Lightweight Channel Attention and Multiscale Feature Fusion Discrimination for Remote Sensing Scene Classification

HUIYAO WAN\textsuperscript{1}, JIE CHEN\textsuperscript{1,2}, (Member, IEEE), ZHIXIANG HUANG\textsuperscript{1}, (Senior Member, IEEE), YUN FENG\textsuperscript{1}, ZHENG ZHOU\textsuperscript{1}, XIAOPING LIU\textsuperscript{2}, Bайдонтюю (Member, IEEE), ZHENZHEN YUAN\textsuperscript{2}, AND TAO XU\textsuperscript{2}

1Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, Hefei 230601, China
238th Research Institute of China Electronics Technology Group Corporation, Hefei 230088, China
Corresponding author: Jie Chen (jiechen@ahu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62001003, in part by the Natural Science Foundation of Anhui Province under Grant 2008085QF284, in part by the China Postdoctoral Science Foundation under Grant 2020M671851, and in part by the New Energy Vehicle and Intelligent Connected Vehicle Innovation Project.

ABSTRACT High-resolution remote sensing image scene classification has attracted widespread attention as a basic earth observation task. Remote sensing scene classification aims to assign specific semantic labels to remote sensing scene images to serve specified applications. Convolutional neural networks are widely used for remote sensing image classification due to their powerful feature extraction capabilities. However, the existing methods have not overcome the difficulties of large-scene remote sensing images of large intraclass diversity and high interclass similarity, resulting in low performance. Therefore, we propose a new remote sensing scene classification method that combines lightweight channel attention and multiscale feature fusion discrimination, called LmNet. First, ResNeXt is used as the backbone; second, a new lightweight channel attention mechanism is constructed to quickly and adaptively learn the salient features of important channels. Furthermore, we designed a multiscale feature fusion discrimination framework, which fully integrates shallow edge feature information and deep semantic information to enhance feature representation capabilities and uses multiscale features for joint discrimination. Finally, a cross-entropy loss function based on label smoothing is built to reduce the influence of interclass similarity on feature representation. In particular, our lightweight channel attention and multiscale feature fusion mechanism can be flexibly embedded in any advanced backbone as a functional module. The experimental results on three large-scale remote sensing scene classification datasets show that compared with the existing advanced methods, our proposed high-efficiency end-to-end scene classification method has reached state-of-the-art. Moreover, our method has a weaker dependence on labeled data and provided better generalization performance.

INDEX TERMS Remote sensing scene classification, convolutional neural network, lightweight channel attention, multiscale feature fusion, label smoothing.

I. INTRODUCTION

With the rapid improvement of remote sensing and intelligent information processing technology, a large number of remote sensing images have been accumulated. Therefore, how to quickly mine the inherent laws and characteristics behind these massive amounts of data has brought new challenges to the in-depth application of the remote sensing field [10], [12]. In particular, remote sensing scene classification is an very important research direction for the intelligent interpretation of remote sensing images [15], [18]. Remote sensing scene classification extract semantic information by model and classify it into a set of meaningful specific labels. In the past few decades, the practical application of remote sensing scene classification has been extensively studied, such as urban planning [21], [24], natural
disaster detection [25], [26], [28], environmental monitoring [29], [30], and vegetation mapping [31], [32].

In recent years, many methods have been proposed for remote sensing scene classification. Although these methods for remote sensing scene classification have made remarkable achievements, it is still challenging to develop a high-performance and robust method because of the high intraclass diversity and interclass similarity.

At present, remote sensing image scene classification is mainly divided into three types from the perspective of research methods: methods based on manual feature extraction, methods based on unsupervised feature extractors, and methods based on deep learning.

Remote scene classification method based on artificial feature extraction: The past few decades, the resolution and size of remote sensing images was low, and the most scene classification methods relied on artificially designed specific feature extractors. Since the manual feature extractor can express the features of the entire image, it is feasible to directly apply the manual feature extractor to the lower resolution scene image. However, methods based on manual feature extraction rely excessively on manual design and lack flexibility, making manual feature extractor methods difficult to widely use in remote sensing scene classification.

Remote scene classification methods based on unsupervised feature extractors: To overcome manual feature extractors’ shortcomings, unsupervised learning has gradually developed into the mainstream method. A large number of scene classification methods based on unsupervised learning have appeared [33]–[36] and substantial progress has been made in scene classification. Compared with manual feature extraction methods, unsupervised learning methods are more efficient and flexible. However, due to the lack of label information, the unsupervised method’s convergence direction during training is uncertain. It easily converges to the local optimum, it is difficult to fit a good model, and robustness cannot be guaranteed. Therefore, unsupervised learning still cannot be reliably applied to remote sensing scene image classification tasks.

Remote scene classification methods based on deep learning: In recent years, with the outstanding achievements of convolutional neural networks (CNNs) in computer vision tasks, researchers have proposed many scene classification methods based on CNNs [23], [37]–[42]. Compared with traditional methods, CNN has an unprecedented superior performance in scene classification due to its powerful feature extraction capabilities.

Although the current CNN-based scene classification method has achieved unprecedented results, from the perspective of algorithmic principles, the following points are worthy of further discussion:

(1) First, most of the current methods use an end-to-end CNN architecture to extract edge combination features layer by layer [1], [43]–[46]. This deep feature represents the global information of the remote sensing scene image. In this case, the local feature information of a specific area related to the class is easily disturbed by redundant and irrelevant information, which leads to classification errors. In addition, most of the current methods based on CNN only use the final deep features for scene classification. They do not reasonably introduce shallow local information, which makes the model not having strong generalization performance for remote sensing scene images of different sizes and resolutions.

(2) Second, due to the characteristics of large intraclass diversity and high interclass similarity of remote sensing images, it is difficult for CNN-based methods to distinguish images of similar scenes [3]. Specifically, for some remote sensing images of the same class, the global information is completely different, and the regions for determining the categories of these images are similar. For some remote sensing images of different classes, images and global information are roughly the same, but the regions that determine the classes of these images are different. This means that the existing methods using the CNN architecture are likely to classify different classes of images with the same global information into one class and classify the same class of images with different global information into different classes.

Therefore, researchers try to introduce the attention mechanism into CNNs to improve the performance of feature representation. The attention mechanism is essentially derived from the human visual attention mechanism and has shown good performance in various computer vision tasks, such as image classification, semantic segmentation, and object detection tasks. However, the existing channel attention mechanism usually uses the FC (fully connected) [2], [40], [47] layer to learn the interrelationship between different channels. The large number of parameters in the FC layer can easily lead to overfitting of the model. In addition, how to embed the attention mechanism into the CNN and which layers of the CNN also have a certain impact on the performance of the model. Most of the current work is to directly embed the attention mechanism into some layers of CNN according to expert experience, without considering whether it is reasonable.

To this end, we propose a remote sensing scene classification model that combines lightweight channel attention and multiscale feature fusion discrimination to improve remote sensing scene classification performance. The main contributions of this work are as follows:

1) A new lightweight channel attention mechanism is constructed: The existing channel attention-based methods usually use the FC layer to learn the relationship between different channels, thereby increasing the computational complexity of the model; moreover, the embedding method is not considered, resulting in limited feature representation performance. For this reason, for large-scale remote sensing images, we have constructed a new lightweight channel attention mechanism to balance the accuracy and calculation speed. The traditional FC layer was replaced with a convolutional layer to reduce the attention calculation complexity, and the shortcut connection module was redesigned. Furthermore, we introduced batch normalization in the shortcut
connection module to accelerate the network’s convergence speed and reduce overfitting. In addition, from the CNN mechanism, we embed lightweight attention into the CNN’s appropriate position to maximize the feature representation performance.

2) Multiscale feature fusion discrimination strategy: To solve the existing hierarchical CNN architecture, only the final global feature information is used for discrimination, and the shallow local information is discarded, resulting in the lack of strong generalization ability for remote sensing images of different resolutions. At present, many scholars introduce multiscale feature fusion to remote sensing scene classification. Most of these methods scale different feature maps to the same size after convolution. It is sent to the classifier for classification, and the added feature map contains both shallow and deep information, which effectively improves the performance of scene classification. However, this kind of multi-scale feature fusion only performs classification on a feature map of one scale not only increases the amount of calculation, but also makes the model unable to directly supervise the update of shallow convolutional layer parameters through backpropagation. To this end, we designed a multiscale feature fusion discrimination strategy. A multibranch design is carried out for different resolution features to fully integrate different scale feature information to improve the model’s multiscale feature representation ability. In addition, to reduce the computational complexity of multiple branches, we did not use the multilayer convolution operation used in existing work but directly used global average pooling in each branch to refine features of different scales. Further, information fusion is carried out in the softmax classifier to improve the discrimination ability in a balanced manner.

3) Cross-entropy loss function based on label smoothing: To solve the problem of high similarity between classes in remote sensing scene image classification tasks, we introduce the label smoothing cross-entropy loss function to reduce the influence of the similarity between remote sensing images on the feature representation and to guide the network to learn significant features with class differences.

II. RELATED WORK

A. REMOTE SENSING SCENE CLASSIFICATION

Due to the continuous development of scene classification in practical applications, it has attracted widespread attention in recent years. The remote sensing scene classification method has roughly experienced the following three periods:

Early methods based on handcrafted features: The method based on handcrafted features focuses on designing local feature descriptors to extract features of remote sensing scenes. Examples include scale-invariant feature transform (SIFT) [48], texture descriptor (TD) [49]–[51], color histogram (CH) [52], histogram of oriented gradients (HOG) and GIST [53]. However, because handmade features are a single type feature, these features cannot fully express all the information of a complex image without considering the details of the remote sensing data, which greatly limits the model to effectively extract category-related information. Scholars have proposed several combinations of handcrafted features to overcome the shortcomings of a single type of handcrafted feature [54]. Although the combined features overcome the shortcomings of a single manual feature to a certain extent, the method’s performance based on manual features is still not ideal.

Scene classification method based on an unsupervised feature extractor: To solve the problem of a single feature extracted by a manual feature extractor and relying too much on expert experience, scholars have proposed many methods based on unsupervised features. Examples include principal component analysis (PCA) [55], k-means clustering [56], sparse coding [57] and autoencoders [58]. The unsupervised feature extraction method learns a function to represent the hidden information of the image through a small amount of labeled raw input. However, due to the lack of labeled information, it is difficult for methods based on unsupervised features to capture class-specific information.

Scene classification method based on deep learning: CNN has become the most common methods for remote sensing scene classification owing to its excellent ability to extract feature. The powerful feature extraction capabilities of CNNs make CNN-based methods much better than traditional manual methods. Since 2015, many researchers [62], [76]–[78] have proposed a CNN-based remote sensing scene classification method. He et al. [59] proposed a multi-layer superimposed covariance pool (MSCP). In this method, they use a pretrained network model to extract multilayer convolutional feature maps and superimpose these feature maps. Wang et al. [22] proposed a Multi-Granularity Canonical Appearance Pooling network, which used a granular framework to progressively cropping the input image to learn multi-grained features and designed a Gaussian covariance matrices replace the CNN features for improving the discriminative power of features. Bi et al. [60] proposed an attention pooling-based convolutional network. As far as we know this is the first time shallow features were used for model supervision training, thereby improving the classification accuracy. In recent years, few shot learning has gradually been applied to the field of remote sensing scene classification. Few shot learning aims to solve the problem of how the model can be better generalized when there are too few sample. In order to make the model have better generalization performance when there are fewer training data sets or new categories in the data set, Li et al. [61] proposed a novel ZSRSSC approach, the proposed named LPDCMENs, which can fully assimilate the pairwise intramodal and intermodal supervision in an end-to-end manner. Han et al. [62] proposed a DLA-MatchNet, which use channel attention and spatial attention modules combined different feature fusion schemes, meanwhile utilize an adaptive matcher to measure similarity scores among samples can adaptively deal with the problems of intraclass difference and interclass similarity.
B. ATTENTION MECHANISM

For understand an image, human can quickly scan the global area to obtain the target area that needs to be focused on and then invest more attention resources in this area to obtain more detailed information about the target that needs to be focused on while suppressing other useless information. The attention mechanism in CNN understands and perceives images in a way that simulates humans and differentially weights global features to highlight key local features. In recent years, many channel attention [4], [6], [63]–[66] mechanisms have been successfully applied to different computer vision tasks such as image classification, semantic segmentation, object detection, and image translation. Hu et al. [2] proposed performing differential weighting on different feature channels using the SE module. Fu et al. [6] proposed a dual attention network, contain a position attention module and a channel attention module, which position attention learn the spatial interdependencies of features and channel attention module to learn channel interdependencies. Tang et al. [66] proposed a multi-channel attention selection, which can automatically select from a set of intermediate diverse generations in a larger generation space to improve the generation quality. Woo et al. [67] combined the channel and spatial attention mechanism to perform differential weighting on the channel and spatial features at the same time to enhance the representation ability of CNN. Haut et al. [64] proposed a deep residual channel attention that can focus on those features extracted from land-cover components. Tong et al. [40] introduced the SE module into the model, using channel attention to enhance the information interaction between channels and improving the model’s ability to extract features. Studies have shown that introducing an attention mechanism into CNNs can effectively improve the performance of scene classification.

C. MULTISCALE FEATURE FUSION

Currently, most of the remote sensing scene classification networks based on CNN only classify on the last layer of feature map, the last feature layer extracts the semantic information of the image, and the shallow edge information of the image does not directly participate in the classification. In the process of image processing, CNNs down-sampling the feature map through the pooling layer, which leads to the loss of fine-grained information related to the category. In order to solve this problem, many researchers introduce multi-scale feature fusion strategy, which combines shallow edge information with deep semantic information to improve classification accuracy. Yang et al. [68] proposed an Enhanced Multiscale Feature Fusion Network, which combined task-wise attention and part complementary learning to extract and fuse the features, and utilize PMN to blend large-scale, middle-scale, and small-scale features in parallel during the three stages, make feature fusion can occur at different scales in all stages. Qu et al. [69] on the basis of FPN, proposed a multiscale convolutional feature fusion. Huang et al. [70] proposed a denoising-based multiscale feature fusion mechanism, the proposed DMSFF mechanism can aggregates multiscale features with the denoising operation at the stage of visual feature extraction.

D. LOSS FUNCTION

The loss function optimizes the model by continuously reducing the gap between the predicted data and the actual data. To better optimize the model, some scholars proposed many powerful loss functions to encourage the model to learn more discriminative features. Ye et al. [71] proposed combine the cross-entropy loss function with the central loss function to simultaneously optimize model. Wei et al. [72] to enlarge the gap between from the similar scene category designed a marginal center loss, which effectively solves the problem of large intraclass differences in remote sensing scene datasets. Cheng et al. [14] introduced metric learning in CNN, all data is converted into a new feature space, so that the distance between the feature vectors of similar samples is small, and the distance between feature vectors of dissimilar samples is large, which can effectively improve scene classification accuracy.

III. MATH

A. OVERALL ARCHITECTURE

The proposed LmNet framework is shown in Fig. 1. The network is mainly composed of four parts:

1. Backbone network for feature extraction: We first use ResNext50 as the feature extraction network, which has been initialized with the weights trained on ImageNet. Previous research [8] showed that the use of pretraining weights for migration learning is not only conducive to the rapid convergence of the network but can also effectively improve the classification performance. It needs to be emphasized that our method can be flexibly embedded in any advanced backbone. In this paper, we use ResNext50, which is commonly used in remote sensing scene classification, as the backbone to facilitate performance comparison with other methods.

2. Lightweight channel attention mechanism: the features extracted by the hierarchical CNN network usually ignore the differences in different channels’ features. To this end, we designed a new lightweight channel attention module to distinguish and weight channel features to improve the role of key features in classification.

3. Multiscale feature fusion discrimination mechanism: Existing methods usually only use the last layer of semantic feature modeling for supervised training, and the shallow edge features are discarded. However, this shallow edge information has certain positive significance for characterizing scene attributes. To take advantage of the feature information of different scales, we fuse the features of the three scales at the same time and learn the multiscale feature information with the supervision model to enhance the multiscale feature recognition ability.

4. Definition of loss function: The high similarity between classes in the remote sensing scene dataset is likely to be misclassified. The traditional cross-entropy loss function only
calculation the loss of the predicted value is the real class. As a result, images with similar features are easily classified incorrectly. Therefore, we introduce label smoothing in the cross-entropy loss function to reduce the similarity between classes.

To achieve remote sensing scene classification, we first need to build a backbone network for feature extraction. The backbone network is only used as a basic module, and we can choose any advanced backbone. Related research works show that the backbone currently used in remote sensing scene classification tasks mainly includes ResNet50, DenseNet121, and ResNeXt50. Through experiments, we compared the three most commonly used backbone networks in object classification, ResNet50, DenseNet121, and ResNeXt50.

In the early stage, we studied and analyzed the performance of the above three backbones in detail. DenseNet121 uses dense connections, each front layer function is used as the input of the latter layer, and the classification performance is better than ResNet50. However, compared with ResNeXt50, the performance is not much improved, but the model inference speed of ResNeXt50 is faster. Therefore, considering the balance between speed and accuracy, we choose ResNeXt50 as the backbone network for feature extraction. It should be noted that the method proposed in this paper can be seamlessly embedded in other advanced backbone networks to further improve the classification performance.

ResNeXt was developed on the basis of ResNet. ResNeXt uses grouped convolution to divide each bottleneck into 32 cardinalities, which reduces parameters and is more conducive to feature extraction. The structural information of ResNeXt50 is shown in Table 1. We divide it into six blocks. The first block contains a convolution kernel with a size of $7 \times 7$ and a step size of 2 for the first feature extraction of the input image, and then a maxpool layer with a convolution kernel size of $3 \times 3$ and a step size of 2 performs the first feature downsampling. The second, third, fourth, and fifth blocks have similar structures and are composed of 3, 4, 6, and 3 bottlenecks, respectively. Each bottleneck is composed of 32 identical cardinalities. Each cardinality contains two $1 \times 1$ convolutional layers and one $3 \times 3$ convolutional layer (as shown in Fig. 2), and the bottleneck is connected by residuals. Between the second, third, fourth, and fifth blocks, the downsampling operation reduces the spatial information of the feature to twice the original. The sixth part is a classification layer composed of an average pooling layer and a fully connected layer.
Table 1: Architecture of ResNeXt50.

| Stage | Output Size | ResNetXt |
|-------|-------------|----------|
| Block1 | 112×112 | 7×7, 64, stride=2 3×3, maxpooling, stride=2 |
| Block2 | 56×56 | 1×1,128 3×3,128, C=32 ×3 1×1,128 |
| Block3 | 28×28 | 1×1,256 3×3,256, C=32 ×4 1×1,256 |
| Block4 | 14×14 | 1×1,1024 3×3,1024, C=32 ×6 1×1,1024 |
| Block5 | 7×7 | 1×1,2048 3×3,2048, C=32 ×3 1×1,2048 |
| Classification | 1×1 | Global average pooling, Fully connected |

Figure 2: Bottleneck of ResNeXt.

B. LIGHTWEIGHT CHANNEL ATTENTION

The backbone network (ResNeXt50) is a hierarchical deep architecture and has powerful feature extraction capabilities in the spatial domain. However, it ignores the information interaction between different channels, which leads to a lack of full utilization of features that are essential for classification. To this end, researchers have proposed the use of channel attention strategies. However, the existing channel attention module uses the FC module for feature weighting, which increases the computational complexity of the model and does not fully extract the salient features; in addition, how to embed the channel attention into the backbone requires in-depth consideration.

Therefore, it is mainly based on three considerations. The first is the extraction of salient features at the channel level, the second is the lightweight, and the second is the appropriate embedding position. We designed a new lightweight channel attention mechanism. The lightweight channel attention mechanism can adaptively adjust the weights of different channel feature maps.

As shown in Fig. 3, we designed a lightweight channel attention mechanism based on the SE module [2]. The SE module uses two FC layers to learn the relationship between different channels after feature global pooling, but a large number of parameters in the fully connected layer not only affects the inference time of the model but also may cause the risk of overfitting. Therefore, to reduce the number of calculations, we used two convolution kernels with a size of 1 × 1 instead of the fully connected layer. After each convolutional layer, a batch normalization layer was used to normalize the input value to readjust the data distribution, which ensured the effectiveness of the gradient in the training process. For the activation function we used ReLU and hard-sigmoid function. Hard-sigmoid has less delay cost than sigmoid function [73].

Similar to the SE module, the lightweight channel attention included three stages, respectively are squeeze, excitation and scale. First, the feature map U performed a squeeze operation through global pooling. This operation aggregates the feature maps from the spatial dimension H×W to generate a channel descriptor. The squeeze operation is achieved as follows:

\[ z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \]  

where \( u_c(i, j) \) is the spatial information of the \( c \)-th channel of the feature U at position (i, j), \( F_{sq}(u_c) \) refers to the squeeze operation, and \( z_c \) is the channel descriptor.

The squeeze operation is followed by the excitation operation. The excitation operation in the SE module uses two fully connected layers to obtain the relationship between the channels. We use two convolutional layers with a kernel size of 1 × 1 instead of the fully connected layer to parameterize the attention mechanism. The first convolutional layer uses ReLU as the activation function after batch normalization. The second convolutional layer uses hard-sigmoid = \( \text{ReLU}(x) / \text{ReLU}(x+1) \) [73] as the activation function after batch normalization. The calculated of the excitation stage is as follows:

\[ s = F_{ex}(z, W) = \sigma(B(W_2B(\delta(W_1z)))) \]  

where z is a one-dimensional vector after the squeeze operation, \( W_1 \) and \( W_2 \) are the weights of two convolution layers, B refers to the batch normalization, \( \delta \) and \( \sigma \) refers to the activation function of ReLU and hard-sigmoid, respectively. \( F_{ex} \) refers to the excitation stage.

Finally, the feature map U is reweighted to generate the output of the channel attention. The reweighted features are directly input into the subsequent layers:

\[ \bar{x} = F_{scale}(u, s) = s_c \cdot u_c \]  

where \( s_c \) refers to the reweighted value of the \( c \)-th channel of the feature maps U, \( F_{scale} \) refers to channel-wise multiplication between the scalar \( s_c \) and the feature map \( u_c \).
The lightweight channel attention module is a dynamic feature extraction mechanism. It weighs the channel to pay more attention to the feature mapping that helps classification. Therefore, adding the channel attention module to the CNN can effectively improve the network’s presentation ability. However, in the experiment, we found that the different channel attention positions significantly impact classification accuracy, even a negative impact. Through the mechanism analysis of CNNs, shallow features usually extract local or edge information, and deep features extract overall or semantic information and construct the complete shape of the target. Therefore, we have redesigned how the attention mechanism is embedded in the backbone network. As shown in Fig. 1, in our network, we use the lightweight channel attention module in each block (Fig. 3). The first and second block layers are added to the channel attention module and passed to the following block. The third, fourth, and fifth block layers use the SE module to perform classification directly through average pooling and no longer pass to the following layers.

C. MULTISCALE FEATURE FUSION DISCRIMINATION

To improve the model’s ability to discriminate multisresolution and multiscale remote sensing scene images, we designed a multiscale feature fusion discrimination strategy, which makes full use of shallow edge features and deep semantic information. As shown in Fig. 1, after the third, fourth and fifth block blocks, a lightweight channel attention module is added for feature enhancement, and then the feature space domain resolution is compressed to a size of $1 \times 1$ through an adaptive average pooling layer. Finally, the feature is converted into a probability distribution through the fully connected layer. We add these three predictions and fuse them together to generate the final prediction. In this case, features from different levels generate predictions and supervise the model training process. The formula is as follows:

$$\hat{y}_k = Fc \left( \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \tilde{x}^k(i, j) \right)$$  \hspace{0.5cm} (4)

where $\tilde{x}^k$ represents the feature of the k-th scale after the attention module operation, $Fc$ represents the fully connected layer, and $\hat{y}_k$ represents the predicted value generated by the k-th scale.

Then, add the predicted values generated by each scale:

$$\hat{y}_i = \text{soft max} \sum_{k=1}^{3} \hat{y}^k_i$$ \hspace{0.5cm} (5)

D. CROSS-ENTROPY LOSS FUNCTION BASED ON LABEL SMOOTHING

A large part of scene classification errors is caused by the similar images in the dataset. The similarity between categories means that they belong to different categories but have similar characteristics. We introduce the label smoothing in cross-entropy loss to optimize the model for enlarge the gap of interclass similarity.
Image classification usually uses a softmax function after the FC layer to calculate the probability of the image belonging to a certain category, and then input the probability to the cross-entropy function to calculate the loss between the true targets $y_i$ and the network’s outputs targets $p_i$. The class vector before input the network is usually converted into an array of length $n$ ($n$ refers to the number of categories); where $y_i$ is “1” for the correct class and “0” for the rest. This characteristic encourage enlarge the gap between the real label and other labels, which means model will promote learning in the direction with the largest difference between the correct label and the wrong label. When the training data is small and between similar categories is relatively small, it will cause the network to overfitting. The traditional softmax formula is as follows:

$$p_i = \frac{e^{x^T w_i}}{\sum_{i=1}^{L} e^{x^T w_i}}$$  \hspace{1cm} (6)

where $p_i$ is the probability of the model prediction class $i_{th}$, $w_i$ refers to the weight and bias of the last layer, and $x$ is the probability value of model prediction from the input data.

The cross-entropy loss is calculated by backpropagation and minimizes the expected value between the actual target and the network predicted target.

$$H(y, p) = -\sum_{i=1}^{I} y_i \log(p_i)$$  \hspace{1cm} (7)

where $y_i$ represent true targets and its one-hot label is “1”, and the rest are “0”. $p_i$ represent network’s outputs value, his range is between 0 to 1. The cross-entropy loss function only calculation the loss of the correct label position, which causes the model to increase the probability of predicting the correct label. Instead of reduce the probability of predicting the error label. It would result in model fits very well on training set, but on other test sets do not perform well. Especially there are a lot of similar data in the scene data set, overfitting is more likely to occur.

Label smoothing is a regularization strategy, which is mainly to add noise through soft one-hot, which reduces the weight of the category of the real label when calculating the loss function, and finally suppresses the effect of overfitting. Therefore, we introduce label smoothing [74] into cross-entropy loss:

$$y' = \begin{cases} 
1 - \varepsilon & \text{if } i = y \\
\varepsilon / (K - 1) & \text{otherwise}
\end{cases}$$  \hspace{1cm} (8)

where $y'$ refers to label after adds the label smoothing, $\varepsilon$ is a small constant, $1/(K - 1)$ can be regarded as the introduction of fixed distribution noise into the probability distribution. Therefore, the cross-entropy loss of adds the label smoothing not only calculation the loss of the correct label position, but also slightly considers other wrong label positions, which leads to an increase in the learning ability of the model. Forcing the model to increase the probability of correct classification and at the same time reduce the probability of incorrect classification.

### IV. EXPERIMENTAL RESULTS

In this section, we use three publicly available remote-sensing image datasets to implement a series of experiments to verification the performance of our proposed method. The following details the experimental settings, experimental results, and analysis of the results. The operating environment of all experiments in this paper is one experimental computer that runs the Ubuntu system with an Intel (R) Core (TM) i5-7500 CPU @ 3.40 GHz and an NVIDIA GeForce GTX2080Ti with 11GB memory.

#### A. EXPERIMENTAL DATASETS

1. The UC Merced dataset [75], released in 2010, contains 21 land use classes. The dataset total have 2100 images, each of the classes includes 100 land-use images and the image resolution is 256 × 256 pixels, the pixel resolution is 0.3 m. The 21 land categories are agricultural land, airplanes, baseball diamonds, beaches, buildings, squares, dense residential areas, forests, highways, golf courses, ports, intersections, medium-sized residential areas, mobile home parks, overpasses, parking lots, rivers, runways, sparse residential areas, storage tanks, and tennis courts.

2. AID dataset [8]. The AID dataset was collected from Google Earth imagery, was released by Wuhan University in 2017 and include 30 scene categories and a total of 10,000 images. Each classes consists of 220 to 420 images, the image resolution is 600 × 600 pixels, and the resolution of image is between 8 and 0.5 meters. The images in AID are actually multisource than UC Merced dataset. The 30 classes are airports, bare land, baseball fields, beaches, bridges, centers, churches, commercial areas, dense residential areas, deserts, farmland, forests, industrial areas, grasslands, medium-sized residential areas, mountainous areas, parks, parking lots, playgrounds, ponds, ports, railway stations, resorts, rivers, schools, sparse residential areas, squares, stadiums, storage tanks, and viaducts.

3. The NWPU-RESISC45 dataset [1]. This dataset include 45 categories, each category contains 700 images, covers more than 100 countries and contains 31,500 images. The size of the image is 256 × 256 pixels. Except for some specific categories with a low spatial resolution (such as islands, lakes, mountains, and icebergs), most scene categories’ pixel resolution ranged from 30 meters to 0.2 meters. The 45 classes are airplanes, airports, baseball diamonds, basketball courts, beaches, bridges, churches, circular farmland, clouds, commercial areas, dense residential areas, deserts, forests, highways, golf courses, ground track fields, ports, industry districts, intersections, islands, lakes, grasslands, medium-sized residential areas, mobile home parks, mountains, overpasses, palaces, parking lots, railways, train stations, rectangular farmland, rivers, circular runways, runways, sea ice, ships, snow castles, sparse residential areas, stadiums, storage tanks, tennis courts, terraces, thermal power...
stations, and wetlands. NWPU-RESISC45 as a big dataset has the following three notable characteristics: large range of scale changes, rich image variations, high interclass diversity and interclass similarity.

B. EXPERIMENTAL DETAILS
To evaluate the performance of our method we used the overall accuracy (OA) and confusion matrix to show the Classification result. The overall accuracy is the ratio between the model’s correct predictions and the number of images on all test sets. The x-axis of the confusion matrix represents the predicted category, and the y-axis represents the actual category. We repeated the experiment ten times and each experiment randomly dividing the dataset and calculated the mean and standard deviation of the precision ten times as the final classification result.

To alleviate the problem that the model is prone to overfitting when the training set sample size is too small, we used a data augmentation strategy. As follows: First, adjust the training image to $288 \times 288$. Then, the image is randomly sampled with a random rotation of $270^\circ$ and a random horizontal mirror image. Finally, the image is randomly cropped to $288 \times 288$ by filling the image with 10 pixels.

We used the PyTorch framework to implement our model, and the input size of the images was resize $288 \times 288$ pixels to model. The model was optimized using the stochastic gradient descent (SGD) algorithm, with a weight decay factor of 0.0001, a momentum of 0.9, and a batch size of 32. The initial learning rate is 0.01 (using the weight of pre-training ResNext50 as the initial value), and after every ten epochs, the learning rate was reduced by 0.5 times. A total of 100 epochs models are trained.

C. EXPERIMENTAL RESULTS
1) CLASSIFICATION RESULTS OF THE UC MERCED DATASET
The UC Merced dataset is the simplest dataset compared with the other two datasets, however its interclass similarity is more obvious. The performance comparison between LmNet and the state-of-the-art method on the UC Merced dataset is shown in Table 2. When the training ratio was set to 80%, the overall accuracy of many methods was close to 99%. Among them, CNN-CapsNet and ARCNet have the highest classification accuracy of 99.05% and 99.12%, respectively, which is close to the classification accuracy of our method of 99.52%. However, when the training ratio was set to 50%, most classification methods were below 97%. Among them, CNN-CapsNet and ARCNet have the highest classification accuracies of 97.59% and 96.81%, respectively. The classification accuracy of 98.57% lower than our method is close to 1% and 1.8%. This shows that our network can still extract more discriminative features even with a small amount of data.

Fig. 4 the confusion matrix shows LmNet on the UC Merced dataset classification result. When the training ratio was set to 50%, the classification accuracy of 17 out of 21 classes was higher than 98%. The category with the lowest accuracy was dense residential areas (0.86). The probability that the dense residential area is mistaken for buildings and medium residential areas is the largest, which is 0.04. Dense residential areas and buildings have similar characteristics. The main difference is the density of buildings in the image. Runways mistaken for freeways are also 0.04 because the main characteristics of runways and freeways are the same. When the training ratio is set to 80%, 20 of the 21 categories have a classification accuracy of 100%. Among them, dense residential areas have a probability of 0.05 being mistaken for buildings and medium residential areas. The characteristics of these three categories are similar and easy to mistake. Therefore, how to improve the model’s ability to discriminate fine-grained features is still the focus of remote sensing scene classification.

2) CLASSIFICATION RESULTS OF THE AID DATASET
For the AID dataset, we set the training ratio 20% and 50%. Table 3 shows the performance of LmNet and the
most advanced method. When the training ratio is set to 50%, the method with the highest classification accuracy is SE-MDPMNet with 97.14%, and our method with 97.12% is slightly lower than SE-MDPMNet. However, when the training setting ratio is reduced to 20%, the classification accuracy of our method is 95.82%, which is 1.14% higher than SE-MDPMNet’s 94.86%. This shows that our method can also obtain better classification results when the amount of data is small.

Fig. 5 confusion matrix shows the result of LmNet on the AID dataset. When the training ratio is set to 20%, as shown in Figure (a), 26 of the 30 categories have an accuracy of more than 92%. Only the four categories of Resort (82%), School (83%), Ceter (81%), and Square (89%) do not exceed 90%. Ceter and Resort are incorrectly classified as Square and Park with a maximum probability of 0.08 because they have similar building distributions. In addition, the probability of school being misclassified as commercial is 0.07 because they have similar housing distributions. As shown in Figure (b) the training ratio is set 50%, the accuracy of the three categories of School (0.89), Ceter (0.96), and Square (0.92) increases by 0.05, 0.15, and 0.03 compared to the 20% training set. The probability of the school being misclassified as commercial is reduced to 0.06. Generally, as the training dataset increases, the accuracy of image classification improves.

3) CLASSIFICATION RESULTS OF NWPU-RESISC45 DATASET

Table 4 shows the performance of our method and the state-of-the-art method in the NWPU-RESISC45 dataset. For this dataset, LmNet also achieved significant performance. When the training ratio was set to 10% and 20%, its overall accuracy reached 93.00% and 94.85%, respectively, higher than SE-MDPMNet. When the training ratio was reduced from 20% to 10%, the classification accuracy of most models was reduced by more than 2.5%, and our model was only reduced by 1.85%, which shows that our method has better
generalization performance under the condition of reducing the amount of training data.

Fig. 6 the confusion matrix shows LmNet on the NWPU-RESISC45 dataset classification result, the training ratio is 10% and 20%. With set 10% training ratio, the lowest accuracy rates are Church and Palace (72% and 73%, respectively), and the accuracies of the other categories are all higher than 85%. Because Church and Palace have similar architectural styles, their probability of being misclassified is higher than 10%. When the training ratio increased to 20%, the Church and Palace’s accuracy rates reached 79% and 81%, respectively. These two categories account for the largest proportion of classification errors. How to effectively improve the classification accuracy of these two categories is the key to improving the overall classification accuracy of the NWPU-RESISC45 dataset.

In summary, a comparison of most methods through three experiments shows that our proposed model has better classification performance.

In particular, observe Tables 2, 3, and 4 to adjust different training ratios. We found that when the training ratio decreases, the classification performance of most methods decreases more than our method. This is a positive result for large-scale scene classification, indicating that our method is less dependent on label data and can greatly reduce the cost of manual labeling. Moreover, while the amount of label training data is reduced, our model can still maintain a high classification accuracy, indicating that our model has good robustness, which is extremely critical for actual scene applications.

Through the analysis of Figs. 4, 5, and 6, we find that images with similar features have the highest probability of being classified incorrectly. Therefore, how to improve the model’s ability to identify fine-grained features is the key to improving the accuracy of remote sensing scene classification.

V. DISCUSSION

A. ABLATION EXPERIMENT

To verify the effectiveness of each module of LmNet, we conducted ablation experiments on the AID dataset. We performed four sets of ablation experiments, including ResNext, ResNext + multiscale fusion discrimination, ResNext + multiscale fusion discrimination + channel attention, and ResNext + multiscale fusion discrimination + channel attention + loss function, to verify the performance of each module. As shown in Table 5, the module’s classification performance with multiscale feature fusion discrimination is improved by 0.23% compared with ResNext, and the classification performance of the module with multiscale feature fusion discrimination and channel attention module is increased by 0.43% compared with ResNext. The classification performance of LmNet is 0.58% higher than that of ResNext. Experiments show that each module we designed can effectively improve the classification performance of the model.
TABLE 5. Ablation study of the LmNet on the AID dataset.

|                      | ResNext | ResNext + Multiscale | ResNext + Multiscale + ChannelAttention | LmNet     |
|----------------------|---------|----------------------|----------------------------------------|-----------|
| AID (Train ratio=50%)| 96.47±0.13 | 96.81±0.07 | 96.97±0.17 | 97.12±0.14 |
| AID (Train ratio=20%)| 95.31±0.11 | 95.44±0.09 | 95.63±0.13 | 95.82±0.25 |
| Average              | 95.89    | 96.12    | 96.30     | 96.47     |

TABLE 6. Inference time of model (LCA refers to lightweight channel attention).

|                      | UCM      | AID      | NWPU-RESISC45 | Average |
|----------------------|----------|----------|---------------|---------|
| ResNext              | 0.3205 ms| 0.2922 ms| 0.3162 ms     | 0.3036 ms|
| ResNext+SE           | 0.3463 ms| 0.3285 ms| 0.3412 ms     | 0.3387 ms|
| ResNext + LCA        | 0.3316 ms| 0.3184 ms| 0.3311 ms     | 0.3270 ms|
| LmNet                | 0.3779 ms| 0.3550 ms| 0.3758 ms     | 0.3696 ms|

TABLE 7. Different channel attention experiment (LCA refers to lightweight channel attention).

|                      | AID (Train ratio=50%) | AID (Train ratio=20%) | Average inference time |
|----------------------|-----------------------|-----------------------|------------------------|
| LmNet-LCA+SE[2]      | 96.78±0.21            | 95.34±0.11            | 0.3805 ms              |
| LmNet-LCA+CA[4]      | 96.93±0.27            | 95.57±0.15            | 0.3671 ms              |
| LmNet-LCA+ACM[6]     | 96.38±0.13            | 95.01±0.21            | 0.7512 ms              |
| LmNet-LCA+PA[9]      | 97.05±0.21            | 95.80±0.17            | 0.4101 ms              |
| LmNet(ours)          | 97.12±0.14            | 95.82±0.25            | 0.3696 ms              |

B. MODEL INFERENCE TIME ANALYSIS

With the need for model deployment on the mobile terminal, very high requirements are placed on the model inference time. To verify the efficiency of our LmNet inference time, we performed two sets of experiments on the UCM, AID, and NWPU-RESISC45 datasets, namely, the inference times of ResNeXt, ResNext + SE, ResNext + channel Attention and LmNet. To eliminate the influence of randomness on the experimental results, each group of experiments was performed ten times, and the average value was taken as the final result. The experimental results are shown in Table 6. The average results on these three datasets show that ResNext’s single image inference time is 0.3036ms, ResNext + SE inference time is 0.3387ms, ResNext + LCA is 0.3270ms and LmNet’s inference time is 0.3696ms. Our channel attention is reduced by 0.011ms compared with traditional SE.

C. CONTRASTIVE EXPERIMENTAL OF DIFFERENT CHANNEL ATTENTION

In order to compare different channel attention to verify the effectiveness of our proposed lightweight channel attention, as shown in Table 7, we conducted five sets of experiments on the AID data set. We only replace the lightweight channel attention module in the LmNet network with SE, CA, ACM, and PA modules without changing the insertion position of the channel attention. Experiments show our lightweight attention network has good performance not only in time but also in accuracy. According to our experiments, we found that the classification performance of SE, CA, ACM and PA modules is very different when they are inserted into the network in different locations. Therefore, further research is needed to achieve the best classification performance when different channel attention are inserted into different locations of the network.

VI. CONCLUSION

We propose a new remote sensing scene classification network, namely, LmNet. First, a new lightweight channel attention module is designed to quickly learn the interaction between channels to distinguish the contribution of different channels. In addition, we designed a multiscale fusion discrimination strategy to fully integrate shallow edge information and deep semantic features to improve the multiscale feature representation ability of the model. Finally, to reduce the impact of similarity between classes on classification performance, we defined a loss function to guide network training, which combined label smoothing and cross-entropy functions. In particular, our proposed lightweight channel attention and multiscale feature discrimination strategy can be flexibly embedded in any advanced backbone and has good applicability.

Experimental results on three large-scale datasets of UCMerced, NWPU-RESISC45, and AID show that our methods all achieve state-of-the-arts. Moreover, in different training sample ratios, our method can still achieve the best performance and shows better generalization performance. At the same time, the results of ablation experiments show that each embedding module of LmNet can stably improve the classification performance and show better robustness. In addition, the statistical results of our method and the classic backbone on the inference time show that our method not only has a...
significant improvement in accuracy but also does not incur much time consumption. In particular, when the amount of labeled training data is reduced, our method still has the best recognition accuracy. This result shows that our method is less dependent on labeled training data, which can effectively reduce the cost of manual data annotation. This is a very meaningful research result for the practical application of large-scale scene classification tasks.

In the future, we will study efficient neural network automatic compression strategies to further improve the inference speed and promote practical applications. In addition, the current transformer-based network architecture shows an emerging trend in the field of computer vision. Its core idea is the block processing mechanism, which can be introduced into large-scene remote sensing image classification tasks for tentative exploration.
T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and
A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic
M. J. Swain and D. H. Ballard, “Color indexing,” Int. J. Comput. Vis.
A. K. Jain, N. K. Ratha, and S. Lakshmanan, “Object detection using
R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for
d. Yu, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image
S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual
K. Jain, N. K. Ratha, and S. Lakshmanan, “Object detection using
A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with
depth convolutional networks for very deep residual neural networks,”
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image
K. Simonny and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” pp. 1–16, 2015, arXiv:1409.1556v6.
X. He, Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, “A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 2021, doi: 10.1109/TNNLS.2021.3062070.
W. Huang, Q. Wang, and X. Li, “Denoising-based multiscale feature fusion for remote sensing image captioning,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 3, pp. 436–440, Mar. 2021.
Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 286–301.
H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, “Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2412–2421, Jun. 2019.
S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.
J. Yang, C. Wu, B. Du, and L. Zhang, “Enhanced multiscale feature fusion network for HSI classification,” IEEE Trans. Geosci. Remote Sens., early access, Jan. 13, 2021, doi: 10.1109/TGRS.2020.3046757.
Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, “A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 2021, doi: 10.1109/TNNLS.2021.3062070.
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. 2019.
Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 286–301.
H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, “Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2412–2421, Jun. 2019.
S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.
J. Yang, C. Wu, B. Du, and L. Zhang, “Enhanced multiscale feature fusion network for HSI classification,” IEEE Trans. Geosci. Remote Sens., early access, Jan. 13, 2021, doi: 10.1109/TGRS.2020.3046757.
Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, “A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 2021, doi: 10.1109/TNNLS.2021.3062070.
W. Huang, Q. Wang, and X. Li, “Denoising-based multiscale feature fusion for remote sensing image captioning,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 3, pp. 436–440, Mar. 2021.
L. Ye, L. Wang, Y. Sun, R. Zhu, and Y. Wei, “Aerial scene classification via an ensemble extreme learning machine classifier based on discriminative hybrid convolutional neural networks features,” Int. J. Remote Sens., vol. 40, no. 7, pp. 2759–2783, Apr. 2019.
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. 2019.
A. Howard, M. Sandler, B. Chen, W. L. C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, “Searching for MobileNetV3,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1314–1324.
R. Müller, S. Kornblith, and G. Hinton, “When does label smoothing help?” pp. 1–13, 2019, arXiv:1906.02629. [Online]. Available: https://arxiv.org/abs/1906.02629.
Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. ACM, 2010, pp. 270–279.
G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, “Remote sensing image scene classification using bag of convolutional features,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
R. Pires de Lima and K. Marfurt, “Convolutional neural network for remote-sensing scene classification: Transfer learning analysis,” Remote Sens., vol. 12, no. 1, p. 86, Dec. 2019.
Y. Li, Y. Zhang, and Z. Zhu, “Error-tolerant deep learning for remote sensing image scene classification,” IEEE Trans. Cybern., vol. 51, no. 4, pp. 1756–1768, Apr. 2021.

HUIYAO WAN was born in Henan, China, in 1995. He received the B.S. degree from the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China, in 2019, where he is currently pursuing the M.S. degree. His research interests include remote-sensing images classification and object detection for SAR image.
JIE CHEN (Member, IEEE) received the Ph.D. degree in computer applied technology from the Hefei Institute of Physical Science, Chinese Academy of Sciences, University of Science and Technology of China, in 2019. He is currently a Lecturer and a Master’s Supervisor with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, China, and also a Postdoctoral Fellow with the 38th Research Institute of China Electronics Technology Group Corporation. His research interests include deep learning, intelligent interpretation of remote sensing images, and autonomous driving. He is a Professional Member of the China Computer Federation.

ZHIXIANG HUANG (Senior Member, IEEE) was born in Anhui, China. He received the B.S. and Ph.D. degrees in computational electromagnetics from Anhui University (AHU), Hefei, China, in 2002 and 2007, respectively. Since 2008, he has been a Full Professor with the School of Electronic Information and Engineering, AHU. From September 2010 to September 2011, he was a Visiting Scholar with Iowa State University, Ames, IA, USA. From August 2013 to October 2013, he was a Visiting Professor with The University of Hong Kong, Hong Kong. From February 2014 to February 2015, he was a Visiting Professor with the Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China. He has co-authored one monograph and three book chapters for CRC Press, Boca Raton, FL, USA, and InTech Publishers, London, U.K. He has co-authored 80 peer-reviewed journal articles included in the Web of Science Core Collection. His current research interests include theoretical and computational research in electromagnetics and image, focusing on the multiphysics and interdisciplinary research, and fundamental and applied aspects in metamaterials and active metamaterials. He is a member of OSA. He was a recipient of the Second Prize of Science and Technology from the Anhui Province Government, China, in 2015. He was also a recipient of the National Science Foundation for Outstanding Young Scholar of China, in 2017.

YUN FENG received the B.S. degree from the School of Electronical and Information Engineering, Anhui Jianzhu University, Anhui, China. She is currently pursuing the M.S. degree with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, Hefei, Anhui. Her research interests include deep learning and remote-sensing images classification.

ZHENG ZHOU received the B.S. degree in electronic information engineering from Anhui University, Hefei, Anhui, China, in 2020, where he is currently pursuing the M.S. degree with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering. His research interests include few-shot object detection, and object detection for SAR image.

XIAOPING LIU received the Ph.D. degree in computer applied technology from the Hefei University of Technology, China, in 2009. From 2009 to 2011, he worked as a Postdoctoral Fellow with the Hefei University of Technology. He is currently a Senior Engineer with the 38th Research Institute of China Electronics Technology Group Corporation, China. His main interests include the system design of synthetic aperture radar and SAR intelligent target recognition.

BAIDONG YAO received the Ph.D. degree in optics from the University of Chinese Academy of Sciences, China, in 2013. He is currently a Senior Engineer with the 38th Research Institute of China Electronics Technology Group Corporation, China. His research interests include SAR system design, SAR remote sensing image processing, and intelligence application technology.

TAO XU received the Ph.D. degree from the Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences, China, in 2016. He is currently a Senior Engineer with the 38th Research Institute of China Electronics Technology Group Corporation, China. His research interests include SAR object detection and recognition, and remote sensing image processing and applications.