


Article

Residual Dense Network Based on Channel-Spatial Attention for the Scene Classification of a High-Resolution Remote Sensing Image

Xiaolei Zhao ¹, Jing Zhang ^{1,2,*} , Jimiao Tian ¹, Li Zhuo ^{1,2} and Jie Zhang ³

¹ Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; zhxl@emails.bjut.edu.cn (X.Z.); tianjm@emails.bjut.edu.cn (J.T.); zhuoli@bjut.edu.cn (L.Z.)

² Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China

³ Institute of Mathematical Geology Remote Sensing, China University of Geosciences, Wuhan 430074, China; zhangjie0130@126.com

* Correspondence: zhj@bjut.edu.cn

Received: 13 May 2020; Accepted: 9 June 2020; Published: 10 June 2020



Abstract: The scene classification of a remote sensing image has been widely used in various fields as an important task of understanding the content of a remote sensing image. Specially, a high-resolution remote sensing scene contains rich information and complex content. Considering that the scene content in a remote sensing image is very tight to the spatial relationship characteristics, how to design an effective feature extraction network directly decides the quality of classification by fully mining the spatial information in a high-resolution remote sensing image. In recent years, convolutional neural networks (CNNs) have achieved excellent performance in remote sensing image classification, especially the residual dense network (RDN) as one of the representative networks of CNN, which shows a stronger feature learning ability as it fully utilizes all the convolutional layer information. Therefore, we design an RDN based on channel-spatial attention for scene classification of a high-resolution remote sensing image. First, multi-layer convolutional features are fused with residual dense blocks. Then, a channel-spatial attention module is added to obtain more effective feature representation. Finally, softmax classifier is applied to classify the scene after adopting data augmentation strategy for meeting the training requirements of the network parameters. Five experiments are conducted on the UC Merced Land-Use Dataset (UCM) and Aerial Image Dataset (AID), and the competitive results demonstrate that our method can extract more effective features and is more conducive to classifying a scene.

Keywords: high-resolution remote sensing image; scene classification; residual dense network; channel-spatial attention

1. Introduction

High-resolution remote sensing images are the basic data of spatial information technologies in geographic information systems, global navigation satellite systems and important basic and strategic information resources of the country [1–6]. With the development of remote sensing satellites, imaging radars and unmanned aerial vehicle technology, remote sensing image further presents the development trend of fine resolution, multi-source, expanded range and quantitative data [7–11]. It has become a research hotspot that analyzing and processing the content of high-resolution remote sensing images in recent years [12–14]. The scene classification of a high-resolution remote sensing image refers to distinguishing specific areas present in the image, such as ocean, land, vegetation, etc., and is widely applied in urban planning, geographic image retrieval and other fields [15–17].

Unlike ordinary natural images, the shooting angle of remote sensing image is top-down so that the remote sensing scene often has complex background and spatial structure [18,19]. In general, the high intra-class differences and inter-class similarity of the scene categories will bring great challenges to the task of remote sensing scene classification [20,21]. As we know, the scene content in remote sensing image has a strong connection with the spatial relationship characteristics [22]. Thus, how to design an effective feature extraction network to fully mine the spatial information of a high-resolution remote sensing image directly determines the quality of classification.

The traditional scene classification approaches mainly adopt the hand-crafted feature extraction with the classifier [23–26]. Common visual features are a color histogram, a gray level co-occurrence matrix (GLCM), scale invariant feature transform (SIFT), a histogram of oriented gradient (HOG), a local binary pattern (LBP) and other manual features [27–29]. Due to the fact that hand-crafted features always produce weaker feature representation, it is difficult to process images with a complex scene. With the emergence of convolutional neural networks (CNNs) that have the powerful feature learning ability, the performance of computer vision tasks has been greatly improved [30–33]. CNN has been also applied into scene classification task of remote sensing image [34–38]. Cheng et al. [39] adopted a variety of deep network structures, such as AlexNet, VGGNet, and GoogLeNet as feature extractors combining metric learning regularization terms to optimize the network, and the scene classification accuracy rated on the Aerial Images Dataset (AID) are 94.47%, 96.89%, and 96.22%, respectively. However, when the deep network is increased to a certain layer, the degradation problem of the gradient disappearance will occur. Furthermore, these classification networks usually only use the last convolutional layer or fully connected layer features of the network for classification, which do not make full use of the information of each convolutional layer in the network so as to ignore the multi-layer features generated by the network. Considering the complex background of a remote sensing image, the variety of scene categories and the different shapes of objects, it is difficult to obtain more detailed and specific image feature information for a single feature. He et al. [40] thought that ResNet could solve the degradation problem of the deep network. Huang et al. [41] proposed the DenseNet to achieve feature reuse with a dense connection for obtaining more complete feature information. Zhang et al. [42] proposed a residual dense network (RDN) for image super-resolution that combines residual connection with the dense connection of ResNet and DenseNet. This connection mode can solve the degradation problem and produce excellent performance by extracting of local dense features and utilizing all layered features. Generally speaking, the existing tasks of remote sensing scene classification are usually to extract features from the whole image. High-resolution remote sensing images have complex spatial feature relationships, and directly extracting the features of the whole image is difficult to reflect main information of the content due to the existence of much redundant information. Human visual attention mechanism scans the global image to obtain the detailed information of the target area that needs to be focused on as well as suppress other useless information. Mnih et al. [43] added an attention mechanism in recurrent neural network (RNN) to reduce the complexity of the image classification task, achieving excellent performance. Wang et al. [44] firstly proposed a novel recurrent attention structure for the scene classification of a remote sensing image, in which a mask matrix is used as the attention weight to multiply the feature map to obtain the attention feature representation of high-level features. But this approach only increases the weight in the spatial dimension, lacking focus on the feature map channel dimension. Woo et al. [45] increased the attention weight in two dimensions of channel and spatial to learn meaningful features in channel-spatial, respectively, which achieved excellent performance in the classification task. In addition, due to the lack of labeled high-resolution remote sensing image data, existing works basically utilize pretrained models as feature extractors that could not significantly improve the structure of classification network, thus limiting the scene classification performance.

In order to solve the above problems, we propose an RDN-based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. First, multi-layer convolutional features are fused with residual dense blocks (RDBs) that are used to form an RDN. Then, the channel-spatial

attention is added after each RDB to obtain more effective feature representation. Finally, softmax classifier is applied to classify the scene after adopting data augmentation strategy for meeting the training requirements of the network parameters.

The main contributions of this article are summarized as follows:

(1) Due to the complex background of the remote sensing image, the variety of scene categories, and the rich scene feature information, multiple RDBs are used to form an RDN to fuse multi-layer convolutional features, including dense connection and residual connection. The input of each layer in the RDB is determined by the output of all the previous layers, and the output will be transmitted to each subsequent layer. In this way, the information of each layer's feature is fully utilized by fusing the multi-layer features generated with RDN to obtain more specific and detailed features in the remote sensing image.

(2) With the complex spatial information of a remote sensing image, there is a large amount of redundant information in the whole image. In order to extract refined features and enhance feature representation capability, channel-spatial attention is added after each RDB by focusing on the useful features of image in both channel and spatial dimensions so as to remove redundant information. Besides that, the channel-spatial attention module can also reduce the computational effort and further improve classification performance.

(3) It is not enough to meet the training requirements of a large number of parameters of the deep network as the amount of labeled remote sensing image data is small. The data augmentation strategy is used to rotate and mirror the image to expand the training data, which is more conducive to network parameter training.

The rest of this paper is organized as follows: Section 2 states the overview of our proposed method. Section 3 illustrates the deep feature extraction of high-resolution remote sensing image, including the structure of RDB and the channel-spatial attention module. Section 4 introduces the scene classification method of a remote sensing image consist of classifier and dataset augmentation strategy. In Section 5, the experimental results of the overall accuracy and confusion matrix are computed and compared to verify the effectiveness of our proposed method. The discussion is presented in Section 6. Finally, some beneficial conclusions have been drawn in Section 7.

2. Structure of the Proposed Residual Dense Network Based on Channel-Spatial Attention

The overall framework of the proposed RDN based on channel-spatial attention is shown in Figure 1. Based on the structure of DenseNet121 [41], the proposed overall framework includes four residual dense blocks (RDB₁, RDB₂, RDB₃, RDB₄), a channel-spatial attention module (CBAM) and softmax classifier. The input image size is 224px × 224px, through convolution, maxpool operation and four RDBs, multilayer features are fused to obtain detailed feature information. The CBAM is added after each RDB. In this way, the network can obtain more effective features by refining features and removing redundant information. Finally, the obtained deep features are pooled by the global avgpool operation, then input to the softmax classifier for classification, and finally get the probability P_1, P_2, \dots, P_n of each category.

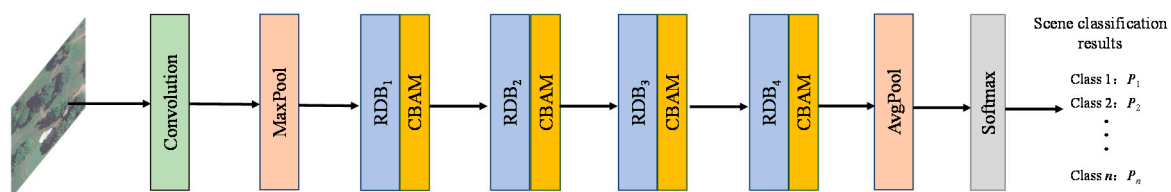


Figure 1. Structure of the residual dense network (RDN) based on channel-spatial attention (RDN+ channel-spatial attention module (CBAM)).

3. Deep Feature Extraction of a High-Resolution Remote Sensing Image

Feature extraction plays an important role in improving the performance of scene classification. This section will introduce the RDN + CBAM to extract deep features of high-resolution remote sensing images. First, RDBs in RDN are used for multilayer feature fusion to make full use of the information of each layer of features for obtaining more specific and detailed features in a high-resolution remote sensing image. Then, a CBAM is added after each RDB to focus on the important features of the image in channel and spatial dimensions and remove redundant information. Next, we will describe the above process in detail.

3.1. Residual Dense Blocks for Multilayer Feature Fusion

The RDN is composed of four residual dense blocks (RDBs). Driven by the idea of RDN [42], we improve the network according to the dense connection block in DenseNet121, and add local residual connection at the end of the dense connection block. Our RDB shown in Figure 2 is composed of bottleneck layers, transition layer, and local residual connection.

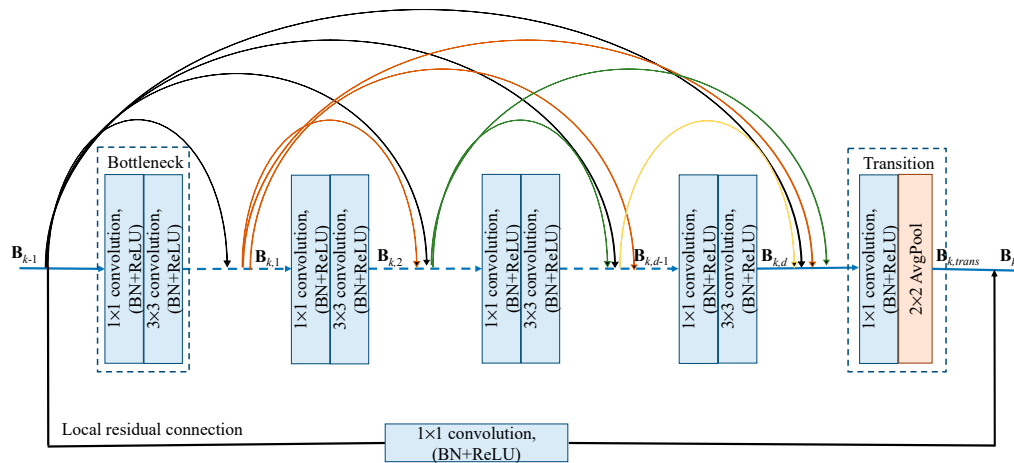


Figure 2. Structure of residual dense block.

(1) Bottleneck Layers

The bottleneck layers consist of 1×1 and 3×3 convolution operations. After each convolution operation, a batch normalization (BN) layer and activation functions of a rectified linear units (ReLU) are added. The input of each bottleneck layer is the superposition of the outputs of all the previous bottleneck layers. The output of each bottleneck layer will be used as the input of all the bottleneck layers afterwards. The features produced by all convolutional layers are connected with dense connection to fuse the multilayer features. These connections can make full use of the feature information of each layer to obtain specific semantic features. We set the number of bottleneck layers in four RDB as 6, 12, 24, 16, respectively, the same as the DenseNet121. Let \mathbf{B}_{k-1} and \mathbf{B}_k be the output of the k th RDB, $\mathbf{B}_{k,d}$ is the output of the d th bottleneck layers in the k th RDB.

$$\mathbf{B}_{k,d} = \Phi([\mathbf{B}_{k-1}, \mathbf{B}_{k,1}, \dots, \mathbf{B}_{k,d-1}]) \quad (1)$$

where Φ is the operation of the convolution, BN and ReLU, $[\mathbf{B}_{k-1}, \mathbf{B}_{k,1}, \dots, \mathbf{B}_{k,d-1}]$ is the connect operation of the feature maps produced by the $k-1$ th RDB and $1, \dots, (d-1)$ th bottleneck layers in k th RDB. The feature map channel of $\mathbf{B}_{k,d}$ is $G_0 + G \times (d-1)$, G_0 is the feature map channel of \mathbf{B}_{k-1} , G is the growth rate of each bottleneck layer.

(2) Transition Layer

After stacking multiple bottleneck layers, the final number of output feature map channels is the sum of the number of channels that all layers generated. In order to reduce the number of

channels of the final feature map, we add transition layers after stacked bottleneck layers, consisting of a 1×1 convolution and avgpool operations, which can reduce the number of feature map channels after the stacked bottleneck layers by half, thereby reducing the amount of subsequent calculations. The feature map of the output of the transition layer is follow as:

$$\mathbf{B}_{k,trans} = \Psi([\mathbf{B}_{k-1}, \mathbf{B}_{k,1}, \dots, \mathbf{B}_{k,d-1}, \dots, \mathbf{B}_{k,d}]) \quad (2)$$

where $\mathbf{B}_{k,trans}$ is the output of transition layer, Ψ is the operation of convolution and avgpool in transition layer.

(3) Local Residual Connection

In order to further ensure that the input RDB information is not lost and improve the information flow of the network, a local residual connection is added between the input of the RDB and the output of the transition layer to better retain the characteristics of the front layer. This skip connection method can solve the problem of the gradient disappearance in deep network, and the fusion of local features of dense connected blocks will be achieved to ensure that the network can extract more accurate features. In order to retain the consistency of the input of RDB and the output of the transition layer, the local residual connection included is a 1×1 convolution, and the output \mathbf{B}_k of k th RDB can be denoted as:

$$\mathbf{B}_k = \mathbf{B}_{k-1} + \mathbf{B}_{k,trans} \quad (3)$$

In the designed network, four RDBs are connected in series, the output of the previous RDB is used as the input of the next RDB, and the obtained multilayer features are passed to the next RDB. The high-resolution remote sensing images always have a complex background and rich scene feature information. Our designed method can enable the network to capture the feature information of each convolutional layer of the high-resolution remote sensing image and merge the local features of RDB to improve the feature propagation efficiency and the features utilization rate. That is to say, it is possible to extract more detailed features as a result of improving the overall performance of scene classification.

3.2. Channel-Spatial Attention Module for Feature Refinement

If fusing multilayer features with RDB, the extracted information will have a lot of redundancy. The attention mechanism plays an important role in the human visual system so that the network can selectively focus on the prominent parts to more effectively extract the important features from a high-resolution remote sensing image while removing redundant information. As mentioned above, Woo et al. [45] proposed the convolutional block attention module (CBAM) to extract meaningful features in two dimensions of channel and spatial, respectively, in order to capture important information and perform adaptive feature refinement well. This module can be embedded in any network to enhance the ability of network feature learning. For a high-resolution remote sensing image with complex spatial information, the CBAM can capture the important information of the remote sensing image in both channel and spatial dimensions and suppress unimportant information that is more conducive to extracting key features. Therefore, we add a CBAM after each RDB to refine the features and enhance the feature representation ability, see Figure 3.

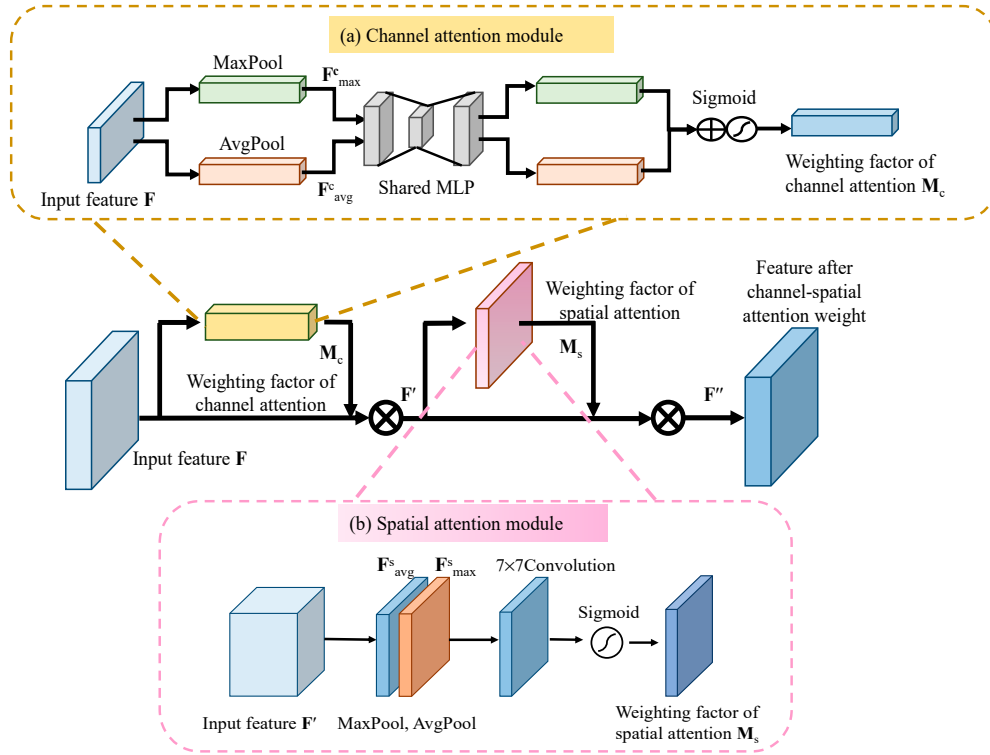


Figure 3. Channel-spatial attention module.

(1) The channel attention process is shown in part (a) of Figure 3.

Step 1: Let the input feature be F of $H \times W \times C$ dimension. F is passed through global AvgPool and MaxPool, respectively, to obtain two $1 \times 1 \times C$ channel descriptions for F^c_{avg} and F^c_{max} .

Step 2: The channel descriptions of F^c_{avg} and F^c_{max} , enter a shared network (MLP) with a hidden layer and multilayer perceptron. The activation size of the hidden layer is $R^{C/r \times 1 \times 1}$, in which r is the compression ratio.

Step 3: The two features obtained after shared network are added, and then the weight coefficient M_c is obtained through the sigmoid activation function.

Step 4: Multiply the weighting factor M_c with the original feature F to obtain the new feature F' after the channel attention. The weighting factor of channel attention M_c can be described as:

$$\begin{aligned} M_c(F) &= \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \quad (4)$$

where σ is the function of sigmoid, $W_0 \in R^{C/r \times C}$, $W_1 \in R^{C \times C/r}$, the MLP weight, W_0 and W_1 are shared for both inputs, the ReLU activation function is followed by W_0 .

(2) The spatial attention process is shown in part (b) of Figure 3.

Step 1: Let the input feature be F' of $H \times W \times C$ dimension. F' is passed through AvgPool and MaxPool operations of one channel dimension to obtain two $H \times W \times 1$ channel description F^s_{avg} and F^s_{max} , and splice the two descriptions together according to the channel.

Step 2: The weighting factor of spatial attention M_s is obtained by a 7×7 convolution and sigmoid activation function.

Step 3: The weighting factor M_s and F' are multiplied to obtain the feature F'' after spatial attention weighting. The weighting factor of spatial attention M_s is as follows:

$$\begin{aligned} M_s(F') &= \sigma(f^{7 \times 7}([\text{AvgPool}(F'); \text{MaxPool}(F')])) \\ &= \sigma(f^{7 \times 7}([F^s_{avg}; F^s_{max}])) \end{aligned} \quad (5)$$

where σ is the function of sigmoid, $f^{7 \times 7}$ denotes the 7×7 convolution operation.

The authors of [45] have proved the order that channel attention first and spatial attention last in a sequential manner can achieve the best performance. By this design, we follow the same arrangement as shown in Figure 3, in which CBAM represents the channel-spatial attention module, which is embedded in each RDB to form an RDN + CBAM.

Through channel-spatial attention, important information can be focused on two dimensions of channel and spatial in high-resolution remote sensing image. It can suppress useless information and remove redundant information in each RDB to perform feature refinement. In this way, the network can extract more effective features of the remote sensing scene as well as drop the amount of network calculation to further improve the performance of scene classification.

4. Scene Classification of High-Resolution Remote Sensing Image

After extracting deep features using the RDN + CBAM, the extracted features are then input to the classifier for scene classification. Due to the fact that a sufficient amount of data is conducive to the training of deep network parameters and softmax classifier can handle multi-classification problems well, we use the data augmentation strategy and softmax classifier to complete the scene classification task of a high-resolution remote sensing image. We will specifically explain these two parts as follow.

4.1. Data Augmentation

Deep learning-based classification networks require a large amount of training data to train complex network structures to achieve the best classification accuracy. A high-resolution remote sensing image are not enough to achieve the training requirements of deep network parameters, as they are difficult to obtain a large number of labeled images. In addition, high-resolution remote sensing image often have multiple directions in the scene due to its characteristics of overhead shooting, so it is difficult to make the features obtained by the network robust to rotation. In order to solve the problem of small amount of data of labeled remote sensing image and enhance the rotation invariance of network learning features, we use data augmentation strategy. We rotate the remote sensing image to 0° , 90° , 180° , and 270° , respectively, and then perform mirror operations on these four angles to expand the data to eight times of the original data, which is beneficial to the training of network parameters, on the other hand, the features learned by the network have rotation invariance.

4.2. Scene Classification

We design the last layer of the network as the fully connected layer in which the number of nodes is the total number of categories and use the softmax as the classifier. The extracted deep features by the network are put into the softmax classifier for the scene classification of a high-resolution remote sensing image, and output is the scene category. Softmax is a classifier used to solve multi-classification problems. It outputs the probability that the scene belongs to each category. Each probability value is between 0~1. The highest probability is the final category of the scene. Let the array of softmax input be x , x_i is the element of the i th element in x , and the class probability of x_i pass through softmax is:

$$P_i = \frac{e^{x_i}}{\sum_{j=1}^N e^{x_j}} \quad (6)$$

where P_i is the probability of class i th, and N is the total number of categories.

5. Experimental Results

In this section, we employ the UC Merced Land-Use Dataset (UCM) and Aerial Image Dataset (AID) to evaluate our proposed methods, and compare the performance of scene classification with other state-of-the-art algorithms.

5.1. Experiment Details

A. Dataset Description

We apply the UC Merced Land-Use Dataset (UCM) and Aerial Image Dataset (AID) that are commonly used on remote sensing scene classification to demonstrate the performance of the proposed method. The characteristics of these two datasets are described in detail below.

(1) UC Merced Land-Use Dataset (UCM)

The UC Merced Land-Use Dataset [28] is the first public dataset obtained from a real remote sensing image. These images are selected from manually extracted aerial orthophotos and downloaded from the United States Geological Survey National Map. It contains 21 scene categories that each consists of 100 remote sensing images and size is $256\text{px} \times 256\text{px}$. The resolution of each pixel space is 0.3 meters in the RGB color space. Some example images are shown in Figure 4. The classification of the UCM dataset is challenging because of the high inter-class similarity among categories such as medium residential and dense residential areas. It is widely used in high-resolution remote sensing scene classification tasks.



Figure 4. UC Merced Land-Use Dataset.

(2) Aerial Image Dataset (AID)

The Aerial Image Dataset [46] is a large-scale dataset for aerial images classification obtained from Google Earth. It consists of almost 10,000 images with a size of $600\text{px} \times 600\text{px}$ and the pixel resolution is 8–0.5 m. The dataset includes 30 scene categories and the number of images in each category is about 220–420. Some example images are shown in Figure 5. There is a strong similarity between various types in the dataset, and it was taken in different regions and at different times, which makes the differences within the categories larger.

B. Experiment Setup

We use the above-mentioned data augmentation method to expand to eight times the original dataset. After augmenting the UCM and AID datasets, we use 60% for training, 20% for validation, 20% for testing. The images are cropped into $224\text{px} \times 224\text{px}$ to be suitable for design of network requirements. The overall accuracy and the confusion matrix are used as performance evaluation metrics for scene classification results. The overall accuracy is defined as the number of correctly classified images divided by the total number of test images, which is a valuable measure to reveal the classification method performance on the whole test images. The value of overall accuracy is in the range of 0 to 1, and a higher value indicates a better classification performance. The confusion matrix is an informative table that can allow direct visualization of the performance on each class and can be used for easily analyzing the errors and confusion between different classes, in which the column represents the instances in a predicted class and the row represents the instances in an actual class.

Thus, each item x_{ij} in the matrix is the proportion of images that are predicted to be the i th class while truly belonging to the j th class.

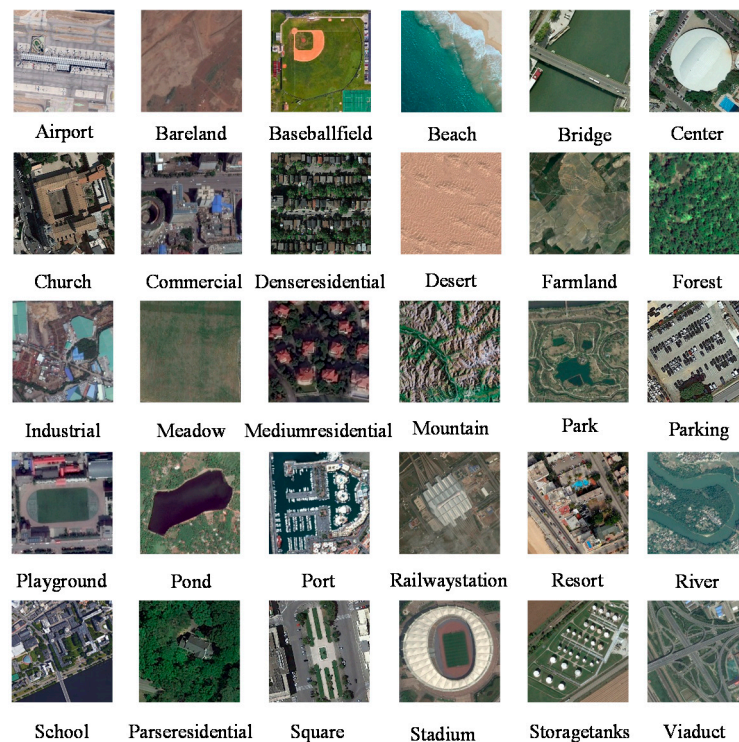


Figure 5. Aerial Image Dataset.

We use the Ubuntu16.04 system in our experiment and GPU is NVIDIA TITAN Xp. The experiment is implemented using Pytorch and is optimized with Adam. The initialization parameters of the dense connection are pretrained module of DenseNet121 on ImageNet. The other network parameters are randomly initialized on the Gaussian distribution. The hyperparameter settings used by the network are shown in Table 1.

Table 1. Hyperparameter settings of the network.

Epoch	Learning Rate	Batch Size	Weight Decay
300(UCM)	0.01 (0–200 epoch)	64	0.0002
400(AID)	0.001 (200–400 epoch)		

5.2. Experiment I: Effect of the Combination of Residual Connection and Dense Connection

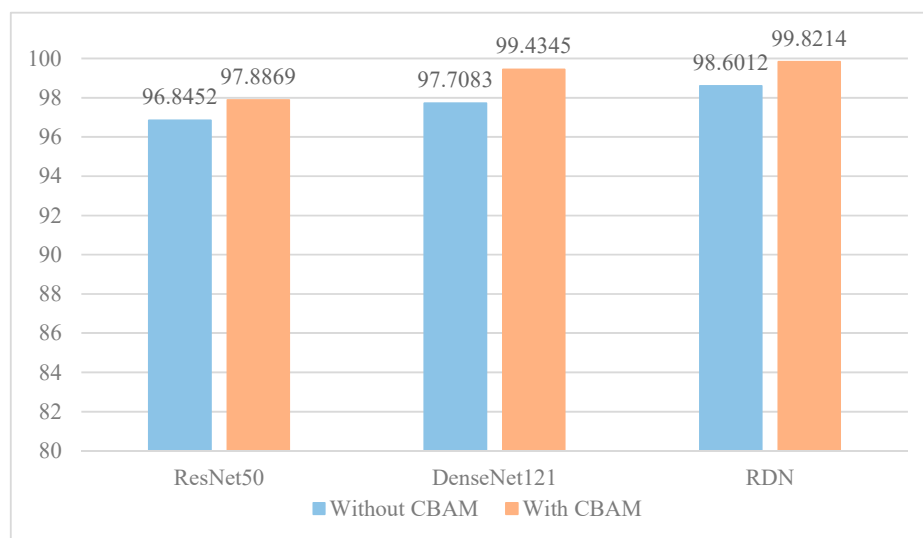
To illustrate the effectiveness of the combination of residual connection and dense connection, we use UCM and AID datasets to conduct experiments on ResNet50, DenseNet121 and the RDN. Table 2 shows the overall accuracy of the two datasets on these three networks. It can be seen that the overall accuracy of the combination of residual connection and dense connection is higher than that of individual residual connection and dense connection. The classification accuracy of RDN on the UCM dataset is 1.756% and 0.8929% higher than ResNet50 and DenseNet121, respectively. For the AID dataset, the classification accuracy of RDN makes an increase of 7.0063% and 2.225% points over ResNet50 and DenseNet121, respectively. RDN makes full use of multilayer features using dense connection, and fuses local features and preserves the integrity of information adopting residual connection. It can confirm that the advantages of residual connection and dense connection are combined to obtain more detailed features than single use, which is helpful to improve the classification performance.

Table 2. Overall accuracy of ResNet50, DenseNet121 and RDN on the UC Merced Land-Use Dataset (UCM) and the Aerial Image Dataset (AID).

Network	UCM (%)	AID (%)
ResNet50	96.8452	87.1500
DenseNet121	97.7083	91.9313
RDN (ours)	98.6012	94.1563

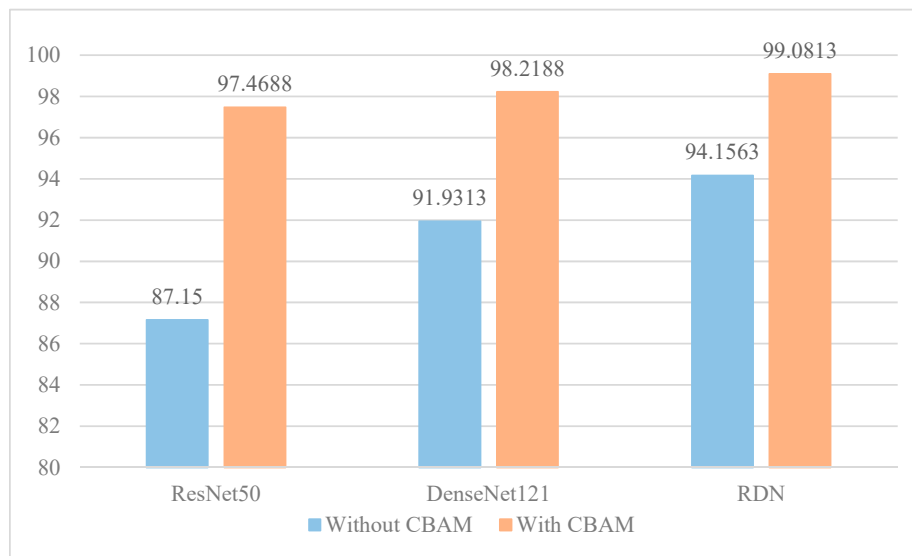
5.3. Experiment II: Effect of the Channel-Spatial Attention Module

In order to demonstrate the effect of the channel-spatial attention module (CBAM) on improving classification performance, we apply the UCM and AID datasets to compare the performance of ResNet50, DenseNet121 and RDN with and without CBAM. As we can see in Figure 6, the performance of the three networks has been greatly improved after adding channel-spatial attention module. For the UCM dataset (Figure 6a), the classification accuracy is improved by 1.0417%, 1.7262%, and 1.2202% on ResNet50, DenseNet121 and RDN after adding CBAM. As for AID datasets (Figure 6b), the classification accuracy is improved by 10.3188%, 6.2875%, and 4.925% on these three networks after adding CBAM. For more the difficult AID dataset, the classification performance has been significantly improved after adding CBAM. The highest accuracy is obtained on two datasets after adding CBAM, with 99.8214% and 99.0813%, respectively, since the more detailed features are acquired by RDN. The experimental results show that CBAM plays a great role in extracting the effective features of remote sensing image and improving the performance of network classification. The features can be focused on two dimensions of channel and spatial. In this way, important features are concerned and the unimportant parts are suppressed.



(a)

Figure 6. Cont.



(b)

Figure 6. Overall accuracy of ResNet50, DenseNet121, RDN on UCM and AID datasets with and without CBAM: (a) UCM dataset; (b) AID dataset.

5.4. Experiment III: Effect of Dataset Augmentation

In order to evaluate the impact of the augmented datasets on classification performance, we conduct the original datasets and the augmented datasets for UCM and AID on the proposed RDN + CBAM. The experimental result is shown in Figure 7. After the augmentation strategy, the overall accuracy of scene classification is increased by 9.4196% and 12.1765% on UCM and AID, respectively. It is illustrated that the amount of augmented data is large enough to meet the training requirements of network parameters and can make the network learn rotation robust features. Therefore, it is necessary to conduct an augmentation strategy if there is not a pre-trained classification model.

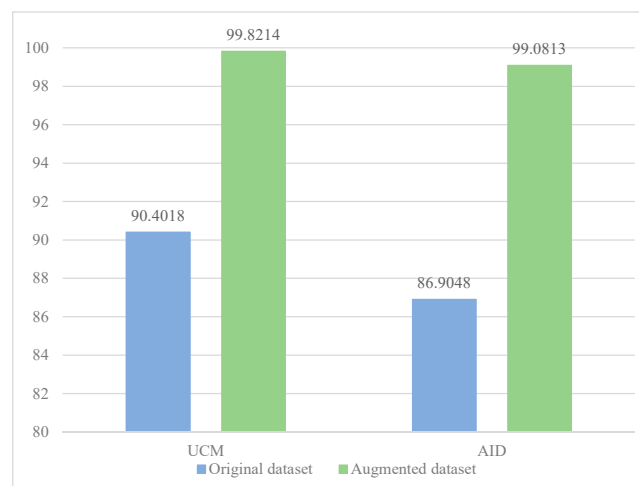


Figure 7. Overall accuracy of RDN + CBAM on the UCM and AID of the original dataset and the augmented dataset.

5.5. Experiment IV: Comparison of the Scene Classification Accuracy of Each Category in Different Networks

In order to specifically analyze the scene classification of each category, we compare the classification accuracy of each scene on RDN + CBAM, RDN, ResNet50 and DenseNet121. The

proposed method RDN + CBAM obviously outperforms the other three networks in accuracy of each category. The result on UCM and AID are shown in Figure 8 and Table 3, Table 4. It can be seen that RDN + CBAM has achieved a classification accuracy of more than 99% in each category on the UCM dataset, which can well distinguish baseball diamond, mediumresidential, storage tanks, overpass, airplane, mobile home park and intersection. These scenes have reached 100% classification accuracy, while the other three networks usually get much lower accuracy. For the AID dataset, our proposed method achieves 100% accuracy for beach, church, mediumresidential, bridge, sparseresidential, stadium, pond, denseresidential, railway station, port and meadow while the accuracy of the other three networks is still low. Compared with the other three networks, the proposed network combines multiple convolutional layer features and local features, and adds a channel-spatial attention module. It can extract more distinguished features for scenes with high inter-class similarity and intra-class difference, so that further improving the classification performance.

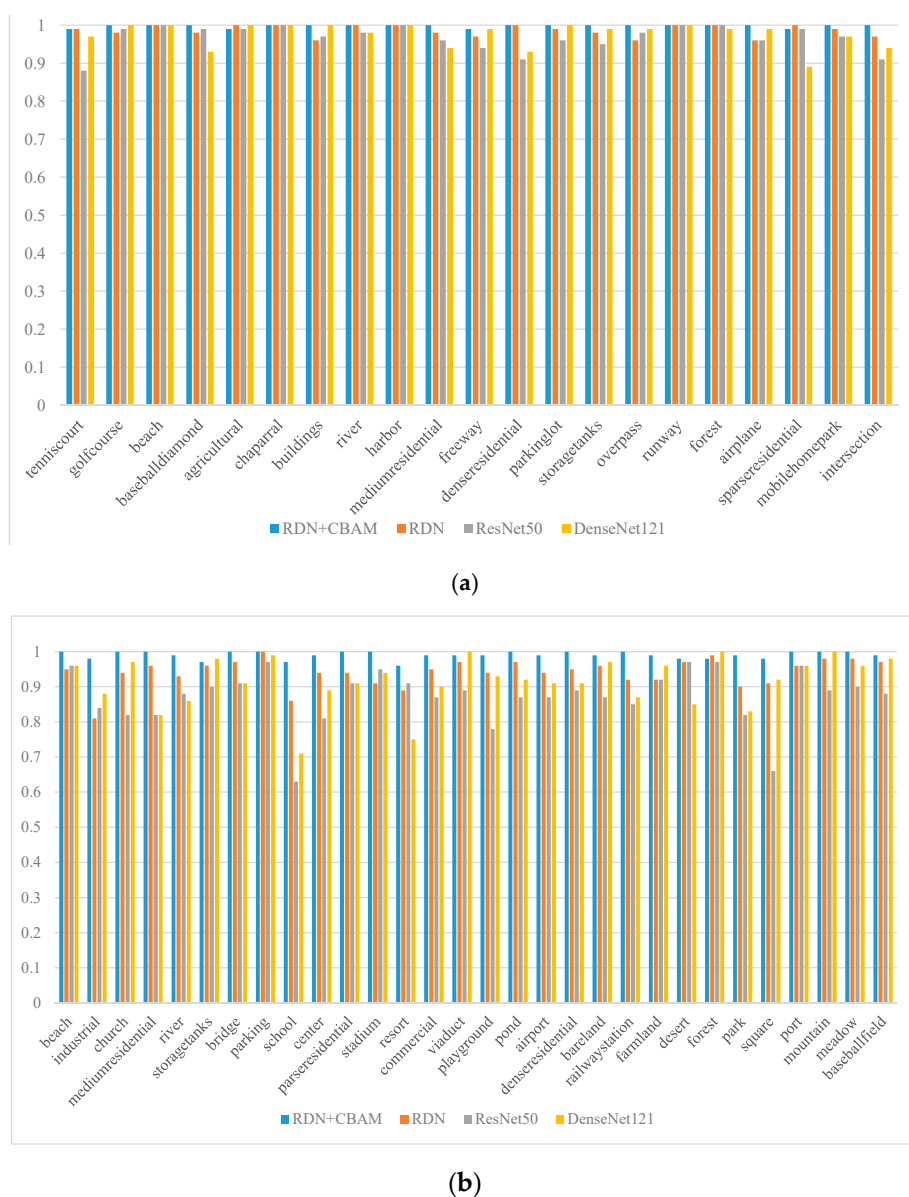


Figure 8. Classification accuracy in each category of RDN + CBAM, RDN, ResNet50, DenseNet121 on UCM and AID: (a) UCM dataset; (b) AID dataset.

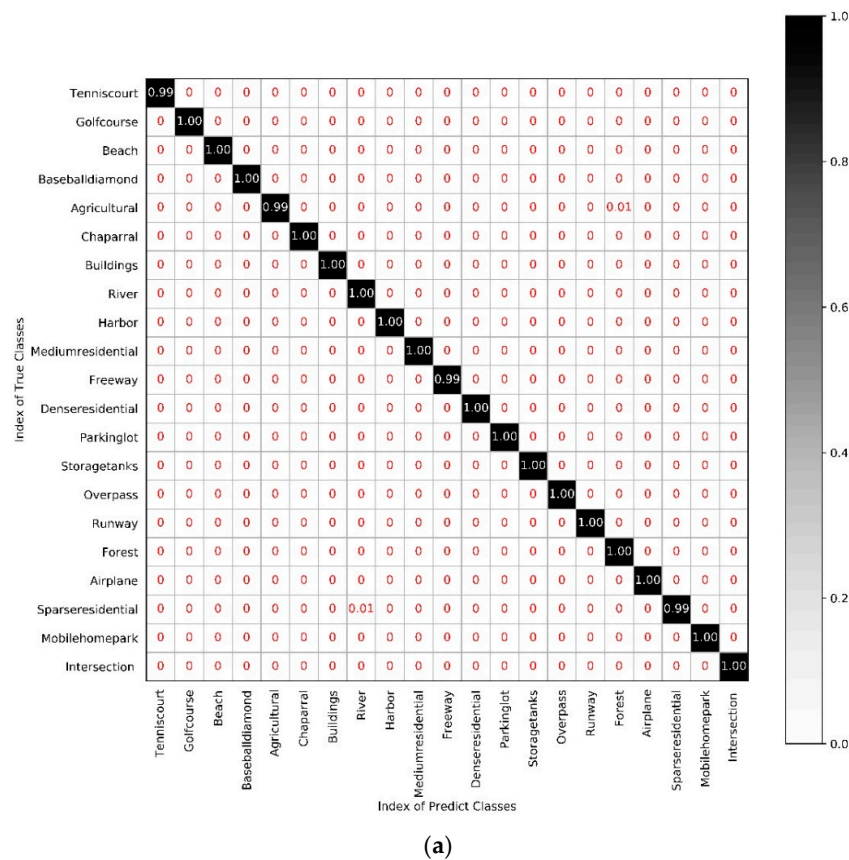
Table 3. Classification accuracy in each category of RDN + CBAM, RDN, ResNet50, DenseNet121 on the UCM dataset.

Class	RDN + CBAM	RDN	ResNet50	DenseNet121
Tenniscourt	0.99	0.99	0.88	0.97
Golfcourse	1	0.98	0.99	1
Beach	1	1	1	1
Baseballdiamond	1	0.98	0.99	0.93
Agricultural	0.99	1	0.99	1
Chaparral	1	1	1	1
Buildings	1	0.96	0.97	1
River	1	1	0.98	0.98
Harbor	1	1	1	1
Mediumresidential	1	0.98	0.96	0.94
Freeway	0.99	0.97	0.94	0.99
Denseresidential	1	1	0.91	0.93
Parkinglot	1	0.99	0.96	1
Storagetanks	1	0.98	0.95	0.99
Overpass	1	0.96	0.98	0.99
Runway	1	1	1	1
Forest	1	1	1	0.99
Airplane	1	0.96	0.96	0.99
Sparseresidential	0.99	1	0.99	0.89
Mbilehomepark	1	0.99	0.97	0.97
Intersection	1	0.97	0.91	0.94

Table 4. Classification accuracy in each category of RDN + CBAM, RDN, Resnet50, Densenet121 on the AID dataset.

Class	RDN +CBAM	RDN	ResNet50	DenseNet121
Beach	1	0.95	0.96	0.96
Industrial	0.98	0.81	0.84	0.88
Church	1	0.94	0.82	0.97
Mediumresidential	1	0.96	0.82	0.82
River	0.99	0.93	0.88	0.86
Storagetanks	0.97	0.96	0.9	0.98
Bridge	1	0.97	0.91	0.91
Parking	1	1	0.97	0.99
School	0.97	0.86	0.63	0.71
Center	0.99	0.94	0.81	0.89
Parseresidential	1	0.94	0.91	0.91
Stadium	1	0.91	0.95	0.94
Resort	0.96	0.89	0.91	0.75
Commercial	0.99	0.95	0.87	0.9
Viaduct	0.99	0.97	0.89	1
Playground	0.99	0.94	0.78	0.93
Pond	1	0.97	0.87	0.92
Airport	0.99	0.94	0.87	0.91
Denseresidential	1	0.95	0.89	0.91
Bareland	0.99	0.96	0.87	0.97
Railwaystation	1	0.92	0.85	0.87
Farmland	0.99	0.92	0.92	0.96
Desert	0.98	0.97	0.97	0.85
Forest	0.98	0.99	0.97	1
Park	0.99	0.9	0.82	0.83
Square	0.98	0.91	0.66	0.92
Port	1	0.96	0.96	0.96
Mountain	1	0.98	0.89	1
Meadow	1	0.98	0.9	0.96
Baseballfield	0.99	0.97	0.88	0.98

We also make a confusion matrix to further analyze the effect of RDN + CBAM on UCM and AID datasets, as shown in above Figure 9. As can be seen from Figure 9a, for the UCM dataset, the classification results of the 17 categories have reached 100%, 1% of the agriculture scene is incorrectly classified as forests, and 1% of the sparseresidential is incorrectly classified as river, because they have the same morphological distribution and are relatively close in color. As can be seen from Figure 9b, for the AID dataset, the classification results of the 27 categories have reached more than 98%, but 2% of the resort is incorrectly classified as denseresidential, and 2% of the desert is incorrectly classified as bareland. The similarities of these scenes are greater, so there are still misclassifications. According to the result of the confusion matrix, the proposed method has achieved excellent performance, which can prove that the method can generally distinguish the remote sensing scene categories with higher complexity.



(a)

Figure 9. Cont.

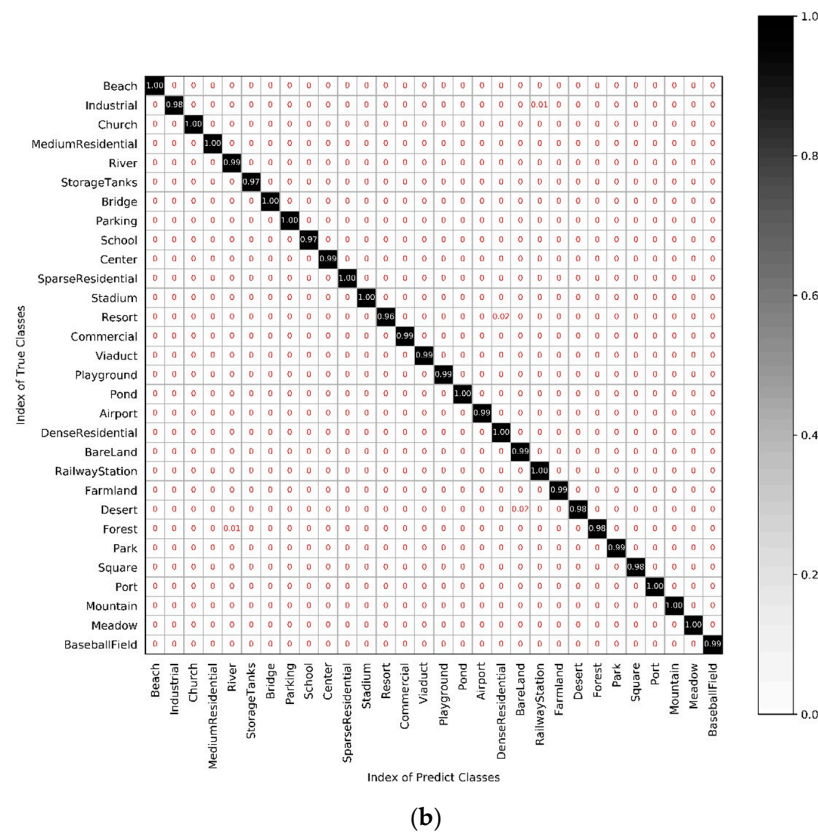


Figure 9. Confusion matrix of the proposed method on the UCM dataset and AID dataset: (a) UCM dataset; (b) AID dataset.

5.6. Experiment V: Comparison with Other State-of-the-Art Methods

In order to prove the effectiveness of our proposed RDN + CBAM, we use the UCM and AID datasets to compare the proposed network with mainstream deep learning methods for remote sensing scene classification methods.

The classification accuracy comparison with other state-of-the-art scene classification methods on UCM and AID datasets is shown in Table 5. CaffeNet, GoogLeNet, and VGG-VD-16 are classic classification networks with different depths, but all use the last fully connected layer as the classification feature. The overall accuracies of CaffeNet on UCM and AID are 95.02% and 89.53%. The overall accuracy of GoogLeNet are 94.31% and 86.39%. The overall accuracy of VGG-VD-16 are 95.21% and 89.64%. The overall accuracy of the network we proposed on UCM is 99.82%, which makes an increase of 1.38%, 1.01%, and 0.7% point over the CTFCNN, VGG-16-CapsNet and ARCNet-VGG16, respectively. The overall accuracy of the network we proposed on AID is 99.08%, that has increased by 4.17%, 4.34%, and 5.98%, respectively, compared with these three remote sensing scene classification networks. CTFCNN, VGG-16-CapsNet, ARCNet-VGG16 are three networks dedicated to remote sensing scene classification, and all have achieved good performance on both datasets. CTFCNN used different types of features, fusing three parts of convolutional layer, fully connected layer and LBP for classification. VGG-16-CapsNet took advantage of CapsNet features, using pre-trained VGG16 to extract features, and then input to CapsNet for remote sensing scene classification tasks. But none of them make full use of the characteristics of each layer in the network. ARCNet-VGG16 adopted a recurrent attention to pay attention to the important information of the image, which greatly improves the accuracy of scene classification of remote sensing image, but only features weighting in the spatial dimension, ignoring the role of channel dimension weighting features. Therefore, it can be confirmed that the proposed method makes a great contribution to the promotion of the classification accuracy by fusing multilayer feature and adding attention module.

Table 5. Accuracy comparison of other state-of-the-art scene classification methods on UCM and AID datasets.

Network	UCM (%)	AID (%)
CaffeNet [46]	95.02	89.53
GoogLeNet [46]	94.31	86.39
VGG-VD-16 [46]	95.21	89.64
CTFCNN [47]	98.44	94.91
VGG-16-CapsNet [48]	98.81	94.74
ARCNet-VGG16 [44]	99.12	93.10
RDN + CBAM (ours)	99.82	99.08

6. Discussion

We have proposed an efficient method for the scene classification of a high-resolution remote sensing image. We improved the classification performance using three methods, including the combination of dense connection and residual connection, the addition of the channel-spatial attention module and the data augmentation in softmax classifier period. In particular, considering the complex background of the remote sensing image, the variety of scene categories, and the rich scene feature information, more specific and detailed features are obtained by RDN, in which multiple RDBs are used to fuse multi-layer convolutional features by the combination of dense connection and residual connection. The input of each layer in the RDB is determined by the output of all the previous layers, and the output will be transmitted to each subsequent layer. In order to reduce the amount of redundant information in the whole remote sensing image that with the complex spatial information, channel-spatial attention is added after each RDB by focusing on the useful features of image in both channel and spatial dimensions, so that the features obtained are more refined and has a more expressive representation capability, which can contribute to the classification performance. In addition, the amount of labeled remote sensing image data is always small, which is not enough to meet the training requirements of a large number of parameters of the deep network. In order to expand the training data, we conducted a data augmentation strategy by rotating and mirroring the image, which is more conducive to network parameter training.

Based on the above, we conducted five experiments to verify the effectiveness of our proposed method. In Experiment I, the effect of the combination of residual connection and dense connection was compared in Table 2. The RDN obtain a 98.6012% and 94.1563% overall accuracy on the UCM and AID dataset, respectively, both of which exceed the accuracy of ResNet50 with residual connection alone and DenseNet121 with dense connection alone. The results show that the advantages of residual connection and dense connection are combined to obtain more detailed features than single use. In Experiment II, the performance of adding channel-spatial attention module was shown in Figure 6. Although the classification accuracy was improved greatly by RDN, the extracted information will have a lot of redundancy. The result in Figure 6 showed that the performance of all the network added CBAM has been greatly improved. The highest accuracy was obtained from RDN added CBAM on two datasets, with 99.8214% and 99.0813%. It can be confirmed that CBAM plays a great role in extracting the effective features of a remote sensing image and improving the performance of network classification. In Experiment III, we compared the effects of dataset augmentation, and the result in Figure 7 shows that our method with augmentation achieves a 9.4196% and 12.1765% improvement compared with the manner without augmentation on UCM and AID, and the performance gain confirmed the important role of dataset augmentation. In Experiment IV, we compared the scene classification accuracy of each category with four networks, including RDN + CBAM, RDN, ResNet50 and DenseNet121. The results on UCM and AID datasets are shown in Figure 8, Tables 3 and 4. The proposed method RDN + CBAM obviously outperforms the other three networks in accuracy of each category. The result demonstrates that our proposed method can extract more distinguished features for scenes with high inter-class similarity and intra-class difference, which can further make contributions to improve the classification performance. Confusion matrixes of RDN + CBAM on UCM and AID are shown in Figure 9; it can

see the incorrectly classified categories of the scene. Although there were still misclassifications due to the similarities of some scenes are greater, it can prove that the method can generally distinguish the remote sensing scene categories with higher complexity. We further compared the result with the other state-of-the-art methods in Experiment V. Table 5 showed that our method achieves optimal accuracy with a 0.7% and 4.17% increase compared with the best result of the six compared methods, which further proved that the proposed method makes a great contribution to the promotion of the classification accuracy by fusing multilayer feature and adding attention module with dataset augmentation strategy.

7. Conclusions

We proposed an RDN based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. First, multi-layer convolutional features are fused with RDBs to achieve feature reuse and obtain detailed image feature information. Then, CBAM is added to obtain more effective feature representation. Finally, softmax classifier is applied to classify the scene after adopting a data augmentation strategy for meeting the training requirements of the network parameters. The proposed network can achieve better performance than state-of-the-art methods on two publicly available scene classification datasets of a high-resolution remote sensing image, and the classification accuracy on UCM and AID datasets reach 99.8214% and 99.0813%, respectively. It shows that our network can extract more representative and distinguishing features, which are more conducive to scene classification tasks. However, for some scenes with high inter-class similarity, there are still some misclassifications. In the future, we will consider integrating multiple types of features to further enhance the feature learning ability for distinguishing these scenes more effectively. In addition, due to the large network parameters and the high amount of training datasets, the network training speed is slower. In the future, we will also improve the performance of the scene classification task in terms of improving the network training speed.

Author Contributions: All the authors made significant contributions to the work. X.Z.'s contribution is conceptualization, investigation, methodology and writing-original draft. J.Z. (Jing Zhang)'s contribution is conceptualization, data curation, funding acquisition, methodology and writing-review & editing. J.T. is working for formal analysis, validation and writing-review & editing. L.Z. is working for funding acquisition, methodology and resources. J.Z. (Jie Zhang)'s contribution is conceptualization, methodology and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (No. 61370189), Beijing Education Committee Cooperation Beijing Natural Science Foundation (No. KZ 201810005002, No. KZ 201910005007).

Acknowledgments: The authors would like to thank the anonymous reviewers and associate editor for their valuable comments and suggestions to improve the quality of the paper.

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

References

1. Cheng, G.; Guo, L.; Zhao, T.; Han, J.; Li, H.; Fang, J. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *Int. J. Remote Sens.* **2012**, *34*, 45–59. [\[CrossRef\]](#)
2. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [\[CrossRef\]](#)
3. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [\[CrossRef\]](#)
4. Wang, Y.; Zhang, L.; Tong, X.; Zhang, L.; Zhang, Z.; Liu, H.; Xing, X.; Mathiopoulos, P.T. A Three-Layered Graph-Based Learning Approach for Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6020–6034. [\[CrossRef\]](#)
5. Plaza, J.; Plaza, J.; Paz, A.; Sanchez, S. Parallel Hyperspectral Image and Signal Processing [Applications Corner]. *IEEE Signal Process. Mag.* **2011**, *28*, 119–126. [\[CrossRef\]](#)

6. Hubert, M.J.; Carole, E. Airborne SAR-efficient signal processing for very high resolution. *Proc. IEEE*. **2013**, *101*, 784–797.
7. Cheriadat, A.M. Unsupervised Feature Learning for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 439–451. [\[CrossRef\]](#)
8. Shao, W.; Yang, W.; Xia, G.-S. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification. *Int. J. Remote Sens.* **2013**, *34*, 8588–8602. [\[CrossRef\]](#)
9. Estoque, R.C.; Murayama, Y.; Akiyama, C.M. Pixel-based and object-based classifications using high- and medium-spatial-resolution imageries in the urban and suburban landscapes. *Geocarto Int.* **2015**, *30*, 1113–1129. [\[CrossRef\]](#)
10. Zhang, X.; Wang, Q.; Chen, G.; Dai, F.; Zhu, K.; Gong, Y.; Xie, Y. An object-based supervised classification framework for very-high-resolution remote sensing images using convolutional neural networks. *Remote Sens. Lett.* **2018**, *9*, 373–382. [\[CrossRef\]](#)
11. Pham, M.-T.; Mercier, G.; Regniers, O.; Michel, J. Texture Retrieval from VHR Optical Remote Sensed Images Using the Local Extrema Descriptor with Application to Vineyard Parcel Detection. *Remote Sens.* **2016**, *8*, 368. [\[CrossRef\]](#)
12. Napoletano, P. Visual descriptors for content-based retrieval of remote-sensing images. *Int. J. Remote Sens.* **2017**, *39*, 1343–1376. [\[CrossRef\]](#)
13. Yang, Y.; Newsam, S. Geographic Image Retrieval Using Local Invariant Features. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 818–832. [\[CrossRef\]](#)
14. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [\[CrossRef\]](#)
15. Wang, S.; Guan, Y.; Shao, L. Multi-Granularity Canonical Appearance Pooling for Remote Sensing Scene Classification. *IEEE Trans. Image Process.* **2020**, *29*, 5396–5407. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Fang, J.; Yuan, Y.; Lu, X.; Feng, Y. Robust Space–Frequency Joint Representation for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7492–7502. [\[CrossRef\]](#)
17. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, J. Remote Sensing Scene Classification Using Multilayer Stacked Covariance Pooling. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6899–6910. [\[CrossRef\]](#)
18. Khan, N.; Chaudhuri, U.; Banerjee, B.; Chaudhuri, S. Graph convolutional network for multi-label VHR remote sensing scene recognition. *Neurocomputing* **2019**, *357*, 36–46. [\[CrossRef\]](#)
19. Liu, N.; Wan, L.; Zhang, Y.; Zhou, T.; Huo, H.; Fang, T. Exploiting Convolutional Neural Networks With Deeply Local Description for Remote Sensing Image Classification. *IEEE Access* **2018**, *6*, 11215–11228. [\[CrossRef\]](#)
20. Jin, P.; Xia, G.-S.; Hu, F.; Lu, Q.; Zhang, L. AID++: An Updated Version of AID on Scene Classification. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 4721–4724. [\[CrossRef\]](#)
21. Hu, F.; Xia, G.S.; Yang, W.; Zhang, L.P. Recent advances and opportunities in scene classification of aerial images with deep models. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 4371–4374.
22. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised Deep Feature Extraction for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1349–1362. [\[CrossRef\]](#)
23. Yu, Y.; Liu, F. Dense Connectivity Based Two-Stream Deep Feature Fusion Framework for Aerial Scene Classification. *Remote Sens.* **2018**, *10*, 1158. [\[CrossRef\]](#)
24. Luo, B.; Jiang, S.; Zhang, L. Indexing of Remote Sensing Images with Different Resolutions by Multiple Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1899–1912. [\[CrossRef\]](#)
25. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.-S.; Zhang, L. Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [\[CrossRef\]](#)
26. Wu, H.; Liu, B.; Su, W.; Zhang, W.; Sun, J. Hierarchical Coding Vectors for Scene Level Land-Use Classification. *Remote Sens.* **2016**, *8*, 436. [\[CrossRef\]](#)
27. Yang, Y.; Newsam, S. Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery. In Proceedings of the IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 1522–4880.

28. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
29. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.; Zhang, L. The bag-of-visual-words scene classifier combining local and global features for high spatial resolution imagery. In Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China, 15–17 August 2015.
30. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Ohio, CO, USA, 24–27 June 2014; pp. 580–587.
31. Yang, H.; Yu, B.; Luo, J.; Chen, F. Semantic segmentation of high spatial resolution images with deep neural networks. *GISci. Remote Sens.* **2019**, *56*, 749–768. [\[CrossRef\]](#)
32. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [\[CrossRef\]](#)
33. Yuan, Q.; Yuan, Q.; Li, J.; Shen, H.; Zhang, L. Hyperspectral Image Denoising Employing a Spatial-Spectral Deep Residual Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1205–1218. [\[CrossRef\]](#)
34. Zhong, Y.; Fei, F.; Zhang, L. Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *J. Appl. Remote Sens.* **2016**, *10*, 25006. [\[CrossRef\]](#)
35. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [\[CrossRef\]](#)
36. Jian, L.; Gao, F.; Ren, P.; Song, Y.; Luo, S. A Noise-Resilient Online Learning Algorithm for Scene Classification. *Remote Sens.* **2018**, *10*, 1836. [\[CrossRef\]](#)
37. Scott, G.J.; Hagan, K.C.; Marcum, R.A.; Hurt, J.; Anderson, D.T.; Davis, C. Enhanced Fusion of Deep Neural Networks for Classification of Benchmark High-Resolution Image Data Sets. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1451–1455. [\[CrossRef\]](#)
38. Liu, Y.; Suen, C.Y.; Liu, Y.; Ding, L. Scene Classification Using Hierarchical Wasserstein CNN. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2494–2509. [\[CrossRef\]](#)
39. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [\[CrossRef\]](#)
40. He, K.; Zhang, X.; Ren, S. 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.
41. Huang, G.; Liu, Z.; Maaten, L.V.D. 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.
42. Zhang, Y.; Tian, Y.; Kong, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
43. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 2204–2212.
44. Wang, Q.; Liu, S.T.; Chanussot, J. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [\[CrossRef\]](#)
45. Woo, S.; Park, J.; Lee, J.Y. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Xia, G.S.; Hu, J.; Hu, F. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [\[CrossRef\]](#)
47. Huang, H.; Xu, K. Combining Triple-Part Features of Convolutional Neural Networks for Scene Classification in Remote Sensing. *Remote Sens.* **2019**, *11*, 1687. [\[CrossRef\]](#)
48. Zhang, W.; Tang, P.; Zhao, L. Remote Sensing Image Scene Classification Using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [\[CrossRef\]](#)

