Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Scene classification of high-resolution remote sensing images based on IMFNet

Xin Zhang
Yongcheng Wang
Ning Zhang
Dongdong Xu
Bo Chen
Guangli Ben
Xue Wang

SPIE.

Xin Zhang, Yongcheng Wang, Ning Zhang, Dongdong Xu, Bo Chen, Guangli Ben, Xue Wang, "Scene classification of high-resolution remote sensing images based on IMFNet," *J. Appl. Remote Sens.* **13**(4), 048505 (2019), doi: 10.1117/1.JRS.13.048505.

Scene classification of high-resolution remote sensing images based on IMFNet

Xin Zhang,[a,b] Yongcheng Wang,[a,*] Ning Zhang,[a,b] Dongdong Xu,[a,b] Bo Chen,[a] Guangli Ben,[a,b] and Xue Wang[a]

[a]Chinese Academy of Sciences, Changchun Institute of Optics, Fine Mechanics and Physics, Changchun, China

[b]University of Chinese Academy of Sciences, College of Materials Science and Opto-Electronic Technology, Beijing, China

Abstract. Currently, due to the limited amount of data and the difficulty of designing a network, there are few papers on constructing a new convolutional neural network for scene classification using the publicly available datasets of high-resolution remote sensing images. Considering the existing problems, the current scene classification methods of high-resolution remote sensing images are summarized, and the IMFNet model is constructed to classify scenes of high-resolution remote sensing images in this paper. The IMFNet is an end-to-end network, which can learn features from data automatically. The main characteristic of the IMFNet network structure is that the Inception module is used to extract the details of remote sensing images and the multifeature fusion strategy is proposed to ensure the integrity of information. In addition, optimization methods are adopted to improve the classification accuracy. In order to verify the effectiveness of the method proposed in this paper, the two benchmark datasets—the UC Merced dataset and the SIRI-WHU dataset were adopted for experiments. The classification accuracy of the two datasets reaches 92.14% and 90.43%, respectively. Experimental results show that the method proposed has certain advantages over the classification methods based on low-level and middle-level visual features and even some classification methods based on high-level visual features. [DOI: 10.1117/1.JRS.13.048505]

Keywords: image processing; remote sensing; artificial intelligence; pattern recognition; scene classification.

Paper 190640 received Aug. 20, 2019; accepted for publication Nov. 22, 2019; published online Dec. 13, 2019.

1 Introduction

With the improvement of remote sensing data acquisition ability and the diversification of imaging methods, the amount and types of remote sensing data increase significantly. In the field of computer vision, image processing technology is also developing, and it is an inevitable trend to use intelligent and automatic technology to analyze remote sensing data.[1] The purpose of automatic scene classification of high-resolution remote sensing images is to classify images containing multiple land-cover or land-use types into different semantic categories. Scene classification of high-resolution remote sensing images is not only a key part of intelligent remote sensing processing but also has important research value.[2] It has a wide range of applications, such as vegetation types mapping,[3] geological disaster monitoring,[4] geospatial target detection,[5] land-use or land-cover determination,[6] geographic image retrieval,[7] and so on.

The classification methods of remote sensing images have developed greatly in the past decade. According to the characteristics of the different scene classification methods, it can be divided into three categories: methods based on low-level visual features, methods based on middle-level visual features, and methods based on high-level visual features.

The scene classification methods of high-resolution remote sensing images are generally based on low-level visual features at an early stage. The low-level visual features are mainly

*Address all correspondence to Yongcheng Wang, E-mail: wangyc@ciomp.ac.cn

1931-3195/2019/$28.00 © 2019 SPIE

the color,[8] shape,[9] texture,[10] and other information of remote sensing images that can completely and objectively reflect the scene content extracted by technicians with a large amount of professional knowledge and engineering experience. The low-level visual features are extracted manually, and the selection of features by human participation is subjective to a certain extent. Moreover, there is a semantic gap between the semantic categories of scenes and the low-level visual features. Therefore, when the scenes of remote sensing images become complex, the scene classification methods based on low-level visual features are often not very effective.

The middle-level visual features are abstractions of the low-level visual features through quantization, coding, and other methods. Methods based on middle-level visual features can be roughly divided into three categories: methods based on semantic objects,[11] methods based on semantic attributes,[12] and methods based on local semantic concept.[13] The study of remote sensing scene classification methods based on middle-level visual features has been a hotspot in the past 10 years and achieved good results. However, with the rise of deep learning, the research based on middle-level visual features has gradually decreased.

The scene classification methods based on high-level visual features refer to the using of deep learning methods, which can extract semantic information of images. These methods based on deep learning can automatically extract the complex structure of high-dimensional data, which have higher classification performance than the methods based on low-level and middle-level visual features. According to whether labels are used or not, the scene classification methods of high-resolution remote sensing images based on deep learning can be roughly divided into two categories: unsupervised classification methods and supervised classification methods. The unsupervised classification methods adopt unlabeled data for classification, such as deep belief network,[14] deep Boltzmann machine,[15] stacked autoencoder,[16] and so on. The supervised classification methods adopt labeled data for classification, such as multilayer perceptron[17] and convolutional neural network (CNN).[18] Among them, the classification method based on CNN has many advantages in image processing. It can solve the invariance of translation, rotation, and scaling of feature images in a certain space. Moreover, the features of CNN are learned from data through training, which avoids the process of artificial feature extraction. In addition, the weights of neurons on the same feature map are the same, which not only plays a parallel role in the self-learning of the network but also reduces the computational complexity of the whole neural network. So, the method based on CNN is adopted for the scene classification of high-resolution remote sensing images in this paper.

Remote sensing images contain more complex information about the arrangement of ground objects than natural images, so it is challenging to classify scenes of high-resolution remote sensing images. In addition, the number of high-resolution remote sensing images available is limited, so most methods based on CNN adopt the pretrained models for transfer learning. However, there are relatively few papers on constructing a new CNN model based on the publicly available datasets of high-resolution optical remote sensing images, and there is a lack of relevant summary of the current research status in this filed, which limits the further development of deep learning in remote sensing image processing.

Based on the existing problems in scene classification of high-resolution remote sensing images, the main contributions of this paper are as follows:

1. This paper summarizes and classifies the current scene classification methods of high-resolution remote sensing images based on CNN, which can promote the development of automatic scene classification on remote sensing images by presenting the current development status of this field.
2. A scene classification method of high-resolution remote sensing images based on IMFNet is proposed. The IMFNet is an end-to-end network based on high-level visual features, which can automatically extract features from data.
3. The Inception module is used and the multifeature fusion strategy is proposed in the IMFNet. In addition, in order to further improve the generalization ability of the IMFNet, the optimization methods of data augmentation, dropout, parameter norm penalty, moving average model, and Adam optimization algorithm are used.
4. The IMFNet was trained and tested on the publicly available datasets of high-resolution remote sensing images—the UC Merced dataset and the SIRI-WHU dataset and obtained satisfactory classification results.

The remainder of this paper is organized as follows: the related work is presented in Sec. 2; the scene classification method of high-resolution remote sensing images based on IMFNet is presented in Sec. 3; Sec. 4 presents the experiments and results; Sec. 5 presents the parameter analysis; and Sec. 6 presents the conclusions.

2 Related Work

The CNN was put forward by Fukushima[19] and revised by LeCun et al.[20] It is a special artificial neural network, whose training process is divided into forward propagation and backpropagation. In the process of forward propagation, information is generally transmitted from the input layer to the output layer through layer-by-layer transformation of the convolutional layer, the pooling layer, and the fully connected layer. Finally, the output layer of CNN formalizes its target task as a loss function. In the process of backpropagation, the loss between the real value and the predicted value is calculated, and it is fed back by the backpropagation algorithm layer-by-layer, so as to update the parameters of each layer. Since other articles have introduced the specific structure of CNN in detail,[21] it is not covered in this article. In recent years, with the improvement of computer performance and the increasing amount of data available, many classic CNN models based on natural images have been put forward one after another, such as LeNet,[22] AlexNet,[23] VGGNet,[24] GoogleNet,[25] ResNet,[26] DenseNet,[27] etc. And it provides new development prospects for the field of computer vision, such as image classification,[28] data reconstruction,[29] image semantic segmentation,[30] image retrieval,[31] object detection,[32] image style transfer,[33] and so on.

As shown in Fig. 1, the scene classification method of high-resolution remote sensing images based on CNN can be roughly divided into two categories: (1) the classification method based on the full-trained CNN and (2) the classification method based on transfer learning. Details are covered in the following sections.

2.1 *Classification Method Based on the Full-Trained CNN*

The core idea of the classification method based on the full-trained CNN is using the publicly available datasets of high-resolution remote sensing images to fully train a CNN model, as shown in Fig. 2. This method can be subdivided into two strategies: (a) the classification method

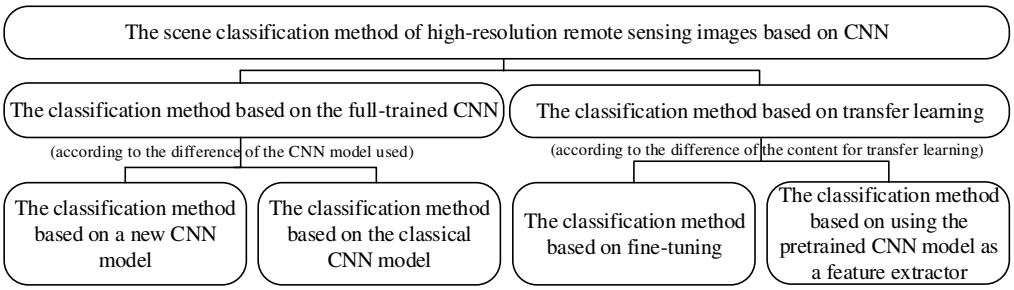


Fig. 1 The schematic of the scene classification method of high-resolution remote sensing images based on CNN.

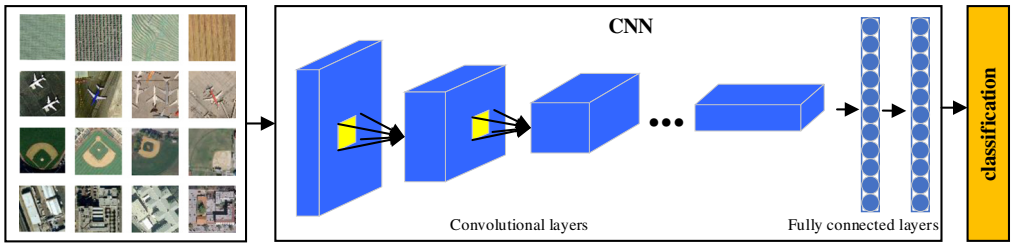


Fig. 2 The schematic of the classification method based on full-trained CNN.

based on a new CNN model and (b) the classification method based on the classical CNN model. The two strategies adopt the target dataset for training, and the difference between them lies in the difference of the CNN model used.

2.1.1 *Classification method based on a new CNN model*

Training a new CNN model requires strong computing equipment, long training time, and rich experience in network architecture design. Due to the limited amount of data in the publicly available datasets of high-resolution remote sensing images, the new CNN model based on the existing datasets is generally designed in a shallow way, which cannot give full play to the performance of the deep CNN model. However, in order to further develop the method of deep learning in remote sensing scene classification, it is necessary to design a new structure of CNN using the publicly available datasets of remote sensing images. Because of the difficulty of training a new CNN model, there are few articles about this method at present.[34–37]

2.1.2 *Classification method based on the classical CNN model*

The target dataset is used to train the weights of the existing classical CNN model without changing the structure and parameters in this method. Due to the limited amount of data in the publicly available datasets of high-resolution remote sensing images, it is impossible to compare with the ImageNet dataset containing tens of thousands of natural images. What is more, the depth of the classical CNN model is usually deep. Therefore, there will be a problem of mismatch been the amount of data and the depth of the model in the full-trained method based on the classical CNN model. The classification method based on the classical CNN model is generally used to compare with other scene classification methods of high-resolution remote sensing images in the existing articles.[38,39]

2.2 *Classification Method Based on Transfer Learning*

Transfer learning is a method of machine learning that applies the existing knowledge to different but related fields, so as to facilitate the learning of knowledge in new fields, as shown in Fig. 3. In deep learning, transfer learning can solve the problem of limited labeled data in the target field. It is worth noting that there must be some similarity between the target dataset and the original dataset when transfer learning is used. This method can also be subdivided into two strategies: (a) the classification method based on fine-tuning and (b) the classification method based on using the pretrained CNN model as a feature extractor. The two strategies use the target data

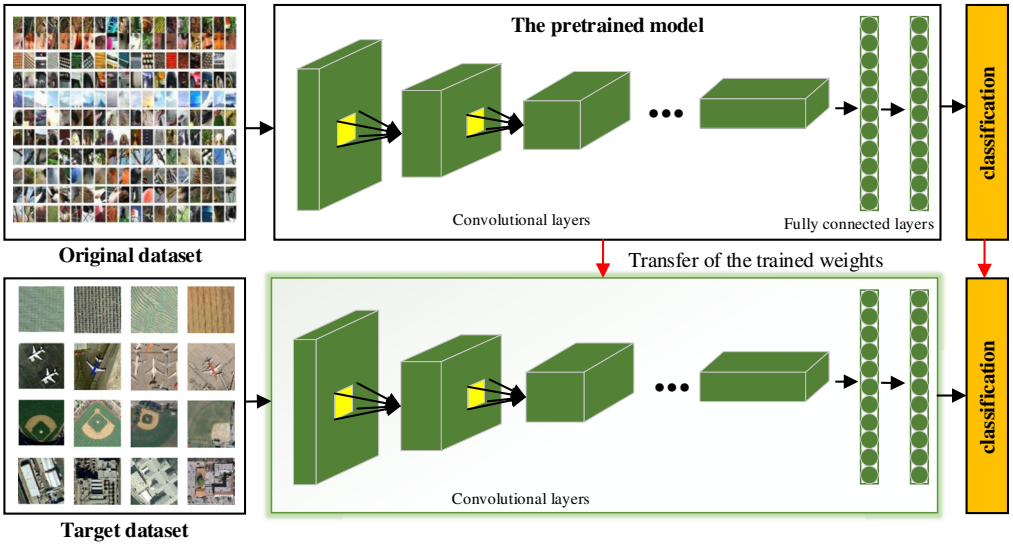


Fig. 3 The schematic of the classification method based on transfer learning.

to adjust the parameters of the classical pretrained CNN model obtained by original large-scale datasets. And the two strategies differ in two respects: (1) The classification method based on fine-tuning needs to finely tune the trained weights of the pretrained model while the classification method based on using the pretrained model as the feature extractor does not need the fine-tuning operation. (2) The classification method based on fine-tuning does not change the category of the classifier but the number of categories, while the type of classifier used in the other strategy changes.

2.2.1 *Classification method based on fine-tuning*

Fine-tuning is proposed according to the theory that the generalized features extracted from the shallow convolutional layers are universal for various tasks, while the deep convolutional layers extract more abstract features than the shallow ones. The classification method based on fine-tuning adjusts the weights of the pretrained model in order to improve the classification accuracy. And it can be roughly divided into two categories: (1) all the layers of the classical pretrained CNN model are fine-tuned and (2) only the top layers of the classical pretrained CNN model are fine-tuned. Currently, the transfer learning method based on fine-tuning has been widely used in the field of remote sensing image scene classification.[40–43]

2.2.2 *Classification method based on using the pretrained CNN model as a feature extractor*

The classification method using the pretrained model as a feature extractor does not need to redesign a new CNN or adjust the structure of the network, but only needs to choose which pretrained model to use, which layer of the network to extract features and which machine learning classification algorithm to select for classification.[42–45] In order to obtain more accurate feature information and higher classification accuracy, many works of literature use this method to extract high-level visual features and conduct feature fusion with low-level, middle-level, or other high-level visual features.[46–50] These fusion methods generally have high classification results due to the integration of multiple feature types rather than just the use of high-level visual features.

3 Scene Classification Method of High-Resolution Remote Sensing Images Based on IMFNet

IMFNet model for scene classification of high-resolution remote sensing images is proposed in this paper. This chapter is divided into two parts. The structure of IMFNet is introduced in Sec. 3.1. The optimization methods to improve the classification accuracy of IMFNet model are described in Sec. 3.2.

3.1 *Structure of IMFNet*

IMFNet is a CNN based on the Inception module and multifeature fusion strategy, and its structure is shown in Fig. 4. The IMFNet consists of four convolutional layers, six maximum pooling layers, two Inception modules, three fully connection layers, and one output layer. The input of the IMFNet model is $256 \times 256 \times 3$ pixels. And in the lower layers of the IMFNet, the convolutional layers and the maximum pooling layers are used alternately to extract the shallow features. The first three convolutional layers of IMFNet model adopt the convolutional kernel of 5×5 to extract the larger features, and the fourth convolutional layer adopts the convolutional kernel of 3×3 to extract more refined features. The filter size of each maximum pooling layer is 2×2 and the stride is 2. After that, two inception modules are used to extract high-level visual features. In addition, the multifeature fusion strategy is adopted to ensure the integrity of information to a certain extent. Finally, a softmax classifier is used in the output layer of the IMFNet. The output size depends on the number of remote sensing scene categories to be classified. Next, the Inception module and multifeature fusion strategy are introduced in detail, respectively.

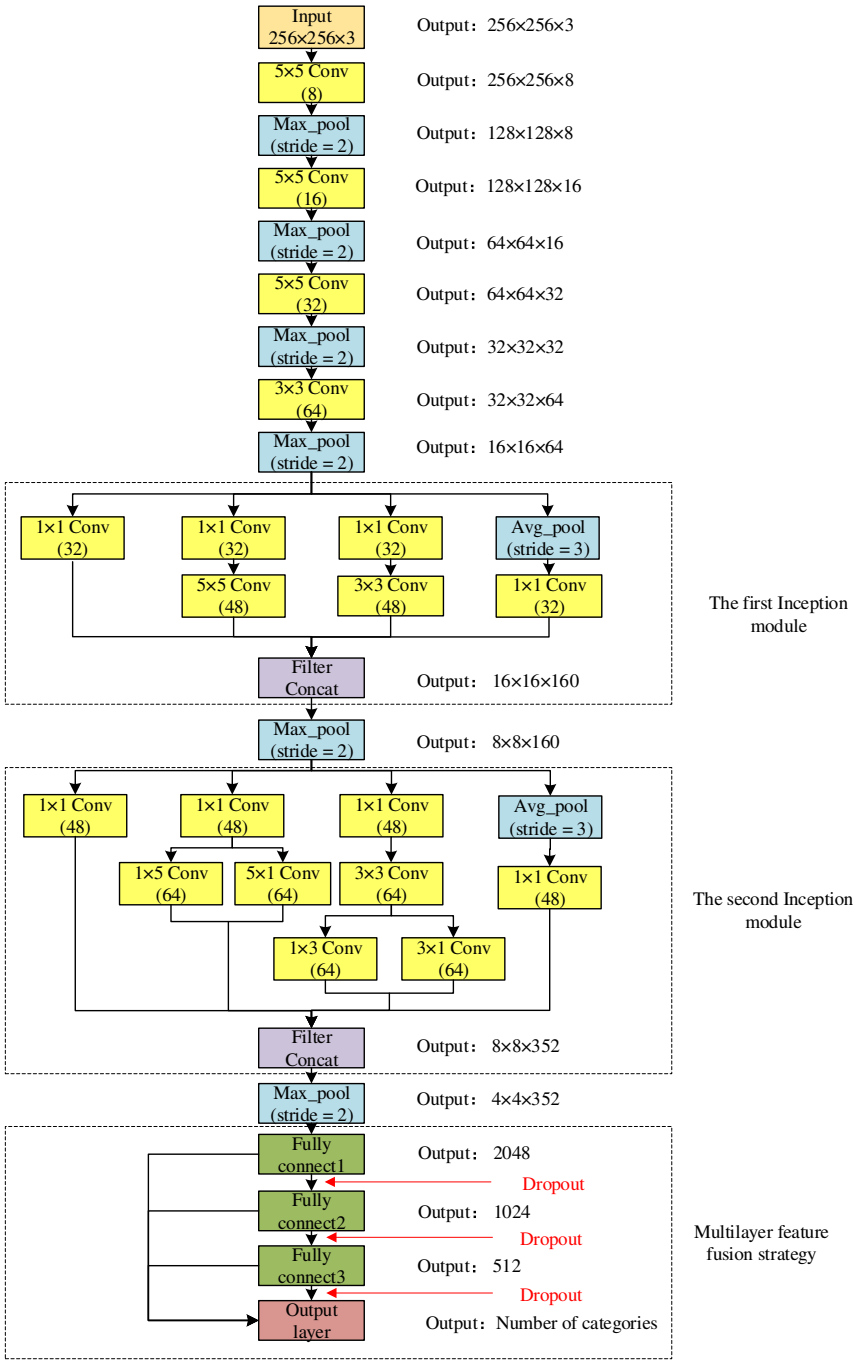


Fig. 4 Structure diagram of IMFNet.

3.1.1 *Inception module*

The Inception module is a combination of different perception fields, which can learn both microfeatures and macrofeatures simultaneously.[25,51–53] The first Inception module used in this article is an improved module of the Inception V1.[25] The second Inception module is an improved module of the Inception V3.[52]

The first Inception module used in the article is shown in Fig. 5. It has four batches, the first branch includes a convolutional layer with a convolution kernel size of 1×1; the second branch is composed of two convolutional layers with convolution kernel sizes of 1×1 and 5×5; the third branch is composed of two convolutional layers with convolution kernel sizes of 1×1 and 3×3; the fourth branch is composed of a pooling layer with a kernel size of 3×3 and a

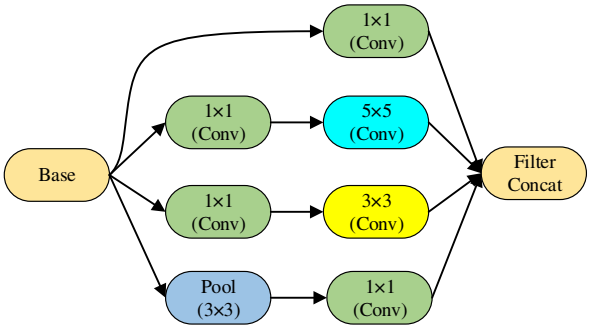


Fig. 5 The first Inception module.

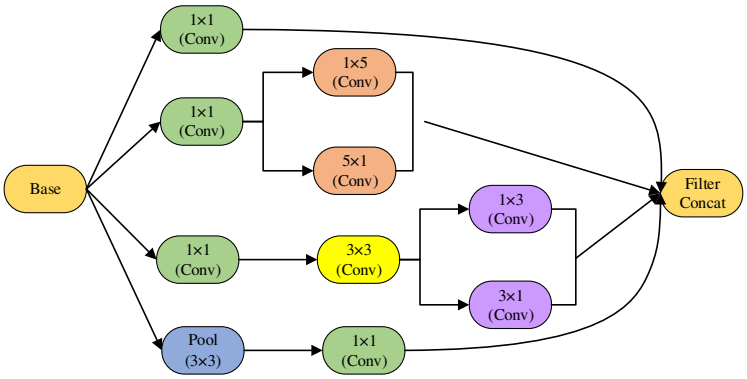


Fig. 6 The second Inception module.

convolutional layer with a convolution kernel size of 1×1. The concatenation of convolution kernels of different sizes is equivalent to the fusion of features with different scales so that the learning ability of the model is improved. The convolution kernel of 1×1 is used for dimensionality reduction, so as to reduce the calculation cost.

The second Inception module used in the article is shown in Fig. 6. Its main characteristic is to decompose a large two-dimensional convolution into two small one-dimensional convolutions, so as to save a large number of parameters, speed up the operation, and alleviate the problem of overfitting. The second Inception module also contains four branches. The first branch includes a convolutional layer with a convolution kernel size of 1×1; the second branch includes a convolutional layer with a convolution kernel size of 1×1 and two parallel convolutional layers with convolution kernel sizes of 1×5 and 5×1. The third branch is composed of a convolutional layer with a convolution kernel size of 1×1, another convolutional layer with a convolution kernel size of 3×3, and two parallel convolutional layers with convolution kernel sizes of 1×3 and 3×1. The fourth branch is composed of an average pooling layer with a kernel size of 3×3 and a convolutional layer with a convolution kernel size of 1×1.

3.1.2 *Multifeature fusion strategy*

CNN has the characteristics of hierarchy, and the features obtained from each layer have gradual transition from the generalization features, such as edge and texture to the high-level semantic representation. Based on this characteristic, the multifeature fusion strategy based on the fully connected layers was proposed in the article, as shown in Fig. 7. The output features of the fully connected layers containing high-level semantic information are cascaded, and then the cascaded features are classified through the output layer, so that the semantic information contained in the features of different layers can complement each other, thus improving the classification accuracy.

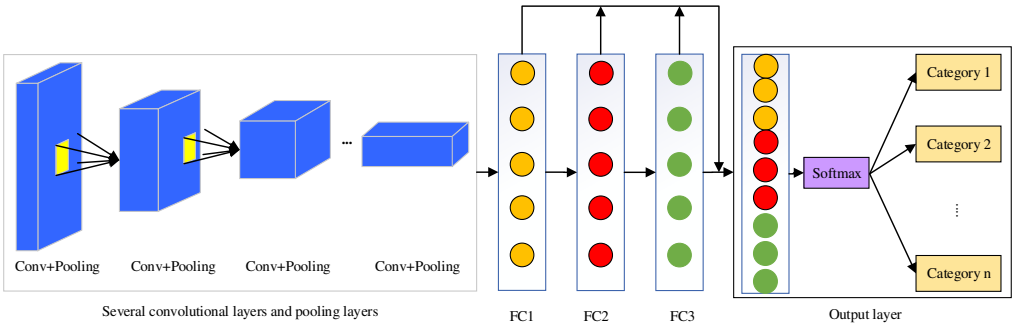


Fig. 7 Schematic diagram of the multifeature fusion strategy.

3.2 *Optimization Methods*

In the training process of CNN, the problem of overfitting occurs easily when the amount of data is too little and the parameters of the model are too vast. The so-called overfitting refers to the phenomenon that the model overlearns the training data during the training process so that it performs well in the training set but poorly in the test set. Therefore, the optimization methods of data augmentation, dropout, parameter norm penalty, moving average model, and Adam optimization algorithm are adopted to prevent the problem of overfitting and make the IMFNet constructed in the paper more robust.

3.2.1 *Data augmentation*

Data augmentation is a method that increases the amount of data in a dataset while keeping the data label unchanged. In this paper, random cropping, scale stretching, rotation, and image standardization are adopted to enlarge the dataset, so as to obtain sufficient training data of high-resolution remote sensing images, improve the generalization ability and robustness of the IMFNet model constructed, and effectively alleviate the problem of overfitting.

3.2.2 *Dropout*

Dropout can significantly reduce overfitting phenomenon and improve network generalization performance.[54] The basic idea of dropout is to randomly make some neurons in the hidden layer temporarily stop working in a certain proportion during network training. In other words, the weights of some neurons in the hidden layer are not updated but retained according to a certain probability, which is equivalent to training on different networks and reducing the complex coadaptive relationship between neurons. Figure 8 shows the comparison chart of the standard neural network and the neural network using dropout.

3.2.3 *Parameter norm penalty*

The main idea of parameter norm penalty regularization is to add a parameter norm penalty $\Omega(\theta)$ to the target function J to obtain the regularized target function $\tilde{J}$, as shown in Eq. (1):

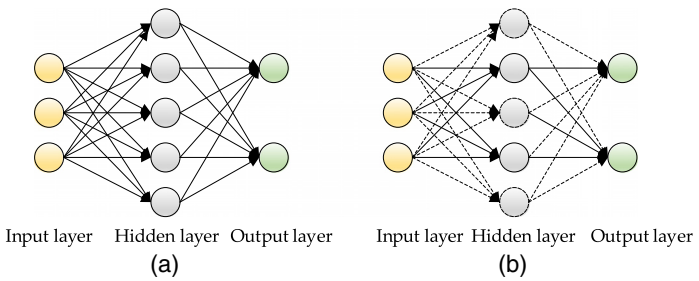


Fig. 8 Schematic diagram of dropout: (a) the standard neural network and (b) the neural network using dropout.

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta), \tag{1}$$

where X is the input data; y is the label; θ is the parameter; $\alpha \in [0, \infty)$ is the hyperparameter used to weigh the relative contribution of the parameter norm penalty $\Omega(\theta)$, the larger the value is, the greater the intensity of the regularization is. This method controls the complexity of the model by adding parameter norm penalties to the target function, thus avoiding the overfitting problem of the neural network. The common regularization methods of parameter norm penalties include L1-norm regularization and L2-norm regularization.

The L1-norm represents the sum of the absolute values of all the elements in the vector, as shown in Eq. (2):

$$\Omega(\theta) = \|W\|_1 = \sum_i |W_i|, \tag{2}$$

L2-norm represents the sum of squares of each parameter in the vector, as shown in Eq. (3):

$$\Omega(\theta) = \frac{1}{2}\|W\|_2^2 = \sum_i W_i^2. \tag{3}$$

In this paper, the L2-norm regularization is used.

3.2.4 *Moving average model*

The main principle of the moving average model is to control the gap before and after the parameter update by controlling the decay rate, so as to reduce the variation of parameters. When training the model, it is very beneficial to keep the moving average of the training parameters. The use of averaged parameters in testing sometimes produces much better results than the use of the final trained parameter values, which can make the model more robust in test data.

3.2.5 *Adam optimization algorithm*

Adam optimization algorithm is a first-order gradient-based optimization algorithm proposed by Kingma and Ba,[55] which is based on adaptive moment estimation and can replace the traditional stochastic gradient descent (SGD). SGD maintains a single learning rate to update all weights, and the learning rate does not change during the training process, whereas Adam designs an independent adaptive learning rate for different parameters by calculating the first-moment estimation and second-moment estimation of gradient.

4 Experiments and Results

This chapter mainly introduces the experimental process and result analysis. In Sec. 4.1, the data selection is introduced; in Sec. 4.2, the experimental setup is described; and in Sec. 4.3, the results and analysis are presented.

4.1 *Dataset Selection*

In recent years, many scholars have made great efforts to construct the datasets of high-resolution remote sensing images. The publicly available datasets of high-resolution remote sensing images are presented in Table 1.

In order to prove the effectiveness of the proposed method, the UC Merced dataset and the SIRI-WHU dataset were used for training and testing in the paper. The two datasets are described in detail below.

Table 1 The publicly available datasets of high-resolution remote sensing images.

Dataset name	Total class	Number per class	Total images	Image size	Bands	Spatial resolution (m)	Sources
Brazilian Coffee Scene[49]	3	1438–36,577	50,000	64×64	R; G; near-infrared	—	SPOT sensor
UC Merced[56]	21	100	2100	256×256	RGB	0.3	United States Geological Survey
WHU-RS19[57]	19	50–61	1005	600×600	RGB	~0.5	Google Earth
Banja-Luka[58]	6	28–178	606	128×128	RGB	—	Aerial images
RSSCN7[14]	7	400	2800	400×400	RGB	—	Google Earth
SAT-4[59]	4	125,000	500,000	28×28	R; G; B; near-infrared	1	National Agricultural Imagery Program
SAT-6[59]	6	67,500	405,000	28×28	R; G; B; near-infrared	1	National Agricultural Imagery Program
SIRI-WHU[60]	12	200	2400	200×200	RGB	2	Google Earth
RSC11[6]	11	~100	1232	512×512	RGB	0.2	Google Earth
AID[61]	30	220–420	10,000	600×600	RGB	0.22–3	Google Earth
NWPU-RESISC45[43]	45	700	31,500	256×256	RGB	0.2–30	Google Earth
PatternNet[62]	38	800	30,400	256×256	RGB	0.062–4.693	Google Earth

4.1.1 *UC Merced dataset*

The UC Merced dataset[56] contains 21 types of land-use categories, and each category contains 100 images. Each image is 256×256 pixels with a pixel resolution of 1 foot. The dataset was manually extracted from the United States Geological Survey (USGS) National Map of US urban areas. The typical images of the UC Merced dataset are shown in Fig. 9.

4.1.2 *SIRI-WHU dataset*

The SIRI-WHU dataset[60] contains 12 scene categories, each category contains 200 images. Each image is 200×200 pixels with a spatial resolution of 2 m. The SIRI-WHU dataset is derived from Google Earth and mainly includes urban areas in China. The typical images of the SIRI-WHU dataset are shown in Fig. 10.

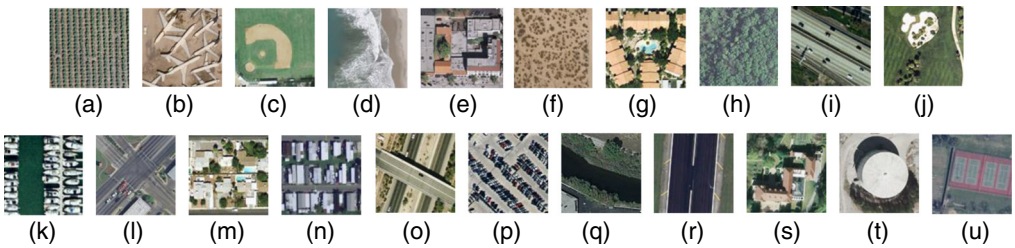


Fig. 9 Category representatives of the UC Merced dataset: (a) agricultural, (b) airplane, (c) baseball diamond, (d) beach, (e) buildings, (f) chaparral, (g) dense residential, (h) forest, (i) freeway, (j) golf course, (k) harbor, (l) intersection, (m) medium residential, (n) mobile home park, (o) overpass, (p) parking lot, (q) river, (r) runway, (s) sparse residential, (t) storage tanks, and (u) tennis court.

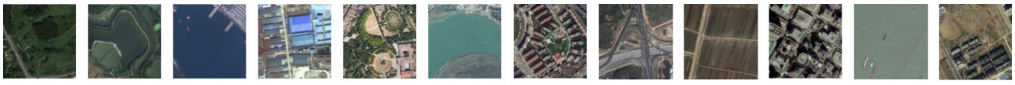


Fig. 10 Category representatives of the SIRI-WHU dataset: (a) meadow, (b) pond, (c) harbor, (d) industrial, (e) park, (f) river, (g) residential, (h) overpass, (i) agriculture, (j) commercial, (k) river, and (l) idle land.

The reasons for choosing the UC Merced dataset and the SIRI-WHU dataset are as follows:

1. These two datasets are universal datasets, on which many scholars have conducted a large number of experiments, so as to facilitate the comparison between the methods proposed in this paper and proposed by others.
2. There is a relatively small number of images in both datasets. If the method proposed in this paper can achieve good results, it can be proved that the method is effective for the case of limited data.
3. Among the datasets with a limited amount of data, the two datasets have a relatively large number of scene categories, and each category has a high degree of overlap. In order to verify the effectiveness of the proposed method, it is persuasive to select datasets with high intraclass diversity and low interclass distance.
4. The two datasets have different data sources and cover different regions. The UC Merced dataset, obtained by the USGS, covers images of different cities in the United States; the SIRI-WHU dataset is obtained through Google Earth, covering urban areas in China.
5. The image resolutions of the two datasets are different. The image resolution of the UC Merced dataset is 0.3 m and that of the SIRI-WHU dataset is 2 m. If the proposed method can achieve satisfactory results under the datasets with different resolutions, the validity of the proposed method can be further demonstrated.

4.2 *Experimental Setup*

In the experiment, 80% of high-resolution remote sensing images in the dataset were randomly selected as the training set and the remaining 20% as the test set. What's more, the hyperparameter used to weigh the relative contribution of the parameter norm penalty of L2-norm regularization was set to 0.0001, the decay rate of moving average model was set to 0.9999, the batch size was set to 64, and the number of iterations was set to 100,000. The experimental hardware platform was based on Intel E5 2665 dual-core processor, four-channel GTX1080Ti GPU, and 32 Gb memory. The experimental software platform was based on Ubantu16.04 version, using CUDA 8.0.61, CUDNN v6, and TensorFlow1.4.0 environment.

4.3 *Results and Analysis*

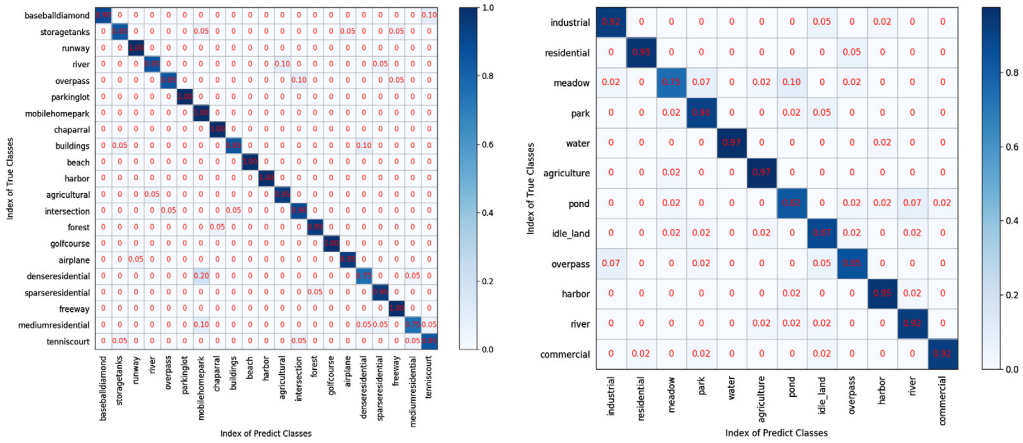
In order to verify the advantages of the proposed method, nine network structures were constructed for the corresponding comparison in this paper. The characteristics of different CNN models constructed are presented in Table 2, where "$\surd$" indicates that the strategy is adopted. For example, FNet represents the CNN that adopts multifeature fusion strategy and all optimization methods mentioned, but only uses one branch of the Inception module—convolutional layer with a convolution kernel size of 1×1 to replace the Inception module. It is worth noting that in the construction of A-IMFNet, the paper uses the traditional SGD to replace the Adam optimization algorithm.

4.3.1 *Analysis of the confusion matrix on the two datasets*

Figure 11 shows the confusion matrix of two datasets. It can be seen from Fig. 11(a) that the classification accuracy of a single category, such as runway, parking lot, mobile home park, chaparral, beach, harbor, golf course and freeway on the UC Merced dataset, has reached 100%. In addition, as shown in Fig. 11(b), the highest classification accuracy for a single category on the SIRI-WHU dataset is already 97%. The above has proved the high performance of the proposed method.

Table 2 The characteristics of different CNN models constructed.

Name	Inception module	Multifeature fusion strategy	Data augmentation	Dropout	Parameter norm penalty	Moving average model	Adam
FNet		√	√	√	√	√	√
IMNet	√		√	√	√	√	√
DA-IMFNet	√	√		√	√	√	√
D-IMFNet	√	√	√		√	√	√
P-IMFNet	√	√	√	√		√	√
M-IMFNet	√	√	√	√	√		√
A-IMFNet	√	√	√	√	√	√	
IMFNet	√	√	√	√	√	√	√



(a) (b)

Fig. 11 Confusion matrix of the two datasets: (a) confusion matrix of the UC Merced dataset and (b) confusion matrix of the SIRI-WHU dataset.

However, there are still categories with lower accuracy, such as dense residential and medium residential on the UC Merced dataset, meadow and pond on the SIRI-WHU dataset, all with an accuracy of 75%. As shown in Fig. 12, categories that are easily misclassified are shown. These categories have high similarity, which increases the difficulty of classification.

4.3.2 *Analysis of the classification accuracy on the two datasets*

In order to verify the advantages of the proposed method, it is compared with other classification methods based on low-level, middle-level, and high-level visual features. Table 3 presents the comparison accuracy of different methods on the UC Merced dataset and the SIRI-WHU dataset. It can be seen from Table 3 that the classification accuracy of the method based on IMFNet on the two datasets is the highest compared with the other methods mentioned. For classification methods based on low-level or middle-level visual features, such as bag of visual words (BoVW),[56] spatial pyramid co-occurrence kernel (SPCK),[63] spatial pyramid match (SPM),[64] randomized spatial partition-based classifier via boosting (BRSP),[65] saliency-guided unsupervised feature learning (SG + UFL),[66] bag of scale-invariant feature transform (Bag of SIFT),[67] Partlets[68] on the UC Merced dataset and latent dirichlet allocation (LDA),[69] SPM,[64] scale-invariant feature transform and bag of visual words (SIFT + BoVW),[70] Integration,[71] probabilistic latent semantic analysis (PLSA),[72] random forests (RF),[73] fisher kernel with the incorporation of the spatial

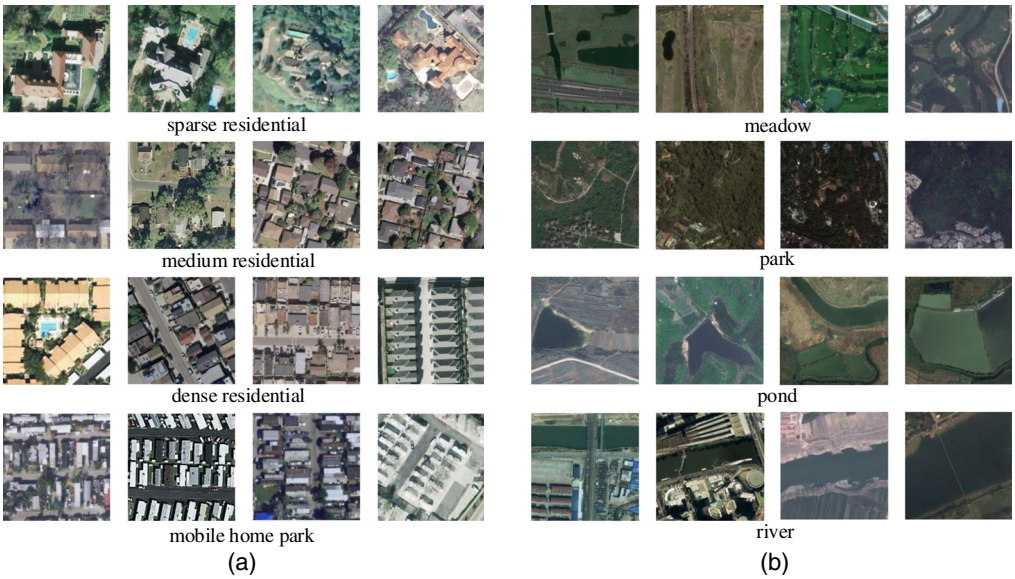


Fig. 12 Categories that can be easily misclassified: (a) categories that can be easily misclassified in the UC Merced dataset and (b) categories that can be easily misclassified in the SIRI-WHU dataset.

Table 3 Comparison accuracy of different methods on the two datasets.

	UC Merced dataset		SIRI-WHU dataset	
Attribute	Classification methods	Classification accuracy (%)	Classification methods	Classification accuracy (%)
Classification methods based on low-level and middle-level visual features	BoVW[56]	76.81	LDA[69]	60.32 ± 1.20
	SPCK[63]	73.14	SPM[64]	77.69 ± 1.01
	SPM[64]	75.29	SIFT + BoVW[70]	75.63
	BRSP[65]	77.8	Integration [71]	88.64
	SG + UFL[66]	82.72 ± 1.18	PLSA[72]	89.60 ± 0.89
	Bag-of-SIFT[67]	85.37 ± 1.56	RF[73]	89.29
	Partlets[68]	88.76	FK-S[74]	90.40
Classification methods based on high-level visual features	LPCNN[34]	89.9	LPCNN[34]	89.88
	FNet	88.33	FNet	87.3
	IMNet	89.52	IMNet	88.67
	DA-IMFNet	75	DA-IMFNet	73.24
	D-IMFNet	88.33	D-IMFNet	86.33
	P-IMFNet	90.95	P-IMFNet	89.65
	M-IMFNet	87.13	M-IMFNet	88.67
	A-IMFNet	40.24	A-IMFNet	53.75
	IMFNet	92.14	IMFNet	90.43

information (FK-S)[74] on the SIRI-WHU dataset, the classification method based on IMFNet has the highest accuracy. This is because the classification method based on high-level visual features can simulate the human brain mechanism to learn features from the data automatically, thus avoiding the incomplete extraction of artificial features, so the accuracy is generally relatively high. Due to the limited amount of data in the publicly available datasets of high-resolution remote sensing images, the classification accuracy of CNN methods based on full training is generally lower than that of transfer learning using the pretrained models, which are trained by the large-scale datasets. Therefore, this paper only compares the classification methods using high-level visual features based on the full-trained CNN, such as large patch convolutional neural networks (LPCNN),[34] FNet, IMNet, DA-IMFNet, D-IMFNet, P-IMFNet, M-IMFNet, A-IMFNet, and IMFNet. And the classification accuracy of the method based on IMFNet is 92.14% and 90.43%, respectively, on the UC Merced dataset and the SIRI-WHU dataset, which is the highest among them. The reasons for the good results achieved by the scene classification of high-resolution remote sensing images based on IMFNet are as follows:

1. The Inception module is used in the classification method based on IMFNet, which enables the network to select appropriate learning parameters by itself using convolution kernel of different sizes for feature extraction. As can be seen from Table 3, the accuracy of the classification method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 3.81% and 3.13% higher than that based on FNet, which replace Inception module with one of its branches, respectively. Through the above data analysis, the validity of the Inception module for scene classification of high-resolution remote sensing images can be proved.
2. Multifeature fusion strategy is adopted in the classification method based on IMFNet to ensure the integrity of information by cascading the fully connected layer features. As can be seen from Table 3, the accuracy of the classification method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 2.62% and 1.76% higher than that based on IMNet, respectively. The above analysis can prove the effectiveness of multifeature fusion strategy.
3. The optimization method of data augmentation is adopted in the classification method based on IMFNet. By means of data augmentation, not only the dataset is amplified but also the interpolation operation in data augmentation process is equivalent to adding noise to the input image, thus improving the antinoise ability of the model. It can be seen from Table 3 that the accuracy of the classification method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 17.14% and 17.19% higher than that based on DA-IMFNet, respectively. The analysis above can prove that the optimization method of data augmentation is effective for scene classification of high-resolution remote sensing images.
4. Dropout optimization method is adopted in the classification method based on IMFNet to reduce the overfitting problem of the model by making some units of the hidden layer not to work at a certain probability. As can be seen from Table 3, the accuracy of the classification method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 3.81% and 4.1% higher than that based on D-IMFNet, respectively. The above analysis can prove the effectiveness of using dropout strategy.
5. The optimization method of parameter norm penalty is adopted in the classification method based on IMFNet to solve the problem of overfitting by adding penalty terms to the target function to control the complexity of the model. It can be seen from Table 3 that the accuracy of the method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 1.19% and 0.78% higher than that based on P-IMFNet, respectively. The above result analysis proves the effectiveness of using the parameter norm penalty.
6. The optimization method of the moving average model is adopted in the classification method based on IMFNet to make the model more robust in test data. As can be seen from Table 3, the accuracy of the classification method based on IMFNet on the UC Merced dataset and on the SIRI-WHU dataset is 5.01% and 1.67% higher than that based on M-IMFNet, respectively. The above result analysis can prove the effectiveness of using moving average model.

7. Adam optimization algorithm is adopted in the classification method based on IMFNet to adjust the learning rate adaptively. Adam optimization algorithm can update the weights of neural network iteratively based on the training data, which can not only accelerate the training speed of the model but also avoid falling into the local optima. It can be seen from Table 3 that the accuracy of the classification method based on IMFNet using Adam optimization algorithm on the UC Merced dataset and on the SIRI-WHU is 51.9% and 36.68% higher than that based on A-IMFNet using traditional SGD with a single learning rate, respectively. The above result analysis can prove the effectiveness of using Adam optimization algorithm.

5 Parameter Analysis

The dropout rate, learning rate, data augmentation factor, and training iterations have a great impact on the scene classification accuracy of high-resolution remote sensing images. Next, the influence of these four parameters on the results will be analyzed in detail.

5.1 *Dropout Rate*

Dropout rate p is a hyperparameter, which means the probability of neurons being discarded, and the optimal values of dropout rate are different for different networks and different application fields. The IMFNet model using dropout rates from 0.1 to 0.7 was trained on the UC Merced dataset and the SIRI-WHU dataset under the condition of keeping other parameters unchanged. The test results are shown in Fig. 13. It can be seen from the figure that the overall trend of classification accuracy corresponding to different dropout rates on the UC Merced dataset and the SIRI-WHU dataset is roughly the same. The classification accuracy of dropout from 0.1 to 0.6 on the UC Merced dataset and the SIRI-WHU dataset is generally on the rise. When the dropout probability is 0.6, the classification accuracy of the UC Merced dataset and the SIRI-WHU dataset is the highest. When the dropout probability is 0.7, the classification accuracy of the UC Merced dataset and the SIRI-WHU dataset slowly declines, which is due to the fact that too many neurons are discarded to make the network unable to learn the characteristics well.

5.2 *Learning Rate*

In order to analyze the influence of learning rates on classification accuracy, the IMFNet model with learning rates of 0.01, 0.001, 0.0001, 0.00001, and 0.000001 was trained and tested on the UC Merced dataset and the SIRI-WHU dataset. The classification accuracy of IMFNet under different learning rates is shown in Fig. 14. It can be seen from the figure that the accuracy trend of IMFNet model with different learning rates is basically the same on the UC Merced dataset and the SIRI-WHU dataset. When the learning rate is set to 0.01, the classification accuracy of the two datasets is <10%. This is because the learning rate is too large, which makes the gradient oscillate back and forth near the optimal value, unable to converge. When the learning rate

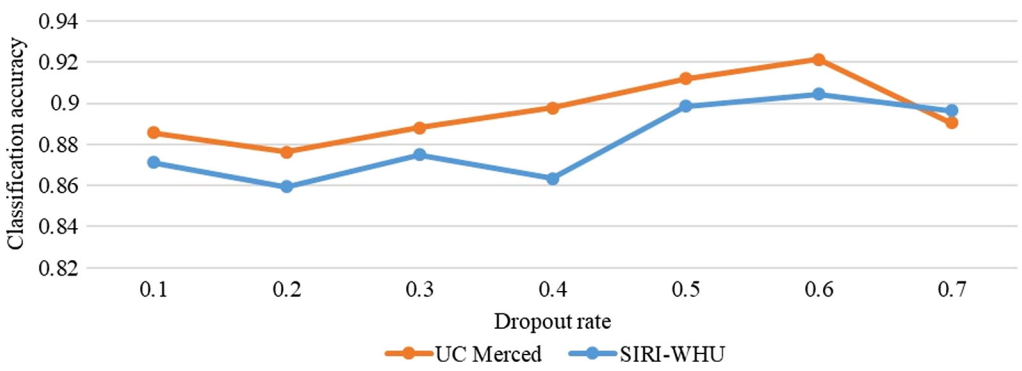


Fig. 13 Schematic diagram of the classification accuracy corresponding to different dropout rates.

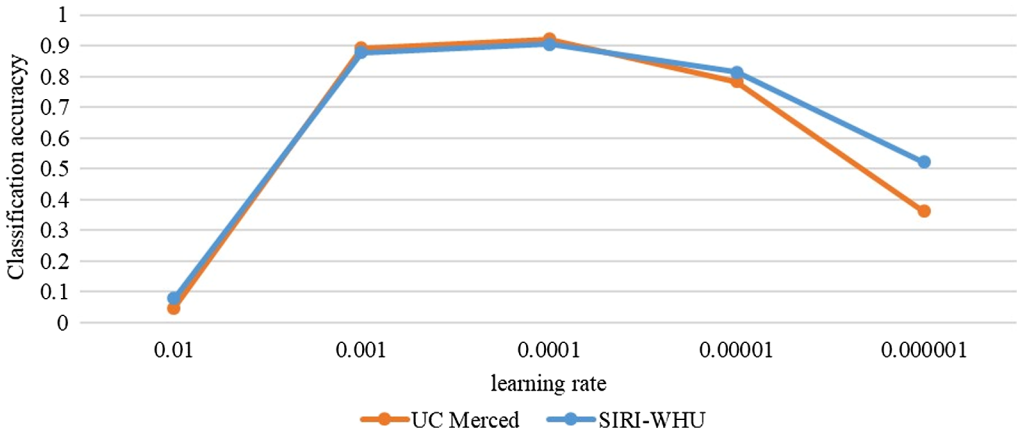


Fig. 14 Schematic diagram of the classification accuracy corresponding to different learning rates.

decreases to 0.001, the classification accuracy of the two datasets is improved. When the learning rate is 0.0001, the classification accuracy of the two datasets is the highest. However, when the learning rate is 0.00001 and 0.000001, the classification accuracy of the two datasets shows a decreasing trend. This is because when the learning rate is too small, the convergence process will become very slow, and the optimal classification accuracy cannot be achieved under the same number of iterations.

5.3 *Data Augmentation Factor*

The data augmentation factor represents the multiple of the original dataset after amplification. Theoretically, the more images in the dataset, the stronger the generalization ability of the network model. However, the classification accuracy depends partly on the quality of the picture. As shown in Fig. 15, for the UC Merced dataset, the accuracy is the highest when the data augmentation factor is 40. For the SIRI-WHU dataset, when the amplification factor is 80, the classification accuracy is the highest. Different from the object-centered natural image classification task, objects in high-resolution remote sensing images are often randomly distributed in the images. In the process of data augmentation, some useful information may be removed, which may lead to the fact that the classification accuracy decreases with the increase of amplification factor.

5.4 *Training Iterations*

The influence of different training iterations on the two datasets is shown in Fig. 16. The accuracy of the UC Merced dataset and the SIRI-WHU dataset shows a significant upward trend

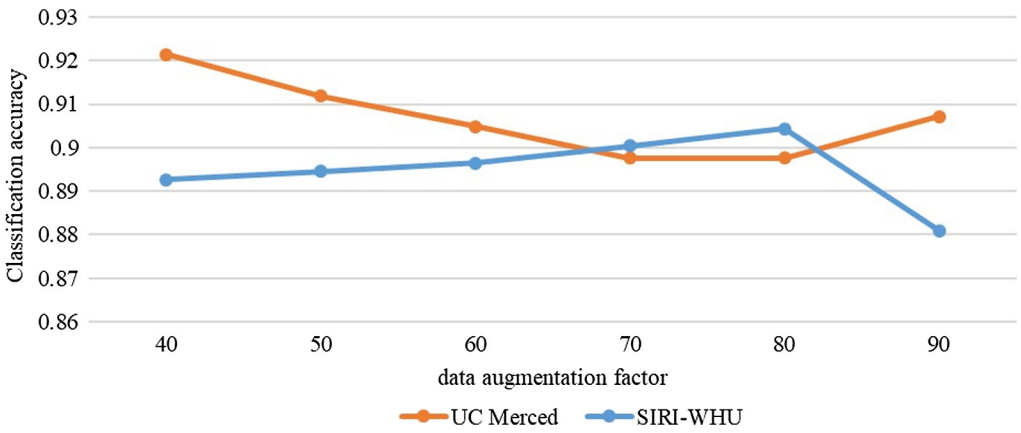


Fig. 15 Schematic diagram of the classification accuracy corresponding to different data augmentation factors.

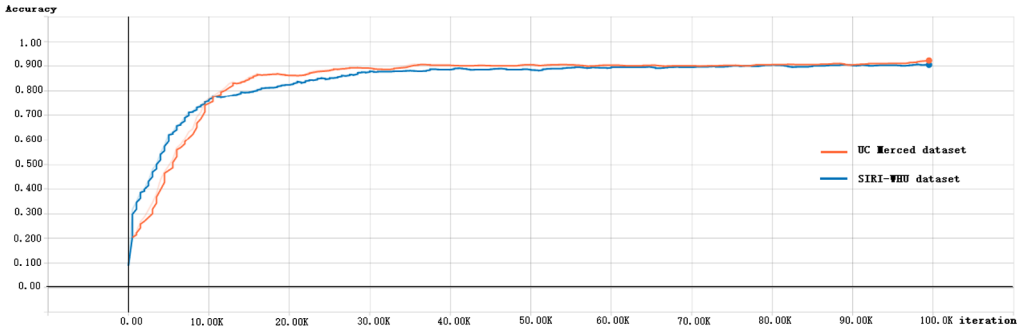


Fig. 16 Schematic diagram of the classification accuracy corresponding to different training iterations.

before 35,000 iterations and a slight upward trend overall after 35,000 iterations. This is because as the number of training iterations increases, the network can learn more features from images. At the same time, the test accuracy of the UC Merced dataset is higher than that on the SIRI-WHU dataset due to the difference of data characteristics in the datasets.

6 Conclusions

This paper not only summarizes the current scene classification methods of high-resolution remote sensing images but also presents a scene classification method of high-resolution remote sensing images based on IMFNet. IMFNet is an end-to-end CNN, which can learn high-level visual features from data automatically. The main characteristic of IMFNet is that Inception module is adopted to extract detail features of remote sensing images, and the multifeature fusion strategy is proposed to ensure the integrity of information. In order to further improve the classification accuracy of the IMFNet, the optimization methods of data augmentation, dropout, parameter norm penalty, moving average model, and Adam optimization algorithm are adopted to optimize the IMFNet.

In order to verify the effectiveness of the proposed method, the paper tested on the UC Merced dataset and the SIRI-WHU dataset, and the classification accuracy of the two publicly available datasets reaches 92.14% and 90.43%, respectively. Compared with the classification methods based on low-level, middle-level, and high-level visual features, the proposed method has certain advantages.

The method proposed in this paper still has some shortcomings. Since the method of supervised learning is adopted in this paper, the classification accuracy also depends on the accuracy of labeled data. When data augmentation is carried out, the operation of random clipping will lead to incomplete or even wrong information. In addition, most remote sensing data in practical engineering applications are unlabeled data. Therefore, an unsupervised scene classification method of high-resolution remote sensing images is planned to adopt in the follow-up research work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant no. 11703027. The authors declare no conflicts of interest.

References

1. L. I. Deren, L. Zhang, and G. Xia, "Automatic analysis and mining of remote sensing big data," *Acta Geod. Cartogr. Sin.* **43**, 1211–1216 (2014).
2. F. Hu et al., "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.* **7**(11), 14680–14707 (2015).
3. N. B. Mishra and K. A. Crews, "Mapping vegetation morphology types in a dry savanna ecosystem: integrating hierarchical object-based image analysis with random forest," *Int. J. Remote Sens.* **35**(3), 1175–1198 (2014).

4. G. Cheng et al., "Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA," *Int. J. Remote Sens.* **34**(1), 45–59 (2013).
5. C. Gong, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **54**(12), 7045–7415 (2016).
6. L. Zhao, T. Ping, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.* **10**(3), 035004 (2016).
7. Y. Li et al., "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sens.* **8**(9), 709 (2016).
8. T. Ojala et al., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987 (2002).
9. P. Mukhopadhyay and B. B. Chaudhuri, "A survey of Hough transform," *Pattern Recognit.* **48**(3), 993–1010 (2015).
10. A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using Gabor filters," *Pattern Recognit.* **30**(2), 295–309 (1997).
11. X. Luo and J. Xu, "Object-based representation for scene classification," in *Can. Conf. Artif. Intell.* (2016).
12. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vision* **42**(3), 145–175 (2001).
13. R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of bag-of-words and bag-of-topics models of EO image patches," *IEEE Geosci. Remote Sens. Lett.* **12**(6), 1357–1361 (2015).
14. Z. Qin et al., "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.* **12**(11), 2321–2325 (2015).
15. R. Salakhutdinov and G. Hinton, "An efficient learning procedure for deep Boltzmann machines," *Neural Comput.* **24**(8), 1967–2006 (2012).
16. X. Han et al., "Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery," *Int. J. Remote Sens.* **38**(2), 514–536 (2017).
17. S. Folving, "Forest mapping from multi-source satellite data using neural network classifiers —an experiment in Portugal," *Remote Sens. Rev.* **12**(1–2), 83–106 (1995).
18. X. Han et al., "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.* **9**(8), 848 (2017).
19. K. Fukushima, "Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.* **36**(4), 193–202 (1980).
20. Y. Lecun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.* **1**(4), 541–551 (1989).
21. W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: a comprehensive review," *Neural Comput.* **29**, 2352–2449 (2017).
22. Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Inf. Process. Syst.*, Vol. **141**, no. 5, pp. 1097–1105 (2012).
24. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.* (2015).
25. C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1–9 (2015).
26. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
27. G. Huang et al., "Densely connected convolutional networks," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2261–2269 (2017).
28. S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *3rd IAPR Asian Conf. Pattern Recognit.* (2016).
29. Q. Zhang et al., "Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* **56**(8), 4274–4288 (2018).

30. J. Dolz et al., "HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imaging* **38**(5), 1116–1126 (2019).
31. A. Qayyum et al., "Medical image retrieval using deep convolutional neural network," *Neurocomputing* **266**, 8–20 (2017).
32. Y. Hu et al., "A sample update-based convolutional neural network framework for object detection in large-area remote sensing images," *IEEE Geosci. Remote Sens. Lett.* **16**(6), 947–951 (2019).
33. L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2414–2423 (2016).
34. Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.* **10**(2), 025006 (2016).
35. F. P. S. Luus et al., "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.* **12**(12), 2448–2452 (2015).
36. F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.* **54**(3), 1793–1802 (2016).
37. Y. Liu, F. Fei, and Q. Zhu, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.* **10**, 444 (2018).
38. M. Castelluccio et al., "Land use classification in remote sensing images by convolutional neural networks," *Acta Ecol. Sin.* **28**(2), 627–635 (2015).
39. W. Han et al., "Adaptive spatial-scale-aware deep convolutional neural network for high-resolution remote sensing imagery scene classification," in *Int. Geosci. and Remote Sens. Symp.*, pp. 4736–4739 (2018).
40. Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* **56**(12), 7109–7121 (2018).
41. X. Liu et al., "Classifying high resolution remote sensing images by fine-tuned VGG deep networks," in *Int. Geosci. and Remote Sens. Symp.* (2018).
42. G. Cheng et al., "Scene classification of high resolution remote sensing images using convolutional neural networks," in *Int. Geosci. and Remote Sens. Symp.*, pp. 767–770 (2016).
43. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: benchmark and state of the art," *Proc. IEEE* **105**(10), 1865–1883 (2017).
44. X. Gong et al., "Deep salient feature based anti-noise transfer network for scene classification of remote sensing imagery," *Remote Sens.* **10**(3), 410 (2018).
45. K. Chatfield et al., "Return of the devil in the details: delving deep into convolutional nets," in *Br. Mach. Vision Conf.* (2014).
46. X. Gong, L. Yuanyuan, and Z. Xie, "An improved bag-of-visual-word based classification method for high-resolution remote sensing scene," in *26th Int. Conf. Geoinf.*, pp. 1–5 (2018).
47. B. Zhou et al., "Learning deep features for scene recognition using places database," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, pp. 487–495 (2014).
48. D. Marmanis et al., "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.* **13**(1), 105–109 (2016).
49. O. A. B. Penatti, K. Nogueira, and J. A. D. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains," in *IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 44–51 (2015).
50. Q. Zhu et al., "A deep-local-global feature fusion framework for high spatial resolution imagery scene classification," *Remote Sens.* **10**(4), 568 (2018).
51. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn.*, pp. 448–456 (2015).
52. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2818–2826 (2016).
53. C. Szegedy et al., "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. Thirty-First AAAI Conf. Artif. Intell.*, pp. 4278–4284 (2016).
54. N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).

55. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Int. Conf. Learn. Represent.* (2015).
56. Y. Yang and S. D. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, pp. 270–279 (2010).
57. D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.* **8**(1), 173–176 (2011).
58. V. Risojevic, S. Momic, and Z. Babic, "Gabor descriptors for aerial image classification," in *Int. Conf. Adapt. and Nat. Comput. Algorithms*, pp. 51–60 (2011).
59. S. Basu et al., "DeepSat: a learning framework for satellite imagery," in *Adv. Geogr. Inf. Syst.*, p. 37 (2015).
60. B. Zhao et al., "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.* **54**(4), 2108–2123 (2016).
61. G. S. Xia et al., "AID: a benchmark dataset for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.* **55**(7), 3965–3981 (2017).
62. W. Zhou et al., "PatternNet: a benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.* **145**, 197–209 (2018).
63. Y. Yi and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *IEEE Int. Conf. Comput. Vision* (2011).
64. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2169–2178 (2006).
65. Y. Jiang, J. Yuan, and Y. Gang, "Randomized spatial partition for scene recognition," *Lect. Notes Comput. Sci.* **7573**, 730–743 (2012).
66. Z. Fan, D. Bo, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.* **53**(4), 2175–2184 (2014).
67. H. Fan et al., "Unsupervised feature coding on local patch manifold for satellite image scene classification," in *IEEE Geosci. and Remote Sens. Symp.* (2014).
68. C. Gong et al., "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **53**(8), 4238–4249 (2015).
69. M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.* **7**(1), 28–32 (2010).
70. G. Csurka et al., "Visual categorization with bags of keypoints," in *Eur. Conf. Comput. Vision* (2004).
71. X. Wang et al., "Integration of heterogeneous features for remote sensing scene classification," *J. Appl. Remote Sens.* **12**(1), 015023 (2018).
72. A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(4), 712–727 (2008).
73. A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Mach. Learn.* **45**(1), 157–176 (2011).
74. B. Zhao et al., "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.* **8**(2), 157 (2016).

Xin Zhang obtained her bachelor's degree from Northeastern University at Qinhuangdao in 2016. Currently, she is studying for her doctorate at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. Her research direction is deep learning and image processing.

Yongcheng Wang is a researcher. He received his bachelor's degree from Jilin University in 2003 and his doctoral degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences in 2010. His research direction is image engineering and space payload embedded systems.

Ning Zhang received her bachelor's degree from Northeastern University at Qinhuangdao in 2017. She is currently a PhD student at Changchun Institute of Optics, Fine Mechanics and

Physics, Chinese Academy of Sciences. Her research interests cover image processing and deep learning.

Dongdong Xu received his bachelor's degree from Shandong University in 2013 and a master's degree from Harbin Institute of Technology in 2015. Currently, he is a PhD student and a research assistant at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include deep learning, image processing, and embedded systems.

Bo Chen is a researcher and a professor. He obtained his bachelor's degree from Jilin University in 1984, and obtained his doctoral degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2003. His research direction is the technology of space optics.

Guangli Ben received his bachelor's degree in 2013 and a master's degree in 2016 from Harbin Engineering University. He is a PhD student and a research assistant at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research direction is digital signal processing and space payload embedded systems.

Xue Wang received the bachelor's degree from Dalian University of Technology in 2008 and her master's degree from Jilin University in 2012. Currently, she is a research assistant in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. Her current research interests include machine learning, software designing and testing, and software maintenance.