A scene recognition algorithm based on deep residual network

Mao Jiafa, Wang Weifeng, Hu Yahong and Sheng Weiguo

Abstract

Scene recognition is quite important in the field of robotics and computer vision. Aiming at providing high performance and universality of feature extraction, a convolutional neural network-based scene recognition model entitled Scene-RecNet is proposed. To reduce parameter space and improve the feature quality, deep residual network is introduced as the feature extractor. A feature adjustment layer composed of a convolutional layer and a fully connected layer is added after the feature extractor to further synthesize and compress the extracted features. Migration learning-based 'pre-training and fine-tuning' mode is used to train Scene-RecNet. The feature extractor is pre-trained by ImageNet, and the overall network performance is fine-tuned on specific data sets. Experiments show that comparing with other algorithms, the features obtained by Scene-RecNet have high generality and robustness, and Scene-RecNet can provide better scene classification accuracy rate.

1. Introduction

Scene image recognition belongs to visual classification (Li, Shi, Dong, & Tao, 2015; Ponce, Hebert, Schmid, & Zisserman, 2006) and it means to determine the scene category of an image by analysing its contents. J. Gray, the Turing Award Winner, raised some difficult issues in the future, where large-scale classification and retrieval of images is included. Scene image recognition provides valid context information in the boot target recognition and it has become a hot topic in many fields of research, such as remote sensing (Yu, Wu, Luo, & Ren, 2017), aerial scene classification (Qayyum et al., 2017) and in the fields of computer vision and robotics (Espinace, Kollar, Roy, & Soto, 2013). Zhou, He, Qing, Wan, and Zheng (2019) proposed a method to recognize under construction building areas in complex scenarios. Li (2018) achieved urban structure types recognition based on remote sensed image scene. In robotics, in order to determine the robot's position and orientation in its environment in real time (Durrant-Whyte & Bailey, 2006; Se, Lowe, & Little, 2001; Wang et al., 2014; Wu, Wang, Xing, Gong, & Liu, 2011), it is necessary to create a system for both map construction and positioning (Choset & Nagatani, 2001) and scene image recognition is the key issue of these systems. In the area of computer vision, it’s more and more important to label images with semantic tags based on image contents so as to analyse and manage complex image data (Mao et al., 2018). Scene image recognition is one effective way to solve this kind of problems (Parizi, Oberlin, & Felzenszwalb, 2012; Zhou, Hu, & Zhou, 2013).

Scenes can be roughly divided into four categories, i.e. natural scenes, urban scenes, indoor scenes, and event scenes, respectively (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). The constituent elements of different scenes are quite different, so distinct effects may be obtained when the same recognition method is used to handle different scene datasets, especially for outdoor and indoor scenes. Current research in scene classification faces many challenges (Roy et al., 2018). Firstly, scenes are complex and diverse, and scene images may have a large difference even captured in the same type of scene. Secondly, external factors may bring interference during shooting. In the same scene, different shooting angles can cause visual differences among scene images. During shooting, light, occlusion, and resolution also make this visual difference even noticeable. Therefore, the key issue for scene recognition is to find suitable image features, and then using these features to obtain structural information which can reflect the scene category. At the same time, insignificant differences such as texture details should be suppressed.

Normally, features used in scene recognition include low-level features and high-level features (Tian, 2013). Usually, low-level features are images’ original attributes extracted based on pixel points. They focus on subtle tex-
ture information of the image. High-level features of an image are extracted by modelling based on low-level features, and they provide rich image semantic information, e.g. object bank feature (Li, Su, Lim, & Fei-Fei, 2010; Li, Su, Xing, & Fei-Fei, 2010), bag of parts feature (Juneja, Vedaldi, Jawahar, & Zisserman, 2013), latent pyramidal regions (Sadeghi & Tappen, 2012), etc. However, features obtained using the methods described above have some limitations. Though they can achieve good results for general scene recognition tasks, the features’ effects deteriorate with increasing complexity of data sets.

Studies have shown that features derived from deep learning have better robustness, efficiency, and lower error rates (Tang, Wang, & Kwong, 2017). In this paper, a residual network-based image feature extraction and classification system entitled Scene-RecNet is proposed. It has been found that performance improvements of CNN are achieved mostly by optimizing the feature extractor structure. A large number of image features can be obtained by extracting the output of the convolutional layers and fully connected layers of CNN. Scene-RecNet is based on a deeper number of ResNet and the feature extraction module of ResNet is improved. A feature adjustment module consisting of one convolutional layer and one fully connected layer is introduced to adjust the features obtained from the feature extractor. The fully connected layer performs feature dimension reduction for easier storage. The experimental results show the effectiveness of Scene-RecNet.

The organization of this paper is as follows. Section 2 reviews the related research work, and Section 3 introduces the proposed scene image recognition model Scene-RecNet. Section 4 describes the experiments and also analyses the system performance. Section 5 gives the summary and further work.

2. Related work

Feature is the most important factor in image recognition, and scene recognition is no exception. Scene recognition technology has evolved from low-level feature based methods to high-level feature based methods. Now, commonly used are feature learning-based methods (Li et al., 2015).

There are lots of low-level features, e.g. HSV (Hue, Saturation, Value) (Wang, Li, Wang, Liu, & Li, 2013), GIST (Oliva & Torralba, 2001; Oliva & Torralba, 2006), SIFT (Scale-Invariant Feature Transform) (Lowe, 2004), HOG (Histogram of Oriented Gradient) (Navneet & Bill, 2005), SPM (Spatial Pyramid Matching) (LaZebnik, Schmid, & Ponce, 2006), BOW (Bag of Words) (Banerji, Sinha, & Liu, 2013) etc. HSV describes colours from the perspective of hue, saturation, and value in human visual system. It is suitable for image processing and recognition. Hue means different colours, saturation refers to various shades and value refers to the degree of light and dark. HOG describes local characteristics of an image through statistical edge gradient information (Andrews, Tsochantaridis, & Hofmann, 2002).

SIFT extracts scale-invariant stable points in images and finishes objects matching by comparing these points. The SIFT descriptor is very robust to deformations such as translation, rotation and affine. In BOW, after all the features have been extracted, similar features form a visual vocabulary. A histogram is created according to the vocabulary dimensions and each dimension represents a visual vocabulary that describes a feature of the image.

SPM algorithm considers spatial information. It divides an image into several blocks and extracts the features of each block separately. Features of each block are put together to form the complete image features. SPM uses a multi-scale blocking method and it has the structure of a hierarchical pyramid.

The basic idea of fisher vector coding is to build a visual dictionary based on GMM (Gaussian Mixture Model) so as to increase the image feature dimensions and to describe an image better (Perronnin & Dance, 2007). Through space description, GIST features can show macro characteristics of an image, such as naturalness, openness, roughness, swell, and steepness.

Feature learning refers to the process of obtaining potential image features through a series of mappings and transformations of the input image pixels, and it belongs to the category of machine learning. Convolutional Neural Network (CNN) is a multi-layer and non-fully connected neural network and the first deep learning algorithm to use multi-level network structures successfully (Matheus, Graham, & Fergus, 2011). CNN has many successful applications in image processing (Ding et al., 2018; Feng, Shen, & Liu, 2018).

CNN generally consist of convolutional layers, pooling layers, and fully connected layers (LeCun, Bottou, Bengio, & Haffner, 1998). Each layer consists of multiple two-dimensional planes, and each plane has multiple independent neurons. A convolutional layer and a pooling layer form a convolution group and multiple convolution groups increase the depth of CNN. Figure 1 shows the structure of a typical CNN.

As a deep neural network, CNN can capture complex features from image data. Several structures of CNN have become into being and they outperform traditional methods in the area of image recognition to a great extent (Kheradpisheh, Ghodrati, Ganjtabesh, & Masouei, 2016; Simonyan & Zisserman, 2014). They have been pre-trained using images from the ImageNet database which contains approximately 1.2 million images of 1000
ordinary objects (Mishkin, Sergievskiy, & Matas, 2017; Russakovsky et al., 2015). Table 1 compares these network structures.

Among all, VGGNet (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014) and GoogLeNet (Szegedy, Liu, et al., 2015) are the most frequently used structures. VGGNet contains many levels of networks with the depth ranging from 11 to 19, and the commonly used ones are VGGNet-11, VGGNet-16 and VGGNet-19. VGGNet divides the network into five segments and each segment concatenates multiple 3 × 3 convolutional networks. Each convolution layer is followed by a maximum pooling layer. The last part of the network is three fully connected layers and one Softmax layer. At ILSVRC 2014, its Top-5 error was 7.32%.

GoogLeNet, proposed by Google in 2014, was the champion of ILSVRC 2014 competition classification task with Top-5 error only 6.67%. It introduces Inception structure, and 1 × 1 convolution kernel is used to reduce the dimension of the feature map and also increase the depth of network indirectly. Inception V1 consists of 22 convolution layers without containing fully connected layers. At the same time, the auxiliary classifier is used to adjust the low-level weights.

With the increase of network depth, network performance may become saturated and then degrade rapidly. To solve this problem, a deep residual learning framework was proposed. Shortcut connections providing identity mapping are introduced and their outputs are added to the outputs of the stacked layers (He et al., 2015). This network won the 1st place in the ILSVRC 2015 classification competition with the Top-5 error rate 3.57%.

3. Scene image recognition technology based on deep residual network

3.1. Scene image recognition model

Based on deep residual network, a scene image recognition model entitled Scene-RecNet is proposed in this paper. The key to scene image recognition is to find feature vectors that can describe the image accurately with less time and space overhead. Therefore, the focus is on how to extract image features with strong representation and at the same time to guarantee a certain recognition rate even when an image is compressed or tampered with.

Figure 2 shows the structure of Scene-RecNet. It consists of three parts, i.e. feature extractor, feature adjustment layer and classifier, respectively.

3.2. Image feature extraction

The process of image feature extraction is as follows. First, the convolutional neural network is trained to finish the parameter adjustment, and then the original classifier is removed and the adjustment layer is inserted. Since the features are learned through network learning, they have good generality, robustness and strong characterization.

3.2.1. Network structure

In order to improve the features’ quality, ResNet is chosen as the feature extraction model. Through residual
connection and block overlay, ResNet is able to increase network depth and also ensure the richness and diversity of features extracted. The pre-residual network consists of 4 convolution block stacks with a total of 50 layers. The input image size is $224 \times 224 \times 3$. The convolutional layers in the block use $1 \times 1$ and $3 \times 3$ small convolution kernels for parameter reduction. The feature adjustment layer consists of a convolution layer $P_1$ (512 outputs) and a fully connected layer $P_2$ (512 outputs). $1 \times 1$ convolution kernels are used in Layer $P_1$. $P_1$ is designed to abstract the general features extracted by ResNet on ImageNet dataset, and then to generate features for the target datasets. The fully connected layer $P_2$ reinforces the features from $P_1$, and at the same time to reduce the end-to-end features nonlinearly for better feature storage. The node number of $P_2$ can be set according to actual storage requirements. Section 4 provides the simulation to investigate the influence of dimension and depth of layer $P_2$ on feature performance.

### 3.2.2. Training

There are two phases in network training, i.e. pre-training phase and feature adjustment phase. The first stage is the training of residual network, and the feature adjustment layer is masked during this phase. The classifier classifications are modified to 1000 and the model is trained on the ILSVRC2012 dataset. The training super-parameters are chosen according to ResNet of MSRA. The second stage is the dataset feature adjustment and dimension reduction by applying the feature adjustment layer. The network output is set to be the scene image dataset categories. Parameters of the feature extractor for this stage are provided by the pre-training model, and the parameter initialization mode for feature adjustment layer and the classifier weights is set to ‘Xavier’.

‘Xavier’ is chosen to accomplish initialization as it can ensure evenly information distribution across the network. The derivation process is quite complicated, and only a brief introduction is given here. For the detailed process, please refer to He, Zhang, Ren, and Sun (2014). Assume the input and output dimension of the layer for the parameters are $m$ and $n$, respectively, then the parameters will be initialized uniformly in the range of $\left[ -\sqrt{6/(m+n)}, \sqrt{6/(m+n)} \right]$. The main idea is to make the variance of a neuron’s input weight (when using back propagated, it can be seen as the output weight) equal to $1/n$.

To ensure the feature stability, convolution kernels of the first three convolutional blocks are not trained, and that of the fourth block is allowed to be adjusted during training.

During the training phase, the size of the input image should be $224 \times 224$. The image width and height need to be scaled in proportion so that the short side satisfies 224 pixels, and the standard image is obtained by central cropping. During training, horizontal flipping is used to increase the number of images, and the mean of the image’s each channel is subtracted during preprocessing.

For the readers’ convenience, the training algorithm of Scene-RecNet is summarized below.

**Algorithm:** Scene-RecNet training  
**Input:** Training dataset (ILSVRC2012 dataset)  
**Output:** a scene recognition network

1. Images in the training dataset are cropped to $224 \times 224$.
2. Split each image to obtain its R, G and B channel separately.
3. For each channel, the mean of the channel is subtracted.
4. Train the residual network while the feature adjustment layer is masked.
5. Enable the feature adjustment layer and the parameter initialization mode for feature adjustment layer and the classifier weights is set to ‘Xavier’.
6. Adjust the network parameter till the loss function satisfy the pre-determined constrain.

Steps (1)–(3) belongs to the image preprocessing. Step (4) is for pre-training and the last two steps are for parameter adjustment.
3.2.3. Generality of the model

The feature extraction network proposed in this paper has quite good generality. In the case of limited computing resources, the existing pre-training models can be used to replace the feature extractors such as VGGNet and Inception directly. The idea of the feature adjustment layer comes from the NOC network (He et al., 2014). It shows that the network detection effect can be improved by adding a convolution layer. The feature adjustment layer can change the network size accordingly, i.e. to improve the feature quality by increasing the network depth and to accelerate convergence by reducing the number of parameters.

3.3. Image recognition

After the image features have been extracted, two kinds of outputs are obtained, i.e. the feature vectors and the image classification labels. The number of feature vectors of an image is determined by the setup of training models. The shorter the extraction interval, the more feature vectors are obtained. The last feature vector obtained has the best effect in identifying the image. Increasing the total number of trainings can also increase the number of feature vectors, which yields to better recognition effect. The image to be recognized is identified according to the feature vectors first, and then classified according to the labels obtained. Finally, the label with the highest probability is output as the image label.

4. Experiments

4.1. Experimental environment

The hardware configuration of the experiments is a computer with CPU of Intel core i7 6700HQ, 8 core 16 lines, 2.60 GHz frequency, 16 GB memory, NVIDIA GTX 1060.

Figure 3. Examples of images in self-made data sets.
6GB. The operating system is Windows10, and the programming language is Python. The experimental framework is Tensorflow, and Toolkits used are Sklearn and Skimage.

Two self-made data sets Scene-21 and Scene-30 are used in the experiments to increase the validity and reliability of the experimental results, and there are 21 scenes in Scene-21 and 30 scenes in Scene-30. Images in Scene-21 and Scene-30 are from Google and Baidu Images, including airports, forests, rivers and other scenes. There are about 100 images for one scene. Figure 3 shows some examples in Scene-21.

### 4.2. Experiments of feature accuracy and quality

The purpose of this experiment is to verify the accuracy and quality of features obtained using Scene-RecNet, and self-made data set Scene-21 is chosen. The comparison algorithm includes traditional low-level feature extraction algorithms such as HSV, SIFT + BOW, ScSPM, HOG, GIST, and neural network models VGGNet and GoogleNet. Table 2 shows the performance comparison results.

In this experiment, random forest classifier RF is chosen to achieve the image feature classification performance evaluation, and the number of decision trees is set to 25. It can be seen from Table 2 that generally the feature dimension of traditional algorithms is lower compared with those of the neural network models. On one hand, PCA is used for linear dimensionality reduction, on the other hand, specific coding methods are used, such as BOW and SPM. In terms of image classification correctness, the neural network feature extraction algorithms have much better performance. The main reason is that in traditional algorithms, features need to be designed manually and prior knowledge is required. Thus the feature generality is not good, and it is not easy to capture the essential features of images in complex scenes. Relying on high dimensional parameter space and the end-to-end learning structure, neural network models are able to abstract and synthesize advanced features gradually from bottom to top. The data-driven self-learning method ensures that CNN has excellent feature extraction capabilities. Based on residual neural network, Scene-RecNet model has deeper layers and residual connection mode, so the quality of its feature extractor is higher than that of shallow neural networks and it obtains higher classification correctness.

### 4.3. Generality analysis of the feature adjustment layer

This experiment is to test the generality of the feature extractor and its influence on the overall network performance. The data set used is Scene-30, and the comparison algorithms are VGGNet-19 and GoogLeNet. In the experiment, the feature extraction model of Scene-RecNet is inserted to both VGGNet-19 and GoogLeNet as follows.

For VGGNet-19, its output layer is removed. FC7, the penultimate layer of the fully connected layer, is used as the feature extractor, and its number of neurons is changed to 1024. The feature adjustment layer network has the structure of FC_P1/512, FC_P2/512.

For GoogLeNet, its auxiliary classifier and main classifier output layer are removed, and a 1 × 1/512 convolution layer is added as the feature extractor. The feature adjustment layer network has the structure of Conv_P1/512, FC_P2/512.

Table 3 gives the performance comparison of these network models.

As can be seen from Table 3, all networks using the feature adjustment layer obtain better feature performance compared with their counterpart original network. The classification accuracy is improved by 2–8% in average. At the same time, the output characteristics of the FC_P2 layer have a significant dimensionality reduction effect. On the premise of accuracy rate guarantee, the features can be compressed greatly. In terms of classification accuracy, among the three network models using adjustment layer, Scene-RecNet outperforms the other two with relatively shallow networks. This experimental result is consistent with that of the previous experiment. In summary, the neural network with feature adjustment layer has good generality and supports multiple feature extraction networks.

### Table 2. Comparison of feature extraction performance.

| Feature extraction algorithm | Feature dimension | Classification accuracy (%) |
|-----------------------------|-------------------|-----------------------------|
| HSV                         | 256               | 29.5                        |
| SIFT-PCA + BOW (1000)       | 1000              | 32.0                        |
| ScSPM (500)+PCA             | 1024              | 36.5                        |
| HOG + PCA                   | 128               | 35.0                        |
| GIST                        | 960               | 41.8                        |
| VGGNet-11                   | 4096              | 86.9                        |
| VGGNet-16                   | 4096              | 89.0                        |
| VGGNet-19                   | 4096              | 88.1                        |
| GoogLeNet (Pool5)           | 1024              | 91.8                        |
| Scene-RecNet                | 512               | 94.4                        |

### Table 3. Comparison of network versatility.

| Network model | With/without feature adjustment layer | Feature dimension | Classification accuracy (%) |
|---------------|---------------------------------------|-------------------|-----------------------------|
| VGGNet-19     | ×                                     | 4096              | 66.2                        |
| VGGNet-19     | √                                     | 512               | 74.8                        |
| GoogLeNet     | ×                                     | 1024              | 70.0                        |
| GoogLeNet     | √                                     | 512               | 76.8                        |
| ResNet        | ×                                     | 1024              | 74.6                        |
| Scene-RecNet  | √                                     | 512               | 76.6                        |
4.4. Analysis of dimension reduction and depth increase effect of feature adjustment layer

This experiment investigates the impact of network dimension reduction and depth increase on the overall feature performance of Scene-RecNet. The data set used is Scene-30. Dimension reduction and depth increases are all applied to the feature adjustment layer FC_P2.

4.4.1. Effect of dimension reduction
In Figure 4, the abscissa represents the dimension of features and the ordinate represents the classification accuracy rate. It can be seen that the dimension reduction of the feature adjustment layer has a very small impact on the overall network feature performance, and the error of the classification accuracy rate is maintained within 1–2%. Therefore, as long as the classification accuracy rate is desirable, the features can be compressed greatly to reduce storage and calculation cost.

4.4.2. Effect of network depth increase
Figure 5 gives the simulation results about the effect of network depth, where the abscissa indicates the number of network layers, and the ordinate indicates the classification accuracy.

It can be seen from Figure 5 that increasing the depth of the feature adjustment layer can increase the classification accuracy rate. However, compared with the increase of the computation time for image classification, the accuracy increase is not that significant. A compromise needs to be found between classification accuracy and its computation time.

5. Conclusion

Traditional image feature extraction algorithms have low versatility and stability, which limit their application. Convolutional neural networks are highly versatile and stable in feature extraction. In order to improve the correct rate of scene image recognition, a novel deep residual network-based model Scene-RecNet is proposed. Scene-RecNet consists of a feature extraction layer and a feature adjustment layer, and the feature adjustment layer is composed of a convolution layer and a fully connected layer. The feature adjustment layer has the function of dimension reduction while guaranteeing the scene recognition accuracy. Experiments show that Scene-RecNet has the advantages of fast calculation speed, small storage space and high recognition accuracy and it is superior to other existing convolutional neural network structures in feature learning.

Now the images in the self-made data sets are all top-view images, and the complexity of this scene is relatively low. In the next step, indoor scenes with higher scene complexity will be handled. The most difficult issue for indoor scene recognition comes from the objects existing everywhere in the room. So the objective of the future research is to minimize the impact of objects in indoor scenes to improve the recognition rate of indoor scenes.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by National Natural Science Foundation of China [grant numbers 61573316, 61873082], and the National Key R&D Program of China [grant number 2018YFB0204003].

References

Andrews, S., Tsochantaridisi, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. In Advances in neural information processing systems (pp. 561–568).
Banerji, S., Sinha, A., & Liu, C. (2013). A new bag of words LBP (BoWL) descriptor for scene image classification. Computer Analysis of Images and Patterns, 8047, 490–497.
Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. Computer Science, arXiv:1405.3531 v4 [cs.CV].

Choset, H., & Nagatani, K. (2001). Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. IEEE Transactions on Robotics and Automation, 17, 125–137.

Ding, X., Luo, Y., Li, Q., Cheng, Y., Cai, G., Munnoch, R., … Wang, B. (2018). Prior knowledge-based deep learning model for indoor object recognition and application. Systems Science & Control Engineering, 6(1), 249–257.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. IEEE Transactions on Robotics and Automation, 17, 125–137.

Espinace, P., Kollar, T., Roy, N., & Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. Robotics and Autonomous Systems, 61(9), 932–947.

Feng, F., Shen, B., & Liu, H. (2018). Visual object tracking: In the simultaneous presence of scale variation and occlusion. Systems Science & Control Engineering, 6(1), 456–466.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. 13th European conference on computer vision.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition.

Heineke, D. Z., Graham, W. T., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In IEEE international conference on computer vision (pp. 2018–2025).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097–1105).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer society conference on computer vision and pattern recognition (pp. 2119–2178).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Li, E. J. (2018). Remotely sensed image scene deep learning and application: A case study of urban structure types recognition. Geography and Geo-Information Science, 34(6), 127–127.

Li, X., Shi, J., Dong, Y., & Tao, D. (2015). Overview of scene image classification techniques. China Science: Information Science, 45(7), 827–848.

Li, L. J., Su, H., Lim, Y., & Fei-Fei, L. (2010). Objects as attributes for scene classification. In European conference on computer vision, Heraklion (pp. 57–69).

Li, L. J., Su, H., Xing, E. P., & Fei-Fei, L. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification. In Advances in neural information processing systems (pp. 1378–1386).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 91–110.

Mao, J., Zhong, D., Hu, Y., Sheng, W., Xiao, G., & Qu, Z. (2018). An image authentication technology based on depth residual network. Systems Science & Control Engineering, 6(1), 57–70.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42, 91–110.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research, 155, 23–36.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211–252.

Szegedy, C., Ioffe, S., & Vanhoucke, V. (2016). Inception-v4, Inception-ResNet and the impact of residual connections on learning. Retrieved from https://arxiv.org/pdf/1602.07261.pdf

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In IEEE conference on computer vision and pattern recognition (pp. 1–9).

Tang, P., Wang, H., & Kwong, S. (2017). G-M52F: GoogleLeNet based multi-stage feature fusion of deep CNN for scene recognition. Neurocomputing, 225, 188–197.
Tian, D. P. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering, 8*(4), 385–395.

Wang, Q., Li, Z., Wang, J., Liu, S., & Li, D. (2013). A fast processing method of foreign fiber images based on HSV color space. *IFIP Advances in Information & Communication Technology, 392*, 390–397.

Wang, B. F., Zhou, J. L., Tang, G. S., Di, K., Wan, W., Liu, C., & Wang, J. (2014). Research on visual localization method of lunar rover. *Scientia Sinica (Informationis), 44*, 452–460.

Wu, W. R., Wang, D. Y., Xing, Y., Gong, X., & Liu, J. (2011). Binocular visual odometry algorithm and experimentation research for the lunar rover. *Scientia Sinica Informationis, 41*, 1415–1422.

Xiao, J. X., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE conference on computer vision and pattern recognition*, San Francisco (pp. 3485–3492).

Yu, X. R., Wu, X. M., Luo, C. B., & Ren, P. (2017). Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing, 54*(5), 741–758.

Zeller, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhou, W. Y., He, X. H., Qing, L. B., Wan, Y. J., & Zheng, X. B. (2019). Recognizing building areas under construction in complex scenarios. *Computer Systems & Applications, 28*(1), 140–146.

Zhou, L., Hu, D. W., & Zhou, Z. T. (2013). Scene recognition combining structural and textural features. *Science China Information Sciences, 56*(78), 1–14.