IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 19, 2021, accepted February 4, 2021, date of publication February 8, 2021, date of current version February 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057868

Urban Remote Sensing Scene Recognition Based on Lightweight Convolution Neural Network

JINGMING XIA[1], YUE DING[1], AND LING TAN[2]
[1]School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

Corresponding author: Ling Tan (cillatan0@nuist.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41505017.

ABSTRACT The use rate of urban land is a significant sign to evaluate urban construction, and scene recognition has important application value in improving urban land use rate. In this paper, a new lightweight model based on VGG16 is proposed to extract distinct features of remote sensing images through five convolution modules. This model uses depthwise separable convolution to reduce the network parameters. An adaptive pooling layer is added to solve the inherent non-adaptive problem of the convolution network. It makes the network insensitive to the size of the input image. The global average pooling layer is used to sum the information to make the input spatial transformation more stable. This paper conducts training and testing on two data sets, NWPU-RESISC45 Dataset and SIRI-WHU Dataset, and the recognition scenarios are 13 and 12 categories. Experimental results show that this method is better than other models in recognition accuracy and model size.

INDEX TERMS Adaptive pooling, lightweight network, land use, scene recognition.

I. INTRODUCTION

Urban land use/ land cover ratio (LULC) is an important evaluation sign of urban construction strategy. The rapid development of cities leads to a change in land cover. Dense population and highland use rate are the main characteristics of cities, so reasonable urban regional identification is the premise of urban management and planning [1]. The change in urban land use rate is an important reason for the rapid development of society. Remote sensing has the ability in the periodically urban area monitoring, and neural network can carry out systematic modeling, so it has become two major tools to significantly promote the related research on urban LULC changes [2].

Remote sensing technology and deep learning are widely used to study the change of LULC rate in cities. The rapid development of remote sensing technology has diversified ways of obtaining urban image data [3]. At the same time, the emergence of deep learning promotes the study of image scene classification [4], semantic segmentation [5], target detection [6], and other fields. Scene image annotation for scene classification costs a lot of labor, so support vector machine (SVM) has been widely concerned by pattern recognition field and applied to remote sensing image recognition. Lai *et al*. [7] studied the method of using SVM for road recognition in remote sensing images with edge features, extracted the edge features of the image, and achieved good results. However, SVM is difficult to train large-scale samples. To solve this problem, a pixel-centric spectral method is proposed. Chen *et al*. [8] first introduced the concept of deep learning to hyperspectral data classification and verified the applicability of automatic encoder through the classification based on spectral information. The framework is a hybrid of principle component analysis (PCA), deep learning architecture, and logistic regression, which has higher accuracy than the traditional pixel-centric spectral method. Huang *et al*. [9] fused the analysis of multi-spectral images into the classification layer, which is called a semi-transferred CNN with a deep structure. The method solved the problem that the DCNN method is not ideal for multi-spectral remote sensing images with more than three channels and the training samples are limited. It shows superior performance on LULC. Convolution neural network usually ignores the texture image contain discriminating information, Huang and Xu [10] proposed a CaffeNet-based method to overcome this problem. The method developed an improved bag-of-view-word (iBoVW) coding method to represent the discriminating information from each convolutional layer, weighted concatenation is

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

employed to combine different features for classification, which can provide high-resolution remote sensing image with distinguishing description.

In recent years, convolution neural networks (CNNs) have been widely used in remote sensing image classification. It is difficult for the existing CNNs to balance network depth and model size. The network structure becomes complicated, but fine features are often lost or even disappeared in the deep convolution process. The training parameters are greatly increased, and more training time is required [11]. However, too few convolutional layers will cause insufficient image features, making it impossible to identify urban scenes quickly and correctly. To solve too many parameters in deep CNNs, a lightweight end-to-end CNN is presented (Lightweight Convolution Neural Network, hereinafter referred to as LW-CNN) to reduces the network model while ensuring the deep convolution of the network. Finally, the lightweight network model is embedded in the Unmanned Aerial Vehicle (UAV) and other mobile devices to identify the urban remote sensing scene. The main contributions of this paper are as follows:

1. An end-to-end lightweight network (LightWeight-CNN) is proposed. This network is improved based on VGG16. It not only utilizes 5 deep convolution modules to extract remote sensing image features, but also uses depthwise separable convolution instead of standard convolution to reduce the parameters during the convolution operation.

2. The adaptive pooling layer is adopted to adjust the image size after the deep feature extraction of the image. It solves the problem that the existing neural network can only input fixed-size images, and makes full use of the image features in the original image, so that the network can handle the problem of different input image sizes.

3. Using the global average pooling layer to replace the flatten layer, extracting the global average feature, and converting the final output feature into a $1 \times 1 \times$ M tensor. Since the global average pooling sums the spatial information, it is more stable in the input spatial transformation, the convolution structure is simpler and avoids network overfitting.

The Sec. II of this paper introduces related work such as LULC and scene classification. In Sec.III, the methods proposed in this paper are introduced in detail. Experiments and analysis are conducted in Sec. IV and conclusions are presented in Sec. V.

II. RELATED WORK

Deep learning-based land use and land cover classification are explored both at pixel-level, object-level, and scene-level. Due to the limited spatial resolution of optical remote sensing images, the pixel-centered spectral method is the mainstream of traditional LULC classification work [12]. Li and Zhang [13] introduced a geostatistics framework based on the Markov chain to classify pixel-centered images. However, the pixel-based classification method has the problem of salinization effect in the classification results. The object-based classification method fully utilizes the local spatial information of irregularly shaped objects in the image, and deals with the problems of the traditional pixel-based classification method [14]. Zhao *et al*. [15] proposed a combination of deep learning strategy and object-based classification to accurately capture the contours of different objects. This method fills the gap between complex image patterns and semantic tags. The rapid development of high spatial resolution remote sensing images has brought new opportunities to land use classification. With the continuous increase of remote sensing image data, more deep learning models have been applied to urban land use. Remero *et al*. [16] proposed the greedy hierarchical unsupervised pre-training learning algorithm, which has good performance in aviation scene classification and high-resolution land use classification. Due to the lack of remote sensing image categories and insufficient labeling data, it is difficult to directly apply deep convolution features to remote sensing images. Yuan *et al*. [17] proposed a pyramid multi-subset feature fusion method. It can effectively fuse the deep features extracted from different pre-trained CNNs and integrate the global and local information of the deep features, thereby obtaining stronger discriminative and low-dimensional features. Ye [18] *et al*. proposed a new classification method based on deep learning and metric learning. The cluster center of each class is preset on the output feature, and the Euclidean distance is used to calculate the average center metric loss. In the feature space, this method improved the classification accuracy by forcing intra-class compactness and inter-class separability.

The CNN models are used in remote sensing scene classification, but training a deep CNN will produce large parameters. In recent years, networks such as AlexNet, VGG16, GoogleNet, and InceptionV3 have shown strong generalization capabilities in scene recognition. The use of ready-made pre-trained CNN models as a general feature extractor has become a method of remote sensing scene classification [19]. However, generating large parameters during the training process leads to higher requirements on hardware devices. To reduce the amount of CNN input data, Zhao *et al*. [20] extended a simple linear iterative clustering algorithm. It makes full use of the spatial spectrum and environmental information of superpixels to segment images and generate superpixels. Wei *et al*. [21] proposed a CNN classification structure based on cube pairs and established a three-dimensional full convolutional network model. This model fully uses the three-dimensional features of hyperspectral images and has fewer parameters than traditional CNN. Liu *et al.* [22] proposed a dense dilated convolution merge network, it utilizes the expansive convolution combination and expands the network's receptive field with fewer parameters. The evolution from the traditional pixel-based method to the model-based method reduces the number of network parameters to a certain extent. Howard *et al*. [23] proposed to use depthwise separable convolution to construct a lightweight neural network. Depthwise separable convolution consists of two layers, one for filtering and one for merging, this factorization can significantly reduce the

computation and model size. Zhang *et al*. [24] improved U-Net by taking depthwise separable convolution, proposed a neural network-based remote sensing image city building extraction model, and optimized the network hyperparameters according to the characteristics of the building. Liu *et al*. [25] proposed an improved full convolutional network (FCN), the depthwise separable convolution is used to replace the original convolution of FCN, so that the whole network has fewer parameters and the performance of the method is better. Yu [26] *et al*. introduced a bilinear convolutional neural network and used MobileNetV2 as a feature extractor, each feature is transformed into two features with different convolutional layers to enhance bilinear features. In this paper, the depthwise separable convolution method is used to reduce the parameters of the CNN network in the training process. Under the premise of ensuring the depth of the network, the network parameters can be reduced to about 67 times.

Due to the different sources and resolutions of remote sensing images, image size has always been a concern. Although the CNN-based remote sensing scene recognition network has deep layers, due to the non-self-adaptability of convolution operation, the CNN still has limitations, that is, regardless of the input image size, the convolution has a fixed weight and cannot be effective in Distinguish all types of images. Fang *et al*. [27] proposed a multi-scale adaptive sparse representation (MASR) model, the multi-scale spatial information is effectively utilized to limit the pixels from different scales to achieve better image representation. Kim *et al*. [28] proposed a characteristic response modulation network based on adaptive convolution, it solves the inherent inadaptability problem of CNN-based network. Since the input image is limited by the pixel size, the performance of the network in large-size image input is affected. To enable the network to handle large-size input well, Wei *et al*. [29] proposed to add an adaptive pooling layer to the network to process large-size image input, breaking the limitation of image input size required by deep CNNs. To taking the non-adaptive problem of the CNN, the algorithm in this paper adds an adaptive pooling layer to the network, and adjusts the size and step length of the convolution kernel through feature mapping.

Deep CNN models are usually trained on ImageNet containing millions of images, while the remote sensing data set NWPU-RESISC45 contains less than 35,000. Too few data sets will lead to over-fitting of CNNs. Zhang *et al*. [30] proposed a feature extraction algorithm for sparse transmission, sparse constraint and transfer constraint are introduced to avoid overfitting caused by too few training samples. Tian *et al*. [31] used regularization, loss and fine-tuning strategies to alleviate the overfitting problem in remote sensing image classification during CNN training. Dai *et al*. [32] adopted a CNN combining multi-scale deep residuals. In the back propagation and forward propagation process, global average pooling is introduced to solve the above problems. This paper enhances the data set and adds a global average pooling layer to the network structure to solve the over-fitting problem of CNNs.

To sum up, the LW-CNN proposed in this paper classifies remote sensing images at the scene level, and uses depthwise separable convolution instead of standard convolution to construct a lightweight network. The adaptive pooling layer is adopted to solve the problem of different sizes of remote sensing images. Finally, the global average pooling layer is used to extract the global average feature. In the CNN training process, the overfitting problem is solved by avoiding the process of parameter optimization.

III. METHODS

A. LW-CNN NETWORK STRUCTURE

VGG16 [33] is a kind of deep learning network model that has attracted much attention in recent years. VGG16 contains 5 convolution modules and 3 fully connected layers. After each convolution module is the maxpooling layer, the number of convolution channels starts from 64 in the first layer and doubles after each maxpooling layer until 512. Each convolution module contains multiple convolutional layers of a 3×3 convolution kernels. On the one hand, parameters are reduced; on the other hand, more nonlinear mapping is carried out to enhance the fitting ability of the network.

VGG16 network is simple and practical, has a deeper feature extraction than AlexNet, and a simpler network structure than GooLeNet and ResNet50. Compared with other lightweight convolutional neural networks, the number of parameters itself is very small, and the operation of parameter reduction will make the identification accuracy of the network decline. Based on the recognition rate and network structure, this paper proposes a lightweight model (LW-CNN) based on VGG16. It solves the problem of CNN with too much computation and the inherent inadaptability of the input image in CNN network. LW-CNN has the following characteristics:

Firstly, the five convolution modules in VGG16 are retained, including the convolution kernel and the number of convolution channels. The depthwise separable convolution method is adopted, and the stride of the first convolution layer of each convolution module is changed from 1 to 2 to replace the maxpooling layer in VGG16. The lightweight network is constructed from two aspects: changing the convolution mode and reducing the number of network layers.

Secondly, an adaptive pooling layer is added. After the deep feature extraction of the image, the image size is adjusted through adaptive pooling, and the image features in the original image are fully utilized, so that the network can handle the size of the input image well.

Finally using the global average pooling layer to replace the flatten layer, extract the global average feature, and convert the final output feature into a $1 \times 1 \times$ M tensor. Since the global average pooling sums the spatial information, it is more stable in the input spatial transformation, the convolution structure is simpler and avoids network overfitting,

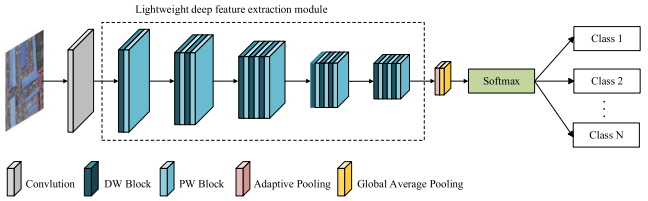


FIGURE 1. LW-CNN network structure.

combining softmax layer to forecast the result of the identification.

The LW-CNN network structure is shown in Fig.1. The lightweight deep feature extraction module is shown in the Fig.1. DW Block includes depthwise convolution and Batch Normalization, and PW Block includes pointwise convolution and Batch Normalization. In this paper, features in remote sensing images are extracted layer by layer through 5 convolution modules of LW-CNN. By changing the convolution mode, the number of network model parameters is reduced from 134M to 1.7M, which is 67 times smaller. After the feature extraction module, an adaptive pooling layer is added to solve the problem that the size of the input image is limited by the CNN. LW-CNN can be used for urban scene recognition of 50m-100m UAV aerial photograph.

B. DEPTHWISE SEPARABLE CONVOLUTION

Depthwise separation is a form of factoring Convolution, it divides the standard convolution into depthwise convolution (hereinafter referred to as DW Convolution) and pointwise convolution (hereinafter referred to as PW Convolution), significantly reduce the convolution computation, and add the batch normalization layer after each convolution. The filters decomposition process of standard convolution is shown in Fig.2. Depthwise separable convolution applies a filter to each input channel, splitting the filtering and combination of standard convolution into two independent modules [23].

The calculation amount of standard convolution X_1 is:

$$X_1 = D \times D \times M \times N \times D_W \times D_H \tag{1}$$

The sum of the computations of DW convolution and PW convolution X_2 is:

$$X_2 = D \times D \times M \times D_W \times D_H + M \times N \times D_W \times D_H \tag{2}$$

The reduced calculation amount X_3 is:

$$X_3 = \frac{D \times D \times M \times D_W \times D_H + M \times N \times D_W \times D_H}{D \times D \times M \times N \times D_W \times D_H} \tag{3}$$

Among them, the input mapping size is (D_W, D_H, M), the kernel of standard convolution is (D, D, M, N), D_W and D_H are the width and height of the input map respectively, M is the number of input channels, D is the height and width of the convolution kernel, and N is the number of output channels.

To verify the effect of depthwise separable convolution and build an appropriate lightweight network, this paper attempts

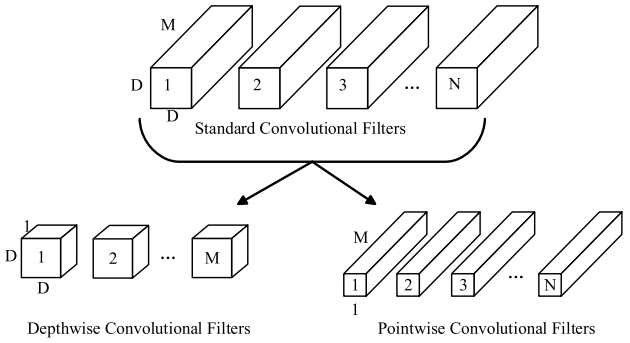


FIGURE 2. Decomposition process of standard convolution filters.

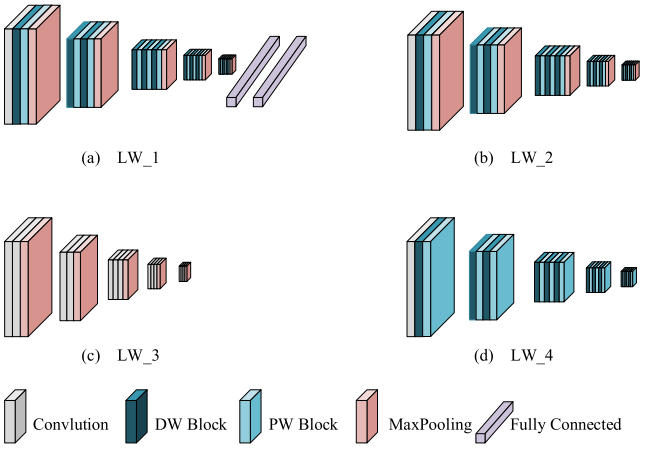


FIGURE 3. Four lightweight modules.

to reduce the network model in four ways. The corresponding four models are shown in Fig.3. First, LW_1 changes the standard convolution in VGG16 to depthwise separable convolution, and found that the two fully connected layers produced a large number of parameters during the model operation. Therefore, LW_2 only deletes two full connection layers based on LW_1, while LW_3 only deletes two full connection layers of VGG16 without changing the convolution mode. Based on LW_2, LW_4 changes the stride of the first convolutional layer of each convolution module from 1 to 2 to replace the maximum pooling layer. See Sec. IV-C for details of the experiment. By comparing the model sizes of the four networks, and the trade-off between the recognition accuracy and the loss value on the two data sets, LW_4 is finally selected as the lightweight network model in this paper.

C. ADAPTIVE POOLING LAYER

With the development of remote sensing technology, the source and resolution of remote sensing images are different, and the image size becomes the focus of image classification. The existing classical convoluted neural network can only input images of fixed size. For example, the FC layer input in VGG16 is $7 \times 7 \times 512$, including 25088 connections, so the input picture is fixed to 224×224 pixels. For large input, resizing the image or randomly cropping it to 224×224

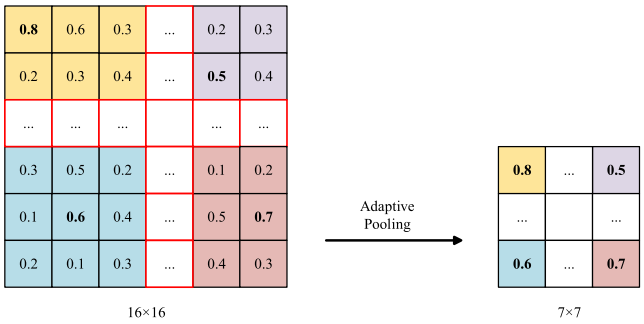


FIGURE 4. Adaptive pooling principle.

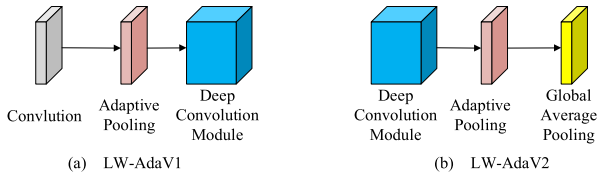


FIGURE 5. Two ways to add adaptive pooling layer.

pixels is not enough to take full advantage of the data set. To make the network insensitive to the input image size, feature mapping level processing is adopted in this paper, and an adaptive pool layer is used to carry out down-sampling for the feature mapping. The adaptive pooling layer converts the original size of the feature into several smaller sizes. The adaptive pooling process is shown in Fig.4. For example, the input image size is set to 256 × 256 pixels, and the input size of the adaptive pool layer is 16 × 16 pixels. Then, the adaptive pooling layer divides the 18 × 18 feature map into 7 × 7 subcells that are approximately 2 pixels or 3 pixels. Next, the maximum value in each subcell is mapped to the corresponding output grid position. Finally, we obtain small-sized features of 7 × 7 pixels.

The adaptive pooling layer mainly solves the limitation of the input image size. This paper uses two methods to add the adaptive pooling layer: LW-AdaV1 is shown in Fig.5 (a). After the input image is subjected to a standard convolution, an adaptive pooling layer is added, the image size is changed and then input to the deep convolution feature extraction module. As shown in Fig.5 (b), LW-AdaV2 performs deep feature extraction on the original image and adds an adaptive pooling layer after the fifth convolution module. This paper compares the indicators of the two methods, and finally chooses the LW-AdaV2 method to add an adaptive pooling layer. For the experimental details, see Sec. IV-D.

D. GLOBAL AVERAGE POOLING

After deep feature extraction, traditional CNN vectorizes the feature map of the last convolutional layer, connects it to the FC layer, and finally classifies the feature through the softmax layer [34]. In this paper, the global average pooling layer is used to replace the full connection layer in CNN. Global processing is carried out for the entire feature graph, and the average value of features is extracted. Finally,

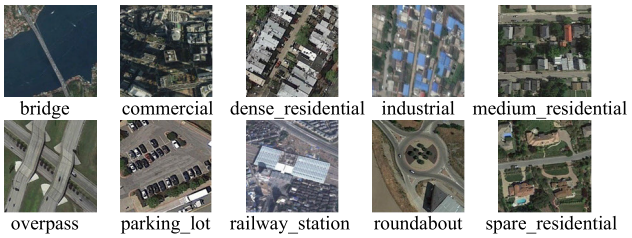


FIGURE 6. Sample image of NWPU dataset.

a 1 × 1× M tensor is generated and directly entered into the softmax layer. The advantage of global average pooling over the fully connected layer is that it reduces network parameters. In addition, there are no parameters to be optimized in the global average pooling, so overfitting is avoided at this layer. And because the global average pooling sums the spatial information, it is more stable in the input spatial transformation. Sec. IV-E experimental results show that the overall recognition accuracy increases after adding the global average pooling layer.

IV. EXPERIMENT AND ANALYSIS

A. DATASET INTRODUCTION

To evaluate the performance of LW-CNN, we conducted a large number of comparison experiments on two remote sensing scene data sets, identifying 13 types and 12 types of scenes on the two data sets respectively. Compared with other methods in recognition accuracy, recognition rate, etc. The two data sets used in this paper are described in detail below.

NWPU-RESISC45 Data Set (hereinafter referred to as NWPU) [35] contains 45 types of remote sensing scenes, each of which consists of 700 images, each with a size of 256 × 256 pixels. For each scene category, the NWPU dataset has rich changes in appearance, spatial resolution, lighting, background, and occlusion, etc. The main application scene of this paper is urban land-use scene identification. Therefore, 13 types of commonly used urban scenes are selected in this paper, including bridge, commercial_area, dense_residential, industrial_area, lake, meadow, medium_residential, overpass, parking_lot, railway_station, river, roundabout, and spare_residential. Fig.6. is an example image of the NWPU data set.

SIRI-WHU Dataset (hereinafter referred to as SIRI) [36] contains 12 categories of remote sensing scene images, a total of 2400 images, each category is composed of 200 images, each with a size of 200 × 200 pixels. The data comes from Google Earth and mainly covers urban areas in China. The 12 land use categories include agriculture, commercial, harbor idle_land, industrial, meadow, overpass, park, pond, residential, river, and water. Fig.7. is an example image of the SIRI dataset.

B. TRAINING DETAILS AND EVALUATION INDICATORS

The LW-CNN model proposed in this paper is built on the Keras library of python, and it traverses the images in

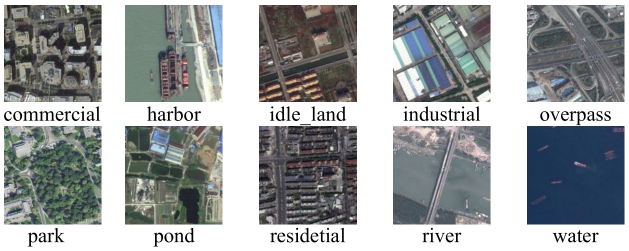


FIGURE 7. Sample image of SIRI dataset.

the dataset through ImageDataGenerator, and enhance each image with random transformation data. The number of epochs was set to 100, and the number of training batch_size was set to 36. Use stochastic gradient descent (SGD) optimizer to train the LW-CNN network model, momentum is set to 0.9, weight decay is set to 0.0002, the initial learning rate is set to 0.001, with 10 batches of recognition accuracy as a learning unit, if the value does not increase, the learning rate will decrease by 10%.

To evaluate the performance of the network in all aspects, this paper uses Accuracy, Precision, Recall, F1 score, Parameter and FPS as the evaluation indicators of the experimental results. Among them, the Parameter is used to measure the size of the network model, and FPS is the number of frames of images identified per second. The calculation formula of other evaluation indexes is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precison} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

In the formula, P represents Positive, N represents Negative, FP is the number of samples that are actually negative but predicted to be positive, TN is the number of samples that are actually negative and predicted to be negative, and TP is actually the number of samples that are predicted to be positive, and FN represents the number of samples that are actually positive but predicted to be negative.

C. LIGHTWEIGHT NETWORK COMPARISON EXPERIMENT

To select a better lightweight network, try four ways to reduce network parameters. Training was conducted on two training sets respectively, with 80% of the data set as the training set and 20% as the test set. The recognition results of the VGG16 and four methods are shown in Table 1. According to the experimental results, a comparison diagram of the accuracy and loss of the verification set is drawn, as shown in Fig.8 and Fig.9. Fig.8 compares LW_4 with the other three algorithms. The network parameters of LW_4 and LW_2 algorithms are about 2M, and the test on the two data sets shows that the accuracy is better than LW_1 and LW_3,

TABLE 1. Recognition results of 4 methods.

Dataset	Models	Parameter	Accuracy
NWPU	VGG16	134.28M	88.76%
	LW_1	121.3M	88.3%
	LW_2	2.02M	90.76%
	LW_3	15.01M	90.84%
	LW_4	2.02M	**91.15%**
SIRI	VGG16	134.28M	90.20%
	LW_1	121.9M	86.45%
	LW_2	2.00M	93.12%
	LW_3	14.99M	90.83%
	LW_4	2.00M	**92.70%**

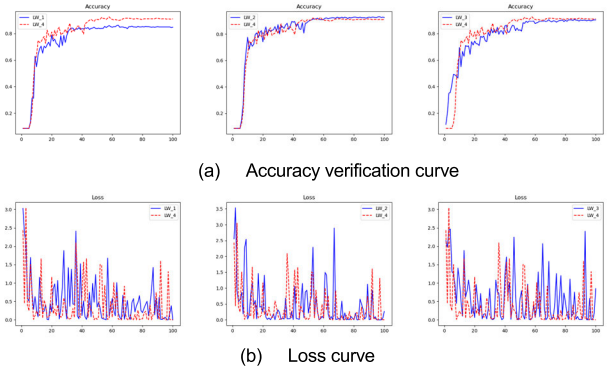


FIGURE 8. Performance comparison between LW_4 and other algorithms (NWPU).

which is due to the effectiveness of the depthwise separable convolution method adopted by LW_4 and LW_2.

The model parameters of LW_4 and LW_2 are the same. LW_4 deletes the 4 maxpooling layers based on LW_2, and reduce the size of the convolution feature map through the convolution step of DW, thus reducing the number of network layers. On the NWPU data set, the accuracy of LW_4 in identifying 13 scenarios was 91.15%, higher than 90.76% for LW_2. On the SIRI dataset, the accuracy of LW_4 in identifying 12 types of scenes is 92.7%, which is lower than the 93.12% of LW_2. Observe the loss function graphs in Fig.9 (b) and Fig.10 (b). After multiple iterations of training, it is found that the loss of the LW_4 network is smaller than the loss of LW_2, the number of layers of LW_4 network was smaller, so LW_4 is finally selected as lightweight network model.

The accuracy of LW_4 in identifying 13 types of scenes on the NWPU dataset is shown in Table 2, the accuracy of dense_residential and parking_lot is as high as 99%, and the accuracy of identifying 12 types of scenes on the SIRI dataset is shown in Table 3. The recognition accuracy of meadow and park is as high as 100%. However, the river recognition accuracy on the two datasets is not ideal, 83% and 77.5% respectively. This is because the convolutional neural network CNN deep convolution module only retains salient features when extracting image features, and river images often contain bridges. Because the river area is too large,

TABLE 2. Recognition results of LW_4 on NWPU.

Dataset	bridge	commercial	dense	industrial	lake	meadow	medium	overpass	parking	railway	river	roundabout	spare
NWPU	92%	90%	99%	86%	92%	94%	87%	88%	99%	91%	83%	86%	98%

TABLE 3. Recognition results of LW_4 on SIRI.

Dataset	agriculture	commercial	harbor	idle_land	industrial	meadow	overpass	park	pond	residential	river	water
SIRI	100%	97.5%	92.5%	95%	95%	100%	80%	100%	85%	92.5%	77.5%	97.5%

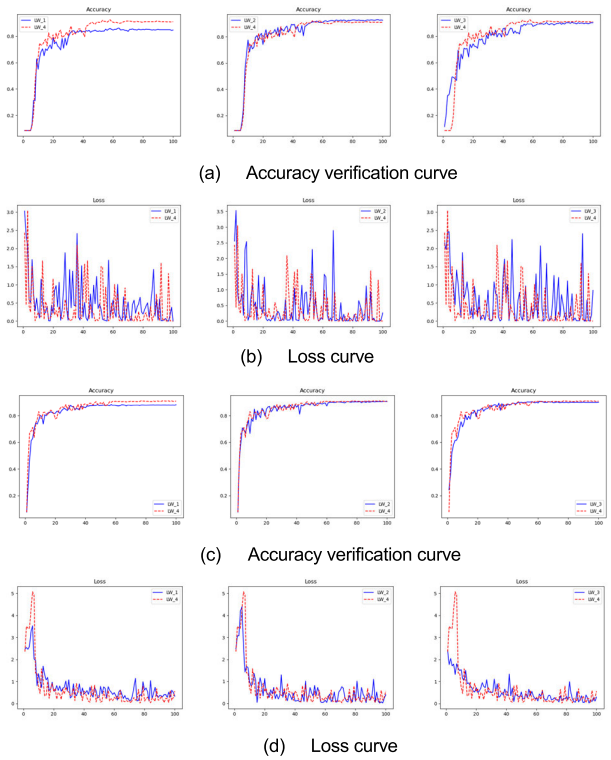


FIGURE 9. Performance comparison between LW_4 and other algorithms (SIRI).

it is easy to be recognized as a background rather than a feature, causing the river image to be incorrectly recognized as a bridge.

D. ADAPTIVE POOLING EXPERIMENT

To solve the problem that only fixed-size images can be input into the convolutional neural network CNN, this paper uses two methods to add an adaptive pooling layer on the LW_4 network. LW_AdaV1 adds an adaptive pooling layer after one standard convolution. LW_AdaV2 performs deep feature extraction on the original image, and adds adaptive pooling layer after the fifth convolution module to fully extract the features of the original image.

In this paper, experiments were carried out on two data sets, and the results of adaptive pooling were shown in Table 4. According to the experimental results, the accuracy verification curve on the verification set was drawn, as shown in Fig.10. The accuracy of LW_AdaV2 on the two data

TABLE 4. Experimental results of adaptive pooling.

Dataset	Models	Accuracy
NWPU	LW_AdaV1	76.46%
	LW_AdaV2	**90.38%**
SIRI	LW_AdaV1	66.04%
	LW_AdaV2	**88.74%**

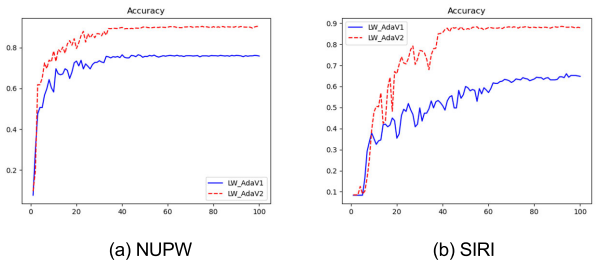


FIGURE 10. Accuracy verification curves of LW_AdaV1 and LW_AdaV2.

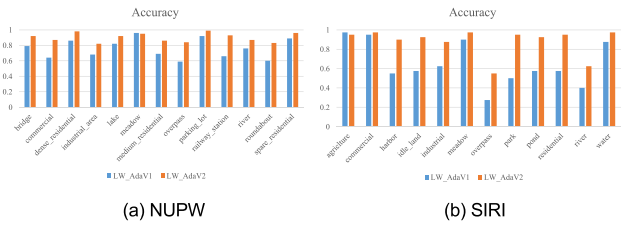


FIGURE 11. Verification accuracy of LW_AdaV1 and LW_AdaV2.

TABLE 5. Recognition results of ablation experiments.

Dataset	Models	Parameter	Accuracy
NWPU	LW_AdaV2	2.0M	90.38%
	LW-CNN	1.70M	**91.23%**
SIRI	LW_AdaV2	2.0M	88.74%
	LW-CNN	1.70M	**93.51%**

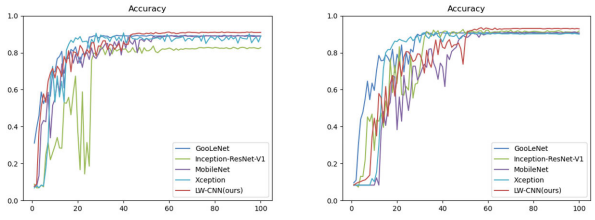
sets was 90.38% and 88.74%, which was far better than LW_AdaV1. The two methods are compared on the two data sets for the recognition accuracy of various scenes. The result is shown in Fig.11. In all scenes, the recognition accuracy of LW_AdaV2 network is higher than that of LW_AdaV1. LW_AdaV2 makes full use of the information contained in the image, while the image features extracted

TABLE 6. Recognition results of ablation experiments on NWPU.

Method	bridge	commercial	dense	industrial	lake	meadow	medium	Overpass	parking	Railway	river	Roundabout	spare
LW_AdaV2	92%	87%	98%	82%	92%	95%	86%	84%	99%	93%	87%	83%	96%
LW-CNN	87%	90%	98%	87%	94%	97%	89%	83%	100%	88%	88%	90%	95%

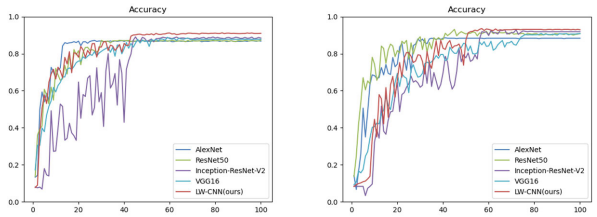
TABLE 7. Recognition results of ablation experiments on SIRI.

Method	agriculture	commercial	harbor	idle_land	industrial	meadow	overpass	park	pond	residential	river	water
LW_AdaV2	95%	97.5%	90%	92.5%	87.5%	97.5%	55%	95%	92.5%	95%	62.5%	97.5%
LW-CNN	97.5%	100%	90%	95%	97.5%	100%	87.5%	92.5%	87.5%	97.5%	80%	97.5%



(a) NUPW (b) SIRI

FIGURE 12. Accuracy verification curves of LW-CNN and lightweight networks.



(a) NUPW (b) SIRI

FIGURE 13. Accuracy verification curves of LW-CNN and non-lightweight networks.

by LW_AdaV1 are not sufficient, leading to the loss of many important image features in the deep extraction module. However, LW_AdaV2 avoids the feature loss of LW_AdaV1, thus achieving better recognition accuracy.

E. ABLATION EXPERIMENT

To verify the performance of the global average pooling layer, LW_AdaV2 does not add a global average pooling layer, and LW-CNN adds. Ablation experiments were performed on two data sets, and the experimental results are shown in Table 5. Validated on the NWPU dataset, the accuracy of 13 types of scene recognition increased from 90.38% to 91.23%, and verified on the SIRI dataset, the accuracy of 12 types of scene recognition increased from 88.74% to 93.51%. The network with the global average pooling layer improves the overall recognition accuracy while reducing the number of parameters.

In the ablation experiment, the identification accuracy of each scene on the NWPU data set is shown in Table 6. The LW-CNN network has better recognition accuracy in scenes such as commercial districts, high-density residential areas, lakes, grasslands, roundabouts, etc. The recognition accuracy of each type of scene in the ablation experiment on the SIRI data set is shown in Table 7. The recognition accuracy of LW-CNN in commercial areas, industrial areas, overpasses, and rivers have been greatly improved.

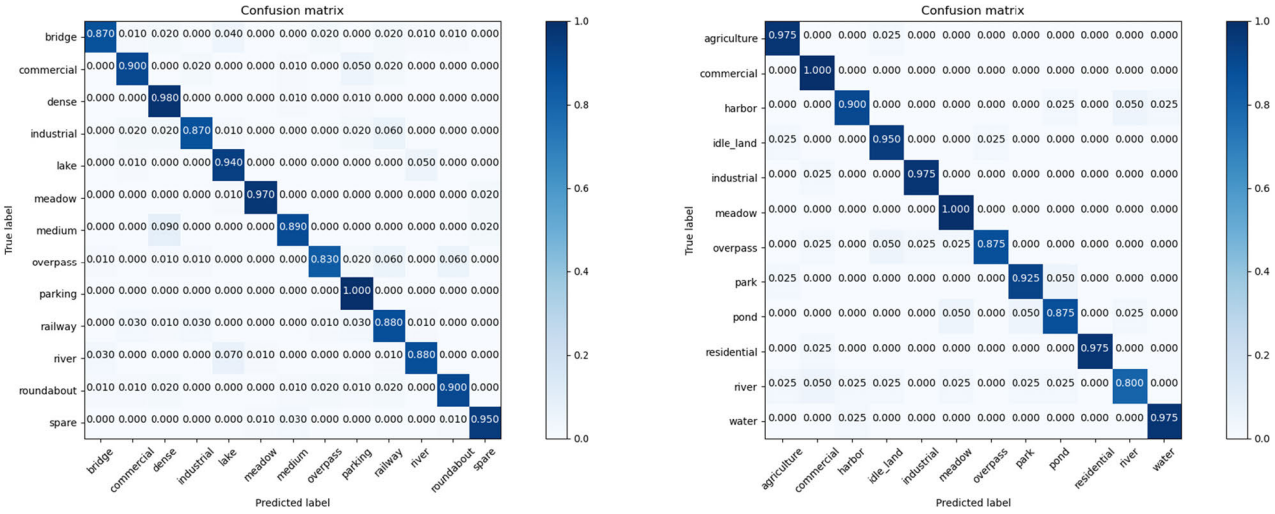
F. PERFORMANCE COMPARISON EXPERIMENTS WITH OTHER ALGORITHMS

To assess the urban scene recognition performance of LW-CNN, the network model was compared with other deep learning models (AlexNet, GoogLeNet, ResNet50, Inception-ResNet-V1, Inception-ResNet-V2, MobileNet, VGG16, Xception) on two data sets. In the experiment, the recognition accuracy and F1 parameters are shown in Table 8. The LW-CNN method proposed in this paper has the highest recognition accuracy, with a recognition accuracy of 91.23% on 13 types of scenes, and recognition accuracy of 93.51% for 12 types of scenes; The model parameters in this paper are 1.7M, which is 67 times less than the VGG16 network parameter, the recognition accuracy is the best and the model parameter is the smallest. The recognition rate of LW-CNN on two datasets is better than that of other networks and only slightly slower than that of AlexNet. By analyzing the network structure, it is found that LW-CNN contains 25 convolution layers to extract image features, while AlexNet contains 5 convolution layers, so it has more advantages in recognition rate. It can be seen from Table 8 that the FPS of VGG16 and LW-CNN are equivalent, but the number of parameters of LW-CNN is significantly reduced, which proves from another aspect that the number of parameters of network is not an important factor affecting FPS, and the difference of network structure has a greater impact on FPS. But compared with other deep network models, the recognition rate of this paper still maintains an advantage. Comparing the method with the lightweight network MobileNet and Xception, LW-CNN model parameters were 1.5M less than MobileNet, and the recognition accuracy was about 2% higher on both data sets. LW-CNN model parameters were 19M less than Xception, and the recognition rate was faster.

Fig.12. is drawn based on the accuracy comparison between LW-CNN and lightweight networks (GooleNet,

TABLE 8. Identification results of different methods.

Dataset	Models	Parameter	Accuracy	Precision	Recall	F1	FPS
NWPU	AlexNet	58.33M	87.15%	0.877	0.871	0.871	170
	GoogLeNet	12.45M	89.53%	0.903	0.895	0.895	49
	ResNet50	45.76M	87.30%	0.879	0.873	0.873	45
	Inception-ResNet-V1	21.06M	81.61%	0.822	0.816	0.814	44
	Inception-ResNet-V2	30.41M	88.92%	0.896	0.889	0.889	40
	MobileNet	3.2M	89.46%	0.899	0.894	0.894	92
	VGG16	134.28M	88.76%	0.893	0.887	0.888	91
	Xception	20.88M	87.53%	0.883	0.875	0.875	62
	LW-CNN(ours)	**1.70M**	**91.23%**	0.914	0.912	0.912	95
SIRI	AlexNet	58.33M	88.54%	0.885	0.885	0.877	107
	GoogLeNet	12.45M	91.45%	0.915	0.914	0.912	30
	ResNet50	45.76M	91.45%	0.916	0.914	0.913	27
	Inception-ResNet-V1	21.05M	91.66%	0.918	0.916	0.914	27
	Inception-ResNet-V2	30.40M	92.91%	0.928	0.929	0.927	25
	MobileNet	3.2M	91.25%	0.914	0.912	0.910	58
	VGG16	134.28M	90.20%	0.902	0.902	0.899	55
	Xception	20.88M	91.45%	0.916	0.914	0.914	40
	LW-CNN(ours)	**1.70M**	**93.51%**	0.936	0.935	0.934	58



(a) NUPW (b) SIRI

FIGURE 14. Confusion matrix of this method on two data set.

Inception- Resnet-V1, MobileNet, Xception). The accuracy comparison of LW-CNN with non-lightweight networks (AlexNet, ResNet, Inception- ResNet-v1, VGG16) is shown in Fig.13. Compared with the lightweight network, the depth feature extraction module of LW-CNN is simple and effective, and the verification curve of LW-CNN tends to rise steadily. Compared with non-lightweight networks, LW-CNN has more advantages in parameters. After 100 batches of training, the verification accuracy of LW-CNN is higher than that of other network structures.

This paper draws confusion matrices on two data sets to show the recognition performance of the network, as shown in Fig.14. On the NWPU dataset, the recognition accuracy of the parking lot reached 100%. Due to the complexity of the NWPU dataset and the defects in the recognition of similar scenes in this network, the recognition accuracy of overpass, industrial_area, and the river is slightly lower. On the SIRI data set, the recognition accuracy of commercial and meadow is 100%, and the recognition accuracy of overpass and river is also lower than the average recognition accuracy.

V. SUMMARY AND PROSPECT

Aiming at remote sensing scene recognition of urban land use, this paper proposes a lightweight model (LW-CNN) based on VGG16, and the number of model parameters is reduced by 67 times compared with VGG16. This method adopts an adaptive pooling layer to solve the neural network that can only input fixed-size images. A large number of experiments show that LW-CNN is superior to other classical networks in the recognition of 13 and 12 scenarios on two data sets of NWPU and SIRI, and has advantages in model size and recognition speed. However, this method still has some shortcomings in river scene identification, as too many convolutional layers cause the feature map of the river not to be obvious, so it cannot be accurately identified. At present, the network model is only applicable to urban land-use scene identification, and we hope that through subsequent research and improvement, this model can be applied to more scene identification. At the same time, there are some defects in the study of applying this network to mobile devices in this paper, and this part of the content will be improved through subsequent research and practice.

REFERENCES

[1] R. Cao, W. Tu, C. Yang, Q. Li, J. Liu, J. Zhu, Q. Zhang, Q. Li, and G. Qiu, "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 82–97, May 2020.

[2] L. Mu, L. Wang, Y. Wang, X. Chen, and W. Han, "Urban land use and land cover change prediction via self-adaptive cellular based deep learning with multisourced data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5233–5247, Dec. 2019.

[3] C. Ma, Q. Dai, J. Liu, S. Liu, and J. Yang, "An improved SVM model for relevance feedback in remote sensing image retrieval," *Int. J. Digit. Earth*, vol. 7, no. 9, pp. 725–745, Oct. 2014.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: https://arxiv.org/abs/1704.06857

[6] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[7] J. Y. Lai, A. Sowmya, and J. Trinder, "Support vector machine experiments for road recognition in high resolution images," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.*, 2005, pp. 426–436.

[8] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[9] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[10] H. Huang and K. Xu, "Combing triple-part features of convolutional neural networks for scene classification in remote sensing," *Remote Sens.*, vol. 11, no. 14, p. 1687, Jul. 2019.

[11] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[12] T. Blaschke, G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. Van der Meer, H. Van der Werff, F. Van Coillie, and D. Tiede, "Geographic object-based image analysis–towards a new paradigm," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 180–191, Jan. 2014.

[13] W. Li and C. Zhang, "A Markov chain geostatistical framework for land-cover classification with uncertainty assessment based on expert-interpreted pixels from remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 8, pp. 2983–2992, Aug. 2011.

[14] Q. Yu, P. Gong, N. Clinton, G. Biging, M. Kelly, and D. Schirokauer, "Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 7, pp. 799–811, Jul. 2006.

[15] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[16] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[17] B. Yuan, L. Han, X. Gu, and H. Yan, "Multi-deep features fusion for high-resolution remote sensing image scene classification," *Neural Comput. Appl.*, pp. 1–17, Jun. 2020.

[18] L. H. Ye, L. Wang, W. W. Zhang, Y. G. Li, and Z. K. Wang, "Deep metric learning method for high resolution remote sensing image scene classification?" *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 6, p. 698, 2019.

[19] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 12, no. 4, p. 701, Feb. 2020.

[20] W. Zhao, L. Jiao, W. Ma, J. Zhao, J. Zhao, H. Liu, X. Cao, and S. Yang, "Superpixel-based multiple local CNN for panchromatic and multispectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4141–4156, Jul. 2017.

[21] W. Wei, J. Zhang, L. Zhang, C. Tian, and Y. Zhang, "Deep cube-pair network for hyperspectral imagery classification," *Remote Sens.*, vol. 10, no. 5, p. 783, May 2018.

[22] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated Convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[24] X. Zhang, Y. Zheng, W. Liu, Y. Peng, and Z. Wang, "An improved architecture for urban building extraction based on depthwise separable convolution," *J. Intell. Fuzzy Syst.*, vol. 38, no. 11, pp. 1–9, Jan. 2020.

[25] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1109–1118, 2020.

[26] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," *Sensors*, vol. 20, no. 7, p. 1999, Apr. 2020.

[27] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.

[28] J.-H. Kim, J.-H. Choi, M. Cheon, and J.-S. Lee, "MAMNet: Multi-path adaptive modulation network for image super-resolution," *Neurocomputing*, vol. 402, pp. 38–49, Aug. 2020.

[29] T. Wei, J. Wang, W. Liu, H. Chen, and H. Shi, "Marginal center loss for deep remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 968–972, Jun. 2020.

[30] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1043, Feb. 2014.

[31] T. Tian, L. Gao, W. Song, K.-K. R. Choo, and J. He, "Feature extraction and classification of vhr images with attribute profiles and convolutional neural networks," *Multimedia Tools Appl.*, vol. 77, no. 14, pp. 18637–18656, 2018.

[32] J. Dai, Y. Du, T. Zhu, Y. Wang, and L. Gao, "Multiscale residual convolution neural network and sector descriptor-based road detection method," *IEEE Access*, vol. 7, pp. 173377–173392, 2019.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[35] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[36] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

JINGMING XIA was born in Nanjing, Jiangsu, China, in 1980. He received the B.S. and M.S. degrees in information engineering from the Nanjing University of Information Science and Technology, Nanjing, China, in 2002 and 2005, respectively, and the Ph.D. degree in atmospheric science from Nanjing University of Information Science and Technology, in 2012.

Since 2013, he has been an Assistant Professor with the Artificial Intelligence Department, Nanjing University of Information Science and Technology. He is the author of two books and more than 30 articles. His research interests include application of machine learning and deep learning in meteorology.

YUE DING was born in Lianyungang, Jiangsu, China, in 1996. She is currently pursuing the master's degree with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China.

Her research interests include image processing, computer vision, and deep learning.

LING TAN was born in Wuxi, Jiangsu, China, in 1979. She received the B.S. and M.S. degrees in information engineering from Nanjing Normal University, in 2002 and 2006, respectively, and the Ph.D. degree in information and network from the Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, in 2012.

Since 2014, she has been an Assistant Professor with the Department of Computer Science, Nanjing University of Information Science and Technology. She is the author of three books and more than 20 articles. Her research interests include application of machine learning and deep learning in image processing.