Self-Attention Network With Joint Loss for Remote Sensing Image Scene Classification

HONGLIN WU, SHUZHEN ZHAO, LIANG LI, CHAOQUAN LU, AND WEN CHEN
School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
Corresponding author: Honglin Wu (honglinwu@csust.edu.cn)

ABSTRACT
Deep learning, especially convolutional neural network (CNN), has been widely applied in remote sensing scene classification in recent years. However, such approaches tend to extract features from the whole image rather than discriminative regions. This article proposes a self-attention network model with joint loss that effectively reduce the interference of complex backgrounds and the impact of intra-class diversity in remote sensing image scene classification. In this model, self-attention mechanism is integrated into ResNet18 to extract discriminative features of images, which makes CNN focus on the most salient region of each image and suppress the interference of the irrelevant information. Moreover, for reducing the influence of intra-class diversity on scene classification, a joint loss function combining center loss with cross-entropy loss is proposed to further improve the accuracy of complex scene classification. Experiments carried out on AID, NWPU-NESISC45 and UC Merced datasets show that the overall accuracy of the proposed model is higher than that of most current competitive remote sensing image scene classification methods. It also performs well in the case of fewer data samples or complex backgrounds.

INDEX TERMS
Remote sensing image, scene classification, convolutional neural network, self-attention, joint loss function.

I. INTRODUCTION
In recent years, with the rapid development of remote sensing technology and sensor system, remote sensing image data are emerging continuously. Remote sensing image scene classification is a fundamental and important task of remote sensing image analysis and interpretation, which is widely applied in urban planning, land resource management, and military investigation [1]–[4]. However, due to the highly complex geometric structure and spatial pattern of remote sensing image, Remote sensing image scene classification is seriously interfered by the redundant background in classification. Moreover, the diversity of objects also makes classification more difficult. These make the classification task quite challenging.

Feature extraction is a key step in remote sensing image scene classification, and most of the existing works focus on how to better describe image features. Most of the traditional classification methods are based on low-level or mid-level features [5], [6], but these features are difficult to effectively describe the semantic information of remote sensing images, which makes the classification results unsatisfactory. Therefore, recent researches focus on finding more advanced methods to automatically learn high-level feature representation. With the development of deep learning, it has been widely used in many computer vision tasks, such as image segmentation [7], [8], object detection [9]–[11], scene recognition [12], [13]. With the outstanding performance, convolutional neural network is utilized by researchers to extract high-level semantic features from remote sensing image scene classification. And most of them use the pretrained convolutional neural network models, such as CaffeNet [14], GoogLeNet [15], and VGGNet [16], as a feature extractors for remote sensing image scene classification. In [17], Chaib et al. utilize the pretrained VGGNet as feature extractor, and further refines the features extracting from VGGNet by discriminant correlation analysis, thereby achieving good classification performance. Similarly, ResNet [18] is the common convolutional neural network commonly used in remote sensing image scene classification. Compared with VGGNet, ResNet

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.
is deeper and can effectively alleviate the problem of network degradation. In general, deep learning-based methods can learn more abstract and advanced semantic features of images, and further effectively represent and recognize image scenes, thereby significantly improving classification accuracy.

Although deep learning-based methods have obtained wide attention, there are still great limitations for remote sensing image scene classification. Because they mainly focus on global feature extraction, but ignore the most critical information of objects in the image. In a natural image, the major object usually occupies most of the image, as shown in Figure 1 (a). Nevertheless, objects are usually small and scattered in the remote sensing image, and the redundant background occupies most of the image. Moreover, not all information in the image is useful for recognizing the image scene. As shown in Figure 1 (b), cars and roads are very important for classifying the freeway scenes, but trees and houses in the background will interfere the classification results. Therefore, we should pay more attention to the most discriminative features of objects when representing features of remote sensing images, and try to reduce the interference of irrelevant background. For this reason, we integrate self-attention mechanism into the convolutional neural network model ResNet18, focusing on the most discriminative and representative features of images, which can effectively reduce the computation and improve the performance. The effect of self-attention mechanism is shown in Figure 1 (c).

In addition, the high intra-class diversity of remote sensing image scene makes the classification very difficult, as shown in Figure 2. Different seasons, locations, or sensors may lead to huge variation of images with the same category. Consequently, we combine the center loss function with the cross-entropy loss function to reduce the influence of intra-class diversity on representing images. Center loss has been widely applied in face recognition [19]. It lessens intra-class differences and improves the discriminative ability of features.

In a word, the major contributions are described as follows:

1. We propose a deep network structure with self-attention, which embeds self-attention mechanism into ResNet18 for remote sensing image scene classification. The self-attention mechanism forces the network to focus on the most salient regions of images and suppresses the interference of redundant background. It can effectively reduce the computation and improve the performance of classification, especially in the case of small datasets or complex image background.
(2) In order to decrease the effect of intra-class diversity on classification, we propose a joint loss function that combines the center loss and the cross-entropy loss. The loss function lessens intra-class differences and improves the discriminative ability of the proposed model.

(3) The proposed model is evaluated on three benchmark datasets and performs well compared with other state-of-the-art methods, even in the case of fewer data samples or complex backgrounds.

The remainder of this article is organized as follows. Section 2 briefly reviews the background. The detailed introduction of the proposed model is carried out in Section 3. Section 4 introduces the experimental datasets, experimental setup, and analyzes the experimental results. The conclusion is presented in Section 5.

II. BACKGROUND

A. FEATURE EXTRACTION

According to the different semantic levels of the extracted features, the previous remote sensing image scene classification methods are mainly divided into three types: low-level features, mid-level features, and high-level features.

Methods based on low-level features represent images by extracting the primary characteristics of images, and then discriminate the scene. Therefore, the remote sensing image scene is usually expressed by feature vectors extracted from low-level visual attributes, which can be divided into global feature vectors and local feature vectors. On the one hand, for describing the complex structure information, local feature descriptors are widely used to model the local structures of remote sensing images, such as scale-invariant feature transform (SIFT) [20] and histogram of oriented gradients (HOG) [21]. On the other hand, in order to describe the spatial arrangements of images, global feature descriptors such as color histogram [22] and texture feature [23]–[25] are also widely applied. Methods based on low-level features can achieve good performance when the structure and spatial arrangements of the scene are evenly arranged, but limit in complex scenes.

Methods based on mid-level features mainly attempt to learn a set of basis functions used for feature encoding, among which the bag of visual words (BoVW) model [26], [27] is one of the most typical methods. In the BoVW-based models, local invariant features are encoded from local image patches into a vocabulary of visual words firstly, and then the image is represented according to the histogram of visual words. With simplicity and effectiveness, BoVW-based models and their improved models are widely used [28]–[30]. However, methods relying on BoVW also have limitations, for example, their large quantization error may lead to the loss of some important information. Therefore, for making better use of spatial information, lots of feature coding methods are proposed, such as improved Fisher kernel (IFK) [31], spatial pyramid matching (SPM) [32], and probabilistic latent semantic analysis (pLSA) [33], [34]. It is worth mentioning that both low-level features-based and med-level features-based methods are mainly based on hand-crafted features, and they lack flexibility and adaptability to different situations.

In recent years, Methods based on high-level features have attracted tremendous attention. Deep learning has gained dramatic development in computer vision [35]–[37], and many achievements have been made in remote sensing image scene classification based on deep features [38]–[40]. Generally, deep learning-based methods regard remote sensing image scene classification as an end-to-end problem, and adopt a multi-level feature learning framework to adaptively learn image features [5]. Compared with methods based on low-level features and mid-level features, methods based on deep learning have more advantages [41]–[43]. For one thing, more abstract and discriminative semantic features can be automatically learned by convolutional neural network from the original images. For another, deep learning-based methods can better distinguish scenes with great similarity.

B. ATTENTION MECHANISM

The human visual system usually pays attention to the salient region of the image and distinguishes the object in the image accordingly. Inspired by this, attention mechanism is proposed and developed rapidly in many fields [44], such as natural language processing, image recognition, and speech recognition. Numerous researches based on attention mechanism also emerged in remote sensing image scene classification recently. In [45], an algorithm based on multi-scale process to extract visual attention features is proposed for fuzzy classification of high-resolution remote sensing scenes. Wang et al. [46] design a recurrent attention structure that adaptively selects some key regions or locations in scenes and discard some noncritical information to improve classification performance. In order to make full use of the global and local information of aerial scenes, an end-to-end global-local attention network is proposed in [47]. In the network, global attention branch and local attention branch are used to replace the fully connected layer in VGGNet to learn global information and local semantic information respectively. Zhu et al. [48] propose a deep feature fusion framework based on attention, which utilizes Gradient-weighted Class Activation Mapping (Grad-CAM) to generate attention maps, thereby forcing the network to focus on the most significant area of the image and further improving the accuracy of classification.

Self-attention mechanism is an improvement of attention mechanism, which decreases the dependence on external information and is better at capturing the internal correlation of data or features. In [49], Wang et al. introduce the self-attention mechanism to solve some specific problems in computer vision. In [50], a self-attention generative adversarial network (SAGAN) is proposed. Different from the traditional convolutional GANs, SAGAN allows attention-driven, long-range dependency modeling for image generation tasks, and achieves good performance on ImageNet dataset. Hoogi et al. [51] introduce self-attention mechanism into medical image classification. In their model, the self-attention mechanism as an integral layer within the Capsule
Network (CapsNet), not only extracts significant features, but also reduces the computation. Self-attention mechanism can make the model focus on more relevant areas in the image, and it can achieve better classification performance in the case of fewer data samples or complex image background [52]. Self-attention mechanism is lack of exploration in remote sensing image scene classification currently. Therefore, we investigate it in this article.

III. MATERIALS AND METHODS

A. OVERALL ARCHITECTURE

The overall framework of the proposed model is shown in Figure 3, the proposed model consists of two parts: the deep network with self-attention and the joint loss function combining center loss and cross-entropy loss. In the deep network with self-attention, ResNet18 is used as the backbone network, and the self-attention mechanism is integrated into it to guide the network to focus on the key objects of remote sensing images, as shown in the gray box in Figure 3. More specifically, we add a self-attention mechanism between L5 and L6 of the ResNet18. This module further processes the features output from L5, focuses on the salient regions and suppresses the interference of the unrelated regions. The features output by self-attention mechanism are more discriminative and contain more high-level semantic information. Besides, there are usually large intra-class differences in remote sensing image scenes, resulting in the loss of classification accuracy. To solve this problem, we propose a joint loss function combining center loss and cross-entropy loss to optimize the model, thereby effectively distinguishing scenes with high similarity, and further improving classification performance.

B. RESNET

Compared with some deep network such as CaffeNet and VGGNet, ResNet is deeper and easier to train. In general, the deeper the network is, the more abstract and semantic features are extracted. Therefore, we adopt ResNet18 as the backbone network to extract more discriminative features for classification. ResNet18 consists of residual blocks, and a residual block is described in Figure 4. The residual block is composed of stacked convolutional layers and cross-layer


calculating features can be reused by cross-layer parameter sharing and intermediate features preserving.

\section{C. CSELF-ATTENTION MECHANISM}

The information in remote sensing images is abundant, and not all information has a positive impact on image representation. Therefore, it is necessary to capture salient objects from complex scenes, thereby decreasing the interference of irrelevant objects to the representation of images. In this article, we integrate self-attention mechanism into ResNet18 to achieve the attention to the salient regions of objects, as shown in the gray box in Figure 3.

The self-attention mechanism models long-distance dependencies by non-local operation, which weights all pixels according to their correlation. The importance of a region can be reflected by its weight. The greater the weight, the more important the region is. Non-local operation is implemented by cross-layer connection and element-wise addition, and then followed by ReLU. Finally, the output of the residual block is $\sigma(Y)$. ResNet18 shortens the distance between non-adjacent layers with cross-layer connection, which makes the gradient better back propagation, and alleviates the gradient degradation caused by increasing the depth in deep neural network. Besides, the redundancy of features can be effectively reduced, and existing features can be reused by cross-layer parameter sharing and intermediate features preserving.

The self-attention mechanism models long-distance dependencies by non-local operation, which weights all pixels according to their correlation. The importance of a region can be reflected by its weight. The greater the weight, the more important the region is. Non-local operation is implemented by cross-layer connection and element-wise addition, and then followed by ReLU. Finally, the output of the residual block is $\sigma(Y)$. ResNet18 shortens the distance between non-adjacent layers with cross-layer connection, which makes the gradient better back propagation, and alleviates the gradient degradation caused by increasing the depth in deep neural network. Besides, the redundancy of features can be effectively reduced, and existing features can be reused by cross-layer parameter sharing and intermediate features preserving.

\section{D. joint loss function}

There are great intra-class differences in remote sensing scenes. Generally, different seasons, locations, or sensors may lead to huge variation of images with the same category. Most remote sensing image scene classification methods optimize the network only by cross-entropy loss function, and the classification accuracy is unsatisfactory. In order to improve the discriminative ability of the model and reduce the influence of intra-class diversity, we propose a joint loss function that combines the center loss with the cross-entropy loss.

The cross-entropy loss function evaluates the difference between the probability distribution of the true label and the predicted label, which is defined as follows:

\[ L_{ce} = \frac{1}{m} \sum_{k=1}^{m} \log \frac{e^{W_{c}y_{k} + b_{k}}}{\sum_{l=1}^{n} e^{W_{c}y_{l} + b_{l}}} \]

where $m$ is the number of training samples, and $n$ is the number of categories. $v_{k} \in \mathbb{R}^{d}$ represents deep features of the $k$-th image belonging to category $c_{k}$, and $d$ is the feature dimension. $W_{c} \in \mathbb{R}^{d}$ represents the weight of the last fully connected layer in column $l$. $b$ is the bias term.

Although the cross-entropy loss function can improve the performance to a certain extent, it may deliver poor performance in classifying some difficult samples with high intra-class diversity and inter-class similarity. Center loss function is common in face recognition, which can reduce the intra-class differences effectively. Remote sensing image scene classification has a great similarity with the face recognition, so we introduce the center loss into the classification model, as shown in Figure 3. After the output feature of ResNet18, we use a fully connected layer to reduce the dimension, for avoiding large fluctuation of training caused by large calculation loss. The center loss function is defined as follows:

\[ L_{c} = \frac{1}{m} \sum_{k=1}^{m} \| v_{k} - a_{c_{k}} \|_{2}^{2} \]

where $a_{c_{k}} \in \mathbb{R}^{d}$ represents the average value of all deep features belonging to category $c_{k}$ in each mini-batch.
Finally, we combine the cross-entropy loss with the center loss to further improve the discriminative ability of network.

\[ L_{\text{joint}} = \alpha L_s + \beta L_c \]  

where \( \alpha \) and \( \beta \) are trade-off parameters, which are used to control the balance between center loss and cross-entropy loss. Under the supervision of joint loss function, the intra-class differences are reduced, the inter-class differences are enlarged, and the classification performance is further improved.

**IV. RESULTS**

**A. DATASET AND EVALUATION INDICATOR**

1) DATASET

In order to verify the performance of the proposed model, we carry out experiments on UC Merced Land-use dataset (UCM), Aerial Image dataset (AID) and NWPU-NESISC45 dataset. Next, we will introduce the details of the three datasets.

UC Merced Land-use dataset [26] contains 21 remote sensing image scene categories, which are agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each category contains 100 images, each of which has the size of 256 × 256, three channels of R, G, and B, and the spatial resolution of 0.3m. Figure 5 shows some samples of this dataset.

There are 10000 remote sensing images in Aerial Image dataset [5], which are divided into 30 categories. They are airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. The number of images in different categories of scenes ranges from 220 to 420, and the size of each image is 600 × 600. Some samples of this dataset are shown in Figure 6.

NWPU-NESISC45 dataset [6] is relatively complex, with a total of 31500 images, divided into 45 categories, and each...
FIGURE 7. Example images of NWPU-NESISC45 dataset. (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) snow berg. (39) sparse residential. (40) stadium. (41) storage tank. (42) tennis court. (43) terrace. (44) thermal power station. (45) wetland.

category contains 700 images of size $256 \times 256$. These 45 categories are airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snow berg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland respectively. Figure 7 shows some samples of this dataset. Due to the small inter-class differences and the large intra-class differences, this dataset is a pretty challenge.

2) EVALUATION INDICATOR

We adopt confusion matrix and overall accuracy to evaluate the effectiveness of the proposed model.

Confusion matrix is a special matrix which is widely utilized to measure the performance of the algorithm in remote sensing image scene classification. Each column of the confusion matrix represents predicted labels, and each row represents true labels. The disparity between the predicted label and the true label can be clearly recognized by the confusion matrix.

Overall accuracy is the ratio of the correct number predicted by the model on all test sets to the total number predicted.

B. EXPERIMENTAL DETAILS

The proposed model in this article is implemented by pytorch framework. In our experiments, the training epoch is set to 200, and the batch size of each epoch is 128. We adopt Adam as the parameter optimizer of the model, the initial learning rate is fixed as 0.0001. Meanwhile, the learning rate decay strategy is used to reduce the learning rate to 0.9 times for every 30 epochs. In order to decrease memory usage, the size of the training image is adjusted to $224 \times 224$. In addition, data augmentation is performed in training to develop a better model. The trade-off parameters of the cross-entropy loss and the center loss are set to 1 and 0.008 respectively. Our experiments are implemented on the GPU server, and the configuration is as follows: CPU Intel (R) Xeon (R) E5-2670 v3, 128GB memory, GeForce GTX TITAN X and 12GB of display memory.

We train and test models on UC Merced dataset, AID dataset, and NWPU-NESISC45 dataset. In the experiment, each dataset is divided into two parts: training set and test set. Since the number of images in three datasets is different, the training set and test set are randomly generated according to different ratios for different datasets. The training and test ratios for each dataset are shown in Table 1. In actual training, the training set is divided into two parts, 50% for training and 50% for validation.

We compare the proposed model with some existing remote sensing image scene classification methods to verify the effectiveness, include Color-Boosted Saliency-Guided
C. ABLATION EXPERIMENT

We verify the performance of each module in the proposed method, and show the visualization results of important areas in different situations in Figure 8. The categories in Figure 8 are “airport”, “church”, “pond”, “resort” and “river” from left to right. The higher the brightness of the color in the image, the more important the corresponding area of the image is. As shown in Figure 8(b), the pre-trained ResNet18 can recognize the important area of the image to a certain extent. As can be seen from the “airport” in Figure 8(c), self-attention mechanism makes the highlighted areas in the image more concentrated, and the redundant background is not displayed as highlighted. The self-attention mechanism enhances the attention to the salient region of the image, suppresses the interference of redundant background, and has a positive impact on scene classification. More importantly, after further optimizing the model by joint loss function, our model improves the discriminative ability and the classification performance.

For reflecting the impact of each module on the classification results more intuitively, we show the corresponding experimental results in Table 2. “” represents that the corresponding module is used. It can be seen from Table 2 that self-attention mechanism makes image features more discriminative, so the classification results on UC Merced dataset and AID dataset are significantly improved. Compared with ResNet18, the overall accuracy of our proposed model is improved by 1.45% on the UC Merced dataset, 2.16% on the AID dataset, and 0.74% on the NWPU-NESISC45 dataset.

D. PERFORMANCE ANALYSIS

1) EXPERIMENTAL RESULTS ON UC MERCED DATASET

We compare the proposed model with some state-of-the-art methods, and the results on UC Merced dataset are shown in Table 3. As can be seen from Table 3, some traditional algorithms such as Color-Boosted Saliency-Guided BOW and BOCF perform the worst, while deep network-based models such as CaffeNet, GoogLeNet, and VGG16 have greatly improved compared with traditional algorithms. Some improved deep learning-based methods,
such as pre-trained ResNet-50 + SRC and VGG-VD16 with DCF, have comparable classification accuracy. Nevertheless, our model achieves remarkable classification accuracy among all algorithms. When training ratios are 50% and 80%, overall accuracies of the proposed model achieve 95.81% and 97.43% respectively. It shows that our model improves the discrimination of features through self-attention mechanism. The joint loss function also reduces the intra-class differences, and further improves the classification ability of the model.

Figure 9 shows the confusion matrix of the proposed model when the training ratio is 80%.
Our model achieves 100% accuracy in 17 categories, and most categories can achieve more than 95% accuracy. The “buildings” and the “medium density residential” have relatively low accuracy, only 90%. This may be due to the similar spatial distribution of buildings in the “buildings” and the “medium density residential”, and the interference of these similar features may easily lead to the loss of classification accuracy.

2) EXPERIMENTAL RESULTS ON AID DATASET

Table 4 shows the overall accuracy and standard deviation of different algorithms on AID dataset, and the experimental results prove the effectiveness of the proposed model. On AID dataset, the overall accuracies of our model reach 92.61% and 95.06% when training ratios are 20% and 50% respectively. When the training set is 50%, the overall accuracy of VGG16-CapsNet is the highest in comparison algorithms, reaching 94.74%. Our algorithm is 0.32% higher than VGG16-CapsNet. Deep network has a great advantage in capturing the discriminative representations of remote sensing images. Among methods based on deep network, in addition to VGG16-CapsNet, CaffeNet with DCF and VGG-VD16 with DCF also perform well. They are 91.35% and 91.57% respectively when the training ratio is 20%, and 93.10% and 93.65% respectively when the training ratio is 50%.

Figure 10 shows the confusion matrix of the proposed model when the training ratio is 50% on AID dataset. Among 30 categories of AID dataset, nearly 85% of the classification accuracies exceed 90%, and even 6 of them reach 100%. The model performs poorly in both the “resort” and the “school”.

3) EXPERIMENTAL RESULTS ON NWPU-NESISC45 DATASET

In order to evaluate the performance of the model more comprehensively, we also carry out experiments on NWPU-NESISC45 dataset, which is more complex than UC Merced dataset and AID dataset. The proportion of training set is divided into 10% and 20%, and the experimental results are shown in Table 5. On this dataset, CaffeNet, GoogLeNet, and

### TABLE 4. The overall accuracy and standard deviation (%) of different algorithms on aid dataset.

| Method                | Training Ratios (%) |                      |
|-----------------------|---------------------|----------------------|
|                       | 20%                 | 50%                  |
| Color-Boosted Saliency-Guided BOW | 86.67 ± 0.38       | 86.67 ± 0.39         |
| CaffeNet              | 86.46 ± 0.47        | 89.53 ± 0.31         |
| GoogleNet             | 85.44 ± 0.40        | 89.39 ± 0.55         |
| VGG16                 | 86.59 ± 0.29        | 89.64 ± 0.36         |
| salm3LBp-CLM          | 86.92 ± 0.35        | 89.76 ± 0.45         |
| VGG-VD16 with DCF     | 91.57 ± 0.10        | 93.65 ± 0.18         |
| CaffeNet with DCF     | 91.35 ± 0.23        | 93.10 ± 0.27         |
| BOCF                  | 85.24 ± 0.33        | 87.63 ± 0.41         |
| VGG16-CapsNet         | —                   | 94.74 ± 0.17         |
| VGG19 + WSPM-CRC      | —                   | 92.57                |
| **Ours**              | **92.61 ± 0.32**    | **95.06 ± 0.36**     |

### TABLE 5. The overall accuracy and standard deviation (%) of different algorithms on nwpu-nesisc45 dataset.

| Method                | Training Ratios (%) |                      |
|-----------------------|---------------------|----------------------|
|                       | 10%                 | 20%                  |
| Color-Boosted Saliency-Guided BOW | 85.16 ± 0.23       | 87.05 ± 0.29         |
| CaffeNet              | 77.69 ± 0.21        | 79.85 ± 0.13         |
| GoogleNet             | 77.19 ± 0.38        | 78.48 ± 0.26         |
| VGG16                 | 77.47 ± 0.18        | 79.79 ± 0.15         |
| salm3LBp-CLM          | 85.32 ± 0.35        | 86.59 ± 0.28         |
| VGG-VD16 with DCF     | 87.14 ± 0.19        | 89.56 ± 0.25         |
| CaffeNet with DCF     | 87.59 ± 0.22        | 89.20 ± 0.27         |
| BOCF                  | 83.65 ± 0.31        | 85.32 ± 0.17         |
| **Ours**              | **88.29 ± 0.38**    | **91.54 ± 0.23**     |

![Figure 10: The confusion matrix of the proposed model on AID dataset when the training ratio is 50%](image-url)
VGG16 methods perform poorly, and the overall accuracy of classification below 80%. VGG-VD16 with DCF and CaffeNet with DCF methods are relatively good. Their overall accuracies of classification are 87.14% and 87.59% respectively when the training ratio is 10%, and 89.56% and 89.20% respectively when the training ratio is 20%. Our model performs best among all algorithms, reaching 88.29% and 91.54% when training ratios are 10% and 20% respectively. In general, the more training data, the better the results.

Figure 11 shows the confusion matrix generated by the classification results of the proposed model when the training ratio is 20% on NWPU-NESISC45 dataset. Among 45 categories, the overall accuracies of 30 categories exceed 90%, and the overall accuracy of the “cloud” even reaches 99%. The proposed model performs the worst in the “palace” and is most easily confused with the “commercial area”. This may be because the buildings in the two scenes are very similar.

V. CONCLUSIONS

In this article, a self-attention network model with joint loss is proposed to learn more discriminative features from remote sensing images and reduce the impact of intra-class diversity on representing images, thereby improving classification performance. The model integrates self-attention mechanism into ResNet18 to make the network pay attention to the most important area of the image and suppress the interference of irrelevant information. Moreover, a joint loss function combining cross-entropy loss and center loss is proposed to further optimize the network. This loss function reduces the differences of intra-class features and better distinguishes the image scenes that are easily confused. The experimental results on several common remote sensing image scene classification datasets verify the effectiveness of the proposed model, and our model achieves better performance with limited training data.

REFERENCES

[1] J. R. Taylor and S. T. Lovell, “Mapping public and private spaces of urban agriculture in Chicago through the analysis of high-resolution aerial images in Google Earth,” Landscape Urban Planning, vol. 108, no. 1, pp. 57–70, Oct. 2012.

[2] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, “Multimodal classification of remote sensing images: A review and future directions,” Proc. IEEE, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[3] Q. Hu, W. Wu, T. Xia, Q. Yu, P. Yang, Z. Li, and Q. Song, “Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping,” Remote Sens., vol. 5, no. 11, pp. 6026–6042, Nov. 2013.

[4] H. Huang, B. Zhao, and Y. Song, “Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery,” Remote Sens. Environ., vol. 214, pp. 73–86, Sep. 2018.

[5] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “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.

[6] W. Sun and R. Wang, “Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 3, pp. 474–478, Mar. 2018.

[7] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “Object detection with deep learning: A review,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[8] Z. Yuan, S. Li, and N. Li, “Multiscale deep features learning for land-use scene recognition,” J. Appl. Remote Sens., vol. 12, no. 1, p. 1, Feb. 2018.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Jun. 2015, pp. 1–9.

[10] M. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[11] S. Chai, 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.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. Eur. Conf. Comput. Vis., in Lecture Notes in Computer Science, vol. 9911, 2016, pp. 499–515.

FIGURE 11. The confusion matrix of the proposed model on NWPU-NESISC45 dataset when the training ratio is 20%.
HONGLIN WU received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently a Lecturer with the School of Computer and Communication Engineering, Changshua University of Science and Technology, Changsha, China. His current research interests include computer vision and image processing.

SHUZHEN ZHAO received the B.S. degree from the Hunan Institute of Science and Technology, Yueyang, China, in 2017. She is currently pursuing the M.S. degree with the Changsha University of Science and Technology, Changsha, China. Her current research interests include computer vision and remote sensing image processing.

IEEE Access
LIANG LI received the B.S. degree from the Hunan University of Science and Engineering, Yongzhou, China, in 2018. She is currently pursuing the M.S. degree with the Changsha University of Science and Technology, Changsha, China. Her current research interests include computer vision and remote sensing image processing.

CHAOQUAN LU received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2017, where he is currently pursuing the M.S. degree. His current research interests include scene classification of remote sensing images, pattern recognition, and deep learning.

WEN CHEN received the B.S. degree from Hunan City University, Yiyang, China, in 2018. He is currently pursuing the M.S. degree with the Changsha University of Science and Technology, Changsha, China. His current research interests include pattern recognition and deep learning.