Remote Sensing Image Scene Classification Based on SURF Feature and Deep Learning

Jinxiang Liang, Jianwu Dang, Yangping Wang, Jingyu Yang, Zhenhai Zhang
School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China
Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphic & Image Processing
Gansu Provincial Key Lab of System Dynamics and Reliability of Rail Transport Equipment
Lanzhou Bocai Technology Co., Ltd.
E-mail: liangjxp@outlook.com Tel: +86-19893170181

Abstract— **Remote sensing image scene classification is one of the key points in remote sensing image interpretation. The traditional remote sensing image scene classification feature performance is not strong, and the deep learning extraction semantic feature process is complex. This paper proposes a fusion feature remote sensing image scene classification method which is based on artificial features and deep learning semantic features. Firstly, the SURF feature of the remote sensing image is extracted and encoded by the VLAD algorithm. The semantic feature of a remote sensing image is extracted by transfer learning. Then the feature reduction is performed by PCA algorithm and feature fusion is performed. Finally, the scene classifier is trained by using the random forest algorithm. The experimental results show that the classification accuracy and Kappa coefficient of this method are higher and the method is effective.**

I. INTRODUCTION

Remote sensing image scene classification is a research focus in the field of remote sensing technology application. It refers to the process of automatically discriminating image scene categories by learning the feature information of image data. Remote sensing image scene classification technology can provide important auxiliary decision support in specific practice applications such as ecological monitoring, engineering supervision, and military reconnaissance. The in-depth study on remote sensing image scene classification has profound research significance and application value.

The core of the remote sensing image scene classification problem is to extract discriminative expression features to maximize the difference between scene classes and minimize the differences between classes. Around this point, the majority of scholars have made many efforts[1][2][3]. In the early days, many studies focused on designing effective artificial features to obtain basic description information on the color, texture, and structure of remote sensing images to achieve the purpose of scene differentiation. The image local feature descriptor (SIFT, Scale-Invariant Feature Transform) proposed by David Lowe has the invariance of scale-space rotation and scaling, and is applied in simple natural scene classification[4][5]. The Local Binary Patterns proposed by Ojala et al., which characterizes images by describing texture features, are used in detection and classification[6][7][8]. Dalal et al. proposed HOG (Histograms of Oriented Gradients), which obtained good experimental results by counting pixel edge gradients and expressing structural features[9]. By designing such artificial features for the classification of remote sensing image scenes, although some effects have been achieved, they are relatively simple, the information-carrying capacity is limited, and the applicable image scenes are relatively simple. With the increasing complexity of remote sensing images, the limitations of artificial feature classification based on artificial features are more obvious. Therefore, the semantic information reorganization around the underlying features and the enhancement of feature expression have become the research trend. Zhu et al. constructed an effective remote sensing image scene classifier through the BOVW (Bag Of Visual Word) model combined with image local and global feature information[11]. Sheng et al. used SC (Sparse Coding) method to classify remote sensing image scenes by combining multiple features and achieved good results[12]. Zhao et al. used the Fisher Kernel coding method to map local low-level features into new feature vectors by the Gaussian mixture model for scene classification, which achieved good results[13]. With the rise of deep learning, remote sensing image processing based on convolutional neural networks has become the focus and hotspot of current research. Zou et al. used the DBN (Deep Belief Network) to construct remote sensing image representation features and achieved the unsupervised method for better scene classification[14]. Hu et al. studied the situation of remote sensing image scene classification by studying the convolutional neural network and analyzed the operation of transfer learning[15]. Wang et al. extracted the hierarchical feature information by transferring the VGG network and the ResNet network, re-encoding and applying it to the remote sensing image scene classification[16]. Zhang et al. obtained the local hierarchical structure features of remote sensing images through convolutional neural networks, and combined with the CapsNet network for classifier training, and achieved good classification results[17].

At present, the use of convolutional neural networks to extract image features for remote sensing image scene classification has become a mainstream research method. However, the deep learning algorithm still has some defects. First of all, a large amount of training data sets are needed, the effect of transfer learning is not obvious, and deep learning is not sensitive to scale changes and rotation changes of remote sensing images. In this paper, a combination of traditional artificial features and deep learning algorithms is used to construct a remote sensing image scene classification model, which has achieved good experimental results.

The main contributions of this paper are: (1) A remote sensing image scene classification model combining SURF features and deep learning semantic features is designed; (2) The introduction of SURF features compensates for the shortcomings of deep learning that are insensitive to scale changes and rotational changes in remote sensing images; (3) The VLAD (Vector Of Locally Aggregated Descriptors) algorithm is used to re-encode the SURF features to obtain enhanced regularization features; (4) Design of remote sensing image scene classifier using random forest algorithm.

The rest of the paper is organized according to the following structure: the second part introduces the method model proposed in this paper in detail, the third part uses the method of this paper to conduct experiments and analysis, and the fourth part summarizes the content.

II. RELATED WORKS

In this paper, for the remote sensing image data, the SURF features are extracted first, then re-encoded by VLAD algorithm to form regularized expression features, and then normalized. The PCA algorithm is used to reduce the features and form artificial design features; At the same time, using the transfer learning method, the VGG16 pre-training network is used to extract the semantic features of the remote sensing image, and after the normalization processing and feature dimension reduction, the deep learning semantic features are formed. Finally, the two features are combined as a fusion feature to train the random forest classifier to obtain the final remote sensing image scene classification model. The complete process is shown in Fig. 1.

A. SURF Features

The SURF (Speed Up Robust Features)[18] feature is a local feature descriptor for image retrieval and matching. The principle is similar to SIFT, and it also has the advantages of rotation invariance and scale invariance. Due to the adoption of the Harr feature and the concept of the integral image, the algorithm execution efficiency is improved. The specific steps are: Firstly, the Hessian matrix is constructed according to the (1), all the points of interest are generated, then the scale-space

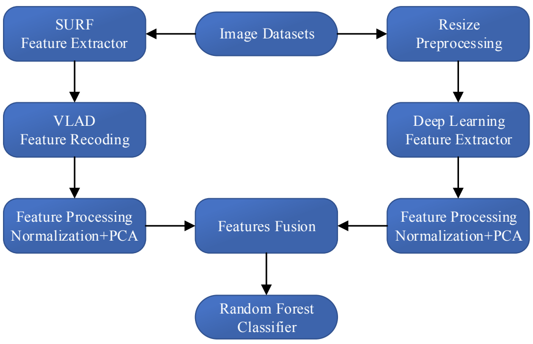

Fig. 1 The schematic diagram of the overall process

is constructed according to the (2), the feature points are determined by non-maximum suppression, and the main direction of the feature points is determined by statistical Harr wavelet features. Finally, generate the corresponding descriptor.

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{1}$$

where $L_{xx}(x,\sigma)$ is convolutional of Gaussian second-order derivative with the image in point x, similar to the others.

$$L(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} * I(x,y) \tag{2}$$

where $*$ is the convolution operation in x and y.

A remote sensing image can generate more SURF feature points. This information can describe the local detail information of the scene and can be used as the information basis for scene discrimination. Moreover, when deep learning is used to extract semantic features, the image is preprocessed, which causes the image to have a certain degree of information loss. The SURF feature can effectively retain the information, thus forming benign information with the deep learning semantic features. Complementary, providing a richer feature base for scene classification.

In order to enhance the information expression ability of SURF features and better feature fusion with deep learning semantic features, this paper re-encodes the SURF features using VLAD[19] algorithm. Characterizing remote sensing images as regularized vectors to better adapt to scene classification problems. The VLAD algorithm first uses the k-means clustering algorithm to perform visual word training on the acquired SURF features, and then assigns the SURF features of each image to the K cluster centers according to the nearest neighbor principle, and then makes residual sum for each clustering center according to (3), and then the residual is subjected to L2 normalization according to equation (4), and finally it is spliced into a long vector, that is, the final coding feature is obtained.

$$v_{i,j} = \sum_{x \text{ such that } NN(x)=c_i} x_j - c_{i,j} \tag{3}$$

where x_j and $c_{i,j}$ respectively denote the j^{th} component of the descriptor x considered and of its corresponding visual word c_i.

$$v := v/\| v \|_2 \tag{4}$$

The description dimension of the feature points of the SURF algorithm is 64. In order to integrate well with the deep learning semantic features, the number K of cluster centers is set to 64, so that the feature dimension after encoding by the VLAD algorithm is 4096, and deep learning semantic features remain the same structure.

B. Deep Learning Semantic Features

Deep learning semantic features are image representation information extracted by nonlinear transformation using a multi-layered perception network. Since the convolutional neural network has good information self-extraction ability in image processing, this paper uses a convolutional neural network to extract the semantic features of remote sensing images. The use of deep learning models requires many annotated samples. However, the amount of data in the datasets of remote sensing images is generally small, so this paper uses the method of transfer learning for feature extraction. In the classical convolutional neural network architecture, VGG[20] has a relatively simple network structure and shallow depth. The 3*3 small convolution kernel reduces the parameter quantity and achieves higher classification accuracy. The ReLU activation function (5) effectively prevents over-fitting, so the VGG16 network model is used for deep learning semantic feature extraction.

$$f(x) = max(0, x) \tag{5}$$

Since the image data input shape processed by the network is [224, 224, 3], before the network learning, the data set needs to be preprocessed, and the image size is adjusted to the corresponding format. Then the training data set is processed by the network, and the feature extraction is performed by connecting the custom fully connected layer to obtain the deep learning semantic feature. In order to obtain more accurate expression features, after initial network transfer, use the fine-tuning to improve the network model. The low-level network generally obtains the basic information of the image, including lines and edges, etc., and the information is universal, so the low-level network can directly adopt the pre-trained parameters. The high-level network learns the semantic information related to the characteristics of the data set through these basic features and can perform fine-tuning training. The fine-tuning strategy will be different for different data sets. In this paper, it is verified by experiments that for the VGG16 network, it is relatively effective to select the first fully connected layer output as the final deep learning semantic feature by fixing the first 14 layers of network parameters. The output feature dimension is 4096.

C. Features Fusion

For the SURF features encoded by the VLAD algorithm and the deep learning semantic features acquired by the transfer learning, this paper uses a joint approach for feature fusion. Since the extracted SURF features and deep learning semantic features are all 4096-dimensional high-dimensional data, direct fusion will lead to double the dimension and higher computational complexity. Therefore, the two kinds of feature information are respectively normalized before being merged. Once processed, the PCA algorithm is used for data dimensionality reduction. After balancing the richness of classification features and computational complexity, the number of principal components after the dimension reduction is chosen to be 2048. Finally, the two features after dimension reduction are combined to form the final fusion classification feature.

D. Scene Classifier

The classifier outputs the class discrimination result by performing operations on the features. In general, classifiers can be divided into two categories, linear discriminant classifiers and nonlinear discriminant classifiers. The linear discriminant classifier separates different categories of space by constructing a linear hyperplane, and the calculation is simple. The nonlinear discriminant classifier constructs a nonlinear hyperplane to separate different types of spaces through mapping, etc., and the classification effect is better. This paper builds a classifier based on a nonlinear discriminant model using a random forest algorithm. The random forest algorithm is calculated based on the decision tree. The decision tree calculates the information entropy according to (6), constructs the decision branch by calculating the information gain, and finally constructs a complete decision tree.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \, log \, p(x_i) \tag{6}$$

The random forest algorithm trains multiple decision trees in a random way, and finally make a unified decision according to (7). This integration strategy can effectively avoid the model bias of a single decision tree, strengthen the generalization ability of the model, and retain the nonlinear discriminant classification. The advantages of the device can enhance the accuracy of remote sensing image scene classification. For the random forest algorithm, the number of weak classifiers is set to 100, and the maximum number of used features is 64.

$$P_c(x) = \frac{1}{t}\sum_{j=1}^{t} P(c|v_j(x)) \tag{7}$$

III. Experimental

In order to verify the effectiveness of the proposed method, this paper uses the UCMerced_LandUse and SIRI-WHU two common remote sensing image data sets for experimental verification. UCMerced_LandUse image data is rural area information collected by the US Geological Survey. The image size is 256*256, including 21 different scenes. Each scene has 100 images. Some pictures are shown in Fig. 2. The SIRI-WHU image data comes from Google Earth and is collected by Wuhan University. The image size is 200*200, including 12 different scenes, each with 200 images. Some images are shown in Fig. 3.

A. UCMerced_LandUse

Fig. **4** is the whole process of deep learning training. In the

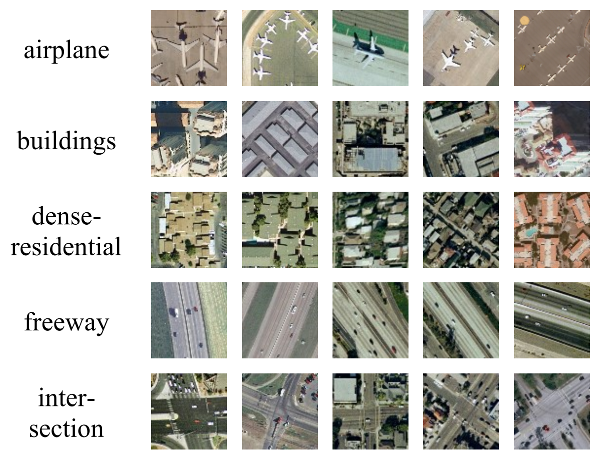

Fig. 2 UCMerced_LandUse dataset examples

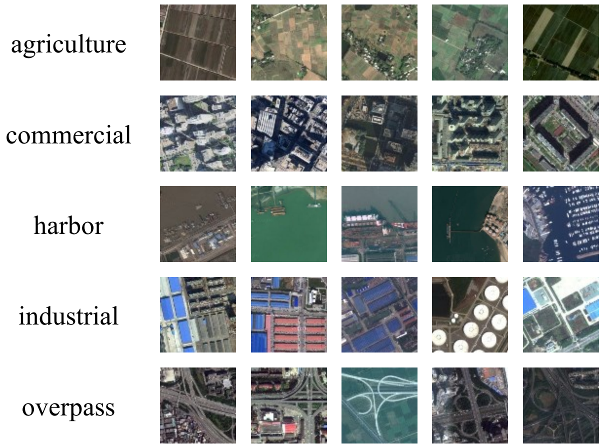

Fig. 3 SIRI-WHU dataset examples

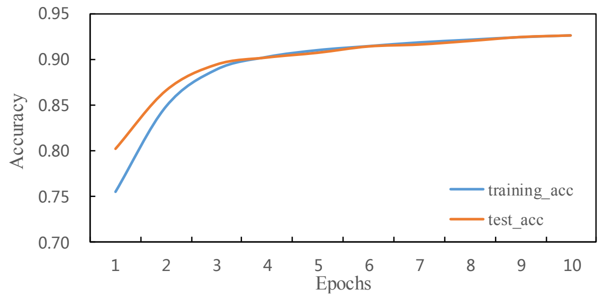

Fig. 4 the process of deep learning training

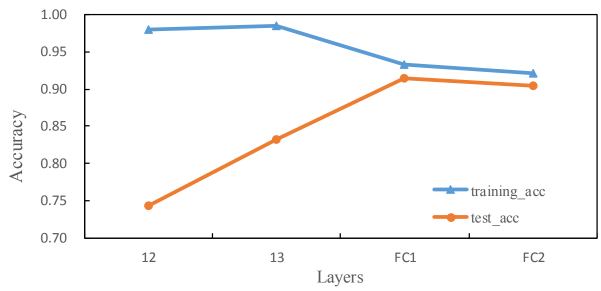

Fig. 5 the fine-tuning experimental result

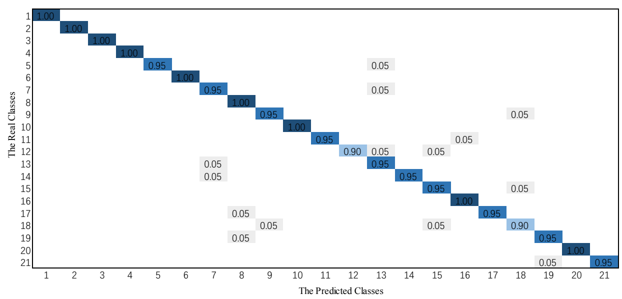

Fig. 6 confusion matrix

epochs of about 10, the classification accuracy rate is close to 92.45% and the classification accuracy rate reaches the saturation level.

Tab. 1 Comparison of different algorithms

algorithm	accuracy	Kappa
SIFT+DL	96.17%	0.958
HOG+DL	92.34%	0.895
CaffeNet[2]	95.02%	0.943
VGG-VD-16[2]	95.21%	0.950
GoogLeNet[2]	94.31%	0.936
CNN+MKL[21]	96.43%	0.962
The Paper	**96.67%**	**0.965**

Fig. **5** shows the fine-tuning effect on the data set. It can be seen that all convolutional layer parameters of VGG16 are fixed, and only the fully connected layer is retrained. Although the training set accuracy rate is good, the test set performance is poor, because the data set sample is insufficient, and the training parameters are too many, caused the over-fitting problem; while fixing the FC2 full connection layer parameters, retraining the classifier has little effect on the accuracy improvement, indicating that the FC1 fully connected layer has obtained distinguishable feature information, so the first fully connected layer is adopted. The output as a deep learning semantic feature.

After the SURF feature and deep learning semantic features are combined, after training, the confusion matrix is shown in Fig. **6**. The final test accuracy is 96.67%. Compared with Fig. **4**, the classification accuracy is improved by about 4%. It shows that the SURF feature can be combined to provide information complementation for deep learning semantic features, and combined with the random forest algorithm, it can effectively classify remote sensing image scenes. By calculating the Kappa coefficient, the result is 0.965, indicating that the classification results are highly consistent.

Compared with other commonly used image local description features, our paper compares the SIFT and HOG features encoded by the VLAD algorithm and compares it with other literature methods. The results are shown in Tab. **1**. The method using SIFT features is similar to our method, but the processing complexity is high. The HOG feature itself has a higher dimension and the accuracy of adoption is lower. In addition, comparison with other literature methods shows that our method is effective.

B. SIRI-WHU

Fig. **7** is a deep learning semantic feature training process under the data set. A closer observation reveals that, under the data set, although the same output layer features are used compared to experiment one, the accuracy of training classification based on deep learning semantic features in this scenario is 95.47%, which is slightly higher than experiment one 3% or so, which indicates that the data set size has a significant impact on the deep learning semantic features. Increasing the data set can effectively enhance the expressive power of deep learning semantic features.

After the SURF feature and the deep learning semantic feature are combined, after training, the confusion matrix is shown in Fig. **8**. The accuracy of the overall classification is

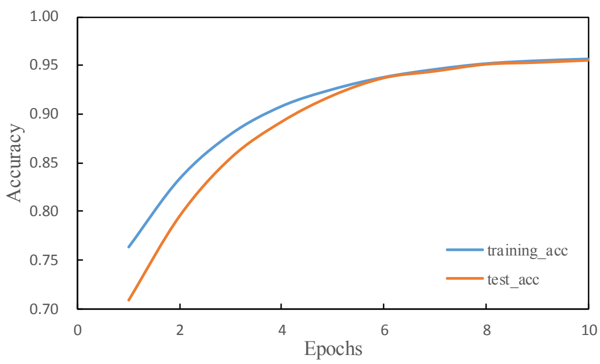

Fig. 7 the process of deep learning training

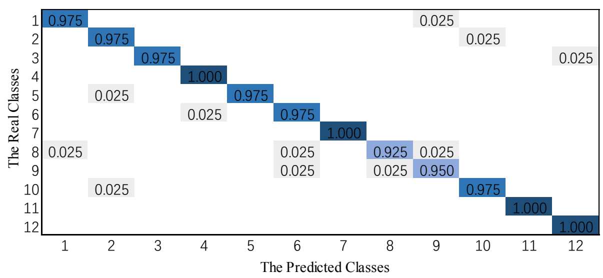

Fig. 8 confusion matrix

Tab. 2 Comparison of different algorithms

algorithm	accuracy	Kappa
SIFT+DL	97.23%	0.965
HOG+DL	93.47%	0.918
CaffeNet[2]	95.74%	0.953
VGG-VD-16[2]	96.85%	0.961
GoogLeNet[2]	95.56%	0.947
CNN+MKL[21]	97.12%	0.966
The Paper	**97.70%**	**0.975**

97.7%, which is due to the decrease in the number of classification scenarios and the increase in the sample number of a single type of scene, indicating that the proposed method has certain robustness when the sample number is increased or the number of scenes is reduced. By calculating the Kappa coefficient of 0.975, the classification results are also highly consistent.

Compared with the methods based on the SIFT and the HOG features and other algorithms, the results are shown in Tab. 2, indicating that our method can be effectively adapted to the classification of remote sensing image scenes.

IV. CONCLUSION AND FUTURE WORK

This paper adopts the method of transfer learning, combines SURF features and deep learning semantic features, and proposes a feasible method for dealing with remote sensing image scene classification. It solves the problem that the traditional artificial feature classification performance is not strong and the deep learning extraction semantic feature is complex. It provides some reference and inspiration for the research of remote sensing image processing combined with traditional methods and deep learning. The experimental analysis shows that traditional artificial features and deep learning semantic features can provide effective information complementation. The joint fusion strategy can enhance the feature expression ability. Combined with the random forest classifier, it can be effectively applied to the remote sensing image scene classification research. However, there are still many improvements in the method proposed in this paper. Firstly, the computational process of this method is relatively complicated, and real-time performance is not high. Subsequent research should improve the computational flow and achieve the parallelization of algorithms to improve computational efficiency. Secondly, due to the limited amount of data used for remote sensing image scene classification, this paper adopts the method of transfer learning. Subsequent research can deeply explore the small sample feature learning. Thirdly, the proposed method is more effective for the classification of simple remote sensing image scenes, but for a relatively complex scene environment, a small number of misjudgments occur, so the classification research work for complex scenes will also be the focus of follow-up research.

ACKNOWLEDGMENT

This research was supported by the National Nature Foundation of China (Grant No. 41761082, 61162016, 61562057), the Gansu Science and Technology Project (Grant No. 18JR3RA104, 1504FKCA038), the Science and Technology Project of Gansu Education Department (Grant No. 2017D-08), Lanzhou Jiaotong University Innovation and Entrepreneurship Course Construction Project in 2019 (2019CXCYKC10), Lanzhou Jiaotong University Experimental Teaching Reform Research Project (2019026) , Lanzhou Jiaotong University Experimental Teaching Reform Research Project (2019007) and Lanzhou Talent Innovation and Entrepreneurship Project(2015-RC-86). all support is gratefully acknowledged.

REFERENCES

[1] Y Tian, W Zhang, Q Zhang, G Lu, X Wu. Review on Image Scene Classification Technology[J]. Acta Electronica Sinica, 2019, 47(04):915-926.

[2] G Xia, J Hu, F Hu, et al. AID: A benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.

[3] M Li, S Zang, B Zhang, et al. A review of remote sensing image classification techniques: The role of spatio-contextual information[J]. European Journal of Remote Sensing, 2014, 47(1): 389-411.

[4] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[5] Everingham M, Zisserman A, Williams C K I, et al. The 2005 pascal visual object classes challenge[C]//Machine Learning Challenges Workshop. Springer, Berlin, Heidelberg, 2005: 117-176.

[6] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002 (7): 971-987.

[7] Y. Li and M. Liu, "Aerial Image Classification Using Color Coherence Vectors and Rotation & Uniform Invariant LBP Descriptors," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2018, pp. 653-656.

[8] Z.Ye and L.Bai. Hyperspectral Image Classification Based on Principal Component Analysis and Local Binary Patterns[J]. Laser & Optoelectronics Progress, 2017,54(11):139-148.

[9] S. Qi, J. Ma, J. Lin, Y. Li and J. Tian, "Unsupervised Ship Detection Based on Saliency and S-HOG Descriptor From Optical Satellite Images," in IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 7, pp. 1451-1455, July 2015.

[10] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//international Conference on computer vision & Pattern Recognition (CVPR'05). IEEE Computer Society, 2005, 1: 886--893.

[11] Q. Zhu, Y. Zhong, B. Zhao, G. Xia and L. Zhang, "Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery," in IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 6, pp. 747-751, June 2016.

[12] G Sheng, W Yang, T Xu, et al. High-resolution satellite scene classification using a sparse coding based multiple feature combination[J]. International journal of remote sensing, 2012, 33(8): 2395-2412.

[13] B Zhao, Y Zhong, L Zhang, et al. The Fisher kernel coding framework for high spatial resolution scene classification[J]. Remote Sensing, 2016, 8(2): 157.

[14] Q. Zou, L. Ni, T. Zhang and Q. Wang, "Deep Learning Based Feature Selection for Remote Sensing Scene Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 11, pp. 2321-2325, Nov. 2015.

[15] F Hu, G Xia, J Hu, et al. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery[J]. Remote Sensing, 2015, 7(11): 14680-14707.

[16] G Wang, B Fan, S Xiang, et al. Aggregating rich hierarchical features for scene classification in remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(9): 4104-4115.

[17] W Zhang, P Tang, L Zhao. Remote Sensing Image Scene Classification Using CNN-CapsNet[J]. Remote Sensing, 2019, 11(5): 494.

[18] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2006: 404-417.

[19] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]//CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2010: 3304-3311.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] X Wang, K Li, C Ning, et al. Remote Sensing Image Classification Method based on Deep Convolution Neural Network and Multi-kernel Learning[J]. Journal of Electronics & Information Technology, 2019, 41(05):1098-1105.