GCSANet: A Global Context Spatial Attention Deep Learning Network for Remote Sensing Scene Classification

Weitao Chen, *Member, IEEE*, Shubing Ouyang, Wei Tong, Xianju Li, Xiongwei Zheng, and Lizhe Wang, *Fellow, IEEE*

***Abstract*—Deep convolutional neural networks have become an indispensable method in remote sensing image scene classification because of their powerful feature extraction capabilities. However, the ability of the models to extract multiscale features and global features on surface objects of complex scenes is currently insufficient. We propose a framework based on global context spatial attention (GCSA) and densely connected convolutional networks to extract multiscale global scene features, called GCSANet. The mixup operation is used to enhance the spatial mixed data of remote sensing images, and the discrete sample space is rendered continuous to improve the smoothness in the neighborhood of the data space. The characteristics of multiscale surface objects are extracted, and their internal dense connection is strengthened by the densely connected backbone network. GCSA is introduced into the densely connected backbone network to encode the context information of the remote sensing scene image into the local features. Experiments were performed on four remote sensing scene datasets to evaluate the performance of GCSANet. The GCSANet achieved the highest classification precision on AID and NWPU datasets and the second-best performance on the UC Merced dataset, indicating the GCSANet can effectively extract the global features of remote sensing images. In addition, the GCSANet presents the highest classification accuracy on the constructed mountain image scene dataset. These results reveal that the GCSANet can effectively extract multiscale global scene features on complex remote sensing scenes. The source codes of this method can be found in https://github.com/ShubingOuyangcug/GCSANet.**

***Index Terms*—Attention mechanism, feature channel, global context information, remote sensing, scene classification.**

I. Introduction

REMOTE sensing image scene classification marks remote sensing scene images with specific high-level semantic categories, which can be effectively analyzed to obtain high-level semantic information [1]. In recent years, it has become a prominent research area in the field of high-resolution remote sensing image classification [2]–[5]. It is widely used in natural resource investigation, land use and land coverage classification, disaster detection, environmental monitoring, and urban planning [6]–[8].

It is well known that remote sensing image features extracted severely limit the precision of the remote sensing image scene classification. Thus, researchers have long been devoted to extracting various effective remote sensing image feature representations to improve the accuracy [2]. These features can be divided into three major categories: manual features, middle-level features, and deep-level features.

1) Manual feature methods. Earlier scene classification methods mainly extract the local or global features of the images based on the manual features. They distinguish between the remote sensing image scenes via the spectrum, texture, and structure features, roughly covering color histograms [2], scale-invariant feature transforms [9], local binary pattern features [2] (LBP), gray level co-occurrence matrices [10], Gabor filters, and histograms of directional gradients [10], or a combination of these features [4]. However, because of a strong subjectivity and a lack of a systematic feature fusion method, the manual feature method results in poor generalization ability and less than ideal classification results on complex remote sensing scenes [4].

2) Middle-level feature method. Starting with the extraction of local attributes of image blocks, these methods map these local attributes to a dictionary or parameter space to obtain the overall feature representations with a stronger discrimination sense [11]–[15]. The bag-of-words model is the most popular middle-level feature encoding method because of its simplicity and interpretability [16]. Sridharan and Cheriyadat [17] extended and improved the visual bag-of-words model on image extraction, and then established the inner connections between manual features and high-level semantic features by gathering and integrating the manual features. However, when modeling local and image features, a potentially severe spatial information loss could happen. These losses are represented by the occurrence time of local features sharing a less closed connection with the spatial structure. Moreover, such methods usually have a high feature dimension, and hence it is difficult to balance the efficiency of the feature extraction and the classification performance.

Manuscript received October 26, 2021; revised December 18, 2021; accepted January 4, 2022. Date of publication January 11, 2022; date of current version January 21, 2022. This work was supported by the Fundamental Research Funds for the Natural Science Foundation of China under Grant U1803117, Grant 41925007, and Grant 42071430. *(Corresponding author: Lizhe Wang.)*

The authors are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: wtchen@cug.edu.cn; oysb@cug.edu.cn; weitong@cug.edu.cn; ddwhlxj@cug.edu.cn; zhengxiongwei@mail.cgs.gov.cn; lizhe.wang@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3141826

3) Deep-level feature method. As computing power continues to grow, deep learning technology has displayed an excellent classification performance in remote sensing image classification [18], remote sensing target detection [19], and hyperspectral unmixing [20]. This depends on its potent feature extraction capacity [21]–[23]. These methods usually conduct the training on a large quantity of labeled data to extract deep image features [8], [18], [24]. Among them, the convolutional neural networks (CNN) are widely used; they can be divided into three categories, depending on whether the CNN is trained and how it is used:

1) New deep network trained from scratch.
2) Pre-trained network: Cheng *et al.* [25] used a VGG [26] network and an Alex [18] network to fine-tune the CNN and introduced a variety of objective functions in metric learning to improve its capability to distinguish deep features.
3) The pre-trained CNN is exploited as a feature extractor to directly extract important deep features and reprocess them to deliver the final deep image feature representation [8], [27], [28].

In addition, a combination of various feature fusion techniques can be used to assemble different and effective deep features [5], [29]–[35] to improve the network performance.

These three methods have their own advantages and disadvantages. The CNN model trained from scratch has a small amount of data; therefore, it often has a small number of network layers, which can be designed by the researcher, and has greater flexibility [36]. However, CNNs normally need a large amount of labeled data during the training process. Thus, in the case of limited data, scene classification networks based on brand-new training tend to have over-fitting problems and poor feature generalization abilities [37], [38]. In response to these problems, which are caused by training the deep network from scratch, researchers have adopted well-trained neural networks in natural scenarios as pre-training networks [5], [30], [33], [35], [39]. This strategy usually fine-tunes the remote sensing scene datasets directly by using a pre-trained CNN model in a bid to produce better parameter initialization effects, faster network convergence, and higher classification accuracy. This classification method, with the pre-trained CNN as the feature extractor, directly conducts the extraction and classification of the remote sensing scene datasets by using the natural image-trained CNN model. Therefore, there is no need for a large amount of additional data for network retraining, thereby reducing the data demands. By fine-tuning the pre-trained CNN, this classification method enables the CNN and the remote sensing scene datasets to be more compatible. While fine-tuning pre-trained CNNs can yield significant performance, relying on pre-trained CNNs has some limitations: the learned features do not exactly suit the characteristics of the target dataset, and modifying pre-trained CNNs is inconvenient for the researcher [68]. The ability of the common CNN to extract local information from remote sensing images is not adequately powerful; however, the local information is important to scene classification. To address the issue, local attention-based networks were designed [62]. Thereafter, researchers found that by combining the global and local information or different local information, the remote sensing scene classification results were enhanced even further [63]–[65]. Moreover, because the potential relationships among scene semantics in the mentioned networks are likely to be ignored, the graph convolutional network structure is built to investigate the class dependencies to further improve the ability of CNN features representation [66], [67]. However, most of those models demand large-scale labeled data. In order to reduce the high cost of labeling, few-shot algorithms have been used to solve remote sensing scene classification problems with datasets that have limited annotation [69], [70].

Despite avoiding the resource consumption of traditional artificially designed operators, it is still difficult for the method with a CNN as the feature extractor to obtain multiscale information and global features among various surface objects. Owing to the complicated backgrounds and variable surface objects, remote sensing scene images often present the characteristics of intraclass diversity and interclass similarity, thereby decreasing the remote sensing scene classification accuracy. Intraclass diversity refers to the tremendous differences among major surface objects appearing in the same scene category. Surface objects usually vary in style, size, shape, and distribution. For example, there are different road forms in the same category of "mountain roads." The challenge of interclass similarity mainly lies in the overlapping of the same surface feature among different scenes. For instance, surface objects involving buildings and roads exist in both dense and medium-dense residential areas; additionally, there are large-scale mountains in the "mountain range" as well as the "mountain road" categories. Thus, one of the key scientific problems for remote sensing scene classification is how to obtain the global feature relationship and lessen the influence of intraclass diversity and interclass similarity.

Traditional deep CNNs are based on local convolution operations; thus, it is difficult to obtain the long-distance dependence of features. Graph convolution methods can obtain global dependencies; nevertheless, they are limited to extracting detailed image features [40]. However, the attention mechanism enables us to acquire important features by simulating how people understand and perceive images, thus providing a significant solution for remote sensing scene classification in feature capture [41]. Currently, attention methods normally carry out sampling on the whole area or other sampling strategies. With regards to the feature, the attention method needs less spatial information. It also needs to mention the connection with our previous proposed attention network (Channel-Attention-Based DenseNet) for remote sensing image scene classification, called CAD, which adopts the channel attention mechanism to give more attention to important features [41]. It has the same backbone of DenseNet121 as the GCSANet. The comparison of GCSANet and CAD can better reflect the different roles of the channel attention mechanism and the spatial attention mechanism for remote sensing scene recognition.

In addition, usually there are two data enhancement methods before training the network. The first is a data generation method based on a generative confrontation network. Although there is no need for prior knowledge, this method still conducts training on the same data type to generate more samples. However, these methods need a larger amount of data and have a more

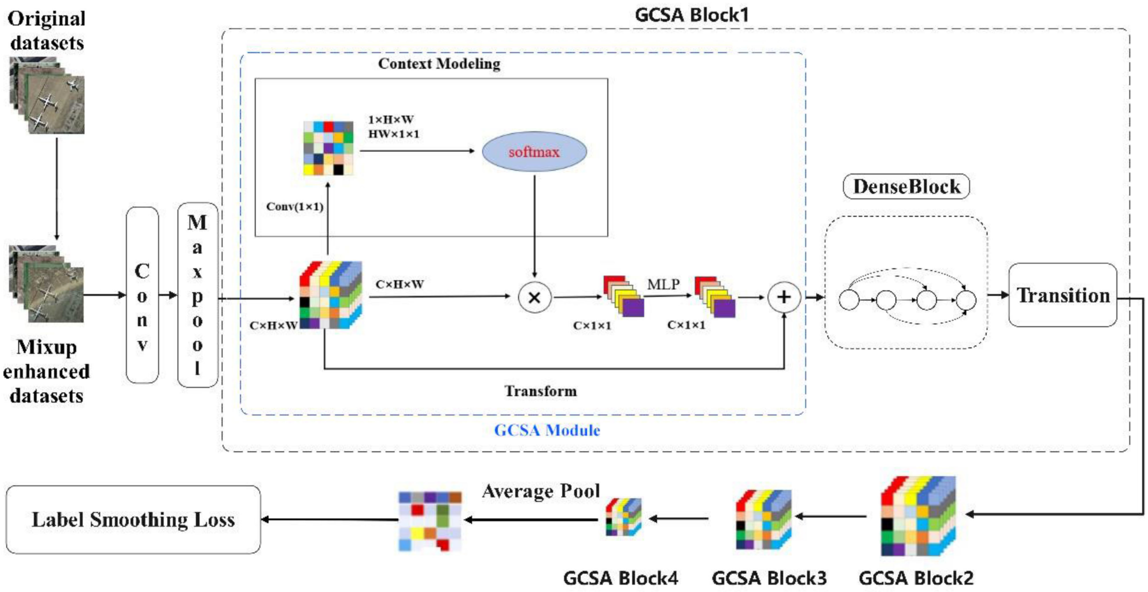


Fig. 1. GCSANet framework based on the GCSA.

difficult training convergence [43]. The second one is based on prior knowledge, which is used to describe the neighborhood around each training sample. The data enhancement method then extracts additional virtual samples from the neighboring distribution to expand the training distribution. However, most of them perform traditional data enhancement on a single image, including image flipping, image scaling, image rotation, etc. In our work, to provide smoother neighborhood dataspace, we use mix-up operations on images and corresponding labels for data enhancement of remote sensing scene [42].

To the best of our knowledge, we are the first to investigate the effectiveness of global context spatial attention, within DenseNet framework (GCSANet), for remote sensing scene classification. The GCSANet network extracts the features of multiscale surface objects, establishes internal feature connections, and introduces an attention mechanism in the spatial domain, which aims to extract the global features of remote sensing scene images. The objectives of this study are as follows.

1) Constructing a method to integrate the GCSA module and the densely connected network to improve the feature extraction ability of the scene classification model.
2) From the aspect of remote sensing scene data enhancement, we first apply a method that will improve the efficiency of data utilization and improve the smoothness of the data space in the neighborhood.
3) Creating a complicated mountain scenes' dataset and validating its generalization ability of the GCSANet.

II. Methods

The proposed GCSANet adopts DenseNet121 as the network backbone and improves the densely connected network using data enhancement and GCSA.

First, a spatially mixed data enhancement is conducted on the remote sensing scene image samples to render the discrete sample space continuous and enhance the smoothness of the data space in the neighborhood. Second, the GCSA module is designed as the transform module, which updated input features by the global context information. Then the output features into the backbone of the densely connected network to form a GCSA block, which encodes the context information of the remote sensing scene image as local features, thereby enhancing the feature extraction capability of the classification model. The GCSA block 1 includes a GCSA module, a dense block, and a transition layer for reducing the size of the feature map. The other GCSA blocks maintain the same structure and are represented by the thumbnails of the gradually decreasing feature maps. The network architecture is shown in Fig. 1.

A. Mixup-Based Data Enhancement

To make data utilization more efficient, the original training data is replaced by a mixup operation and then fed into the network. It proportionally performed a weighted summation of randomly selected images at the pixel level and integrated the existing samples in the dataset into a mixed sample. Simultaneously, through a proportionally weighted summation performance of the corresponding labels, an augmented mixed sample data was used as the virtual sample for training. For two random samples (x_i, y_i) and (x_j, y_j), where x_i and x_j represent two randomly selected training sample images from the training sets, and y_i and y_j stand for the labels of the two training samples, we have

$$\begin{aligned}\tilde{x} &= \lambda x_i + (1-\lambda)\, x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)\, y_j\end{aligned} \tag{1}$$

where $\lambda \in [0, 1]$ represents the weight of the sample. $\widetilde{x}$ represents the result of the mixing operation for two training sample images. $\tilde{y}$ represents the result of the mixing operation for two training sample labels. The linear weighting process of

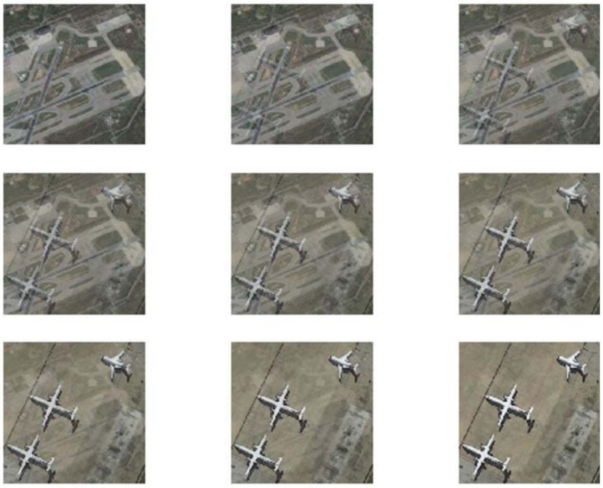


Fig. 2. Schematic diagram of the new mixed category of "Airplane" and "Airport" upon the mixup operation.

intersample regions performed by mixup enables the model to learn extra samples besides the training sample, thus diminishing the inadaptability of the data prediction targeting beyond training samples and providing a smoother uncertainty evaluation. Although it constructs virtual scenes that partially do not exist in reality, it can really mix the characteristics of different scenes in the remote sensing image by weight, as observed from the mixed images. Moreover, the data enhancement is for all training data. Although in an epoch, there is no guarantee that all training data can become mixup enhanced images after mixup operation, because when λ equals 0 or 1, the image has not been changed and remains the original image. However, in each of all epochs, the value of λ is taken randomly, and the probability that λ takes the same value for the same image is small, and the probability that it takes 0 or 1 at the same time is even smaller. In the case of augmentation of all training data, the training model upon the mixup operation is more stable than the traditional one or two models with identical network structures, training processes, and datasets. This reveals that the model is an effective data enhancement solution to over-fitting [42]. Taking the "airplane" and the "airport" categories as examples, the calculation method for the mixup on remote sensing scenes is as follows: fuse the images in the current and next input batches and send them into the neural network to obtain the $\tilde{y}$ in the form of a one-hot. Subsequently, apply the $\tilde{y}$ to perform the loss function on the labels corresponding to the two images used during the fusion. Finally, the loss is fused in the same way by the beta distribution calculation; we then adopt this loss as the final loss. The new mixed category of "airplane" and "airport" upon the mixup operation is shown in Fig. 2.

B. Backbone Network Selection

Remote sensing scene classification usually adopts CNNs such as AlexNet, VGGNet, GoogLeNet, as the network backbones [2]. Although CNNs display a remarkable ability for feature representation in scene classification, there still exist two issues: one is the potential over-fitting phenomenon, which is a result of insufficient training data; the other is a limited high-level feature extraction ability for relatively shallow network layers. In response to the latter issue, network layers deepening can be used to learn more latent and robust features. However, as the number of layers increases, issues such as gradient disappearance tend to arise and undermine the classification effect of the network model to some extent. Regarding the gradient disappearance, researchers [44]–[46] usually adopt jump connections from the front convolutional layer to the back layer to create a network model. Furthermore, a densely connected layer has been proposed to ensure full interlayer information utilization in the network [47].

In this study, a densely connected network was adopted as the backbone network. Not only can this network be used to extract the multiple features of different receptive fields, but also the features that can be reused in the interlayer. This can further integrate the features of the receptive fields on different scales, thus presenting a further complicated semantic relationship between multiscale objects in remote sensing scene images. In addition, the same DenseNet121 backbone as CAD was used for the purpose of better comparing the capabilities of the different attention mechanisms between GCSANet and CAD.

C. Global Context Information Module Establishment

The CNN model is generally based on local operations. It is difficult to learn the relationship among nonadjacent pixels in the image. To capture the long-distance dependence of nonadjacent pixels, convolutional layer stacking is used [48]. Three problems occur during the process of continuously repeated convolutions.

1) The computational efficiency is very low, and deepening the convolutional layer requires more training parameters.
2) Parameter optimization is difficult; therefore, the parameter optimization process must be carefully designed.
3) Network modeling is difficult, specifically for multilevel dependencies; information must be transmitted at different distances.

In contrast to the progressive behavior of ordinary convolution operations, nonlocal operations capture the long-distance dependence relationship by calculating the interactive features between two random locations; it does not use the notion of distance [48]. The nonlocal operation in the convolutional layer considers the weighted sum of features of all positions on a specific position as the response value of the position. The generated nonlocal features can better identify the spatial layout and object distribution of complex remote sensing scenes. The nonlocal operation stands out in several tasks, including remote sensing image classification and segmentation [34], [49], [50].

In this study, the nonlocal computing and CNN framework were integrated into a nonlocal block, as shown in Fig. 3. The input of the feature map is $T \times H \times W \times 1024$ and both mappings $\theta, \emptyset$ are in a convolutional form of $1 \times 1 \times 1$.

The calculation for the specific global context information of each location in the nonlocal network is fused into the part of feature conversion [51]. The parts of global context attention and feature conversion of the nonlocal network were improved in this article (see Fig. 3). This module reduces the number of

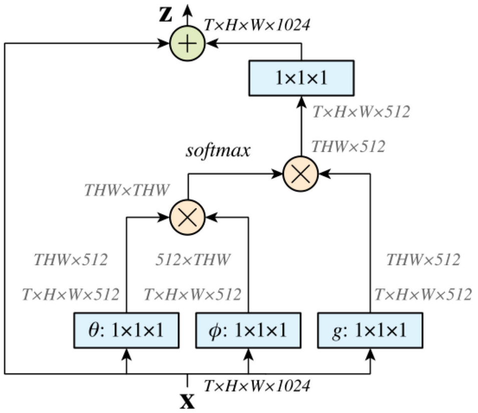


Fig. 3. Schematic diagram of the nonlocal block. ⊗ represents the matrix multiplication operation and ⊕ is the element addition operation [48]. "T×H×W×512" and "THW×512" represent the data dimensions after two consecutive data operations, the first operation is to reduce the number of channels of data from 1024 to 512, and the second operation is to reshape the data from four dimensions to two dimensions.

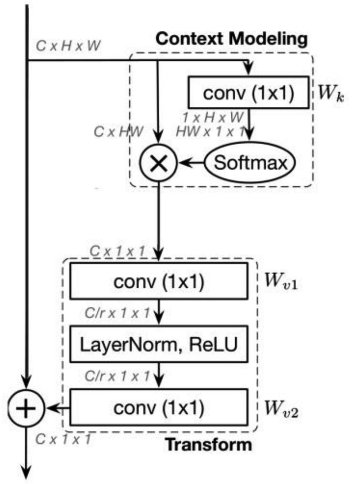


Fig. 4. Network structure schematic diagram of the global context information module. The context modeling module represents the global context information module [51].

parameters of traditional nonlocal networks and makes network training easier. As shown in Fig. 4, context modeling is used for extracting global context features which were obtained from aggregating all position features via the convolution module weight. The transform module was used to capture interchannels dependencies, and original features were updated by the global context information.

III. Dataset and Experiment Settings

The proposed GCSANet model was conducted on three open remote sensing scene datasets to verify its performance: the UC Merced (UCM) land use dataset, the AID dataset, and the NWPU-RESISC45 dataset.

A. Adopted Remote Sensing Scene Dataset

1) UCM Dataset: The UCM dataset is widely used in remote sensing image scene classification [52]. These images of dataset were manually extracted from large images of the USGS National Map Urban Area Imagery series for use in urban areas across the country. It consists of 21 land use scenes, each of them covers 100 pieces of 256 × 256-pixel images, which are RGB images with a spatial resolution of 0.3 m per pixel. Fig. 5 shows a schematic diagram of these 21 types of datasets.

2) AID Dataset: Compared to the UCM dataset, AID collected from Google Earth is more challenging and is widely used to evaluate by various remote sensing image scene classification methods [53]. First, it is a large-scale image dataset featuring a greater number of scene types and images. It contains 10 000 images, which can be divided into 30 categories, each has a size of 600 × 600 pixels. The number of images in different scene categories ranges from 220 to 420. Furthermore, there are more interclass differences, including images collected at different times and seasons under different imaging conditions across the world. The resolution of the AID data set is from about 0.5–8 m. The scene classes include airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, farmland, forest, industrial, meadow, square, stadium, storage tanks, and viaduct.

3) NWPU-RESISC45 Dataset: The NWPU-RESISC45 data set is a large-scale image dataset published in 2017. This dataset is even more complicated than the UCM and AID datasets. It contains a total of 31 500 images and 45 scenes; the images have an RGB color space. Each category contains 700 images that are 256 × 256 pixels, whose spatial resolution is about 0.2–30 m. It was created using Google Earth and it covers more than 100 regions around the world. The scene categories in NWPU-RESISC45 dataset are airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, intersection, island, lake, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, meadow, medium residential, mobile home park, mountain, overpass, palace, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station parking lot, railway, railway station, rectangular farmland, and wetland.

B. Evaluation Criteria

The overall accuracy (OA) and confusion matrix were used as the accuracy evaluation criteria in this article. The OA refers to the total number of correctly classified images divided by the number of images in the test set; this indicates the classification performance in predicting actual images. In the confusion matrix, each column represents the predicted category, and each row represents the actual data category. Therefore, the confusion matrix can directly show each distribution of the categories and simply analyze the misjudgments among different categories.

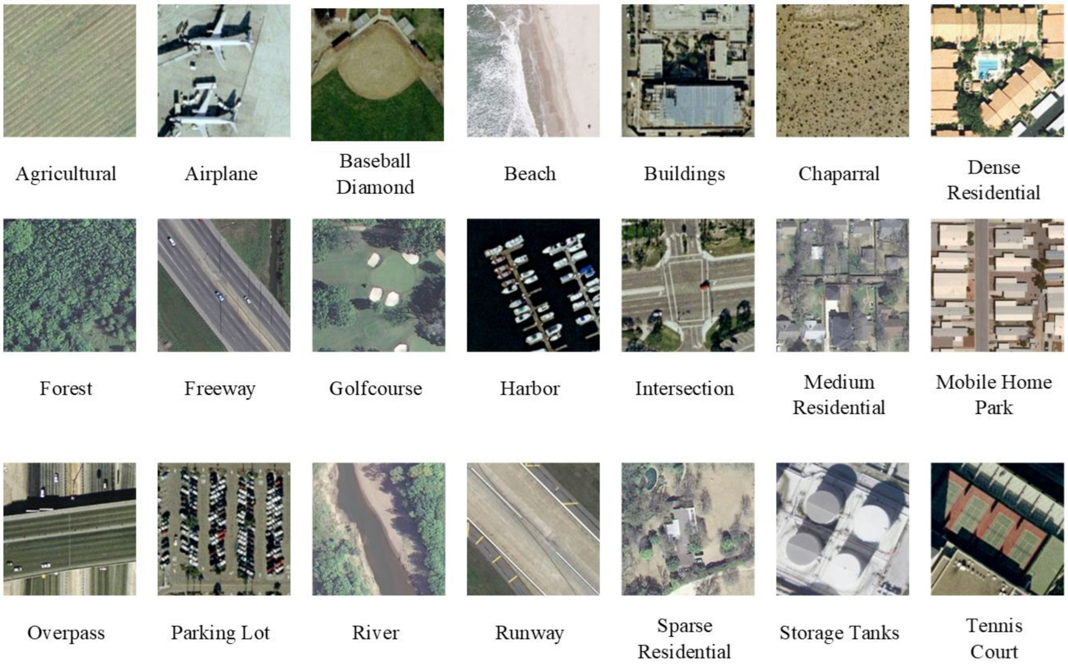


Fig. 5. Example diagram of UCM dataset.

C. Experimental Setup and Environment

To reduce the influence of randomness on the results, the dataset was randomly divided in this study, and the experiment was repeated 10 times, taking the average and standard deviation of the total accuracy as the result. In addition, to fairly evaluate the proposed GCSANet, we maintained the same training test ratio as previous studies on the open dataset; two different ratios for training versus validation were set for each dataset. The ratios for training versus validation were set at 50% and 80%, 20% and 50%, and 10% and 20% on the UCM, AID, and NWPU-RESISC45 datasets, respectively. Comparative experiments were all evaluated on the same dataset. Among them, the CAD network and GCSANet network have the same backbone to further compare the effects of two different attention mechanisms on remote sensing scene recognition.

The Pytorch deep learning framework was used to build the network model in this article. The adopted network parameters and settings were as follows: the images of the training set were input and the middle size of 288 × 288 for the three open datasets was chosen to ensure that less information about the image is lost while minimizing the amount of computation. In the training process, each dataset had 16 samples in each batch; the dataset was then inputted into the neural network for training. All our experiments were run for 100 epochs utilizing the ReduceLRonPlateau schedule with the patience of five epochs and decay factor of 0.2 to ensure that the network was fully fitted. The cross-entropy loss, combined with label smoothing [60], reduces the effect of interclass similarity in representing scene images. Whereas conventional cross-entropy loss only considers the optimal result for a single true label, label smoothing takes multiple categories of losses into account in the loss function to reduce the incidence of overfitting. The experiment was conducted on a computer equipped with dual Intel Xeon E5-2620v4 processors, two GeForce RTX 2080Ti, and 128 GB RAM.

IV. Results and Analysis

A. Experimental Results and Analysis on the UCM Dataset

To evaluate the efficiency of the GCSANet, 15 of the latest methods were used for comparison on the UCM dataset. The results are shown in Table I.

Compared to the other 15 models, the GCSANet displays a significant improvement in classification accuracy. The results show a 4.29% and 4.34% higher classification accuracy than that of the CaffeNet method at a training versus validation ratio of 80% and 50%, respectively. The performance of the CAD network is the best on the UCM data set, and the OAs of training versus validation ratio of 80% and 50% are 99.66% and 98.57%, respectively, while the corresponding results of the GCSA network are 99.31% and 98.32%, a decrease of 0.35% and 0.25%, respectively. This can be because there are less than 100 images in each dataset and there being more network parameters in the GCSANet network model than in the CAD network could result in a relatively low precision because of the under-fitted model.

The 50% training versus validation ratio in terms of the confusion matrix is generated by the classification results in the GCSANet network. The accuracy for each category is higher than 92%, and it reaches 100% in 10 of them. This shows that the classification performance of the GCSANet is stable. The ratio of the category of "dense dwellings" with the greatest

TABLE I
ACCURACY EVALUATION COMPARISON OF THE GCSANET AND 15 OTHER METHODS ON THE UCM DATASET

Method	80% ratio of training versus validation（%）	50% ratio of training versus validation (%)
CaffeNet [2]	95.02 ± 0.81	93.98 ± 0.67
GoogLeNet [2]	94.31 ± 0.89	92.70 ± 0.60
VGG-16 [2]	95.21 ± 1.20	94.14 ± 0.69
salM3LBP-CLM [54]	95.75 ± 0.80	94.21 ± 0.75
TEX-Net-LF [30]	96.62 ± 0.49	95.89 ± 0.37
LGFBOVW [16]	96.88 ± 1.32	/
Fine-tuned GoogLeNet [55]	97.1	/
Fusion by addition	97.42 ± 1.79	/
CCP-net [56]	97.52 ± 0.97	/
Two-Stream Fusion [57]	98.02 ± 1.03	96.97 ± 0.75
DSFATN [58]	98.25	/
Deep CNN Transfer [8]	98.49	/
GCFs+LOFs [59]	99.00 ± 0.35	97.37 ± 0.44
Inception-v3-CapsNet [23]	99.05 ± 0.24	97.59 ± 0.16
CAD [41]	99.66 ± 0.27	98.57 ± 0.33
GCSANet (ours)	**99.31 ± 0.56**	**98.32 ± 0.71**

TABLE II
ACCURACY EVALUATION COMPARISON OF THE GCSANET NETWORK AND 11 OTHER POPULAR MODELS ON THE AID DATASET

Method	50% ratios of training versus validation (%)	20% ratios of training versus validation (%)
CaffeNet [2]	89.53 ± 0.31	86.86 ± 0.47
GoogLeNet [2]	86.39 ± 0.55	83.44 ± 0.40
VGG-16 [2]	89.64 ± 0.36	86.59 ± 0.29
salM3LBP-CLM [54]	89.76 ± 0.45	86.92 ± 0.35
TEX-Net-LF [30]	92.96 ± 0.18	90.87 ± 0.11
Fusion by addition [27]	91.87 ± 0.36	/
Two-Stream Fusion [57]	94.58 ± 0.25	92.32 ± 0.41
GCFs+LOFs [59]	96.85 ± 0.23	92.48 ± 0.38
VGG-16-CapsNet [23]	94.74 ± 0.17	91.63 ± 0.19
Inception-v3-CapsNet [23]	96.32 ± 0.12	93.79 ± 0.13
CAD [41]	97.16 ± 0.26	95.73 ± 0.22
GCSANet（Ours）	**97.53 ± 0.32**	**95.96 ± 0.38**

chance of misclassification to be classified as “medium-dense dwellings” is 8%, owing to the extremely similar surface objects of the buildings and streets. By the same token, categories of “medium-dense dwellings” and “sparse dwellings” tend to be misclassified.

B. Experimental Results and Analysis on the AID Dataset

The results of the GCSANet and 11 other popular models on the AID dataset are shown in Table II. We can see that when the training versus validation ratio is 50% and 20%, the OAs of the GCSA network are 97.53% and 95.96%, respectively, 1.21% and 2.17% higher than the Inception-v3-CapsNet network. Compared to the CAD network, the GCSANet network performs better on the AID dataset, with an advantage of 0.37% and 0.23%, respectively. Owing to the densely connected network backbones of both methods, the GCSANet performs better than the channel attention mechanism in the CAD network and on large datasets such as the AID. Its highest classification accuracy also reveals that the GCSANet has an excellent remote sensing scene classification capability. The confusion matrix generated from the GCSANet with training versus validation ratio of 20% is shown in Fig. 6. It can be seen that the classification accuracy in 26 categories is higher than 90% and exceeded 83% in the other ones. Several categories with difficult identification on the UCM data set, involving “sparse residential,” “medium residential,” and “dense residential,” obtained a 99.6%, 97.4%, and 97.6% classification accuracy, respectively. The classification accuracies of the “school” and “central area” were relatively low at 83.8% and 88%, respectively. The reason is that they share similar architectural arrangements. Similarly, it is easy to confuse

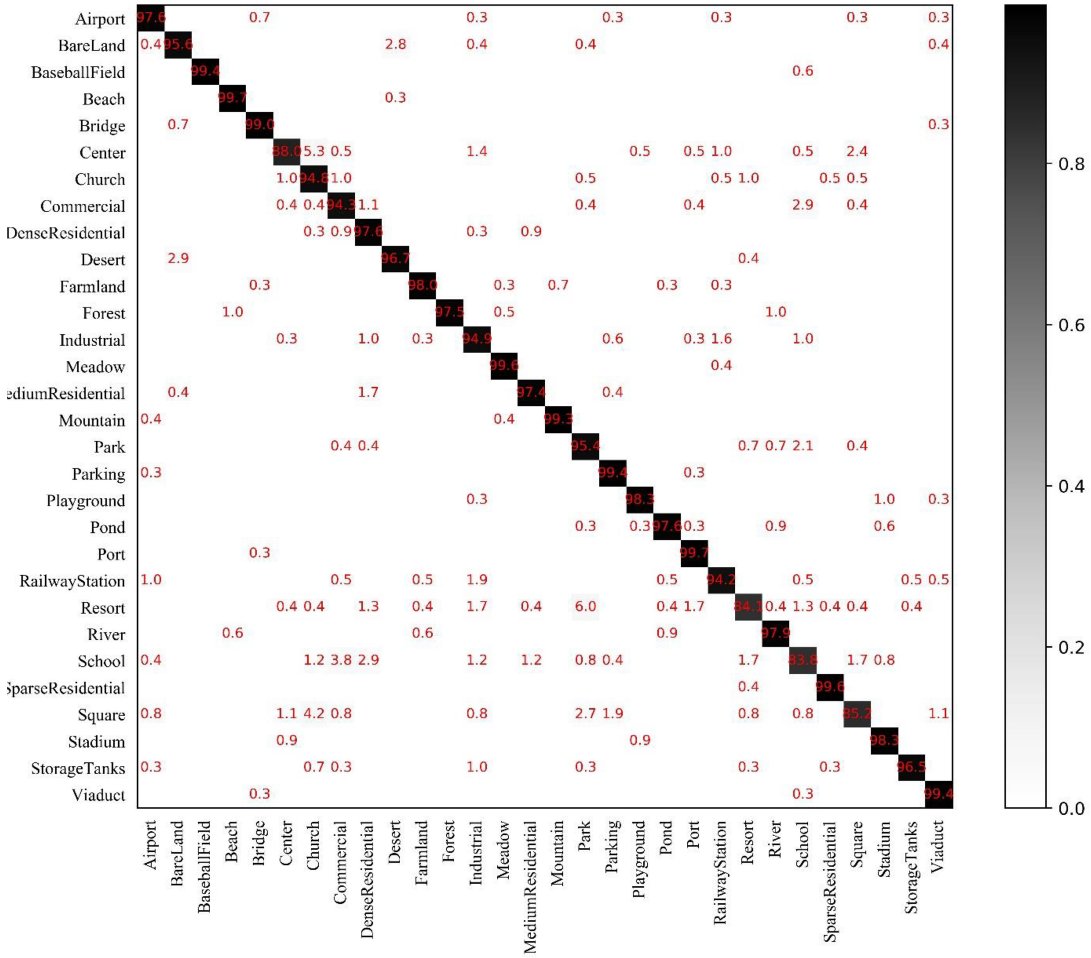


Fig. 6. Confusion matrix of the GCSANet on the AID dataset.

"school" and "commercial area," "square" and "church," and "resort" is easily confused with the "park" because considerable vegetation and house coverage exist in both images. Nevertheless, when compared to the 49% and 60% classification accuracy of the VGGNet-16 [2], the GCSANet network made remarkable progress, indicating that it can effectively identify images on the AID remote sensing scene dataset with interclasses similarity.

C. Experimental Results and Analysis on the NWPU-RESISC45 Data Set

The experimental results on the most challenging dataset, NWPU-RESISC45, comparing the GCSANet with 12 other popular models are listed in Table III. With ratios for training versus validation at 20% and 10%, the classification accuracy of the GCSANet is 94.95% and 93.39% respectively, 2.35% and 4.36% higher than that of the Inception-v3-CapsNet network. The results indicate that the GCSANet is more effective for large-scale remote sensing scene data. A 3.57% difference was recorded between the 20% and 10% ratios of training versus validation for the Inception-v3-CapsNet network, while the gap was only 1.56% for the GCSANet. This indicates that the GCSANet has a powerful learning capacity for image feature representations on large-scale datasets.

For the confusion matrix generated from the GCSANet with training versus validation ratio of 20% on the NWPU-RESISC45 dataset, the accuracy of the "palace" scene is only 83%. The reason is that some samples are mistakenly classified as "church," because of certain similarities in the buildings and streets. Furthermore, parts of the "rectangular farmland" are identified as "terraced fields" because of the similar surface cover characteristics. Overall, the classification accuracy of 40 categories was higher than 90%, while the remainder exceeded 83%; however, the classification accuracy of only three categories was higher than 90% in the CAD network, which proves the superior robustness and classification effect of the GCSANet.

V. Discussion

A. Ablation Experiment on the GCSANet

To further analyze the role of each module in the GCSANet, we carried out ablation experiments on the AID and NWPU datasets, including the attention mechanism without the global context and the mixup operation.

Table IV shows the classification results. The average accuracy of the GCSANet was 1.69% higher than that of the GCSANet without the global context attention mechanism. This indicates that the GCSANet framework has a potent feature

TABLE III
ACCURACY RESULTS OF THE GCSANET AND 12 OTHER POPULAR METHODS ON THE NWPURESISC45 DATASET

Method	20% ratio of training versus validation（%）	10% ratio of training versus validation（%）
GoogLeNet [2]	78.48 ± 0.26	76.19 ± 0.38
VGG-16 [2]	79.79 ± 0.15	76.47 ± 0.18
AlexNet [2]	79.85 ± 0.13	76.69 ± 0.21
Two-Stream Fusion [57]	83.16 ± 0.18	80.22 ± 0.22
BoCF [7]	84.32 ± 0.17	82.65 ± 0.31
Fine-tuned AlexNet [2]	85.16 ± 0.18	81.22 ± 0.19
Fine-tuned GoogLeNet [2]	86.02 ± 0.18	82.57 ± 0.12
Fine-tuned VGG-16 [2]	90.36 ± 0.18	87.15 ± 0.45
Triple networks [61]	92.33 ± 0.20	/
VGG-16-CapsNet [23]	89.18 ± 0.14	85.08 ± 0.13
Inception-v3-CapsNet [23]	92.60 ± 0.11	89.03 ± 0.21
CAD [41]	94.58 ± 0.26	92.7 ± 0.32
GCSANet（Ours）	**94.95 ± 0.36**	**93.39 ± 0.39**

TABLE IV
ABLATION EXPERIMENT RESULTS OF THE GCSANET ON BOTH THE AID AND NWPU DATASETS

	GCSANet without Global Context Attention (%)	GCSANet without Mixup operation (%)	GCSANet (ours)（%）
AID (50%)	95.99 ± 0.49	96.88 ± 0.42	97.53 ± 0.32
AID (20%)	94.71 ± 0.32	95.52 ± 0.51	95.96 ± 0.38
NWPU (20%)	93.14 ± 0.28	94.49 ± 0.38	94.95 ± 0.36
NWPU (10%)	91.22 ± 0.41	92.75 ± 0.47	93.39 ± 0.39
AVERAGE	93.77	94.91	95.46

extraction capability. This is especially true in the case of a backbone network with high accuracy, where the GCSANet is still able to achieve an accuracy improvement. In addition, the average classification accuracy of the GCSANet only using the mixup operation increased by 0.55%, showing the effect of continuous-discrete spatial samples in this article.

B. Generalization Ability of GCSANet Network

1) Classification Performance on Complicated Mountain Scene Data Sets: Currently, owing to the lack of public mountain remote-sensing scene datasets, we created a dataset in a mountain area covering 2589 km^2 using China's Ziyuan-3 satellite images to further verify the generalization ability of the proposed GCSANet. The Ziyuan-3-02, whose imaging time was December 17, 2018, and was obtained without cloud coverage, was used to generate the dataset used in this article. First, we used the ENVI 5.3 software to extract the digital terrain model of the area based on the Ziyuan-3-02 stereo pair data. Second, we extracted a digital terrain model to ortho-rectify the multispectral and panchromatic images. Thereafter, we fused the rectified multispectral image with the panchromatic image using the pan-sharpening method to produce a fused image with a resolution of 2.1 m.

The total number of image patches was 1060, including six types of scene images: farmland, residential area, mountains, mountain roads, rivers, and terraces; these are listed in Table V. The size of the patch was 256 × 256 pixel. There are patchy towns distributed along the river, roads connecting towns, farmland, and rivers, as well as overlapping terraces and farmland.

To examine the validity of the dataset, we refer to Google Earth and compare it to the data of the third national land survey. The mountain scene datasets can be downloaded.[1] Some typical patches are shown in Fig. 7. Unlike the land cover images of the public dataset, first, the image was collected from Zhiyuan-3. Second, the dataset exhibits a high degree of similarity between classes in the background of the same complex mountain scenario, for example, "mountain roads" are often existed in the background of "mountains." It is also a challenge to verify the generalization ability of the GCSANet network.

[1] [Online]. Available: https://pan.baidu.com/s/18NW5syly4WnrHOXDO8uSmQ, watchdog: cug0

TABLE V
NUMBER OF EACH CATEGORY IN THE MOUNTAIN SCENE DATASET

Category	Farmland	House	Mountain	Mountain Road	River	Terrace
Number	180	200	200	100	180	200

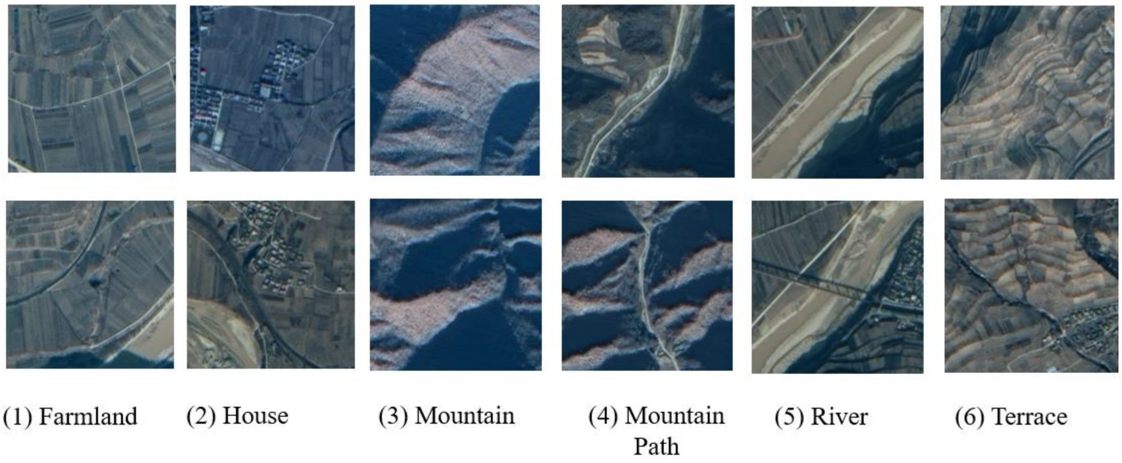


Fig. 7. Sample image of a mountainous scene data set.

TABLE VI
PERFORMANCE COMPARISONS OF THE GCSA AND SIX OTHER POPULAR MODELS UNDER DIFFERENT RATIOS OF TRAINING VERSUS VALIDATION ON THE MOUNTAIN REMOTE SENSING SCENE DATASET

Models	ACCURACY		FLOPs (G)	Params (M)
	70% ratios of training versus validation（%）	50% ratios of training versus validation（%）		
AlexNet[18]	92.52 ± 0.97	90.21 ± 0.89	**1.86**	61.1
GoogLeNet[2]	93.73 ± 0.72	91.38 ± 0.93	3.94	**6.62**
VGG-11[2]	94.51 ± 0.81	92.70 ± 0.96	19.82	132.86
VGG-16[2]	94.53 ± 0.59	92.45 ± 0.83	40.36	138.36
ResNet-18[45]	94.63 ± 0.84	92.99 ± 0.59	4.75	11.69
CAD[41]	95.67 ± 0.68	93.85 ± 0.79	7.51	14.1
GCSANet (ours)	**95.85 ± 0.49**	**94.13 ± 0.52**	7.50	8.11

The results of the GCSANet which applying the mixup operation and other six popular methods on the mountainous scene data set are shown in Table VI. For the 70% and 50% ratios of training versus validation, the OAs of the GCSANet are 95.85% and 94.13%, respectively, 1.2% and 1.1% higher than that of the ResNet-18. The average accuracy of the GCSANet is higher than that of the CAD network with a feature channel attention mechanism [41]; meanwhile, the variances of the GCSANet in each of these cases were 0.49 and 0.52, respectively, a 0.19 and 0.27 reduction than that of the CAD network. These indicate that the GCSANet is more stable in terms of classification than the feature channel-based attention mechanism on the mountain scenes dataset. Moreover, the GCSANet of 8.11M parameters is 1.49% higher than the GoogLeNet of the lowest parameters on all methods. This demonstrates that the GCSANet is a relatively lightweight network.

2) Prediction Analysis on Results of Remote Sensing Scene Datasets in Mountainous Areas: In the 70% and 50% ratios of training versus validation, the prediction accuracy of the GCSANet is 95.85% and 94.13%, respectively. However, the “mountain road” category is only 79.6%. And 20.4% of them are mistakenly classified as “mountains” as a result of a relatively high level of confusion between them as shown in Fig. 8. We can see that the misclassified mountains tend to lie in the images with mountain features. When the road is small in proportion to the mountain patches, it becomes more difficult to distinguish it owing to the large interclasses similarity.

Furthermore, to reveal the effect of sample spatial independence on the classification accuracy, we carried out an experiment based on sample validation and test dataset in discontinuous regions. The results show that the average accuracy of prediction is 93.36% and 91.53% when the ratios of training

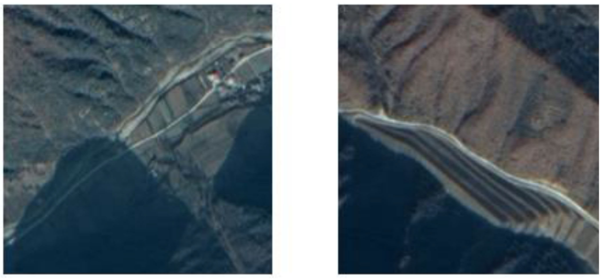


Fig. 8. Image feature of “Mountain Road” scene was mistakenly divided into the “mountain” scene.

versus validation are 70% and 50%, respectively, which is a small decrease. This indicates that the spatial independence of samples has a certain bearing on the mountain scenes classification of the GCSANet.

VI. Conclusion

The ability to extract multiscale features of local and global features of ground objects is insufficient for remote sensing image scene classification. Thus, the GCSANet based on GCSA was proposed in this article. First, experiments were carried out on three open remote sensing scene datasets to evaluate the performance of the GCSANet. Second, to validate the generalization ability of the GCSANet, experiments were carried out on the self-built challenging mountain remote sensing scene dataset. The main conclusions are as follows. First, the GCSANet can effectively extract the global features of remote sensing scene images compared to other popular models. Furthermore, it has a stronger learning ability for large and complex scene datasets. The variance of accuracy is smaller under different proportions of training sets versus validation sets, and its robustness and stability are better. Second, the mixup operation can effectively enhance the smoothness and classification accuracy of the data space in the neighborhood and improve the utilization efficiency of the remote sensing scene sample data. Finally, the GCSANet presents the highest classification accuracy among other popular models on the mountain image scene dataset and is more stable than the feature channel attention network with high precision. Owing to certain limitations, such as inefficient dataset, we will further develop an improved GCSANet framework to ensure stronger transferability and add more comparative experiments to demonstrate the strong classification accuracy of the model.

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] C. Gong, 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, Apr. 2017.

[3] G. Yue, S. Jun, L. Jun, and W. Ruoyu, “Remote sensing scene classification based on high-order graph convolutional network,” *Eur. J. Remote Sens.*, vol. 54, no. sup1, pp. 141–155, Feb. 2021.

[4] B. Luo, S. Jiang, and L. Zhang, “Indexing of remote sensing images with different resolutions by multiple features,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.

[5] 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, doi: 10.1080/2150704X.2017.1415477.

[6] 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, Jan. 2017, Art. no. 23, doi: 10.3390/rs10010023.

[7] C. Gong, Z. Li, X. Yao, G. Lei, 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, Aug. 2017.

[8] F. Hu, G. 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, Nov. 2015, doi: 2015 10.3390/rs71114680.

[9] Y. Yang and S. Newsam, “Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery,” in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, 2008, pp. 1852–1855.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, “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.

[11] Z. Li, Z. Zhou, and D. Hu, “Scene classification using a multi-resolution bag-of-features model,” *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, Jan. 2013, doi: 10.1016/j.patcog.2012.07.017.

[12] L. Zhao, T. Ping, 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. 5/6, pp. 2296–2310, Mar. 2014, doi: 10.1080/01431161.2014.890762.

[13] L. Zhao, P. Tang, and L. Huo, “Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model,” *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 7, no. 12, pp. 4620–4631, Aug. 2014.

[14] S. Chen and Y. Tian, “Pyramid of spatial relations for scene-level land use classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[15] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[16] Q. Zhu, Y. Zhong, Z. Bei, G. Xia, and L. Zhang, “The bag-of-visual-words scene classifier combining local and global features for high spatial resolution imagery,” in *Proc. 12th Int. Conf. Fuzzy Syst. Knowl. Discov.*, 2015, pp. 717–721.

[17] H. Sridharan and A. Cheriyadat, “Bag of lines (BoL) for improved aerial scene representation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 676–680, Mar. 2015.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[19] G. Xia *et al.*, “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[20] X. Lu, H. Wu, and Y. Yuan, “Double constrained NMF for hyper-spectral unmixing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2746–2758, May 2014.

[21] E. Othmana, 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, May 2016.

[22] 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.

[23] W. Zhang, P. Tang, and L. Zhao, “Remote sensing image scene classification using CNN-CapsNet,” *Remote Sens.*, vol. 11, no. 5, pp. 494, Feb. 2019, doi: 10.3390/rs11050494.

[24] A. Ma, Y. Wan, Y. Zhong, J. Wang, and L. Zhang, “SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search,” *ISPRS J. Photogramm.*, vol. 172, pp. 171–188, Feb. 2021, doi: 10.1016/j.isprsjprs.2020.11.025.

[25] 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.*, vol. 56, no. 5, pp. 2811–2821, Jan. 2018.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014.

[27] 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.

[28] 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, Jun. 2017.

[29] 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, Jun. 2014, doi: 10.1080/17538947.2014.925517.

[30] R. M. Anwer, F. S. Khan, J. van de Weijer, and M. Molinier, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm.*, vol. 138, pp. 74–85, 2018, doi: 10.1016/j.isprsjprs.2018.01.023.

[31] 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, May 2017.

[32] 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, Jan. 2018, doi: 10.1109/ACCESS.2018.2798799.

[33] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.

[34] H. Wang, Q. Hu, C. Wu, J. Chi, and X. Yu, "Non-locally up-down convolutional attention network for remote sensing image super-resolution," *IEEE Access*, vol. 8, pp. 166304–166319, 2020.

[35] 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.

[36] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.-Basel*, vol. 10, no. 3, pp. 444, Mar. 2018, doi: 10.3390/rs10030444.

[37] K. Nogueira, O. Penatti, and J. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2016, doi: 10.1016/j.patcog.2016.07.001.

[38] J. Chen, C. Wang, Z. Ma, J. Chen, D. He, and S. Ackland, "Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters," *Remote Sens.*, vol. 10, no. 2, pp. 290, Feb. 2018, doi: 10.3390/rs10020290.

[39] B. Yuan, L. Han, X. Gu, and H. Yan, "Multi-deep features fusion for high resolution remote sensing image scene classification," *Neural Comput. Appl.*, vol. 33, no. 6, pp. 2047–2063, Mar. 2021, doi: 10.1007/s00521-020-05071-7.

[40] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Comput. Vis. Image Understand*, vol. 184, pp. 22–30, 2019, doi: 10.1016/j.cviu.2019.04.004.

[41] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, Jul. 2020, doi: 10.1109/JSTARS.2020.3009352.

[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, Feb. 2018.

[43] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "BAGAN: Data augmentation with balancing GAN," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2018.

[44] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[46] H. Gao, S. Yu, L. Zhuang, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 646–661.

[47] G. Huang, Z. Liu, L. Van Der Maten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[48] X. L. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[49] R. Lei, C. Zhang, S. Du, C. Wang, and M. Yu, "A non-local capsule neural network for hyperspectral remote sensing image classification," *Remote Sens. Lett.*, vol. 12, no. 1, pp. 40–49, Jan. 2021, doi: 10.1080/2150704X.2020.1864052.

[50] M. Zhang, Q. Cheng, F. Luo, and L. Ye, "A triplet non-local neural network with dual-anchor triplet loss for high resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2711–2723, Feb. 2021, doi: 10.1109/JSTARS.2021.3058691.

[51] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1971–1980.

[52] Y. Yi and N. Shawn, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.

[53] G. 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.

[54] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Apr. 2017.

[55] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *Acta Ecologica Sinica*, vol. 28, no. 2, pp. 627–635, 2015, doi: 10.1016/S1872-2032(08)60029-3.

[56] K. Qi, Q. Guan, Y. Chao, 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, Jun. 2018, Art. no. 934, doi: 10.3390/rs10060934.

[57] Y. Yu and F. Liu, "Dense connectivity based two-stream deep feature fusion framework for aerial scene classification," *Remote Sens.*, vol. 10, no. 7, Jul. 2018, Art. no. 1158, doi: 10.3390/rs10071158.

[58] 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, doi: 10.3390/rs10030410.

[59] 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, doi: 10.3390/rs10050734.

[60] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.

[61] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Oct. 2017.

[62] Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, and D. Li, "Global-local attention network for aerial scene classification," *IEEE Access*, vol. 7, pp. 67200–67212, 2019.

[63] R. Fan, L. Wang, R. Feng, and Y. Zhu, "Attention based residual network for high-resolution remote sensing imagery scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1346–1349.

[64] J. Shen, T. Zhang, Y. Wang, R. Wang, Q. Wang, and M. Qi, "A dual-model architecture with grouping-attention-fusion for remote sensing scene classification," *Remote Sens.*, vol. 13, 2021, Art. no. 433, doi: 10.3390/rs13030433.

[65] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, Jan. 2021, doi: 10.1109/JSTARS.2021.3051569.

[66] Y. Gao, J. Shi, J. Li, and R. Wang, "Remote sensing scene classification based on high-order graph convolutional network," *Eur. J. Remote Sens.*, vol. 54, pp. 141–155, 2021, doi: 10.1080/22797254.2020.1868273.

[67] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, Apr. 2021, doi: 10.1109/TNNLS.2021.3071369.

[68] G. Cheng, X. Xie, J. Han, L. Guo, and X. GS, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020, doi: 10.1109/JSTARS.2020.3005403.

[69] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7844–7853, Sep. 2021.

[70] G. Cheng *et al.*, "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–11, Jul. 2021, doi: 10.1109/TGRS.2021.3099033.

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

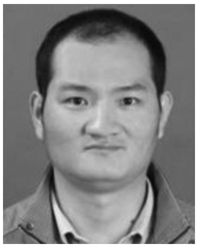
He is a Professor with the School of Computer Science, China University of Geosciences. He has authored or coauthored more than 30 papers. His main research interests include machine learning and remote sensing of environment.

Shubing Ouyang was born in Minhou County, Fuzhou City, Fujian Province, China, in 1990. She received the B.S. degree in geology from the Wuhan University of Engineering Science, Wuhan, China, in 2012, and the M.S. degree in mineral resource prospecting and exploration from the China University of Geosciences, Wuhan, China, in 2015. She is currently working toward the Ph.D. degree in geoscience information engineering with the School of Computer Science, China University of Geosciences, Wuhan, China.

Her research interests include geoscience information processing, remote sensing image processing, and deep learning.

Wei Tong received the B.S. degree in electronic information engineering from the Wuhan University of Technology, Wuhan, China, in 2018, and the M.S. degree in computer technique with the School of Computer Science, China University of Geosciences, Wuhan, China, in 2021.

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



Xianju Li received the B.S. degree in geomatics 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 authored or coauthored more than ten papers. His main research interests include remote sensing image processing and analysis, computer vision, and machine learning.

Xiongwei Zheng born in Tianmen, Hubei Province, China. He received the bachelor's degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2003. He is currently working toward the doctor's degree in geoscience information engineering from China University of Geosciences, Wuhan, China.

He is a professor level Senior Engineer. He is the Director of Big Data Center of China Airborne Geophysical and Remote Sensing Center for natural resources. His research interests include data acquisition and processing of satellite multi spectral, hyperspectral, laser sounding, and radar remote sensing.

Lizhe Wang (Fellow, IEEE) received the B.E. and M.E. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1998, 2001, respectively, and the Doctor of Engineering degree from University Karlsruhe (Magna Cum Laude), Germany.

He is a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include HPC, e-Science, and remote sensing image processing.

Prof. Wang is a Fellow of IET and British Computer Society.