IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 18, 2019, accepted September 1, 2019, date of publication September 4, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939488

Multi-Component Fusion Network for Small Object Detection in Remote Sensing Images

JING LIU[1,2,3], SHUOJIN YANG[1,2], LIANG TIAN[1,2], WEI GUO[2], BINGYIN ZHOU[2], JIANQING JIA[2], AND HAIBIN LING[4]

[1]College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang 050024, China
[2]Key Laboratory of Augmented Reality, College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, China
[3]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[4]Center for Data Analytics and Biomedical Informatics, Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

Corresponding authors: Liang Tian (tianliang@chmiot.net) and Wei Guo (guowei@chmiot.net)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802109, in part by the Science and Technology Foundation of Hebei Province Higher Education under Grant QN2019166, in part by the Natural Science Foundation of Hebei Province under Grant F2017205066, and in part by the Science Foundation of Hebei Normal University under Grant L2017B06, Grant L2018K02, and Grant L2019K01.

ABSTRACT Small object detection is a major challenge in the field of object detection. With the development of deep learning, many methods based on deep convolutional neural networks (DCNNs) have greatly improved the speed of detection while ensuring accuracy. However, due to the contradiction between the spatial details and semantic information of DCNNs, previous deep learning methods often meet problems when detecting small objects. The challenge can be more serious in complex scenes involving similar background objects and/or occlusion, such as in remote sensing imagery. In this paper, we propose an end-to-end DCNN called the multi-component fusion network (MCFN) to improve the accuracy of small object detection in such cases. First, we propose a dual pyramid fusion network, which densely concatenates spatial information and semantic information to extract small object features via encoding and decoding operations. Then we use a relative region proposal network to adequately extract the features of small objects samples and parts of objects. Finally, to achieve robustness against background disturbance, we add contextual information to the proposal regions before final detection. Experimental evaluations demonstrate that the proposed method significantly improves the accuracy of object detection in remote sensing images compared with other state-of-the-art methods, especially in complex scenes with the conditions of occlusion.

INDEX TERMS Small object, remote sensing, multi-component, dual pyramid fusion, occlusion, complex scene.

I. INTRODUCTION

With the improvements in earth observation technology and the diversity of remote sensing platforms, object detection on remote sensing images has attracted more and more attention [1]–[3]. However, due to the complex backgrounds, small objects, the uneven distributions of training samples in terms of size and quantity, illumination and occlusion detection tasks are challenging. Existing object detection innovations [4]–[10] can be divided into two main categories, traditional machine learning methods and deep learning methods.

Traditional machine learning methods include the scale-invariant feature transform (SIFT) [4] and histogram of oriented gradients (HOG) [5]. These methods first use the traditional filters to extract features and then perform feature fusion and dimension reduction to concisely extract features. Finally, the features are fed into a classifier like Support Vector Machine (SVM) [11] or AdaBoost [12], which rely on hand-engineered features; however, these classifiers have difficulty to efficiently processing remote sensing images in the context of big data. In addition, hand-engineered features

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

FIGURE 1. The object can not be accurately detected in complex scene and in the occlusion situation. The aircraft (red box) in the complex scene (object and background have the same color and texture) shows in (a). the aircraft (blue box) which is partly exposed shows in (b).

can detect only specific targets, when applying them to other objects, the detection results are unsatisfactory.

In recent years, deep learning algorithms based on DCNNs including region proposals with a convolution neural network (faster RCNN) [13], and one-stage networks such as You Only Look Once (YOLO) [14] and the single-shot multibox detector (SSD) [15], have achieved good performance in object detection tasks. However, these innovations usually fail to detect very small objects because small object features are lost during the downsampling processes of DCNNs; i.e., the DCNNs can not accurately extract the features of small objects. Objects in optical remote sensing images usually have small objects, and the objects are usually blurry, which has created considerable challenges in normal object detection with no good solutions to date.

To alleviate the issues of small object detection, many methods such as feature pyramid network (FPN) [16], deeply supervised object detectors [17], and scale normalization for image pyramids [18] have been proposed. To a certain extent, these methods strengthen the feature extraction of small objects. However, they do not perform well when detecting remote sensing objects because many objects in remote sensing images have complex backgrounds due to terrain or illumination factors, and the above methods cannot easily distinguish them.

In the field of object detection in remote sensing images, many DCNN-based methods have been proposed to improve the detection accuracy. Many of object detection algorithms [3], [19], [20] consider only the features of the objects themselves. The DNN-based method [21] used CNN features from combined layers to perform orientation-robust aerial object detection. A position-sensitive balancing (PSB) framework [7] based on the ResNet [22] and a novel end-to-end adaptively aspect-ratio multi scale network (AARMNet) [8] can significantly improve detection accuracy. Figure. 1, (a) shows an aircraft (red box) in a complex scene in which the object and background have the same color and texture. Due to the diversity and complexity of the objects in remote sensing images, in many cases, which the object and the background have similar texture and color features, these methods can not detect them accurately.

Some existing works [9], [23] take local contextual information into account and obtain good performance. The relationships among objects play an important role in improving detection. Therefore, in addition to the use of local contextual information, the proposed method takes object-object relationship contextual information into consideration. In the condition that many objects are occluded by buildings, trees or because of the shooting angle; for example Fig. 1 (b) shows an aircraft (blue box) that is partly exposed. Thus, there are not always enough features to detect the objects.

To address these issues, we propose a framework called the multi-component fusion network (MCFN) to extract the essential features of objects from remote sensing images and accurately detect them.

The major contributions of this paper are summarized as follows:

1) First, we propose the *dual pyramid fusion network*(DPFN) to extract features for small objects. We concatenate the shallow feature map and the deep feature map in every scale through a series of encoders and decoders. Because the object in the remote sensing image is very small, with the deepening of the neural network, the features of small objects will be considerably decreased. DPFN can extract not only shallow spatial location features of small objects, but also deep semantic features. Multiscale feature concatenation can strengthen the features of objects, especially those of small objects.
2) Second, we propose the *relative region proposal network* (RRPN), which maximally uses the small object features and makes the network learn local information

for detecting occluded objects. The intersection over union (IoU) is an important indicator for determining whether anchors are used for training. Previous methods cannot fully use small object samples, because the small object always have low IoU values. To solve this problem, we not only use the IoU as the classification indicator for positive and negative samples but also consider small object properties. We introduce the concept of *relative intersection over union* (RIoU) and use the anchors that satisfy the condition of RIoU for gradient backpropagation.

3) Third, we interpolate contextual information to increase the scale of the proposal region before Region of Interesting Alignment [45], while learning the relationship between background information and object information. In general, background and object information are strongly correlated. Thus, this step is helpful for detecting small objects, especially when the objects are similar to the background. After the relative region proposal network, we divide the proposals into four parts according to the scale of the proposals, and use the *contextual information network* (CIN) for the final classification and regression.

Compared to previous state-of-the-art methods, the proposed method improves the results obtained for remote sensing object detection dataset [24] by 2%, especially for complex scenes and objects that are occluded. The algorithm also performs well on the Microsoft Common Objects in Context (COCO) [25] dataset, especially for small object detection, which indicates that the proposed method is generally accurate.

The remainder of this paper is organized as follows. Section II provides a summary of the related work. We present the proposed method and details in Section III. Section IV introduces the training details. The experimental results are presented in Section V. We discuss about the deficiency of the proposed method in Section VI. Section VII provides some conclusions with suggestions for future work.

II. RELATED WORK

Object detection in remote sensing images has been widely researched in recent years. Deep learning models have received increased attention and being applied to various tasks related to remote sensing images. The most popular deep learning methods are CNN-based models. CNNs do not require handcrafted features, and they require fewer parameters than other networks because they share weights for the same filter. The CNN-based model can learn the essential features of input images based on its network structure. CNNs have been widely used in object classification, object detection, and speech recognition. Many theoretical studies concerning CNN [26] have been conducted. As computer technology has advanced, deeper and more efficient CNN models have been proposed.

AlexNet, developed by Krizhevsky *et al.* [27], was a groundbreaking CNN architecture. The main feature of the GoogleNet [28] is its improved utilization of the computing resources inside the network. This improvement was achieved through a carefully crafted design that allowed the depth and width of the network to increase while keeping the computational demands constant. The VGG models proposed by Simonyan and Zisserman [29] were used to investigate the relationship between the depth of a convolutional network and its accuracy in large-scale image recognition regardless of the size or scale of the image, thus eliminating the requirement for a fixed-size input image. ResNet [22] was reformulated to learn residual functions with reference to the layer inputs instead of learning unreferenced functions to ease the training of networks that are substantially deeper than those used previously. DenseNet [30] based on the ResNet uses dense connections to enhance the feature propagation, and greatly reduce the numbers of parameters.

With the application of DCNN in object detection, increasingly efficient detection algorithms, such as RCNN [18], the spatial pyramid pooling network [31], and fast-RCNN [32], have been proposed. faster-RCNN uses a region proposal network (RPN) structure and improves the detection efficiency while achieving end-to-end training. Instead of relying on regional proposals, YOLO [14] and SSD [15] directly estimate the object region and truly enable real-time detection. The FPN adopts the multiscale feature pyramid form and makes full use of the feature map to achieve better detection results. Region-based fully convolutional networks (R-FCN) [33] build a fully convolutional network, which greatly reduces the number of parameters, improves the detection speed, and has a good detection effect.

More and more CNN based methods were used in the field of remote sensing images, many solutions use the CNN-based model to train the very high resolution remote sensing image datasets [34], [35]. Wu *et al.* [36] proposed an efficient way to automatically learn the presentations from the passive image data and increase the computational efficiency of aircraft detection. Ding et al. [37] investigated the capabilities of a CNN model combined with data augmentation operations in SAR target recognition. Zhang *et al.* [38] designed a network with a deconvolution layer after the last convolution layer of base network for small object detection on high resolution remote sensing images. Ševo and Avramović [39] presented an automatic content-based analysis of aerial imagery in order to detect and mark arbitrary objects or regions in high-resolution images. Zhang *et al.* [40] presented a hierarchical oil tank detector with deep surrounding features combined with local features to describe oil tanks and then applied gradient orientation to select candidate regions from satellite images. Salberg *et al.* [41] investigated an algorithm for automatic detection of seals in aerial remote sensing images using features extracted from a pre-trained deep convolutional neural network.

Zhang *et al.* [19] constructed an iterative weakly supervised learning framework to automatically mine and augment the training dataset from the original image and combined the frame work with the candidate RPN to locate aircraft

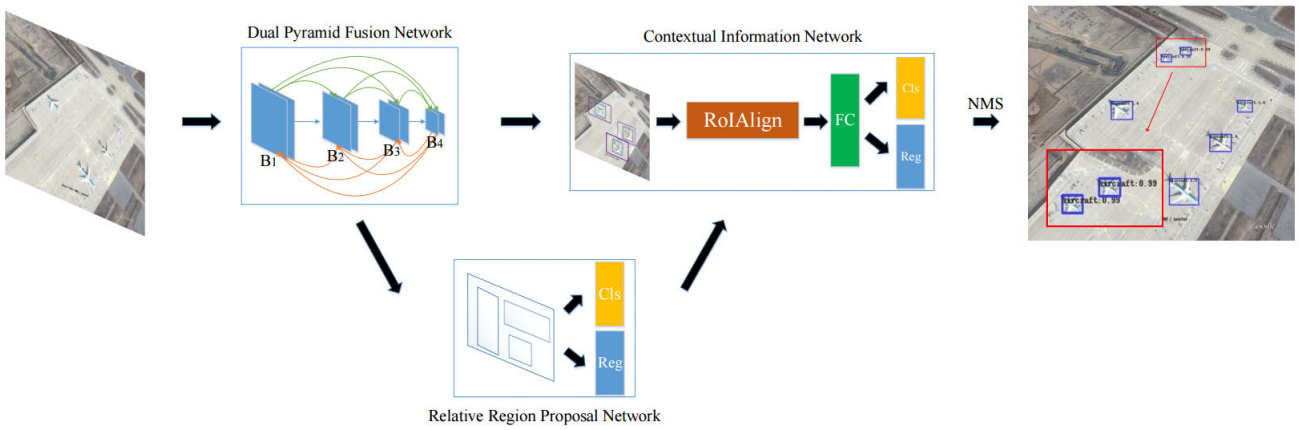


FIGURE 2. The architecture of MCFN framework containing three main parts: DPFN, RRPN, and CIN. In DPFN, the green lines represent encoding operation, and the orange lines represent decoding operation. We use dense concatenation operation to concatenate the feature maps with same scale. In RRPN and CIN, the green box is full-connect layer. The output of yellow box is class probability, and the output of blue box is the probably coordinates. Finally, we use NMS to choose the best bounding box. as the last image show(red box), the proposed MCFN can detect the small object in the complex scene.

in large-scale very high-resolution images. Jiang *et al.* [42] proposed a vehicle detection method in satellite images using DCNNs based on superpixel segmentation. Zhu *et al.* [43] used CNN features from combined layers to perform orientation-robust aerial object detection. Yang *et al.* [3] proposed a dense FPN builds high-level semantic feature maps for all scales by means of dense connections and adopts rotation anchors to avoid the side effects of non-maximum suppression to solve the problems resulting from the narrow width of ships in aerial remote sensing images. To address small remote sensing objects that usually renders poor performance, Ren *et al.* [23] investigated modified faster R-CNN for the task of small object detection in optical remote sensing images.

Methods such as FPN [16], DFPN [3], RetinaNet [44], and modified faster-RCNN [23], which use a pyramidal representation and combine features of DCNNs, have access to high-level semantic information layers with shallow convolutional layers that have access to high-level location information, which has a certain effect on small object detection. In the proposed method, we strengthen the fusion of deep and shallow layers. In addition, we presented a RRPN to maximize the extraction of small object features and local information and interpolate contextual information to improve the accuracy of remote sensing image object detection.

III. PROPOSED WORK

Existing object detection methods have limitation in remote sensing images. First, because of the limitations of DCNN, 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. The last, when the object is in a complex scene the accuracy of previous algorithms will be decreased.

Therefore, we proposed a multi-component fusion network to enhance the accuracy of the small object. The figure 2 shows the architecture of the multi-component fusion network (MCFN) framework, which includes three main parts: dual pyramid fusion network (DPFN), relative region proposal network (RRPN) and contextual information network (CIN). The DPFN includes a series of encoder and decoder to adequately extract the features of the small object. RRPN maximally extracts the small object features and makes the network learn the local information for detecting the occluded object. In the CIN, we interpolate the contextual information to learn the relationship between the background information and the object information. Finally, we use non-maximum suppression (NMS) operations and output the categories and the best bounding box of the object from the input image. The key layers of the network are detailed and shown in Table 1.

A. DUAL PYRAMID FUSION NETWORK

In general, the feature maps in lower resolutions depicts more global scene layout, while the feature maps in higher resolutions contain more structure details. Due to pooling operations and convolution operations with strides in CNNs, a large amount of low-level visual features are lost especially in small objects.

As Fig. 3 shows, B_1, B_2, B_3, and B_4 are four-scale convolutional layers in backbone network. The heatmaps of the B_1, B_2, B_3, and B_4 is (a), (b), (c) and (d) showing in Fig. 4. From the subgraph we can see that (a) contains more edge information, the (b) contains more profile information and the (c) and (d) contain more local and global information.

It is conclude that the first few layers contain more information of small objects, on the contrary, the heatmaps collected from latter layers contain strong semantic information but less information of small objects. That is to say, the distinctive features of small objects in remote sensing imagery

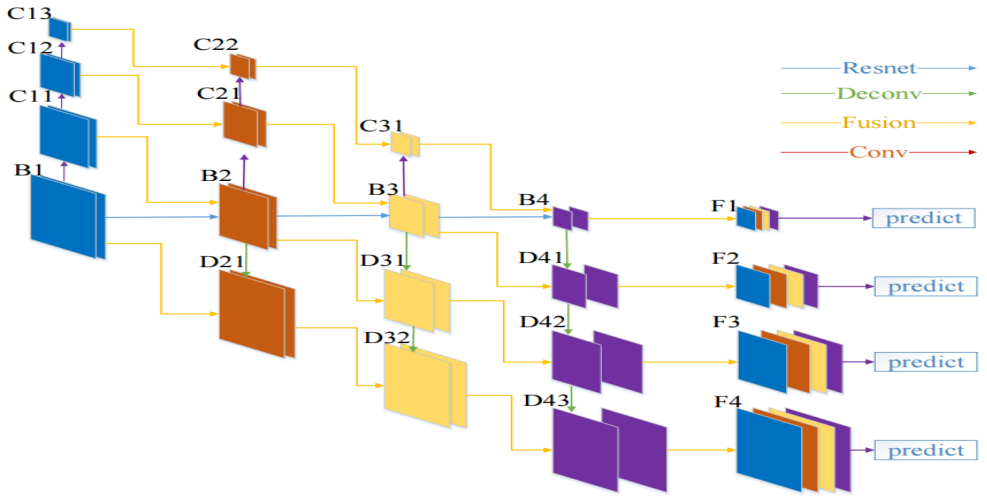


FIGURE 3. The details of DPFN network. C_{11}, C_{12} and C_{13} are the feature maps of different scales obtained by the convolutional of B_1. Similarly, D_{41}, D_{42}, D_{43}, are the feature maps of different scales obtained by deconvolution of B_4. Other feature maps are analogously obtained by convolutional and deconvolutional operations. Finally we concat the same scale feature map and use 3 × 3 convolutional layer to prevent aliasing effects.

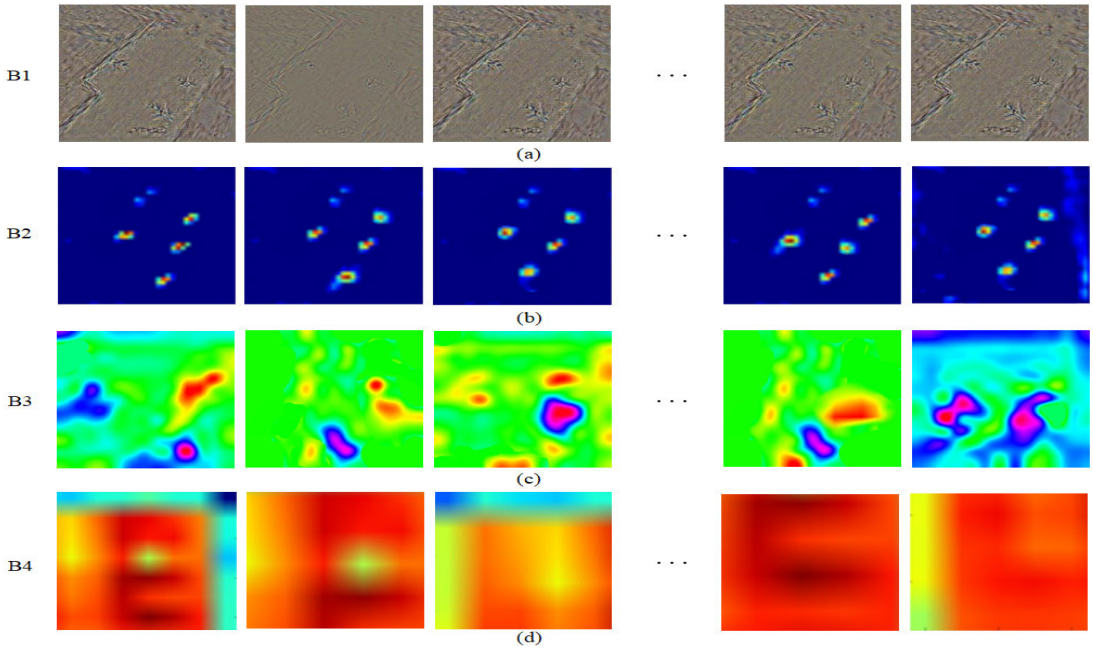


FIGURE 4. The heatmaps (a), (b), (c) and (d) are respectively collected from the B_1, B_2, B_3 and B_4 feature maps in backbone network. From the subgraph we can see that (a) (64 feature maps) contains more edge information, the (b) (128 feature maps) contains more profile information and the (c) (256 feature maps) and (d) (512 feature maps) contain more local and global information. It is obvious that the first few layers contain more information of small objects, on the contrary, the heatmaps collected from the latter layers contain strong semantic information but less information of small objects.

are mainly preserved in the forepart of the whole network based on CNN. So it is difficult for the detector to detect the small object without the lost low-level structure details. As a result, both low-level features and high-level features are all useful to detect small objects.

In order to obtain sufficient information for the prediction in small objects, motivated by FPN [16] we proposed the dual pyramid including a series of convolutional and deconvolutional processes. The convolutional operation can obtain the better feature map in the low-level convolutional layers. The deconvolutional operation can obtain the better in the high-level convolutional layers.

In Fig 3, C_{11}, C_{12} and C_{13} are the feature maps of different scales obtained by the convolutional of B_1. They have more position information in three scales. Similarly, D_{41}, D_{42}, D_{43}, are the feature maps of different scales obtained by deconvolution of B_4. They have more semantic information.

TABLE 1. Architectures for MCFN. In the DPFN Downsampling is performed by C_{11}, C_{12}, C_{13}, C_{21}, C_{22}, and C_{31} with a stride of 2.Upsampling is performed by D_{21}, D_{31}, D_{32}, D_{41}, D_{42} and D_{43} with a rate of 2. F_1, F_2, F_3, F_4 is the result which concat every same scale feature map. In RRPN we use full-convolution to output 2 classes(object or background) and 4 cordinates. In CIN, F_{c1}, F_{c2}, F_{c3}, F_{c4} are the feature maps contain the contextual region. we use RoIAlign, and output k classes (depends on dataset) and 4 cordinates.

Input(dimension)	Layer setting	Output(dimension)
Image(800x800x3)	Resnet-101	B1,B2,B3,B4
Dual Pyramid Fusion Network		
B1(400x400x64)	3x3x64	C11
C11(200x200x64)	3x3x64	C12
C12(100x100x64)	3x3x64	C13(50x50x64)
B2(200x200x128)	Deconv, 3x3x128	D21(400x400x128)
B2(200x200x128)	3x3x128	C21
C21(100x100x128)	3x3x128	C22(50x50x128)
B3(100x100x256)	Deconv, 3x3x256	D31
D31(200x200x256)	Deconv, 3x3x256	D32(400x400x256)
B3(100x100x256)	3x3x256	C31(50x50x256)
B4(50x50x512)	Deconv, 3x3x512	D41
D41(100x100x512)	Deconv, 3x3x512	D42
D42(200x200x512)	Deconv, 3x3x512	D43(400x400x512)
C13,C22,C31,B4	Concat, 3x3x512	F1(50x50x512)
C12,C21,B3,D41	Concat, 3x3x512	F2(100x100x512)
C11,B2,D31,D42	Concat, 3x3x512	F3(200x200x512)
B1,D21,D32,D43	Concat, 3x3x512	F4(400x400x512)
Relative Region Proposal Network		
F1,F2,F3,F4	Full-conv, 1x1	2(classes),4(coordinates)
Contextul Information Network		
Fc1,Fc2,Fc3,Fc4	RoIAlign, 7x7	k(classes),4(coordinates)

Other feature maps are analogously obtained by convolutional and deconvolutional operations.

These concatenation feature maps have both position information and semantic information which are effective in predicting the object, especially small objects. To prevent aliasing effects, we use 3×3 convolutional layers after concatenating every four-scale convolutional layer. Finally, we obtain a four-scale feature map F_1, F_2, F_3, F_4. The four-scale feature map is partly used in the RRPN for prediction.

B. RELATIVE REGION PROPOSAL NETWORK

IoU is an important concept in object detection. It depends on whether the sample is actually used for training. In general, the anchor size we chose was suitable for most objects. However, if there were some relatively small objects the anchors were not used for gradient backpropagation because of the low IoU, but the information still helps the model learn features. Figure 5 shows two situations in which the anchors do not use for training. In the situation I, the object is small, and the IoU is lower than the threshold that we usually use. However, the information in the anchor is useful for training the model. In the situation II, the object is big, and the IoU is much lower than the threshold that we usually use, however, the information in the anchors is important for the network to learn the local features to detect occluded objects.

To maximally uses the small object features, the proposed work uses the RRPN. We redefined the way the foreground

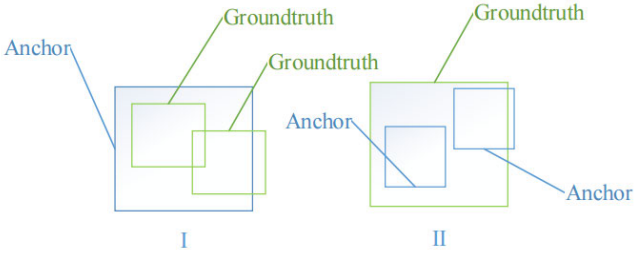


FIGURE 5. The two situations of the IoU. In the situation I, the object is small, the IoU is lower than the threhold we usually use. In the situation II, the object is big, the IoU is much lower than the threshold we usually use.

and background labels were assigned. Firstly, we chose the anchors with IoU values greater than a threshold (such as 0.3). Secondly, we chose RIoU values greater than another threshold (such as 0.7). We assigned these anchors with label "1". Similarly, for the background definition, first, we chose an IoU of less than a threshold (such as 0.3). Second, we chose an RIoU of less than another threshold (such as 0.3). We assigned these anchors with the label "0". Others that were not used for training were assigned the label "−1". For training, the same number (such as 128) of positive samples and negative samples were chosen according to the value of the IoU and RIoU. The RIoU is defined as follows:

$$RIoU = \frac{IoU}{s} \tag{1}$$

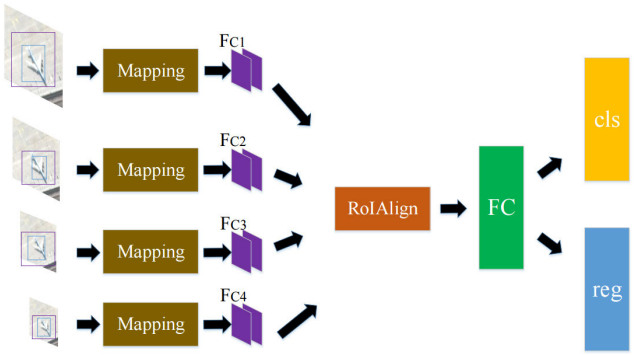


FIGURE 6. The arhciteture of the context information network. after generating the object proposals in the RRPN stage, we use a feature mapping process to choose a suitable feature maps and sends it to the RoIAlign.

$$s = \frac{\max[A_a, A_g]}{\min[A_a, A_g]} \tag{2}$$

where, A_g and A_a are the square of the groundtruth and anchors.

C. CONTEXTUAL INFORMATION NETWORK

The feature maps corresponding to small candidate proposals, whose spatial resolution are very small after enduring multiple convolution processes and RoIAlign [45] which is a pooling operation, are used to obtain feature maps of the same scale. Moreover, contextual information contributes to object detection and can learn the relationship between small remote sensing objects and the background.

In this section, we focus on leveraging the contextual information to enlarge the spatial resolution of small remote sensing objects. It is effective to distinguish the foreground from the background in the complex scene and improve the small object detection accuracy. In Fig. 6, after generating the object proposals in the RRPN stage, the proposed model begins a feature mapping process to choose a suitable feature maps and transfer it to RoIAlign [45]. The feature maps corresponding to each proposal are then transformed into a fixed-dimensional representation with a predefined spatial resolution. Then, several convolutional layers and fully connected layers are fed with these feature maps for classification and class-specific bounding box regression. Then NMS is used to select the best bounding box and output the category label and probabilities, as shown in Fig. 6. The index "N" of the four-scale feature to select is defined as:

$$M = \left\lceil \frac{\sqrt{w \bullet h}}{32} \right\rceil \left(\frac{\max[w, h]}{\min[w, h]} < k \right) \tag{3}$$

$$N = \begin{cases} 4, & M = 1 \\ 3, & M = 2 \\ 2, & M = 3, 4 \\ 1, & \text{others} \end{cases} \tag{4}$$

where, w and h is weight and height of the proposals with contextual region. k is the hyper-parameter which limit the proposal not so slender or flat which may reduce the network performance. For example if the proposal has a width of 30 and a height of 20, the "M" can be calculated to be "1"(Eq.3), and the "N" is "4"(Eq.4). So we select the corresponding region of interesting in the F_4 feature map.

D. LOSS FUNCTION

The RRPN loss function is defined as:

$$L_{RRPN}(p_i, t_i) = \frac{1}{N} \sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N} \sum_i p_i^* \bullet L_{reg}(t_i, t_i^*) \tag{5}$$

$$p_i^* = \begin{cases} 1, & \text{positive samples} \\ 0, & \text{negative samples} \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

where, i is the index of an anchor in a minibatch, and p_i is the predicted probability of anchor i being an object. The groundtruth label p_i^* is "1" if the anchor is positive, and "0" if the anchor is negative. t_i (Eq.6) is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t_i^* (Eq. 6) represents the ground truth box associated with a positive anchor. The classification loss L_{cls} (Eq. 6) is the log loss over two classes (object vs. not object). The regression loss L_{reg} is the $smooth_{L1}$ (Eq. 9) loss with four location parameters. The two terms are normalized by N the size of mini batch, in our experiment N=256 For classification. The specific loss function of classification is defined as:

$$L_{cls}(p_i, p_i^*) = -\log\left[p_i^* p_i + \left(1 - p_i^*\right)(1 - p_i)\right] \tag{7}$$

For bounding box regression, we adopt the parameterization of the 4 following coordinates:

$$\begin{aligned} t_x &= \frac{(x - x_a)}{w_a} \quad t_y = \frac{(y - y_a)}{h_a} \\ t_w &= \log\left(\frac{w}{w_a}\right) \quad t_h = \log\left(\frac{h}{h_a}\right) \\ t_x^* &= \frac{(x^* - x_a)}{w_a} \quad t_y^* = \frac{(y^* - y_a)}{h_a} \\ t_w^* &= \log\left(\frac{w^*}{w_a}\right) \quad t_h^* = \log\left(\frac{h^*}{h_a}\right) \end{aligned} \tag{8}$$

where x and y denote the box's center coordinates and w and h denote its width and height, respectively. Variables x, x_a, and x^* represent the predicted box, anchor box, and ground truth box, respectively (likewise for y, w, h), which can be considered bounding box regression from an anchor box to a nearby ground truth box. The specific loss function of bounding box regression is defined as :

$$\begin{aligned} L_{reg}(t_i, t_i^*) &= \text{smooth}_{L1}(t_i - t_i^*) \\ &= \begin{cases} 0.5(t_i - t_i^*)^2, & if\,|t_i - t_i^*| < 1 \\ |t_i - t_i^*| - 0.5, & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

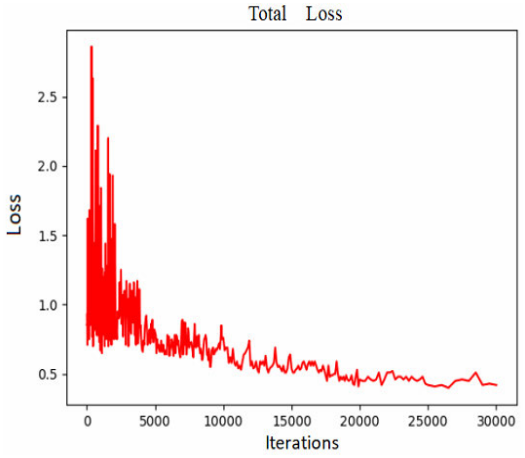


FIGURE 7. The total loss in the proposed method. After 20k iterations, the total loss (relative region proposal network loss and contextual information network loss) tends to be stable.

The contextual information network loss function defined as:

$$L_{CIN}(p_i, c_i) = \frac{1}{N}\sum_i L_{cls}\left(p_i, p_i^*\right) + \frac{1}{N}\sum_i p_i^* \bullet \mathrm{L_{reg}}\left(c_i, \mathrm{c}_i^*\right) \quad (10)$$

where, p_i is the predicted class scores that include the contextual information, p_i^* is the ground truth class label, c_i is the predicted coordinates including the contextual information, and c_i^* is the ground truth coordinates label. The loss functions are the same as the RRPN loss function and are cross-entropy and $smooth_{L1}$ (Eq.9).

IV. TRAINING

A. DETAILS

The proposed experiments were conducted based on the MCFN. Ren *et al.* [23] experimentally proved that choosing the pretrained ResNet model enables better performance than other pretrained models, such as VGG and Inception. Accordingly, we chose the ResNet-101 model as the backbone network. The model was initialized by the ImageNet classification model and fine-tuned on the remote sensing object detection dataset. We randomly split the samples into 80% for training and 20% for testing.

In all experiments, we trained and tested the proposed method model based on the TensorFlow deep learning framework. We resized the images to 800 × 800 pixels, and applied stochastic gradient descent for 30k iterations to train our model. The learning rate was 0.001 and decreased to 0.0001 after 20k iterations. We adopted only one scale anchor for one scale predicted with areas of 32 × 32 pixels, 64 × 64 pixels, 128 × 128 pixels, and 256 × 256 pixels, and three ratios: 1:1, 1:2 and 2:1, which performed well on the dataset.

The evaluation used the average precision of each object instance and the mean average precision with an IoU threshold of 0.3 and an RIoU threshold of 0.7. To reduce redundancy, NMS was adopted on the proposal regions based on their box-classification scores. The IoU threshold was fixed for NMS at 0.7. From the loss pict. in Fig. 7, we can see that after 20k iterations, the total loss(RRPN loss and CIN loss) tended to be stable.

B. COMPUTATIONAL TIME ANALYSIS

The proposed method was implemented on a PC with a Core i7-6700 3.30 GHZ CPU, four Nvidia 2080Ti GPUs and 64 GB of RAM. The detection time of the proposed method is approximately 25-30 FPS. The time is basically the same as the detection time of the mainstream algorithm, which can meet real-time requirements and has high practical value.

V. EXPERIMENTAL RESULTS

In this section, a series of experiments were performed to verify the effects of the proposed method. We performed a series of experiments on the remote sensing image datasets. Our method achieved state-of-the-art performance: **96.6**% for aircraft and **95.6**% for cars.

We firstly conducted evaluations on the remote sensing object detection dataset to compare the performance with other state-of-the-art object detection methods. We also evaluated the robustness of the proposed method using the COCO dataset. We used mean average precision (mAP), average precision (AP) and recall (R) defined as follows:

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$AP = \frac{1}{N}\sum_{i=1}^{N} P \quad (12)$$

$$mAP = \frac{1}{M}\sum_{i=1}^{M} P \quad (13)$$

$$R = \frac{TP}{TP + FN} \quad (14)$$

where N is the number of the pictures in one class and M is the number of class as indicators to compare with other methods. TP is the number of true positive samples which means the positive samples be predicted positively. FP is the number of false postive samples which means the negative samples be predicted positively. FN is the number of false negative samples which means positive samples predicted negtively. Higher map and recall mean better network performance.

A. EVOLUTION USING THE REMOTE SENSING OBJECT DETECTION DATASET

In Fig. 8 the results of the detection of small objects show that the tranditional algorithms (a), (b), (f), (g) result in many instances of misdetection, and the proposed algorithm can detect more small objects than other deep learning methods. The proposed dual pyramid network concatenates the shallow feature maps and the deep feature maps in four scales that can extract the small object features adequately. Other small object detection methods, such as the FPN and modified

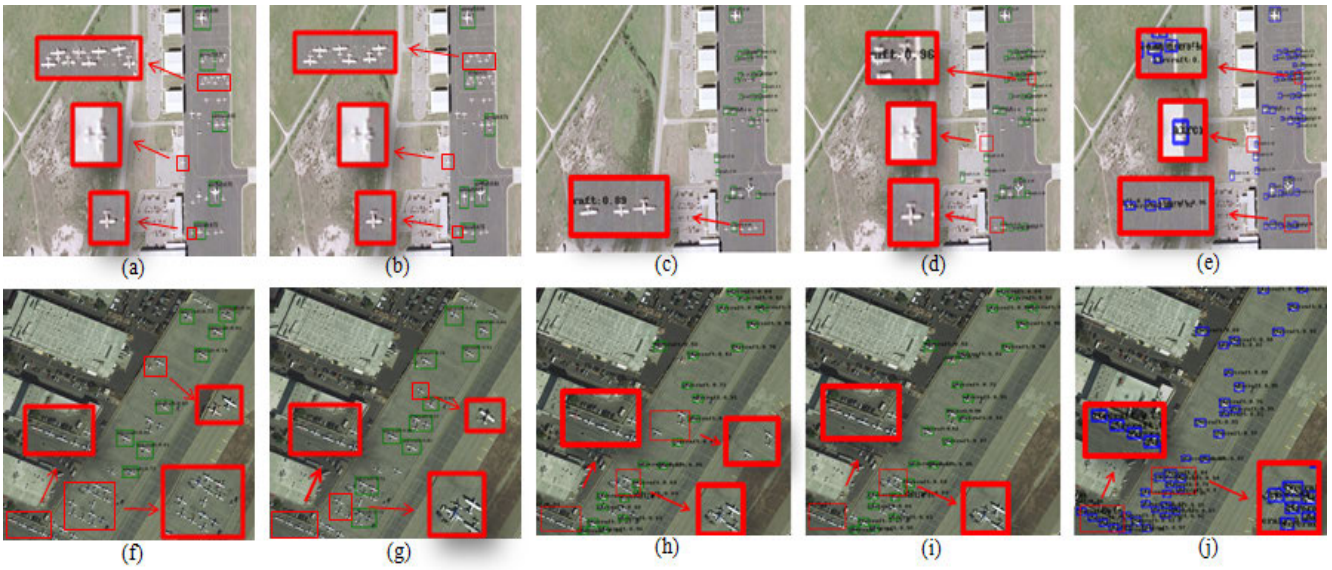


FIGURE 8. The qualitative comparison in the small object detection between the proposed method and other state-of-the-art methods: (a) and (f) are the HOG, (b) and (g) are the LBP (c) and (h) are the FPN, (d) and (i) are the modified faster-RCNN. The red box shows the missing detection, (e) and (j) are the result of proposed method. It is obvious that proposed method had better results in detecting small objects.

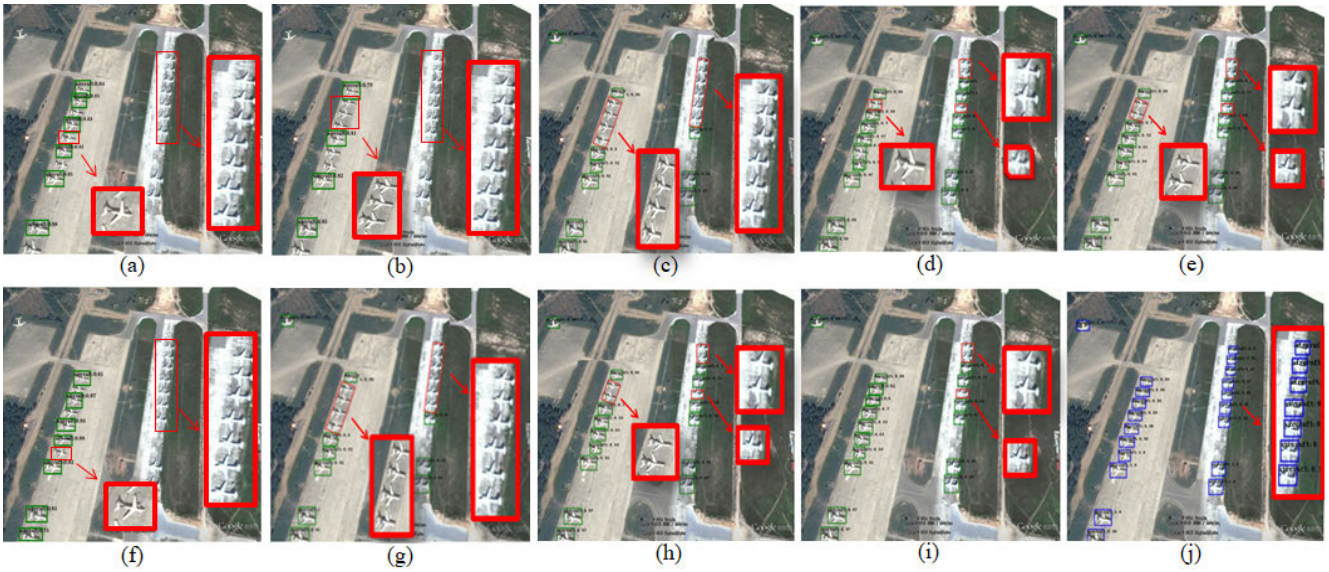


FIGURE 9. A visualization of the comparison results in a complex background situation with other methods: (a) LBP, (b) HOG, (c) faster-RCNN, (d) FPN, (e) YOLOv3, (f) SIFT, (g) Coupled-CNN, (h) SSD, (i) modified faster-RCNN, and the red box shows the missed detection. The proposed method (j) not only interpolate contextual information but also changes the traditional anchor chosen mechanism in training to have better results.

faster-RCNN, only add the deep convolutional layer feature maps to the shallow feature map and do not have enough features. Thus, the proposed method had better results in detecting small objects.

Figure 9 shows the results in a complex scene in which the object has the same color and texture as the background. Traditional methods can not accurately detect objects against complex backgrounds. Other methods, such as the faster-RCNN, SSD, FPN and YOLOv3, do not make the best use of contextual information, and therefore cannot detect objects in complex backgrounds. From (c), (d), (e) and (h) in Fig. 9, the modified faster-RCNN method concatenates the contextual information and the object features in (i), thus it performs better than the previous methods. The proposed method not only interpolate contextual information but also changes the traditional anchor chosen mechanism in training so it has the better results.

There are some objects that are only partially exposed. Because the proposed method can learn the local information in the RRPN and the contextual information in the context

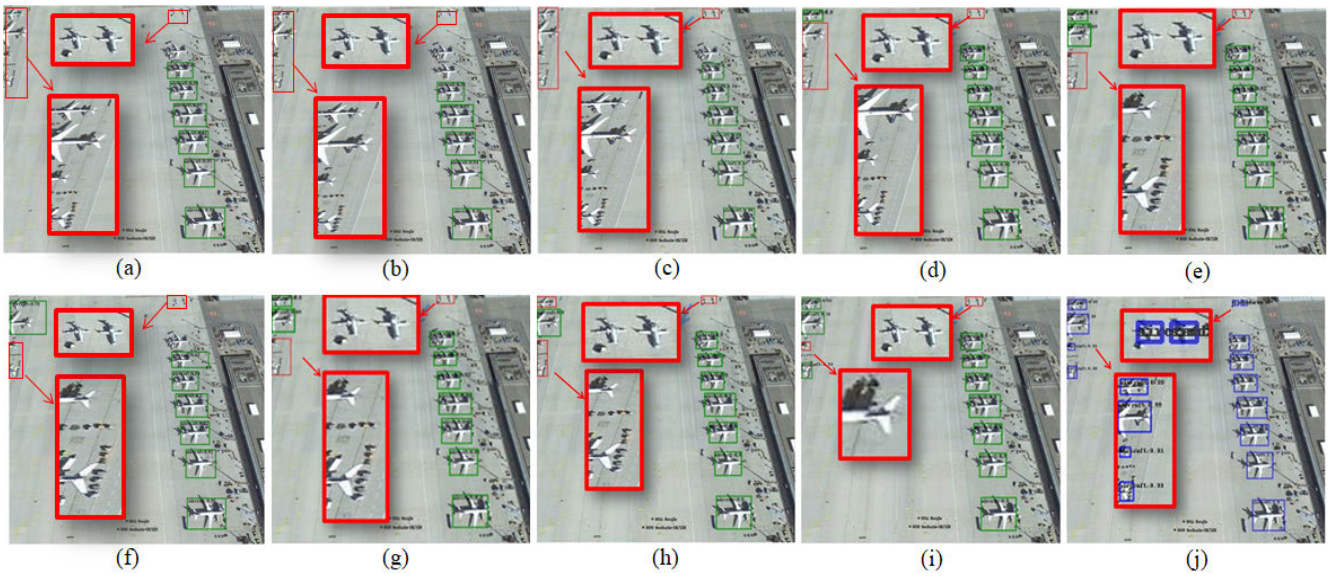


FIGURE 10. The comparison results of occlusion situation: (a) LBP, (b) HOG, (c) faster-RCNN, (d) FPN, (e) YOLOv3, (f) SIFT, (g) Coupled-CNN, (h) SSD, (i) modified faster-RCNN. The red box shows the miss detection, and (j) is the proposed method. it can easily detect when the object is occluded which is obviously superior to other state-of-the-art methods.

information network, the method can detect the partially exposed objects. Figure 10 shows ten figures. The tranditional method is useless when the object is partly exposed. The FPN algorithm (d) in Fig. 10 has a certain effect on detecting individual small objects. However, when the distance between the objects is close, the network does not sufficiently learn the small objects and background information, resulting in these objects not being easily detected. The one-stage methods SSD and YOLOv3 do not perform well in the occlusion situation, as shown in (h) and (e) in Fig. 10. The proposed method can sufficient learn the small object information and background information, and it can easily detect when the object is occluded, which makes it superior to other state-of-the-art methods shown in (j).

To further verify the effect and robustness of the proposed method, we applied the proposed method to detect the cars. The cars in the remote sensing dataset are more difficult to detect than the aircraft; they have a smaller scale and fewer features, and the scene is more complex. As shown in Fig. 11, from the red box in (a) and (b) the cars are occluded in the trees and only slightly exposed, experimental results indicate that the proposed method can detect the cars accurately, especially in complex scenes.

The quantitative evaluation results with the remote sensing object detection dataset are listed in Table 2 and Table 3. We first compared the performance with that of the traditional object detection methods, such as HOG and LBP. The traditional methods did not perform well because the detection method based on the sliding window has high computational redundancy and time complexity. It is difficult to address occlusion problems. Moreover, because the HOG and LBP do not have scale invariance, it is difficult to detect small objects. Although the SIFT have scale invariance it is useless for the object in the complex background. In summary, the accuracy of traditional methods in small target detection is much lower than that of the proposed method as shown in Table 2.

TABLE 2. Compares the proposed method to the traditional methods.

Methods	Aircraft AP (%)	Car AP (%)
LBP	58.13	56.18
HOG	64.04	62.12
SIFT	66.23	64.54
Proposed method	96.64	95.61

We compared the proposed approach with seven common object detection state-of-the-art methods and five remote sensing object detection state-of-the-art methods. As shown in Table 3, the deep learning method is obviously better in accuracy than the traditional methods like HOG and LBP. because of the superiority of the proposed method in small object detection and the strong ability to resist complex scene and occlusion, the proposed method has higher accuracy and robustness when detecting aircraft and cars. From the mAP and recall shown in Table 3 We can observe that in terms of AP over all the aircraft and the car categories, the proposed approach outperforms the previously most accuracy USB-BBR [20] method and AARMN [8] method by 2% and 1% respectively.

B. ABLATION STUDY IN EACH COMPONENT

We did an ablation study to validate the specific contribution of each component of the proposed method for detection. In each experiment, we omitted one part of our method and retained the remaining parts. The AP are listed in Table 4. We can see that the change in network structure increased by 2.9%. because of the DPFN fused more features of the object.

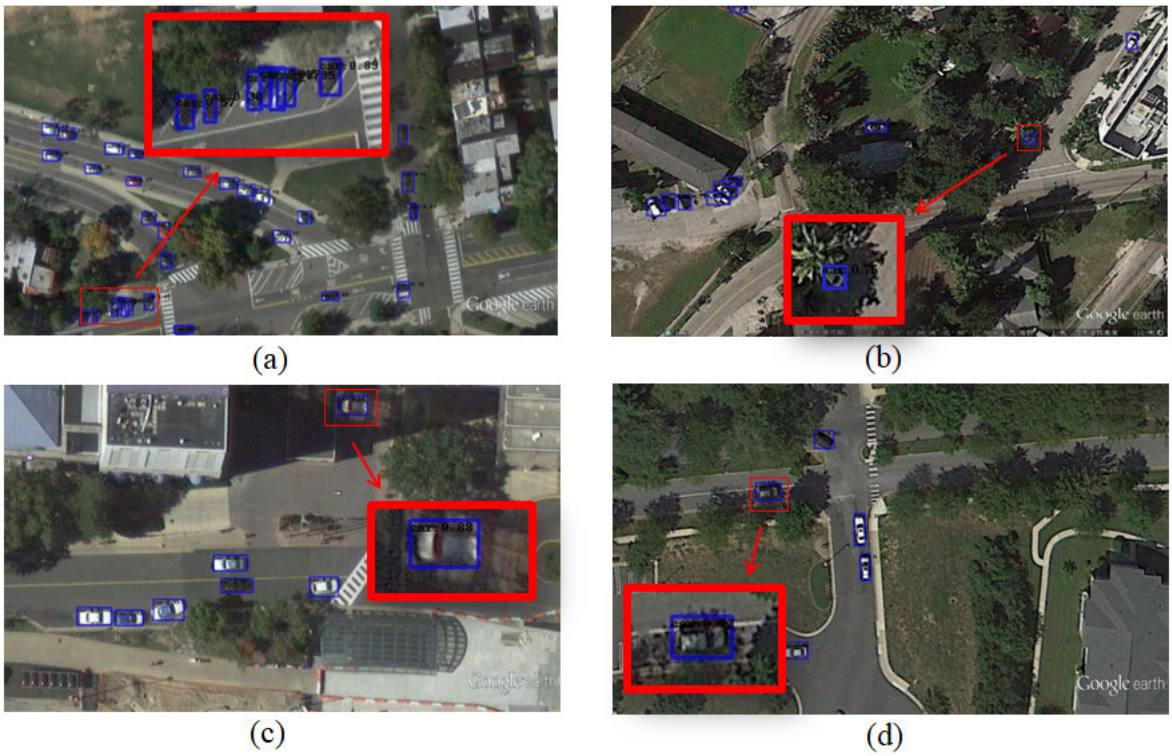


FIGURE 11. The visualization results for detecting cars in the remote sensing image. The proposed method can detect the cars under the trees and only slightly exposed in the complex scene.

FIGURE 12. The result of the proposed method in an extremely complex scene. The proposed method can detect the people in the mountain pass.

Second, we only used the proposed RRPN to take the place of the traditional RPN, the result improved by 3.9%. This result indicates that the small objects that have relatively low IoU are useful for training. Finally, we interposed only the contextual information to the proposals, and the AP improved by 3.3%. The contextual information is effective for the network to detect the object. We can see that when all terms were applied, the proposed method can improve 3.3% of the average precision.

For all results, it can be easily illustrated: due to the use of the DPFN the proposed method obtains more spatial structural information about small objects, so that more semantic information can be obtained to enhance the feature representation in small objects.The RRPN make more small

TABLE 3. Compares the proposed method to the deeplearning methods. The symbol "–" represents that the method does not provide relevant results.

Methods	Aircraft AP (%)	Aircraft R (%)	Car AP (%)	Car R (%)
AARMN[8]	94.27	94.18	94.65	92.16
NEOON[9]	94.49	-	93.22	-
NMMDPN[10]	-	-	91.36	90.56
faster-RCNN[13]	80.15	78.91	82.52	81.21
FPN[16]	91.22	90.18	85.12	86.33
Coupled CNN[19]	89.13	88.26	-	-
USB-BBR[20]	94.69	93.09	-	-
MFast-RCNN[23]	84.5	85.1	80.1	80.5
WSL[28]	61.94	-	57.74	-
R-FCN[33]	84.3	95.26	89.3	88.2
D-RCNN[38]	55.6	53.78	-	-
YOLOv3[46]	84.92	82.81	71.73	70.93
Proposed method	96.64	95.92	95.61	93.05

TABLE 4. The result of ablation study about the proposed framework in the aircraft detection.

Methods	DPFN	RRPN	Contextual Information	AP (%)
MCFN-1	no	no	no	91.22
MCFN-2	yes	no	no	94.13
MCFN-3	no	yes	no	95.30
MCFN-4	no	no	yes	94.5
MCFN	yes	yes	yes	96.5

TABLE 5. Compares the proposed method in the COCO dataset to the state-of-the-art methods.

Methods	COCO mAP-50 (%)
R-FCN [34]	51.9
FPN [16]	59.1
SSD [15]	45.4
YOLOv3 [48]	57.9
Proposed method	60.9

samples to use in the training,so it can make the model learn more local information and small objects information; the CIN containing the object and the background relationship features, the proposed method obtains a discriminative feature representation ability to effectively recognize objects in spite of the diversity and complexity of object appearance.

C. EVOLUTION USING THE COCO DATASET

Finally, we evaluated the performance of the proposed method on the COCO dataset. As shown in Table 5, because the COCO dataset has more classes and considerably smaller objects, many methods do not perform well. Compared to other methods, the proposed method was 3.8% better in detecting small objects. Thus, the proposed algorithm is effective in detecting small objects.

We also use more complex images to verify the robustness of the proposed method. As shown in Figure 12, the resulution of the image is 6000 × 4000. the size of people is about one thousandth of the image. the results of the detection shows the proposed algorithm have good performance under the complex scene which cannot be distinguished by human eyes.

FIGURE 13. The serried objects (in the red rectangle) which the distance of the objects is so near even many parts of the objects are linked together cannot be detected accurately.

VI. DISCUSSION

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, as shown in the red rectangle of Fig. 13, in the serried scene which the distance of the objects is so near even many parts of the objects are linked together the proposed method can not accuratly detect them.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed the MCFN method to address the small object detection problem in optical remote sensing images. We designed a DPFN with a series of encoding and decoding operations to better extract the features of the object, which is critical for detecting small remote sensing objects. At the same time, we used RRPN to

maximally extract the small object features and local features. Furthermore, we leveraged the contextual information enclosing an object proposal to further improve small object detection performance, especially in occlusion and complex scene situations. Finally, we conducted a wide range of experiments and provided a comprehensive analysis of the performance of MCFN on the task of small object detection in optical remote sensing images. Our future work will focus on improving the accuracy in the situation of many objects are serried.

Compliance With Ethical Standards: **Conflict of interest** All authors declare that they have no conflict of interest.

REFERENCES

[1] 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.

[2] L. Zhang and Y. Zhang, "Airport detection and aircraft recognition based on two-layer saliency model in high spatial resolution remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1511–1524, Apr. 2017.

[3] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.

[4] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[5] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, "Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618–644, 2014.

[6] R. Strickland and H. I. Hahn, "Wavelet transform methods for object detection and recovery," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 724–735, May 1997.

[7] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 281–294, Apr. 2018.

[8] H. Q. Qiu, H. L. Li, Q. B. Wu, and F. M. Meng, "A2RMNet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images," *Remote Sens.*, vol. 11, p. 1594, Jan. 2019. doi: 10.3390/rs11131594.

[9] W. Y. Xie, H. N. Qin, Y. S. Li, Z. Wang, and J. Lei, "A novel effectively optimized one-stage network for object detection in remote sensing imagery," *Remote Sens.*, vol. 11, no. 11, p. 1376, 2019. doi: 10.3390/rs11111376.

[10] W. Ma, Q. Guo, Y. Wu, W. Zhao, X. Zhang, and L. Jiao, "A novel multi-model decision fusion network for object detection in remote sensing images," *Remote Sens.*, vol. 11, p. 737, Apr. 2019. doi: 10.3390/rs11070737.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[12] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, 1995.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural. Inf. Process. Syst.*, Montreal, QC, Canada. Cambridge, MA, USA: MIT Press, Dec. 2015, pp. 91–99.

[14] 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.*, Boston, MA, USA, Jun. 2015, pp. 779–788.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, C. Fu, and A. C. Berg, "SSD: Single-shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*. [Online]. Available: https://arxiv.org/abs/1612.03144

[17] Z.-Y. Shen, Z. Liu, J. G. Li, Y. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1919–1927.

[18] 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.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[19] 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.

[20] Y. Long, Y. Gong, Z. F. 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.

[21] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: https://arxiv.org/abs/1512.03385

[23] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Appl. Sci.*, vol. 8, no. 5, p. 813, 2018.

[24] 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.

[25] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2015, *arXiv:1405.0312*. [Online]. Available: https://arxiv.org/abs/1405.0312v3

[26] C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. *Red Hook*. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Sep. 2015, pp. 1–14.

[30] G. Huang, Z. Liu, L. van der Maaten, and K.-Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*. [Online]. Available: https://arxiv.org/abs/1608.06993

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," 2014, *arXiv:1406.4729*. [Online]. Available: https://arxiv.org/abs/1406.4729

[32] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1440–1448.

[33] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*. [Online]. Available: https://arxiv.org/abs/1605.06409

[34] 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.

[35] 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.

[36] H. Wu, H. Zhang, J. Zhang, and F. Xu, "Typical target detection in satellite images based on convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 2956–2961.

[37] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

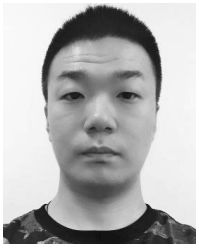
[38] W. Zhang, S. H. Wang, S. Thachan, J. Z. Chen, and Y. T. Qian, "Deconv R-CNN for small object detection on remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2483–2486.

[39] I. Ševo and A. Avramović, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, May 2016.

[40] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4895–4909, Oct. 2015.

[41] A. B. Salberg, "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2015, pp. 1893–1896.
[42] Q. Jiang, L. Cao, M. Cheng, C. Wang, and J. Li, "Deep neural networks-based vehicle detection in satellite images," in *Proc. Int. Symp. Bioelectron. Bioinf.*, Oct. 2015, pp. 184–187.
[43] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3735–3739.
[44] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 2980–2988.
[45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[46] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

JING LIU was born in Hebei, China, in 1985. He received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2016. He is currently a Lecturer with Hebei Normal University. His main research interests include augmented reality, medical image analysis, and computer vision.



SHUOJIN YANG received the B.Sc. degree from the School of Hebei Normal University, in 2015, where he is currently pursuing the M.Sc. degree, under the supervision of Prof. W. Guo. His research interests include computer vision and deep learning, specifically for object detection and augmented reality.

LIANG TIAN was born in Hebei, China, in 1981. He is currently pursuing the Ph.D. degree with the Hebei Key Laboratory of Computational Mathematics and Application, College of Mathematics and Information Science, Hebei Normal University. His research interests include computer vision, image processing, deep learning, and augmented reality.

WEI GUO is currently a Professor and a Ph.D. Supervisor with the Hebei Key Laboratory of Computational Mathematics and Application, College of Mathematics and Information Science, Hebei Normal University. Her research interests include wavelet analysis, image processing, and augmented reality.

BINGYIN ZHOU received the Ph.D. degree from Hebei Normal University, Shijiazhuang, China, in 2012, where he is currently an Associate Professor with the College of Mathematics and Information Sciences. His main research interests include wavelet analysis, tensor analysis, high-dimensional data processing, and image processing.

JIANQING JIA received the B.Sc. and M.Sc. degrees from the Hebei Key Laboratory of Computational Mathematics and Application, College of Mathematics and Information Science, Hebei Normal University, in 2015 and 2019, respectively. He is currently pursuing the Ph.D. degree with Syracuse University. His research interests include image processing, wavelet analysis, and augmented reality.

HAIBIN LING received the B.S. degree in mathematics and the M.S. degree in computer science from Peking University, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from the University of Maryland, College Park, in 2006. From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia. From 2006 to 2007, he was a Postdoctoral Scientist with the University of California, Los Angeles. He joined Siemens Corporate Research as a Research Scientist. Since 2008, he has been with Temple University, where he is currently a Professor. His research interests include computer vision, augmented reality, medical image analysis, and human–computer interaction. He served or will serve as an Area Chair for CVPR 2014, 2016, 2019, and 2020. He serves as an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition* (PR), and *Computer Vision and Image Understanding* (CVIU).