A Lightweight and Discriminative Model for Remote Sensing Scene Classification With Multidilation Pooling Module

Bin Zhang, Yongjun Zhang, and Shugen Wang

***Abstract*—With the growing spatial resolution of satellite images, high spatial resolution (HSR) remote sensing imagery scene classification has become a challenging task due to the highly complex geometrical structures and spatial patterns in HSR imagery. The key issue in scene classification is how to understand the semantic content of the images effectively, and researchers have been looking for ways to improve the process. Convolutional neural networks (CNNs), which have achieved amazing results in natural image classification, were introduced for remote sensing image scene classification. Most of the researches to date have improved the final classification accuracy by merging the features of CNNs. However, the entire models become relatively complex and cannot extract more effective features. To solve this problem, in this paper, we propose a lightweight and effective CNN which is capable of maintaining high accuracy. We use MobileNet V2 as a base network and introduce the dilated convolution and channel attention to extract discriminative features. To improve the performance of the CNN further, we also propose a multidilation pooling module to extract multiscale features. Experiments are performed on six datasets, and the results verify that our method can achieve higher accuracy compared to the current state-of-the-art methods.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), dilated convolution, remote sensing image, scene classification.**

I. Introduction

REMOTE sensing technology has developed rapidly in recent years, and a variety of remote sensing platforms and sensors are used to observe the earth. As a result, the volume of image has increased dramatically and the spatial resolution has been continuously improved. High spatial resolution (HSR) images can provide abundant information about the shape, texture, and other features of the object of interest, which are helpful for improving the accuracy of object recognition. Nowadays, many satellites can provide remote sensing images with spatial resolution up to submeters, which triggers an important need: whether land use and coverage categories can be identified intelligently from remote sensing images. Due to the need for remote sensing applications, it is particularly important to understand the semantic content of images effectively. In this paper, we mainly discuss remote sensing scene classification, which automatically assigns a semantic tag to each remote sensing image [1]. In the past few decades, scene classification has played an important role in a wide range of applications, such as land use and cover classification, geographic target detection, geographic image retrieval, and urban planning [2].

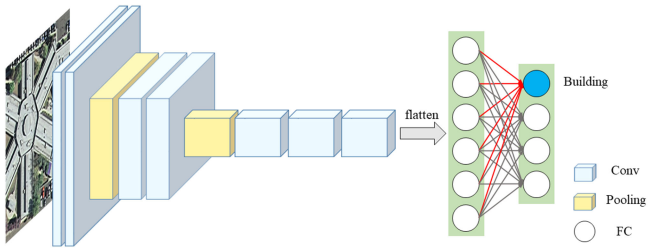


Fig. 1. Scene classification based on a CNN.

Considerable progress has been made in the past in the development of scene classification methods [1]–[23]. In order to obtain the high accuracy results, effective and discriminative features representation play an important role. Existing scene classification methods can be divided into three categories according to the features they use: handcrafted feature-based methods, unsupervised-feature-learning-based methods, and deep-learning-based methods [2]. These three categories are not necessarily independent of each other and sometimes a method involves two or even three of them.

Since the study [24] has shown that convolutional neural networks (CNNs) can extract mid-to-high-level abstract features from the original images, CNNs have become the preferred model for many applications. Thus, CNNs have been introduced into remote sensing image scene classification (RESISC) task [3], [4]. As shown in Fig. 1, the hidden layers of a CNN typically consist of the convolutional layers, pooling layers, and fully connected (FC) layers. A CNN has strong generalization ability and can be well generalized to the task of RESISC even in domains considerably different from the ones for which they were trained [4]. There are typically three empirically possible strategies to take advantage of the capabilities of existing CNNs

Manuscript received September 2, 2018; revised January 15, 2019 and March 6, 2019; accepted May 17, 2019. Date of publication June 25, 2019; date of current version September 16, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003 and in part by the National Natural Science Foundation of China under Grant 41571434. *(Corresponding author: Yongjun Zhang.)*

The authors are with the Department of Photogrammetry, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: bin.zhang@whu.edu.cn; zhangyj@whu.edu.cn; wangsg@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2019.2919317

in different scenarios for the scene in which they are trained: full training, fine tuning, and using CNNs as feature extractors [8]. Due to the small data volumes of available HSR scene datasets, which is far less than that of the natural image datasets, such as ImageNet dataset [25], it is infeasible to fully design and train a new CNNs. Training a new CNN usually requires a considerable amount of labeled data and demands high computational costs. Therefore, most scholars use CNNs through fine tuning or feature extractors or both [3]–[5], [8], [10], [12]. In addition, the past literature includes studies that train a new network through data augmentation or other tricks [7], [19], [22], [26]. In recent years, various feature fusion and feature encoding methods have been used to improve the classification accuracy [13], [15]–[18], [20], [21].

However, these models have two drawbacks: 1) have a large number of parameters and become very complex; and 2) cannot extract more effective features for complex geometrical structures and spatial patterns. To solve these problems, we construct a lightweight and discriminative CNN named SE-multidilation pooling network (SE-MDPMNet). Recently, MobileNet v2 [27], an improved version of MobileNet [28], proposed an inverted residual block and improved the state-of-the-art performance of mobile models on multiple tasks and benchmarks with a good tradeoff between accuracy and the number of parameters. Therefore, we use MobileNet V2 as a base network. Also, the dilated convolution and channel attention in SENet [29] have been added to extract discriminative features. The dilated convolution can enlarge the receptive field of filters without increasing the number of parameters. The channel attention focuses on the channel relationship and can produce significant performance improvements at a minimal additional computational cost. To improve the performance of the CNN further, we propose a multidilation pooling module to extract multiscale features. Previous works were performed with only a few datasets. To verify the validity of our method, extensive experiments are performed on six datasets: UC Merced [30], WHU-RS19 [31], RSSCN7 [32], SIRI-WHU [33]–[35], AID [1], Northwestern Polytechnical University (NWPU)-RESISC45 [2]. The results show that our method can achieve higher accuracy compared with the state-of-the-art methods.

The major contributions of this paper are as follows.

1) Using the dilated convolution preserves the spatial resolution and enlarges the receptive field of filters in the classification task. Our results show that this method does not require extra parameters but effectively improved the performance of CNN.
2) A multidilation pooling module is proposed to extract multiscale features. Our results show that this module can effectively increase CNN accuracy.
3) A lightweight end-to-end deep network is proposed. Our network provides a new baseline for remote sensing scene classification.

The remainder of this paper is organized as follows. In Section II, the recent methods based on CNNs and the progress of remote sensing scene classification are addressed. In Section III, the detailed methods of the proposed scene classification model are discussed. In Section IV, the datasets and the analysis of the experimental results are presented. Section V discusses the proposed method and Section VI concludes this paper.

II. Related Work

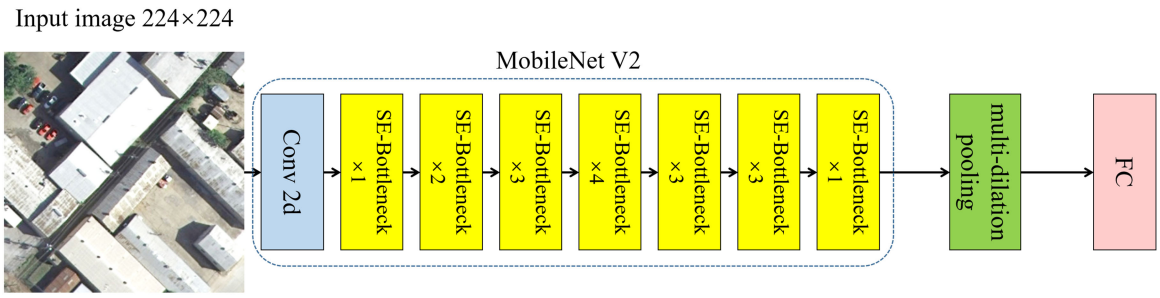
The early research works based on handcrafted features for scene classification mainly focused on using a considerable amount of engineering skills and domain expertise to design various handcrafted features, such as color, texture, shape, spatial and spectral information, or combinations. These features are the primary feature representation of an image and hence can map images into feature spaces. Some handcrafted features have been frequently used in the past: Gabor feature [36]; local binary pattern (LBP) [37]; GIST [38]; scale-invariant feature transform [39]; histogram of oriented gradients [40]; and bag-of-visual words [41]. In addition, many feature encoding methods have been proposed in the past few years, such as Fisher vector coding [34] and spatial pyramid matching [42].

Unsupervised-feature-learning-based methods can infer a function or feature to describe the hidden structures from "unlabeled" data automatically. Thus, for the tasks that do not have labeled data or only little labeled data, unsupervised learning performs better than supervised learning. Typical unsupervised feature learning methods include principal component analysis (PCA), k-means clustering, sparse coding [43], and autoencoder [11], [44].

Deep learning architectures such as deep CNNs, recurrent neural networks, and generative adversarial networks [45] have been applied in computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics and drug design, where they have produced results comparable to human and in some cases superior to human. In 2012, deep learning achieved amazing results in the field of image classification. AlexNet [46] achieved a top-5 test error rate of 15.3% in the ImageNet competition, exceeding the classification performance of the second place competitor by a large margin. As CNNs become increasingly deeper [47], [48], residual networks [49] have surpassed the 100-layer barrier.

Because deep learning has achieved amazing results in the field of natural images, these architectures have also been successfully applied to the field of remote sensing image processing, such as scene classification [5], [8], [10], [12], semantic segmentation [50]–[52], geographic object detection [53], hyperspectral image classification [54], [55], and so on. Existing scene classification methods based on CNN can be divided into two categories according to how they used CNN: 1) use CNN through fine-tuning or feature extractors or both; and 2) propose a new structure or network. Both of these two methods include some feature fusion methods or feature encoding methods to improve classification accuracy.

Hu *et al*. [3] proposed two scenarios for generating image features via extracting CNN features from the fully connected layers and the last convolutional layer at multiple scales respectively. Marmanis *et al*. [5], Li *et al*. [9], and Chaib *et al*. [13] exploited pretrained CNN as deep feature extractor to extract



Output size	112	112	56	28	28	28	28	28
Channels	32	16	24	32	64	96	160	320
Output stride	2	2	4	8	8	8	8	8
Dilation	1	1	1	1	1,2,3,4	2,3,4	2,3,4	1

Fig. 2. The general architecture of the proposed network.

informative features from original HSR images to construct the final representation of the HSR image scenes. Li *et al*. [16], Liu *et al*. [21], and Wang *et al*. [15] explored the benefits of multilayer features for improving scene classification. Qi *et al*. [14], Liu *et al*. [17], Yu and Liu [20], and Liu *et al*. [22], [56] integrated spatial information at multiple scales for land-use scene classification. Anwer *et al*. [23] used LBPs for encoding CNN models, called TEX-Nets, which provide complementary texture information to the standard RGB deep models.

Zhang *et al*. [6] proposed a gradient boosting random convolutional network (GBRCN) framework for scene classification, which can effectively combine many deep neural networks. Zhong *et al*. [57] proposed an agile CNN architecture, called SatCNN, which used smaller kernels to build an effective CNN architecture. Zhong *et al*. [69] proposed a practical CNN architecture, called the large patch CNN (LPCNN), which was used to generate hundreds of possible scene patches for the feature learning. Han *et al*. [12] improved a pretrained AlexNet architecture called pretrained AlexNet-SPP-SS, which incorporated scale pooling—spatial pyramid pooling (SPP) and side supervision (SS) to improve accuracy. Wang *et al*. [10] designed a linear PCA network to synthesize spatial information of remote sensing images in each spectral channel which shortened the spatial "distance" of target and source datasets for pretrained deep CNNs. Yu and Liu [18] proposed a two-stream deep fusion framework which combines saliency detection images and RGB images. Liu and Huang [58] proposed a novel scene classification method via triplet networks, which used weakly labeled images as network inputs. Yang *et al*. [19] proposed "DropBand," which was a simple and effective method of promoting the classification accuracy of CNNs for very-high-resolution remote sensing image scenes, whereby training samples are generated by dropping certain spectral bands out of the original images. Gong *et al*. [59] first introduced deep structural ML into the literature of remote sensing scene classification to specifically capture and use the structural information in training.

Because previous methods do not take full consideration of spatial context information and multiscale features, those network structures cannot extract more effective and discriminative features. To solve these problems, we propose a lightweight and discriminative CNN, which can achieve higher accuracy compared with the current state-of-the-art methods.

III. Proposed Method

A. General Architecture

The general architecture is shown in Fig. 2. The proposed network mainly comprises two parts: a convolutional network based on MobileNet v2 [27] and a multidilation pooling module. We use MobileNet v2 as a base network to extract deep features. Due to the small data volumes available for HSR remote sensing scene datasets, MobileNet V2 was pretrained on ImageNet [25]. Our network is a lightweight CNN. Although Zhong *et al*. [57] proposed a lightweight CNN architecture, which used smaller kernels to build a lightweight CNN architecture, the network cannot have a good feature representation. Furthermore, the author experimented on the SAT-4 and SAT-6 dataset [60] only. Unlike the original MobileNet v2 [27], we make two major changes to the network. First, the dilated convolution is used to enlarge the receptive field of filters without increasing the number of parameters. Therefore, our network encodes image of 224 × 224 pixel size into 28 × 28 feature maps. Second, the channel attention is used to perform dynamic channel-wise feature recalibration, which only increases a few parameters. These modifications can enhance the ability of the network to extract discriminative features. Then, the multidilation pooling module is applied to extract features at multiple scales and pool the feature maps to learn multiscale contextual information by concatenating the multiscale features. Finally, the classification result is obtained by a FC layer on the concatenated feature vector. Therefore, our model can be trained directly by an end-to-end manner. In Section IV, the experimental evaluations on the six remote sensing scene classification datasets demonstrate that our network is superior to the state-of-the-art methods. In the following parts, details about the MobileNet v2, dilated convolution, channel attention, and multidilation pooling module are illustrated.

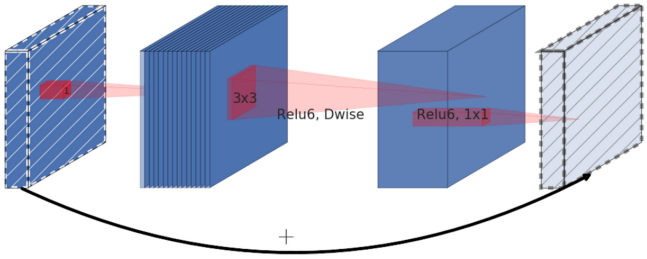


Fig. 3. Inverted residual block. Figure reproduced from [27].

B. MobileNet V2

MobileNet V2 [27] is based on many bottleneck blocks (see Fig. 2). Bottleneck block is a residual structure (see Fig. 3). The architecture of bottleneck block consists of two branches: a shortcut connection branch and a residual branch. The shortcut connection branch is the identity mapping (black line in Fig. 3). Moreover, the residual branch can decompose three convolution layers, a 1×1 convolution, a 3×3 depthwise convolution, and another 1×1 convolution (red cone in Fig. 3). The four blue blocks in Fig. 3 represent feature maps. In this residual structure, due to the number of channels is fewer at start and end of the structure and there are more channels at the middle, the shape is like a "bottleneck," so it is called the bottleneck layer. The 3×3 depthwise convolution and latter 1×1 convolution have another name, depthwise separable convolution (denote Dwise in Fig. 3). Namely, depthwise separable convolution is a form of factorized convolution, which factorizes a standard convolution into a depthwise convolution and a 1×1 convolution called pointwise convolution [28] (see Fig. 4). The depthwise separable convolution is lightweight and is a key building block for MobileNet V2. Thus, many efficient neural network architectures use the depthwise separable convolution in their networks [28], [61], [62].

A standard convolutional layer has N different convolution kernels K of $D_K \times D_K \times M$ size. It takes a $D_F \times D_F \times M$ feature map F as input. D_F is the spatial width and height of a square input feature map F, M is the number of input channels, N is the number of output channels and D_K is the spatial dimension of the kernel. If the output feature map is of the same size, the computation will be reduced to

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2}. \quad (1)$$

The architecture of MobileNet V2 [27] contains the initial fully convolution layer with 32 filters, followed by the 19 bottleneck layers described in Table I. MobileNet V2 uses 3×3 depthwise separable convolutions, which reduce eight to nine times computation than standard convolutions. MobileNet V2 can achieve the top-1 accuracy of 72.0% on ImageNet, which only has about 3.5 M parameters and 300 M multiply adds. Therefore, MobileNet V2 is chosen as our basic model. In our network, we take out the last 1×1 convolution. In our intuition, the 1×1 convolution just nonlinearly maps features to 1280 dimensions; thus, we take it out to reduce the number of parameters. Our experiment proves this modification has a very small effect on the network performance.

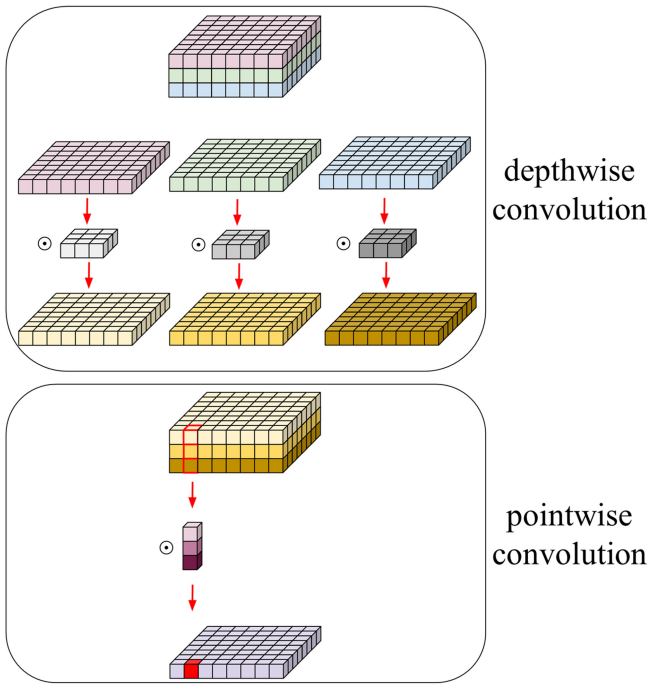


Fig. 4. Depthwise separable convolution can decompose into a depthwise convolution and a 1×1 convolution called pointwise convolution.

TABLE I
ARCHITECTURE OF MOBILENET V2 IN DETAIL [27]

Input	Operator	Expansion factor	output channels	Repeated times	stride
224×224×3	conv2d 3×3	-	32	1	2
112×112×32	bottleneck	1	16	1	1
112×112×16	bottleneck	6	24	2	2
56×56×24	bottleneck	6	32	3	2
28×28×32	bottleneck	6	64	4	2
14×14×64	bottleneck	6	96	3	1
14×14×96	bottleneck	6	160	3	2
7×7×160	bottleneck	6	320	1	1
7×7×320	Conv2d 1×1	-	1280	1	1
7×7×1280	Avgpool 7×7	-	-	1	-
1×1×1280	Conv2d 1×1	-	k	-	-

C. Dilated Convolution

In image classification task, convolutional networks usually progressively reduce resolution until the feature maps are tiny. Therefore, the spatial structure of the scene in the feature map is no longer discernible. However, preserving the contribution of small and thin objects may be important for correctly understanding the content of the image. Such loss of spatial location can limit image classification accuracy. For complex natural

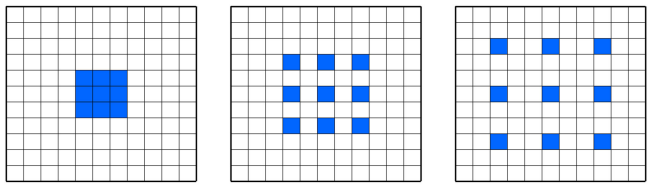


Fig. 5. Two-dimensional dilated convolution with different dilation rate r. (a) $r = 1$. (b) $r = 2$. (c) $r = 3$.

scenes, multiple objects and their relative position must be considered. In remote sensing images, scenes are more varied and diverse. This problem can be alleviated by the dilated convolution, which increases the resolution of output feature maps and expands receptive fields without losing resolution [63]. However, it requires additional memory and time. In recent years, the dilated convolution has been widely used in the semantic segmentation networks, such as [63]–[70]. To our best knowledge, this is the first time to use the dilated convolution in remote sensing scene classification.

In 1-D, the dilated convolution is defined as

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k] \tag{2}$$

where $y[i]$ is the output of dilated convolution, $x[i]$ is the input signal of the dilated convolution, and $w[k]$ is a filter of length$K$. The parameter r is the dilation rate, which corresponds to the stride with which we sample the input signal. Obviously, in standard convolution, $r = 1$.

In 2-D, the dilated convolution is constructed by inserting "holes" (zeros) between each pixel in the convolutional kernel (see Fig. 5). For a convolution kernel with size $k \times k$, the dilated convolution with rate r introduces $r - 1$ zeros between consecutive filter values, effectively enlarging the kernel size to $k_d = k + (k - 1) \cdot (r - 1)$ without increasing the number of parameters. State-of-the-art CNNs typically employ spatially small convolution kernels (typically 3×3) in order to keep both computation and the number of parameters. In Fig. 5, we show three 3×3 2-D dilated convolution kernels with different dilation rates. In Fig. 5, the convolution kernels have receptive field of 3×3, 5×5, and 7×7, respectively.

The dilated convolution is used to maintain high resolution of feature maps through replacing the max-pooling layer or stride convolution layer while maintaining the receptive field (or field of view) of the corresponding layer. For example, if a convolution layer has a stride $s = 2$, then the stride is set to 1 to remove downsampling, and the dilation rate r is set to 2 for all convolution kernels of subsequent layers. This process is applied iteratively through all layers that have a downsampling operation; thus, the feature map in the output layer can maintain the same resolution. It thus offers an efficient mechanism to control the receptive field and finds the best tradeoff between localization and context.

However, the using of dilated convolutions can cause gridding artifacts [67], [69], [70]. Since the dilated convolution introduces zeros in the convolutional kernel, the actual pixels that participate in the computation from the $k_d \times k_d$ region are just $k \times k$, with a gap of $r - 1$ between them. As a result, a kernel can only view information in a checkerboard fashion, and loses a large portion of information. When r becomes large in the higher layers due to additional downsampling operations, the sample from the input can be very sparse, which may not be good for learning, because 1) local information is completely missing; and 2) the information can be irrelevant across large distances [67].

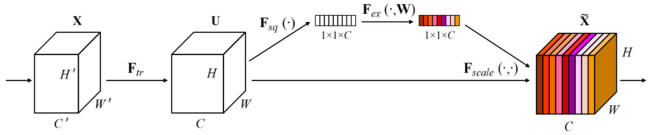


Fig. 6. SE block. Figure reproduced from [29].

In order to avoid grid effects, we use the hybrid dilated strategy [67], [69], [70] in our network. The key of the hybrid dilated strategy is not having a common factor. Instead of using the same dilation rate for all layers after the downsampling occurs, we use the different dilation rate for each layer. Specifically, the original Mobilenet V2 [27] has seven groups of bottleneck block (see Table I). We begin using the dilated convolution form the fourth bottleneck block. The stride of the fourth bottleneck block is set to 1 and the dilation rates of four depthwise convolutions in the fourth bottleneck block are set to 1, 2, 3, and 4, respectively. Therefore, for an input image with a size of 224×224, the output shape of the fourth bottleneck block 28×28 is the same as previous bottleneck block (see Fig. 2). Similarly, in the fifth and the sixth bottleneck block, the dilation rates are set to 2, 3, and 4 and the stride in the sixth bottleneck block is set to 1. Finally, the output shape of our network before multidilation pooling module can maintain 28×28 in size.

D. Channel Attention

To help the network get a more robust feature representation, we add a reweighting layer to tackle this issue and exploit channel dependencies efficiently. SENet [29] focuses on the channel relationship and proposes a novel architectural unit, squeeze-and-excitation (SE) block (see Fig. 6), which adaptively recalibrates the channel-wise feature maps by explicitly modeling the interdependences between channels. The feature maps U are passed through a squeeze operation, which aggregates the feature maps across spatial dimensions $H \times W$ to produce a channel descriptor. This descriptor embeds the global distribution of the channelwise feature responses, enabling information from the global receptive field of the network to be leveraged by its lower layers. This is achieved by using global average pooling to generate channelwise responses. Formally, the squeeze operation is achieved by

$$z_c = F_{\mathrm{sq}}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3}$$

where $u_c(i, j)$ refers to the pixel value at the position (i, j) in channel c, F_{sq} refers to the squeeze operation, and z_c refers to response in channel c.

Then, the squeeze operation is followed by an excitation operation, in which the activations, learned for each channel by attention mechanism, produces the weight of each channel. To reduce block complexity and increase generalization ability, the attention mechanism is parameterized by forming two FC layers. The first FC layer uses ReLU as the activation function and the second layer uses sigmoid

$$s = F_{ex}(z, W) = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \tag{4}$$

where W_1 and W_2 refer to weight in FC layer, σ refers to sigmoid activation, F_{ex} refers to the squeeze operation, and s refers to channelwise weights.

Finally, the feature maps U are reweighted to generate the output of the SE-block [see (5)]. The reweighted feature maps are fed directly into subsequent layers

$$\tilde{x} = F_{\text{scale}}(u, s) = s_c \cdot u_c \tag{5}$$

where s_c and u_c refer to weight and feature maps in channel c, F_{scale} refers to reweighted operation, and $\tilde{x}$ refers to reweighted feature maps.

The entire SE-block [29] can be seen as a dynamic feature extraction mechanism, which handles some layers through channelwise attention to pay more attention to the feature maps which are helpful for classification. Thus, the SE-block [29] can improve the representational capacity of a network, which is useful for remote sensing scene classification. For example, when a scene is predicted as aircraft, the channelwise attention increases the weight of the feature map which is highly related to aircraft's characteristics. The generation of high-level feature maps depends on the low-level feature maps; therefore, it is natural to consider multilayer feature maps. For example, only the low-level convolution kernel extracts more aircraft edge features, and the high-level maps can better abstract the characteristics of the aircraft.

In our network, we use the SE-block in every bottleneck block (see Fig. 2). Fig. 7 (left) shows the original bottleneck block and Fig. 7 (right) depicts the schema of the SE-bottleneck module. SE-block modules are added after each bottleneck where SE-block transformation F_{tr} is taken by a bottleneck module. Both SE are employed after summation with the identity branch. Our SE-bottleneck module is different from the original SE-block [29]. After the first FC layer, batch normalization [71] is added, which can accelerate training of SE-block by reducing internal covariate shifts.

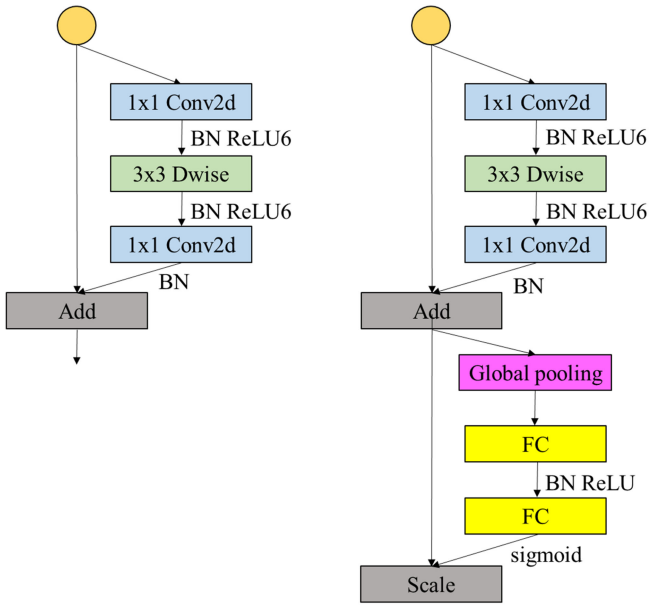


Fig. 7. Schematic of the original bottleneck block (left), SE-bottleneck block (right).

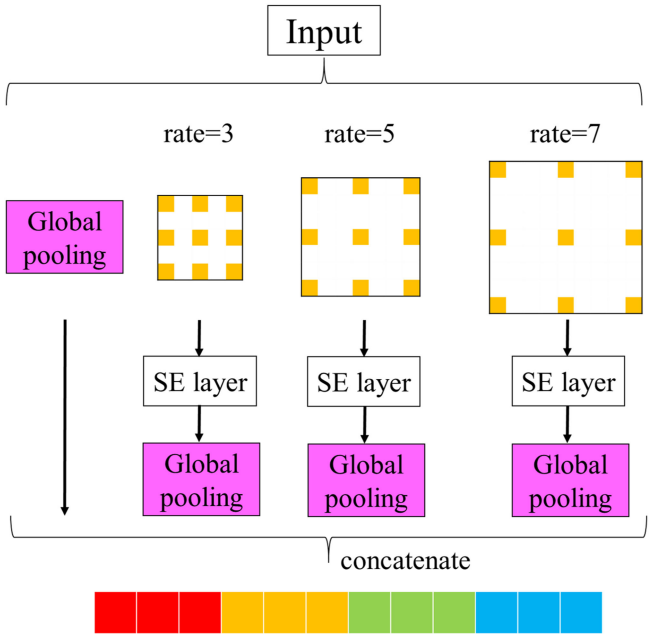


Fig. 8. Structure of the multidilation pooling module.

E. Multidilation Pooling Module

To achieve a multiscale feature representation, we add a multidilation pooling module. Detail of the multidilation pooling module structure is illustrated in Fig. 8. The multidilation pooling module is a pyramid pooling module. The idea of the pyramid pooling starts from SPPNet [72]. In SPPNet, the SPP layer is introduced to remove the fixed-size constraint of the network, so the FC layer can get the fixed-size input. The SPP layer can pool features extracted at variable scales thanks to the flexibility of input scales. Therefore, SPPNet shows great strength in object detection. In semantic segmentation, the pyramid pooling module is also used in DeepLab [65] and PSPNet [66], which can well handle scale variability in semantic segmentation.

As shown in Fig. 8, the multidilation pooling module in our network has multiple branches, which extract the features of different scales, respectively. One branch uses the global average pooling directly. The other branches comprise three layers: dilated convolution, SE layer, and global average pooling.

TABLE II
COMPARISON OF DIFFERENT DATASETS

Datasets	Images per class	Class	Total images	Spatial resolution (m)	Image sizes
UC Merced	100	21	2100	0.3	256×256
WHU-RS19	~50	19	1005	0.5	600×600
RSSCN7	400	7	2800	-	400×400
SIRI-WHU	200	12	2400	2	200×200
AID	200~400	30	10000	0.5-0.8	600×600
NWPU-RESISC45	700	45	31500	~30-0.2	256×256

TABLE III
OVERALL ACCURACY (%) COMPARISON WITH THE UC MERCED DATASET

Method	Training ratios	
	50%	80%
LPCNN [7]	-	89.90
GBRCN [6]	-	94.53
CaffeNet [1]	93.98±0.67	95.02±0.81
VGG-VD-16 [1]	94.14±0.69	95.21±1.20
GoogLeNet [1]	92.70±0.60	94.31±0.89
SICNN [57]	-	96.00
Pre-trained-AlexNet-SPP-SS [12]	-	96.67±0.94
Fusion by addition [13]	-	97.42±1.79
MDDC [14]	-	96.92±0.57
Integrating multilayer features of CNN [16]	-	98.81±0.38
Two-Stream Deep Fusion Framework [18]	96.97±0.75	98.02±1.03
TEX-Net-LF [23]	96.91±0.36	97.72±0.54
SRSCNN [22]	-	95.57
Fine-tune MobileNet V2	**97.88±0.31**	**98.13±0.33**
SE-MDPMNet	**98.57±0.11**	**98.95±0.12**

Finally, the features of multiple branches are concatenated to obtain global multiscale feature representation. In the experiment, we use four branches. We set multiple dilation rates of 3, 5, and 7, respectively, in three dilated convolutions, which can extract the features of different scales. In order to emphasize the importance of different channels adaptively, we also use the SE layer after the dilated convolution. The experiment in Section V shows using the SE layer in this module can have a better performance.

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Setup

In order to test the performance of our method, the following commonly used datasets are employed: UC Merced [30], WHU-RS19 [31], RSSCN7 [32], SIRI-WHU [33]–[35], the challenging large-scale AID [1] and NWPU-RESISC45 [2]. In Table II, the six publicly available datasets are compared about the number of images per class, the number of scene class, total images, the spatial resolution of images and image size.

To compute overall accuracy (OA), following Xia *et al*. [1], two different settings are adopted for UC-Merced, WHU-RS19, RSSCN7, and AID datasets. For the UC-Merced dataset, the training ratios are set at 50% and 80%; for the WHU-RS19 dataset, the ratios are fixed at 40% and 60%; for the RSSCN7 dataset and AID dataset, the ratios are fixed at 20% and 50%. In addition, for the SIRI-WHU datasets, the ratios are fixed at 50% and 80%. Following Cheng *et al*. [2], the ratios are fixed at 10% and 20% for NWPU-RESISC45 dataset.

In order to alleviate the overfitting problem, we use the data augmentation as follows: the training images first are resized to 256×256. Then, the images are randomly sampled with randomly horizontal flip, randomly vertical flip, and randomly rotation. After that, the images are randomly sampled with randomly changing the brightness, contrast, and saturation. Finally, the sample images are cropped to 224×224 randomly. The test images only are resized to 224×224. Both the training and test images are subtracted from the mean and divided by the standard deviation of the dataset.

We use stochastic gradient descent with a minibatch size of 32. The initial learning rate starts at 0.01 and is divided by 10 when epoch reaches 30 and 60. The models are trained for 100 epochs. We use a weight decay of 0.0001 and a momentum of 0.9. All the CNN models are implemented on a PC with a 2.10 GHz 8 core CPUs and 32-GB memory. In addition, a GTX Titan X GPU is also used for acceleration. In order to test the performance of our method, we also fine-tune the MobileNet V2 on each dataset as a baseline.

To compute the OA, following Xia *et al*. [1], we randomly split the datasets into training sets and testing sets for evaluation, and repeat the process ten times to reduce the influence of the randomness and obtain reliable results. The OA is computed for each run, and the results are reported as the mean and standard deviation of the OA from the individual runs.

To compute the confusion matrix, we use the best model in training process. For dataset ratio, we choose the ratio of training sets of the UC-Merced, WHU-RS19, RSSCN, SIRI-WHU, AID, and NWPU-RESISC45 to be the commonly used ones at 80%, 60%, 50%, 80%, 50%, and 20%, respectively.

B. Experiment 1: The UC Merced Dataset

This dataset [30] has 21 land use classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. For each of the classes, there are 100 images, each measuring 256×256 size. The images are manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 foot.

The experimental results of the existing methods and our method for the UC Merced dataset are listed in Table III. Because previous works [1], [2] have shown that the deep-learning-based methods have far surpassed handcrafted feature-based methods,

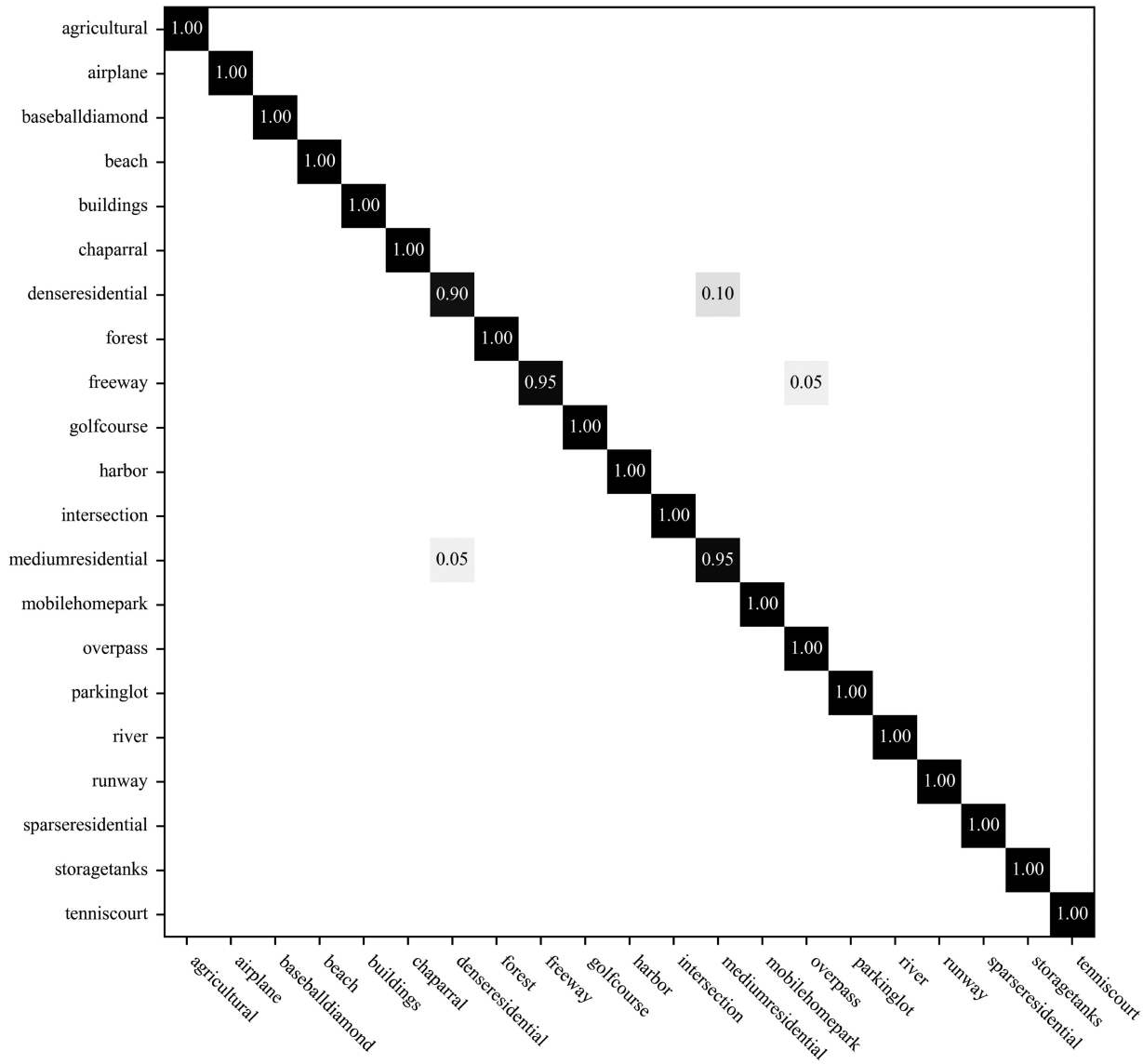


Fig. 9. Confusion matrix of our proposed network with the UC Merced dataset.

we do not compare with the traditional handcrafted feature-based methods. As can be seen in Table III, when the training ratio is 50%, the fine-tuned MobileNet V2 has already surpassed all the previous methods, which indicates that the residual structure in MobilNet V2 is crucial as it prevents non-linearity from destroying too much information. By introducing the SE-block, dilated convolution and multidilation pooling module, out network obtains higher accuracy. When the training ratio is 80%, our method narrowly surpasses the previous best method [16]. The previous best method is able to achieve 98.81%, because they integrate multilayer features of a pre-trained CNN model for scene classification. Moreover, those methods use fusion strategies to improve accuracy and their models are more complicated. However, our method does not use any fusion strategy and the accuracy of our model can achieve 98.95%. The authors [12] also utilize pyramid pooling module in AlexNet [46], but the performance of our network is 2.28% higher than theirs, which indicates that multidilation pooling module could extract better multiscale features than their pyramid pooling module. Compared with the fine-tuned MobileNet V2, our network is 0.69% and 0.82% higher for the ratio of 50% and 80%, respectively.

In Fig. 9 we display the confusion matrix of our result. As can be seen, most of the scene categories are fully recognized by our model with the exception three categories: dense residential, freeway, and medium residential. We believe that there is major confusion between dense residential and medium residential, because the images of dense residential and medium residential all have some similarities in the distribution of buildings.

C. Experiment 2: WHU-RS19

This dataset [31] was constructed by the Computational and Photogrammetric Vision team. All the scenes in the dataset were extracted from a set of satellite images exported from Google Earth with spatial resolution up to 0.5 m. The whole dataset contained 19 classes of high-resolution remote sensing scenes

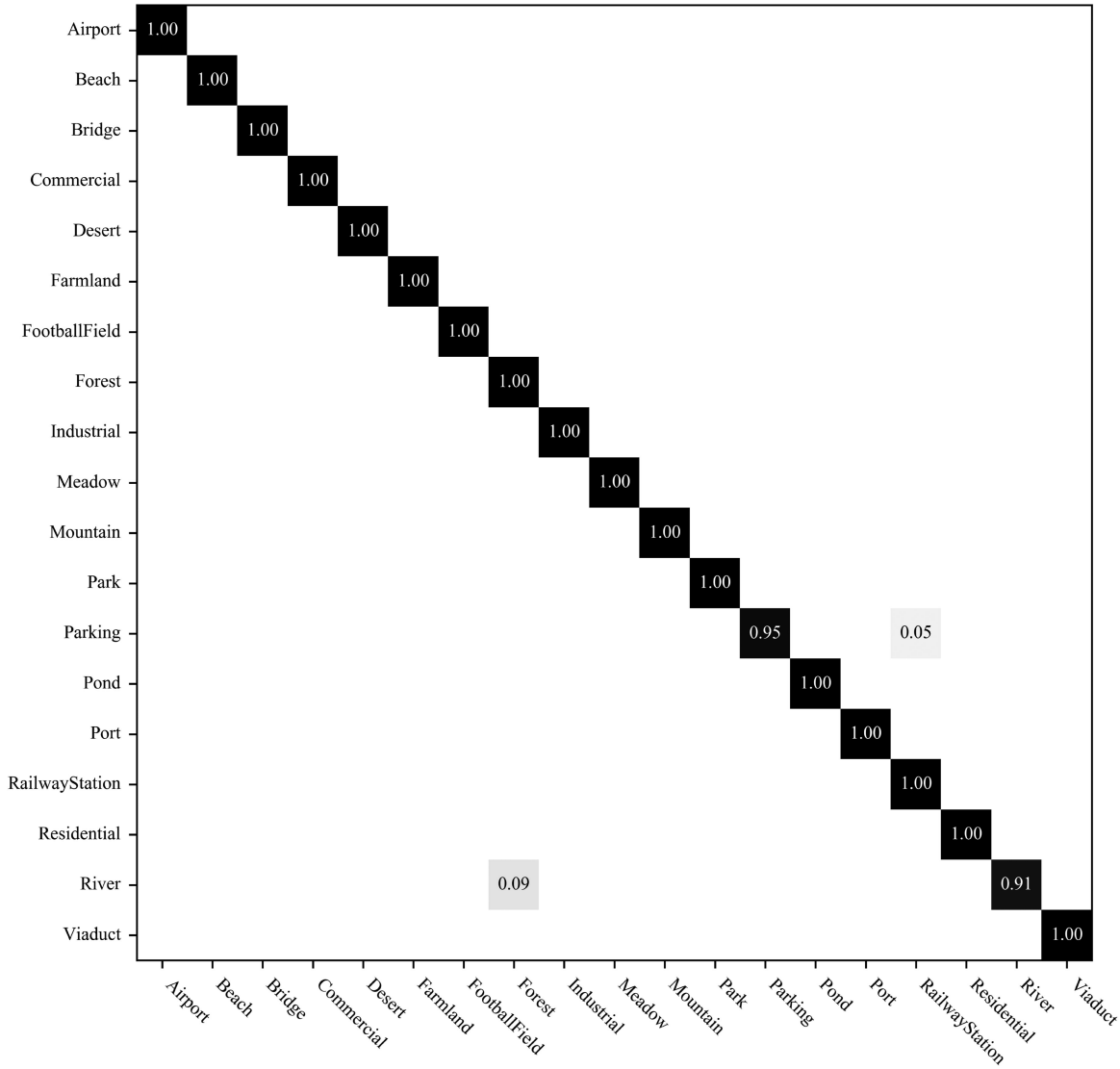


Fig. 10. Confusion matrix of our proposed network with the WHU-RS19 dataset.

including airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking lot, pond, port, railway station, residential area, river, and viaduct. For each scene category, there are about 50 images, with 1005 total images in the entire dataset. The image sizes are 600 × 600. This dataset is very challenging due to the changes in resolution, scale, orientation, and illumination of the images.

The experimental results of existing methods and our method for the WHU-RS19 dataset are listed in Table IV. As can be seen in Table IV, when the training ratio is 40%, the classification result of fine-tuned MobileNet V2 is better than CaffeNet [1], VGG-VD-16 [1], and GoogleNet [1], meaning that MobileNet V2 has a stronger feature representation power. Moreover, our network achieves 98.46% accuracy, which is worse than the previous best method TEX-Net-LF [23]. This may due to TEX-Net-LF [23] has good feature representation power on a small dataset. When the training ratio is 60%, our model can achieve up to 98.97%, outperforming all the previous methods. Compared with the feature fusion-based methods, such as the

TABLE IV
OVERALL ACCURACY (%) COMPARISON WITH THE WHU-RS19 DATASET

Method	Training ratios	
	40%	60%
CaffeNet [1]	95.11±1.20	96.24±0.56
VGG-VD-16 [1]	95.44±0.60	96.05±0.91
GoogLeNet [1]	93.12±0.82	94.71±1.33
DCA by addition [13]	-	98.70±0.22
MDDC [14]	-	98.27±0.53
Two-Stream Deep Fusion Framework [18]	98.23±0.56	98.92±0.52
TEX-Net-LF [23]	98.48±0.37	98.88±0.49
Fine-tune MobileNet V2	**96.82±0.35**	**98.14±0.33**
SE-MDPMNet	**98.46±0.21**	**98.97±0.24**

TABLE V
OVERALL ACCURACY (%) COMPARISON WITH THE RSSCN7 DATASET

Method	Training ratios	
	20%	50%
CaffeNet [1]	85.57±0.95	88.25±0.62
VGG-VD-16 [1]	83.98±0.87	87.18±0.94
GoogLeNet [1]	82.55±1.11	85.84±0.92
Two-Stage Deep Feature Fusion [21]	-	92.37±0.72
TEX-Net-LF [23]	92.45±0.45	94.0±0.57
Fine-tune MobileNet V2	**89.04±0.17**	**92.46±0.66**
SE-MDPMNet	**92.65±0.13**	**94.71±0.15**

fusion of saliency detection features [18], our classification result is better, which may indicate that our network can extract more discriminative feature than the feature fusion-based methods. Moreover, compared with the fine-tuned MobileNet V2, our network is 1.64% and 0.83% higher. We display the confusion matrix in Fig. 10. As can be seen, most of the scene categories are fully recognized except the following two classes: forest and river. These samples are misclassified perhaps due to the smaller interclass dissimilarity. For example, the scenes with a river are classified as a forest scene, which may be explained by the fact that there are more trees on both sides of the river.

D. Experiment 3: RSSCN7

This dataset [32] was also collected from Google Earth and included 2800 remote sensing images, which were from seven typical scene categories—grass land, forest, farm land, parking lot, residential region, industrial region, and river and lake. For each category, there are 400 images, which are sampled on four different scales, and each image is 400×400 in size. This dataset is rather challenging due to the wide diversity of the scene images, which are captured under changing seasons and varying weathers condition and are sampled with different scales.

The experimental results of our method applied to the RSSCN7 dataset are listed in Table V. As can be seen, the accuracy of our network can achieve 92.65% and 94.71%, respectively, when the training ratios are 20% and 50%. When the training ratio is 20%, our model is only 0.2% higher than TEX-Net-LF [23]. Nevertheless, when the training ratio is 50%, the accuracy of our model can gain 0.71%. In addition, our network is more lightweight compared to TEX-Net-LF [21]. Compared with the fine-tuned MobileNet V2, our network is 3.61% and 2.25% higher. We display the confusion matrix in Fig. 11. As can be seen, it achieves the best accuracy for the forest scene, which means that this class has higher interclass dissimilarity. On the other hand, the results of grass and industry experience poor results. We argue that the field and grass scenes are similar. They clearly have smaller intraclass variation, which results in more misclassification samples in the grass class.

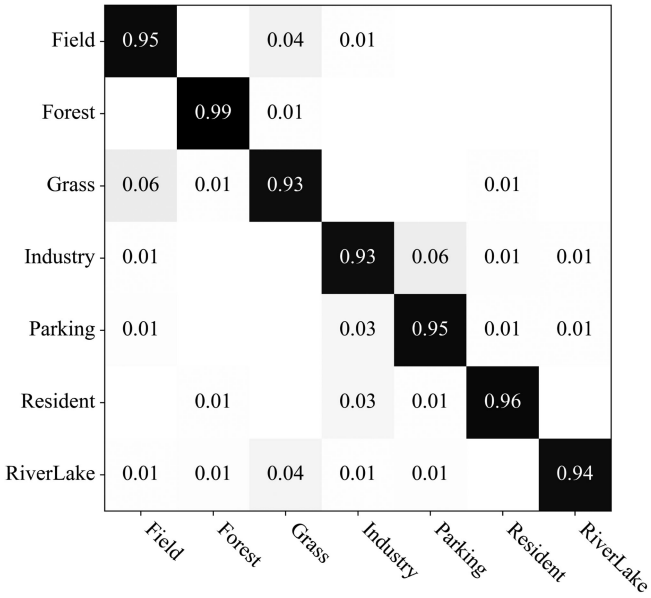


Fig. 11. Confusion matrix of our proposed network with the RSSCN7 dataset.

TABLE VI
OVERALL ACCURACY (%) COMPARISON WITH THE SIRI-WHU DATASET

Method	Training ratios	
	50%	80%
DMTM [35]	91.52	-
LPCNN [7]	-	89.88
SICNN [57]	-	93.00
Pre-trained-AlexNet-SPP-SS [12]	-	95.07±1.09
SRSCNN [22]	93.44	94.76
Fine-tune MobileNet V2	**95.77±0.16**	**96.21±0.31**
SE-MDPMNet	**96.96±0.19**	**98.77±0.19**

E. Experiment 4: SIRI-WHU

This dataset [33]–[35] also was acquired from Google Earth and mainly covers urban areas in China. The scene dataset was designed by RS_IDEA Group in Wuhan University (SIRI-WHU) and served as a 12-class Google image dataset for research purposes. There are 200 images for each of the following classes: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water. Each image is 200×200 with a 2-m spatial resolution.

The experimental results of our method are shown in Table VI. The classification results of our method are better than all the existing methods, which indicates that our CNN can extract better global features than the other models. Our network reaches 96.96% and 98.77% at training set ratios of 50% and 80%, respectively, which confirms that the proposed model is an effective approach for remote sensing scene classification. Moreover, the experimental result of our network is better than fine-tuned MobileNet V2, 1.19% and 2.56%, respectively. We display the confusion matrix of our network in Fig. 12. Three-fourths of the scene categories are correctly classified, and other

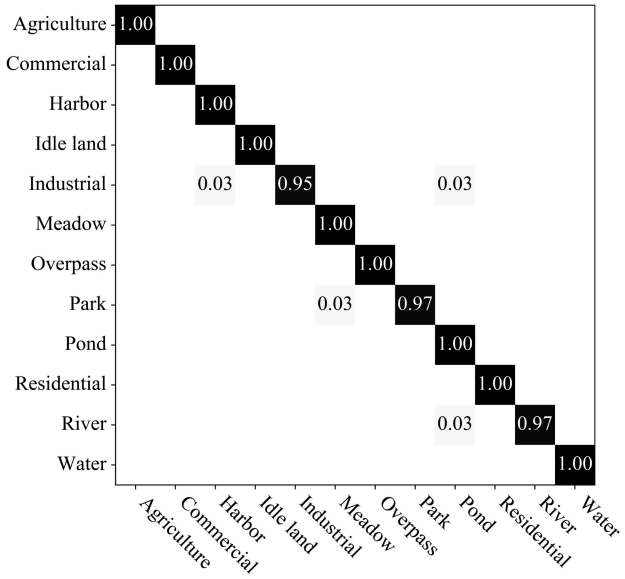


Fig. 12. Confusion matrix of our proposed network with the SIRI-WHU dataset.

misclassified categories industrial, park, and river reach at least 0.95. Among the misclassified scene categories, the industrial scene is poorly classified. Some river scene images are misclassified as water, which is likely because the images of river scene contain more water. So, these images may be misclassified as water scene by the network.

E. Experiment 5: AID

This dataset [1] was collected from Google Earth imagery including the following 30 aerial scene types: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. All the images were labeled by specialists in the field of remote sensing image interpretation. The number of sample images varies a lot with different aerial scene types from 220 up to 420. In all, the AID dataset contains 10 000 images in the 30 classes. The images in AID are actually multisource, which brings more challenges for scene classification than the single source images. In contrast with the existing remote sensing image datasets, e.g., UC-Merced dataset [30] and WHU-RS19 dataset [31], AID dataset has the following properties: higher intraclass variations, smaller interclass dissimilarity, and relatively large scale.

The experimental results of our method for the AID dataset are listed in Table VII. As can be seen, when the training ratio is 20%, our model achieves the best performance on the AID dataset. Our model achieved 94.68%, which is about 8.09% higher than VGG-VD-16 [1] and it also exceeded by 2.36% and 0.87% for the two-stream deep fusion framework [18] and TEX-Net-LF [23], respectively. Moreover, the proposed network is 0.55% higher than fine-tuned MobileNet V2. When the ratio is fixed at 50%, our model also achieves the highest classification accuracy. In addition, our model is even 1.18% higher than the fine-tuned MobileNet V2, which verifies that our model has stronger feature expression ability and can learn more discriminative features without adding any handcraft features or fusion strategies.

TABLE VII
OVERALL ACCURACY (%) COMPARISON WITH THE AID DATASET

Method	Training ratios	
	20%	50%
CaffeNet [1]	86.86±0.47	89.53±0.31
VGG-VD-16 [1]	86.59±0.29	89.64±0.36
GoogLeNet [1]	83.44±0.40	86.39±0.55
Fusion by addition [13]	-	91.87±0.36
Two-Stream Deep Fusion Framework [18]	92.32±0.41	94.58±0.25
Two-Stage Deep Feature Fusion [21]	-	94.65±0.33
Multilevel Fusion(fine-tune SVM) [20]	-	95.36±0.22
TEX-Net-LF [23]	93.81±0.12	95.73±0.16
Fine-tune MobileNet V2	**94.13±0.28**	**95.96±0.27**
SE-MDPMNet	**94.68±0.17**	**97.14±0.15**

We display the confusion matrix in Fig. 13. Its classification accuracy for most of the scene categories reaches 90%. Even the results of baseball fields, forest, mountains, port, and viaducts reach 100%. However, the worse scene category square has accuracy with 89%, which greatly affected the final overall classification accuracy. We think that the most notable confusion is resort and park, because they contain similar structures, e.g., buildings, plants, and ponds.

F. Experiment 6: NWPU-RESISC45

The NWPU-RESISC45 dataset [2] is a publicly available benchmark for RESISC, which was created by the NWPU. This dataset contains 31 500 images, covering 45 scene classes with 700 images in each class. The size of each image is 256 × 256. These 45 scene classes are: airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snow berg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. The spatial resolution varies from about 30 to 0.2 m per pixel for most of the scene classes except for island, lake, mountain, and snow-berg, which have lower spatial resolutions. The NWPU-RESISC45 dataset has the following three notable characteristics: large scale, rich image variations, high within-class diversity, and between-class similarity.

The experimental results for existing methods and our method for the NWPU-RESISC45 dataset are listed in Table VIII. Our model achieves 91.8% and 94.11% when the training set ratios are 10% and 20%, which is 4.65% and 3.75% higher than

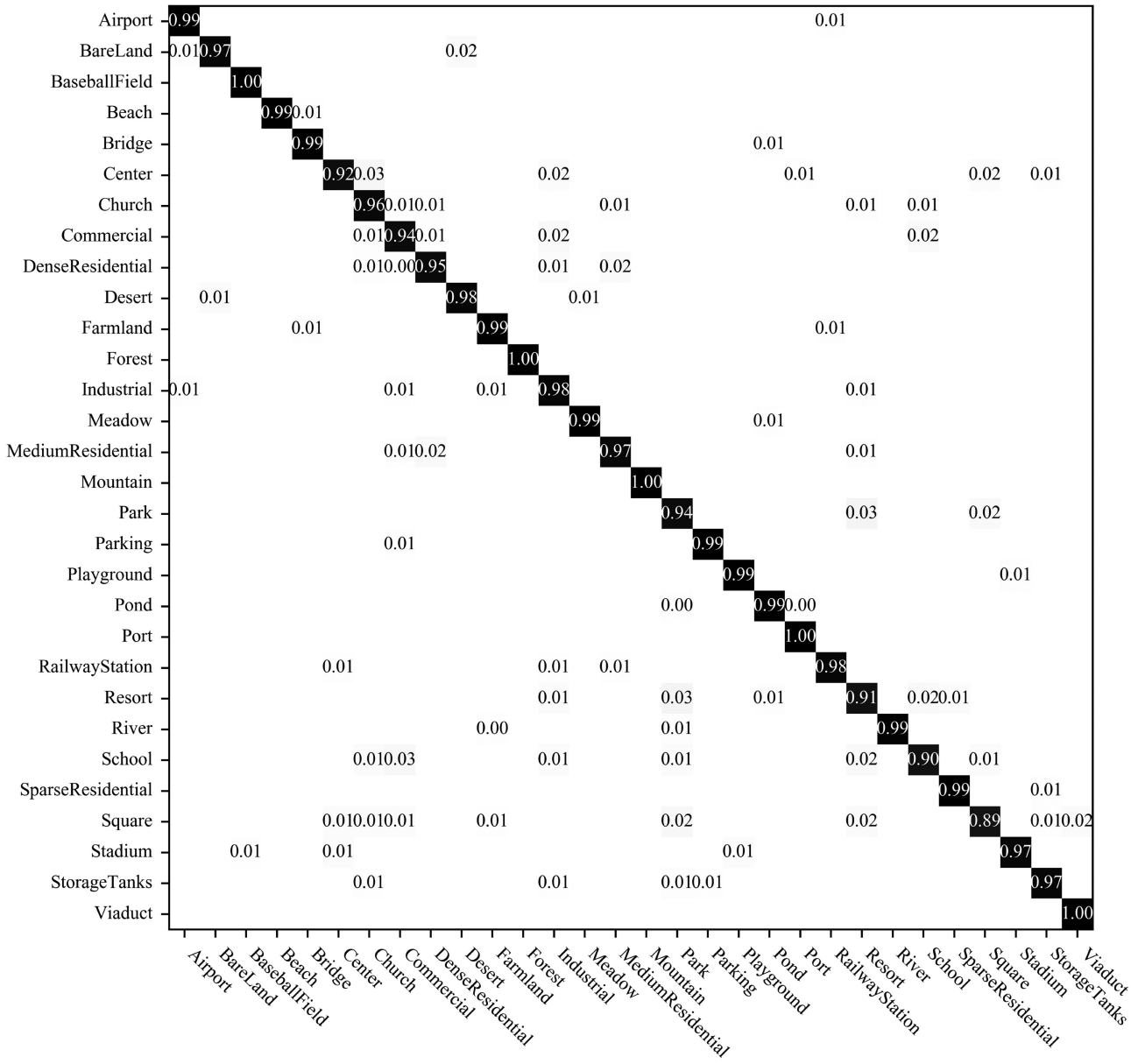


Fig. 13. Confusion matrix of our proposed network with the AID dataset.

TABLE VIII
OVERALL ACCURACY (%) COMPARISON WITH THE NWPU-RESISC45 DATASET

Method	Training ratios	
	10%	20%
Fine-tune AlexNet [2]	81.22±0.19	85.16±0.18
Fine-tune VGGNet-16 [2]	87.15±0.45	90.36±0.18
Fine-tune GoogLeNet [2]	82.57±0.12	86.02±0.18
BoCF [74]	82.65±0.31	84.32±0.17
Two-Stream Deep Fusion Framework [18]	80.22±0.22	83.16±0.18
Fine-tune MobileNet V2	**90.16±0.12**	**93.00±0.18**
SE-MDPMNet	**91.80±0.07**	**94.11±0.03**

the highest accuracy of previous work, 87.15% and 90.36%, respectively. For the 10% training ratio, our model achieves 1.64% higher than the fine-tuned MobileNet V2. For the 20% training ratio, our model is 1.11% higher than the fine-tuned MobileNet V2. The two-stream deep fusion framework [18] method performs worse on this dataset than other datasets, which achieves 80.22% and 83.16%. However, our model performs well on this dataset. This shows that for rich image variations, high within-class diversity, and between-class similarity, our model can extract more discriminative features.

We display the confusion matrix in Fig. 14. As can be seen, similar to the AID dataset results, the classification accuracy of most of the scene categories reaches 90%. However, because this dataset is challenging and contains 45 scene categories, none of the categories are correctly classified completely. The worst

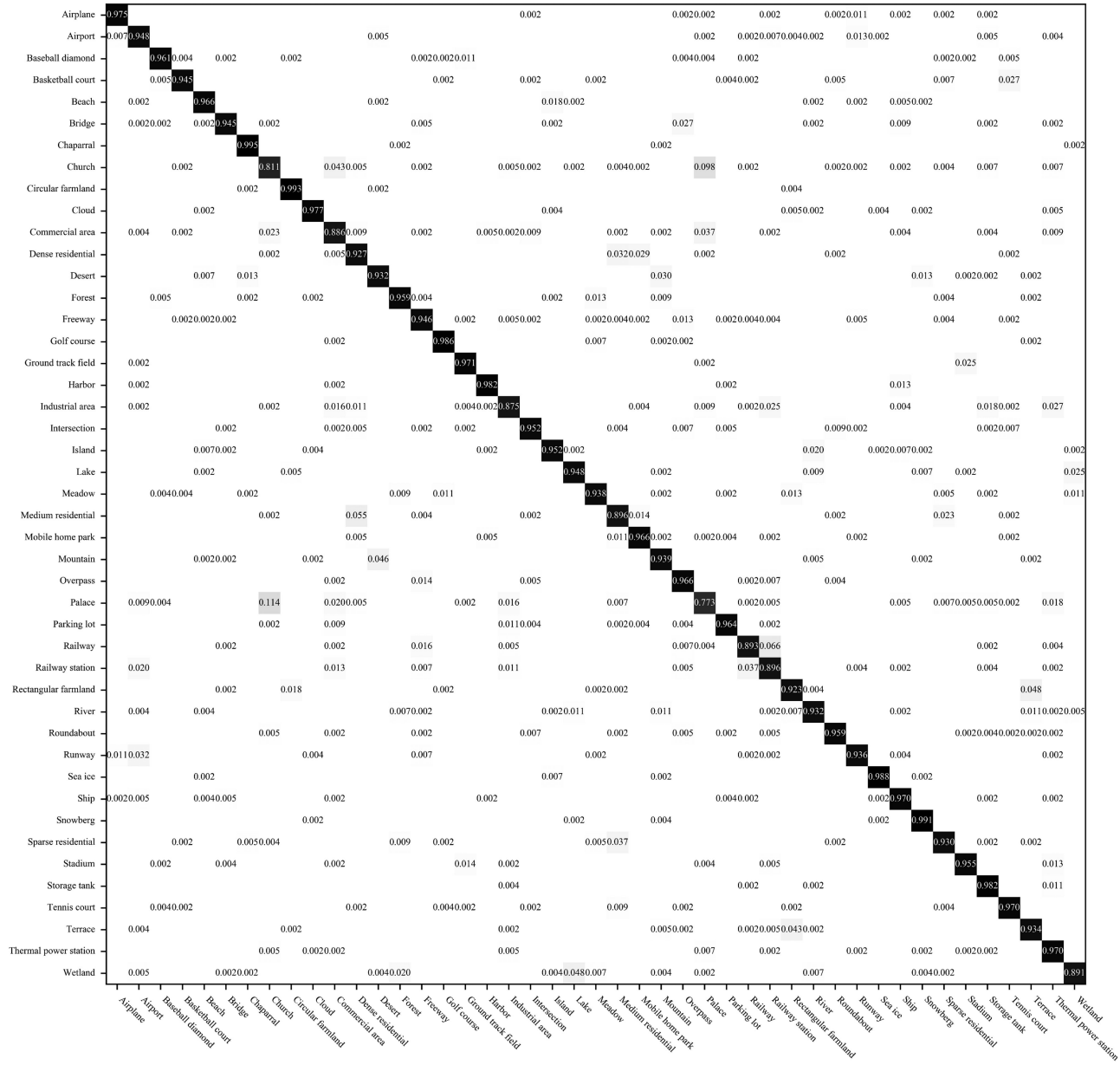


Fig. 14. Confusion matrix of our proposed network with the NWPU-RESISC45 dataset.

results are church and palace, which reach 81.1% and 77.3%, respectively. This is understandably confusing due to the very similar architecture and shape of churches and palaces.

V. Discussion

In this section, five factors, data augmentation, the effect of modified MobileNet v2, dilated convolution, channel attention and pyramid pooling module, are tested to analyze how these factors affect classification accuracy. In addition, we also use the class activation maps (CAM) [74] and t-distributed stochastic neighboring embedding (t-SNE) algorithm [75] to visualize internal mechanism. In all experiments, the AID [1] dataset of 50% training ratio is chosen for the analysis of the above factors. The experimental setup is the same with previous experiments. For comparison, we use the average accuracy of the last ten epochs as the evaluation indicator.

A. Evaluation of Data Augmentation

In deep learning, a large dataset is crucial to train an effective model. However, in the remote sensing scene classification community, there are small data volumes of available scene classification datasets. Moreover, small data volumes are easier to overfitting. Therefore, to alleviate this problem, we use the data augmentation in the training process as follows: randomly horizontally and vertically flip, randomly rotation, randomly scale from 0.8 to 1.2, randomly crop and randomly changing the brightness, contrast, and saturation. As we can see in Table IX second row, after using the data augmentation, the

TABLE IX
ACCURACY (%) COMPARISON WITH DIFFERENT FACTORS

Augmentation	Dilated convolution	SE	Remove 1×1 conv	Acc
				95.89
√				96.30
√	√ (same)			96.50
√	√ (hybrid)			96.71
√		√		96.73
√	√ (hybrid)	√		97.00
√	√ (hybrid)	√	√	96.96

accuracy of the fine-tuned MobileNet V2 is 96.3%, which shows that the data augmentation is very effective.

B. Evaluation of Dilated Convolution and Channel Attention

To evaluate the effect of dilated convolution, we use dilated convolutions in the original MobileNet V2 [27]. Specifically, we experiment with several variants of the dilated convolution (see Table IX).

1) Standard convolution: for all convolutions, we set their dilation rates to 1 (see Table IX second row).
2) Same dilation rate: for seven groups bottleneck blocks (see Table I), we begin using the dilated convolution form the fourth bottleneck block. The stride of the fourth and the fifth bottleneck blocks are set to 1 and the dilation rates are set to 2. Similarly, in the sixth and seventh bottleneck blocks, the dilation rates are set to 4 and the stride in the sixth bottleneck block is set to 1 (see Table IX third row).
3) Hybrid dilation rate: we begin using the dilated convolution from the fourth bottleneck block. The stride of the fourth bottleneck block is set to 1 and the dilation rates of four depthwise convolutions in the fourth bottleneck block are set to 1, 2, 3, and 4, respectively. Similarly, in the fifth and sixth bottleneck blocks, the dilation rates are set to 2, 3, and 4 and the stride in the sixth bottleneck block is set to 1 (see Table IX fourth row).

The experimental results show that using the dilated convolution in the network is beneficial to classify scene images. However, for the same dilation rate strategy, due to the grid effect, the classification accuracy is greatly limited, only increasing 0.2%. However, using the hybrid dilation rate strategy can alleviate this problem to a great extent and the final result can promote 0.41%.

SE-block has an attention mechanism, which can make the discriminative feature maps have larger weights. To evaluate the effect of channel attention, we use SE-block in every bottleneck layer. As we can see in the fifth row of Table IX, adding the SE-block can improve accuracy the accuracy to 96.73%. Therefore, we combine the dilated convolution and channel attention together. In Table IX sixth row, the performance of network can reach 97.0%.

TABLE X
ACCURACY (%) COMPARISON WITH SEVERAL VARIANTS OF MULTIDILATION POOLING MODULE

Module	Acc	Parameters	FLOPs
3×3, 5×5, 7×7 convolutions	96.58	6.80M	5.33G
multi-dilation pooling module (r=2, 4, 6)	96.61	2.21M	1.73G
multi-dilation pooling module (r=3, 5, 7)	96.85		
multi-dilation pooling module (r=4, 8, 12)	96.52		
multi-dilation pooling module (r=6, 12, 18)	96.52		
multi-dilation pooling module (r=2, 4, 6, 8)	96.88	2.95M	2.31G
multi-dilation pooling module (r=3, 5, 7, 9)	96.9		
multi-dilation pooling module (r=4, 8, 12, 16)	96.75		
multi-dilation pooling module (r=6, 12, 18, 24)	96.72		
multi-dilation pooling module (our proposed)	97.14	2.24M	1.73G

C. Evaluation of the Effect of Modified MobileNet V2

In the original MobileNet V2 [27], the number of channels of the last feature maps is 1280 and the kernel size of the last convolution is 1×1. We think that the 1×1 convolution just nonlinearly maps features to 1280 dimensions. Therefore, in our network, we take out the last 1×1 convolution. Removing the last 1×1 convolution can reduce the number of parameters by 387 840 ($320 \times 1280 + 1280 = 387\,840$). In Table IX, the accuracy after removing the last 1×1 convolution can reach 96.96%, which only drops 0.04%. This is acceptable and the result proves that this modification does not hinder the network performance a lot.

D. Evaluation of Multidilation Pooling Module

To evaluate the effect of multidilation pooling module, we use the multidilation pooling module on the basis of dilated convolution and channel attention. We try several variants of this module: different number of branches and different dilation rates. And the 3×3, 5×5, and 7×7 convolutions are baseline.

As can be seen in Table X, using the 3×3, 5×5, and 7×7 convolutions to replace the dilated convolution in the multidilation pooling module would hinder the network performance. At the same time, using large convolutional kennels will increase the number of parameters (about three times parameters of three 3×3 convolutions), which is not conducive to optimize the network. Comparing with using 3×3, 5×5, and 7×7 convolutions, the multidilation pooling module (without SE-layer) can extract multiscale features with fewer parameters. To get more

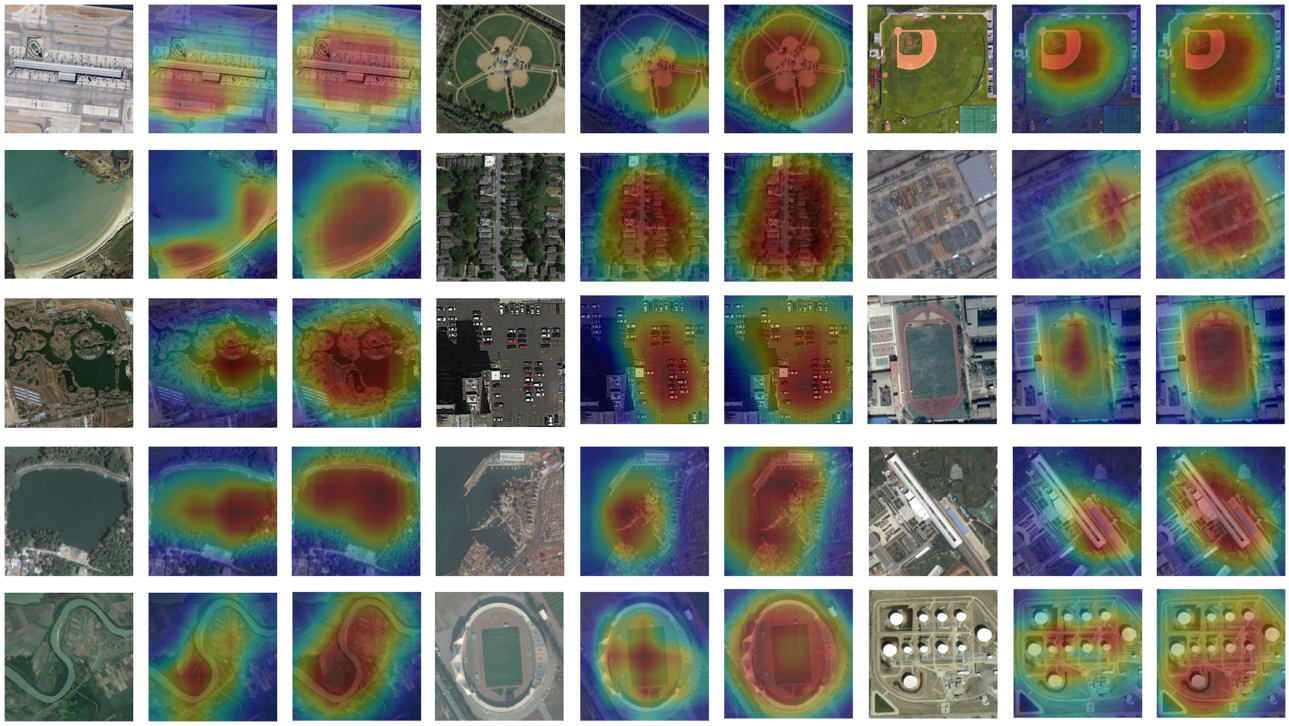


Fig. 15. Examples of the CAMs generated from the AID dataset. We show the original images, CAMs generated from baseline model and our proposed model, respectively.

meaningful and context-sensitive analysis results of the multidilation pooling module, we attempt four different settings about dilation rate and two kinds of number of branches. For a 224 × 224 image, the best dilation rate is 3, 5, and 7 for three dilation convolution branches and 3, 5, 7, and 9 for four dilation convolution branches, respectively. For the large dilation rate, the result gets worse. When rates become large, the sample locations become very sparse, which may not be good for learning, because the local information is completely missing and the information can be irrelevant across large distances. The multidilation pooling module with rate 3, 5, and 7 can reach 96.85%, which has a good tradeoff for accuracy and FLOPs. To move forward a single step, we use the channel attention mechanism in the multidilation pooling module. The experiment shows that it is beneficial to training and can make this module to obtain more powerful feature representation. On the basis of dilated convolution and channel attention, our network with the multidilation pooling module can get 97.14% finally.

CAMs [74] can highlight the discriminative object parts detected by the CNN. For a better understanding of our model, we use CAM to visualize whether the network can recognize correct parts of the image corresponding to the true class. We show the CAMs generated from the baseline model (MobileNet V2) and our model in Fig. 15. The original images are from the AID dataset. As can be seen, the baseline model and our model both can highlight the semantic object corresponding to the true class, which indicates that CNN has the capability of object localization and recognition. In addition, it is interesting that the CAMs generated from our model can better cover the semantic objects and have a wider range of highlights. We believe that this is due to the widespread using of dilation convolutions in our network, which enables the network to utilize context information.

Besides, we also use t-SNE algorithm [75] to visualize global feature representations learned by the baseline model (MobileNet V2) and our model in Fig. 16. It is noted that we only use the test data, because these two networks have trained on training data. We use the features after global pooling layer as input. It is clear that some classes are well separated by the baseline model and our model. But the center part of the baseline model result seems to have a little confusion for similar categories. However, our model's result increases the separability and relative distance between the individual semantic clusters, which means that our model has better global feature representations and can prepare better features for the fully connected layer.

E. Evaluation of Size of Model

We also compare the size of model with different methods. The CaffeNet [1], VGG-VD-16 [1], GoogLeNet [1], MobileNet V2 [27], and our model are compared about the number of parameters and FLOPs. The number of parameters stands for the size of model and the number of FLOPs stands for the computation complexity. The results are listed in Table XI. It is clear that our model is superior to CaffeNet and VGG-VD-16 about the model size and computation complexity. And when compared with lightweight models, GoogLeNet and MobileNet V2, our model has a better tradeoff about the accuracy and model size. However, due to introducing the dilated convolution, the size of subsequent feature map does not reduce. We can see that the computation complexity of our model increases about ten times.

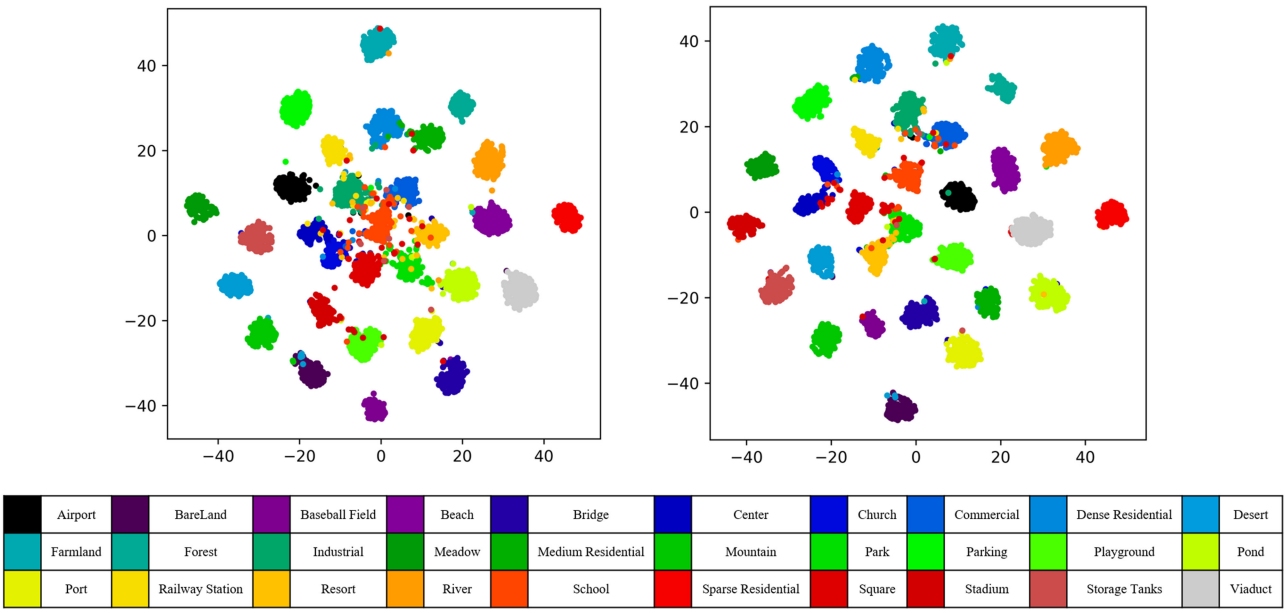


Fig. 16. Two-dimensional feature visualization of image global feature representations learned from the AID dataset using the t-SNE algorithm. Baseline model (left), our proposed model (right).

TABLE XI
ACCURACY (%) COMPARISON WITH DIFFERENT METHODS

Methods	Acc	Parameters	FLOPs
CaffeNet [1]	89.53	60.97M	715M
VGG-VD-16 [1]	89.64	138.36M	15.5G
GoogLeNet [1]	86.39	7M	1.5G
MobileNet V2[27]	95.96	3.5M	334M
SE-MDPMNet	97.14	5.17M	3.27G

But compared with the extra high-computation VGG, our model is still lightweight.

VI. CONCLUSION

In this paper, a lightweight end-to-end deep network is proposed, which combines the advantages of the dilated convolution, channel attention, and multidilation pooling for HSR remote sensing imagery scene classification. We introduce the dilated convolution, and channel attention to MobileNet V2 to extract more robust and discriminative features. To improve the performance of the CNN further, the multiscale features are also considered by adding the multidilation pooling module. Experiments are performed on six datasets, and the results verify that our method is robust and can achieve higher accuracy compared with the current state-of-the-art methods. We believe that our network provides a new baseline for remote sensing scene classification.

The limitations of our research include the following. We only focus on channel attention, and the future research should consider adding spatial attention. For example, the saliency detection can be incorporated into the CNN model. The spatial attention can assign more weight to the key part. In addition, various feature fusion strategies also can promote the accuracy of scene classification.

REFERENCES

[1] 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.
[2] G. Cheng, J. W. Han, and X. Q. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[3] F. Hu *et al.*, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.
[4] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.
[5] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
[6] F. Zhang, B. Du, and L. P. 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.
[7] Y. F. Zhong, F. Fe, and L. P. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, Apr. 25, 2016, Art. no. 025006.
[8] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.
[9] H. Li *et al.*, "Scene classification in remote sensing images using a two-stage neural network ensemble model," *Remote Sens. Lett.*, vol. 8, no. 6, pp. 557–566, Feb. 2017.
[10] J. Wang *et al.*, "Transferring pre-trained deep CNNs for remote scene classification with general features learned from linear PCA network," *Remote Sens.*, vol. 9, no. 3, Mar. 2017, Art. no. 225.
[11] E. Z. Li, P. J. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 1068–1081, Mar. 2017.

[12] X. B. Han *et al.*, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, Aug. 2017, Art. no. 848.

[13] S. Chaib, H. Liu, Y. F. 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.

[14] K. L. Qi *et al.*, "A multiscale deeply described correlatons-based model for land-use scene classification," *Remote Sens.*, vol. 9, no. 9, Sep. 2017, Art. no. 917.

[15] G. L. Wang, B. Fan, S. M. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4104–4115, Sep. 2017.

[16] E. Z. Li, J. S. Xia, P. J. 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.

[17] Q. S. Liu, R. L. Hang, H. H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

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

[19] N. S. Yang, H. Tang, H. Q. Sun, and X. Yang, "DropBand: A simple and effective method for promoting the scene classification accuracy of convolutional neural networks for VHR remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 257–261, Feb. 2018.

[20] Y. L. Yu, and F. X. 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.

[21] Y. S. Liu, Y. B. Liu, and L. W. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.

[22] Y. Liu *et al.*, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 444.

[23] R. M. Anwer *et al.*, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–44, May 2015.

[25] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[26] X. Yu *et al.*, "Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework," *Gisci. Remote Sens.*, vol. 54, no. 5, pp. 741–758, May 2017.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[28] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.

[31] G. S. Xia *et al.*, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC VII Symp. 100 Years ISPRS*, vol. 38, 2010, pp. 298–303.

[32] Q. Zou, L. H. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[33] Q. Q. Zhu, Y. F. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "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.

[34] B. Zhao *et al.*, "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, Feb. 2016, Art. no. 157.

[35] B. Zhao, Y. F. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-Derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[36] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on Gabor texture feature classification," in *Proc. 7th Int. Conf. Signal Process.*, vol. 1, 2004, pp. 733–736.

[37] C. Song, F. Yang, and P. Li, "Rotation invariant texture measured by local binary pattern for remote sensing image classification," in *Proc. 2nd Int. Workshop Educ. Technol. Comput. Sci.*, vol. 3, 2010, pp. 3–6.

[38] 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.

[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2005, pp. 886–893.

[41] M. Shahriari and R. Bergevin, "Land-use scene classification: A comparative study on bag of visual word framework," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 23059–23075, Nov. 2017.

[42] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 2169–2178.

[43] G. F. Sheng *et al.*, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[44] X. Han *et al.*, "Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery," *Int. J. Remote Sens.*, vol. 38, no. 2, pp. 514–536, 2016.

[45] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[46] 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.

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

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

[49] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[50] D. Marmanis *et al.*, "Classification with an edge: Improving semantic with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[51] N. Audebert, B. Le Saux, and S. Lefevre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[52] D. Marcos *et al.*, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[53] Z. P. Deng *et al.*, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018.

[54] Q. S. Liu *et al.*, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, Dec. 2017, Art. no. 1330.

[55] F. Zhou *et al.*, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, 2018.

[56] Y. F. Liu, Y. F. Zhong, J. Zhao, A. Ma, and Q. Qin, "Scene semantic classification based on scale invariance convolutional neural networks," in *Proc. 2017 IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 4754–4757.

[57] Y. Zhong *et al.*, "SatCNN: Satellite image dataset classification using agile convolutional neural networks," *Remote Sens. Lett.*, vol. 8, no. 2, pp. 136–145, 2016.

[58] Y. S. 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.

[59] Z. Q. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 371–390, Jan. 2018.

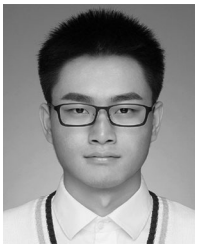
[60] S. Basu *et al.*, "DeepSat—A learning framework for satellite imagery," in *Proc. 23rd ACM Sigspatial Int. Conf. Adv. Geograph. Inf. Syst.*, 2015.

[61] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[62] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6848–6856.

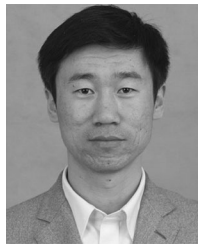
[63] L.-C. Chen *et al.*, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[64] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[65] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[66] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
[67] P. Q. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.
[68] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1743–1751.
[69] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636–644.
[70] L.-C. Chen *et al.*, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[71] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 448–456.
[72] K. He *et al.*, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.
[73] G. Cheng, Z. P. Li, X. W. 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.
[74] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[75] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



Bin Zhang received the B.S. degree in remote sensing science and technology from Liaoning Technical University, Fuxin, China, in 2017, and is currently working toward the M.S. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

His research interests include high spatial resolution remote sensing image processing, computer vision, and pattern recognition.



Yongjun Zhang received the B.S. and M.S. degrees in geodesy from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1997 and 2000, respectively, and the Ph.D. degree in geomatics from Wuhan University, Wuhan, China, in 2002.

He is currently a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include space, aerial, and low-attitude photogrammetry, image matching, combined bundle adjustment with multisource datasets, and 3-D city reconstruction.

Dr. Zhang received the First Prize for the second class National Science and Technology Progress Award (2017). He has been supported by the Yangtze River Scholar Program from the Ministry of Education of China (2017), the China National Science Fund for Excellent Young Scholars (2013), and the New Century Excellent Talents in University from the Ministry of Education of China (2007).

Shugen Wang received the B.S. degree in aerial photogrammetry from the Wuhan College of Surveying and Mapping, Wuhan, China, in 1984, the M.S. degree in photogrammetry and remote sensing from the Wuhan University of Surveying and Mapping Science and Technology, Wuhan, in 1994, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2003.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His major research interests include digital photogrammetry, high spatial resolution remote sensing image processing, and computer vision.