Hierarchical Attention and Bilinear Fusion for Remote Sensing Image Scene Classification

Donghang Yu, Haitao Guo, Qing Xu, Jun Lu, Chuan Zhao, and Yuzhun Lin

***Abstract*—Remote sensing image scene classification is an important means for the understanding of remote sensing images. Convolutional neural networks (CNNs) have been successfully applied to remote sensing image scene classification and have demonstrated remarkable performance. However, with improvements in image resolution, remote sensing image categories are becoming increasingly diverse, and problems such as high intraclass diversity and high interclass similarity have arisen. The performance of ordinary CNNs at distinguishing increasingly complex remote sensing images is still limited. Therefore, we propose a feature fusion framework based on hierarchical attention and bilinear pooling called HABFNet for the scene classification of remote sensing images. First, the deep CNN ResNet50 is used to extract the deep features from different layers of the image, and these features are fused to boost their robustness and effectiveness. Second, we design an improved channel attention scheme to enhance the features from different layers. Finally, the enhanced features are cross-layer bilinearly pooled and fused, and the fused features are used for classification. Extensive experiments were conducted on three publicly available remote sensing image benchmarks. Comparisons with the state-of-the-art methods demonstrated that the proposed HABFNet achieved competitive classification performance.**

***Index Terms*—Bilinear pooling, channel attention, hierarchical feature fusion, remote sensing image, scene classification.**

I. Introduction

THE tremendous progress in Earth observation technology has provided a steady stream of remote sensing image data for observing and understanding changes on the Earth's surface [1]. Full utilization of massive remote sensing data to promote effective analysis and understanding has become a popular and critical issue that must be urgently investigated. Remote sensing image scene classification is one of the important components of remote sensing image understanding. Its main task is to assign predefined category information to crops in large-scale remote sensing images [1], such as airports, ports, farmland, or residential areas. The category information is usually determined according to the function of the ground area, so that the same type of remote sensing image scene usually contains multiple types of objects. Unlike pixel-level and object-level information, a remote sensing image scene contains semantic-level information, which has a wide and important application value [1]–[4] in forest and farmland coverage investigation, geological disaster monitoring, object detection and recognition, and urban planning and environment evaluation.

The key to remote sensing image scene classification is to extract the discriminative features of the input image. In the past few decades, many handcrafted features have been applied in the field of computer vision and significantly developed in the task of remote sensing image scene classification [1], [2]. Examples include texture descriptors [5], color histogram [6], scale-invariant feature transformation [7], histogram of oriented gradients [8], and GIST features [9]. Encoding local features to obtain a global image description has been the mainstream method of remote sensing image scene classification; these features include the improved Fisher kernel [10], vector of locally aggregated descriptors [11], spatial pyramid matching [12], and bag-of-visual-words [13]. When remote sensing image categories are simple and easy to distinguish, these methods can achieve acceptable results. However, with the improvements in the resolution of remote sensing images, the amount of image data continues to expand, and the types of scenes have become increasingly diverse. These handcrafted features can no longer meet the needs of a high-precision scene classification. Fortunately, breakthroughs in deep learning have provided efficient solutions for image recognition. Krizhevskey [14] proposed the first deep convolutional neural network (CNN) called AlexNet and applied it to the large-scale image dataset ImageNet [15] in 2012. Its performance in image classification tasks far exceeded traditional handcrafted features. Since then, higher-performance CNNs, such as VGGNet [16], GooLeNet [17], ResNet [18], and DenseNet [19], have been proposed and successfully applied to image classification, object detection, and semantic segmentation. The essence of scene classification is image classification; therefore, some scholars have begun to apply deep CNNs to remote sensing images for scene classification and object detection, even weakly supervised methods for scene classification [20] and object detection have been applied [21]–[23]. These applications have greatly improved the accuracy and efficiency of scene classification, and currently, CNNs are the mainstream methods for remote sensing image scene classification. However, when the scenes become fine-grained, ordinary CNNs cannot achieve satisfactory results, either. This is because remote sensing images have three characteristics [1], [2] compared with natural images like ImageNet: Large intra-class differences, high

Manuscript received July 23, 2020; revised September 7, 2020; accepted October 2, 2020. Date of publication October 13, 2020; date of current version November 2, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41601507. *(Corresponding author: Haitao Guo.)*

The authors are with PLA Strategic Support Force Information Engineering University, Zheng Zhou 450001, China (e-mail: dong_hang@aliyun.com; ghtgjp2002@163.com; 13937169139@139.com; ljhb45@126.com; hehe 549124@outlook.com; lyz120218@163.com).

Digital Object Identifier 10.1109/JSTARS.2020.3030257

inter-class similarities, and multidirectional and multiscale objects on the images. Because of these characteristics, achieving a high-precision and fine-grained classification of remote sensing images with large datasets is still highly challenging.

Like remote sensing image scene classification, the task of fine-grained visual categorization involves distinguishing similar images. In addition, the attention mechanism is crucial for improving the performance of CNNs. Inspired by the fine-grained visual categorization and the attention mechanism in CNNs, for the scene classification task of a remote sensing image, we propose a feature fusion algorithm called hierarchical attention and bilinear fusion net (HABFNet), based on channel attention and hierarchical bilinear pooling (HBP) [24], which has shown excellent performance in fine-grained visual categorization. We adopt ResNet50 to extract the features of different layers from the input image, and then use the channel attention to enhance the features. Finally, cross-layer bilinear pooling and feature connection are utilized for fusion. In summary, the main contributions of this article are as follows.

1) We introduce HBP into the remote sensing image scene classification task. This approach has been verified by experiments that demonstrate that the HBP model can extract a more discriminative feature representation than ordinary CNNs.
2) We improve the channel attention mechanism for remote sensing image scene classification, and the improved channel attention has a more obvious enhancement effect on features.
3) We design an end-to-end remote sensing image scene classification method using feature cross-layer bilinear fusion and the channel attention mechanism. In experiments using three challenging datasets, compared with state-of-the-art methods, our method achieves higher accuracies.

The remainder of this article is organized as follows. In Section II, related work is summarized, including the latest research progress in remote sensing image scene classification, fine-grained visual categorization, and the attention mechanism in CNNs. Our proposed method is introduced in detail in Section III. Section IV displays the experimental data and implementation details, and comparisons with the state-of-the-art methods. Section V contains a discussion about our method. Finally, the conclusion is drawn in Section VI.

II. Related Work

A. Scene Classification of Remote Sensing Images Using CNNs

Currently, there are three main methods used to recognize categories of remote scene sensing images using CNNs: Transferring learning, improving existing CNNs, and designing loss functions with strong discriminability.

Transferring learning mainly uses a pretrained model as a feature extractor to extract high-level and robust semantic features of remote sensing images. These features can be used for classification directly or through postprocessing. Penatti *et al*. [25] applied CNN to the remote sensing image scene classification and demonstrated its feasibility. Hu *et al.* [26] adopted the pre-trained CNN as a feature extractor, which extracted features of the convolutional layer and the fully connected layer respectively and a strong robust global representation of the image was obtained. Nogueira *et al*. [27] fine-tuned six popular CNNs. Compared with traditional methods, they achieved remarkable classification accuracies with limited data. Zhou *et al*. [28] utilized a pretrained CNN to extract deep features for image retrieval. Cheng *et al*. [29] re-encoded the features extracted by CNN to construct a bag of convolutional features. Li *et al*. [30] utilized a pretrained CNN to extract features and adopted a fusion mechanism for different layers of features to improve the accuracy of scene classification. Zhu *et al*. [31] designed a classification algorithm by fusing high-level, middle-level, and low-level features. Xue *et al*. [32] connected the features extracted by multiple CNNs from the same image, which improved the robustness of the features.

In addition to directly using the pretrained CNNs as feature extractors, a more popular method is to improve the structure of the existing CNNs according to the characteristics of remote sensing images or design an end-to-end feature fusion method to improve the robustness of features. Wang *et al*. [33] adopted active rotating filters based on VGG16 to enhance directional information. Xie *et al*. [34] proposed a scale-free CNN (SF-CNN) that can support images of any size. Wang *et al*. [35] introduced recurrent attention in CNN to enhance useful information and suppress useless information. Liu *et al*. [36] applied the Siamese CNN to the scene classification of a remote sensing image. He *et al*. [37] adopted multilayer stacked covariance pooling in CNN. Sun *et al*. [38] adopted the gated bidirectional connection method for feature fusion. Zhang *et al*. [39] used a multidilation pooling module, inverse residual, and channel attention in CNN, and designed a lightweight scene classification model.

In addition, inspired by face recognition, some scholars began to adopt a strong discriminative loss function in CNN training. Ye *et al*. [40] used the cross-entropy loss function and center loss function together to optimize the CNN. Wei *et al*. [41] used the marginal center loss to minimize the divergence between features from the same scene. Cheng *et al*. [42] added metric learning constraints to the features extracted by CNN, so that in the new feature space, the feature distribution of the same scene was more compact, which could effectively improve the accuracy of scene classification.

B. Fine-Grained Visual Categorization

Fine-grained visual categorization has already become an exceedingly important research topic in the fields of computer vision and pattern recognition in recent years. Its purpose is to classify images from the same basic category into more detailed subcategories [43], such as identifying different kinds of birds, dogs, and cats. Therefore, there is a high similarity between different subcategories and a large difference between image from the same category. This is the reason fine-grained visual categorization is so challenging. We found that the characteristics of fine-grained visual categorization are similar to those of remote sensing image scene classification. The categories of

high-resolution remote sensing images are becoming increasingly fine-grained. For example, the categories for residential areas consist of three subcategories: dense, medium, and sparse residential areas. Even for images of the same category, many images have different scales and perspectives.

To realize fine-grained visual categorization, Lin *et al*. [44] proposed bilinear CNN models by introducing the concept of bilinear pooling and achieved excellent classification performance. The form of bilinear pooling is simple and facilitates gradient back-propagation and end-to-end training. Gao *et al*. [45] proposed compact bilinear pooling to reduce the dimension of features. Fukui *et al*. [46] proposed multimodal compact bilinear pooling and applied it to visual question answering. Kim *et al*. [47] proposed the use of element-wise multiplication to perform low-rank bilinear pooling in feature fusion. Yu *et al*. [24] proposed HBP that utilized the cross-layer feature to enhance the representation ability of the model. Yu *et al*. [48] introduced bilinear pooling to the scene classification of remote sensing images, and the second-order information obtained by bilinear pooling significantly improved the accuracy of scene classification. Therefore, certain fine-grained visual categorization methods can help improve the accuracy of remote sensing image scene classification.

C. Attention in CNNs

The core idea of attention in CNNs is to design a mechanism that can automatically ignore irrelevant information and focus on important information. There have been many attention mechanisms successfully applied to image classification, semantic segmentation, and object detection tasks. Hu *et al*. [49] proposed squeeze-and-excitation (SE) networks by utilizing the SE module to obtain the importance of different feature channels through training and won the championship of the ImageNet 2017 image classification task. Woo *et al*. [50] added an attention module in spatial and channel maps simultaneously and applied it to boost the representation power of CNNs. Wang *et al*. [51] designed a nonlocal structure that aimed to gather information from other locations to enhance the features of the current location and overcome problems in expressing local receptive field features. Nonlocal structures can be easily integrated into CNNs to improve the performance of tasks such as object detection, semantic segmentation, and pose estimation. Fu *et al.* [52] proposed the use of spatial attention and channel attention to obtain spatial and channel dependency relationships. The output of the two attention modules was fused by multiplication to improve the accuracy of the semantic segmentation results. Bello *et al*. [53] proposed the use of self-attention as an alternative to convolution to improve the ability to express global information. Many experiments have verified that the introduction of the attention mechanism into CNNs effectively improves the performance. Inspired by the attention mechanism, we introduced an improved channel attention approach in our remote sensing image scene classification model to enhance the features extracted by the CNN.

III. Proposed Method

Through the current research on bilinear CNN models, we have learned that the bilinear model can obtain the second-order information of the image (realization by the outer product or dot product of the features), which is much more robust and discriminative than general features. Using second-order information, excellent performance can be achieved in fine-grained visual categorization. In fine-grained visual categorization, not only images of different categories are similar, but images of the same category also have diversity in perspectives and scales, which are consistent with the characteristics of remote sensing image scene classification tasks. Introducing the idea of the dot product on features to obtain second-order information from remote sensing images can improve the accuracy of scene classification. In addition, the attention mechanism is another technology that can significantly improve the performance of CNNs, which is based on a simple idea and structure. Inspired by the ideas of fine-grained visual categorization and attention in CNNs, the proposed HABFNet framework is shown in Fig. 1. HABFNet mainly includes three stages: Extraction of hierarchical features, enhancement of hierarchical features using channel attention, and bilinear pooling and fusion of hierarchical features. In the extraction of the hierarchical features stage, the image is input into ResNet50, which has been pretrained by ImageNet, and fine-tuning is performed based on the weight required to extract the deep features. Features from three layers are utilized for fusion. Then, channel attention is adopted to enhance the extracted features. Finally, the enhanced features are cross-layer bilinearly pooled and fused for classification.

A. ResNet50 for Feature Extraction

ResNet50 introduced residual learning to CNNs, thus, increasing the depth of the network to hundreds of layers, and its performance in image classification tasks exceeded that of humans, for the first time. ResNet50 is currently one of the most widely used networks in image classification and object detection tasks. Subsequently, many high-performance CNNs have been further improved based on ResNet50; thus, ResNet50 was used to extract features in this study. ResNet50 contains 49 convolutional layers and a fully connected layer. The architecture information of ResNet50 is shown in Table I. The 49 convolutional layers can be divided into five stages. Stage 1 contains only one convolutional layer and the number of convolution kernels is 64, the size of the convolution kernels is 7×7, and the stride is 2. Stages 2–5 have repeated similar units. The difference between each stage is the number of units and the number of convolution kernels in each unit. In each stage, except for the first unit, the other units adopt short-cuts across layers. Each unit has three convolutional layers. In every unit, the kernel size of the first convolutional layer is 1×1, the kernel size of the second convolutional layer is 3×3, and the kernel size of the third convolutional layer is 1×1. Stage 5 consists of three units, denoted as Stage 5-1, Stage 5-2, and Stage 5-3. The size of the output feature maps of each unit in Stage 5 is $2048 \times 7 \times 7$. Most methods also adopted ResNet50 to extract the features of remote sensing images, but only utilized

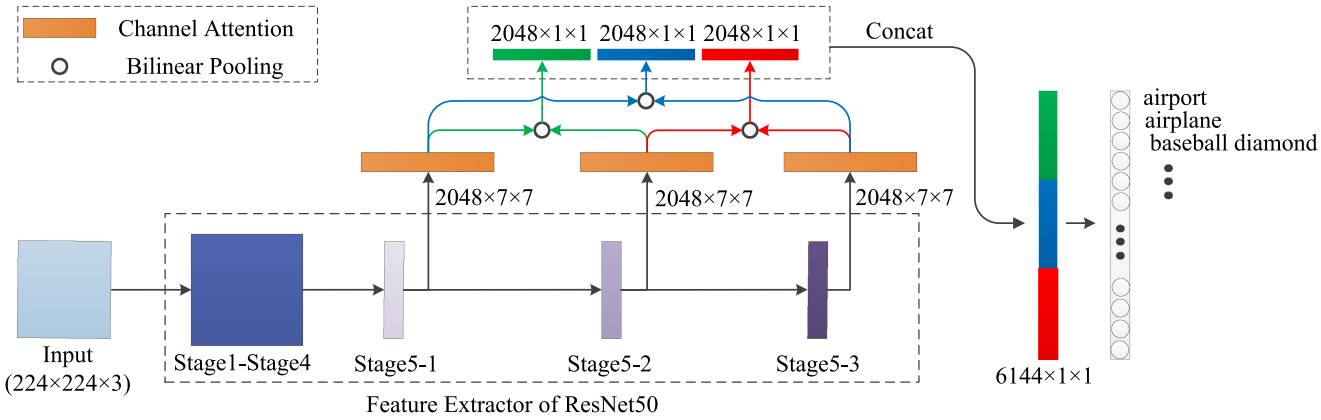


Fig. 1. Framework of HABFNet.

TABLE I
ARCHITECTURE OF RESNET50

Layer	Output size	Operation
Stage1	64 × 112 × 112	64@Conv (7 × 7)
Stage2-x	256 × 56 × 56	[64Conv(1 × 1); 64@Conv(3 × 3); 256@Conv(1 × 1)] ×3
Stage3-x	512 × 28 × 28	[128@Conv(1 × 1); 128@Conv(3 × 3); 512@Conv(1 × 1)] ×4
Stage4-x	1024 × 14 × 14	[256@Conv(1 × 1); 256@Conv(3 × 3); 1024@Conv(1 × 1)] ×6
Stage5-x	2048 × 7 × 7	[512@Conv(1 × 1); 512@Conv(3 × 3); 2048@Conv(1 × 1)] ×3
Classification	1000 × 1 × 1	[Average pooling layer; Fully connected layer]

the features from the Stage 5-3 layer. We wanted to fully utilize the features from different layers to obtain more discriminative features. Therefore, features from Stages 5-1, 5-2, and 5-3 were exploited for enhancement and fusion.

B. Channel Attention

We have improved the SE module and used the improved SE to enhance the features. SE uses the relationship between different channels of the feature maps to boost the performance of CNN. The SE module mainly includes three operations, namely, squeeze, excitation, and scale, to realize the recalibration of feature maps, which can be regarded as reweighting the feature maps.

In the SE module, the input feature maps $\mathbf{U}$ are squeezed to a channel descriptor. Generally, each channel with a spatial dimension of $H \times W$ is compressed into a channel descriptor with a global receptive field. The squeeze operation is usually realized by global mean pooling, and the implementation is as follows:

$$z_c = F_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

where $\mathbf{u}_c$ is the c_{th} channel of the feature maps $\mathbf{U}$, $u_c(i, j)$ is the value of the c_{th} channel at position (i, j), F_{sq} represents the squeeze operation, and z_c is the channel descriptor.

To fully utilize the information extracted by the squeeze operation, in the excitation operation, a two-layer fully connected neural network is used for nonlinear transformation. After the first fully connected layer, ReLu is used as the activation function, and after the second fully connected layer, Sigmoid is used as the activation function to obtain the weight values of different channels. The implementation of the excitation operation is as follows:

$$\mathbf{s} = F_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{2}$$

where $\mathbf{z}$ is the output of the squeeze operation, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights in the fully connected layer, δ represents the activation function of ReLu, σ represents the activation function of Sigmoid, F_{ex} represents the excitation operation, and $\mathbf{s}$ is the output of the excitation operation.

In the scale operation, $\mathbf{s}$ is used to reweight different channels of the feature maps $\mathbf{U}$ as follows:

$$\tilde{\mathbf{x}}_c = F_{\text{scale}}(\mathbf{U}, \mathbf{s}) = s_c \cdot \mathbf{u}_c. \tag{3}$$

In the formula, s_c represents the weight of the c_{th} channel of the feature maps $\mathbf{U}$, F_{scale} represents the scale operation, and $\tilde{\mathbf{x}}_c$ represents the reweighted feature.

The structure of the SE module is simple. Adding the SE module to CNN can improve the performance of natural image classification. However, using the SE module directly in remote sensing image scene classification tasks could not achieve an effective improvement in accuracy within the experiments. Therefore, based on the SE module, this study designed an improved SE module, as shown in Fig. 2. This was performed by adding a batch normalization [54] layer after each fully connected layer of the SE module, and replacing all activation functions in the SE module with HardTanh. Batch normalization is a

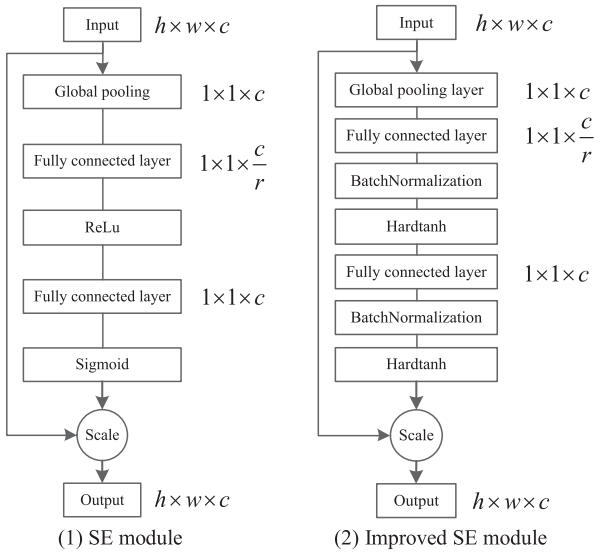


Fig. 2. Schematic of channel attention (1) SE module (2) Improved SE module.

widely used technology to enhance the generalization of CNNs. The batch normalization layer normalizes the input value to a distribution with a mean of 0 and a variance of 1, which ensures the effectiveness of the gradient during training. At the same time, the model can use a larger learning rate, thereby avoiding local extreme values, enhancing the ability of generalization, and greatly speeding up training. HardTanh is a linear piecewise approximation of the Tanh activation function and it is easier to calculate than existing methods, which improves learning speeds. The SE module has a reduction factor r that is used to adjust the number of neurons in the fully connected layers. The value of r in the SE module is 16, which can achieve a better balance between accuracy and calculation. The value of r in the improved SE module in this study is set to 2.

C. Cross-Layer Bilinear Pooling and Fusion

The core idea of bilinear pooling is to use two CNNs to obtain feature **A** and feature **B** of the same image and perform a fusion operation (outer product) on the two features to generate a new feature. Through pooling on the new feature, a bilinear vector can be obtained. Then, the signed square root and L2 normalization operations on the bilinear vector are used for fine-grained image classification. If feature **A** and feature **B** come from two different feature extractors, this type of bilinear pooling is called multimodal bilinear pooling; if feature **A** and feature **B** come from the same feature extractor, i.e., $\mathbf{A} = \mathbf{B}$, this type of bilinear pooling is called homogeneous bilinear pooling or second-order pooling. Yu *et al*. [24] gave the general form of the bilinear model and designed a HBP model. If the feature of the image **I** extracted by a single CNN is $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$, the output $\mathbf{Z_{out}}$ of the homologous bilinear pooling can be calculated as follows:

$$\boldsymbol{Z}_{\mathbf{out}} = \boldsymbol{P}^T(\boldsymbol{Q}^T\boldsymbol{X} * \boldsymbol{R}^T\boldsymbol{X}) \tag{4}$$

where $\boldsymbol{Q} \in \mathbb{R}^{c\times d}$ and $\boldsymbol{R} \in \mathbb{R}^{c\times d}$ are projection matrices that can change the dimension of the feature. If the features of the image **I** extracted by two different feature extractors, $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{Y} \in \mathbb{R}^{h\times w\times c}$, the output $\mathbf{Z_{out}}$ of the heterogeneous bilinear model can be calculated as follows:

$$\boldsymbol{Z}_{\mathbf{out}} = \mathbf{P^T}(\mathbf{Q^T}\mathbf{X} * \mathbf{R^T}\mathbf{Y}) \tag{5}$$

where $\boldsymbol{Q} \in \mathbb{R}^{c\times d}$ and $\boldsymbol{R} \in \mathbb{R}^{c\times d}$ are projection matrices that can transform and reduce the dimension of the feature maps, $\boldsymbol{P} \in \mathbb{R}^{d\times o}$ is the classification matrix, $*$ is the bilinear pooling using the dot product, h, w, and c are the height, width, and channel number of feature maps $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, d is the channel number of the feature maps transformed by the projection matrix, and o is the number of output categories.

If $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$, $\boldsymbol{Y} \in \mathbb{R}^{h\times w\times c}$, and $\boldsymbol{Z} \in \mathbb{R}^{h\times w\times c}$ are the features of different layers from the same CNN, these three features can be cross-layer bilinearly pooled and feature fused, i.e., cross-bilinear pooling is performed between features of different layers, and the results are connected and fused for classification. The output result $\boldsymbol{Z}_{\mathrm{HBP}}$ can be expressed as follows:

$$\begin{aligned}\boldsymbol{Z}_{\mathrm{HBP}} = \boldsymbol{P}^T\mathrm{concat}(&\boldsymbol{Q}^T\boldsymbol{X} * \boldsymbol{R}^T\boldsymbol{Y}, \boldsymbol{Q}^T\boldsymbol{X} \\ &* \boldsymbol{S}^T\boldsymbol{Z}, \boldsymbol{R}^T\boldsymbol{Y} * \boldsymbol{S}^T\boldsymbol{Z}).\end{aligned} \tag{6}$$

In the formula, $\boldsymbol{P}$ is the classification matrix, and $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{S}$ are the projection matrices of the features from different layers. In the HBP algorithm, before cross-layer bilinear pooling, three convolutional layers are used to perform projection transformation on $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, respectively. Ordinary convolutional layers cannot effectively enhance the features. Therefore, we proposed the introduction of channel attention into the HBP algorithm. Before bilinear pooling of the three feature maps of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, the improved SE module was adopted to recalibrate the feature maps respectively, and the output $\boldsymbol{Z}_{\mathrm{HABF}}$ can be expressed as follows:

$$\begin{aligned}\boldsymbol{Z}_{\mathrm{HABF}} = \boldsymbol{P}^T\mathrm{concat}(&\boldsymbol{Q}_{ca}^T\boldsymbol{X} * \boldsymbol{R}_{ca}^T\boldsymbol{Y}, \boldsymbol{Q}_{ca}^T\boldsymbol{X} \\ &* \boldsymbol{S}_{ca}^T\boldsymbol{Z}, \boldsymbol{R}_{ca}^T\boldsymbol{Y} * \boldsymbol{S}_{ca}^T\boldsymbol{Z})\end{aligned} \tag{7}$$

where $\boldsymbol{Q}_{ca}^T \in \mathbb{R}^{c\times d}$, $\boldsymbol{R}_{ca}^T \in \mathbb{R}^{c\times d}$, and $\boldsymbol{S}_{ca}^T \in \mathbb{R}^{c\times d}$ are the projection matrices formed by the improved SE module.

IV. Results and Analysis

A. Experimental Data

In this article, three publicly available and challenging remote sensing image datasets are used for evaluating the proposed approach.

The first dataset is the UC Merced dataset [13], which was released in 2010. The images in this dataset were selected from the United States Geological Survey National Map, covering more than 20 regions of the United States. The UC Merced dataset contains 2100 RGB color images, 100 images for each category. The image size is 256 × 256 pixels, and the pixel resolution is 30 cm. The UC-Merced dataset has 21 scene categories, namely agricultural land, airplane, baseball diamond, beach, buildings,

chaparral, dense residential area, forest, freeway, golf course, harbor, intersection, medium residential area, mobile home park, overpass, parking lot, river, runway, sparse residential area, storage tanks, and tennis courts.

The second dataset is the AID dataset [55], which was released in 2017. The images in the AID dataset were selected from Google Earth images and covered a wide range of countries and regions. The AID dataset contains 30 categories and a total of 10 000 RGB color images. However, the number of images in each category varies from 220 to 420. The image size in AID dataset is 600 × 600 pixels, and the pixel resolution ranges from 8 to 0.5 m, which further increases the classification challenge. The 30 categories are airport, bare land, baseball field, beach, bridge, center, church, commercial area, dense residential area, desert, farmland, forest, industrial area, meadow, medium residential area, mountain, park, parking area, playground, pond, port, railway station, resort, river, school, sparse residential area, square, stadium, storage tanks, and viaduct.

The third dataset is the NWPU-RESISC45 dataset [2], which was released in 2017 and contains 45 scene categories, covering more than 100 countries and regions, with a total of 31 500 RGB color images. Each category contains 700 images with a size of 256 × 256 pixels. The spatial resolution ranges from 30 to 0.2 m. The 45 categories are airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential area, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential area, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential area, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. NWPU-RESISC45 is one of the largest datasets, with the most categories and largest image differences, making it the most challenging dataset for remote sensing image scene classification.

B. Implementation Details

Before conducting the experiment, we divided the data into two parts: training set and test set. The data in the training set was used to train the model. When the training loss and accuracy of the model stabilized, the data in the test set was used to verify the model generalization. To fully utilize the limited training data and avoid the problem of overfitting, we rotated the training images clockwise by 90°, 180°, and 270°, and flipped the images horizontally and vertically to expand the data six times. To facilitate comparisons with other methods, we set the training ratio in accordance with published papers. For the UC-Merced dataset, we randomly selected 20%, 50%, and 80% of the data as the training images, and left the remaining images as the test images. For the AID dataset, the training ratios were set to 20% and 50%. For the NWPU-RESISC45 dataset, the training ratios were set to 10% and 20%. The overall accuracy and confusion matrix were used to evaluate the performance of our method. The overall accuracy calculation method comprised the number of correctly classified images divided by the number of all test images. To reduce the influence of random factors in the data division, the images in the training set were randomly selected from all the images five times on each dataset at each training ratio. The average and standard deviation of the overall accuracy were calculated.

TABLE II
OVERALL CLASSIFICATION ACCURACY (%) COMPARISON FOR THE UC MERCED DATASET

Method	Training Ratio		
	20%	50%	80%
Fusion by Addition [57]	92.96 ± 0.58		97.42 ± 1.79
salM^{3}LBP–CLM [58]		94.21 ± 0.75	95.75 ± 0.80
MSCP [37]			98.40 ± 0.34
ARCNet [34]		96.81 ± 0.14	99.12 ± 0.40
D-CNNs [42]			98.93 ± 0.10
MCNN [59]			96.66 ± 0.90
CNN-CapsNet [55]		97.59 ± 0.16	99.05 ± 0.24
VGG16-DF [60]			98.97
Siamese ResNet50 [36]	76.50	90.95	94.29
Fine-tuned ResNet50 [36]	74.11	89.43	91.90
FACNN [61]			98.81 ± 0.24
GBNet [38]		97.05 ± 0.19	98.57 ± 0.48
DDRL-AM [62]			99.05 ± 0.08
HABFNet (Ours)	**96.95 ± 0.59**	**98.47 ± 0.47**	**99.29 ± 0.35**

The highest accuracy appears in boldface font.

We used the deep learning framework PyTorch to build the model. The input size of the model was 224 × 224 × 3 pixels. The model was optimized using a stochastic gradient descent (SGD) algorithm, and the batch size was 32. The initial learning rate of the feature extraction was 0.01 (the weight of pretrained ResNet50 was loaded as the initial value), and the initial learning rate of the attention enhancement layer and classification layer was 0.1. The learning rate was reduced to 0.5 times after every 10 epochs. The model was trained with a total of 60 epochs, the weight decay factor was 0.00001, and the momentum was 0.9. The model training was implemented on a workstation equipped with an i7-6900K, 3.2 GHz, 64 GB memory processor. A graphics processing unit NVIDIA GeForce GTX1080Ti with 11 GB memory was used for acceleration.

C. Comparison With State-of-the-Art Methods

1) Classification of the UC Merced Dataset: For the UC Merced dataset, the performance comparison between HABFNet and the state-of-the-art methods is shown in Table II. Compared with the other two datasets, the UC Merced dataset has fewer categories, and the difference between the categories is also more obvious. When this dataset was released, it was mainly used to evaluate the performance of methods using handcrafted features. When using this dataset for experiments, most methods set the training ratio to 80%, and the overall accuracies have been close to 99%. Under the training ratio of 80%, the overall accuracy of our method reached 99.29%, exceeding most scene classification methods. Although the overall

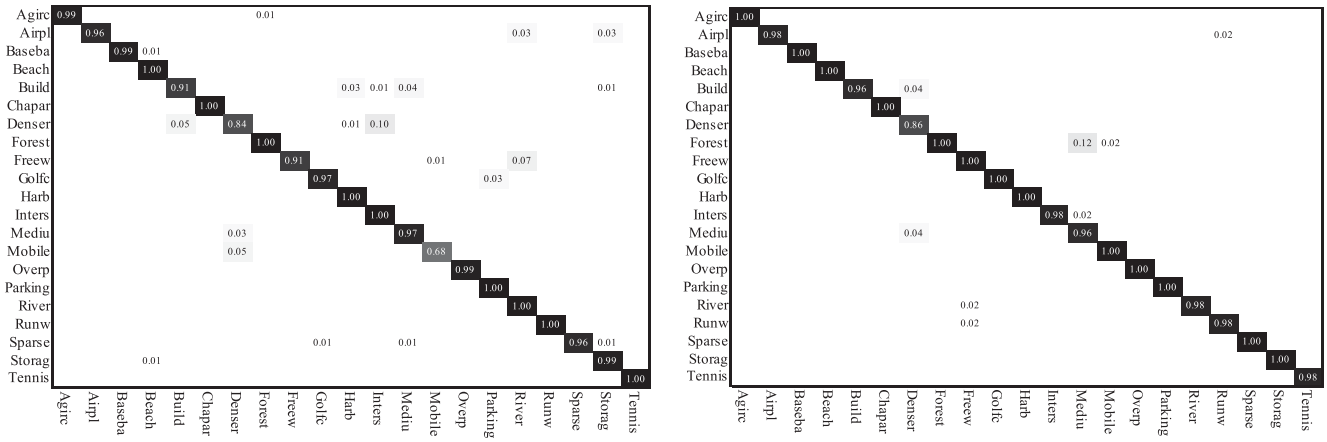


Fig. 3. Confusion matrices for HABFNet on the UC Merced dataset with a training ratio of 20% (left) and 50% (right).

accuracies of ARCNet [35], CNN-CapsNet [56], and GBNet [38] were close to ours when the training ratio was 80%, the accuracy of our method exceeds them when the training ratio was 50%. Compared with ARCNet, CNN-CapsNet, and GBNet, the overall accuracy of our method was improved by 1.66%, 0.88%, and 1.42%, respectively. There are a few methods that set the training ratio to 20%, mainly because the training dataset is so small that it could easily overfit. For example, Liu *et al*. [36] used 20% of the image to train ResNet50, and the overall accuracy was only 74.11%. When training Siamese ResNet50, the overall accuracy was only increased to 76.50%. Chaib [57] used VGG16 as a feature extractor instead of training and fused the extracted features. At a training rate of 20%, they achieved an overall accuracy of 92.96%. Under the same training ratio, the overall accuracy of HABFNet reached 96.95%. This demonstrates that the feature fusion by HABFNet is more efficient than addition.

Fig. 3 shows the confusion matrices of HABFNet on the UC Merced dataset when the training ratios were 20% and 50%, respectively. When the training ratio was 20%, 19 of the 21 categories had a classification accuracy above 91%. The categories with lowest accuracies were dense residential area (0.84) and mobile home park (0.86). Some images from the dense residential area category can be mistaken for buildings and medium residential area. The main ground objects of the three scenes (dense residential area, building, and medium residential area) are buildings, and the only difference is the building density in the image. When the training ratio was 50%, 20 of 21 categories had a classification accuracy above 96%, and the accuracy of dense residential area (0.86) was still the lowest. Under the same training ratio, the classification accuracy of dense residential area by CNN-CapsNet was only 80%. Obviously, our method can better learn the spatial differences of similar images and more effectively recognize fine-grained scenes.

2) Classification of the AID Dataset: When using the AID dataset for the experiments, the training ratios were set to 20% and 50%. The performance comparison between HABFNet and the state-of-the-art methods is shown in Table III. The table shows that when the training ratios were 20% and 50%, the overall classification accuracies of our method reached 95.48% and 96.95%, respectively, which surpasses those of most of the latest methods. The advantages of our method are most obvious at the training ratio of 20%. When comparing the six competitive algorithms of MSCP, D-CNN, SF-CNN, CNN-CapsNet, MSDFF, and DDRL-AM, their overall accuracies were close to those of our method when the training ratio was 50%, but when the training ratio was 20%, the overall accuracy of our method was improved as compared to its counterparts by approximately 3.27%, 4.66%, 1.88%, 1.69%, 2.01%, and 3.12%, respectively. Chaib [57] fused features together, and at a training ratio of 50%, their overall accuracy was 91.87%, which is lower than the 96.95% of HABFNet. MSDFF adopted three popular CNNs as feature extractors and fused features by connection. At a training rate of 20%, the overall accuracy of MSDFF was 93.47%, which is lower than the 95.48% accuracy of HABFNet. GBNet also used hierarchical features for feature fusion, but under the training ratios of 20% and 50%, the overall accuracies were lower than those of our method by 3.28% and 2.37%,

TABLE III
CLASSIFICATION ACCURACY (%) COMPARISON FOR AID

Method	Year	Training Ratio	
		20%	50%
Fusion by Addition [57]	2017		91.87 ± 0.36
salM^{3}LBP–CLM [58]	2017	86.92 ± 0.35	89.76 ± 0.45
MSCP [37]	2018	92.21 ± 0.17	96.56 ± 0.18
D-CNN [42]	2018	90.82 ± 0.16	96.89 ± 0.10
MCNN [59]	2018		91.80 ± 0.22
SF-CNN [34]	2019	93.60 ± 0.12	96.66 ± 0.11
CNN-CapsNet [56]	2019	93.79 ± 0.13	96.63 ± 0.12
FACNN [61]	2019		95.15 ± 0.11
MSDFF [32]	2020	93.47	96.74
DDRL-AM [62]	2020	92.36 ± 0.10	96.25 ± 0.05
GBNet [38]	2020	92.20 ± 0.23	94.58 ± 0.12
HABFNet (Ours)	2020	**95.48 ± 0.26**	**96.95 ± 0.17**

The highest accuracy appears in boldface.

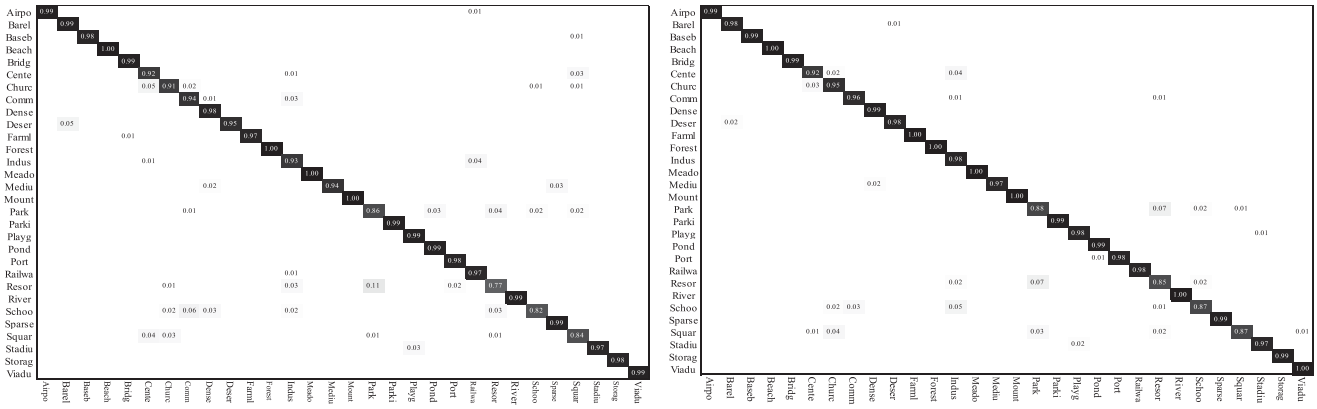


Fig. 4. Confusion matrices for HABFNet on AID with a training ratio of 20% (left) and 50% (right).

respectively. FACNN aggregated the features of multiple layers for fusion. At a training rate of 50%, its overall accuracy was 1.80% lower than that of HABFNet. The performance of all four feature fusion methods were significantly inferior to our method, which demonstrated that the method of feature fusion in HABFNet achieved the strongest performance and significantly improved the classification accuracies in the AID dataset.

Fig. 4 shows the confusion matrixes of the proposed method under different training ratios on the AID dataset. When the training ratio was 20%, 27 of the 30 categories had an accuracy of more than 91%, whereas CNN-CapsNet only has 24 categories. The categories with the lowest accuracy in our method were resort (0.77), school (0.82), and park (0.86), whereas the three classes with the lowest accuracies in CNN-CapsNet were school (0.68), resort (0.73), and center (0.81), which are all lower than those of our method. The resort, school, and park images mainly incorporate buildings and vegetation, and the texture information is relatively similar, thus, explaining the difficulty in distinguishing these classes effectively. When the training ratio was 50%, the accuracies of the three categories (resort, school, and park) were improved to 85%, 87%, and 88%, but the accuracies were still lower than those of the other categories. Therefore, the confusion of the three categories with other categories was the main reason for the limited improvement of the overall accuracy. Therefore, to further improve the performance of scene classification, future research should focus on improving the classification accuracies of these categories.

3) Classification of the NWPU-RESISC45 Dataset: For the NWPU-RESISC45 dataset, the performance comparison between our method and the existing state-of-the-art methods is shown in Table IV. For this dataset, HABFNet also achieved a remarkable performance, with its overall accuracies reaching 92.75% and 94.54%, respectively. When the training ratio was 20%, the performance of our method was better than that of most methods, being only weaker than that of FDPResNet [66]. FDPResNet uses ResNet101 to extract features and combines shallow and deep features via PCA, which cannot realize end-to-end training and prediction, and the calculation amount and complexity are much higher than in our method. Compared with the latest three methods (GLANet, DDRL-AM, and MSDFF), the overall accuracy of our method showed improvements of 1.09%, 3.08%, and 1.99%, respectively, with a training ratio of 20%. With a training ratio of 10%, compared with the latest four methods (GLANet, FDPResNet, DDRL-AM, and MSDFF), the overall accuracy of our method was improved by 1.62%, 0.43%, 0.58%, and 1.19%, respectively. Therefore, with less training data, the advantages of our method are more obvious.

Fig. 5 shows the confusion matrices of our method on the NWPU-RESISC45 dataset with different training ratios. We can observe the different categories being misclassified in detail. Using a training ratio of 10%, there were 11 categories whose accuracies were less than 90%: church (0.74), commercial area (0.88), dense residential area (0.87), freeway (0.87), medium

TABLE IV
OVERALL CLASSIFICATION ACCURACY (%) COMPARISON ON THE NWPU-RESISC45 DATASET

Method	Year	Training Ratio	
		10%	20%
BoCF [29]	2017	82.65 ± 0.31	84.32 ± 0.17
Fine-tuned VGG16 [1]	2017	87.15 ± 0.45	90.36 ± 0.18
D-CNN [42]	2018	89.22 ± 0.50	91.89 ± 0.22
IORN [33]	2018	87.83 ± 0.16	91.30 ± 0.17
ADSSM [63]	2018	91.69 ± 0.22	94.29 ± 0.14
MSCP [37]	2018	88.07 ± 0.18	90.81 ± 0.13
HW-CNNs [64]	2018		94.38 ± 0.17
SF-CNN [34]	2019	89.89 ± 0.16	92.55 ± 0.14
VGG16-DF [60]	2019		89.66
CNN-CapsNet [56]	2019	89.03 ± 0.21	92.60 ± 0.11
Siamese ResNet50 [36]	2019		92.28
GLANet [65]	2019	91.03 ± 0.18	93.45 ± 0.17
FDPResNet [66]	2020	92.32 ± 0.32	**95.40 ± 0.11**
DDRL-AM [62]	2020	92.17 ± 0.08	92.46 ± 0.09
MSDFF [32]	2020	91.56	93.55
HABFNet (Ours)	2020	**92.75 ± 0.18**	94.54 ± 0.06

The highest accuracy appears in boldface.

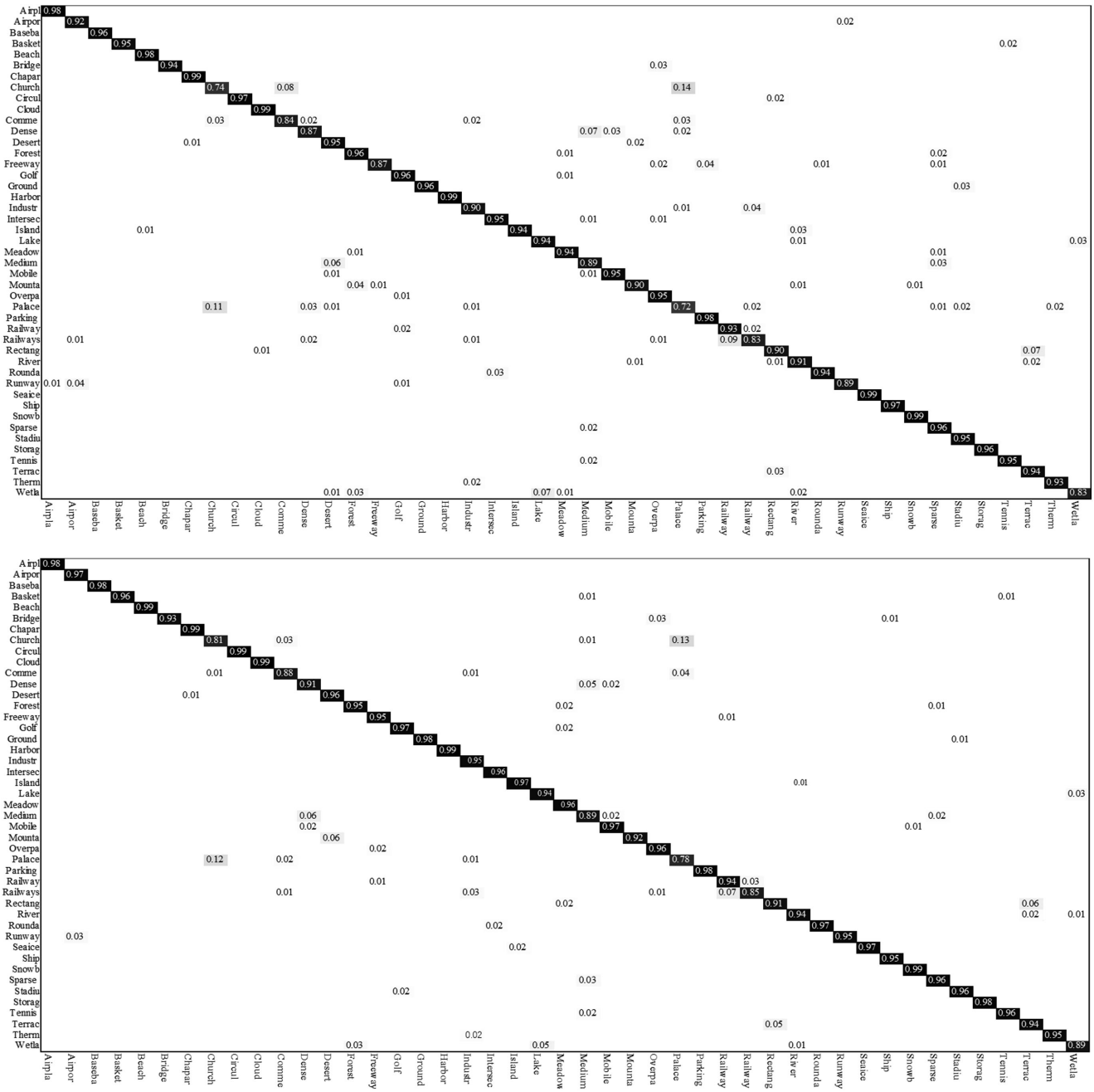


Fig. 5. Confusion matrices for HABFNet on NWPU-RESISC45 with a training ratio of 10% (up) and 20% (down).

residential area (0.89), palace (0.72), railway station (0.83), runway (0.89), and wetland (0.83). The highest mis-categorization occurred in the palace and church categories (at 14% and 11%, respectively) because they both had similar styles of buildings. Even when the training ratio was 20%, the accuracies of the church and palace categories were 81% and 78%, compared with 79% and 68% in CNN-CapsNet. There were six categories whose accuracy was less than 90%, whereas there were nine categories in CNN-CapsNet. The performance on the NWPU-RESISC45 dataset further verified the effectiveness of bilinear attention pooling and hierarchical feature fusion.

V. Discussion

The goal of the method in this study was to apply HBP with improved channel attention to the scene classification of remote sensing images. To further verify the effect of the attention mechanism on feature enhancement, the performances of the HBP, HBP with the SE module, and HABFNet (HBP with improved SE module) were compared for the three datasets, as shown in Tables V–VII, respectively. The HBP algorithm adopts bilinear pooling to obtain second-order information of the image, and cross-layer features are connected to form a

TABLE V
OVERALL ACCURACY (%) COMPARISON FOR THE UC MERCED DATASET WITH DIFFERENT ATTENTIONS

Method	Training Ratio		
	20%	50%	80%
HBP	96.31 ± 0.52	**98.48 ± 0.25**	98.94 ± 0.57
HBP with SE	95.38 ± 0.72	98.20 ± 0.41	98.75 ± 0.66
HABFNet	**96.95 ± 0.59**	98.47 ± 0.47	**99.29 ± 0.35**

The highest accuracy appears in boldface.

TABLE VI
OVERALL ACCURACY (%) COMPARISON FOR THE AID DATASET WITH DIFFERENT ATTENTIONS

Method	Training Ratio	
	20%	50%
HBP	94.84 ± 0.08	96.68 ± 0.27
HBP with SE	94.08 ± 0.14	96.25 ± 0.26
HABFNet	**95.48 ± 0.26**	**96.95 ± 0.17**

The highest accuracy appears in boldface.

TABLE VII
OVERALL ACCURACY (%) COMPARISON FOR THE NWPU-RESISC45 DATASET WITH DIFFERENT ATTENTIONS

Method	Training Ratio	
	10%	20%
HBP	92.20 ± 0.30	94.21 ± 0.11
HBP with SE	91.17 ± 0.16	93.63 ± 0.13
HABFNet	**92.75 ± 0.18**	**94.54 ± 0.06**

The highest accuracy appears in boldface.

strong robust feature. Without any improvement, this approach achieved a remarkable performance for all three datasets. With a training ratio of 20%, the overall accuracies of HBP for the UC Merced, AID, and NWPU-RESISC45 datasets were 96.31%, 94.84%, and 92.75%, respectively, which exceeded those of most state-of-the-art methods. SE is a module that is effective in boosting model performance for natural image classification. However, directly applying the SE module to the HBP algorithm will reduce the performance of HBP. For the three datasets with different training ratios, the overall accuracy of HBP with the SE module was significantly lower than that of the original HBP algorithm and our method. As a comparison, the introduction of the improved SE module in the HBP algorithm further improved the accuracy of the HBP in remote sensing image scene classification to a certain extent. For the UC Merced, AID, and NWPU-RSEISC45 datasets, when the training ratios were 20% and 80%, 20% and 50%, and 10% and 20%, respectively, the proposed method improved the overall accuracy of HBP by 0.64% and 0.35%, 0.64% and 0.27%, and 0.55% and 0.33%, respectively. When the training ratio was 20%, the improved SE module boosted the overall accuracies by 1.57%, 1.40%, and 1.58% as compared with the SE module, respectively. Moreover, when there were less training data, the performance improvement was more obvious.

TABLE VIII
OVERALL ACCURACY (%) COMPARISON BASED ON RESNET18

Method	Dataset		
	UC Merced	AID	NWPU-RESISC45
HBP	95.57 ± 0.67	94.47 ± 0.09	91.18 ± 0.37
HBP with SE	95.46 ± 0.67	94.33 ± 0.11	91.13 ± 0.30
HABFNet	**95.65 ± 0.82**	**94.66 ± 0.12**	**91.35 ± 0.22**

The highest accuracy appears in boldface.

TABLE IX
OVERALL ACCURACY (%) COMPARISON BASED ON RESNET34

Method	Dataset		
	UC Merced	AID	NWPU-RESISC45
HBP	95.97 ± 0.57	94.69 ± 0.18	91.58 ± 0.18
HBP with SE	95.42 ± 0.66	94.40 ± 0.19	91.34 ± 0.22
HABFNet	**96.64 ± 0.53**	**94.93 ± 0.12**	**91.94 ± 0.29**

The highest accuracy appears in boldface.

To verify the generalizability of the feature fusion method in this article to other CNNs, ResNet18, and ResNet34, which have a similar structure to ResNet50, were also selected for experiments. ResNet18 and ResNet34 were used to extract the hierarchical features of the image like ResNet50. Then, the features were enhanced by channel attention and fused. The results for the UC Merced (20% training ratio), AID (20% training ratio), and NWPU-RESISC45 (10% training ratio) datasets are shown in Tables VIII and IX. The tables show that no matter which CNN was used, the HBP algorithm with the SE module had the lowest accuracies, while HABFNet had the highest accuracies among the three methods. These results were consistent with those of ResNet50 for feature extraction, demonstrating that our improvement of the SE module is effective. In addition, ResNet50 had the best performance and, as the backbone, it can improve the accuracy of SE by about 1%. ResNet34 comes next and improved the accuracy of SE by $\sim$0.5%. ResNet18 only improved the accuracy by $\sim$0.3%. The stronger the performance of the backbone, the more obvious the advantages of the improved SE module.

VI. CONCLUSION

In recent years, developments in deep learning, especially CNNs, have provided efficient and simple solutions for scene classifications of remote sensing images. However, with improvements in resolution, the scene categories of remote sensing images are becoming increasingly diversified and fine-grained. Images of different categories are becoming increasingly similar in spatial distribution, which makes classification increasingly challenging. Existing methods still cannot completely solve the problems of intraclass diversity and interclass similarity, mainly because they cannot extract enough robust and discriminative features. Fine-grained visual categorization, which has been extensively studied, has a strong similarity with remote sensing image scene classification. However, there are few studies that regard remote sensing

image scene classification as fine-grained classification. Bilinear pooling is a simple and efficient approach for fine-grained image classification that can obtain second-order information from the image. Compared with first-order information, second-order information has better representation capability. For remote sensing image scene classification, second-order information is also applicable. Inspired by the fine-grained image classification and attention mechanism, this study proposes an end-to-end hierarchical attention and bilinear feature fusion algorithm for the scene classification of remote sensing images. Experiments were conducted on three popularly used datasets. The results prove the effectiveness of the attention and future fusion mechanism in this study. Future research should focus on studying the application of high-order information and fine-grained classification and design lightweight and high-precision classification algorithms for remote sensing images.

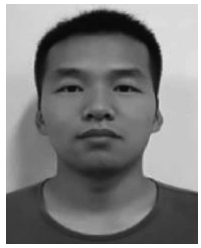
Acknowledgment

The authors would like to thank the anonymous Reviewers and Associate Editor for their valuable comments and suggestions to improve the quality of this article.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," 2020, *arXiv:2005.01094*.
[2] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state-of-the-art," *Proc. IEEE*, vol. 105, no. 5, pp. 1865–1883, Oct. 2017.
[3] Q. Hu *et al.*, "Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.
[4] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.
[5] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[6] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
[7] D. G. Lowe, "Distinctive image features from scale-invariant key-points," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, 2005, vol. 1, pp. 886–893.
[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[10] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[11] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Actions Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2011.
[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.
[13] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[17] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[20] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[21] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2985989.
[22] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.
[23] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2991407.
[24] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 574–589.
[25] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.
[26] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.
[27] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.
[28] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 489.
[29] 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.
[30] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.
[31] Q. Zhu, Y. Zhong, Y. Liu, L. Zhang, and D. Li, "A deep-local-global feature fusion framework for high spatial resolution imagery scene classification," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 568.
[32] W. Xue, X. Dai, and L. Liu, "Remote sensing scene classification based on multi-structure deep features fusion," *IEEE Access*, vol. 8, pp. 28746–28755, 2020.
[33] J. Wang, W. Liu, L. Ma, H. Chen, and L. Chen, "IORN: An effective remote sensing image scene classification framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1695–1699, Nov. 2018.
[34] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.
[35] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[36] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.
[37] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.
[38] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
[39] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.

[40] 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, 2019.
[41] 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.
[42] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
[43] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 119–135, 2017.
[44] T.Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[45] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 317–326.
[46] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.
[47] J. Kim, K. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv: 1610.04325*.
[48] 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, 2020, Art. no. 1999.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[50] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[51] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[52] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[53] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3286–3295.
[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[55] G.-S. Xia *et al.*, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[56] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.
[57] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
[58] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vvol. 10, no. 6, pp. 2889–2901, Jun. 2017.
[59] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.
[60] Y. Boualleg, M. Farah, and I. R. Farah, "Remote sensing scene classification using convolutional features and deep forest classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1944–1948, Dec. 2019.
[61] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.
[62] J. Li *et al.*, "Deep discriminative representation learning with attention map for scene classification," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1366.
[63] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.
[64] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, "Scene classification using hierarchical Wasserstein CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2494–2509, May 2019.
[65] Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, and D. Li, "Global-local attention network for aerial scene classification," *IEEE Access*, vol. 7, pp. 67200–67212, 2019.
[66] R. Dong, D. Xu, L. Jiao, J. Zhao, and J. An, "A fast-deep perception network for remote sensing scene classification," *Remote Sens.*, vol. 12, no. 4, 2020, Art. no. 729.



Donghang Yu received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Information Engineering University, Zhengzhou, China, in 2016 and 2019, respectively. He is currently working toward the Ph.D. degree with PLA Strategic Support Force Information Engineering University.

His research interests include remote sensing image classification and object detection.

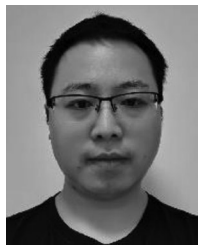
Haitao Guo received the B.S., M.S., and Ph.D. degrees from the Information Engineering University, Zhengzhou, China, in 1998, 2002, and 2008, respectively.

He is currently an Associate Professor with the PLA Strategic Support Force Information Engineering University. His research interests include remote sensing image processing and understanding, and change and object detection in remote sensing images.

Qing Xu received the B.S., M.S., and Ph.D. degrees from the Information Engineering University, Zhengzhou, China, in 1985, 1990, and 1995, respectively.

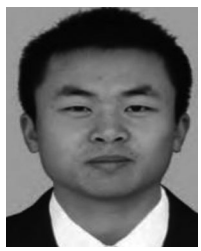
He is currently a Professor and a Doctoral Supervisor with the Information Engineering University. His research interests include photogrammetry and deep space remote sensing mapping.

Prof. Xu has been a member of the Professional Committee of the Photogrammetry and Remote sensing of the Chinese Society for Geodesy, Photogrammetry and Cartography, the Professional Committee of the Deep Space Exploration of the Chinese Society of Astronautics, and the Expert Committee of the Scientific Application of the Exploring Project around the Moon since 2000, 2004, and 2006, respectively.



Jun Lu received the B.S., M.S., and Ph.D. degrees from the Information Engineering University, Zhengzhou, China, in 2004, 2008, and 2015, respectively.

He is currently an Associate Professor with the PLA Strategic Support Force Information Engineering University. His research interests include remote sensing image matching and deep learning.



Chuan Zhao received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Information Engineering University, Zhengzhou, China, in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree with PLA Strategic Support Force Information Engineering University.

His research interests include airborne LiDAR point cloud classification and deep learning.

Yuzhun Lin received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Information Engineering University, Zhengzhou, China, in 2015 and 2018, respectively.

He is currently working as a Teaching Assistant with PLA Strategic Support Force Information Engineering University. His research interests include building extraction and deep learning.