A Multiscale Attention Network for Remote Sensing Scene Images Classification

Guokai Zhang, Weizhe Xu, Wei Zhao, Chenxi Huang, Eddie Ng Yk, Yongyong Chen, and Jian Su, *Member, IEEE*

***Abstract*—The remote sensing scene images classification has been of great value to civil and military fields. Deep learning models, especially the convolutional neural network (CNN), have achieved great success in this task, however, they may suffer from two challenges: first, the sizes of the category objects are usually different, but the conventional CNN extracts the features with fixed convolution extractor, which could cause the failure in learning the multiscale features; second, some image regions may not be useful during the feature learning process, therefore, how to guide the network to select and focus on the most relevant regions is crucially vital for remote sensing scene image classification. To address these two challenges, we propose a multiscale attention network (MSA-Network), which integrates a multiscale (MS) module and a channel and position attention (CPA) module to boost the performance of the remote sensing scene classification. The proposed MS module learns multiscale features by adopting various sizes of sliding windows from different depths' layers and receptive fields. The CPA module is composed of two parts: the channel attention (CA) module and the position attention (PA) one. The CA module learns the global attention features from channel-level, and the PA module extracts the local attention features from pixel-level. Thus, fusing both of those two attention features, the network is apt to focus on the more critical and salient regions automatically. Extensive experiments on UC Merced, AID, NWPU-RESISC45 datasets demonstrate that the proposed MSA-Network outperforms several state-of-the-art methods.**

***Index Terms*—Remote sensing scene, multi-scale, attention, feature fusion.**

I. Introduction

THE explosion of high-resolution remote sensing imaging technology has unleashed a veritable data deluge in investigating the land-use and land-cover scenes [9], [16], [18], [19]. Especially, the recognition and classification of the remote sensing scene images have been of great value to civil and military fields, due to its plentiful spatial and semantic information. Specifically, in this classification task, it uses pixel-based [3] or different levels of features to identify and label the images based on the image contents. However, the remote sensing scene image usually contains complicated ground objects with different spatial distributions, such as roads, buildings, and rivers, and that makes it difficult to classify the specific category of the whole scene image.

To tackle this challenge, during the past years, many works have been proposed for scene classification. The early attempts for this task mainly focused on the hand-crafted features such as texture, shape, color, and spatial representations, which are combined and fed into a classifier for prediction. For example, Yang *et al.* [65] designed a category of image descriptors based on saliency for remote sensing scene images classification. Luo *et al.* [40] proposed to extract the radiometric, Gaussian wavelet, Gabor, and Gray level co-occurrence matrix features with different spatial resolutions for indexing of remote sensing scene images. To extract more spatial features, the work in [77] developed an effective approach by fusing the local and global spatial features with multiscale feature learning mode. Meanwhile, Huang *et al.* [24] utilized the patch-based multiscale local binary pattern features and a Fisher vector for remote sensing scene images classification. Some other previous works also used the color features [31], [49], [51], [57], histogram of oriented gradients features [6], [7], [41], and bag-of-visual-words [12], [52], [74] for classification. However, the above methods based on the hand-crafted features may yield unsatisfactory performance since they require subjective and empirical feature definition and selection. Furthermore, these features are usually low-level and mid-level, which may become limited and inadequate for the complex remote sensing scene image feature learning.

Since the remarkable performance of the convolutional neural network (CNN) in many computer vision fields [5], [17], [22], [25], [45], [61], [73], [76], many researchers also used CNN to extract high-level features from the remote sensing scene images[15], [20]. Compared with the hand-crafted based feature learning, CNN could extract more semantic features by deep network layers in an end-to-end learning manner. Especially,

Manuscript received June 3, 2021; revised July 24, 2021 and August 10, 2021; accepted August 20, 2021. Date of publication September 2, 2021; date of current version October 4, 2021. This work was supported in part by the Shenzhen College Stability Support Plan under Grant GXWD20201230155427003-20200824113231001 and in part by the General Program of National Natural Science Foundation of China (NSFC) under Grant 62106063 and Grant 62102259. *(Corresponding authors: Yongyong Chen; Jian Su.)*

Guokai Zhang is with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: zhangguokai_01@163.com).

Weizhe Xu is with the School of Computer Science, University of Manchester, M13 9PL Manchester, U.K. (e-mail: weizhe.xu@postgrad.manchester.ac.uk).

Wei Zhao is with the School of Software Enginnering, Tongji University, Shanghai 200092, China (e-mail: zhaoweilh@tongji.edu.cn).

Chenxi Huang is with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: supermonkeyxi@xmu.edu.cn).

Eddie Ng Yk is with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore (e-mail: mykng@ntu.edu.sg).

Yongyong Chen is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: yongyongchen.cn@gmail.com).

Jian Su is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sj890718@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3109661

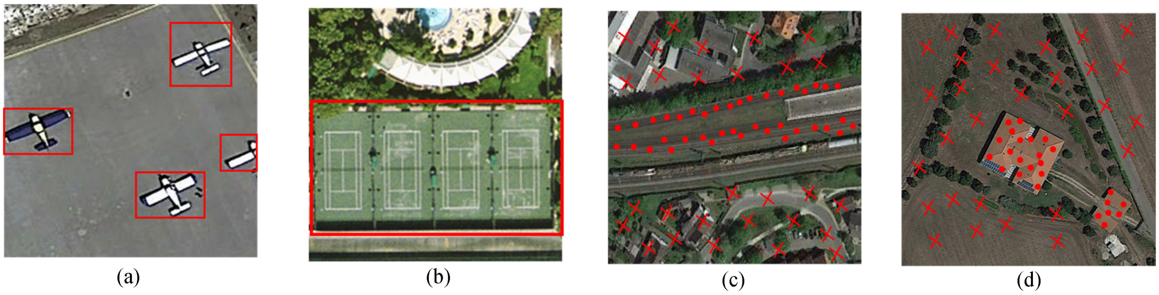

Fig. 1. Size of the target category airplane in scene image (a) is quite different from that of the target category tennis court in scene image (b). In the image (c) and (d), the red "." denotes the target category key regions, and the red "×" represents the useless regions of other objects, it is obvious that not all the image regions are useful during the feature leaning process.

the feature representations from the deeper layers could provide more abstract and semantic information, which are more applicable to the remote sensing scene image classification task. There are many baseline CNN models such as AlexNet [26], VggNet [54], GoogleNet [56], ResNet [13], which have achieved impressive performance for the natural image classification on ImageNet dataset [27]. Inspired by these works, many works tried to explore the huge power of CNN for remote sensing scene image classification. For example, Han *et al.* [11] used the AlexNet as the backbone architecture and incorporated the spatial pyramid pooling layer to learn the multiscale information of the remote sensing scene image. Castelluccio *et al.* [3] explored the CaffeNet [28] and GoogleNet [56] with widely different network settings for this classification task, in [43] and [50], the deeper network architecture such as the ResNet was used to learn more semantic features. Meanwhile, some other hand-crafted features such as the texture, color, and local binary patterns (LBP) [1] were combined with the deep features to further improve the classification performance. The previous works in [4], [23], [39], [58], [69] proposed different feature fusion strategies with various network structures to explore the effectiveness of different fusion modes. To alleviate the influence of limited remote sensing scene images, transfer learning, or pretrained methods [3], [4], [10], [21], [23], [32], [34], [44], [58], which pretrain the network on large amounts of datasets such as the ImageNet had been used for the improvement of the classification performance.

Although these works have achieved promising performance for remote sensing scene classification, they may encounter two main challenges. First, the category objects of remote sensing scene images usually have different sizes as shown in Fig. 1(a) and (b), in which the size of the target category airplane in scene image (a) is quite different from that of the tennis court in (b). However, the conventional CNN extracts the features with a fixed convolution extractor, which could be a handicap to learn the multiscale features; second, as shown in Fig. 1(c) and (d), the red "." denotes the target category key regions, and the red "×" represents the useless regions of other objects, it is obvious that not all the image regions are useful during the feature leaning process. Therefore, how to guide the network to select and focus on the most relevant regions is crucially vital for remote sensing scene image classification. To address these two challenges, in this article, we propose a multiscale attention network (MSA-Network) to achieve remote sensing scene images classification tasks. The proposed MSA-Network uses ResNet [13] as the backbone network and integrates a multi-scale (MS) module and a channel and position attention (CPA) module to further boost the classification performance. Inspired by the previous work [56], the designed MS module extracts multiscale features from different receptive fields with various sizes of sliding windows. Besides, since different depths of layers may contain pyramidal scale features, we add the MS module behind each stage's last residual block to extract the multiscale features hierarchically. The CPA modules consist of two parts, i.e., the channel attention (CA) module, and position attention (PA) module, respectively. During the feature learning process, the CA module extracts the attention features from channel-level globally, while the PA module learns the attention features from pixel-level locally. By integrating those two attention features, it could guide the network to focus on more informative and critical regions globally and locally. The main contributions of this article can be summarized as follows:

1) A novel MS module has been proposed to improve the network ability to capturing multiscale features during the feature learning process.
2) We design a CPA module, which guides the network to focus on the informative critical regions globally and locally.
3) Extensive experiments on three (UC Merced, AID, NWPU-RESISC45) datasets demonstrate our proposed MSA-Network has achieved competitive results over other state-of-the-art methods.

The rest of this article is organized as follows. Section II introduces the proposed MSA-Network. The extensive experiments and results are presented in Section III. In Section IV, we conduct a qualitative analysis of our designed model. Finally, Section V, concludes this article.

II. Proposed MSA-Network

In this section, we give a detailed description of the designed model. As illustrated in Fig. 2, the main backbone of our model is based on ResNet, which has achieved great success in many computer visions tasks. Meanwhile, for better extracting more multiscale and discriminative features, we propose an MS module and a CPA module. Specifically, we add the MS module

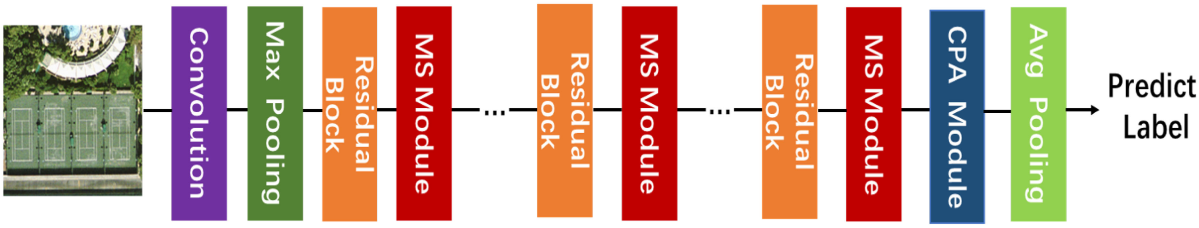

Fig. 2. Overview of the proposed architecture for remote sensing scene images classification. The main backbone of our network is based on ResNet, and we integrate the MS module and the CPA module with the network. The MS module is added behind each stage's last residual block to extract the multiscale features hierarchically. The CPA module is added behind the last MS module to guide the feature learning process to focus on more informative and critical regions. For different depth of ResNet backbones, we list detailed parameters and network settings in Table I.

TABLE I
PARAMETERS SETTING OF THE NETWORK

layer name	18-layer	34-layer	50-layer	101-layer
Conv1	7×7, 64, stride 2			
Pooling	3×3, max pool, stride 2			
Conv2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
	MS module layer			
Conv2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
	MS module layer			
Conv2	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$
	MS module layer			
Conv2	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	MS module layer			
	CPA module layer			
Output-layer	average pool, fc, softmax			

after each residual block to extract the multiscale features hierarchically. Since the deeper layer contains more high-level and semantic features, we integrate the CPA module after the last MS module to focus on more informative and critical regions. For different depths of ResNet backbones, we list the parameters and network settings in Table I. The more detailed introductions of the proposed modules will be discussed in the following sections.

A. Multiscale Feature Learning

The remote sensing scene images usually compose of complex and diverse objects in the real world, and the main category of remote sensing scene object is often with various sizes. However, the conventional ResNet usually uses the fixed size convolutional layers to extract local features, which could be a handicap for the network to learn multiscale features. Thus, inspired by the previous work in [56], we use an MS feature learning module as a basic unit of the network to enhance the multiscale feature learning. The detailed structure of the proposed MS module is illustrated in Fig. 3. Denote the F_{i-1} as the input feature from the previous layer, and F_i as the output feature from the designed module. Instead of directly passing F_{i-1} into the next layer, we first apply multiscale feature learning with different convolution kernel sizes. Here, we use $\{1 \times 1, 3 \times 3, 6 \times 6, 9 \times 9\}$ as the basic units to generate four scale-level features. Next, an interlaced feature learning strategy is adopted for aggregating more contextual multiscale information from the input features. Specifically, for the sizes of 1×1 and 3×3 kernels, they aim to learn more precise

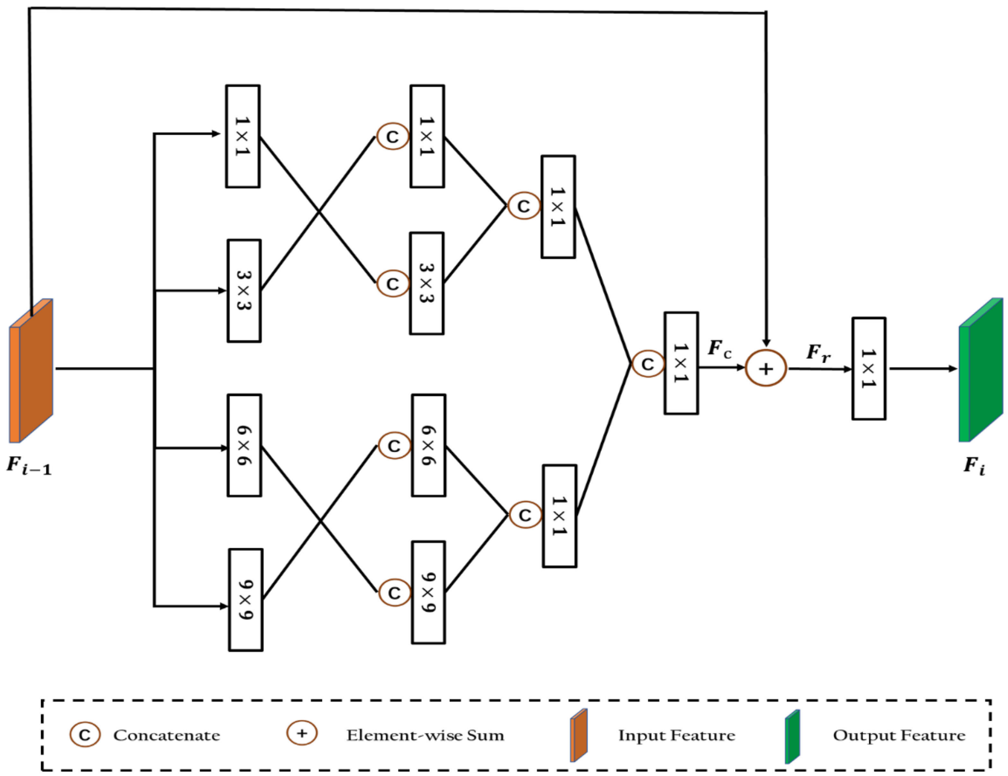

Fig. 3. Detailed structure of the MS module, F_{i-1} and F_i are the input and output feature maps of the ith layer, respectively.

and subtle information. While for the sizes of 6×6 and 9×9 kernels, they are more likely to extract global and large-scale representations. The different sizes of convolution operations f_s can be formulated as follows:

$$f_s = \sum_{s \in S} (k_s * F_{i-1} + b_s) \tag{1}$$

where k_s is the kernel for the scale of $s \in \{1, 3, 6, 9\}$, and b_s is the bias for f_s. After the interlaced feature learning, two 1×1 convolution layers are utilized to aggregate the global and local features and then input to another 1×1 convolution layer to gain the squeezed feature F_c. Subsequently, we perform an element-wise sum F_c with F_{i-1} by residual learning to further improve the convergence ability of the network, and the output feature maps F_r can be given as follows:

$$F_r = F_c + F_{i-1}. \tag{2}$$

After that, the final F_i is gained by applying a 1×1 convolution operation on F_r. Notably, inspired by the fact that different depths of layers could contain pyramidal scale features [33], for our MAS-Network, we insert our MS module behind the last residual block of each stage to learn the scale-relevant feature pyramidally, and the detailed inserted position of this module is illustrated in Table I.

B. CPA Modules

Since the remote sensing scene images are captured from an overhead view, they usually contain complex and diverse objects. Thus, many objects are not useful for the image classification task. To handle this problem, we use the attention mechanism to guide the feature learning process to focus on more informative and critical regions. Conventionally, the attention mechanism has two categories: one is the hard attention, which restricts the regions to 0 or 1, the other one is the soft attention, which calculates the weight of the specific region. In this article, we use the soft attention mechanism with two attention modules to select and learn more global and local representations. Denote the output feature map from the last residual block as $F \in \mathbb{R}^{C \times H \times W}$, where C, H, W are the channel, height, and width of the output feature, separately. F_C is the CA feature map, and F_P is the PA feature map. In the following sections, we will give a detailed introduction to these two attention modules.

1) CA Module: Our CA module is inspired by the previous work squeeze and channel excitation block [29] and the spatial pyramid pooling [14]. The feature map of each channel contains different global and semantic responses, which is essential for remote sensing scene images understanding. Thus, in order to improve the ability to learn discriminative features from the channel level, we design a CA module to encode a wider of contextual and global representations from the channel level. The detailed structure of the proposed CA module is illustrated in Fig. 4. For the input feature map $F \in \mathbb{R}^{H \times W \times C}$, we first perform three max-pooling operations with sizes of $\{2 \times 2, 4 \times 4, 8 \times 8\}$ to learn the contextual and global information, and it could be formulated as $\{F_{\text{sp}}, p \in \{2, 4, 8\}\}$. Then, a SE Block is utilized on F_{sp} to enable the network to focus on the most salient representations globally from the channel-level.

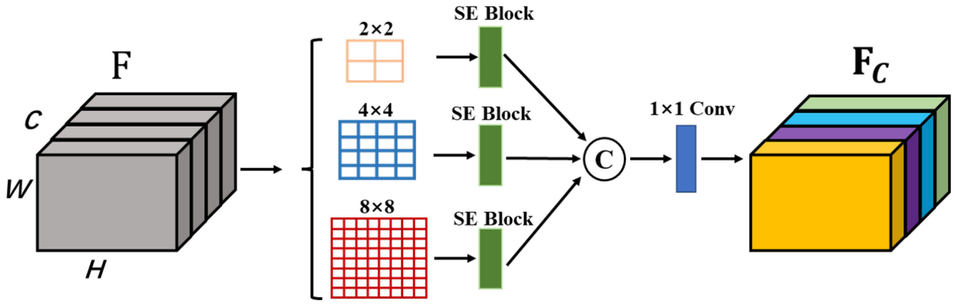

Fig. 4. Structure of our CA module, F denotes the input feature map with dimension of $C \times H \times W$, and F_C is the final output from the CA module.

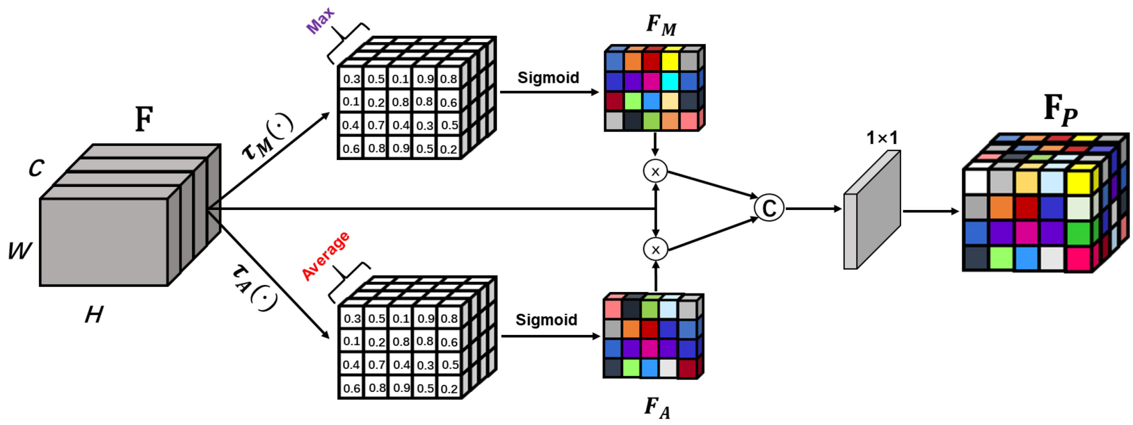

Fig. 5. Structure of PA module, F denotes the input feature maps with the dimension of $C \times H \times W$, F_P is the final output from PA module.

Here, we denote each single F_{sp}^j as the jth channel feature map of F_{sp} where $j \in [1, C]$. In SE Block, a global-average pooling over F_{sp}^j is used to generate the global weight a_j of the jth channel. And the operation of the global-average pooling for the jth channel can be formulated as follows :

$$a_j = \frac{1}{H \times W} \sum_x^H \sum_y^W F_{\text{sp}}^j(x, y). \tag{3}$$

For better improving the generalization of the module, a fully connected (FC) layer is applied to the weight vector a ($a = \{a_1, a_2, ..., a_j, ..., a_C\}$) with a ReLU operator δ. Then, another FC layer with sigmoid activation σ is used to normalize outputs of the previous layer. These two FCs operations are defined as follows:

$$A_W = \sigma(W_2 \delta(W_1 a)) \tag{4}$$

where $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$ and $W_2 \in \mathbb{R}^{\frac{C}{r} \times C}$ are the respective weights of the two FCs, and the gained value of A_W represents the importance weight of channel feature map from the global-average level. Additionally, the variable of r is the bottleneck of the channel excitation, and we set it four empirically. Next, the CA output of $\hat{F_{\text{sp}}}$ is gained by applying an muplication between A_W and F_{sp}, which could be formulated as follows:

$$\hat{F_{\text{sp}}} = A_W \cdot F_{\text{sp}}. \tag{5}$$

After that, we concatenate those three CA outputs and apply a convolution layer with a size of 1×1 to squeeze the channel value to C. Therefore, the final output of CA module $\hat{F}$ is calculated as follows:

$$F_C = \text{Conv}(\text{Concat}(\hat{F_{s2}}, \hat{F_{s4}}, \hat{F_{s8}})). \tag{6}$$

With the designed CA module, it could guide the network to focus on the globally crucial representations, which further improve the classification performance.

2) PA Module: Different from the CA module, the goal of our PA module is to extract more subtle features from the pixel level. The detailed structure of the PA module is illustrated in Fig. 5. In order to learn the importance of each feature pixel position, we first apply two position operations $\tau_M(.)$ and $\tau_A(.)$, which are utilized to calculate the max-value and average-value of each feature pixel position of the whole feature channels. After that, two sigmoid activations are employed to those two obtained features and gained the position weighted feature map F_M and F_A, respectively. Then, we multiply the input feature F with F_M and F_A to gain the position enhanced feature F'_M and F'_A, separately. And it could be formulated as follows:

$$F'_M = F \cdot F_M, F'_A = F \cdot F_M. \tag{7}$$

Finally, the output feature map of the PA module is obtained by concatenating the F'_M and F'_A, and then applying a (1×1)

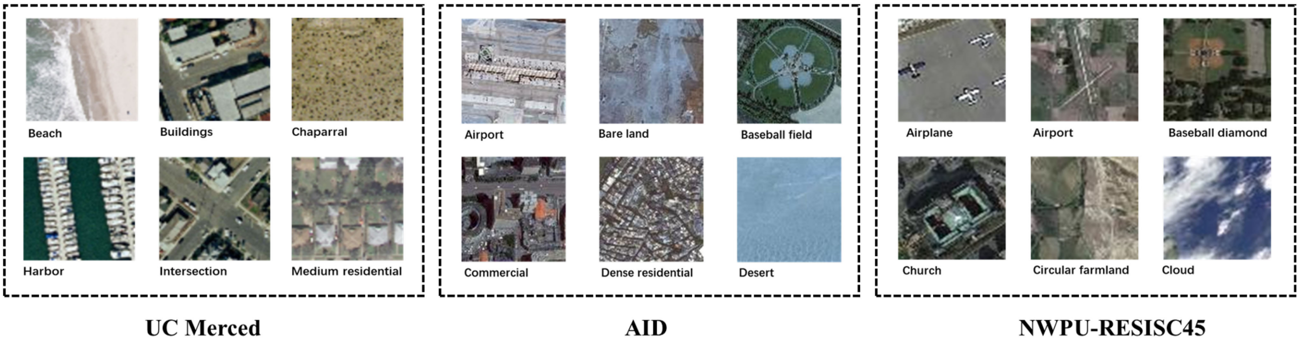

Fig. 6. Example images of UC Merced/AID/NWPU-RESISC45 dataset.

convolution layer to aggregate those two features, and the aggregation process can be defined as follows:

$$F_P = \text{Conv}(\text{Concat}(F'_M, F'_A)). \tag{8}$$

3) Channel and Local Feature Fusion: In our designed model, the CA module could generate more global representations over the whole feature map, while the PA module extracts the location attention feature from the pixel level. To incorporate more discriminative features from those two modules, we explore three combination methods to boost the remote sensing scene classification performance.

1) *Concatenation:* The concatenation operation is to concatenate two features at the same location x, y along the specific channel direction d into a vector

$$y^{x,y,d}_{\text{concat}} = \text{concat}(F^{x,y,d}_C, F^{x,y,d}_P) \tag{9}$$

where $d \in \{0, 1, 2\}$, and $1 \leq x \leq H, 1 \leq y \leq W$. $y^{x,y,d}_{\text{concat}}$ is the output feature by concatenation operation. Although the concatenation operation increases the complexity with stacking more channels, there is no information loss during the fusion process.

2) *Addition:* The addition operation is to compute the sum of two input features at the same location. The addition operation is formulated as follows:

$$y^{x,y,c}_{\text{add}} = F^{x,y,c}_C + F^{x,y,c}_P \tag{10}$$

where $1 \leq x \leq H, 1 \leq y \leq W, 1 \leq c \leq C$, and $y^{x,y,c}_{\text{add}}$ is the output feature by addition operation.

3) *Nonlinear Fusion:* The strategy of nonlinear fusion is similar to the concatenation fusion, except that we apply an FC layer with the nonlinear activation function to each input feature before the concatenation operation:

$$y^{x,y,d}_{\text{nonlinear}} = \text{concat}(W_C(F^{x,y,d}_C), W_P(F^{x,y,d}_P)). \tag{11}$$

Here, $1 \leq x \leq H, 1 \leq y \leq W$, and $d \in \{0, 1, 2\}$ is the specific channel direction, W_C and W_P is the corresponding weight of the FC layer, $y^{x,y,c}_{\text{nonlinear}}$ is the output feature by nonlinear fusion operation.

III. Experiments and Results

In this section, we evaluate the proposed MSA-Network model on three different aerial scene public datasets. We first give a brief description of all datasets, then the detailed experimental setting of the model is introduced. Finally, we conduct extensive experiments to further validate the performance of the proposed MSA-Network model.

A. Introduction of all Datasets

1) UC Merced Dataset: The UC Merced dataset, which was extracted from the USGS National Map Urban Area Imagery collection, consists of 21 land-use scene classes, as shown in Fig. 6. There are 100 images for each class, and each image has a one-foot spatial resolution, measuring 256 × 256 pixels.

2) AID Dataset: The AID dataset is a new large-scale image dataset that contains sample images collected from Google Earth imagery. There are 10 000 images within 30 classes, as shown in Fig. 6. Each image of the AID dataset has a spatial resolution ranging from 8 m to half a meter, and there are about 220–400 samples measuring 600 × 600 pixels in each class. For each class of the AID dataset, all the images are selected from different countries and regions around the world. As a consequence, the dataset has high intraclasses diversities.

3) NWPU-RESISC45 Dataset: The NWPU-RESISC45 dataset, created by Northwestern Polytechnical University (NWPU), contains 31 500 images for remote sensing image scene classification. The dataset consists of 45 scene classes, and there are 700 images measuring 256 × 256 pixels in each class. The 45 classes are shown in Fig. 6.

From Fig. 6, we can see that the sizes of the categories are various with different distributions, which could be a challenge for the conventional neural network to learn the semantic features.

B. Implementation Details

In this article, we use Tensorflow as the basic framework to implement the proposed MSA-Network model. The main backbone of our network is based on the ResNet [13]. Following the experimental setting of [47], [59], [71], for the UC Merced dataset, we use 80% and 50% as the training sample ratios, respectively. For the AID dataset, we use 50% and 20% as the

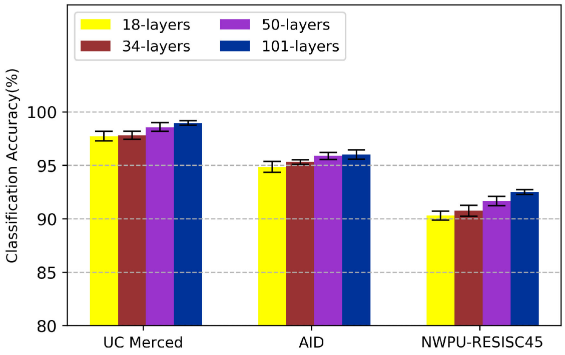

Fig. 7. Comparison results with different network depths.

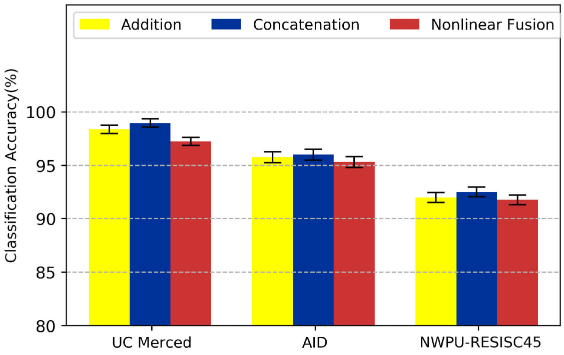

Fig. 8. Comparison results with different fusion methods.

training sample ratios, respectively. For the NWPU-RESISC45 dataset, we use 20% and 10% training sample ratios, respectively. During the training process, we use Adam optimization [30] to make the network convergent. The initial learning rate is set as 1.0×10^{-4}, and then we reduce it by 0.1 factor after the val-loss not improved 10 epochs. The batch-size of the proposed MSA-Network is 32, and we use real-time data augmentation on the training dataset such as the random rotation, flip, and cropping. We pretrain our model on the ImageNet dataset, and dropout is adopted to avoid the network being overfitting. We conduct our experiments on Ubuntu 14.04 operating system with 64 GB memory, and an NVIDIA GTX 1080 graphics processing unit has been used to accelerate the training process.

C. Evaluation Protocol

The overall accuracy (OA) and confusion matrix are selected as the criterion to evaluate the performance of the MSA-Network. The value of OA is calculated by the ratio between the correct numbers of classified images and the total number of images. It is one of the basic evaluation metrics for the classification task, and the higher value of OA denotes the more accurate classification performance. The confusion matrix is a table layout that describes the errors and confusion of each class. Each row of the table denotes the predicted category instance, and the column represents the actual category instance. In addition, the diagonal of the table are the numbers of all the classes correctly classified. The final confusion matrix is calculated by the best classification result of all the training ratios for each dataset.

D. Performance of Different Network Depths

Different depths of networks could extract various representations. For a deeper network, it is liable to learn more semantic and high-level features, for a shallower network, it would extract tinier and more detailed information. Thus, in this section, we first explore the performance of four network depths (18-layer, 34-layer, 50-layer, 101-layer) on the three datasets with our designed architecture. The depth of the network is deepened by adding the convolution and residual blocks as previous work [13]. The results are reported in Fig. 7. We use the 80%, 50%, and 20% of training data for the UC Merced dataset, AID dataset, and NWPU-RESISC45 dataset, respectively. On the UC Merced dataset, the proposed MSA-Network achieves $(97.73 \pm 0.45\%)$, $(97.82 \pm 0.38\%)$, $(98.59 \pm 0.42\%)$, and $(98.96 \pm 0.21\%)$ classification accuracy for 18-layers, 34-layers, 50-layers, and 101-layers, separately. On the AID dataset, the results are $(94.86 \pm 0.51\%)$, $(95.32 \pm 0.21\%)$, $(95.89 \pm 0.33\%)$, and $(96.01 \pm 0.43\%)$ for 18-layers, 34-layers, 50-layers, and 101-layers, respectively. The performance of 18-layers, 34-layers, 50-layers, and 101-layers on NWPU-RESISC45 dataset is $(90.31 \pm 0.42\%)$, $(90.78 \pm 0.51\%)$, $(91.67 \pm 0.43\%)$, and $(92.52 \pm 0.23\%)$. The experimental results demonstrate that the deeper network layers could achieve better performance compared with shallower network layers. The best result on three datasets is by using the 101-layers.

E. Effectiveness of Different Modules

In this article, we propose an MS module and two attention modules (CA module and PA module) to extract the multiscale, global, and local representations more effectively. In order to evaluate the effectiveness of different modules, we conduct relevant experiments on those three datasets. The detailed comparison result is shown in Table II. From the experimental results, we can observe that adding one of the MS, CA, and PA module could efficiently improve the classification performance of the network on different datasets, that further proves the effectiveness of our proposed modules. Overall, the best performance is achieved by using the MS module, CA module, and PA module simultaneously, and it further proves that the multiscale, CA, and PA features are crucially important in the remote sensing images classification task.

F. Compared With Different Fusion Methods

The designed CA module generates the features from the channel level, and the PA module generates the features from pixel level. Thus, fusing both of the two features could enrich the discrimination of the feature learning model. In this section, we explore the effectiveness of three fusion methods (addition, concatenation, nonlinear fusion) on all datasets. The detailed results are shown in Fig. 8. On the UC Merced dataset, the OA

TABLE II
EFFECTIVENESS OF DIFFERENT MODULES ON THE UC MERCED, AID AND NWPU-RESISC45 DATASET

Dataset	MS module	CA module	PA module	Accuracy(%)
UC Merced	×	×	×	96.88 ± 0.32
	✓	×	×	97.09 ± 0.22
	×	✓	×	97.21 ± 0.63
	×	×	✓	97.13 ± 0.32
	✓	✓	×	98.16 ± 0.13
	✓	×	✓	97.83 ± 0.25
	×	✓	✓	98.35 ± 0.42
	✓	✓	✓	**98.96 ± 0.21**
AID	×	×	×	93.02 ± 0.27
	✓	×	×	93.13 ± 0.26
	×	✓	×	94.33 ± 0.37
	×	×	✓	93.87 ± 0.44
	✓	✓	×	95.31 ± 0.51
	✓	×	✓	94.73 ± 0.25
	×	✓	✓	95.75 ± 0.32
	✓	✓	✓	**96.01 ± 0.43**
NWPU-RESISC45	×	×	×	89.83 ± 0.31
	✓	×	×	90.01 ± 0.34
	×	✓	×	90.32 ± 0.45
	×	×	✓	90.89 ± 0.22
	✓	✓	×	91.25 ± 0.37
	✓	×	✓	91.74 ± 0.65
	×	✓	✓	92.03 ± 0.33
	✓	✓	✓	**92.52 ± 0.23**

is $(98.37 \pm\ 0.39\%)$, $(98.96 \pm\ 0.21\%)$, and $(97.25 \pm\ 0.64\%)$ for addition, concatenation, and nonlinear fusion methods, respectively. Meanwhile, the OA on the AID dataset is $(95.77 \pm\ 0.51\%)$, $(96.01 \pm\ 0.43\%)$, and $(95.32 \pm\ 0.37\%)$. On NWPU-RESISC45 dataset, it achieves $(92.01 \pm\ 0.46\%)$, $(92.52 \pm\ 0.23\%)$, and $(91.79 \pm\ 0.56\%)$ with the three fusion methods, respectively. The performance on these three datasets demonstrates that the concatenation fusion method performs best compared with the other two fusion ones. Furthermore, the addition method is slightly better than the nonlinear fusion method, and it could be the reason that some spatial correspondences between the two generated attention features are decreased at the FC layers. Since the concatenation fusion method performs best, we use it as the final fusion method of our model.

G. Classification of the UC Merced Dataset

In order to evaluate the effectiveness of the proposed MSA-Network, we compare it with other state-of-the-art methods on the UC Merced dataset. The detailed results are presented in Table III. In [3], [4], [10], [21], [23], [32], [34], [44], [58], researchers used pretrained CNNs to boost the performance of the models. In [23], [47], [64], [74], bag-of-visual-words (BoVW) was applied to land-use classification. In [58] and [68], the extreme learning machine classifier was applied for final classification with fused features. In [34] and [66], they constructed a pyramid image to characterize both the photometric and geometric aspects of an image. In [1], [23], [60], images with the LBP based texture, local, and global information was employed to convey the comprehensive message to the models. From this table, we can see that the proposed MSA-Network has gained $(98.96 \pm\ 0.21\%)$ and $(97.80 \pm\ 0.33\%)$ classification accuracy for 80% and 50% training ratios, separately. It is noteworthy that the OA of 80% training ratio is better than the 50% training data ratio, it can be explained that with more training data samples, the network could learn more high-level and abstract features, which further enhance the classification performance. Overall, the experimental results demonstrate that our MSA-Network could gain state-of-the-art classification performance on the UC Merced dataset. The detailed confusion matrix with the training ratio of 80%

TABLE III
COMPARISONS OF OA (%) ON THE UC MERCED DATASET

Method	80% Training Ratio	50% Training Ratio
SPCK [66]	73.14	/
SPCK+ [66]	76.05	/
SPCK++ [66]	77.38	/
BoVW [3]	76.81	/
BoVW + SCK [64]	77.71	/
Saliency-Guided [72]	82.72 ± 1.18	/
VGG-VD-16 [59]	95.21 ± 1.20	94.14 ± 0.69
CaffeNet [59]	95.02 ± 0.81	93.98 ± 0.67
SRSCNN [37]	95.57	/
CNN-ELM [58]	95.62	/
salM^{3}LBP-CLM [60]	95.75 ± 0.80	94.21 ± 0.75
TEX-Net-LF [1]	97.72 ± 0.54	96.91 ± 0.36
Appearance-based [48]	96.05 ± 0.62	/
LGFBOVW [74]	96.88 ± 1.32	/
GoogLeNet [59]	94.31 ± 0.89	92.70 ± 0.60
Fusion by addition [4]	/	97.42 ± 1.79
DCA fusion [4]	/	96.90 ± 0.09
CCP-net [47]	97.52 ± 0.97	/
Two-Stream Fusion [68]	98.02 ± 1.03	96.97 ± 0.75
DSFATN [10]	98.25	/
Deep CNN Transfer [21]	98.49	/
ResNet [50]	98.50 ± 1.40	/
Aggregate strategy [34]	97.40	96.25
SHHTFM [75]	98.33 ± 0.98	/
VGG-VD16+AlexNet [32]	98.81 ± 0.38	/
FACNN [39]	98.81 ± 0.24	/
CTFCNN [23]	98.44 ± 0.58	/
SAFF [2]	97.02 ± 0.78	/
SSRL [63]	94.05 ± 0.12	/
GBNet [55]	98.57 ± 0.48	97.05 ± 0.19
InceptionV3+Xception[46]	97.86 ± 0.59	96.60 ± 0.65
MSA-Network	**98.96 ± 0.21**	**97.80 ± 0.33**

is illustrated in Fig. 9. The confusion matrix result shows that most of the categories have achieved 100% classification accuracy.

H. Classification of the AID Dataset

Table IV shows the results with other state-of-the-art methods on the AID dataset. The FACNN in [39] combined the feature learning, feature aggregation, and classifier for joint training. In [67], they utilized a multilevel fusion method, which can make a judgment by incorporating different levels' information. In [4], combined with the SIFT feature, the deep learning feature can get a discriminative image presentation which overcoming the scale and rotation variability. The result in Table IV shows that our model has achieved $(93.53 \pm 0.21\%)$ classification accuracy with 20% training data, and $(96.01 \pm 0.43\%)$ classification accuracy with 50% training data, which is the best performance compared with all competing methods. Overall, the result demonstrates that our MSA-Network could achieve state-of-the-art classification performance on the AID dataset. The detailed confusion matrix with a fixing training ratio of 50% is illustrated in Fig. 10. The result shows that most of the categories have gained high classification accuracy, which is above 90%.

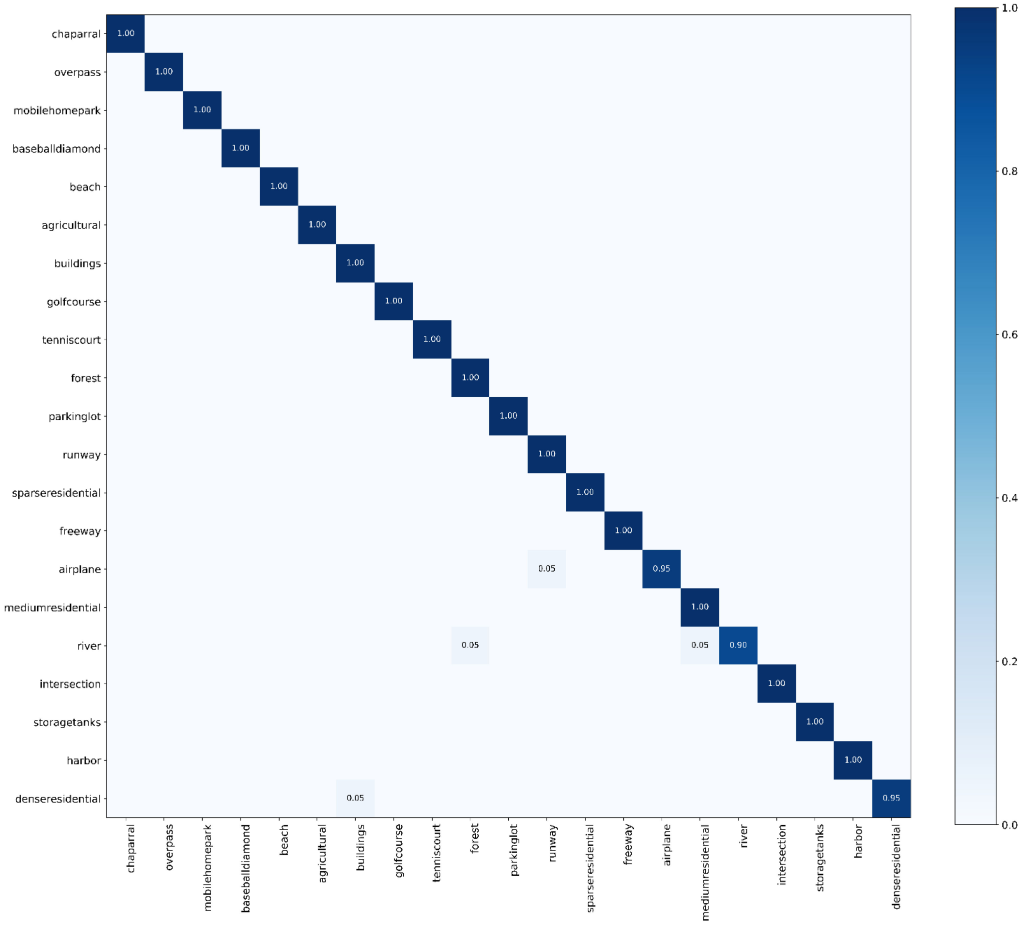

Fig. 9. Confusion matrix on UC Merced dataset with training data of 80%.

TABLE IV
COMPARISONS OF OA (%) ON THE AID DATASET

Method	50% Training Ratio (%)	20% Training Ratio
BoVW [60]	/	78.66 ± 0.52
MS-CLBP+FV [60]	/	86.48 ± 0.27
CaffeNet [59]	89.53 ± 0.31	86.86 ± 0.47
GoogLeNet [59]	86.39 ± 0.55	83.44 ± 0.40
VGG-VD-16 [59]	89.64 ± 0.36	86.59 ± 0.29
DCA fusion [4]	89.71 ± 0.33	/
$salM^3LBP - CLM$ [60]	89.76 ± 0.45	/
TEX-Net-LF [1]	95.73 ± 0.16	/
Fusion by addition [4]	91.87 ± 0.36	/
Bidirectional adaptive feature fusion [38]	93.56	/
Multilevel fusion [67]	94.17 ± 0.32	/
Two-Stream Fusion [69]	94.58 ± 0.25	92.32 ± 0.21
FACNN [39]	95.45 ± 0.11	/
CTFCNN [23]	94.91 ± 0.24	/
LCNN-BFF [53]	94.62 ± 0.16	91.66 ± 0.48
SAFF [2]	93.83 ± 0.28	90.25 ± 0.29
GBNet [55]	95.52	/
MSA-Network	**96.01 ± 0.43**	**93.53 ± 0.21**

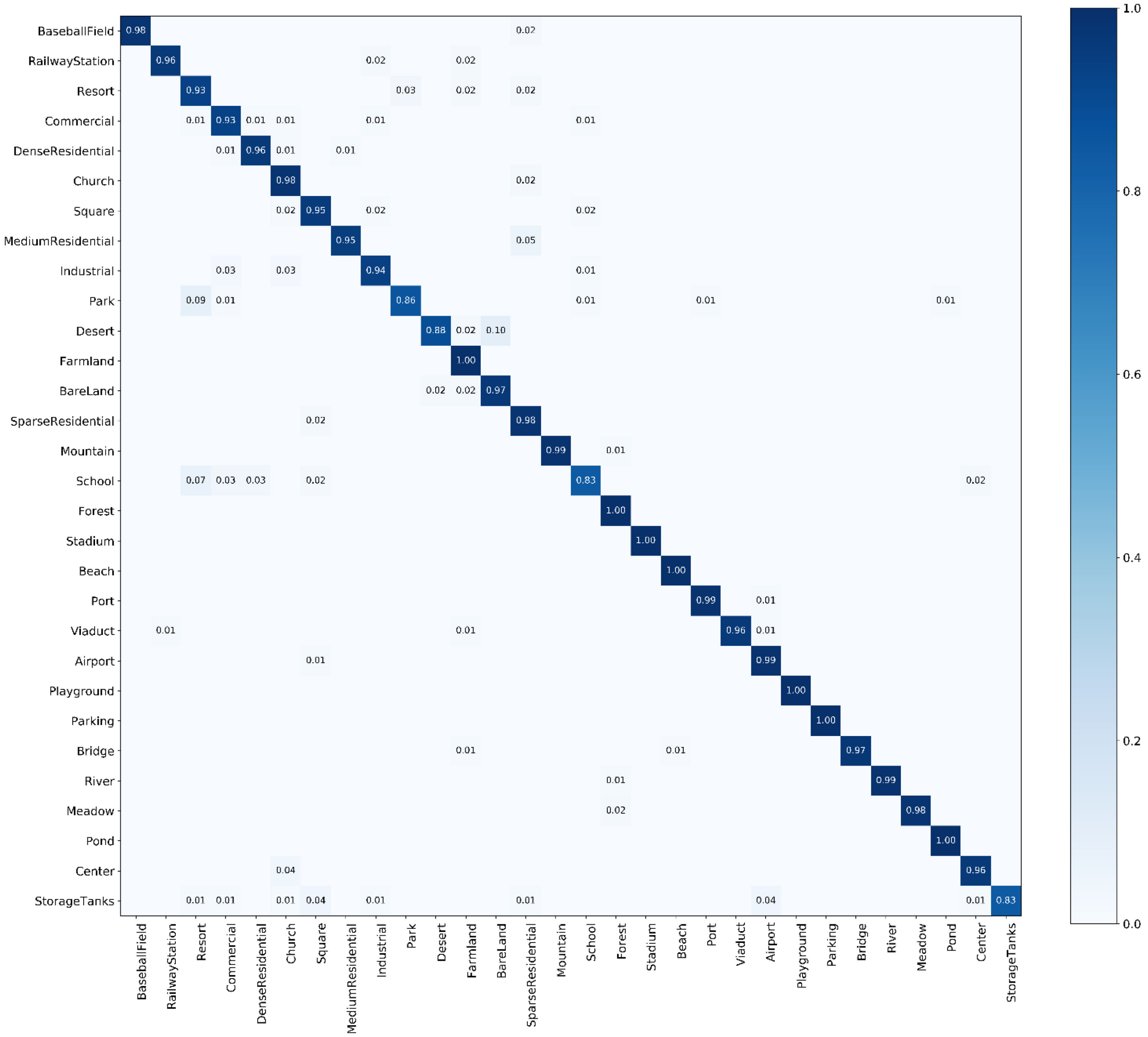

Fig. 10. Confusion matrix on AID dataset with training data of 50%.

I. *Classification of the NWPU-RESISC45 Dataset*

For the NWPU-RESISC45 dataset, the result over other state-of-the-art methods is shown in Table V. In [36], the lecture focused on the four new loss functions to achieve better performance. Based on a residual network and dense convolutional networks in [42], it achieved a competitive result. Using a metric learning regularization term, the Siamese CNN proposed in [35] was also very robust and effective. it notes that the proposed architecture in [35] replaces the final FC layer with a convolutional layer to predict the corresponding label of each class. The result demonstrates that our method achieves the highest classification accuracy compared with other state-of-the-art methods. For the 20% training data, our model achieves $(93.52 \pm \ 0.21\%)$ classification accuracy, and for the 10% training data, it has gained the classification accuracy of $(90.38 \pm \ 0.17\%)$. Especially, compared with the other state-of-the-art network architectures (e.g., AlexNet, GoogLeNet, VGG-16), the proposed MSA-Network could gain better performance. It could be explained from two folds: first, our designed network is based on ResNet, which has a deeper network architecture to learn more high-level and semantic features; second, with our designed MS module and CPA module, the model is apt to extract more multiscale and salient features, which further improve the classification performance of the model. The confusion matrix of 20% training data is illustrated in Fig. 11, indicating that our model achieves competitive classification results in most of the categories.

IV. Qualitative Analysis

A. *Multiscale Feature Map Visualization*

In this section, we visualize some examples of the MS module feature maps on different datasets in Fig. 12. Since different depths of layers could contain various semantic features, we select the feature maps from the four depth layers of our designed architecture. Here, we denote the “MS_module layer_1,” “MS_module layer_2,” “MS_module layer_3,” nd “MS_module

TABLE V
COMPARISONS OF OA (%) ON THE NPWU-RESISC45 DATASET

Method	20% Training Ratio (%)	10% Training Ratio
LASC-CNN[70]	84.30	81.37
Two-Stream Fusion [69]	83.16 ± 0.18	80.22 ± 0.22
BoCF [8]	84.32 ± 0.17	82.65 ± 0.31
AlexNet [8]	85.16 ± 0.18	81.22 ± 0.19
GoogLeNet [8]	86.02 ± 0.18	82.57 ± 0.12
VGG-16 [8]	90.36 ± 0.18	87.15 ± 0.45
Triple networks [36]	92.33 ± 0.20	/
ResNet [42]	91.96 ± 0.71	89.24 ± 0.75
Siamese_ResNet [35]	92.28	/
LCNN-BFF [53]	91.73 ± 0.17	86.53 ± 0.15
SAFF [2]	87.86 ± 0.14	84.38 ± 0.19
SSRL [63]	83.12 ± 0.26	/
MF^{2}Net [62]	92.73 ± 0.21	90.17 ± 0.25
MSA-Network	**93.52 ± 0.21**	**90.38 ± 0.17**

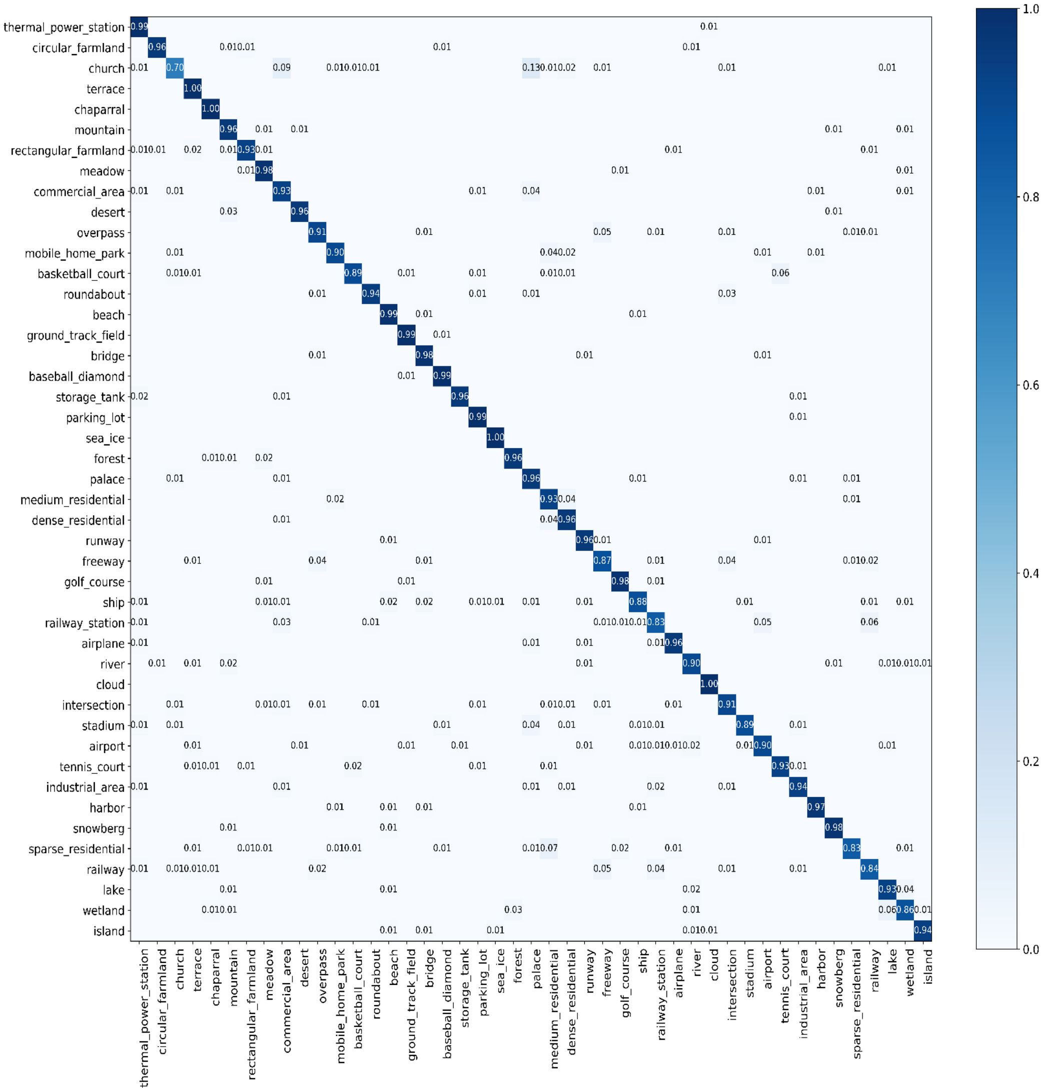

Fig. 11. Confusion matrix on NPWU-RESISC45 dataset with training data of 20%.

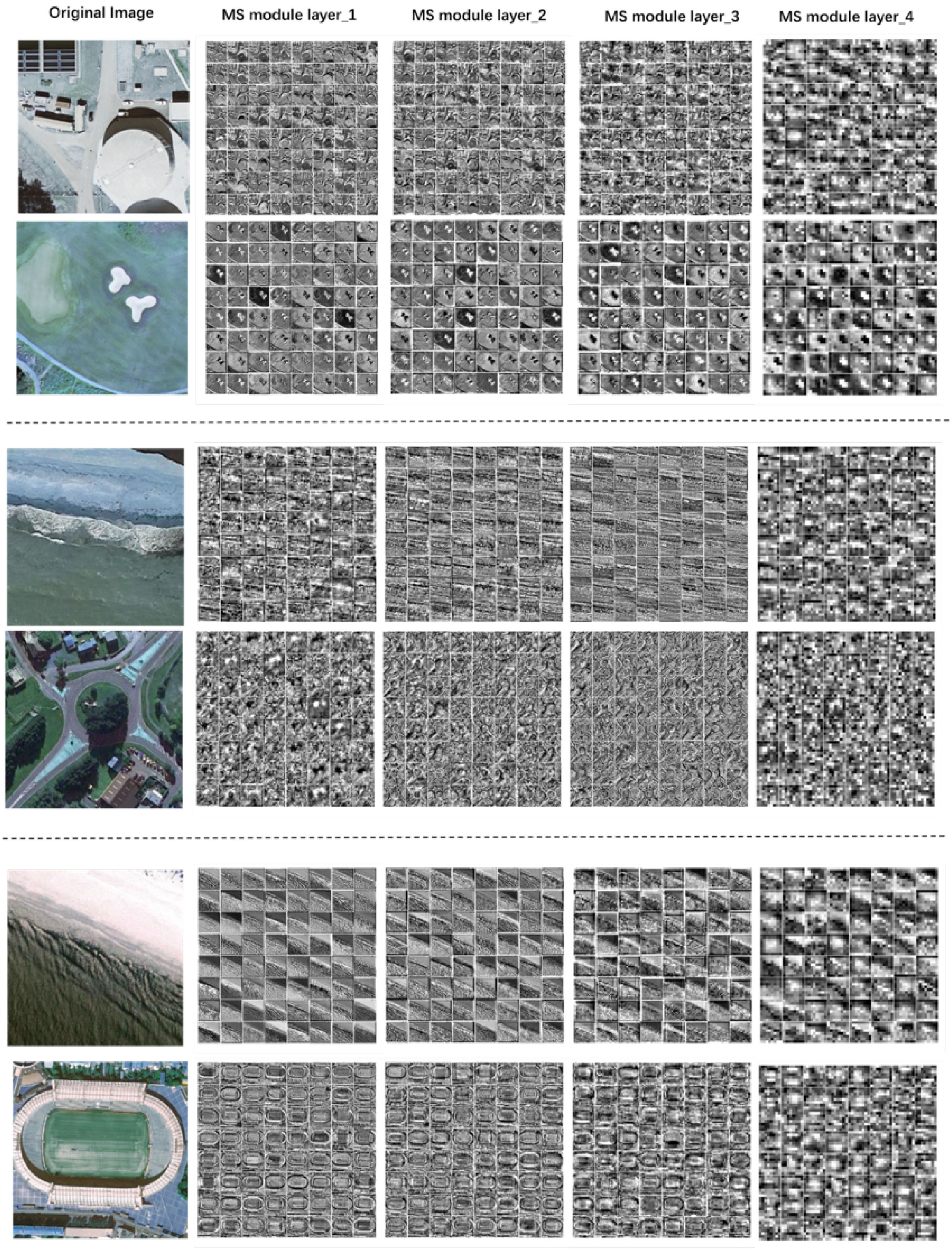

Fig. 12. Some examples of the MS module feature maps on different datasets, the images on the top two rows are samples from the UC Merced dataset, the images on the middle two rows are samples from the NPWU-RESISC45 dataset, and the images on the bottom two rows are the samples from AID dataset. The "MS_module layer_1," "MS_module layer_2," "MS_module layer_3," and "MS_module layer_4" as the layers from the first, second, third, and fourth of the MS module layer, respectively.

layer_4" as the layers from the first, second, third, and fourth of the MS module layer, respectively. From the visualization results, we can see that the MS module could extract diverse features through different sizes of filters, and the shallower depth layer tends to extract more edge and subtle representations, while the deeper depth layer is liable to learn more high-level and abstract features. Overall, with the designed MS module, the designed network could encode more discriminative features from different levels, which further improve the classification performance of the model.

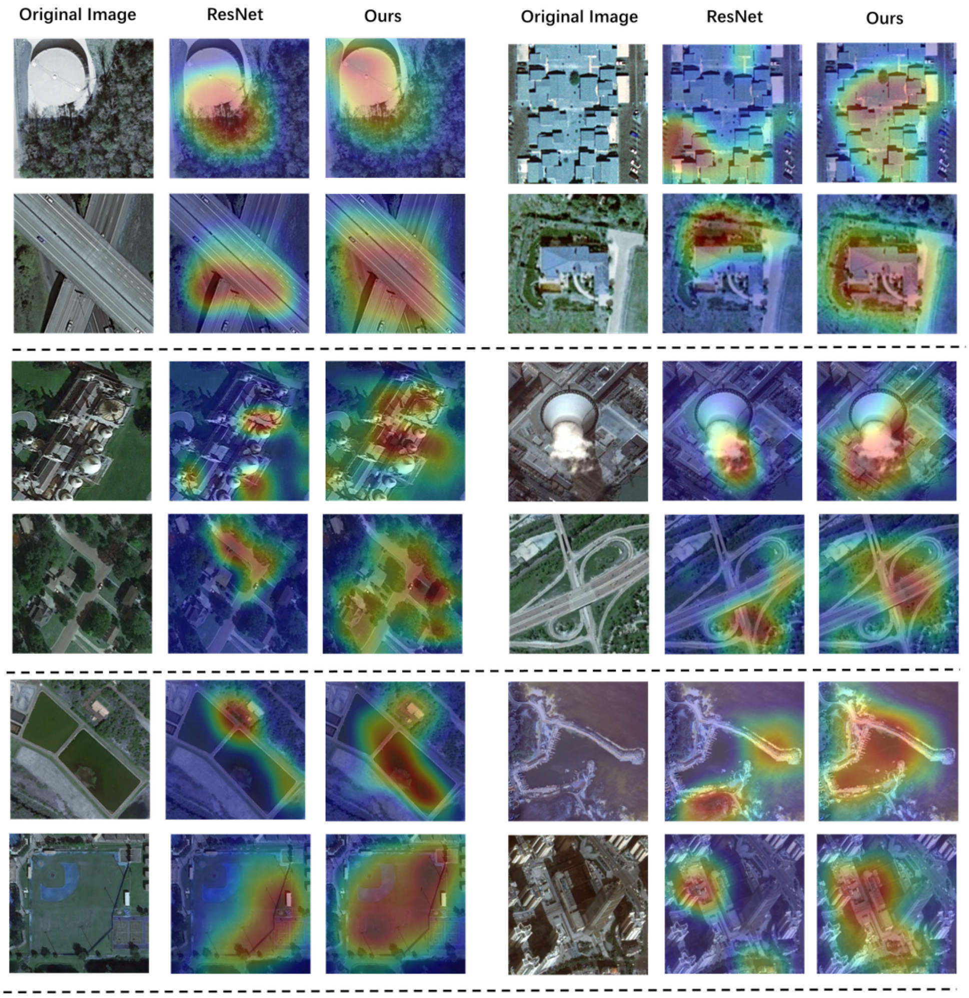

Fig. 13. Some examples of the attention maps on different datasets, the images on the top two rows are samples from the UC Merced dataset, the images on the middle two rows are samples from the NPWU-RESISC45 dataset, and the images on the bottom two rows are the samples from the AID dataset.

B. *Visualization of the Attention Map*

In this section, we show some examples of attention maps on different datasets. For better comparison, we compare our model with the ResNet architecture (depth with 101). The detailed visualization results are illustrated in Fig. 13. Compared with the ResNet visualization results, it is obvious that our designed model can attend to the more crucial regions, especially on the category relevant regions. Meanwhile, the designed model could provide more diverse and detailed region features, which could further enhance the final classification performance. We suggest that adding the CA module, could guide the model to focus on more global regions while adding the PA module, tends to guide the model to focus on more subtle regions, thus, fusing both the CA module and PA module could efficiently improve the model ability to learn more crucial and salient features.

V. Conclusion

In this article, we propose an MSA-Network to handle the remote sensing scene image classification task, in which the MSA-Network uses the ResNet and incorporates an MS module and CPA module to further improve the performance of the designed architecture. The proposed MS module extracts multiscale features from different receptive fields with various sizes of sliding windows. Moreover, we add the MS module behind each stage's last residual block to extract the multiscale features hierarchically. The CPA module consists of two parts: the CA module and the PA module. The CA module aims to extract the attention features from channel-level globally, while the PA module learns the attention features from pixel-level locally. With those two attention modules, the proposed MSA-Network could focus on more informative and critical regions globally and locally. Experimental results on UC Merced, NWPU-RESISC45, and AID datasets demonstrate that the proposed MSA-Network could achieve better performance over several state-of-the-arts on the overall classification accuracy. Furthermore, we also conduct relevant experiments on our designed modules to further analyze the effectiveness of each module. In future work, we will try to explore the designed MS module and CPA module with deeper network architecture such as DenseNet to validate the effectiveness of our modules. Meanwhile, some other techniques such as ensemble more models with early fusion or late fusion may further improve the classification accuracy.

ACKNOWLEDGMENT

The authors would like to thank the paper reviewers. And also thank the associate editor for supplying valuable comments and suggestions.

REFERENCES

[1] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 74–85, 2018.

[2] Ran Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2021.

[3] C. Marco, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.

[4] 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.

[5] L. Chen, Y. Yang, J. Wang, W. Xu, and A. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[6] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[7] G. Cheng, P. Zhou, J. Han, L. Guo, and J. Han, "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images," *Comput. Vis. Letter*, vol. 9, no. 5, pp. 639–647, 2015.

[8] 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.

[9] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[10] X. Gong, Z. Xie, Y. Liu, X. Shi, and Z. Zheng, "Deep salient feature based anti-noise transfer network for scene classification of remote sensing imagery," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 410.

[11] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 848.

[12] S. Hao, S. Xian, H. Wang, L. Yu, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 1, pp. 109–113, Jan. 2012.

[13] 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.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[15] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[16] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[17] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 35–49, 2019.

[18] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[19] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. Zhu, "Learnable manifold alignment (LEMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 193–205, 2019.

[20] D. Hong, N. Yokoya, G. Xia, J. Chanussot, and X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.

[21] F. Hu, G. 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.

[22] Y. Hu, "Design and implementation of abnormal behavior detection based on deep intelligent analysis algorithms in massive video surveillance," *J. Grid Comput.*, vol. 18, no. 2, pp. 227–237, 2020.

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

[24] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 483.

[25] Z. Huang, J. Tang, G. Shan, J. Ni, Y. Chen, and C. Wang, "An efficient passenger-hunting recommendation framework with multitask deep learning," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7713–7721, Oct. 2019.

[26] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50× fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[27] D. Jia, D. Wei, R. Socher, L. Li, L. Kai, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[29] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[31] L. Li and N. Shu, "Object-oriented classification of high-resolution remote sensing image using structural feature," in *Proc. 3rd Int. Congr. Image Signal Process.*, 2010, pp. 2212–2215.

[32] 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.

[33] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[34] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[35] X. Liu, Y. Zhou, J. Zhao, R. Yao, Bing Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Let.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.

[36] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Jan. 2018.

[37] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 444.

[38] X. Lu, W. Ji, X. Li, and X. Zheng, "Bidirectional adaptive feature fusion for remote sensing scene classification," *Neurocomputing*, vol. 328, pp. 135–146, 2019.

[39] 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.

[40] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.

[41] M. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.

[42] R. Minetto, M. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for Geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 2018.

[43] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for Geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 22019.

[44] L. Na, X. Lu, L. Wan, H. Hong, and F. Tao, "Improving the separability of deep features with discriminative convolution filters for RSI classification," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 3, 2018, Art. no. 95.

[45] M. Nauman, H. U. Rehman, G. Politano, and A. Benso, "Beyond homology transfer: Deep learning for automated annotation of proteins," *J. Grid Comput.*, vol. 17, no. 2, pp. 225–237, 2019.

[46] B. Petrovska, E. Zdravevski, P. Lameski, R. Corizzo, I. Štajduhar, and J. Lerga, "Deep learning for feature extraction in remote sensing: A. case-study of aerial scene classification," *Sensors*, vol. 20, no. 14, 2020, Art. no. 3906.

[47] K. Qi, Q. Guan, Y. Chao, F. Peng, S. Shen, and H. Wu, "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 934.

[48] K. Qi, C. Yang, Q. Guan, H. Wu, and J. Gong, "A multiscale deeply described correlatons-based model for land-use scene classification," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 917.

[49] J. Santos, O. Penatti, and R. Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, Volume 2: VISAPP, Angers, France, 2010, pp. 203–208.

[50] G. Scott, M. England, W. Starms, R. Marcum, and C. Davis, "Training deep convolutional neural networks for land-cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.

[51] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, 2014, Art. no. 083584.

[52] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.

[53] C. Shi, T. Wang, and L. Wang, "Branch feature fusion convolution network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5194–5210, Aug. 2020.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] 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.

[56] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[57] P. Tokarczyk, J. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jan. 2015.

[58] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 704–708, May 2017.

[59] G. Xia, J. Hu, F. Hu, B. Shi, and L. Zhang, "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.

[60] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.

[61] G. Xu, M. Liu, Z. Jiang, D. Söffker, and W. Shen, "Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning," *Sensors*, vol. 19, no. 5, 2019, Art. no. 1088.

[62] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1894–1898, Nov. 2020.

[63] P. Yan, F. He, Y. Yang, and F. Hu, "Semi-supervised representation learning for remote sensing image classification based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 54135–54144, 2020.

[64] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Sigspatial Int. Conf. Adv. Geographic Inf. Syst.*, 2010, paper no. 270.

[65] Y. Yang and S. Newsam, "Comparing sift descriptors and gabor texture features for classification of remote sensed imagery," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1852–1855.

[66] Y. Yi and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.

[67] Y. Yu and F. Liu, "Aerial scene classification via multilevel fusion based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 287–291, Feb. 2018.

[68] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, 2018, no. 4–5, pp. 1–13, 2018.

[69] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, vol. 2018, 2018, Art. no. 8639367.

[70] B. Yuan, S. Li, and N. Li, "Multiscale deep features learning for land-use scene recognition," *J. Appl. Remote Sens.*, vol. 12, no. 1, 2018, Art. no. 015010.

[71] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[72] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[73] S. Zhao, D. Zhang, and H. Huang, "Deep learning-based image instance segmentation for moisture marks of shield tunnel lining," *Tunnelling Underground Space Technol.*, vol. 95, 2020, Art. no. 103156.

[74] Q. Zhu, Y. Zhong, B. Zhao, and G. Xia, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[75] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous-heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.

[76] T. Zia and S. Razzaq, "Residual recurrent highway networks for learning deep sequence prediction models," *J. Grid Comput.*, vol. 18, no. 1, pp. 169–176, 2020.

[77] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Inf. Sci.*, vol. 348, pp. 209–226, 2016.