Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification

Wei Tong, Weitao Chen, Wei Han, Xianju Li, and Lizhe Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing image scene classification has been widely applied and has attracted increasing attention. Recently, convolutional neural networks (CNNs) have achieved remarkable results in scene classification. However, scene images have complex semantic relationships between multiscale ground objects, and the traditional stacked network structure lacks the ability to effectively extract multiscale and key features, resulting in limited feature representation capabilities. By simulating the way that humans understand and perceive images, attention mechanisms can be beneficial for quickly and accurately acquiring key features. In our study, we propose a channel-attention-based DenseNet (CAD) network for scene classification. First, the lightweight DenseNet121 is selected as the backbone for the spatial relationship between multiscale ground objects. In the spatial domain, densely connected CNN layers can extract spatial features at multiple scales and correlate with each other. Second, in the channel domain, a channel attention mechanism is introduced to strengthen the weights of the important feature channels adaptively and to suppress the secondary feature channels. Third, the cross-entropy loss function based on label smoothing is used to reduce the impact of interclass similarity upon feature representations. The proposed CAD network is evaluated on three public datasets. The experimental results demonstrate that the CAD network can achieve performance comparable to those of other state-of-the-art methods. The visualization through the Grad-CAM ++ algorithm also reflects the effectiveness of channel attention and the powerful feature representation capabilities of the CAD network.**

***Index Terms*—Attention mechanism, deep learning, DenseNet, remote sensing, scene classification.**

I. Introduction

REMOTE sensing image scene classification involves categorizing remote sensing image scene patches into meaningful labels according to the image contents and providing high-level semantic concepts [1]. With the rapid development of satellite remote sensing technology, it is possible to obtain large amounts of remote sensing scene images. In recent years, remote sensing image scene classification has attracted increasing attention. It has always been an active research topic [2], [3] because of its important applications in land use and land cover classification and determination [4], [5], vegetation mapping [6], and urban functional analysis [7].

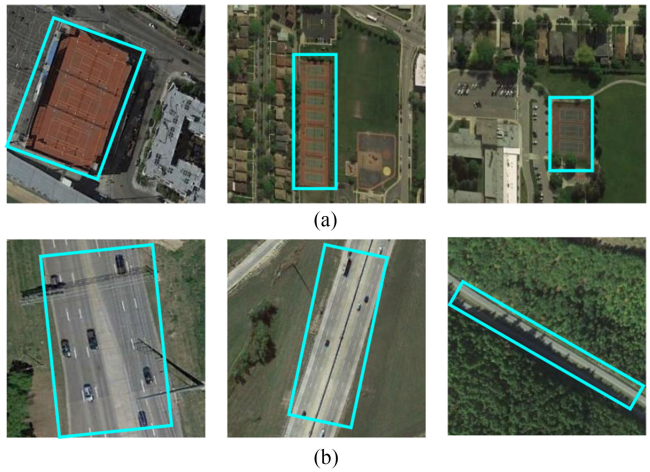


Fig. 1. Remote sensing scene images containing multiscale objects. Blue boxes represent the attention regions. (a) Tennis court. (b) Freeway.

In the past few decades, researchers have concentrated on extracting various effective feature representations to improve the performance of remote sensing image scene classification [3]. They can be roughly divided into three main categories: handcrafted features, midlevel features, and deep features. The related literature for scene classification is described in Section II. Benefiting from the deep learning methods for natural images, a series of deep convolutional neural network (CNN) models [2], [8], and [9] was proposed, which are superior to traditional handcraft features and midlevel features in terms of feature representation.

However, for remote sensing image scene classification, how to obtain multiscale spatial features is highly appealing. Some remote sensing scene image categories contain multiple classes of objects at different scales. As shown in Fig. 1, for the "tennis court" category, the image contains tennis courts, lawns, and roads on different scales. For the "freeway" category, the image contains freeways, grassland, and cars on different scales. Additionally, compared with natural images, remote sensing scene images appear to be limited in number, with only hundreds of images per category, and exhibit interclass similarity. Previously developed CNNs [2], [8], and [9] are usually not deep and require large parameters. For example, CaffeNet has five

Manuscript received February 19, 2020; revised April 17, 2020 and June 11, 2020; accepted July 9, 2020. Date of publication July 15, 2020; date of current version July 30, 2020. This work was supported by the National Natural Science Foundation of China under Grant U1803117, Grant 41701516, Grant U1711266, and Grant 41925007. *(Corresponding authors: Weitao Chen; Lizhe Wang.)*

The authors are with the School of Computer Science and Hubei Key Laboratory of Intelligent Geo-Information Processing, the China University of Geosciences, Wuhan 430074, China, and also with the Key Laboratory of Geological Survey and Evaluation of Ministry of Education, the China University of Geosciences, Wuhan 430074, China (e-mail: weitong@cug.edu.cn; wtchen@cug.edu.cn; weihan@cug.edu.cn; ddwhlxj@cug.edu.cn; lizhe.wang@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.3009352

convolution layers with 61.2M parameters, and VGG16 has 13 convolution layers and 138.3M parameters. When the amount of training data is relatively small, CNNs with shallow layers and numerous parameters tend to overfit, whereas CNNs with deep layers and few parameters do not. Moreover, deeper layers are more conducive to the extraction of high-level information [10].

On the other hand, the importance of image features is not the same for judging scene categories. The features of the tennis court and freeway play major roles, whereas those of the lawns and cars are secondary. The human eye prioritizes the primary features and pays less attention to the secondary features [11]. Similarly, during this study, we paid more attention to the key image features for scene categorization.

We propose a novel channel-attention-based DenseNet (CAD) network to better extract multiscale and key features. The CAD network has lightweight parameters (8.25M) and a deep network (121 layers). In the spatial domain, densely connected CNN blocks can extract and associate image features at multiple scales, which is beneficial for extracting high-level features. The attention mechanism is used to adaptively adjust feature weights. In the channel domain, the channel attention mechanism is integrated into DenseNet. The cross entropy loss function, based on label smoothing, is used for backpropagation, which can reduce the effect of interclass similarity in representing scene images.

The major contributions of this study are as follows.

1) A lightweight network with powerful spatial representation capabilities is utilized. To consider the small amount of remote sensing scene image data and the complex semantic relationships of multiscale objects, the CAD network uses lightweight and deep DenseNet121. In the spatial domain, densely connected CNN blocks in the network can extract multiscale image features and establish connections.
2) A network design that includes a channel attention mechanism is used. To make the network pay more attention to important image features, in the channel domain, the channel attention mechanism is integrated into DenseNet. The attention mechanism increases the weights of important features and suppresses the weights of secondary features to improve the scene classification performance.
3) An effective loss function is employed. To reduce the effect of interclass similarity on representing scene images, a cross-entropy loss function based on label smoothing is used for backpropagation. The traditional cross-entropy function only considers the optimal result on a single real label, and the label smoothing considers the loss of multiple categories on the loss function, reducing the occurrence of overfitting.

The rest of this article is organized as follows. Section II gives a brief review of related works. Section III describes the proposed CAD network for the remote sensing scene classification in detail. The experimental results are presented in Section IV. Section V provides a discussion, and finally, Section VI concludes this article.

II. Related Work

The previous feature representations [3] for scene classification can be roughly divided into three types: handcrafted features, midlevel features, and deep features. In addition, the development of attention mechanism [12]–[14] has improved the representation capability of deep features.

The early scene classification methods were based primarily on handcrafted features, focusing on designing local or global shallow features, such as texture, color, shape, and spatial information [15]. These handcrafted features include color histogram (CH) [3], scale-invariant feature transform (SIFT) [16], local binary pattern (LBP) [17], gray level co-occurrence matrixes [18], Gabor filters, and histograms of oriented gradients [3] or combinations of these features [19]. In addition, multiple handcrafted feature fusion methods perform well on images with neat texture or spatial structures. However, the participation of human ingenuity in feature design significantly affects the representation ability and effectiveness. When the scene becomes more complicated and challenging, these methods offer very limited use [3]. Compared to the methods used for handcrafted features, midlevel feature methods attempt to compute local visual features to represent overall images. In general, midlevel features [20]–[24] are built to extract the local properties of the image patches first and then encode them to obtain a more advanced feature representation. The visual word bag (BoVW) is one of the most popular midlevel coding methods for scene classification due to its simplicity and excellent performance. BoVW-based methods [25] require large amounts of prior information, and the image is represented by the number of occurrences of local features. Thus, the spatial information in the image is severely lost, which greatly limits the ability to describe the image.

Recently, with the fast development of deep learning technology, impressive achievements have been made in many research areas, including image classification [26], hyperspectral unmixing [27], image superresolution [28], and image caption generation [29]. For remote sensing scene classification, deep learning demonstrates powerful feature extraction capabilities and has also achieved state-of-the-art performance [3], [8], [15], [30]–[32]. The CNN-based methods in deep learning dramatically outperform previous techniques strongly relying on handcrafted features. CNNs are typically trained on large quantities of labeled datasets and can extract image features without the need for engineering skill and domain expertise. When CNN models train from scratch for scene classification, the performance is poorer than those of pretrained CNN models due to the limited training data [33], [34]. A few works [33], [34] directly fine-tune the pretrained CNNs on remote scene datasets and obtain promising results. In certain works [2], [8], [9], pretrained CNNs have been employed as feature extractors without the training process and important layer features have been reprocessed for the final image representation. Some methods [35]–[41] involve the use of fusion technologies to combine various effective features and achieve impressive performance.

The development of attention mechanisms in deep learning has also provided new technologies for the field of remote sensing. Attention mechanisms, by simulating the way that humans

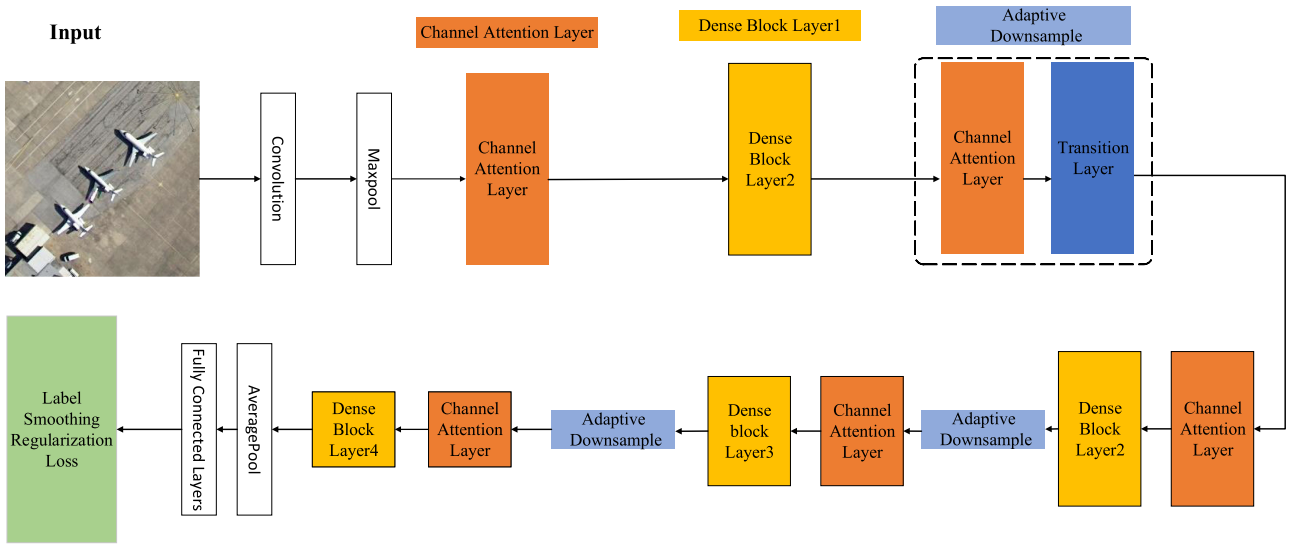


Fig. 2. Overall architecture of the CAD network. The novel CAD network consists of channel attention layers, dense block layers, the adaptive downsample, and the label smoothing regularization loss function.

understand and perceive images, can be beneficial for quickly and accurately acquiring key features [11]. Early works were mainly focused on the saliency detection methods. For example, Zhang *et al.* [42] employed a saliency sampling strategy to extract salient areas used for acquiring important graphical information. Hu *et al.* [12] utilized the salient region and keypoint methods for scene classification. However, saliency detection cannot be used to judge the importance of different objects adaptively, because the saliency area is obtained through texture information. Recently, some works have been proposed to adaptively extract attention-aware features during network training. Wang *et al.* [13] employed long short-term memory networks as attention extractor. Haut *et al.* [14] utilized generated attention masks to multiply corresponding areas to obtain attention features. Unlike previous works [13], [14] that were focused on the spatial attention with plenty of weight parameters, the attention method proposed in this work is focused on channel attention with few weight parameters for scene classification.

III. Proposed Method

As shown in Fig. 2, the novel CAD network consists of channel attention layers, dense block layers, adaptive downsample, and label smoothing regularization loss function. Section III-A provides the details of the backbone DenseNet and the dense block layers. Section III-B describes the channel attention layers and adaptive downsample. In Section III-C, the label smoothing regularization loss function is introduced with regard to its contribution toward building the network.

A. Spatial Feature Extraction Backbone

This backbone can be utilized to extract features of multiple different scales, and more importantly, this densely connected structure further cross links the features of these different scales, which performs better than traditional CNNs in representing the complex semantic relationships of multiscale objects in remote sensing scene images. CNNs have shown remarkable feature representation capabilities for scene classification. However, remote sensing scene datasets have small numbers of images, and some traditional CNN methods have numerous parameters and shallow network layers. On the one hand, overfitting often occurs due to small training data, and on the other hand, the shallow network layer results in limited extraction of high-level information. In addition, a deeper network produces more robust and discriminative features by acquiring higher-level feature mapping, which makes it easier to identify potential and inherent features [10]. However, deep networks tend to have problems, such as gradient disappearance, which partially negate the effects of network deepening. Many researchers [43]–[45] have addressed these problems, commonly by sharing a key characteristic: creating short paths from previous layers to subsequent layers.

DenseNet builds deeper densely connected networks and solves the above-mentioned problems. The most important architecture, the dense block layer, is designed to ensure maximum information flow between layers in the network [46]. In this architecture, each layer uses inputs from all previous layers and passes its corresponding feature map to all subsequent layers. These short connections between layers close to the input and output allow the previous features to be passed to the back effectively for automatic feature reuse. Thus, this network structure can be used to extract more global and important features, and can be more accurate and efficient to train. Some researchers have reprocessed features extracted from all layers and subsequently used the fused features for classification [2], [8], [9]. This method is a simple connection of multiple feature maps and is not intended to encourage features reuse between each layer. Thus, instead of integrating all feature maps, all previous layers are treated as input layers, as illustrated in Fig. 3.

Therefore, unlike in some conventional network structures, there are $L(L+1)/2$ connections instead of only L connections in an L-layer. In this way, the input feature maps of the ℓth

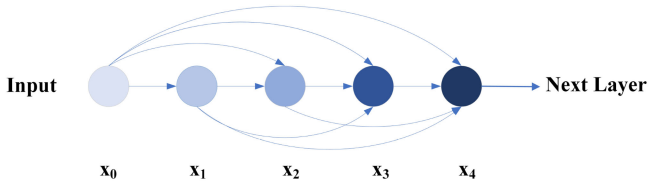


Fig. 3. Operation of densely connected convolution, the dense block layer.

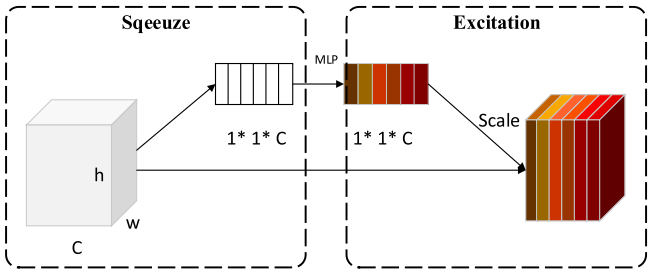


Fig. 4. Architecture of the channel attention layer.

layer can be calculated based on the preceding layers, including $\mathbf{x}_0, \ldots, \mathbf{x}_{l-1}$

$$\mathbf{x}_\ell = T_\ell\left([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]\right) \tag{1}$$

where $[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]$ denotes the concatenation of feature maps from layers $0, \ldots, \ell-1$. Moreover, the concatenation operation concatenates feature maps dimensionally, instead of referring to the pointwise sum. The nonlinear transformation $T_\ell(\cdot)$ in (1) can contain operations, such as convolution, pooling, [47], and rectified linear units [48]. Each dense block includes multiple sets of the 1×1 and 3×3 convolution layers with the same padding for the concatenation operation. Although this structure uses a densely connected pattern, it requires fewer parameters than a traditional convolutional network. In fact, this network architecture eliminates the necessity of learning redundant information, and the number of feature maps required by the network layer is reduced. Hence, parameter efficiency is significantly improved. On the other hand, continuous concatenations of different layers require each layer to access the gradients from the original input data and loss function. This fast access improves the flow of information between layers and reduces the gradient disappearance problem. This feature reuse method is conducive to building a deeper network architecture and extracting deep semantic relationships. To prevent overfitting and construct a lightweight network, we chose DenseNet121 as the backbone of the CAD network to extract complex spatial features. DenseNet121 contains a minimum of four dense blocks and corresponding transition layers with downsampling.

B. Channel Feature Attention Network

Since DenseNet has strong feature extraction capabilities in the spatial domain, we introduced an attention mechanism [11], called "squeeze-and-excitation blocks" into the feature channel domain. This attention mechanism adaptively selects important features by adjusting the weights of different feature maps in the channel domain. To make full use of the channel attention module without adding numerous parameters, only the dense block and the transition layers are combined in the channel attention mechanism. Moreover, the channel attention network itself is a small and effective architecture, and the parameter it adds is only 0.22M, so it will not cause overfitting. The transition layer consists of the 1×1 convolution layer and an average pooling with a stride of 2 to reduce the size of the feature map. The channel attention module that is integrated with the transition layer is called an adaptive downsample. As can be seen in Fig. 4, the processing of feature channels by this channel attention mechanism is divided into two stages: "squeeze" and "excitation."

In the squeeze stage, the input features are compressed into a 1-D vector whose length is the number of channels. The original input feature U size is $H \times W \times C$, the spatial domain size is $H \times W$, and the number of channels is C. Using global average pooling to compress each spatial domain $H \times W$ to a value, the resulting output is a $1 \times 1 \times C$ vector. Formally, the c_{th} element of the squeeze output z_c is calculated using

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i, j). \tag{2}$$

In the excitation stage, dependencies between channels are captured through the gate mechanism of two nonlinear fully connected layers. To limit the complexity of the model, the dimensions of the two fully connected layers are $C/16$ and C. The excitation output is denoted as s_c and is calculated using

$$s_c = F_{\text{ex}}(z, W) = \sigma(g(z, W)) = \sigma\left(W_2 \delta\left(W_1 z\right)\right) \tag{3}$$

where σ is the Sigmod function, δ is the ReLU function, and W_1 and W_2 are the parameters of $C/16$ and C layers. Then, each feature channel is assigned a corresponding weight. The inputs are the original feature maps u_c and weight vector s_c. The final output, feature maps u_c', is obtained by channelwise multiplication

$$u_c'(i, j) = s_c \cdot u_c(i, j). \tag{4}$$

The channel attention module assigns adaptive weights to different features by squeezing and expanding the feature channels. Compared with the attention model for the feature maps, this module has fewer parameters, reducing the risk of overfitting.

C. Cross-Entropy Loss Function Based on Label Smoothing

Interclass similarity of remote sensing scene datasets occurs when the main features in different types of scenes are the same or very similar. To reduce the effect of interclass similarity, the cross-entropy loss is combined with label smoothing to form the proposed cross-entropy loss function based on label smoothing.

For remote sensing scene image and classification, a softmax function is usually added to the last layer to calculate the probability that the input data belong to each category, and the highest probability is then used as the input to the cross-entropy function to calculate the loss value. The category vector is usually converted into a one-hot vector, where for an array of length n, only one element is 1 and the rest are 0. This characteristic causes the full probability and zero probability to encourage the gap between the real category and other categories to be as large as

possible, which means the model rewards the most correct cases and punishes the least correct cases. However, the feature gap between similar classes of ground objects is relatively small, and there is a hidden danger of overfitting when classifying these classes. Therefore, label smoothing [49] is used to improve the original cross-entropy loss function. The traditional softmax formula is as follows:

$$p_i = \frac{e^{\boldsymbol{x}^T \boldsymbol{w}_i}}{\sum_{l=1}^{L} e^{\boldsymbol{x}^T \boldsymbol{w}_l}} \tag{5}$$

where p_i is the likelihood the model assigns to the i_{th} class, w_i represents the weights and biases of the last layer, and x is the vector containing deep features extracted from the image. The expected value of the cross-entropy between the real target y_i and network output p_i is calculated and minimized by backpropagation

$$H(\boldsymbol{y}, \boldsymbol{p}) = \sum_{i=1}^{I} -y_i \log(p_i) \tag{6}$$

where y_i is "1" for the correct class and "0" for the rest. It can be found that the loss with label smoothing only considers the loss of the correct label position. Not considering the loss of wrong label positions causes the model to pay too much attention to increase the probability of predicting the correct label, instead of reducing the probability of predicting the wrong label. The final result is that the model fits its own training set well but yields poor results for other test sets. In particular, when the scene categories are similar and the loss of similar category labels is not considered, overfitting is more likely to occur.

To consider the loss of the correct label positions (the positions at which the one-hot label is "'1") and the loss of other incorrect label positions (the positions at which the one-hot label is "0") in the training sample, label smoothing is introduced

$$y' = (1 - \epsilon)y + \epsilon u(I) \tag{7}$$

where y' is the sample label after the label smoothing operation, ϵ is the smoothing factor, $u(I)$ obeys the uniform distribution about the classes I.

Therefore, the cross-entropy loss function based on label smoothing not only pays attention to the loss of the correct category, but also considers the loss of other categories. When the scene categories are similar, it is more important to pay attention to the losses of other categories, which also reduces the impact of the interclass similarity on feature representations.

IV. Experiments and Analysis

A series of experiments was conducted to evaluate the performance of the proposed CAD network using three public remote sensing scene image datasets. The experimental settings and experimental results, along with an analysis of the results, are presented in detail below.

A. Experimental Datasets

Three well-known public datasets for remote sensing scene classification are chosen to evaluate the CAD network. They are the popular UC Merced Land-Use Dataset [50], the more challenging AID Dataset [51], and the NWPU-RESISC45 Dataset [3] with large scale classes. These datasets are introduced briefly in Sections IV-A1–IV-A3.

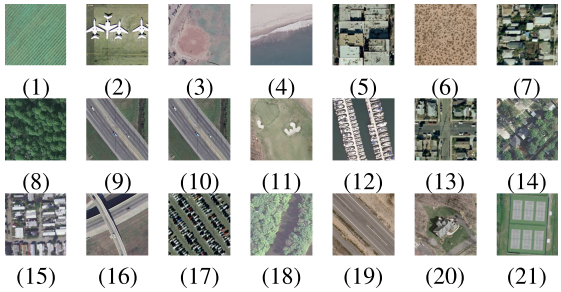


Fig. 5. Class representatives of the UC Merced Land-Use dataset.

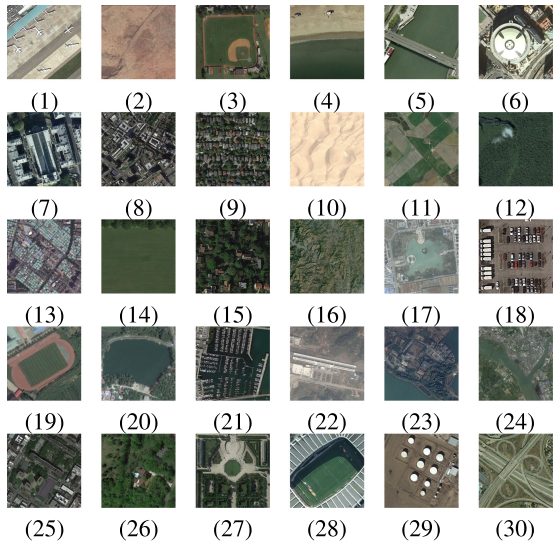


Fig. 6. Class representatives of the AID dataset.

1) UC Merced Land-Use Dataset: The UC Merced Land-Use Dataset consists of 2100 HR scene images, divided into 21 categories on average (see Fig. 5). There are 100 (256 × 256 pixel) images in each category. Each image is selected from aerial orthoimagery with a pixel spatial resolution of 30 cm in the RGB color space. This dataset is widely used to evaluate different methods for remote sensing image scene classification. As shown in Fig. 5, the scene classes include: (1) agricultural; (2) airplane; (3) baseball diamond; (4) beach; (5) buildings; (6) chaparral; (7) dense residential; (8) forest; (9) freeway; (10) golf course; (11) harbor; (12) intersection; (13) medium residential; (14) mobile home park; (15) overpass; (16) parking lot; (17) river; (18) runway; (19) sparse residential; (20) storage tanks; and (21) tennis court.

2) AID Dataset: The AID is a more challenging dataset than the UC Merced Land-Use dataset for the following reasons. First, it is a large-scale image dataset with more scene types and images. It contains 10 000 images, each of size 600 × 600 within 30 classes (see Fig. 6), and the number of images in the different scene classes ranges from 220 to 420. Furthermore, this dataset has more intraclass diversities as it includes images collected at various times and seasons, under various imaging conditions,

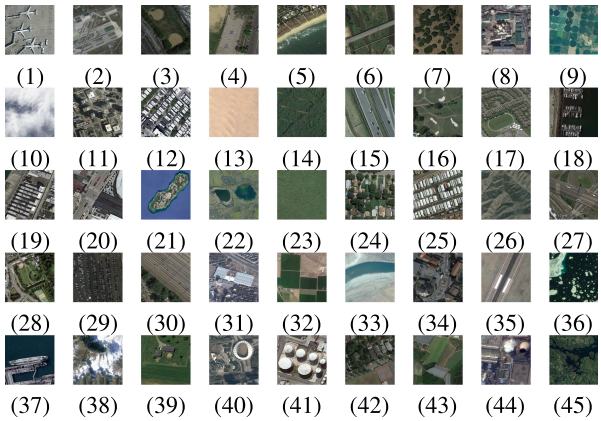


Fig. 7. Class representatives of the NWPU-RESISC45 dataset.

and from different regions and countries all over the world. Last but not least, the AID dataset has multiple resolutions, which range from approximately 0.5 to 8 m. This dataset is also widely used to evaluate different remote sensing image scene classification methods. As shown in Fig. 6, the scene classes include: (1) airport; (2) bare land; (3) baseball field; (4) beach; (5) bridge; (6) center; (7) church; (8) commercial; (9) dense residential; (10) desert; (11) farmland; (12) forest; (13) industrial; (14) meadow; (15) medium residential; (16) mountain; (17) park; (18) parking; (19) playground; (20) pond; (21) port; (22) railway station; (23) resort; (24) river; (25) school; (26) sparse residential; (27) square; (28) stadium; (29) storage tanks; and (30) viaduct.

3) NWPU-RESISC45 Dataset: This dataset is a new large-scale image dataset and is more complex than the UC Merced Land-Use and AID datasets. It contains a total of 31 500 images spread over 45 scene types. There are 700 (256 × 256) pixel images in the RGB color space per category. The spatial resolution of each image ranges from approximately 0.2 to 30 m. This dataset is selected from Google Earth and covers more than 100 regions throughout the world. As shown in Fig. 7, the scene classes include: (1) airplane; (2) airport; (3) baseball diamond; (4) basketball court; (5) beach; (6) bridge; (7) chaparral; (8) church; (9) circular farmland; (10) cloud; (11) commercial area; (12) dense residential; (13) desert; (14) forest; (15) freeway; (16) golf course; (17) ground track field; (18) harbor; (19) industrial area; (20) intersection; (21) island; (22) lake; (23) meadow; (24) medium residential; (25) mobile home park; (26) mountain; (27) overpass; (28) palace; (29) parking lot; (30) railway; (31) railway station; (32) rectangular farmland; (33) river; (34) roundabout; (35) runway; (36) sea ice; (37) ship; (38) snowberg; (39) sparse residential; (40) stadium; (41) storage tank; (42) tennis court; (43) terrace; (44) thermal power station; and (45) wetland.

B. Experimental Setup

To evaluate our proposed method fairly, we maintained the same training–testing ratios as in previous experiments [3], [8], [15], [51]–[56]. For the UC Merced Land-Use Dataset, the training ratios of 50% and 80%, respectively, were set, and the rest were used for testing. For the AID dataset, the training ratios are 20% and 50%, respectively. For the NWPU-RESISC45 dataset, the training ratios were 10% and 20%, respectively. In addition, two different training–testing ratios were considered for each dataset to comprehensively evaluate the CAD network.

In this work, the Pytorch framework was used to implement the proposed method. The network parameters and settings were as follows. All images were resized to 288×288 pixel as the input. Then, the batch size was set as 16, and stochastic gradient descent with dynamic learning rates was used as the optimizer. The training epochs lasted until the network converges. The experiments were implemented on a computer with dual Intel Xeon E5-2620 v4 processors, two 1080Ti GPUs, and 128 GB RAM.

C. Accuracy Assessment

The performance of the CAD network is evaluated using the overall accuracy (OA) and the confusion matrix. The OA is defined as the total number of correctly classified images divided by the number of images in the test set, which elucidates the classification performance with regard to predicting actual images. In the confusion matrix, each column represents the predicted instances, and each row represents the actual attribution of the class data. Therefore, the confusion matrix can directly display the distribution of each class, and can be employed to simply analyze the misclassification between different classes. To reduce the influence of randomness for a reliable result, we repeat the experiment ten times by randomly dividing the dataset for the above ratios. The mean and standard deviation of the overall accuracies for the ten repetitions are reported as the final performance.

D. Experiment Results

1) Classification of the UC Merced Land-Use Dataset: To demonstrate the effectiveness of the CAD network, the results of this work are compared with those of some state-of-the-art methods for the UC Merced Land-Use dataset. Table I shows that CaffeNet and VGG-16 outperform GoogLeNet, but the performance of GoogLeNet in natural scenes is the best among the three networks. Because GoogLeNet is a deeper and wider network, it is easier to learn some of the more detailed features. If it is not fine-tuned by retraining, it extracts many detailed features of the natural scenes for classification. Therefore, in the classification of remote sensing scenes, GoogLeNet does not provide the effect of a traditional network, such as CaffeNet and VGG-16. However, fine-tuned GoogLeNet demonstrates the powerful feature extraction capabilities of deeper and wider networks, and achieves better performance. The performance of deep CNN transfer also proves that deep nets and transfer learning could improve the classification effect. In addition, our proposed method is much deeper than CaffeNet and VGG-16, and is more susceptible to the gradient disappearance problem. However, the CAD network shows the best performance, with OAs of 99.66% and 98.57% for 80% and 50% training ratios, respectively. This indicates that the CAD network can provide discriminative image representations for scene classification and the lightweight network is less prone to overfitting.

Fig. 8 displays the confusion matrix generated from the best classification results obtained the CAD network with the training

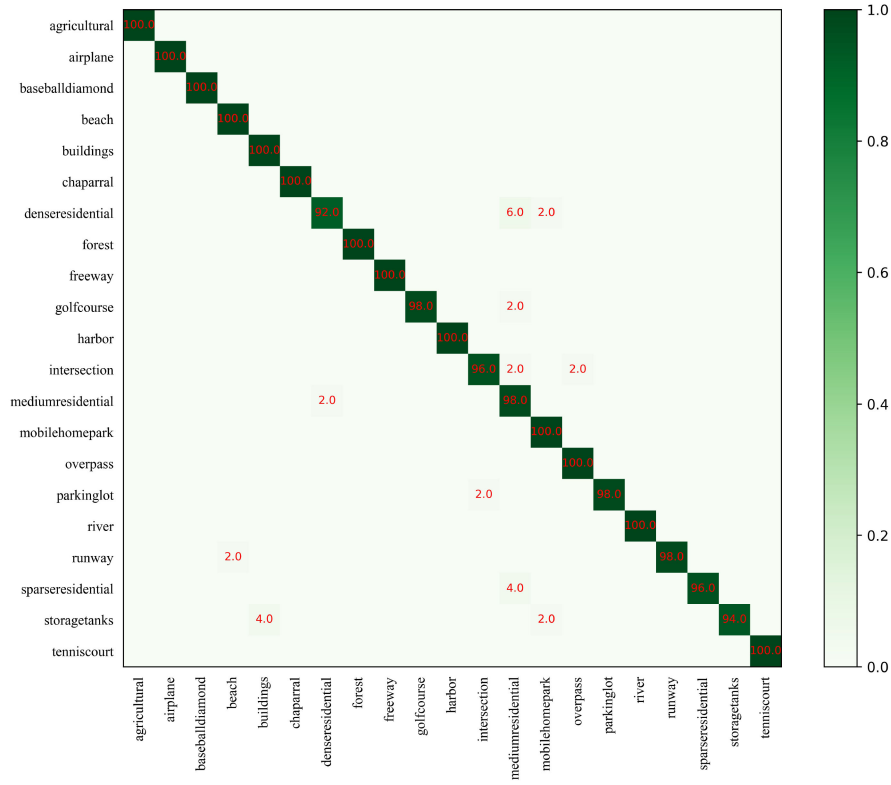


Fig. 8. Confusion matrix of the proposed method on UC Merced Land-Use dataset in the training ratio of 50%.

TABLE I
OA (%) AND STANDARD DEVIATIONS COMPARISON WITH THE UC MERCED DATASET

Method	80% Training Ratio	50% Training Ratio
CaffeNet [51]	95.02 ± 0.81	93.98 ± 0.67
GoogLeNet [51]	94.31 ± 0.89	92.70 ± 0.60
VGG-16 [51]	95.21 ± 1.20	94.14 ± 0.69
salM3LBP-CLM [54]	95.75 ± 0.80	94.21 ± 0.75
TEX-Net-LF [32]	96.62 ± 0.49	95.89 ± 0.37
LGFBOVW [22]	96.88 ± 1.32	/
Fine-tuned GoogLeNet [30]	97.10	/
Fusion by addition [8]	97.42 ± 1.79	/
CCP-net [55]	97.52 ± 0.97	/
Two-Stream Fusion [12]	98.02 ± 1.03	96.97 ± 0.75
DSFATN [52]	98.25	/
Deep CNN Transfer [2]	98.49	/
GCFs+LOFs [53]	99 ± 0.35	97.37 ± 0.44
Inception-v3-CapsNet [27]	99.05 ± 0.24	97.59 ± 0.16
D-CNN [57]	98.93 ± 0.10	/
CAD (ours)	**99.66 ± 0.27**	**98.57 ± 0.33**

TABLE II
OA (%) AND STANDARD DEVIATIONS COMPARISON WITH THE AID DATASET

Method	50% Training Ratio	20% Training Ratio
CaffeNet [51]	89.53 ± 0.31	86.86 ± 0.47
GoogLeNet [51]	86.39 ± 0.55	83.44 ± 0.40
VGG-16 [51]	89.64 ± 0.36	86.59 ± 0.29
salM3LBP-CLM [54]	89.76 ± 0.45	86.92 ± 0.35
TEX-Net-LF [32]	92.96 ± 0.18	90.87 ± 0.11
Fusion by addition [8]	91.87 ± 0.36	/
Two-Stream Fusion [12]	94.58 ± 0.25	92.32 ± 0.41
GCFs+LOFs [53]	96.85 ± 0.23	92.48 ± 0.38
Inception-v3-CapsNet [27]	96.32 ± 0.12	93.79 ± 0.13
D-CNN [57]	96.89 ± 0.10	90.82 ± 0.16
CAD (ours)	**97.16 ± 0.26**	**95.73 ± 0.22**

ratio of 50%. The figure shows that all categories achieved accuracies greater than 92%; thus, more than half of the scene classes can be fully identified by the CAD network. The maximum misclassification probability is 6%; "dense residential" is misclassified as "medium residential." This misclassification may be due to the fact that these two types have great similarities, such as the style of building construction. The main differences lie in the building density and roads. These characteristics lead to the misclassification of "dense residential."

2) Classification of the AID Dataset: Table II compares the classification performance of the CAD network with those of the existing state-of-the-art methods for the AID dataset. The CAD network shows the best performance with OAs of 97.16% and 95.73% for the 50% and 20% training ratios, respectively. Comparing Tables I and II, the classification accuracy of the same method on the AID dataset is lower than that on the UC Merced dataset. Due to the complexity of the AID dataset, for previous single networks (e.g., CaffeNet, GoogLeNet, and

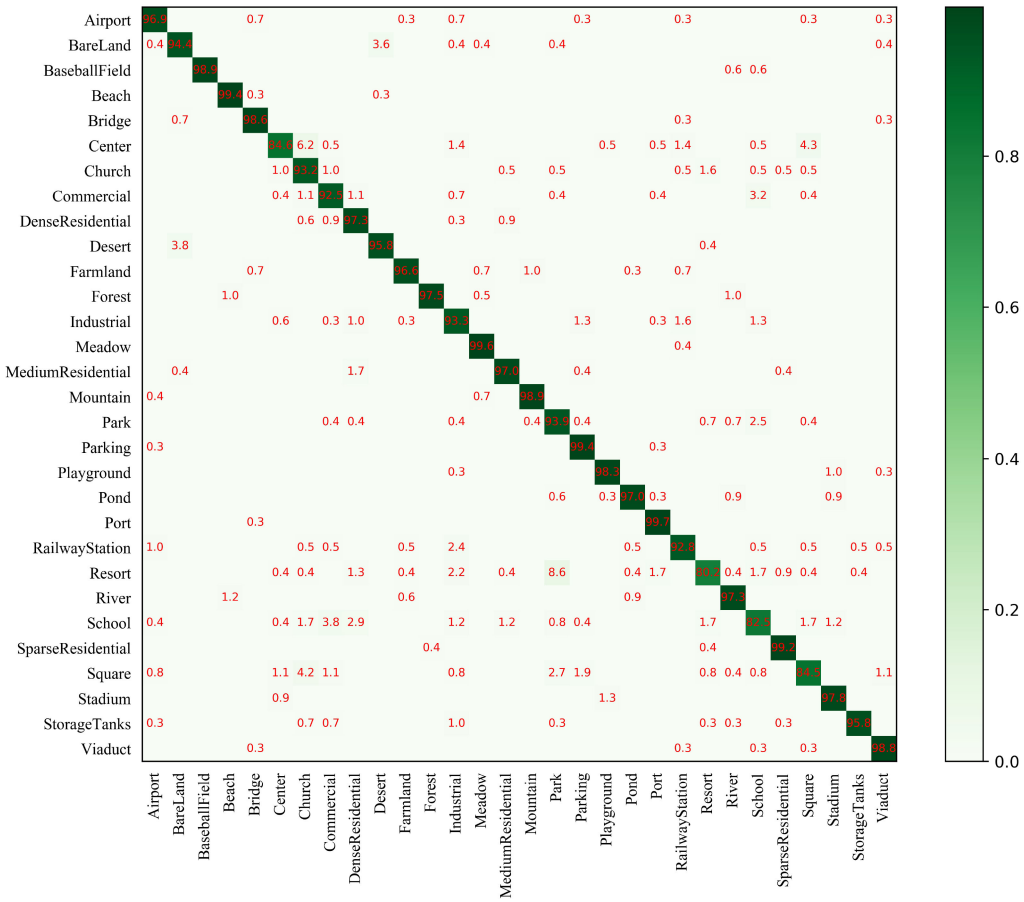


Fig. 9. Confusion matrix of the proposed method on the AID dataset in the training ratio of 20%.

VGG-16), the classification accuracy is not higher than 90%. For the fusion by addition and two-stream fusion, the fusion and reuse of deep network features have achieved the higher accuracy. In addition, the CAD network has the essence of multiscale feature connections and selection of effective features in network design. Therefore, the proposed method comprehensively produces state-of-the-art results for the AID dataset.

The confusion matrix obtained by the CAD network for the AID dataset with a 20% training ratio is shown in Fig. 9. The figure shows that 26 categories have classification accuracies greater than 90%, and the remainder have accuracies exceeding 80%. Some categories with similar content distributions, such as "sparse residential," "medium residential," and "dense residential" also have OAs of 99.2%, 97.0%, and 97. 3%, respectively. Similar to the results of a previous work[30], [51], the classification accuracies of "school" and "resort" are relatively low, at 82.5% and 80.2%, respectively. Specifically, "school" and "commercial" are misclassified because they have similar housing distributions. Likewise, the "resort" class is confused with "park" as they both are characterized by considerable vegetation and house coverage. Nonetheless, the proposed method shows considerable improvements compared to the previously reported classification accuracies of 49% and 60%, respectively [51]. Therefore, the proposed method can effectively recognize such categories with interclass similarity.

TABLE III
OA (%) AND STANDARD DEVIATIONS COMPARISON WITH THE NWPU-RESISC45 DATASET

Method	20% Training Ratio	10% Training Ratio
GoogLeNet [3]	78.48 ± 0.26	76.19 ± 0.38
VGG-16 [3]	79.79 ± 0.15	76.47 ± 0.18
AlexNet [3]	79.85 ± 0.13	76.69 ± 0.21
Two-Stream Fusion [12]	83.16 ± 0.18	80.22 ± 0.22
BoCF [58]	84.32 ± 0.17	82.65 ± 0.31
Fine-tuned AlexNet [3]	85.16 ± 0.18	81.22 ± 0.19
Fine-tuned GoogLeNet [3]	86.02 ± 0.18	82.57 ± 0.12
Fine-tuned VGG-16 [3]	90.36 ± 0.18	87.15 ± 0.45
Triple networks [56]	92.33 ± 0.20	/
Inception-v3-CapsNet [27]	92.60 ± 0.11	89.03 ± 0.21
D-CNN [57]	91.89 ± 0.22	89.22 ± 0.50
CAD (ours)	**94.58 ± 0.26**	**92.70± 0.32**

3) Classification of the NWPU-RESISC45 Dataset: The most challenging NWPU-RESISC45 dataset is also employed to demonstrate the effectiveness of the CAD network compared with other advanced methods (see Table III). With 10% and 20% training ratios, the classification accuracies for the CAD network are the best, with OAs of 92.70% and 94.58%, respectively, which are 3.67% and 1.98%, respectively, higher than

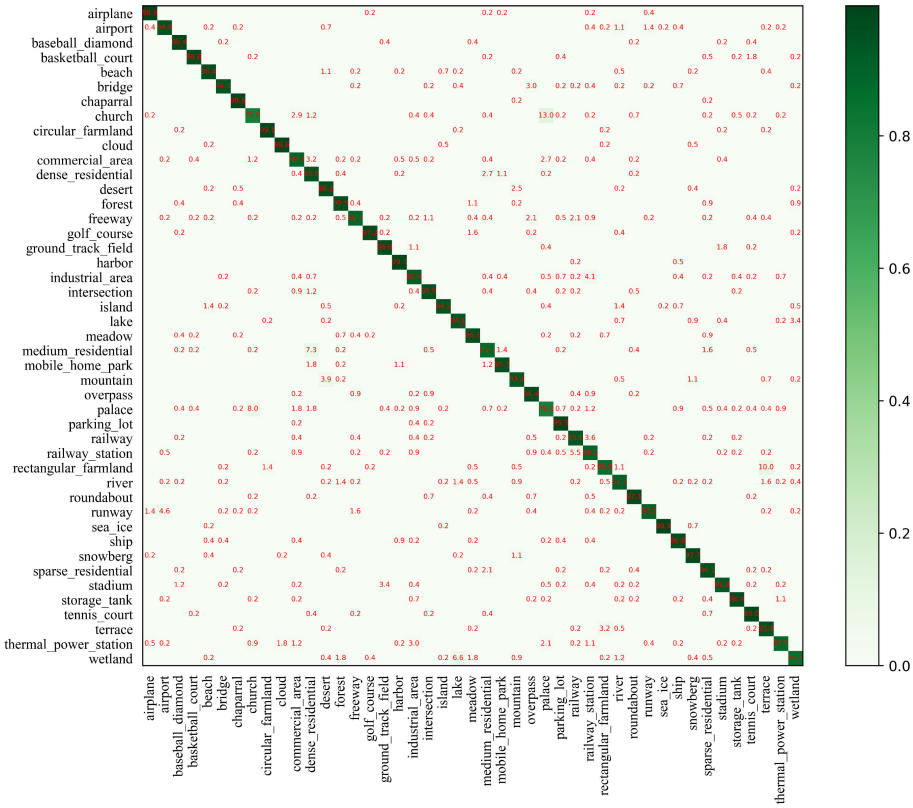


Fig. 10. Confusion matrix of the proposed method on the NWPU-RESISC45 dataset in the training ratio of 20%.

those obtained using Inception-v3-CapsNet. Compared to the other methods, the improvement is more obvious. Especially with small training ratios, the CAD network is significantly better than the other methods. In this challenging dataset, our proposed method can achieve such high precision, confirming the superiority of the proposed CAD network.

Fig. 10 shows the confusion matrix yielded from the best performance results with the CAD network and a training ratio of 20%. As shown in the confusion matrix, the classification accuracy of 36 of the 45 categories is greater than 90%. Although the classification accuracies of "palace" and "church" are relatively low at 79.3% and 79.5%, respectively. There is still improvement compared to the previous OAs of 75% and 64% [3]. Moreover, our model can distinguish each class well, with the lowest class precision being 79.5%. The model with close OAs [30] shows the lowest class precision value of only 68%. All these comparisons demonstrate the effectiveness and stability of the CAD network.

V. Discussion

A. Ablation Studies for the Proposed CAD Network

To demonstrate the performance of the CAD network on complex datasets, we performed ablation studies on the AID and the NWPU datasets. Table IV presents the classification performance of the CAD network and its different parts, including the CAD network without channel-attention and without label smoothing. In general, the CAD network without channel attention yielded competitive results, exceeding most methods. Although it is difficult to improve on such a high accuracy, the accuracy of the CAD network with channel attention is generally improved by 0.94%. This improvement was due to the channel attention enhancing the representation of important image features and suppressing the representation of less useful features. Label smoothing also had a positive effect on classification, which increased by 0.33%, because of the strong feature extraction capability of the CAD network. In the VGGNet and AlexNet, label smoothing can be used to obtain more obvious classification accuracy improvement.

TABLE IV
Ablation Study of the CAD Network on the AID and the NWPU-RESISC45 Dataset

	Without Channel-Attention	Without label Smoothing	The CAD Network
AID(50%)	95.99 ±0.49	96.72 ± 0.32	97.16 ± 0.26
AID(20%)	94.71 ± 0.32	95.43 ± 0.28	95.73 ± 0.22
NWPU(20%)	93.14 ± 0.28	94.21 ± 0.35	94.58 ± 0.26
NWPU(10%)	91.22 ± 0.41	92.46 ± 0.21	92.7 ± 0.32
AVERAGE	93.77	94.71	95.04

B. Attention Maps on the Proposed CAD Network

Fig. 11 shows the attention maps generated by the Grad-CAM++ algorithm. The brighter regions in the feature map represent the greater importance of the corresponding regions for classification. We chose four scenes: an airplane, a church, a

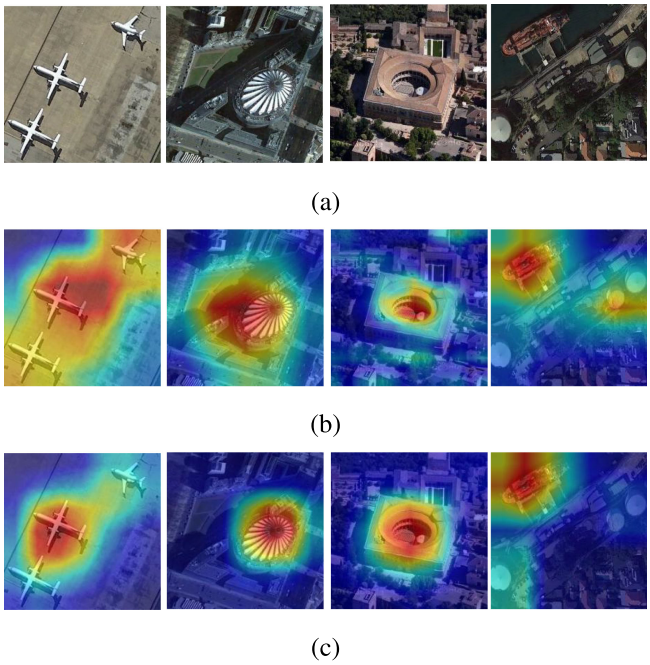


Fig. 11. Original scene images and the attention maps derived from Grad-CAM++. (a) Original scene images. (b) Grad-CAM++ visualization results without the channel attention. (c) Grad-CAM++ visualization results with the channel attention.

palace, and a ship. It can be seen from Fig. 11 that the network with the channel attention extracts important information well for these scenes. The main objects in each scene are accurately captured, significantly aiding scene classification. For the airplane and ship, the important areas of the CAD network with channel attention are more accurately located without interference from other low-priority features. Moreover, although palaces and churches are categories that are easy to confuse, the CAD network with the channel attention visually shows the control of important areas for these two types of features, reflecting the effectiveness of the channel attention and the powerful image feature representation capabilities of the CAD Network.

VI. Conclusion

Remote sensing scene images have complex semantic relationships between multiscale ground objects, whereas the traditional stacked network structure lacks the ability to effectively extract multiscale and key features. In our study, we proposed a novel CAD network to better extract multiscale and key features for remote sensing scene classification. First, the CAD network is a lightweight and deep network, which is more resistant to overfitting for small remote sensing scene datasets than traditional neural networks. Moreover, it has strong feature representation capabilities in the spatial domain, because densely connected CNN structures can effectively extract spatial information on multiple scales and fuse them together. Second, to enhance the extraction of important features, we further improved the channel domain by using a channel attention mechanism to increase the weight of important information and reduce the weight of low-priority information. Third, an improved loss function, a cross-entropy loss function based on label smoothing, is used for backpropagation. The loss function can consider the relationship between different classes, and can reduce the effect of interclass similarity on scene image representation.

Experiments were executed using three well-known open benchmark datasets and revealed the superiority of the CAD network in scene classification. The ablation studies also demonstrated the effectiveness of each module. Moreover, the attention maps generated by the Grad-CAM++ algorithm showed the classification performance improvement provided by the channel attention mechanism from the perspective of visualization. In future work, we will further explore more attention mechanisms and construct new models to apply to remote sensing.

References

[1] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[2] F. Hu, G. S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

[3] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state-of-the-art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[4] W. Chen, X. Li, H. He, and L. Wang, "Assessing different feature sets' effects on land cover classification in complex surface-mined landscapes by ZiYuan-3 satellite imagery," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 23.

[5] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[6] Q. Feng, J. Liu, and J. Gong, "UAV remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sens.*, vol. 7, no. 1, pp. 1074–1094, 2015.

[7] Z. Lv *et al.*, "Managing big city information based on WebVRGIS," *IEEE Access*, vol. 4, pp. 407–415, 2016.

[8] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[9] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[10] Y. Tao, M. Xu, Z. Lu, and Y. Zhong, "Dense net-based depth-width double reinforced deep learning neural network for high-resolution remote sensing image per-pixel classification," *Remote Sens.*, vol. 10, no. 5, p. 779, 2018.

[11] Y. Hu, G. Wen, M. Luo, D. Dai, J. Ma, and Z. Yu, "Competitive inner-imaging squeeze and excitation for residual network," Jul. 2018. [Online]. Available: http://arxiv.org/abs/1807.08920

[12] J. Hu, G. S. Xia, F. Hu, H. Sun, and L. Zhang, "A comparative study of sampling analysis in scene classification of high-resolution remote sensing imagery," in *Proc. Int. Geosci. Remote Sens. Symp.*, Nov. 2015, pp. 2389–2392.

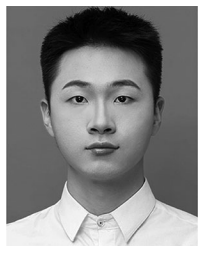
[13] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[14] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[15] Y. Yu and F. Liu, "Dense connectivity based two-stream deep feature fusion framework for aerial scene classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1158.

[16] Y. Yang and S. Newsam, "Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1852–1855.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[18] L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 780–795, Mar. 1999.
[19] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.
[20] L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, 2013.
[21] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.
[22] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.
[23] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.
[24] L. Zhao, P. Tang, and L. Huo, "A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *Int. J. Remote Sens.*, vol. 35, no. 6, pp. 2296–2310, 2014.
[25] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[27] X. Lu, H. Wu, and Y. Yuan, "Double constrained NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2746–2758, May 2014.
[28] X. Lu, Y. Yuan, and P. Yan, "Alternatively constrained dictionary learning for image superresolution," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 366–377, Mar. 2014.
[29] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
[30] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.
[31] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *Int. J. Remote Sens.*, vol. 37, no. 10, pp. 2149–2167, 2016.
[32] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.
[33] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.
[34] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.
[35] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, 2018.
[36] L. Ye, L. Wang, Y. Sun, L. Zhao, and Y. Wei, "Parallel multi-stage features fusion of deep convolutional neural networks for aerial scene classification," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 294–303, Mar. 2018.
[37] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.
[38] Y. Yu and F. Liu, "Aerial scene classification via multilevel fusion based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 287–291, Feb. 2018.
[39] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.
[40] C. Qiao, J. Wang, J. Shang, and B. Daneshfar, "Spatial relationship-assisted classification from high-resolution remote sensing imagery," *Int. J. Digit. Earth*, vol. 8, no. 9, pp. 710–726, Sep. 2015.
[41] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.
[42] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
[43] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 646–661.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[45] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 2017, pp. 2261–2269.
[47] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[48] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[49] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.
[50] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
[51] G. S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[52] X. Gong, Z. Xie, Y. Liu, X. Shi, and Z. Zheng, "Deep salient feature based anti-noise transfer network for scene classification of remote sensing imagery," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 410.
[53] D. Zeng, S. Chen, B. Chen, and S. Li, "Improving remote sensing scene classification by integrating global-context and local-object features," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 734.
[54] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.
[55] K. Qi, Q. Guan, C. Yang, F. Peng, S. Shen, and H. Wu, "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 934.
[56] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Jan. 2018.
[57] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
[58] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.



Wei Tong received the B.S. degree in electronic and information engineering from the Wuhan University of Technology, Wuhan, China, in 2018. He is currently working toward the M.S. degree in computer technology with the School of Computer Science, China University of Geosciences, Wuhan, China.

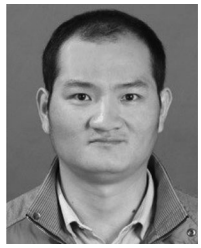
His research interests include remote sensing image processing, computer vision, and deep learning.

Weitao Chen was born in Wugang, Henan province in China. He received the B.E. degree in land resource management from Jiaozuo Institute of Technology, Jiaozuo, China, in 2003, and the M.E degree in quaternary geology, and Ph.D. degree in environmental science and engineering from China University of Geosciences (CUG), Wuhan, China, in 2006 and 2012, respectively.

He is an Associate Professor at the School of Computer Science, CUG. His main research interests include machine learning and remote sensing of environment.

Wei Han received the B.S. degree in network engineering and M.S. degree in computer science and technology from the China University of Geosciences, Wuhan, China, in 2016 and 2018, respectively, where he is currently working toward the Ph.D. degree in geoscience information engineering with the School of Computer Science, China University of Geosciences, Wuhan, China.

His research interests include data management, high-performance computing, and deep learning.



Xianju Li received the B.S. degree in surveying and mapping engineering, M.S. degree in geodesy and survey engineering, and Ph.D. degree in surveying and mapping from the China University of Geoscience, Wuhan, China, in 2009, 2012, and 2016, respectively.

Since 2016, he has been an Associate Professor with the School of Computer Science, China University of Geosciences. He has published more than ten papers. His main research fields include remote sensing image processing and analysis, computer vision, and machine learning.

Lizhe Wang (Senior Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D degree (*Magna Cum Laude*) in computer application from the University of Karlsruhe, Karlsruhe, Germany, in 2007.

He is currently the Dean and "ChuTian" Chair Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include remote sensing data processing, digital earth, and big data computing.

Prof. Wang is currently a Fellow of the Institution of Engineering and Technology and British Computer Society and an Associate Editor for *Remote Sensing, International Journal of Digital Earth, ACM Computing Surveys*, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING, etc. He is the recipient of Distinguished Young Scholars of NSFC, National Leading Talents of Science and Technology Innovation, and 100-Talents Program of Chinese Academy of Sciences.