IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 10, 2020, accepted April 22, 2020, date of publication April 30, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991439

Dilated Convolution and Feature Fusion SSD Network for Small Object Detection in Remote Sensing Images

JUNSUO QU[1,2], CHANG SU[3], ZHIWEI ZHANG[3], AND ABOLFAZL RAZI[4]

[1]School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]Xi'an Key Laboratory of Advanced Control and Intelligent Process, Xi'an 710121, China
[3]School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[4]School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA

Corresponding author: Junsuo Qu (qujunsuo@xupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51875457, in part by the International Cooperation and Exchange Program of Shaanxi Province under Grant 2018KW-026, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM6120 and Grant 2019JM-606, and in part by the Xi'an Science and Technology Projects under Grant 201805040YD18CG24(6).

ABSTRACT Noting the shortcomings of current methods in detecting small objects in image-based remote sensing applications, in this paper, we propose a novel implementation of single shot multibox detector (SSD) networks based on dilated convolution and feature fusion. We call this algorithm dilated convolution and feature fusion single shot multibox detector (DFSSD). This algorithm removes the random clipping steps of data preprocessing layers in conventional SSD networks and utilizes the structure of feature pyramid network (FPN) network to fuse the low-level feature map with high resolution and the high-level feature map with rich semantic information. It also enhances the receptive field of the third-level feature map of the DFSSD network by using dilated convolution. In the data processing step of the model, we use the image segmentation of the feature point region proposals to improve the training sample size. The mean average precision (mAP) value of the proposed DFSSD network, when tested on remote sensing datasets, achieves 76.51%, which is significantly higher than that of the SSD model (69.81%).

INDEX TERMS Small object detection, feature fusion, dilated convolution, DFSSD network.

I. INTRODUCTION

Object detection has always been a research hotspot in the field of computer vision [1]. Detecting objects from general classes is the supporting technology for a large number of applications including intelligent monitoring [2], intelligent robotics [3], and many other applications. For instance, several methods are developed for face detection [4] and pedestrian detection [5] for surveillance systems and self-driving cars that are mature and achieve reasonable performances. However, the detection accuracy for a general class of objects with heterogeneous shapes, sizes, patterns, colors, and morphology is far from satisfactory. The source of difficulty is developing a unified method to capture object-specific features of an object with diverse size, shape, color and etc. It is difficult to find common features, especially for traditional machine learning methods that rely on manually designed feature extraction methods. Recently, more and more researchers have turned their attention to deep learning (DL) methods [6]. There exist many excellent object detection methods based on deep learning architectures and platforms such as AlexNet [7], ZFnet [8], VGGNet [9], GoogleNet [10], R-CNN [11], Faster R-CNN [12], SSD [13], and etc...Among them, the single shot multibox detector (SSD)model is a network architecture based on convolutional neural networks (CNN) with relatively high accuracy and near real-time performance. In this paper, the size of remote sensing image data set is from 800×800 to 4000×4000, the pixel value of small object is between 10×10 and 50×50, and the pixel value of medium object is between 50×50 and 300×300. However, the feature map of SSD model used for prediction is not reused, lack of sufficient semantic information, and the detection effect of overlapping object and small object is poor.

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

Many scholars have carried out research on improving the small object detection capability of the SSD model. Li *et al.* [14] proposed the feature fusion single shot multibox detector (FSSD) model, which reconstructs the multi-scale features of the model through feature fusion and down-sampling operation, and enriches the feature details to improve the detection performance on small objects. Liu *et al.* [15] proposed the DeepSat classification framework based on the ''hand-made'' features and deep belief network (DBN). The framework augments a CNN with handcrafted features (instead of using DBN-based architecture) for classification. This method achieves superior performance on sat-4 and sat-6 datasets with the accuracies of 99.90% and 99.84% respectively. Zhou *et al.* [16] proposed a multi-level feature extraction method to solve the problem of object loss of discontinuous object tracking in the image-based remote sensing systems. Chen *et al.* [17] proposed an improved semantic segmentation Neural Network Based on DeepLabv3, which adopts dilated convolution, a fully connected (FC) fusion path and pre-trained encoder for the semantic segmentation task of HRRS imagery, reaching the classification accuracy of 91%. Duarte *et al.* [18] proposed three multi-resolution CNN feature fusion methods to improve the classification accuracy of building damage in the remote sensing images, reaching the accuracy of 88.7% on the satellite and aerial (unmanned) cases. Ni *et al.* [19] proposed a learnable framework of CNN based on the multi-layer energized locality constrained affine subspace coding (MELASC), which improved the accuracy of scene classification for image-based remote sensing applications.

In this paper, we propose an image double segmentation method based on feature point region, which segments the original remote sensing images, maximally retains all the information of an image, and reduces the adverse effects of the segmented image. More specifically, we propose the dilated convolution and feature fusion single shot multibox detector (DFSSD) network, which combines the high-level feature map and the low-level feature map to improve the spatial semantic information of the low-level feature map. At the same time, the dilated convolution [20] is used to allow the third-layer features to participate directly in the prediction, further enriching the detailed information of the network features. The performance of [the proposed] DFSSD network on remote sensing datasets including700 aircraft and 938 car remote sensing images is not inferior to that of the same type of the networks, while the mAP is increased by 4%. compared with the original SSD network model.

II. RELATED WORK

Compared with the traditional image-based object detection methods, object detection using aerial images encounters problems such as the difficulty of detecting small objects, and the lack of sufficient representative features for objects. Traditional methods generally use scale-invariant feature transform (SIFT) [21], speeded up robust features (SURF) [22], features from accelerated segment test (FAST) [23], Binary robust invariant scalable keypoints (BRISK) [24] and so forth to detect objects. Although these methods yield reasonable performance for object detection under plain backgrounds, they do not achieve good results under complex backgrounds. Therefore, deep learning methods with higher capabilities in capturing intricate patterns received increased attention by the research community and have been applied to various tasks related to image-based remote sensing applications such as military guidance, object tracking, urban planning and so forth. The most popular and successful deep learning methods are based on CNN architecture due to their enhanced performance in modeling visual information. At present, DL-based object detection methods can be divided into two main categories including the object detection algorithms based on the candidate region and object detection algorithms based on regression models.

The calculation process of the object detection algorithm based on candidate region includes the following steps. Firstly, n regions of interest (ROI) are extracted from the input image according to the region selection algorithm. The commonly used selection algorithms are selective search, edge boxes, region proposal network (RPN) and so on. Then, a multi-layer convolution neural network is used to extract the above regions of interest and classify the extracted features. Finally, the bounding box regression is used to correct the output window and provide the final result. Some implementations of the object detection algorithm based on the candidate region include R-CNN [11], Fast R-CNN [25], Faster R-CNN [12], R-FCN [26].

Although the above-mentioned object detection algorithms based on the candidate region can provide high accuracy, they cannot detect the moving object in real-time videos. Indeed, the object detection methods without RPN networks have more advantages in terms of operation speed. The deep learning object detection algorithm based on regression models can identify the objects in multiple locations of the original image or the feature map, and directly obtain the type and the location of the object. Some successful implementations include Yolo [27], SSD [13], YOLO9000 [28], DSSD [14], RSSD [16], FSSD [14].

More and more CNN-based methods were used in the field of image-based remote sensing [29], [30]. Zhang *et al.* [31] constructed an iterative weakly supervised learning framework, which can automatically mine and augment the training datasets from the original images. This method combines the framework with the candidate RPN to locate an aircraft in large-scale and extremely high-resolution images. Cai B *et al.* [32] designed an end-to-end convolution neural network to realize the detection of airport objects. The authors of this work proposed a method of mining difficult samples to train the end-to-end deep convolutional neural network for airport detection in complex situation, reaching the accuracy of 83.02% on a optical remote sensing dataset acquired from Google Earth and integrated them into the network architecture. Pang *et al.* [33] proposed a unified and self-reinforced network called remote sensing

region-based convolutional neural network (R2-CNN) to detect small and medium objects in remote sensing images. The network is composed of backbone Tiny-Net, intermediate global attention block, and final classifier and detector, having the high recall and precision in GF-1 images and GF-2 images. Zhao *et al.* [34] proposed a method of aircraft detection based on the Block-Level F-CNN remote sensing images, combining the image block-level fully convolutional neural network model and the multi-scale structure for object detection, reaching the accuracy of 83.02% on an aircraft dataset from the Beijing capital international airport. Liu *et al.* [2] Proposed an end-to-end multi-component fusion network (MCFN) to realize the small airport objects detection of remote sensing images, composing of dual pyramid fusion network (DPFN), relative region proposal network (RRPN) and contextual information network (CIN).

In short, existing object detection methods have some limitation in remote sensing images. First, because of the limitations of CNN, the low-level feature map semantic information is relatively scarce but accurately presents the object location. In contrast, high-level feature semantic information is rich but imprecisely presents the object location. In addition, previous methods cannot adequately extract the features of a small object. Finally, when the object is in a complex scene the accuracy of previous algorithms will be decreased.

III. PROPOSED WORK

A. IMAGESEGMENTATION OF FEATUREPOINT REGION PROPOSALS

When facing complex scenes, the human visual system can quickly perceive different interest objects and give preference to them. This is the perception ability of the human visual attention mechanism independent of the detection environment, which operates solely based on detecting the contrast between the desired objects and the background [35]. In the field of computer vision, using this biology-inspired feature, we can reduce the redundancy of information and quickly detect the object information under various environmental interference [36], [37].

Since the scale of remote sensing images is typically too large, the direct processing of the original image by the convolution networks may cause the training of the network to diverge. Even if it converges, the accuracy of the subsequent object detection and the generalize ability of the model may not be ideal. Therefore, inspired by the characteristics of the human vision system, in this paper, we propose an image double segmentation method based on feature point region, which are used to generate the feature point map of the remote sensing images, as shown in Figure1.

In particular, the human visual attention mechanism is based on quickly perceiving the objects with large color changes. In this paper, we use feature points to represent the objects in the pictures, and automatically select the regions with a relatively large number of feature points as the training data set. The original size of the utilized remote sensing image data set is from 800 × 800 to 4000 × 4000. After the image size test step, as shown in Table 1, the size of the segmented image is reduced to 512 x 512. To find the feature points, we use the binary robust invariant scalable keypoints (BRISK) method [24], which has desirable properties like rotation invariance, scale invariance, robustness, and relatively fast speed. Finally, we use double segmentation on the remote sensing image with feature points to get segmented images of different regions, count the number of feature points in each small image, keep pictures with at least 30 to 60 feature points in each small image, generate an XML file labeled with the object data in each small image, and the number of pictures is increased from 1638 to 31560.

TABLE 1. Image size test.

Image Size	mAP(%)	Fps
300x300	72.18	38
512x512	76.51	27

The method of double segmentation is to cut the original image twice from two directions. For the first time, it starts from the lower left corner of the original image to the upper right corner, and the segmentation size is 512. If the length of the original image is less than 512, it will not be segmented in the horizontal direction; similarly, if the width of the original image is less than 512, it will not be segmented in the vertical direction; if the length and width of the original image are less than 512 when it is segmented to the rightmost and topmost sides, this part of the image will be discarded. The second time starts from the upper left corner to the lower right corner, and the segmentation method is the same as the first time. If there is no object in the segmented image, the image will not be used as training data and be removed directly.

According to the annotation mapping formula as shown in formula (1), the corresponding sub segmented image data annotation XML file is generated:

$$
\begin{aligned}
f(x_{\min}) &= \begin{cases} 0 & x_{\min} \le l_s, \quad n \ge 30 \\ x_{\min} - l_s & x_{\min} > l_s, \quad n \ge 30 \end{cases} \\
f(y_{\min}) &= \begin{cases} 0 & y_{\min} \le w_s, \quad n \ge 30 \\ y_{\min} - w_s & y_{\min} > w_s, \quad n \ge 30 \end{cases} \\
f(x_{max}) &= \begin{cases} C & x_{\max} \ge C, \quad n \ge 30 \\ x_{max} - l_{\text{end}} & x_{\max} < l_{end}, \quad n \ge 30 \end{cases} \\
f(y_{max}) &= \begin{cases} C & y_{\max} \ge C, \quad n \ge 30 \\ y_{max} - w_{end} & y_{\max} < w_{end}, \quad n \ge 30 \end{cases}
\end{aligned} \tag{1}
$$

where $x_{\min}, y_{\min}, x_{max}, y_{\max}$ is the information of the annotation box, l_s is the starting position of the abscissa of the cut image on the original image, l_{end} is the ending position of the abscissa of the cut image on the original image, w_s is the starting position of the ordinate of the cut image on the original image, w_{end} is the ending position of the ordinate

FIGURE 1. Extracted feature points from an exemplary remote sensing image.

of the cut image on the original image, C is the size of the cut image (C = 512 in this paper), n is the number of characteristic points in the cut area.

If $n \geq 50$, the annotation information of this object will be kept completely, and the difficult item in the XML file will be set to 0.If $30 \leq n < 50$, the annotation box will be kept, but the difficult item in the XML file will be set to 1, if $n < 30$, the annotation information of this object will be removed. In this way, we can not only keep the segmented object information as much as possible, but also eliminate the increase of false detection caused by too little information and too much background information.

In this paper, we propose an image segmentation method based on feature points. This method has two advantages: (1) retaining the useful information of the original image. In a remote sensing image, the background information accounts for the most part, and the object information only accounts for a small portion of the image. Using this method, we can reduce the background information as much as possible to keep useful information, and reduce the redundancy; (2) improving the diversity and quantity of the dataset. In this paper, the result of image segmentation using this method is shown in Fig.2.

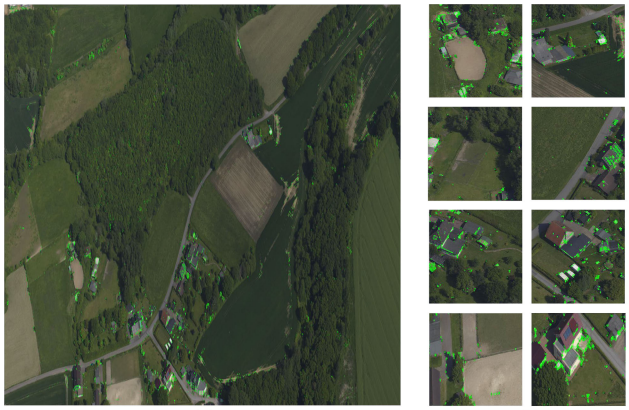

FIGURE 2. Image segmentation results: (a):original image, (b):six images as the result of segmentation stage.

B. NETWORKINFRASTRUCTURES

The proposed DFSSD network can be viewed as an improved version of the SSD network. The benchmark SSD network uses vgg-16 as the basic feature extraction layer, replacing the full connection layers FC6 and fc7 of vgg-16 network structure with two convolution layers, removing the dropout layer and the classification layer in vgg-16, and adding four additional groups of convolution layers. Each group uses 3×3 convolution kernel and 1×1 convolution core to reduce the channel number of the feature map. Different levels of the feature maps are used for the border offset of differently-scaled objects as well as the prediction of different class scores [38]. Finally, the detection results are obtained by the non-maximum suppression (NMS) [39] applied to prediction layer. The feature pyramid network (FPN) network uses the characteristics of multi-scale feature map to detect small objects with low-level and high-resolution feature maps and large objects with high-level and large receptive field feature maps to ensure that objects of different scales can be detected.

C. DFSSDNETWORKSTRUCTUREDESIGN

In this paper, we note the property of small object sizes in remote sensing images. The data enhancement of the SSD network can be considered mainly a random clipping process, which makes the small object with less information lose parts of the information randomly and potentially lead to the decreased detection ability of the final training model for

TABLE 2. Horizontal base layer parameters.

Horizontal base layer	Kernel size	Kernel numbers	Strides	Padding	Feature map
h4_3	1 × 1	256	1	0	64 × 64
h7	1 × 1	256	1	0	32 × 32
h8_2	1 × 1	256	1	0	16 × 16
h9_2	1 × 1	256	1	0	8×8
h10_2	1 × 1	256	1	0	4 × 4

small objects. To avoid this issue, we propose an image segmentation method based on the feature point region proposals method. This method not only compensates for the lack of sample richness, but also improves the detection accuracy.

When using the SSD network, we find out that the prediction layer of the SSD network does not make full use of the local and global semantic features of the lower layer, which leads to the poor detection ability of small objects. For the remote sensing images, the conv4 layer in the SSD network has undergone three down-sampling operations, and the resolution of the resulting feature map is not enough to detect small objects. Therefore, we consider using the Conv3_3 layer feature map. If we use the large convolution kernel or several small convolution kernels to convolute the Conv3_3 layer directly, the semantic information of the feature graph is increased at the cost of increased computation load of the network model training. In order to reduce the computation complexity and accelerate the speed of the training phase, we propose to use dilated convolution to operate the Conv3_3 layer, and combine it with the FPN network structure. We call this method as dilated convolution and feature fusion single shot multibox detector (DFSSD), which improves the size of the receptive field of the feature layer, and increases the semantic information. The detailed network structure of the proposed DFSSD method is shown in Fig. 3.

The proposed DFSSD network model comprises an improved SSD layer, a horizontal base layer, an up-sampling layer, a fusion layer, and a prediction layer. The improved SSD layer is based on the original SSD model, and the detailed parameters following some prior works in the literature including [13]. In this study, we use two dilated convolutions with dilation rates of 12 and 18 for the features of Conv3_3 layer, and then fuse the feature maps of different receptive fields obtained by the convolution operation/layer. At the same time, the convolution kernel of 3 × 3 is used to eliminate the aliasing effect caused by the fusion of different feature maps. The horizontal base layer, the up-sampling layer, the fusion layer, and the prediction layer are improved according to the FPN network structure. Table 2 shows the size of the convolution kernels, the number of the convolution kernels, the strides and the padding of the convolution layers, and the size of the convoluted feature maps. The purpose of this layer is to reduce the number of channels and prepare them to be fused with the latter feature map. The fused feature map not only retains the high-resolution of the lower-level feature map, but also represents better semantic information.

TABLE 3. Feature layer default box scale comparison table.

Feature map	Anchor box size
128 × 128	[15,30]
64 × 64	[30,51]
32 ×32	[51,133]
16 × 16	[133,215]
8 × 8	[215,297]
4 × 4	[297,379]
2×2	[379,461]

The up-sampling layer enlarges the feature map to twice the original size. In the process of the feature map enlargement, there will be many vacancies without pixel values. The vacancies are filled with bilinear interpolation. The number of channels of the feature map is 256. The sizes of the output feature maps in the up-sampling layer are 64 × 64, 32× 32, 16× 16, 8 × 8, and 4 × 4. The purpose of the up-sampling layer is to obtain a feature map of the size needed for the fusion layer. The prediction box scale of the DFSSD network prediction layer is calculated as shown in formula 2:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m - 2}(k - 1) \quad k \in [2, m] \tag{2}$$

where $s_{\min}$ represents the minimum scale of the default frame in the original image, $s_{\max}$ represents the maximum scale of the default frame in the original image, and M represents the number of characteristic maps in the prediction layer. In this paper, $s_{\min} = 0.1$, $s_{\max} = 0.9$, m = 7 are set. The default frame size of conv3 is set manually. When k = 1, set = 0.06, and the default frame size of conv3 layer feature map is 3% - 6% of the original map. Table3 shows the default box scale for each feature map.

The fusion layer implements the pixel addition operation between the feature map obtained from the horizontal base layer and the feature map obtained from the up-sampling layer. The sizes of output feature maps are 64 × 64, 32 × 32, 16 × 16, 8 × 8 and 4 × 4.

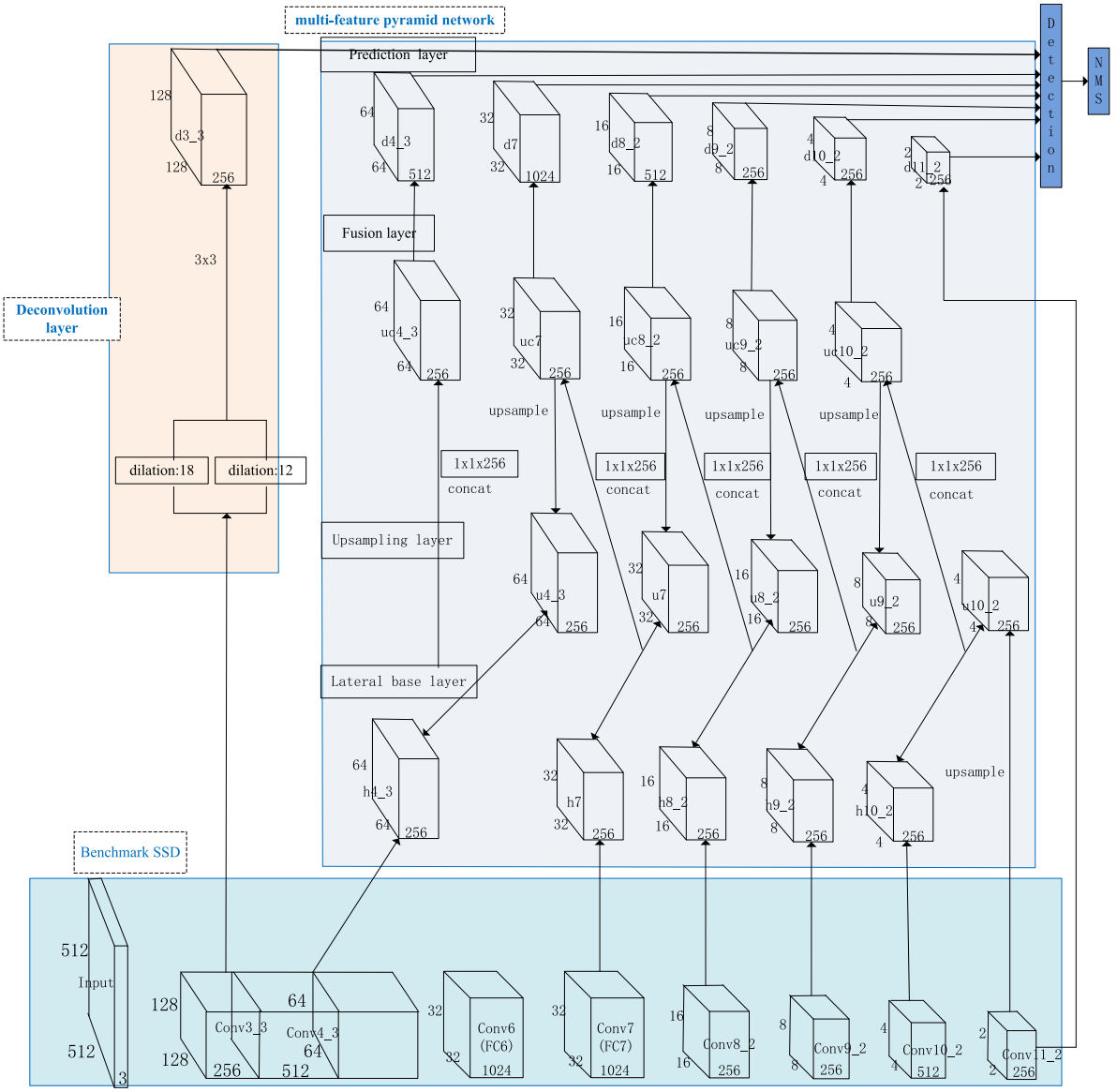

FIGURE 3. The architecture of the proposed DFSSD framework including three main parts: Benchmark SSD, Deconvolution Layer, and multi-feature pyramid network (MFPN). The basic features are extracted by the Benchmark SSD network. In the Deconvolution Layer, we use the dilated convolution to deconvolute the Conv3_3 layer of the SSD to obtain rich low-level feature information. The MFPN module comprises there three parts: the horizontal convolution layer, the up-sampling layer, and the fusion layer. We use the horizontal convolution layer to reduce the dimension of the feature map, fuse the feature map of the upper sampling layer, and obtain the final feature map of the prediction layer through the fusion layer. Finally, we use the non-maximum suppression (NMS)to choose the best bounding box. The proposed method can detect small objects in complex scenes.

Table 4 shows the parameters of the prediction layer. The prediction layer is obtained by the normal convolution of the fusion layer and the dilated convolution of the features of Conv3_3 layer. The purpose of using the convolution kernel of size 3 x 3 is to deblur the feature map of the fusion layer. In the feature image enlargement step, a bilinear interpolation method is used to fill in the vacancies, which may cause the pixel values of the blocks to be similar. This effect may undermine the clarity of the contours around the target objects and make the objects to appear fuzzy, which highlights the need for this operation.

D. LOSS FUNCTION

When training the detection network, it is necessary to save the true value box information for each vehicle position marked in the input image. For each candidate box, the offset of the center point of the candidate box from the center point of the truth box as well as the confidence of the object encompassed by the candidate box should be calculated at the same time. In the training phase, all candidate boxes and the two truth value boxes are first matched according to the Jaccard matching algorithm [40]. The candidate boxes are regarded as matching boxes, whose matching coefficients with the truth

TABLE 4. Parameters of the prediction layer.

Prediction layer	Kernel size	Kernel numbers	Strides	Padding	Feature map
d3_3	3 × 3	512	1	1	128 × 128
d4_3	3 × 3	512	1	1	64 × 64
d7	3 × 3	1024	1	1	32 × 32
d8_2	3 × 3	512	1	1	16 × 16
d9_2	3 × 3	256	1	1	8 × 8
d10_2	3 × 3	256	1	1	4 × 4
d11_2	3 × 3	256	1	1	2 × 2

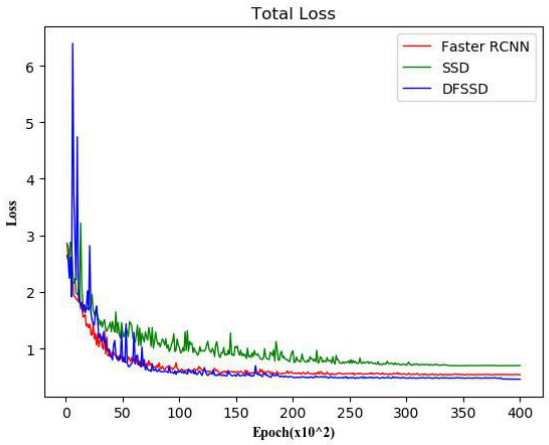

FIGURE 4. The loss comparison results of the DFSSD, SSD and faster RCNN methods.

value boxes are greater than 0.5. They will be marked as positive samples denoted by c^1, and other candidate boxes that do not satisfy the minimum matching rate are considered negative samples, denoted by c^0.

In the process of network training, the total loss function includes the classification loss and the location regression loss, calculated as

$$L(p, d, g) = \left(\frac{1}{N}\right)(L_{cls}\ (p) + L_{loc}(d, g)) \tag{3}$$

where p represents the confidence of the category, d represents the candidate box, g represents the true value box, n represents the number of positive samples, L_{cls} represents the classification loss function, and L_{loc} represents the position regression loss function. The resulting total loss function for the image comparison is shown in Fig. 4.

Note that the large fluctuations in the loss function occur only at the beginning of the training phase for epochs below 20. This behavior is normal for deep learning methods, so it is consistently observed across all methods including the SSD, Faster RCN, and DFSSD methods. Therefore, it is sufficient to use epoch numbers above 50 to achieve stable results.

The classification loss L_{cls} is based on the two-class softmax loss. When classifying, the confidence degree belonging to the automobile category is expressed by p^1, and the confidence degree belonging to the background category is expressed by p^0. Therefore, the classification loss function is

$$L_{cls}(p) = -\sum\nolimits_{i\in c^1}^{N} \ln(\hat{p}_i^1) - \sum\nolimits_{i\in c^0}^{N} \ln(\hat{p}_i^0) \tag{4}$$

where $\ln(\hat{p})$ denotes the natural logarithm and we have $\hat{p}_i^1 = \frac{\exp(p_i^1)}{\exp(p_i^0)+\exp(p_i^1)}$ and $\hat{p}_i^0 = 1 - \hat{p}_i^1$.

The position regression loss function $L_{loc}(d, g)$ is the smooth L1 loss of the matching between candidate box d and the truth value box g [21], which can not only ensure that when the difference between the prediction box and the ground truth is too large, the gradient value is not too large, but also ensure that when the difference between the prediction box and the ground truth is very small, the gradient value is small enough. Following the position regression algorithm in benchmark SSD, we calculate the coordinates of the center points of the matching candidate box and the truth value box, and the migration regression of the width and height as

$$L_{loc}(d, g) = \sum\nolimits_{i\in c^1}^{N} \sum SmoothL1(d_i^k - \hat{g}_j^k) \tag{5}$$

where i represents the i^{th} matched candidate box and j represents the j^{th} true value box.

IV. EXPERIMENTALRESULTS ANDANALYSIS

A. INTRODUCTIONOF EXPERIMENTAL ENVIRONMENT

In this study, the experimental environment is centos7 system, the processor model is Inter (R) Xeon (R) CPU e5-2670 V3 @ 2.30 GHz x 12, the graphics card model is NVIDIA GeForce GTX 1080 Ti, the video memory is 11g, the memory is 32g, the experimental framework is Pytorch deep learning framework. Also, the learning rate parameter is 0.001, the weight attenuation parameter is 0.0005. We use the small batch gradient descent algorithm and the momentum optimization algorithm to optimize the parameters, with the mini-batch size of 16. The epoch number of iterations is 400, the number of steps of each epoch is 1000 and the momentum factor is 0.9. The loss function of the DFSSD is basically similar to that of the SSD. The location information of a category is obtained by the regression function, and the classification confidence is predicted by the softmax function.

FIGURE 5. Different size of datasets objects.

TABLE 5. Map result comparison table.

Network Model	SSD	Faster RCNN	DFSSD
Average Precision (%)	69.81	74.57	76.51

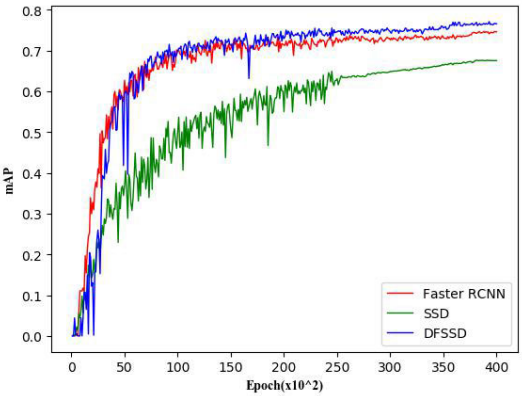

FIGURE 6. mAP curves for different models.

In this work, we have made a remote sensing image dataset about car. According to the format of Pascal voc2007 and 2012 data sets, we use an unmanned aerial vehicle (UAV) remote sensing system to collect vehicle objects in different environments, with a total of 2045 pictures in the dataset including 1138, 500, and 407 images, respectively, for training, verification, and test phases. The actual size of the object in the remote sensing image is shown in Fig. 5.

B. EXPERIMENTALRESULTS

In order to verify the efficacy of the proposed object detection method, we carry out different experiments. We compare the performance of the proposed DFSSD method with the benchmark SSD and Faster RCNN methods. To realize a fair comparison, we train and test the three networks using the same dataset. The achieved results are shown in Fig. 5, Fig. 6, and Table 5.

From the mAP comparison results between the different models in Table 5 and Fig. 6, it can be seen that the mAP of the DFSSD model is more accurate than the other methods. The achieved detection accuracy is 2% higher than the Faster RCNN method and about 7% higher than the baseline SSD method. The superior performance of the DFSSD method in terms of the high detection rate is mainly due to the use of deconvolution step, which enriches the low-level features. At the same time, the fusion of different levels of feature maps can effectively improve the detection accuracy.

TABLE 6. Comparison of experimental results.

Data set	Network Model	mAP(%)	FPS
NWPU	SSD	44.77	18
	DFSSD	65.35	11
RSOD	SSD	31.30	18
	DFSSD	51.78	12

Fig. 7 illustrates the enhanced detection accuracy of the DFSSD model compared with the SSD and Faster RCNN methods. Since the SSD and Faster RCNN methods are primarily designed for object detection in natural scenes, they cannot accommodate the requirements of the small-scaled vehicle detection in remote sensing images. The DFSSD model fully covers the requirements of the remote sensing image through designing different levels of feature fusion methods and enabling different candidate frame scales; therefore, deems effective for vehicle object detection in remote sensing images.

C. COMPARETIVEEXPERIMENTOFOTHERREMOTESENSINGDATASETS

In order to further verify the usability of the proposed method, we have carried out comparative experiments for the DFSSD and SSD methods using the nwpuvhr-10 dataset [41] and rsod dataset [42]. The two networks use the same hyper parameters for learning. The initial learning rate is set to 0.001 and the number of epochs is set to 400. The SSD method does not use our proposed data processing method, whereas the DFSSD method uses the proposed preprocessing method for training and testing. The experimental results are shown in Table 6.

(a) (b) (c)

FIGURE 7. Experimental results for different methods using three different images. Columns a-c represents the results obtained by the SSD, Faster RCNN, and DFSSD methods, respectively.

D. SPEEDTEST

In order to confirm the feasibility of using the DFSSD method in real-time applications, we carry out comparative experiments on different networks. The input size of the SSD network is 512, which uses vgg-16 basic network, and the input size of the Faster RCNN network is1000 x 600, which uses vgg16 basic network. Because of the large-scale difference of the remote sensing images, we test the image in different scale ranges, and set the resolution of image to 0-1000,1000-2000,2000-3000, and 3000-4000. The test results are shown in Fig. 8.

As can be seen in Fig. 9, the single forward inference time for the DFSSD method is 59ms, which is 11ms longer than that of the SSD. The main reason for the slightly decreased speed is the more complex network structure used in the DFSSD method. The DFSSD method apparently shows a big advantage over the Faster RCNN in terms of the operation speed. Overall, we conclude that the DFSSD method does not compromise on the operation time while considerably improving the detection accuracy.

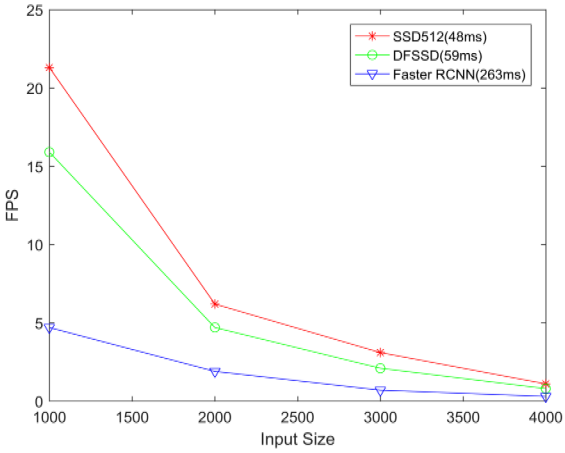

FIGURE 8. Speed test.

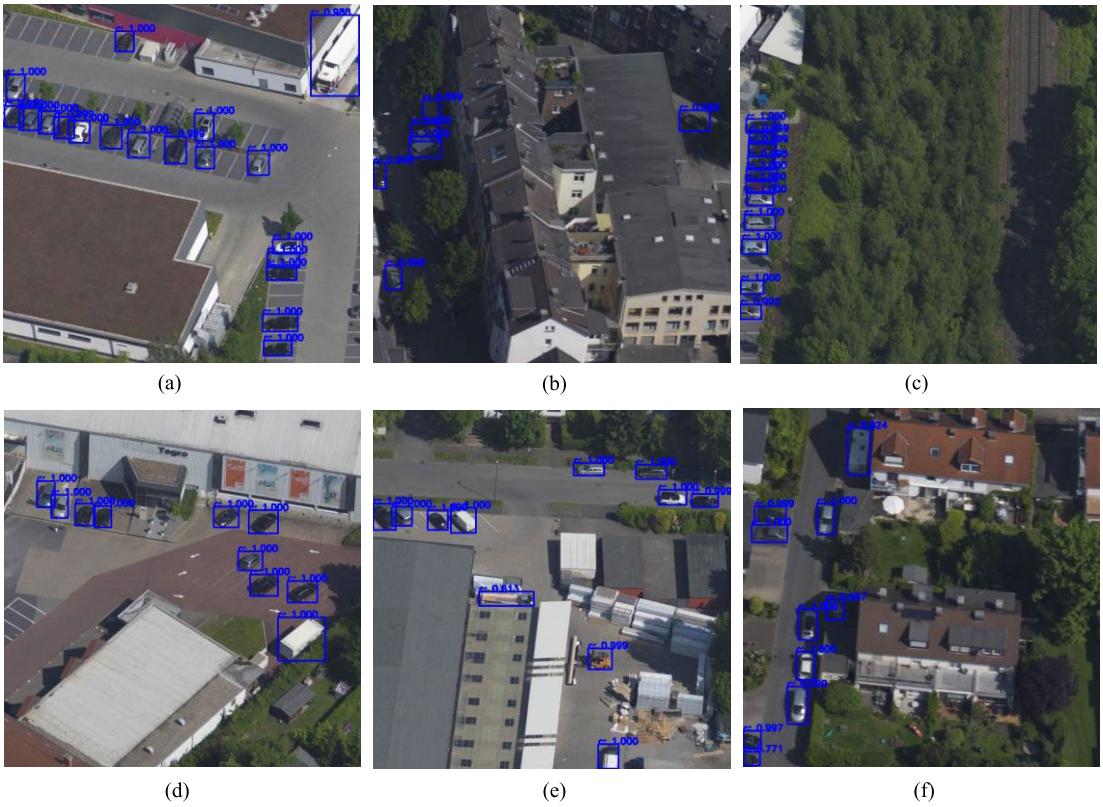

FIGURE 9. The object (vehicle) detection accuracy of the proposed method in six different scenes.

Fig. 9 present the final results of the proposed DFSSD algorithm in terms of detecting vehicles at six different scenes in aerial images. It can be seen that the proposed DFSSD method obtains sufficient spatial structural information about small objects. The more semantic information obtained by the DFSSD method enhances the feature representation of small objects. The proposed method identifies more small samples to use in the training, so it can make the model learn more local information and small objects information. In different scenes, the experimental results confirm the effectiveness of the proposed method.

Compared with other methods, the proposed method is more accurate, especially in the detection of small objects. It is also very effective for objects with complex scenes and occlusion. However, in some scenes, the distance between the objects is very close, that is, many parts of the objects are also connected together, so the proposed method cannot detect them correctly.

V. CONCLUSION

In this paper, we proposed a novel deep learning method, named DFSSD for small-scaled object detection with applications to aerial remote sensing images. The proposed method significantly improves upon an already successful method, called SSD. The key idea is an enhanced image segmentation processing approach based on the extracted feature points of remote sensing images, so that the resulting image segments retain maximal information for the small-scaled objects after scaling. The proposed method replaces the random clipping step of the SSD network, and hence alleviates the adverse effects of the random clipping on small objects in the training phase. For small object detection in remote sensing images, DFSSD uses two different dilated rate convolution kernels to perform multi-scale fusion on the Conv3_3 layer, which expands the receptive field of the feature map. At the same time, based on the FPN network structure, a DFSSD network prediction layer is designed, and feature maps of different layers are integrated to capture multi-scale context information, which improves the network ability to detect small objects. In the prediction phase, the overlapping object frame is removed by non-maximum suppression of the original image. The proposed method, on the premise of ensuring the real-time detection speed of DFSSD network as much as possible, improves the object detection accuracy of remote sensing images by 4% compared with the benchmark SSD network.

REFERENCES

[1] W. Suyu and S. Lansun, "Intelligent visual surveillance technology: A survey," *China Southern Agricult. Machinery*, vol. 12, no. 9, pp. 1505–1514, 2017.

[2] J. Liu, S. Yang, L. Tian, W. Guo, B. Zhou, J. Jia, and H. Ling, "Multi-component fusion network for small object detection in remote sensing images," *IEEE Access*, vol. 7, pp. 128339–128352, Sep. 2019.

[3] M. Qingchun, Q. Yong, and Z. Shujun, "Intelligent robots and development," *J. Ocean Univ. Qingdao*, vol. 2018, no. 4, pp. 123–124, 2018.
[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[5] D. Tom, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detection," *Signal Process., Image Commun.*, vol. 47, pp. 482–489, Sep. 2016.
[6] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, no. 2, pp. 1097–1105, 2012.
[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 818–833.
[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Comput. Sci.*, vol. 12, pp. 1–13, Dec. 2013.
[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[13] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Sep. 2016, pp. 21–37.
[14] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2018, *arXiv:1712.00960*. [Online]. Avaliable: https://arxiv.org/abs/1712.00960
[15] Q. Liu, S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat V2: Feature augmented convolutional neural nets for satellite image classification," *Remote Sens. Lett.*, vol. 11, no. 2, pp. 156–165, 2020, doi: 10.1080/2150704X.2019.1693071.
[16] B. Zhou, X. Duan, D. Ye, W. Wei, M. Wozniak, D. Polap, and R. Damaševicius, "Multi-level features extraction for discontinuous target tracking in remote sensing image monitoring," *Sensors*, vol. 19, no. 22, p. 4855, 2019, doi: 10.3390/s19224855.
[17] G. Chen, C. Li, W. Wei, W. Jing, M. Wozniak, T. Blažauskas, and R. Damaševicius, "Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation," *Appl. Sci.*, vol. 9, no. 9, p. 1816, 2019, doi: 10.3390/app9091816.
[18] D. Duarte, F. Nex, N. Kerle, and G. Vosselman, "Multi-resolution feature fusion for image classification of building damages with convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, p. 1636, 2018, doi: 10.3390/rs10101636.
[19] K. Ni and Y. Wu, "Scene classification from remote sensing images using mid-level deep feature learning," *Int. J. Remote Sens.*, vol. 41, no. 4, pp. 1415–1436, 2020, doi: 10.1080/01431161.2019.1667551.
[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.
[21] M. Guerrero, "A comparative study of three image matcing algorithms: Sift, surf, and fast," Ph.D. dissertation, Sep. 2011. [Online]. Available: https://digitalcommons.usu.edu/etd/1040
[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[23] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.
[24] J. He, Y. Li, and H. Lu, "Research of UAV aerial image mosaic based on SIFT," *Opto-Electron. Eng.*, vol. 2, p. 021, Feb. 2011.
[25] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[26] K. J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," NIPS, 2016.
[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[29] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 4379–4382.
[30] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.
[31] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.
[32] B. Cai, Z. Jiang, H. Zhang, Y. Yao, and J. Huang, "Training deep convolution neural network with hard example mining for airport detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 862–865.
[33] J. Pang, C. Li, and J. Shi, "R2-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Mar. 2019.
[34] W. Zhao, W. Ma, L. Jiao, P. Chen, S. Yang, and B. Hou, "Multi-scale image block-level F-CNN for remote sensing images object detection," *IEEE Access*, vol. 7, pp. 43607–43621, 2019.
[35] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 4379–4382.
[36] S. E. Petersen and M. I. Posner, "The attention system of the human brain: 20 years after," *Annu. Rev. Neurosci.*, vol. 35, no. 1, pp. 73–89, Jul. 2012.
[37] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *Neuroscientist*, vol. 20, no. 5, pp. 509–521, Oct. 2014.
[38] H. Tang, Z. Sun, J. Wang, and M. Qian "Clothing image recognition based on VGG-19 hybrid migration learning model," *J. Xi'an Univ. Posts Telecommun.*, vol. 23, no. 6, pp. 87–93, 2018.
[39] S. Qiu, G. Wen, Z. Deng, J. Liu, and Y. Fan, "Accurate non-maximum suppression for object detection in high-resolution remote sensing images," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 237–246, Mar. 2018.
[40] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
[41] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.
[42] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2155–2162, doi: 10.1109/CVPR.2014.276.

JUNSUO QU received the B.S. degree in telecommunication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 1991, and the M.S. degree in communication and information systems from Xidian University, Xi'an, China, in 1998. He is currently a Full Professor with the School of Automation, Xi'an University of Posts and Telecommunications and a member of the China Institute of Communications. He is also the Director of the Xi'an Key Laboratory of Advanced Control and Intelligent Process. He is leading an IoT Research Team with the School of Automation.

CHANG SU received the B.S. degree in electronics science and technology from Ningxia Normal University, Guyuan, China, in 2017. He is currently pursuing the master's degree with the Xi'an University of Posts and Telecommunications. He is 25 years old and his main research topic is the technology and application of the Internet of Things.

ZHIWEI ZHANG received the B.S. degree in communication engineering from Xi'an Shiyou University, Xi'an, China, in 2017. He is currently pursuing the master's degree with the Xi'an University of Posts and Telecommunications. His current main research topic is the deep learning and computer vision. He has a strong interest in this direction.

ABOLFAZL RAZI received the B.Sc. degree in electrical engineering from Sharif University, the M.Sc. degree from Tehran Polytechnic, and the Ph.D. degree in electrical engineering from the University of Maine. He is currently an Assistant Professor of electrical engineering with the School of Informatics, Computing and Cyber Systems (SICCS) and the Director of Wireless Networking and Smart Health (WiNeSH) Research Laboratory, Northern Arizona University. Prior to joining NAU, he held postdoctoral position with the Electrical and Computer Engineering Department, Duke University, where he developed novel information-theoretic methods for dictionary learning, compressive sensing, and inverse problems. He also held a postdoctoral associate position with Case Western Reserve University, where he developed computational methods based on Bayesian inference for integrative analysis of cancer omics data. In addition to his academic service, he served about seven years in wireless industry holding several positions including the Project Manager of value added services, the Research and Development Researcher, the Network Optimization and Integration Engineer, and the Smart Card Design, and Test Engineer.