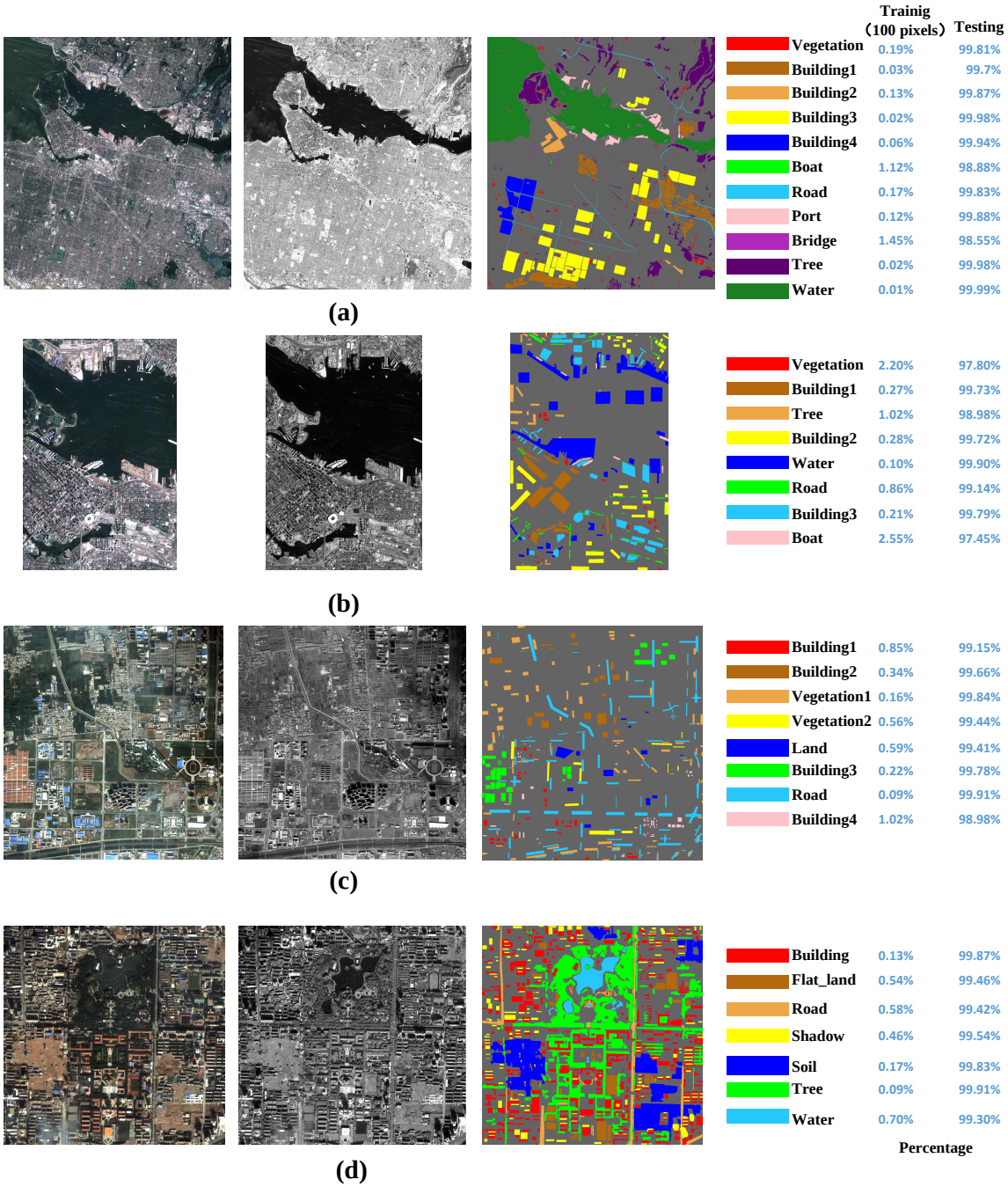

Figure 3. Datasets: (**a**) Level 1B dataset (from left to right: False color image of MS image, PAN image, ground-truth map, and class information) (**b**) Level 1C dataset. (**c**) Xi'an Suburban dataset. (**d**) Xi'an Urban dataset.

Xi'an Suburban dataset and Xi'an Urban dataset were acquired by QuickBird satellites on 30 May 2008 [33].The spatial resolutions of PAN and MS image are 0.61-m and 2.44-m respectively. For the Xi'an Suburban dataset, the size of MS image is 1650 × 1550 with four bands, and the size of PAN image is 6600 × 6200 with one band. The 8 available categories are used for classification. The Xi'an Urban dataset consists of an MS image with size 800 × 830 and a PAN image with size 3200 × 3320, and 7 available categories are used for classification. Figure 3c,d show the Xi'an Suburban and Xi'an Urban datasets and their ground-truth maps.

4.2. Experimental Setup

The detail parameters of the proposed network are shown in Table 1. The size of PAN image is four times larger than that of MS image, and the scene size is the same as MS image. Therefore, the MS image is classified pixel-by-pixel, and the PAN image is classified by interval 4 pixels. For each dataset in the experiments, 100 pixels per class are randomly selected for training, and the rest pixels are used for testing. For collecting the spatial information, each pixel is taken as the center to obtain the sample patch. The size of the MS sample patch is 32 × 32 × 4 and the size of the PAN sample patch is 128 × 128. In addition, the learning rate is set as 0.005, the iteration number is 10,000, and the batch size is 64. The experimental results are mean values over 10 experiments by selecting the training samples randomly.

Table 1. Parameter setting of the proposed AGCNet.

Network Structure/Operator		**Convolution/Full Connection Size**	**[Stride Padding Poolin Activation]**	**Convolution/Full Connection Size**	**[Stride Padding Pooling Activation]**
Initial Feature Extraction		–	–	3 × 3 × 16	[1 1 Max-pooling(2) ReLU]
		3 × 3 × 32	[1 1–ReLU]	3 × 3 × 32	[1 1 Max-pooling(2) ReLU]
		3 × 3 × 64	[1 1 Max-pooling(2) ReLU]	3 × 3 × 64	[1 1 Max-pooling(2) ReLU]
Share split network	Shared branch	3 × 3 × 128	[1 1 Max-pooling(2) ReLU]	**Shared Parameters**	
	Task specific attention unit	–	[– – Avg-pooling –]	–	[– – Avg-pooling –]
		8 × 128	Activation = ReLU	8 × 128	Activation = ReLU
		128 × 8	Activation=Sigmoid	128 × 8	Activation=Sigmoid
Selective classification network	Selective kernel unit	–	[– – Avg-pooling –]	–	[– – Avg-pooling –]
		8 × 128	Activation = ReLU	8 × 128	Activation = ReLU
		128 × 8/128 × 8	Softmax	128 × 8/128 × 8	Softmax
	classification	Softmax			

4.3. Comparison Results

In this subsection, seven state-of-the-art methods are compared to verify the effectiveness of the proposed AGCNet, including extended multi-attribute profiles (EMAP) [55], convolutional auto-encoder (CAE) [18], recurrent neural network (RNN) [22], spatial-channel progressive fusion residual network (SCPF-ResNet) [49], convolutional neural network based on MS images (CNN-MS) [53], convolutional neural network based on PAN images (CNN-PAN) [53] and stacked fusion network (SFNet) [32]. In particular, EMAP, CAE and RNN are verified on MS images. SCPF-ResNet is a very related study, which combines the spatial and channel attentions for MS and PAN image classification. The parameters of CAE, RNN and SCPF-ResNet are set as default values in their papers. CNN-MS and CNN-PAN mean that the CNN is used to classify the MS and PAN images respectively. For a fair comparison, the parameters settings of CNN are consistent with the proposed method, including the number of layers and filter size. In SFNet, the features of MS and PAN are extracted by CNN respectively, and the two features are concatenated for classification; the feature fusion strategy adopts the method in [32], and the parameter setting of CNN is still consistent with Table 1 for a fair comparison. The comparison results

on the four datasets are shown and analyzed below, and overall accuracy (OA), average accuracy (AA) and kappa coefficient (kappa) are used for quality metrics.

1. Experimental Results with Level 1B and Level 1C Datasets

Level 1B and Level 1C (Tables 2 and 3) are very challenging datasets due to the complex scene information. As can be seen from the ground-truth maps in Figure 3 that there exist many building groups. Usually there is a lot of interference information in the building group, such as roads and trees. In addition, the characteristics of different land-covers are highly similar, such as Classes 2, 3, 4, and 5 (Building 1, Building 2, Building 3, and Building 4) in Level 1B dataset, and Classes 2, 4, and 7 (Building 1, Building 2, and Building 3) in Level 1C dataset. These land-covers have similar attributes, but belong to different classes. The insignificant difference increases the difficulty of classification. The classification results of Level 1B and Level 1C datasets are shown in Figures 4 and 5.

RNN obtains the worse classification results than the other comparison methods. The main advantage of RNN is that only the spectral information is considered and the spatial dependence is ignored. Therefore, the classification results are affected by noise, especially for the building areas. The classification accuracies of Building 1 are only 17.38% for Level 1B dataset and 28.93% for Level 1C dataset.

Table 2. Classification accuracy on Level 1B dataset.

Class	EMAP	CAE	RNN	SCPF-ResNet	CNN-MS	CNN-PAN	SFNet	AGCNet
1 (Vegetation)	0.9422	0.9493	0.9178	0.8642	**0.9894**	0.9444	0.9751	0.9351
2 (Building1)	0.5993	0.8160	0.1738	0.3039	0.8582	0.7538	0.9006	**0.9217**
3 (Building2)	0.6831	0.9694	0.2841	0.6673	**0.9821**	0.9697	0.9748	0.9805
4 (Building3)	0.6692	0.8863	0.4932	0.6019	0.9258	0.8947	0.9184	**0.9628**
5 (Building4)	0.7528	0.9524	0.4987	0.6637	0.9492	0.8825	0.9358	**0.9742**
6 (Boat)	0.7962	**0.9810**	0.5602	0.8127	0.9941	0.9800	0.9636	0.9765
7 (Road)	0.3883	0.7225	0.5186	0.5580	**0.8252**	0.8189	0.8125	0.6940
8 (Port)	0.5034	0.8703	0.4025	0.2836	0.9066	0.8615	**0.9356**	0.8639
9 (Bridge)	0.6893	0.9303	0.2724	0.8916	**0.9605**	0.9662	0.9477	0.9589
10 (Tree)	0.9173	0.9288	0.9136	0.4574	0.9278	0.8947	0.9544	**0.9709**
11 (Water)	**0.9895**	0.9802	0.9872	0.9823	0.9876	0.9836	0.9806	9864
OA	0.8378	0.9297	0.7377	0.7152	0.9475	0.9190	0.9507	**0.9633**
Kappa	0.7865	0.9071	0.6560	0.6288	0.9304	0.8928	0.9347	**0.9512**
AA	0.7210	0.9079	0.5475	0.6442	**0.9370**	0.9046	0.9363	0.9300

Table 3. Classification accuracy on Level 1C dataset.

Class	EMAP	CAE	RNN	SCPF-ResNet	CNN-MS	CNN-PAN	SFNet	AGCNet
1 (Vegetation)	0.8698	0.9865	0.9206	0.9280	0.9775	0.9294	**0.9931**	0.9851
2 (Building1)	0.5241	0.9601	0.2897	0.5980	0.9604	0.9331	0.9515	**0.9626**
3 (Tree)	0.8315	0.9735	0.8356	0.8679	**0.9805**	0.9349	0.9339	0.9414
4 (Building2)	0.3973	0.8058	0.3366	0.5555	0.8534	0.8578	0.8926	**0.9248**
5 (Water)	**0.9963**	0.9077	0.9924	0.9346	0.9865	0.9810	0.9755	0.9772
6 (Road)	0.7889	0.7009	0.6410	0.7102	0.7013	0.7428	**0.8024**	0.7533
7 (Building3)	0.4607	0.8256	0.4229	0.6052	0.8296	0.7890	0.8534	**0.9072**
8 (Boat)	0.5849	0.9911	0.5092	0.9198	0.9668	0.9694	0.9910	**0.9936**
OA	0.7111	0.8806	0.6515	0.7525	0.9191	0.9051	0.9225	**0.9405**
Kappa	0.6297	0.8465	0.5557	0.6853	0.8943	0.8763	0.9034	**0.9254**
AA	0.6817	0.8939	0.6185	0.7649	0.9070	0.8922	0.9242	**0.9307**

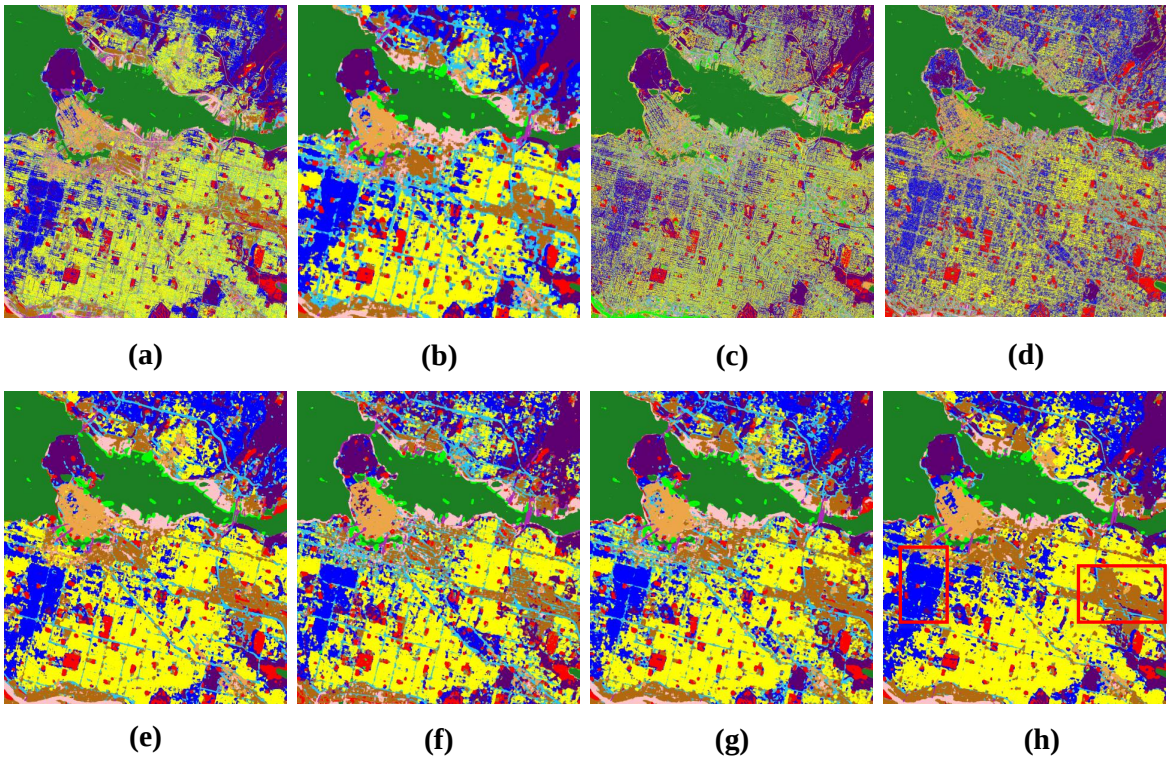

Figure 4. Classification maps with different methods on Level 1B dataset. (**a**) EMAP. (**b**) CAE. (**c**) RNN. (**d**) SCPF-ResNet. (**e**) CNN-MS. (**f**) CNN-PAN. (**g**) SFNet. (**h**) AGCNet.

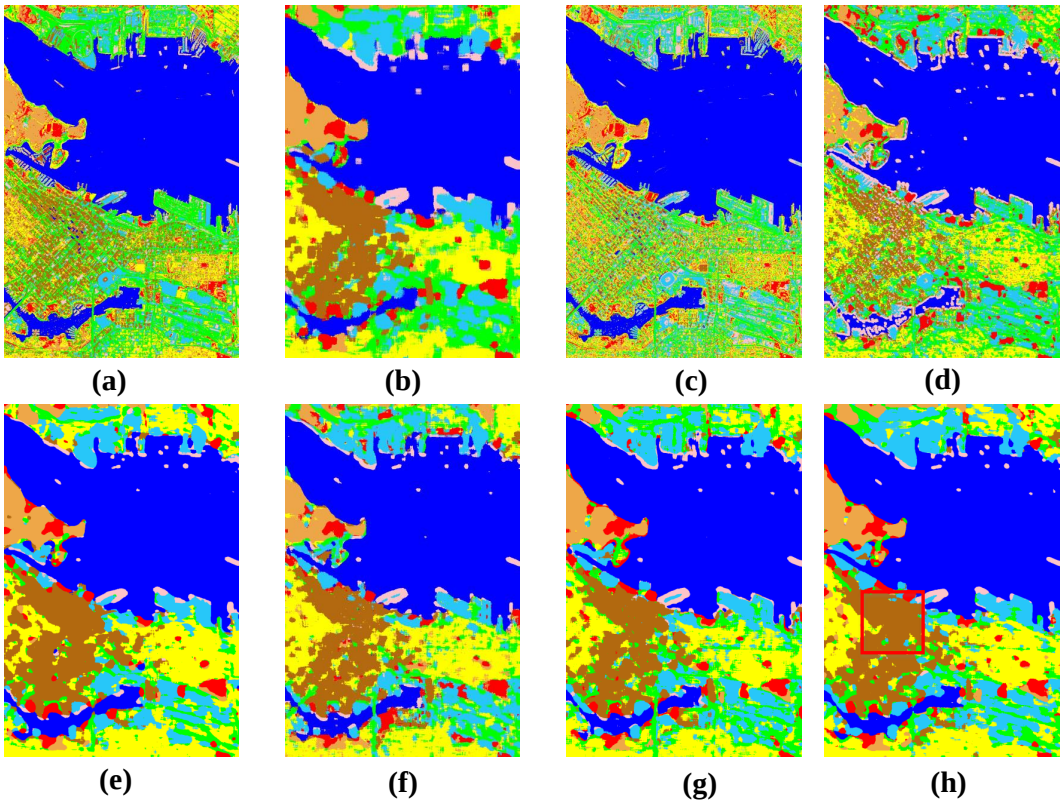

Figure 5. Classification maps with different methods on Level 1C dataset. (**a**) EMAP. (**b**) CAE. (**c**) RNN. (**d**) SCPF-ResNet. (**e**) CNN-MS. (**f**) CNN-PAN. (**g**) SFNet. (**h**) AGCNet.

EMAP is a typical spatial contexture classification model. Although EMAP is a shallow classification model, the classification accuracies are higher than that of RNN. The OA values of EMAP are about 10% and 6% higher than RNN on the two datasets, which illustrates the importance level of spatial information on classification performance.

Spectral-spatial model achieves superior classification performance than the single-spectral and single-spatial models. CAE is an enhanced model of autoencoder (AE), and its implementation considers both the spatial and spectral information. Different from CNN, CAE pays more attentions to image reconstruction rather than classification [56], hence its classification performance is lower than the CNN model. However, compared with RNN and EMAP models, the classification performance has a significantly improved by combining the spectral and spatial information.

CNN model is used to classify the MS and PAN image respectively. Although the quality metrics of CNN-MS are higher than CNN-PAN in all the OA, AA, and Kappa values, CNN-PAN still contains some advantages on the classification maps. The classification map obtained by CNN-PAN method has more detail information, while the classification map obtained by CNN-MS method has better regional consistencies. Therefore, although CNN-MS method obtains higher accuracies than CNN-PAN method in most categories, the CNN-PAN method still achieves a higher classification accuracy on some small object categories, i.e., Class 9 (Bridge) in Level 1B dataset and Class 6 (Road) in Level 1C dataset.

SCPF-ResNet combines the MS and PAN images for classification, but the classification effect is not ideal. The possible reasons are twofold: few training samples and complex network structure. Ref. [48] designed a dual-branch network to improve the classification accuracy of MS and PAN images. However, the designed network introduced a large number of learn-able parameters. Therefore, more training samples were required for effectively training the network. In this experiment, only 100 samples per each class are selected, which may cause ineffective learning with a lower classification accuracy. Therefore, SCPF-ResNet method may not effective for the classification with limited training sample.

SFNet and AGCNet are also designed for MS and PAN image classification. In these two datasets, the classification accuracy of SFNet is only slightly higher than that of CNN-MS and CNN-PAN methods. Therefore, the stacked concentrate is insufficient for feature fusion, and the proposed AGCNet can exploit more effective information from the two images. In most categories, the classification accuracies of AGCNet are higher than that those of CNN-MS, CNN-PAN, and SFNet. As shown in the rectangular areas in Figures 4h and 5h (Building 1 and Building 4 for Level 1B dataset, and Building 1 for Level 1C dataset), the classification maps have better regional consistency for more complex land covers. However, this study also found that the proposed method has some advantages, the classification results of some details are not satisfactory, such as Class 7 (Road) of Level 1B dataset and Cass 6 (Road) of Level 1C dataset. The extraction and fusion of detailed information still needs to be further studied.

2. Experimental Results with Xi'an Suburban and Xi'an Urban Datasets

Different from Level 1B and Level 1C datasets, Xi'an Suburban and Xi'an Urban datasets (Tables 4 and 5) contain more independent objects, especially for the Xi'an Urban dataset, the classification object is single building, rather than building group. The classification results of Xi'an Suburban and Xi'an Urban datasets are shown in Figures 6 and 7, respectively. The proposed AGCNet achieves superior performance on relatively big areas, such as Class 2 (Building 2), Class 5 (Land), and Class 6 (Building 3) of Xi'an Suburban dataset, and Class 5 (Soil), Class 6 (Tree), and Class 7 (Water) of Xi'an Urban datasets. Also shown in Figure 6h, vegetation 1 within the red rectangular area obtains a better classification effect; and in Figure 7h, the building marked in red color are segmented more completely. Therefore, the classification accuracies of these areas are significantly higher than other comparison methods. However, the improvement of accuracy on some small areas is limited, such as Class 4 (Vegetation 2) and Class 7 (Road) of Xi'an Suburban dataset and Class 6 (Shadow) of Xi'an Urban dataset. Therefore, the proposed AGCNet

still achieves a superior classification performance, but there is still room for improvement on small object classification.

Table 4. Classification accuracy on Xi'an Suburban dataset.

Class	EMAP	CAE	RNN	SCPF-ResNet	CNN-MS	CNN-PAN	SFNet	AGCNet
1 (Building1)	0.9986	0.9991	0.9976	0.8636	1.0000	0.9741	0.9990	**1.0000**
2 (Building2)	0.7956	0.8663	0.5506	0.7498	0.9796	0.9911	0.9951	**0.9960**
3 (Vegetation1)	**0.9127**	0.8615	0.8480	0.7185	0.8244	0.8055	0.8780	0.9100
4 (Vegetation2)	0.8808	0.9147	0.8168	0.6860	**0.9488**	0.8726	0.9340	0.9421
5 (Land)	0.9135	0.9742	0.8516	0.5820	0.9944	0.9210	0.9917	**0.9991**
6 (Building3)	0.6717	0.8832	0.7468	0.8954	0.9935	0.9895	**0.9981**	0.9975
7 (Road)	0.5850	0.7072	0.4990	0.4291	**0.9135**	0.8657	0.9014	0.9114
8 (Building4)	0.9820	0.9835	0.9643	0.9545	0.9990	0.9900	**0.9990**	0.9985
OA	0.7495	0.8268	0.6840	0.6448	0.9258	0.8951	0.9334	**0.9448**
Kappa	0.6945	0.7863	0.6169	0.5753	0.9062	0.8675	0.9160	**0.9301**
AA	0.8425	0.8987	0.7843	0.7361	0.9566	0.9262	0.9620	**0.9693**

Table 5. Classification accuracy on Xi'an Urban dataset.

Class	EMAP	CAE	RNN	SCPF-ResNet	CNN-MS	CNN-PAN	SFNet	AGCNet
1 (Building)	0.6603	0.8142	0.4668	0.7607	0.7927	0.7508	**0.8261**	0.8086
2 (Flat land)	0.5485	0.8504	0.6192	0.5739	0.9231	0.8892	0.9220	**0.9274**
3 (Road)	0.7349	0.8786	0.7012	0.6096	0.8969	0.8993	**0.9112**	0.8825
4 (Shadow)	0.8789	**0.9248**	0.7954	0.9190	0.9233	0.8583	0.9079	0.8894
5 (Soil)	0.9318	0.9502	0.8794	0.5169	0.9644	0.8696	0.9564	**0.9750**
6 (Tree)	0.8677	0.8954	0.8169	0.8158	0.8729	0.7952	0.8599	**0.9089**
7 (Water)	0.9300	0.9717	0.8739	0.9106	0.9857	0.9754	0.9654	**0.9944**
OA	0.8076	0.8880	0.7280	0.7261	0.8836	0.8222	0.8843	**0.8994**
Kappa	0.7583	0.8577	0.6620	0.6533	0.8524	0.7759	0.8535	**0.8716**
AA	0.7932	0.8979	0.7361	0.7151	0.9084	0.8626	0.9070	**0.9123**

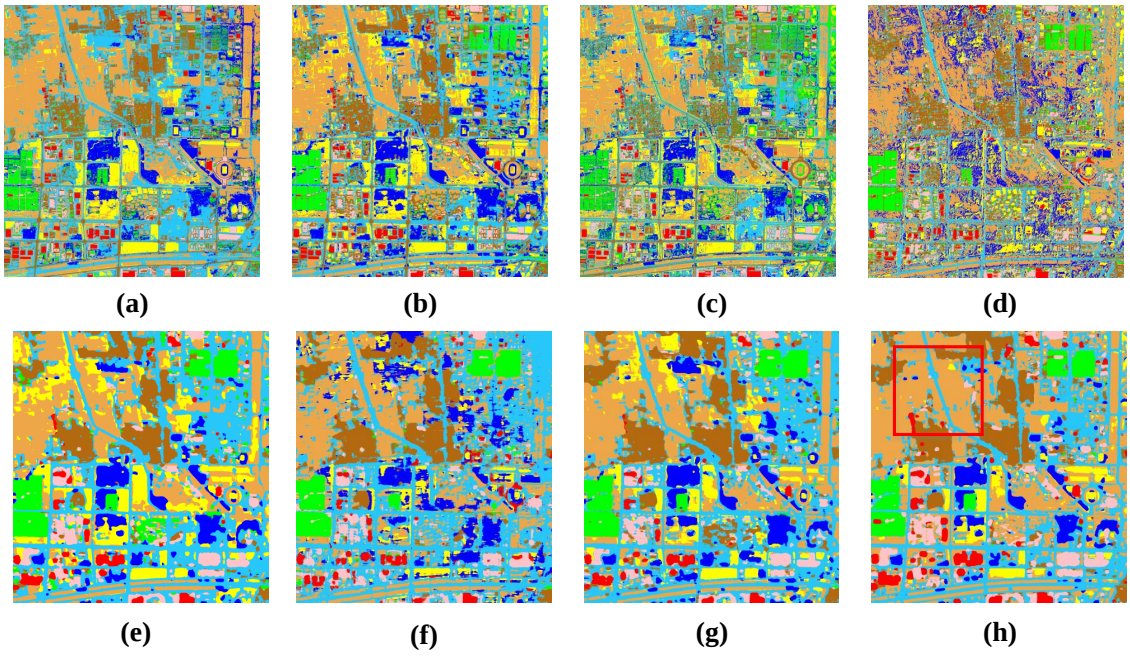

Figure 6. Classification maps with different methods on Xi'an Suburban dataset.(**a**) EMAP. (**b**) CAE. (**c**) RNN. (**d**) SCPF-ResNet. (**e**) CNN-MS. (**f**) CNN-PAN. (**g**) SFNet. (**h**) AGCNet.

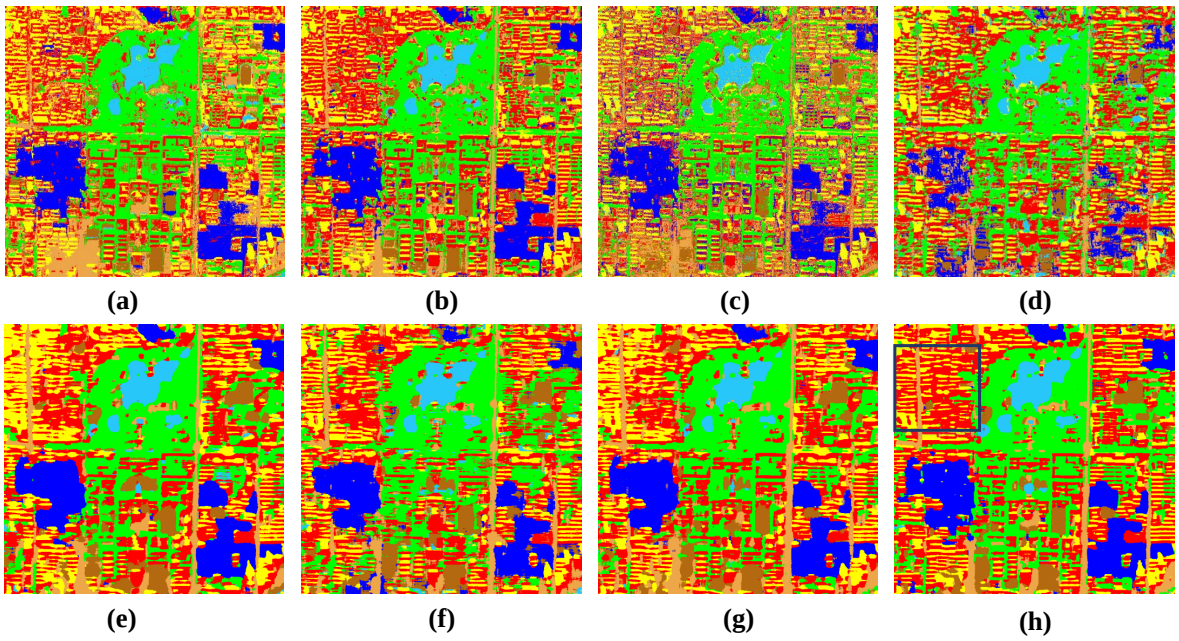

Figure 7. Classification maps with different methods on Xi'an Urban dataset. (**a**) EMAP. (**b**) CAE. (**c**) RNN. (**d**) SCPF-ResNet. (**e**) CNN-MS. (**f**) CNN-PAN. (**g**) SFNet. (**h**) AGCNet.

4.4. Discussion

Performance of SSNet and SCNet. In this section, an ablation study is added to verify the effectiveness of these two parts. The experiments are carried out in the following four steps.

(a) The features of MS and PAN images are extracted by CNN structure respectively, and then fused with an stacked concentrated form (SFNet).

(b) The features of MS and PAN images are extracted with SSNet, and then fused with an attacked concentrated form (SSNet).

(c) The features of MS and PAN images are extracted by CNN structure respectively, and then fused with SCNet (SCNet).

(d) The features of MS and PAN images are extracted with SSNet, and then fused with SCNet (Proposed, AGCNet).

Figure 8 shows the classification accuracy of the four datasets. We can notice that except the Level 1C dataset, the SFNet and SSNet obtain similar classification accuracy on other datasets. Therefore, the SSNet can reduce the trainable parameters of the network without reducing the classification accuracy. In addition, the classification accuracy of SCNet is close to that of AGCNet, therefore, in terms of accuracy improvement, it mainly depends on the adaptive fusion strategy in SCNet. Hence, AGCNet can obtain higher classification accuracy with less time cost.

Learn-able parameter statistics. The purpose of the proposed AGCNet is to better balance the classification performance and time cost. The classification performance is verified in Section 4.3 by comparing with the state-of-the-art methods. In this subsection, the time cost is analyzed by counting the number of learn-able parameters. Table 6 shows the parameter statistics for the five related methods, including CNN-MS, CNN-PAN, SF-Net, SCPF-Net, SGCNet and AGCNet. In addition to the SCPF-ResNet, the other models are as consistent as possible in network parameter setting, e.g., the number of layers and the size of the filter.

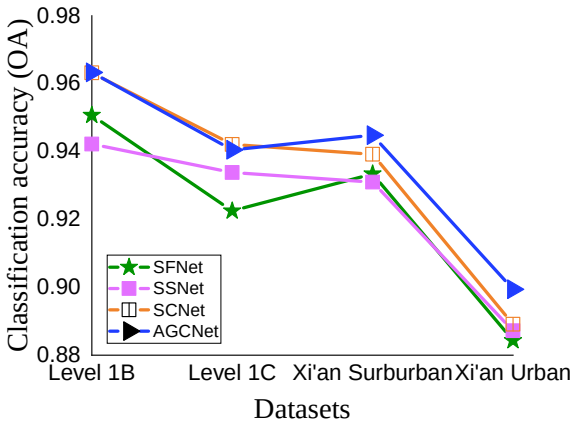

Figure 8. The classification accuracies with different number of training samples.

Since CNN-MS and CNN-PAN are constructed for single-sensor image classification, the number of learn-able parameter is relatively fewer than the other networks. The SFNet, SCPF-ResNet and AGCNet are all designed for combining the MS and PAN images for classification. The SFNet extracts the deep-level features for each sensor image respectively, and then performs a stacked concatenated for feature fusion. Therefore, the number of learn-able parameters in SFNet is about twice that of CNN-MS and CNN-PAN. On the contrary, the parameters of the proposed AGCNet are only slightly increased than the single-sensor image classification networks. Therefore, on the one hand, the statistic results indicate that the SSNet can effectively reduce the network parameters; on the other hand, the SCNet does not introduce a large number of parameters in the calculation of fusion weights. The SCPF-ResNet designs a very complex network structure, leading to an increase in learn-able parameters. The number of learn-able parameters in SCPF-ResNet is about 10 times that of the proposed AGCNet. Therefore, the proposed AGCNet improves the classification accuracy with fewer learn-able parameters.

Table 6. Learn-able parameter statistic.

Methods	CNN-MS	CNN-PAN	SFNet	SCPF-ResNet	AGCNet
The number of learn-able parameters	3.90567×10^5	3.94215×10^5	7.84775×10^5	48.10929×10^5	4.32340×10^5

Performance with different training number. In this subsection, the performance of the proposed AGCNet is investigated by the OA value with the different numbers of training samples, which is shown in Figure 9. The classification accuracy is always highly related with the number of training samples; hence this study performs an analysis on the four datasets by selecting 50, 100, 300, 500, 700, and 900 training samples per class in experiments. Figure 9 compares the OA values obtained by the compared and the proposed methods. The analyses of the experimental results are summarized as follows: (1) The classification methods that only use spectral or spatial information obtain lower classification accuracies, such as RNN and EMAP methods. (2) The classification performance of the SCPF-ResNet is greatly improved with the increased number of training samples; hence the SCPR-ResNet model can achieve higher classification performance with a large number of training samples, but not suitable for the limited training samples. (3) The proposed AGCNet obtains superior classification performance, especially when the number of training samples is fewer.

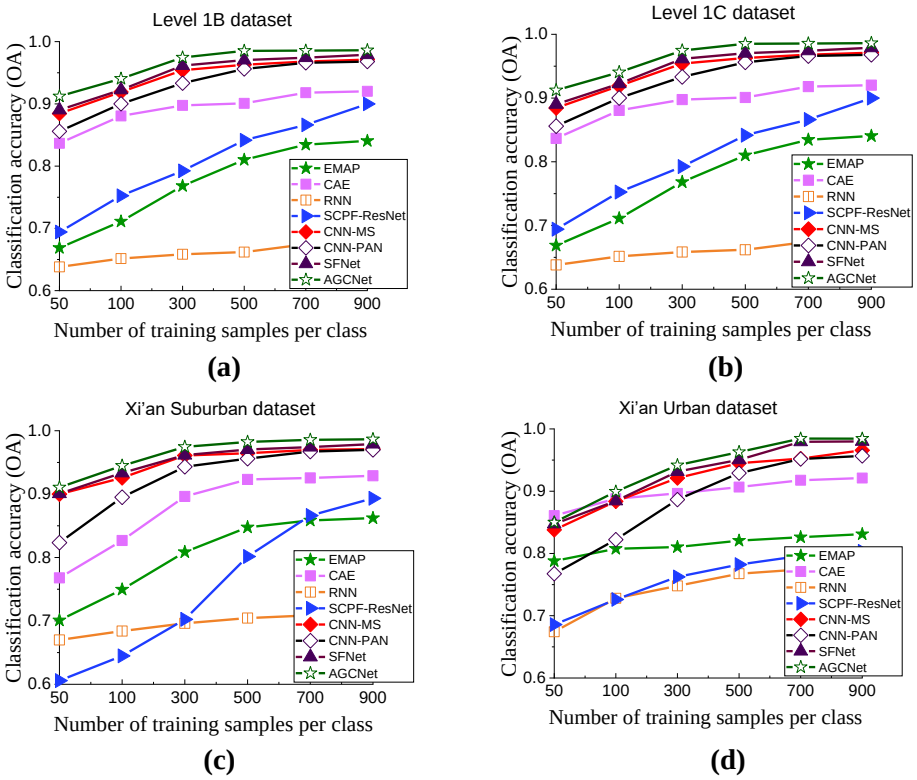

Figure 9. The classification accuracies with different number of training samples.

5. Conclusions and Future Work

In this paper, a lightweight multi-sensor classification network is proposed by combining the channel attention information. The proposed AGCNet mainly consists of a share-split network and selective classification network, which is to better balance the classification performance and time cost. In addition, the network has an end-to-end form and can be trained easily. The experiments are designed to compare the classification performance in two ways which include classification accuracy and time cost. For evaluating the classification accuracy of the proposed AGCNet, seven state-of-the-art methods are used for comparisons on four datasets, including the traditional texture extraction method (i.e., EMAP), spectral or spatial-spectral-based classification methods (i.e., RNN, CAE, CNN-MS and CNN-PAN), and joint multi-sensor classification methods (i.e., SFNet and SCPF-ResNet); the experimental results show that the proposed AGCNet obtains the best performance among all the four datasets. For analyzing the time cost, the learn-able parameters of four related methods are counted for comparisons. The experimental results show that the time cost of the proposed AGCNet is two times less than the SFNet and about ten times less than the SCPF-ResNet. Therefore, the proposed AGCNet is lightweight and effective. The proposed network can be easily extended to other multi-sensor and multi-scale classification, and its effectiveness will be further verified.

Author Contributions: C.S. and Y.D. was primarily responsible for the original idea and experimental design. L.F. contributed to the experimental analysis. Z.L. and H.S. provided important suggestions for improving the quality of the paper. 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 (61902313, 61973250, 61701396, 42101359) and the Natural Science Foundation of Shaan Xi Province (2018JQ4009).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Deimos Imaging for acquiring and providing the data used in this paper, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

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

References

1. Lin, J.; Yu, T.; Mou, L.; Zhu, X.; Wang, Z.J. Unifying top–down views by task-specific domain adaptation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4689–4702. [CrossRef]
2. Lin, J.; Qi, W.; Yuan, Y. In defense of iterated conditional mode for hyperspectral image classification. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14–18 July 2014.
3. Lv., Y.; Liu, T.F.; Benediktsson, J.A.; Falco, N. Land cover change detection techniques: Very-high-resolution optical images: A review. *IEEE Trans. Remote Sens. Mag.* **2021**. [CrossRef]
4. Lv, Z.; Liu, T.; Cheng, S.; Benediktsson, J.A. Local histogram-based analysis for detecting land cover change using VHR remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1284–1287. [CrossRef]
5. Wu, Y.; Huang, M.; Li, Y.; Feng, S.; Wu, D. A Distributed Fusion Framework of Multispectral and Panchromatic Images Based on Residual Network. *Remote Sens.* **2021**, *13*, 2556. [CrossRef]
6. Zhao, W.; Jiao, L.; Ma, W.; Zhao, J.; Zhao, J.; Liu, H.; Cao, X.; Yang, S. Superpixel-Based Multiple Local CNN for Panchromatic and Multispectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4141–4156. [CrossRef]
7. Feng, J.; Li, D.; Gu, J.; Cao, X.; Jiao, L. Deep reinforcement learning for semisupervised hyperspectral band selection. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1–19. [CrossRef]
8. Cheng, S.; Li, F.; Z, L.; M, Z. Explainable scale distillation for hyperspectral image classification. *Pattern Recognit.* **2021**, *122*, 108316.
9. Garcia, N.; Morerio, P.; Murino, V. Learning with privileged information via adversarial discriminative modality distillation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2581–2593. [CrossRef]
10. Zhang, Z.; Li, J.; Shao, W.; Peng, Z.; Zhang, R.; Wang, X.; Luo, P. Differentiable Learning-to-Group Channels via Groupable Convolutional Neural Networks. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.
11. Wang, X.; Kan, M.; Shan, S.; Chen, X. Fully Learnable Group Convolution for Acceleration of Deep Neural Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9041–9050.
12. Howard, A.; Chen, B.; Kalenichenko, D.; Weyand, T.; Zhu, M.; Andreetto, M.; Wang, W. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
13. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2011–2023. [CrossRef]
14. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
15. Lin, J.; Liang, Z.; Li, S.; Ward, R.; Wang, Z.J. Active-learning-incorporated deep transfer learning for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4048–4062. [CrossRef]
16. Lin, D.; Lin, J.; Zhao, L.; Wang, Z.J.; Chen, Z. Multilabel aerial image classification with a concept attention graph neural network. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1–12. [CrossRef]
17. Ma, X.; Wang, H.; Geng, J. Spectral–Spatial Classification of Hyperspectral Image Based on Deep Auto-Encoder. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4073–4085. [CrossRef]
18. Kemker, R.; Kanan, C. Self-Taught Feature Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2693–2705. [CrossRef]
19. Chen, Y.; Zhao, X.; Jia, X. Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2015**, *8*, 2381–2392. [CrossRef]
20. Li, Y.; Xie, W.; Li, H. Hyperspectral image reconstruction by deep convolutional neural network for classification. *Pattern Recognit.* **2017**, *63*, 371–383. [CrossRef]
21. Lu, Y.; Xie, K.; Xu, G.; Dong, H.; Li, C.; Li, T. MTFC: A Multi-GPU Training Framework for Cube-CNN-based Hyperspectral Image Classification. *IEEE Trans. Emerg. Top. Comput.* **2020**. [CrossRef]
22. Hang, R.; Liu, Q.; Hong, D.; Ghamisi, P. Cascaded recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5384–5394. [CrossRef]
23. Lei, T.; Li, L.; Lv, Z.; Zhu, M.; Du, X.; Nandi, A.K. Multi-modality and multi-scale attention fusion network for land cover classification from VHR remote sensing images. *Remote Sens.* **2021**, *13*, 3771. [CrossRef]
24. Wang, D.; Du, B.; Zhang, L.; Xu, Y. Adaptive Spectral-Spatial Multiscale Contextual Feature Extraction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2461–2477. [CrossRef]
25. Zhang, M.; Li, W.; Du, Q. Diverse Region-Based CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2018**, *27*, 2623–2634. [CrossRef]
26. Lin, Z.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063.

27. Zhang, Y.; Liu, K.; Dong, Y.; Wu, K.; Hu, X. Semisupervised Classification Based on SLIC Segmentation for Hyperspectral Image. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1440–1444. [CrossRef]
28. Feng, J.; Yu, H.; Wang, L.; Cao, X.; Zhang, X.; Jiao, L. Classification of Hyperspectral Images Based on Multiclass Spatial–Spectral Generative Adversarial Networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5329–5343. [CrossRef]
29. Lin, J.; Mou, L.; Yu, T.; Zhu, X.; Wang, Z.J. Dual adversarial network for unsupervised ground/satellite-to-aerial scene adaptation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
30. Xu, Y.; Du, B.; Zhang, L.; Cerra, D.; Saux, B. Advanced Multi-Sensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 1709–1724. [CrossRef]
31. Ramirez, J.; Arguello, H. Spectral Image Classification From Multi-Sensor Compressive Measurements. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 626–636. [CrossRef]
32. Hinojosa, C.; Ramirez, J.; Arguello, H. Spectral-Spatial Classification from Multi-Sensor Compressive Measurements Using Superpixels. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 3143–3147.
33. Liu, X.; Jiao, L.; Zhao, J.; Zhao, J.; Zhang, D.; Liu, F.; Yang, S.; Tang, X. Deep Multiple Instance Learning-Based Spatial–Spectral Classification for PAN and MS Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 461–473. [CrossRef]
34. Xu, K.; Huang, H.; Deng, P.; Shi, G. Two-stream feature aggregation deep neural network for scene classification of remote sensing images. *Inf. Sci.* **2020**, *539*, 250–268. [CrossRef]
35. Wang, Z.; Zou, C.; Cai, W. Small Sample Classification of Hyperspectral Remote Sensing Images Based on Sequential Joint Deeping Learning Model. *IEEE Access* **2020**, *8*, 71353–71363. [CrossRef]
36. Feng, J.; Feng, X.; Chen, J.; Cao, X.; Yu, T. Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification. *Remote Sens.* **2020**, *12*, 1149. [CrossRef]
37. Luo, F.; Zhang, L.; Du, B.; Zhang, L. Dimensionality Reduction with Enhanced Hybrid-Graph Discriminant Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5336–5353. [CrossRef]
38. Bell, S.; Zitnick, C.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2847–2883.
39. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
40. Newell, A.; Yang, K.; Deng, J. Stacked hourglass net-works for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
41. Wang, N.; Ma, S.; Li, J.; Zhang, Y.; Zhang, L. Multistage attention network for image inpainting. *Pattern Recognit.* **2020**, *106*, 107448. [CrossRef]
42. Guo, D.; Xia, Y.; Luo, X. Scene Classification of Remote Sensing Images Based on Saliency Dual Attention Residual Network. *IEEE Access* **2020**, *8*, 6344–6357. [CrossRef]
43. Tong, W.; Chen, W.; Han, W.; Li, X.; Wang, L. Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 4121–4132. [CrossRef]
44. Zhu, P.; Tan, Y.; Zhang, L.; Wang, Y.; Wu, M. Deep Learning for Multilabel Remote Sensing Image Annotation with Dual-Level Semantic Concepts. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *58*, 4047–4060. [CrossRef]
45. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 310–314. [CrossRef]
46. Zhao, Z.; Li, J.; Luo, Z.; Li, J.; Chen, C. Remote Sensing Image Scene Classification Based on an Enhanced Attention Module. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1926–1930. [CrossRef]
47. Lin, J.; Yuan, K.; Ward, R.; Wang, Z.J. Xnet: Task-specific attentional domain adaptation for satellite-to-aerial scene. *Neurocomputing* **2020**, *406*, 215–223. [CrossRef]
48. Roy, S.; Chatterjee, S.; Bhattacharyya, S.; Chaudhuri, B.; Platos, J. Lightweight Spectral-Spatial Squeeze-and-Excitation Residual Bag-of-Features Learning for Hyperspectral Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5277–5290. [CrossRef]
49. Zhu, H.; Ma, M.; Ma, W.; Jiao, L.; Hong, S.; Shen, J.; Hou, B. A spatial-channel progressive fusion ResNet for remote sensing classification. *Inf. Fusion* **2020**, *70*, 72–87. [CrossRef]
50. Jin, N.; Wu, J.; Ma, X.; Yan, K.; Mo, Y. Multi-task learning model based on Multi-scale CNN and LSTM for sentiment classification. *IEEE Access* **2020**, *8*, 77060–77072. [CrossRef]
51. Cavallaro, G.; Bazi, Y.; Melgani, F.; Riedel, M. Multi-Scale Convolutional SVM Networks for Multi-Class Classification Problems of Remote Sensing Images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 875–878.
52. Zhang, E.; Liu, L.; Huang, L.; Ng, K. An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery. *Remote Sens. Environ.* **2021**, *254*, 112265. [CrossRef]
53. 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]
54. NYU Computer Science. 2016. Available online: https://cs.nyu.edu/home/index.html (accessed on 1 November 2021).

55. Benediktsson, J.; Palmason, J.; Sveinsson, J. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
56. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Benediktsson, J. Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep: Overview and Toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88. [CrossRef]