Front. Earth Sci. 2024, 18(3): 538−553
https://doi.org/10.1007/s11707-022-1079-x

RESEARCH ARTICLE

A Grad-CAM and capsule network hybrid method for remote sensing image scene classification

Zhan HE[1,4], Chunju ZHANG (✉)[2], Shu WANG (✉)[3], Jianwei HUANG[4], Xiaoyun ZHENG[1], Weijie JIANG[4], Jiachen BO[4], Yucheng YANG[4]

1 Shenzhen Data Management Center of Planning and Natural Resources, Key Laboratory of Urban Land Resources Monitoring and Simulation (Ministry of Natural Resources), Shenzhen 518000, China
2 Key Laboratory of Jianghuai Arable Land Resources Protection and Eco-restoration (Ministry of Natural Resources), Hefei 230088, China
3 State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
4 School of Civil Engineering, Hefei University of Technology, Hefei 230009, China

Abstract Remote sensing image scene classification and remote sensing technology applications are hot research topics. Although CNN-based models have reached high average accuracy, some classes are still misclassified, such as "freeway," "spare residential," and "commercial_area." These classes contain typical decisive features, spatial-relation features, and mixed decisive and spatial-relation features, which limit high-quality image scene classification. To address this issue, this paper proposes a Grad-CAM and capsule network hybrid method for image scene classification. The Grad-CAM and capsule network structures have the potential to recognize decisive features and spatial-relation features, respectively. By using a pre-trained model, hybrid structure, and structure adjustment, the proposed model can recognize both decisive and spatial-relation features. A group of experiments is designed on three popular data sets with increasing classification difficulties. In the most advanced experiment, 92.67% average accuracy is achieved. Specifically, 83%, 75%, and 86% accuracies are obtained in the classes of "church," "palace," and "commercial_area," respectively. This research demonstrates that the hybrid structure can effectively improve performance by considering both decisive and spatial-relation features. Therefore, Grad-CAM-CapsNet is a promising and powerful structure for image scene classification.

Keywords image scene classification, CNN, Grad-CAM, CapsNet, DenseNet

Received September 29, 2022; accepted February 15, 2023

E-mails: zcjtwz@sina.com (Chunju ZHANG)
wangshu@igsnrr.ac.cn (Shu WANG)

1 Introduction

Remote sensing image scene classification aims to classify different scene images into different sections with explicit classes, that is, to understand different images and assign fine-grained specific semantics to different patches (Cheng et al., 2017a). Fine-grained specific semantics in images provide detailed information on objects, types, spatial relationships, and sequential relationships, which can greatly enhance fine-grained management and decision-making abilities in various fields, such as transportation, military, disaster monitoring, land resource management, and urban construction planning (Pan et al., 2020; Raiyani et al., 2021). Thus, remote sensing image scene classification and remote sensing technology applications are hot research topics.

Image scene classification has undergone two key development stages: the traditional machine learning stage and the deep learning stage (Cheng et al., 2017b; Pires de Lima and Marfurt, 2019). The traditional machine learning stage treats image scene classification in two independent parts: feature extraction and classification. In feature extraction, handcrafted features are selected with common methods, such as scale-invariant feature transform (SIFT), sparse representations, and histogram of oriented gradient (HOG) (Sheng et al., 2012; Vo et al., 2015; Gan et al., 2016). In the classification, logistic regression (LR), random forest (RF), and support vector machines (SVM) are used as classifiers (Knorn et al., 2009; Bai, 2016; Ahmed et al., 2020). Combined with handcrafted features and classical classifiers, complex scene images with abstract semantic information are not satisfactory for classifying images because handcrafted

features, for example, texture features, are at a low level for complex scene classification (Zhao et al., 2016; Sun et al., 2021). Thus, high-level features must be considered in the next stage.

In the deep learning stage, a breakthrough method, convolutional neural networks (CNN), stands out with its automatic high-level feature generation mechanism, which has made great achievements in the remote sensing scene classification field (Cheng et al., 2016; Cheng et al., 2018). Improvements in the structure, such as multiple network connections, feature fusion, attention mechanisms, and other relevant issues (Marmanis et al., 2016; Chaib et al., 2017; Tong et al., 2020; Zhao et al., 2020). CNN-based models have reached high accuracy, but the ability of CNNs to identify the relative spatial position information between features is insufficient. The existence of a pooling layer and a fully connected layer in the CNN leads to the inadequacy of the ability to capture relative spatial position information between features. CNN has translation invariance after integrating features through a pooling layer. This ability to focus only on features while ignoring location information between features is applicable to tasks such as target segmentation and detection. However, in the face of complex remote sensing image scene classification tasks, the CNN's ability significantly decreases. The fully connected layer reduces the multidimensional features to one-dimensional features and loses the spatial position information of the features. Therefore, CNN-based models have reached high accuracy, however, the high interclass similarity still leads to misclassification, making scene classification a challenging task. For example, the classification accuracy of 'freeway' is only 73%, while 7% are misclassified as 'railway,' and the classification accuracy of 'medium residential' is 77%, while 8% are misclassified as 'sparse residential' and 'dense residential' (Yu and Liu, 2018b). As shown in Fig. 1, nine images (the above three are labeled 'freeway,' 'railway,' and 'runway'; the middle three are labeled 'dense residential,' 'medium residential,' and 'sparse residential;' the bottom three are labeled 'church,' 'palace,' and 'commercial area') are taken from the NWPU-RESISC45 data set. The above three images have similar components, such as roads and vegetation, however there are decisive details in each scene, such as cars on the 'freeway' and trains on the 'railway'. The middle three images have a high degree of similarity to common objects such as houses, vegetation, and roads, but there may be rich and complex relationship details

Fig. 1 Scene images taken from the NWPU-RESISC45 data set showing the similarity of different land covers, (a) freeway; (b) railway; (c) runway; (d) dense residential; (e) medium residential; (f) sparse residential; (g) church (h) palace; (i) commercial area.

such as density, spacing, and arrangement of the houses in these scenes. The bottom three images have both decisive details and complex relationship details. According to the analyses of the two conditions above, two kinds of features caught our attention that may be the reason for the misclassification results. One is the decisive features around the target objects because the discriminative ability of the decisive feature among several similar images is not strong enough in the CNN-based model. Another is the spatial relationship features between different target objects because the fully connected layers in the CNN model ignore the relative spatial information. For example, the distance and arrangement between buildings are important factors for correctly classifying 'sparse residential,' 'medium residential,' and 'dense residential.' How to effectively solve these two problems together and improve the classification accuracy is what we discuss in this paper.

To improve the decisive features and spatial relationship features, two kinds of relevant studies have inspired us. The first is the attention mechanism, which can make the model focus on more effective information and ignore invalid information. Many attention mechanism applications have been used in deep learning models with improved classification performances (Zhao et al., 2017; Chen et al., 2018; Mei et al., 2019). Notably, a recently adjusted attention mechanism algorithm called gradient-weighted class activation mapping (Grad-CAM) achieved state-of-the-art performance (Li et al., 2020). Grad-CAM can generate an attention map through a pre-trained model. The pixel values in each attention map condenses more decisive features of the corresponding images. Thus, Grad-CAM can achieve more accurate classification results. Many researchers have explored adding Grad-CAM to neural networks to achieve better results for different tasks, such as image-level weakly supervised semantic segmentation (Wang et al., 2020), mobile network design (Hou et al., 2021), and real-time semantic segmentation (Yu et al., 2021). The second method uses the capsule network (CapsNet), which is a novel method with a more effective encoding ability in spatial information (Sabour et al., 2017). It uses a group of neurons as a capsule to replace the traditional single neuron. A group of neurons constitutes a vector. These vectors represent specific spatial features of specific entities in the image on different attributes. This mechanism makes the network analyze and recognize the relationship between different features in images. Therefore, it can effectively improve the final classification accuracy (Zhang et al., 2019; Lei et al., 2021, 2022; Zhang et al., 2022).

It is suspected that when a combination of the Grad-CAM and CapsNet is used, the misclassification results with lower decisive features and spatial relationship features will be fixed. Although the D-CapsNet model attempts to combine the attention mechanism and the CapsNet structure, it uses an attention mechanism in convolutional layers that ignores the original decisive details and complex relationship details (Raza et al., 2020). Thus, it still has lower accuracies in the classes with decisive details and complex relationship details, such as commercial area, medium residential, and church. To address this issue, this paper proposes a Grad-CAM and CapsNet hybrid method for remote sensing image scene classification.

The remainder of this paper is organized as follows. Section 2 introduces the proposed Grad-CAM and CapsNet hybrid method. Section 3 presents the experimental data sets, implementation details, and experimental results. Section 4 discusses the relevant factors to the proposed method. The conclusion is given in Section 5.

2 Grad-CAM and CapsNet hybrid method

For the problem of extracting decisive features, the attention mechanism can help the model focus more on the salient part, which makes the model capture the more useful features while ignoring the relatively useless features. Additionally, CapsNet, which considers the relationship between different features compared to traditional CNN, helps to solve the problem of discovering relationship details. In this paper, we generate an attention map by using Grad-CAM and then use CapsNet rather than fully connected layers to combine the attention mechanism and CapsNet (Grad-CAM-CapsNet) to solve the two problems simultaneously.

2.1 Overall framework

As shown in Fig. 2, the proposed Grad-CAM-CapsNet model is composed of an attention block, feature fusion block, and CapsNet block. First, in the attention block, we transfer the pre-trained CNN model that has been fully trained for the ImageNet1000 data set and use the Grad-CAM algorithm to generate attention maps based on it. In the feature fusion block, the convolution operation is performed on the original input image and the attention map in parallel to obtain two feature maps. Then, the two feature maps are fused by a multiplication operation, and the fused result is input to the CapsNet block. In the CapsNet block, we use the CapsNet model to accept the fused features of the previous part and perform further feature extraction to obtain the classification results. The whole process of the Grad-CAM-CapsNet method can be described as follows.

1) Attention block: First, the Grad-CAM algorithm and the pre-trained model are used to generate the attention map. Note that the generated attention map needs to be upsampled to obtain the same size as the original input

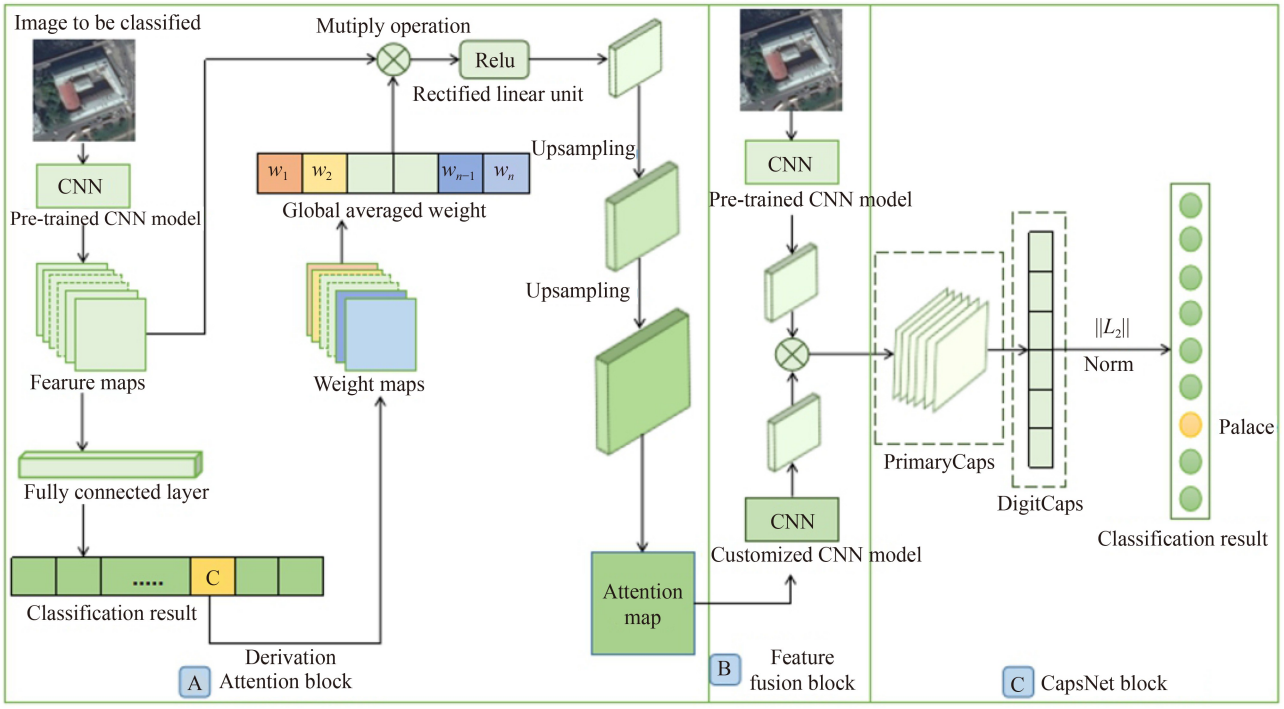

Fig. 2 The Grad-CAM-CapsNet architecture contains three parts: an attention block, a feature fusion block, and a CapsNet block.

image. In our method, the bilinear upsampling method is used.

2) Feature fusion block: Convolution operations on the input image and the attention map are performed in parallel. The former uses a pre-trained CNN model in which the weight of the last several layers is frozen as needed, and the latter uses a customized lightweight CNN model. The feature maps extracted by the two operations are multiplied to obtain the feature map masked by attention.

3) CapsNet block: Taking the feature map extracted in the previous step as input, the features through the PrimaryCaps and DigitCaps layers are extracted, and the vector mold length of the output capsule in each category is calculated. The category with the largest mold length is the classification result. Note that the traditional capsule model also has a convolutional layer before PrimaryCaps, and we removed this convolutional layer to avoid feature redundancy caused by excessive convolution.

Table 1 summarizes the complete Grad-CAM-CapsNet process. In the first step, we input image X into the pre-trained CNN to obtain the attention map that is associated with the network prediction by using Grad-CAM. In the second step, we use the pre-trained CNN and customized CNN to extract features of input image X and attention map Xam, respectively. Then, the fused feature maps XA are generated by multiplying X and Xam. In the last step, the fused feature map XA is input to CapsNet, and CapsNet returns the probability P of the input image X belonging to each class. The class with the highest probability is the classification result of this image.

2.2 Grad-CAM structure

Grad-CAM aims to use the gradient of any target concept to generate a rough positioning map in the last

Table 1 The whole process of the Grad-CAM-CapsNet

Procedure 1: Grad-CAM-CapsNet
Step1: Attention block
Input: Image X
Output: Attention image XA
• **Substep 1**: Calculate the weight coefficients α_i^c in the Grad-CAM according to Eq. (1)
• **Substep 2**: Calculate the attention map Xam according to Eq. (2)
• **Substep 3**: Xam is resized to match the size of input image X by upsampling
• **Substep 4**: Input X to the pre-trained CNN model to obtain feature map Fp
• **Substep 5**: Input Xam into the customized CNN model to obtain feature map Fc
• **Substep 6**: Fuse Fp and Fc by multiplying them to obtain attention masked image XA
• **Substep 7**: Return XA
Step2: CapsNet block
Input: Attention masked image XA
Output: The probability P of the input image
• **Substep 1**: The attention image XA is converted to capsule form by the PrimaryCaps layer
• **Substep 2**: After processing the DigitCaps layer, the category capsules can be obtained
• **Substep 3**: Obtain the probability P by computing the length of each category capsule according to Eq. (4)
• **Substep 4**: Return the probability P of the input image

convolutional layer, highlighting those areas with higher weight for correctly predicting the concept. In scene classification, a common problem is that there are always several images belonging to different classes with highly similar image content, which means that finding the salient parts that play a decisive role in correct classification is the key point for improving accuracy. By using Grad-CAM, we can effectively generate attention maps that highlight the salient parts. As opposed to CAM, it does not need to modify the original model structure or retrain the model. It can be used with various types of CNNs. The main premise of Grad-CAM is to use the global average of the gradient to calculate the weight of each feature map and then perform a weighted summation on all the feature maps to obtain the attention map.

The calculation process for generating the attention map using Grad-CAM is shown in Fig. 3. First, y^c is the probability that the image belongs to the cth class, which is output by the softmax classifier, and A_{ij}^k is the pixel value at position (i, j) in the kth feature map. For every feature map, we calculate the derivative of y^c with respect to A_{ij}^k and then add them up and divide by the total number of pixels Z to obtain the weight a_k^c, which represents the relative importance coefficient between the kth feature map and the cth class according to Eq. (1):

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \tag{1}$$

After obtaining the a_k^c of the cth class corresponding to all feature maps on the last layers in the pretrained model, every feature map A^k is multiplied by its corresponding weight a_k^c, and all the results are added up to obtain the attention map X_{am} by Eq. (2):

$$X_{\text{am}} = ReLU\left(\sum_k \alpha_k^c A^k\right). \tag{2}$$

A rectified linear unit (*ReLU*) is an activation function commonly used in CNN models to eliminate all negative values that can be calculated according to Eq. (3):

$$ReLU = \max(0, x). \tag{3}$$

Figure 4 shows the whole Grad-CAM implementation. Note that we selected the model with good classification accuracy on the ImageNet1000 data set (Abai and Rajmalwar, 2019) as the pre-trained CNN model. We resized the generated attention map (8 × 8) through upsampling to make it consistent with the shape of the input image (256 × 256). This is because our subsequent parallel feature extraction operations require the output feature to be the same size. By visualizing the attention map with the same size as the original image, we can specifically understand how the attention mechanism works.

Figure 5 shows the effect of superimposing the original input image and the attention map. It can be seen that the attention map considers church building, palace building, and financial building to be the more useful part of the classification result for the three types. Therefore, when the model performs feature extraction, the attention map helps it focus more on the information of each representative building while ignoring other information that is not useful.

2.3 CapsNet structure

The capsule model is a novel deep learning network model proposed in recent years to solve the shortcomings of traditional CNNs. While retaining the advantages of CNNs, it considers the relative position information between features that CNNs ignore so that images can be processed more efficiently. The effectiveness of CapsNet has made it widely used in different research areas, such as object detection, semantic segmentation, brain tumor classification, and image classification (Tian et al., 2019a, 2019b).

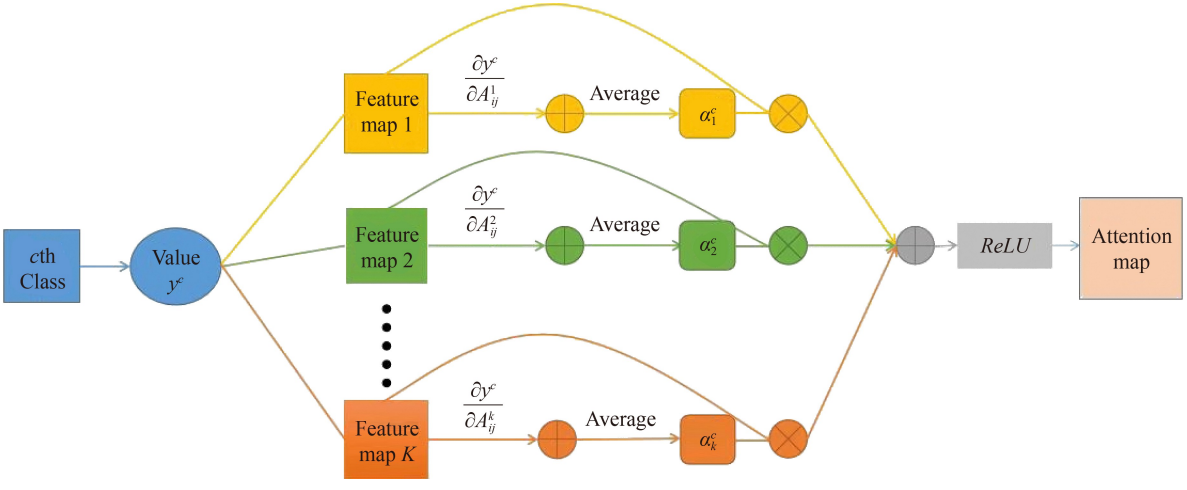

Fig. 3 Calculation process of generating the attention map using Grad-CAM.

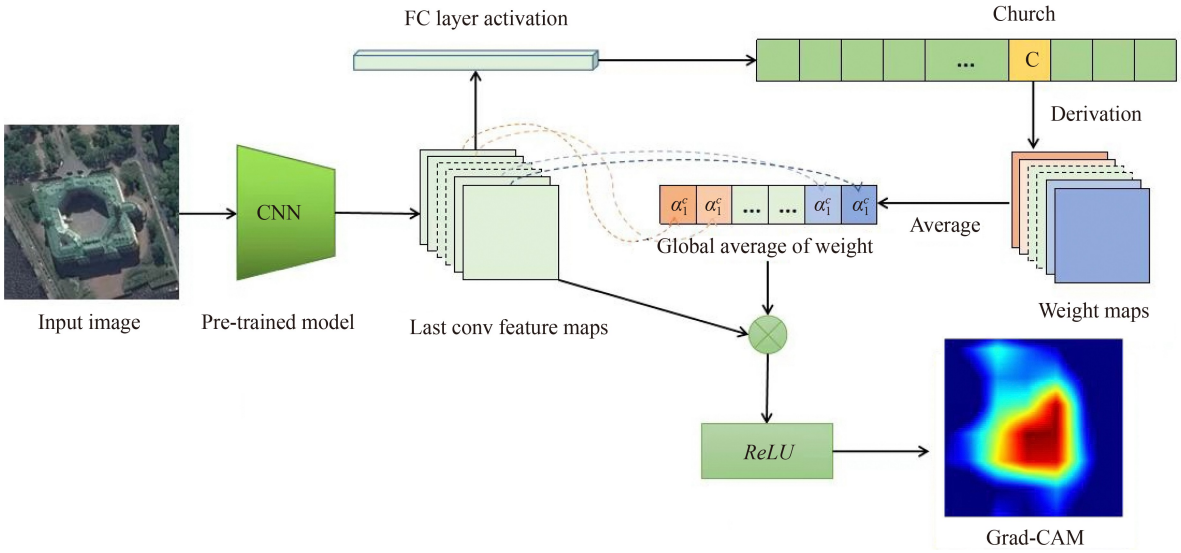

Fig. 4 Grad-CAM implementation.

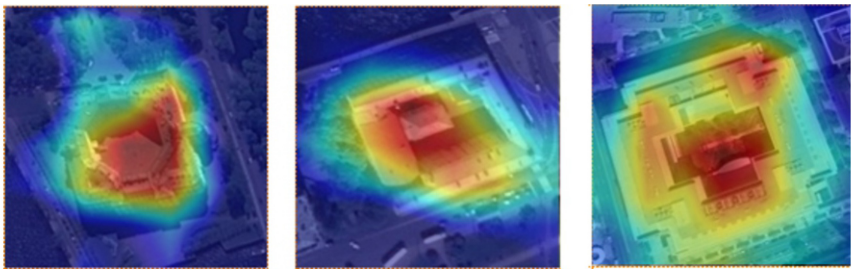

Fig. 5 Input images masked with attention maps.

As shown in Fig. 6, CapsNet contains three parts: the convolution layer, the PrimaryCaps layer, and the DigitCaps layer. First, the convolution layer performs feature extraction on the original image to obtain local convolution feature output ($H \times W \times L$; H, W, L denote the height, width, and channels of the image). Then, the PrimaryCaps layer is a convolution capsule layer that first converts the output of the previous layer into capsule form ([$H \times W$] $D1 \times L$-D Vector, where $D1$ denotes the dimension of the capsule in the PrimaryCaps layer). By a reshape function, the output of the PrimaryCaps layer ([$H \times W \times L$] $N1$-D Vector) can be computed. Next, the DigitalCaps layer is a fully connected capsule network layer whose output is category capsules ([S] $D2$-D Vector, S denotes the number of all predicted classes and $D2$ denotes the dimension of the capsule in the DigitCaps layer]). Finally, by computing the length of each category capsule using the $L2$ norm function according to Eq. (4), we can obtain the probability that the input belongs to every category:

$$L = \sqrt{(\alpha_1)^2 + (\alpha_2)^2 + \cdots + (\alpha_i)^2 \cdots + (\alpha_{D_2})^2}, \tag{4}$$

where α_i is the value in each dimension of one category capsule. The category corresponding to the max length is the final classification result.

The dynamic routing mechanism used to connect the lower-level capsules in the PrimaryCaps layer with all higher-level capsules in the DigitCaps layer is the core of CapsNet. By this connection, CapsNet routes the data from the lower layer to the higher layer while keeping the relative spatial information constant throughout the process. The output of higher-level capsules is predicted by multiplying the output of lower-level capsules with the transformation matrix, and the capsules at higher levels will be active when predictions are consistent. The details of dynamic routing can be illustrated as follows.

First, all prediction vectors of lower-level capsules $\hat{u}_{j|i}$ can be computed by $\hat{u}_{j|i} = W_{ij}u_i$, where W_{ij} is the weight matrix between the lower layer and higher layer refined by back propagation and u_i is the output of lower-level capsules. Then, in higher-level capsules, the input vector s_j can be obtained via the weighted summation of the prediction vectors from all capsules in the lower layers as follows:

$$s_j = \sum c_{ij}\hat{u}_{j|i}, \tag{5}$$

where c_{ij} represents the coupling coefficient determined by the iterative dynamic routing process. c_{ij} indicates the strength of the connection between the lower-level capsules and higher-level capsules and increases when the

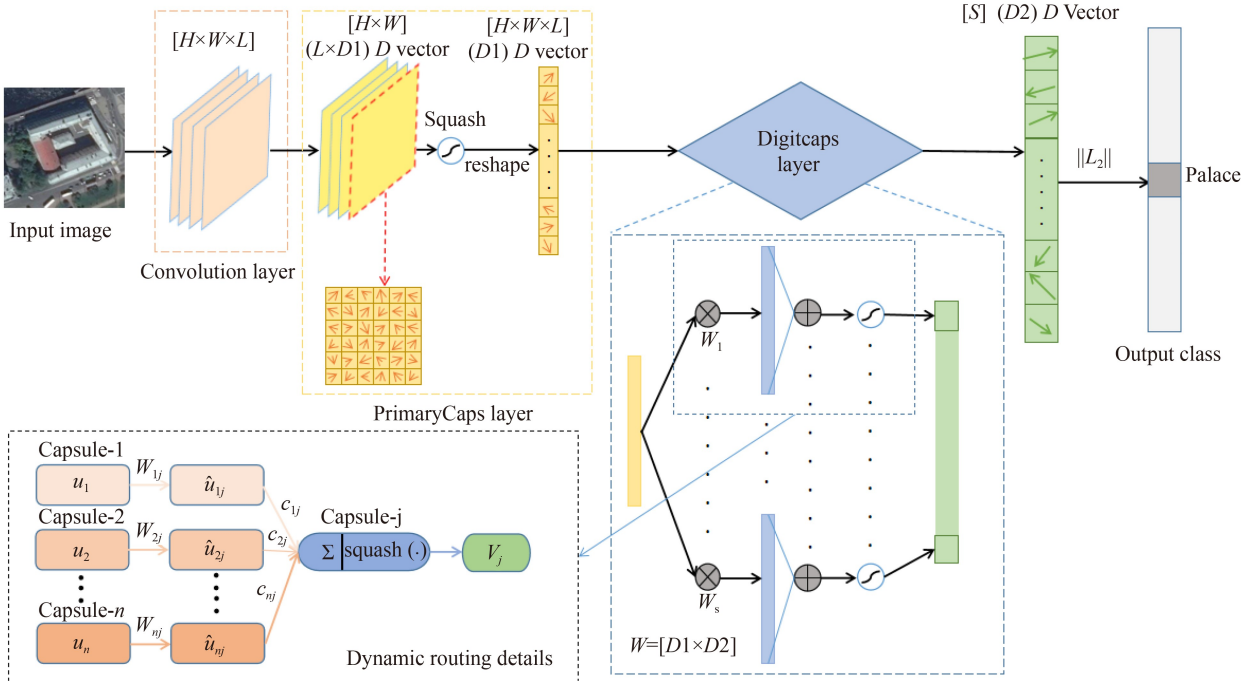

Fig. 6 The architecture of CapsNet.

prediction vector of lower-level capsules has a high agreement with the output of higher-level capsules. The sum of the coupling coefficients between the lower-level capsule i and all the capsules in the higher level is 1 and can be calculated by using the softmax function in Eq. (6):

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}, \tag{6}$$

where b_{ij} is the log prior probability of whether capsule i should be coupled with capsule j. b_{ij} is usually set to 0 as the initial value and then refined according to $b_{ij} \leftarrow b_{ij} + a_{ij}$, where a_{ij} denotes the consistency between the current output v_j of each capsule j in the higher level and the prediction vector $\hat{u}_{j|i}$ of capsule i in the lower level and can be computed by $a_{ij} = v_j \cdot \hat{u}_{j|i}$.

The existence probability of the entity in the current input is represented by the length of the output vector v_j of the capsule j in higher levels, so the capsule model uses a nonlinear squeeze function (squashing) to ensure that the short vectors are compressed to a length close to 0, and the long vectors are compressed to a length close to 1. The squashing function can be calculated according to Eq. (7):

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|}. \tag{7}$$

All the equations mentioned above make up one routing process, as shown in the dynamic routing details in Fig. 6. The routing algorithm consists of multiple iterations of the routing process, and the routing number represents the number of iterations.

The commonly used margin loss L_k is selected as the loss function in CapsNet, as shown in Eq. (8):

$$\begin{aligned} L_k =& T_k \max(0, m^+ - \|v_k\|)^2 \\ &+ \lambda(1-T_k)\max(0, \|v_k\| - m^-)^2, \end{aligned} \tag{8}$$

where k represents the class and T_k represents the classification indicator function (the value is 1 when class k exists and 0 when it does not exist). m^+ is the lower limit of correct classification. When $\|v_k\| \in [m^+, 1]$, the model considers that the current input image belongs to class k. Similarly, m^- is the upper limit of classification error; that is, when $\|v_k\| \in [0, m^-]$, the model considers that the current input image does not belong to class k ground objects. In this paper, we refer to the literature (Lei et al., 2021) and set m^+, m^-, λ as 0.9, 0.1, and 0.5. The total loss is the sum of the loss of all capsules in the last layer.

The lengths and directions of the output vectors of lower-level capsules represent the existence probability and properties of the specific entities corresponding to them, respectively. For example, when the network handles a ‘palace’ scene image, the lengths of the output encode the existence probability of different components in palace buildings, such as the main hall, eaves and square. The directions of the output encode properties of the palace, such as position, size, and orientation. Then, this information is routed to the higher-level capsules. The higher-level capsule (palace), whose output is an active vector that encodes the scene contexts that the ‘palace’ represents, only receives the information that has high similarity with its prediction. Therefore, capsules can make full use of the relative position information between different entities with this scene image.

3 Experiments and results

3.1 Experimental data sets

Three famous baseline data sets are used in our experiments: the UC Merced Land-Use data set (Yang and Newsam, 2010), the AID data set (Xia et al., 2017), and the NWPU-RESISC45 data set (Gong et al., 2017).

The UC Merced Land-Use data set contains 21 scene classes, where each class includes 100 scene images. The size of each scene image is 256 × 256 pixels, and the pixel resolution of each scene image is 0.3 m in the red-green-blue (RGB) color space. What makes it difficult to classify the data set is the severe overlap among some classes, such as sparse residential, dense residential, and medium residential.

The AID data set contains 30 scene classes, where each class includes 220 to 420 scene images. It is a large-scale data set for each scene image that is 600 × 600 pixels compared with the UC Merced Land-Use data set, which makes it more difficult to classify. The pixel resolution of each scene image varies approximately from 0.5 m to 8 m.

The NWPU-RESISC45 data set contains 45 classes, where each class includes 700 scene images. The size of each scene image is 256 × 256 pixels, and the pixel resolution of each scene image varies from approximately 0.3 m to 8 m in the RGB color space.

Note that the order of difficulty for classifying these data sets is NWPU-RESISC45, AID, and UC Merced Land-Use because the NWPU-RESISC45 data set contains the most diverse image types. Thus, the experimental results on these data sets can test the performances of the models hierarchically.

3.2 Implementation details

Implementation details include the setting details about the hyperparameters, hardware, and two core evaluation norms: overall accuracy and confusion matrix.

3.2.1 Settings

In the Grad-CAM stage, the pre-trained DenseNet121 (Zhou et al., 2016) is used to generate the attention map, and the weights in the first 312 layers are frozen. In the CapsNet stage, the hyperparameters are adjusted by using the Adam optimizer (Kingma and Ba, 2014), and the batch size is set as 50. The learning rate is set as 0.001, updating by multiplying 0.9 by the power of epochs, which is set as 40. The margin loss in Eq. (8) is used for the loss function, and we set m^+, m^-, and λ as 0.9, 0.1, and 0.5, respectively.

To demonstrate that our method is fairly effective compared with other state-of-the-art methods, we use the same split ratio of these compared methods according to the data set distribution in (Yang and Newsam, 2010; Gong et al., 2017; Xia et al., 2017). For the UC Merced Land-Use data set, 80% and 20% split ratios are set for training and testing. For the AID data set, 50% and 20% split rations are set for training, and the rest are used for testing. In the NWPU-RESISC45 data set, the experiments use 20% and 10% training ratios. The image data argument that is used for expanding the data and the rotation, width shift, and height shift range are set as 30, 0.1, and 0.1, respectively.

In terms of hardware, all the implementations are performed on a Windows 10 operating system with a 2.4 GHz 4-core i5-9300H CPU and 16 GB memory. An NVIDIA GTX1660Ti GPU with CUDA 10.2 is used to accelerate the computing.

3.2.2 Evaluation norms

The overall accuracy (OA) and confusion matrix are used to measure the designed experiments. OA is defined as the number of correctly classified images divided by the total number of labeled images, which belongs to the range of 0 to 1. The OA can measure the overall performance of the models. The confusion matrix is an informative table that can help readers analyze the errors and conclusions between each class, which directly shows the classification performance of each class in the table. This means that the confusion matrix of the experiment shows the performance on separate classes.

In the calculation process, reliable OAs are achieved by ten repeated experiments according to the different training ratios for the three baseline data sets. The means and standard deviations of the OAs are also reported simultaneously. The confusion matrices are generated by using the classification results according to the ratios of 80%, 20%, and 10% for the UC Merced Land-Use data set, AID data set, and NWPU-RESISC45 data set.

3.3 Experimental results

To illustrate the validity of our proposed model, recent classification models with high image scene classification accuracy rates were selected. The models and their core features are listed in Table 2. The experiments follow from the easiest data set to the hardest data set: UC Merced Land-Use data set, AID data set, and NWPU-RESISC45 data set.

3.3.1 UC Merced Land-Use Data set

The results of the UC merced land-use data set are listed in table 3. Although the experimental models can be divided into two types (with/without attention or CapsNet), there are slight differences in the results between different models. Most of the models perform well on the UC Merced Land-Use data set, with an accuracy of over 95% and some results of over 98% accuracy rate. In particular, the accuracy of our proposed

Table 2 List of experimental comparative models

Model ID	Model	Attention	CapsNet	Core features
1	CaffeNet (Xia et al., 2017)	No	No	Dropout structure
2	GoogLeNet (Szegedy et al., 2017)	No	No	Inception structure
3	VGG16 (Liu et al., 2017)	No	No	Multiple small convolution kernels
4	CNN-ELM (Weng et al., 2017)	No	No	ELM structure
5	Fine-tuned GoogLeNet (Weng et al., 2017)	No	No	Fine-tune
6	Fine-tuned VGG19 (Castelluccio et al., 2015)	No	No	Fine-tune
7	Deep CNN Transfer (Marmanis et al., 2016)	No	No	Different scale features
8	Triple networks (Liu and Huang, 2018)	No	No	Label training replacement
9	Attention-based residual network (Fan et al., 2019)	Yes	No	CNN with attention
10	Two-stream fusion (Yu and Liu, 2018a)	Yes	No	Separate spatial and temporal features
11	VGG16-CapsNet (Zhang et al., 2019)	No	Yes	CapsNet
12	D-CapsNet (Raza et al., 2020)	Yes	Yes	Spatial attention & CapsNets
13	Grad-CA –CapsNet (our proposed model)	Yes	Yes	Pretrained attention & CapsNets

Table 3 The performances of different models on the UC Merced Land-Use data set

Model ID	Model	Types	Accuracy and Standard Deviation
1	CaffeNet	Without attention or CapsNet	95.02±0.81
2	GoogLeNet		94.31±0.89
3	VGG16		95.21±1.20
4	CNN-ELM		95.62
7	Deep CNN transfer		98.49
9	D-CNN with VGGNet16		98.93±0.10
5	Fine-tuned GoogLeNet		97.10
6	Fine-tuned VGG19		98.1
10	Two-stream fusion	With attention or CapsNet	98.02±1.03
9	attention-based residual network		98.81±0.30
12	Ours		99.05±0.15

model is over 99%, and all the models with attention mechanisms or CapsNet are over 98%. Even parts of the models without an attention mechanism or CapsNet had accuracy rates over 98%. Although the accuracy is slightly higher than that of the other models, it cannot highlight the advantages of our proposed model. In other words, the UC Merced Land-Use data set lacked the difficulty needed to represent the advantages of our proposed model.

The confusion matrix of our proposed model also represents this inference (Fig. 7). In the confusion matrix, both the horizontal and vertical coordinates represent the surface feature category; the value on the diagonal represents the correct classification ratio of the surface feature category on the vertical axis; other numbers represent the proportion of the incorrect classifications of this type of surface feature; and the horizontal coordinate of the location of the number represents the incorrect classification category. For example, in Fig. 7, the classification accuracy of the category “agricultural” is 95 percent, with 5 percent misclassified as “forest.” Other confusion matrices have the same meaning. All accuracies of the classes of the UC Merced Land-Use data set with our proposed model are nearly 1.00. Thus, further comparative experiments are presented in the following baseline data sets.

3.3.2 AID data set

The AID data set contains 30 scene classes, which is more than the UC Merced Land-Use data set with 21 classes. Extra scene classes mainly belong to complex building scenes with mixed decisive details and complex relationship details, such as commercial areas. Thus, outstanding classic models without attention or CapsNet structures, models with attention, models with CapsNet, and models with both attention and CapsNet are compared.

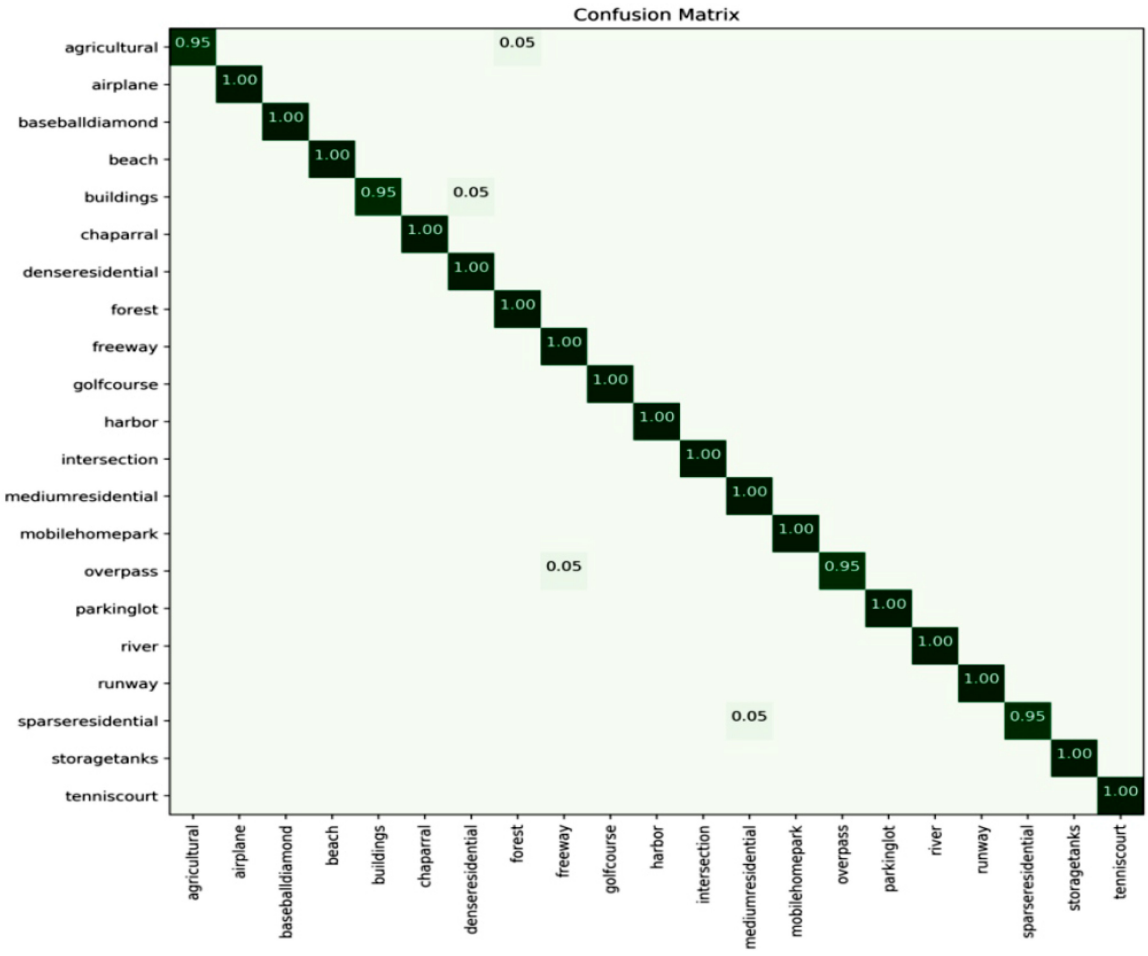

Fig. 7 Confusion matrix of the proposed Grad-CAM-CapsNet model on the UC Merced Land-Use data set with a training ratio of 80%.

Table 4 shows the results of different models on the AID data set. Obviously, three groups are distinguished both on 50% training ratios and 20% training ratios. The first group includes all three selected models without attention or the CapsNet structure, which has accuracies of 86%−89% and 83%−86% in the 50% and 20% training ratios, respectively. The second group includes the two-stream fusion model and VGG-16-CapsNet model. Both of these models achieve approximately 94% and 92% accuracy in 50% and 20% training ratios, respectively. This demonstrates that attention and the CapsNet structure are useful for classifying image scenes. The third group includes the models with both attention and CapsNet structures. The accuracy intervals of the 50% and 20% training ratios are approximately 96% and 92%, respectively. There is a nearly 2% increase in performance compared with the single structure. This means that the combination of the attention mechanism and CapsNet structure can effectively improve the performance of image scene classification.

Table 4 The performances of different models on the AID data set

Model ID	Model	Types	Accuracy (50% training ratio)	Accuracy (20% training ratio)
2	GoogLeNet	Without attention or CapsNet	86.39±0.55	83.44±0.40
1	CaffeNet		89.53±0.31	86.86±0.47
3	VGG16		89.64±0.36	86.59±0.29
12	Two-stream fusion	Attention	94.58±0.25	92.32±0.41
13	VGG-16-CapsNet	CapsNet	94.74±0.17	91.63±0.19
14	D-CapsNet	Attention & CapsNet	96.15±0.14	92.73±0.15
15	Ours		96.43±0.12	93.68±0.14

Note that the interpolations of accuracies between our proposed method and the D-CapsNet model in 50% training ratios and 20% training ratios are different. In 50% training ratios, the proposed method is 0.28% (96.43%−96.15%) higher than the D-CapsNet model. In 20% training ratios, the proposed method is 0.95% (93.68%−92.73%) higher than the D-CapsNet model. This shows that our method can obtain higher accuracy with a lower training ratio than the previous D-CapsNet model indicating that our proposed method has a better convergence performance. This is because the attention map of our model generated by the pre-trained model has a higher efficiency than the attention map of the D-CapsNet model generated by convolutional layers.

For specific classes, the results are shown in the confusion matrix generated by the AID data set with a 20% training ratio (Fig. 8). In addition, the meanings of the horizontal and vertical coordinates in Fig. 8 are the same as those in Fig. 7. It shows that most of the classes have good performance with their classification accuracy of over 90%, especially in the dense residential, sparse residential, and medium residential classes. These three classes have high similarity meaning our proposed method can effectively identify these small interclass dissimilarities. Although the accuracies of school and resort are lower than 80% (72.8% and 77.0%), they are much higher than the scores of the D-CapsNet model (nearly 50%), which demonstrates that the performance of our proposed model surpasses the previous models both in the overall accuracy and the accuracy of specific classes.

3.3.3 NWPU-RESISC45 data set

The NWPU-RESISC45 data set contains 45 classes which are twice as large as the classes of the UC Merced

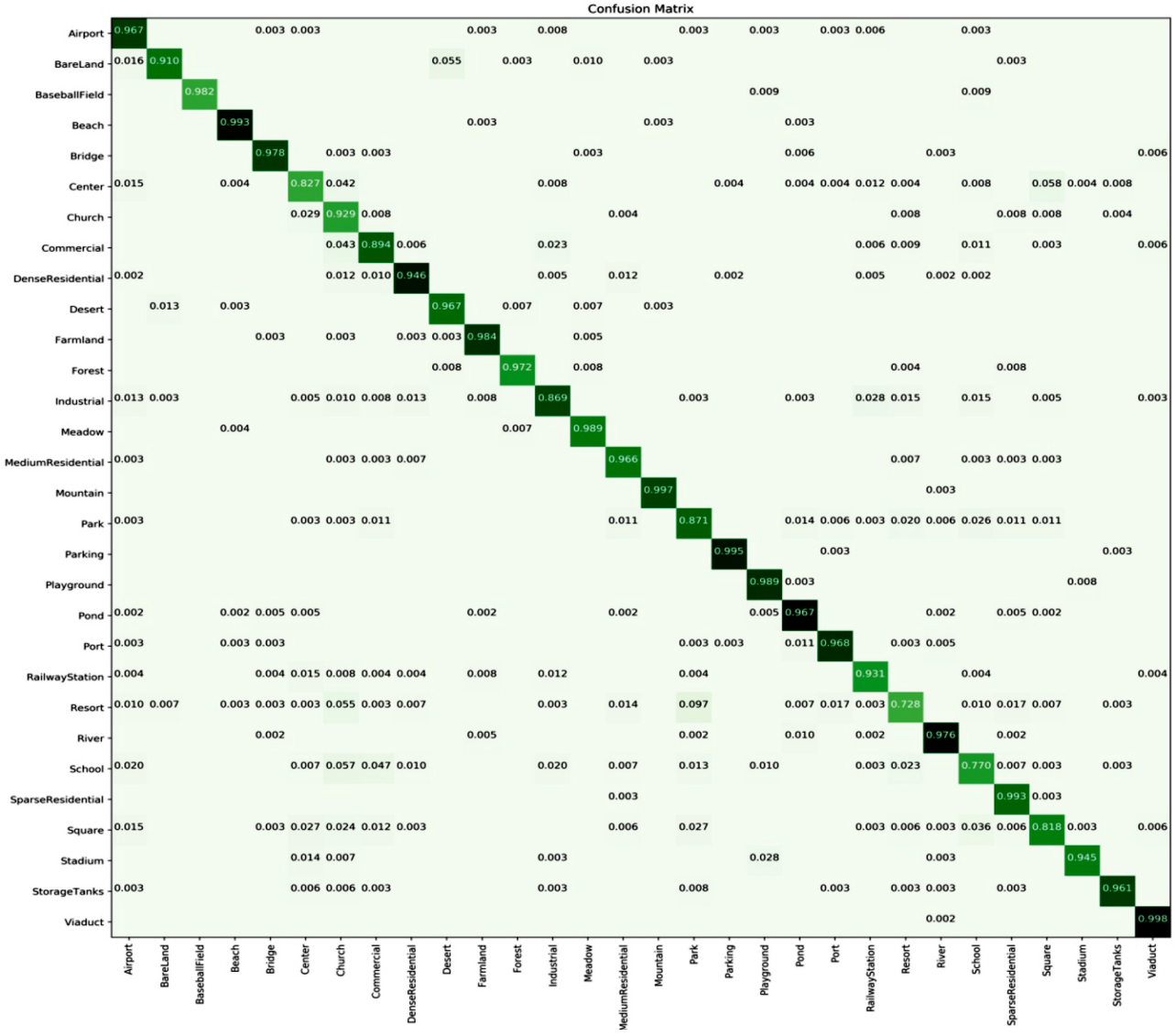

Fig. 8 Confusion matrix of our proposed model on the AID data set with a 20% training ratio.

Land-Use data set. Thus, it is a challenging task to test the models. In the experiment of the NWPU-RESISC45 data set, four groups of models (without attention or CapsNet group, attention group, CapsNet group, and attention & CapsNet group) are selected and compared to verify the effectiveness of our proposed model.

Table 5 lists the results of different models on the NWPU-RESISC45 data set with different training ratios. In general, our proposed model obtains the highest accuracies of 92.67±0.08 and 89.34±0.20 on both 20% and 10% training ratios, respectively. This demonstrates that our model achieves state-of-the-art performance in image scene classification.

Specifically, the group "without attention or CapsNet" has a gap to the top performance level except for the "triple networks" model. The results not only verify that the hybrid structure is effective but also prove that the pre-trained model applied in the "triple networks" model can improve the performance among the complex images. This is also the reason why our proposed model achieves higher performance than the D-CapsNet model in the same group. Moreover, we found that the results of the attention group and the CapsNet group are not ideal. This demonstrates that the model with a single attention mechanism or CapsNet structure cannot handle complex images very well. In contrast, our proposed model with the hybrid structure and the pretrained model can achieve the highest accuracy with the most difficule data set. To analyze the specific details in each class, the confusion matrix is given in Fig. 9, and the meaning of the confusion matrix is the same as above.

According to the challenging experiment on the NWPU-RESISC45 data set with a 20% training ratio, our proposed model can still obtain high accuracies in all classes. The accuracies of the 89% (40/45) classes exceed 0.88. Note that the classes (church, palace, and commercial area) with decisive details and complex relationship details have a comprehensive ascension (Table 6).

According to this comparison, the "attention & CapsNet" structure can combine the strengths of the attention mechanism and CapsNet structure. Both the "church" and "commercial area" classes reach the highest performance. Note that the "palace" class obtains a 75% accuracy. After further investigation, the reason was found to be that the "palace" and "church" images have highly similar architectural styles and diverse spatial distributions. The surrounding environmental features exhibit different mechanisms to classify these complex scenes, which is a potential task for future research. Overall, the Grad-CAM and CapsNet hybrid method is a promising and powerful model for classifying the image scene in different tasks. The following discussion analyzes the effectiveness of different parts of the model.

4 Discussion

Although the experimental results demonstrate that the Grad-CAM and CapsNet hybrid method achieves outstanding performances on three public benchmark data sets, it is still difficult to explain the improved effectiveness of each part of the structure in the proposed method. Thus, the discussion section sets ablation experiments to analyze the improved effectiveness of three key factors: the Grad-CAM mechanism, CapsNet, and the pretrained model.

4.1 Effectiveness of the Grad-CAM mechanism

The Grad-CAM mechanism generates attention maps to make our model concentrate on the salient part while ignoring the useless parts. To verify the effectiveness of the Grad-CAM mechanism, the models are compared in different data sets. The results show that the Grad-CAM mechanism has stable increments of +1.03%, +0.44%, and +0.33% in different data sets (Table 7).

Furthermore, the CapsNet in the above experiment is replaced to exclude the effect of CapsNet. According to experimental universality, an FC layer classifier is used to replace CapsNet. Table 8 compares the accuracy rate without the Grad-CAM mechanism and using the Grad-CAM mechanism. All the increments in Table 8 support that the Grad-CAM mechanism in our model has a

Table 5 The performances of different models on the NWPU-RESISC45 data set

Model ID	Model	Types	Accuracy (20% training ratio)	Accuracy (10% training ratio)
2	GoogLeNet	Without attention or CapsNet	86.39±0.55	83.44±0.40
3	VGG-16		89.53±0.31	86.86±0.47
6	Fine-tuned VGG-16		90.36±0.18	87.15±0.45
8	Triple networks		92.33±0.20	87.15±0.45
10	Two-stream fusion	Attention	83.16±0.18	80.22±0.22
11	VGG-16-CapsNet	CapsNet	89.18±0.14	85.08±0.13
12	D-CapsNet	Attention & CapsNet	92.46±0.14	88.18±0.19
13	Ours		92.67±0.08	89.34±0.20

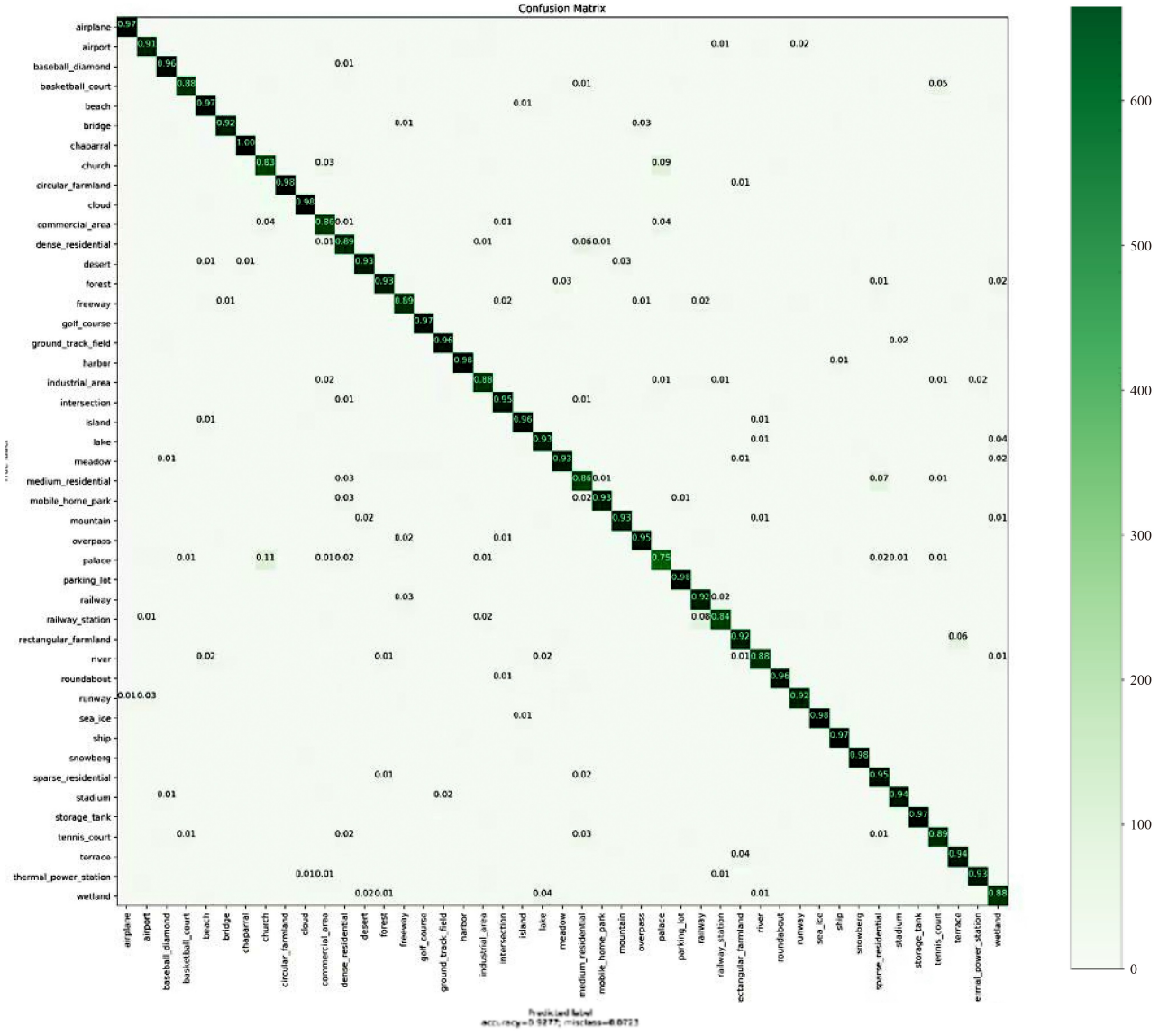

Fig. 9 Confusion matrix of our proposed model on the NWPU-RESISC45 data set with a 20% training ratio.

Table 6 The performances of different models with specific classes on the NWPU-RESISC45 Data set

ID	Model type	Church	Palace	Commercial area	Total accuracy
1	Without attention or CapsNet	64%	61%	76%	87.65%
2	Attention	79%	68%	86%	83.16%
3	CapsNet	74%	85%	80%	89.18%
4	Attention & CapsNet	83%	75%	86%	92.67%

positive effect. Note that the pre-trained model in the experiment uses a unified data set (DenseNet).

This proves that after we add the Grad-CAM attention mechanism to the network, the weight location map generated by the model can highlight the parts that have a decisive impact on the correct classification, help the model to learn the significant features of different ground objects in the remote sensing image classification task, and enhance the classification ability of the model for images with high similarity. Therefore, when feature extraction is carried out in the model, attention should be given to the representative information of different types of ground features, while other useless information should be ignored. This is exactly what the CNN using pooling layers lacks.

4.2 Effectiveness of CapsNet

To verify the effectiveness of CapsNet, two groups of comparative experiments are set. The first comparative experiment ("Grad-CAM-CapsNet" vs. "Grad-CAM-FC layer") is designed to test whether the CapsNet affects performance. The left results in Table 9 show that CapsNet can provide stable increments of +1.43%, +1.07%, and +1.61% in different data sets. This means that CapsNet is effective for our proposed method. To

Table 7 The performances of different models with specific classes on the NWPU-RESISC45 data set

Data set	Accuracy of pretrained model (using Grad-CAM)-CapsNet/%	Accuracy of pretrained model (without Grad-CAM)-CapsNet/%	Increment
UC Merced	99.05	98.02	+ 1.03%
AID	96.43	95.99	+ 0.44%
NWPU-RESISC45	92.67	92.34	+ 0.33%

Table 8 The effectiveness of the Grad-CAM mechanism with an FC layer in different data sets

Data set	Accuracy of pretrained model (using Grad-CAM)- FC layer/%	Accuracy of pretrained model (without Grad-CAM)- FC layer/%	Increment
UC Merced	97.62	95.95	+ 1.67%
AID	95.36	94.82	+ 0.54%
NWPU-RESISC45	91.06	88.21	+ 2.85%

eliminate the effect of Grad-CAM on the previous experiment, the second comparative experiment is set to determine whether CapsNet is effective in the structure without the Grad-CAM mechanism. The right results in Table 9 list the positive scores (+2.07%, +1.17%, and +4.13%) in different data sets.

According to the above results, even though the attention mechanism causes the model to focus more on the core features when these parts are still highly similar, a single attention mechanism is not enough to solve the problem. The capsule network uses output vectors to represent different attributes, including spatial information and accounts for the spatial location information ignored by Grad-CAM, which plays an important role in distinguishing complex and similar image scenes. The capsule model can capture this information very well, which gives the capsule model a strong ability to retain, learn and discriminate spatial information to process images more effectively.

These comparative experiments not only verify the effect of the CapsNet structure on performance in different data sets and different structures but also support the effect of the Grad-CAM mechanism by comparing the results of "Grad-CAM-CapsNet" and "without Grad-CAM-CapsNet."

4.3 Effectiveness of the pre-trained model

To demonstrate the effectiveness of the pre-trained model, further comparative experiments are designed between the methods with and without pre-trained DenseNet121. Specifically, the experiment uses a CNN composed of four convolution layers and one max-pooling layer. The size of the output of the CNN is the same as the output of the pre-trained DenseNet121 ($8 \times 8 \times 1024$). In addition, the experiment compares the pretrained DenseNet121 with and without weight freezing. In the case without weight freezing, the net freezes the weights of the first 312 layers (the first three dense blocks) and sets the weights of the remaining layers (the last dense block) as trainable. Note that all the methods use an attention mechanism.

The OAs of different models are shown in Table 10. DenseNet-CapsNet without weight freezing provides a

Table 9 The effectiveness of CapsNet in different data sets

Data set	Accuracy/%		Increment	Accuracy/%		Increment
	Grad-CAM-CapsNet	Grad-CAM-FC layer		without Grad-CAM-CapsNet	without Grad-CAM-FC layer	
UC Merced	99.05	97.62	+ 1.43%	98.02	95.95	+ 2.07%
AID	96.43	95.36	+ 1.07%	95.99	94.82	+ 1.17%
NWPU-RESISC45	92.67	91.06	+ 1.61%	92.34	88.21	+ 4.13%

Table 10 The effectiveness of the pre-trained model in different data sets

Data set	CNN-CapsNet	DenseNet-CapsNet with weight freeze	DenseNet-CapsNet without weight freeze
UC Merced	61.02	98.12	99.05
AID	64.38	95.21	96.43
NWPU-RESISC45	45.67	91.04	92.67

higher OA than DenseNet-CapsNet with weight freezing. Both methods perform much better than the CNN-CapsNet. This means that the pretrained model improves the performance of the structure. This is because the CNN model cannot directly obtain the various features without the pre-trained model.

Furthermore, it is noted that all the performances of "without weight freeze" are higher than the performances of "with weight freeze." This is because the original task of the pre-trained model may be different from the classified scene images. Thus, adapted weights of the pre-trained model can be more suitable for model usage.

5 Conclusions

This paper proposes a Grad-CAM and capsule network hybrid method for remote sensing image scene classification to accurately identify objects with both decisive details and complex relationship details. It uses the Grad-CAM mechanism to enhance the ability of the method to identify the decisive details and uses a capsule network to enhance the ability of the method to identify the complex relationship details. After detailed experiments and discussions, three core conclusions are found.

1) The hybrid structure of the Grad-CAM and the capsule network can effectively improve the performance of the classifier. The results show that the proposed Grad-CAM and capsule network hybrid method achieves a state-of-the-art performance of 99.05 ± 0.15 under an 80% training rate.

2) The hybrid structure of the Grad-CAM and the capsule network can improve the ability of the classifier on both decisive details and complex relationship details. According to the confusion matrices of different data sets, "dense residential," "medium residential," and other types with decisive details and complex relationship details achieve nearly 0.95 accuracy, which is much higher than the previous studies of approximately 0.80.

3) There are three factors affecting the performance of the hybrid structure: the Grad-CAM mechanism, capsule network, and pre-trained model. This means that the upgraded attention mechanism, capsule network, and pre-trained model may improve the performance step by step. Thus, different pre-trained models, capsule networks, and attention maps generated by other mechanisms will be analyzed in the future to reveal the complementary mechanism of the hybrid structure. If it can be explained, the performance will be controlled by users on different applications, and it must be improved further.

Acknowledgments This research was funded by the open fund of the Key Laboratory of Jianghuai Arable Land Resources Protection and Eco-restoration (Ministry of Natural Resources) (No. 2022-ARPE-KF04), and the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation (Ministry of Natural Resources) (No. KF-2020-05-084).

Competing interests The authors declare that they have no competing interests.

References

Abai Z, Rajmalwar N (2019). DenseNet models for tiny imagenet classification. arXiv preprint arXiv: 1904.10429

Ahmed A, Jalal A, Kim K (2020). A novel statistical method for scene classification based on multi-object categorization and logistic regression. Sensors (Basel), 20(14): 3871

Bai S (2016). Growing random forest on deep convolutional neural networks for scene categorization. Expert Systems with Applications, 71: 279–287

Castelluccio M, Poggi G, Sansone C (2015). Land use classification in remote sensing images by convolutional neural networks. arXiv preprint arXiv: 1508.00092

Chaib S, Liu H, Gu Y, Yao H (2017). Deep feature fusion for VHR remote sensing scene classification. IEEE Trans Geosci Remote Sens, 55(8): 4775–4784

Chen J, Wang C, Ma Z, Chen J, He D, Ackland S (2018). Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters. Remote Sens (Basel), 10(2): 290

Cheng G, Han J, Lu X (2017a). Remote sensing image scene classification: benchmark and state of the art. Proc IEEE, 105(10): 1865–1883

Cheng G, Li Z, Yao X, Guo L, Wei Z (2017b). Remote sensing image scene classification using bag of convolutional features. IEEE Geosci Remote Sens Lett, 14(10): 1735–1739

Cheng G, Yang C, Yao X, Guo L, Han J (2018). When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans Geosci Remote Sens, 56(5): 2811–2821

Cheng G, Zhou P, Han J (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans Geosci Remote Sens, 54(12): 7405–7415

Fan R, Wang L, Feng R (2019). Attention based residual network for high-resolution remote sensing imagery scene classification. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 1346–1349

Gan J, Li Q, Zhang Z, Wang J (2016). Two-level feature representation for aerial scene classification. IEEE Geosci Remote Sens Lett, 13(11): 1626–1630

Gong C, Han J, Lu X (2017). Remote sensing image scene classification: benchmark and state of the art. In: Proceedings of the IEEE, 105(10): 1865–1883

Hou Q, Zhou D, Feng J (2021). Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13713–13722

Kingma D P, Ba J (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv: 1412.6980

Knorn J, Rabe A, Radeloff V C, Kuemmerle T, Kozak J, Hostert P (2009). Land cover mapping of large areas using chain

classification of neighboring Landsat satellite images. Remote Sens Environ, 113(5): 957–964

Lei R, Zhang C, Liu W, Zhang L, Zhang X, Yang Y, Huang J, Li Z, Zhou Z (2021). Hyperspectral remote sensing image classification using deep convolutional capsule network. IEEE J Sel Top Appl Earth Obs Remote Sens, 14: 8297–8315

Lei R, Zhang C, Zhang X, Huang J, Li Z, Liu W, Cui H (2022). Multiscale feature aggregation capsule neural network for hyperspectral remote sensing image classification. Remote Sens (Basel), 14(7): 1652

Li J, Lin D, Wang Y, Xu G, Zhang Y, Ding C, Zhou Y (2020). Deep discriminative representation learning with attention map for scene classification. Remote Sens (Basel), 12(9): 1366

Liu Y, Cheng M M, Hu X (2017). Richer convolutional features for edge detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5872–5881

Liu Y, Huang C (2018). Scene classification via triplet networks. IEEE J Sel Top Appl Earth Obs Remote Sens, 11(1): 220–237

Marmanis D, Datcu M, Esch T, Stilla U (2016). Deep learning earth observation classification using imagenet pretrained networks. IEEE Geosci Remote Sens Lett, 13(1): 105–109

Mei X, Pan E, Ma Y, Dai X, Huang J, Fan F, Du Q, Zheng H, Ma J (2019). Spectral-spatial attention networks for hyperspectral image classification. Remote Sens (Basel), 11(8): 963

Pan Z, Xu J, Guo Y, Hu Y, Wang G (2020). Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net. Remote Sens (Basel), 12(10): 1574

Pires de Lima R, Marfurt K (2019). Convolutional neural network for remote-sensing scene classification: transfer learning analysis. Remote Sens (Basel), 12(1): 86

Raiyani K, Gonçalves T, Rato L, Salgueiro P, Marques da Silva J R (2021). Sentinel-2 image scene classification: a comparison between Sen2Cor and a machine learning approach. Remote Sens (Basel), 13(2): 300

Raza A, Huo H, Sirajuddin S, Fang T (2020). Diverse capsules network combining multiconvolutional layers for remote sensing image scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens, 13: 5297–5313

Sabour S, Frosst N, Hinton G E (2017). Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 3859–3869

Sheng G, Yang W, Xu T, Sun H (2012). High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int J Remote Sens, 33(8): 2395–2412

Sun X, Zhu Q, Qin Q (2021). A multi-level convolution pyramid semantic fusion framework for high-resolution remote sensing image scene classification and annotation. IEEE Access, 9: 18195–18208

Szegedy C, Ioffe S, Vanhoucke V, Alemi A A (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 4278–4284

Tian T, Liu X, Wang L (2019a). Remote sensing scene classification based on res-capsnet. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019: 525–528

Tian X, An J, Mu G (2019b). Power System Transient Stability Assessment Method Based on CapsNet. In: 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia). IEEE, 2019: 1159–1164

Tong W, Chen W, Han W, Li X, Wang L (2020). Channel-attention-based DenseNet network for remote sensing image scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens, 13: 4121–4132

Vo T, Tran D, Ma W (2015). Tensor decomposition and application in image classification with histogram of oriented gradients. Neurocomputing, 165: 38–45

Wang Y, Zhang J, Kan M (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12275–12284

Weng Q, Mao Z, Lin J, Guo W (2017). Land-use classification via extreme learning classifier based on deep convolutional features. IEEE Geosci Remote Sens Lett, 14(5): 704–708

Xia G S, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L, Lu X (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans Geosci Remote Sens, 55(7): 3965–3981

Yang Y, Newsam S (2010). Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010: 270–279

Yu C, Gao C, Wang J, Yu G, Shen C, Sang N (2021). Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis, 129(11): 3051–3068

Yu Y, Liu F (2018a). A two-stream deep fusion framework for high-resolution aerial scene classification. Comput Intell Neurosci, 2018: 8639367

Yu Y, Liu F (2018b). Dense connectivity based two-stream deep feature fusion framework for aerial scene classification. Remote Sens (Basel), 10(7): 1158

Zhang W, Tang P, Zhao L (2019). Remote sensing image scene classification using CNN-CapsNet. Remote Sens (Basel), 11(5): 494

Zhang X, Wang G, Zhao S G (2022). CapsNet-COVID19: Lung CT image classification method based on CapsNet model. Math Biosci Eng, 19(5): 5055–5074

Zhao B, Zhong Y, Zhang L, Huang B (2016). The Fisher kernel coding framework for high spatial resolution scene classification. Remote Sens (Basel), 8(2): 157

Zhao D, Chen Y, Lv L (2017). Deep reinforcement learning with visual attention for vehicle classification. IEEE Trans Cogn Dev Syst, 9(4): 356–367

Zhao X, Zhang J, Tian J, Zhuo L, Zhang J (2020). Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. Remote Sens (Basel), 12(11): 1887

Zhou B, Khosla A, Lapedriza A (2016). Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921–2929