

Received December 20, 2021, accepted January 4, 2022, date of publication January 7, 2022, date of current version January 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3141059

# Multi-Size Object Detection in Large Scene Remote Sensing Images Under Dual Attention Mechanism

JINKANG WANG<sup>ID</sup>, (Student Member, IEEE), XIAOHUI HE, (Senior Member, IEEE),  
SHAOFAMING<sup>ID</sup>, (Senior Member, IEEE), GUANLIN LU<sup>ID</sup>, (Student Member, IEEE),  
QUNYAN JIANG, (Student Member, IEEE), AND RUIZHE HU, (Student Member, IEEE)

Department of Mechanical Engineering, College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Xiaohui He (gcbhxh@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671470, and in part by the Key Research and Development Program of China under Grant 2016YFC0802900.

**ABSTRACT** The remote sensing images in large scenes have a complex background, and the types, sizes, and postures of the targets are different, making object detection in remote sensing images difficult. To solve this problem, an end-to-end multi-size object detection method based on a dual attention mechanism is proposed in this paper. First, the MobileNets backbone network is used to extract multi-layer features of remote sensing images as the input of MFCA, a multi-size feature concentration attention module. MFCA employs an attention mechanism to suppress noise, enhance effective feature reuse, and improve the adaptability of the network to multi-size target features through multi-layer convolution operation. Then, TSDFD (two-stage deep feature fusion module) deeply fuses the feature maps output by MFCA to maximize the correlation between the feature sets and especially improve the feature expression of small targets. Next, the GLCNet (global-local context network) and the SSA (significant simple attention module) distinguish the fused features and screen out useful channel information, which makes the detected features more representative. Finally, the loss function is improved to truly reflect the difference between the candidate frames and the real frames, enhancing the network's ability to predict complex samples. The performance of our proposed method is compared with other advanced algorithms on NWPU VHR-10, DOTA, RSOD open datasets. Experimental results show that our proposed method achieves the best AP (average precision) and mAP (mean average precision), indicating that the method can accurately detect multi-type, multi-size, and multi-posture targets with high adaptability.

**INDEX TERMS** Deep learning, dual attention mechanism, multi-size object detection, remote sensing images.

## I. INTRODUCTION

With the development of remote sensing satellites, unmanned aerial vehicles, and other technologies, the amount of remote sensing image data that can be obtained has exploded. Meanwhile, with the development of Earth observation technology, more and more attention has been paid to object detection in remote sensing images. Multi-size object detection in large-scene remote sensing images aims to automatically, accurately, and efficiently detect interesting targets at different scales and identify the categories and positions of targets

The associate editor coordinating the review of this manuscript and approving it for publication was Khoa Luu<sup>ID</sup>.

at the same time. It plays a vital role in many practical applications such as military operations, national defense construction, urban planning, and environmental monitoring.

Because of the special location of remote sensing observation platforms, the imaging characteristics of remote sensing images are different from those of natural scenes captured by digital cameras. Remote sensing images often contain a large number of complex ground background objects, but the types, scales, and postures of targets to be detected are often uncertain. Object detection of remote sensing images still has many problems and challenges. First of all, remote sensing images are mainly captured at a high altitude, so they cover a wide range of ground objects and complex image backgrounds.

This may lead to more false-positive targets and increase the false alarm rate; Secondly, due to the dense distribution and small size of targets, as well as the different types, scales, and postures of the targets to be detected, many positive samples will not be detected, increasing the false-negative rate; Besides, the imaging quality of remote sensing images is not as good as that captured by digital cameras, and the resolution is low. This further increases the difficulty of object detection. If the existing deep learning detection framework is directly applied to remote sensing images for object detection, the ideal detection accuracy cannot be achieved.

Considering the characteristics of object detection in remote sensing images, an object detection algorithm is proposed based on the dual attention mechanism of MobileNets, which is used for multi-type, multi-size, multi-posture small object detection in large-scene remote sensing images. The contributions of this paper are summarized as follows:

1. MFCA improves the network's feature expression ability without excessively increasing the number of model parameters. By adding an attention mechanism, every region on the feature map is considered in different degrees.
2. TSDFF is exploited to deeply fuse the feature maps output by MFCA, which maximizes the correlation between feature sets and especially improves the feature expression of small targets.
3. GLCNet and SSA are introduced to distinguish the fused features and screen out useful channel information, which makes the features to be detected more characteristic.
4. The experimental results indicate that the proposed network architecture has significantly improved the object detection AP (average precision) and mAP (mean average precision) on the datasets including NWPU VHR-10, DOTA, and RSOD.

This paper is organized as follows. Section II briefly reviews the related works on object detection in remote sensing images using deep learning methods. In Section III, the network model and techniques used in this paper are introduced. In Section IV, the experimental results are analyzed to verify the effectiveness of our method in improving the comprehensive performance of object detection in remote sensing images. Section V summarizes this paper and presents the future work.

## II. RELATED WORKS

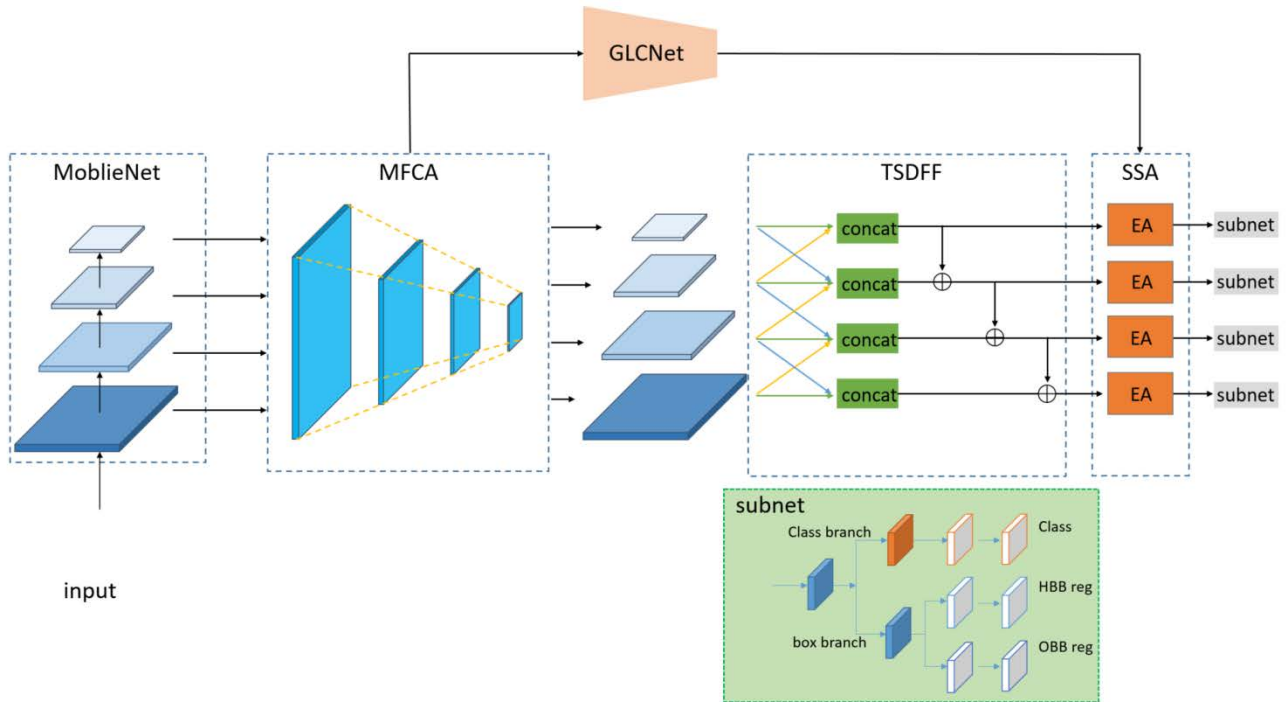
Remote sensing image object detection is a branch of object detection. In the field of object detection in remote sensing images, traditional object detection algorithms, such as circle frequency filtering method [1], edge extraction method [2], sparse representation method [3], and deep Boltzmann machine [4] mainly focus on the use of shallow and middle layer features. These algorithms have poor robustness and tedious detection process, and the detection results are easily affected by the quality of artificially designed fea-

tures. So, these algorithms are not suitable for multi-size object detection in remote sensing images. At present, deep learning is developing rapidly, and the convolutional neural network (CNN) has become a powerful tool for object detection because of its powerful feature extraction ability. The CNN-based method can fit a large number of complex data. Also, it can automatically learn the most useful features in images and fully extract the image information. Therefore, the deep convolutional neural network has advantages over traditional methods in object detection.

The current mainstream object detection algorithms can be divided into two categories, i.e., two-stage algorithms and single-stage algorithms. As for two-stage algorithms, the candidate frames are first generated by region proposal and then regressed and classified. Typical two-stage object detection algorithms include R-CNN [5], fast R-CNN [6], and Faster R-CNN [7]. Although the detection accuracy is high, these algorithms involve a large number of convolution operations, so the calculation cost is high and the speed cannot meet the real-time requirements. The one-stage algorithms use the whole picture as the input of the network and directly regresses the target frames and the category. The representative one-stage object detection algorithms include SSD [8] and YOLO [9], etc. Although the detection speed of these algorithms is fast, the detection results are not good because remote sensing images have low resolution and usually contain many very small targets.

To solve the problem that the traditional object detection algorithms cannot handle multi-size small targets, many researchers put forward improved methods and frameworks. For example, the Perceptual GAN algorithm proposed by Li *et al.* [10] reduced the representation gap between small targets and large targets and enhanced the feature expression of small targets. Liu and Huang [11] proposed the RFB structure, which reduced the down-sampling rate of the network and increased the receptive field by introducing dilated convolution [12]. Kisantal *et al.* [13] exploited an oversampling strategy to handle the samples containing small targets, which improved the detection accuracy of small targets. SNIP (Scale Normalization for Image Pyramids) [14] only selected the targets within a certain scale for learning in the training process, which reduced the influence of domain-shift. Image pyramid [15] scaled pictures at different degrees and extracted features of different scales from each layer of pictures, which achieved high detection accuracy but slow speed; Trident-Net [16] parallelized three different receptive field networks to better cover multi-size object distribution; The FPN [17] (Feature Pyramid Network) algorithm used the high resolution of low-level features and the rich semantic information of high-level features at the same time. It achieved a good prediction effect by fusing these different layers of features.

To shorten the information path and enhance the feature pyramid with low-level accurate positioning information, PANet [18] created a bottom-up path enhancement based on FPN. ThunderNet [19] simplified the FPN structure and introduced the pooling operation to integrate local and global



**FIGURE 1.** The pipelines of DADFFNet. DADFFNet consists of five blocks, including multi-size feature concentration attention module MFCA, two-stage depth feature fusion module TSDFF, global-local context network GLCNet, significant simple attention module SSA, and subnet module.

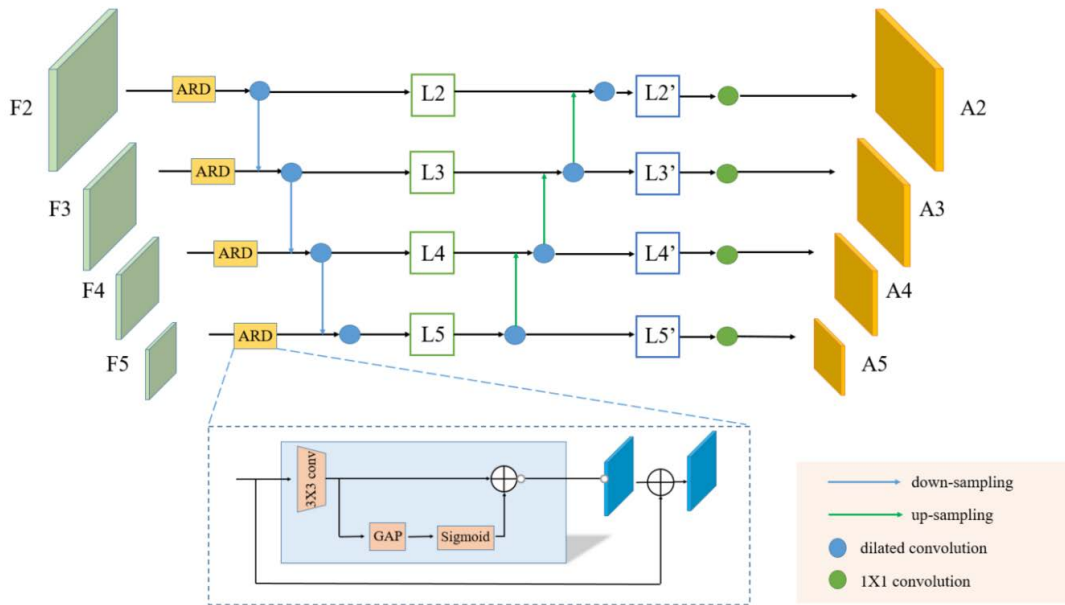
features to enhance the network feature expression ability. These improved methods have significantly improved the accuracy of small object detection.

For object detection in remote sensing images, R2CNN [20] pooled each text box proposed by RPN (region proposal network) with different pooled sizes ( $7 \times 7$ ,  $11 \times 3$ ,  $3 \times 11$ ). Meanwhile, it predicted text/non-text scores, axially aligned boxes, and inclined minimum area boxes simultaneously by using the characteristics of connections. Based on RPN, RoI Transformer [21] converted HRoI (horizontal region of interest) output into RRoI (rotating region of interest). In this way, the number of anchor points was not increased, and accurate RRoI can be obtained. CAD-Net [22] designed and integrated GCNet (global context network) and PLCNet (pyramid local context network) to extract context information at the global scene level and local target level, respectively. SCRDet [23] designed a sampling fusion network, which integrated multi-layer features into effective anchor sampling. Also, it designed a supervised multi-dimensional attention module MDA-Net, which improved the detection sensitivity of small targets; SCRDet++ [24] specified a novel InLD component to approximately decouple the features of different target categories into their respective channels. In this way, the features of objects were enhanced, and the features of background in the spatial domain were weakened. Gliding Vertex [25] proposed that a quadrilateral can be located by learning the offset of four points on a non-rotating rectangle to represent an object. Besides, to overcome the

defects of deep learning in satellite image object detection, an improved fine-grained object detection network structure called YOLT was proposed in [26], and a lot of data enhancement was made to solve the problem of detection invariance.

### III. OVERVIEW OF OUR METHOD

The end-to-end CNN model proposed in this paper is shown in Figure. 1. The network consists of five modules, including the feature extraction backbone network of MoblieNets [27], multi-size feature concentration attention module MFCA, two-stage depth feature fusion module TSDFF, global-local context network GLCNet, significant simple attention module SSA, and subnet module. In the network, the features of different scales extracted from MoblieNets are input to MFCA, which pays attention to various regions in the feature map of the original CNN to reduce the interference of the background and negative sample information. Especially in the shallow feature maps, MFCA can effectively focus on small target objects. Then, the output of MFCA is deeply fused by TSDFF to maximize the correlation between feature sets. Next, the fused features and two groups of memory features learned by GLCNet are input to SSA together. In SSA, different channels of feature maps are distinguished, and the useful channel information is screened out to make the detected features more representative. Finally, the feature maps of each scale are cascaded with the subnet for multi-branch classification and regression. Generally, our dual attention deep feature fusion network DADFFNet



**FIGURE 2.** The structure of MFCA. MFCA takes the output of the second, third, fourth, and fifth stages of the backbone network MblieNets as input. After ARD denoising, it performs a series of up-sampling and down-sampling operations to generate an attention map.

can effectively remove complex background noise, enhance the feature representation of different resolutions, especially small-sized targets, and greatly improve the detection accuracy of remote sensing images.

### A. MULTI-SIZE FEATURE CONCENTRATION ATTENTION MODULE

The visual attention mechanism is a unique vision signal processing mechanism of the human brain. By scanning the global image quickly, the human brain obtains the target areas that need to be focused on, which is commonly called attention focus. Then, more attention resources are put into these areas to obtain more detailed information of the targets while ignoring other useless information. This mechanism is formed by human beings in long-term evolution. It provides a means for human beings to quickly screen out high-value information from a large amount of information by using limited attention resources. The human visual attention mechanism greatly improves the efficiency and accuracy of visual information processing. Similar to the selective visual attention mechanism of human beings [28], the attention mechanism in deep learning selects the information that is more critical to the current task from a large amount of information, so as to maximize the usage of limited computing resources [29].

Following the idea of attention mechanism, MFCA improves the network’s feature expression ability without excessively increasing the number of model parameters. It mainly includes ARD (attention residual denoising) blocks, dilated convolution blocks, up-sampling operation, etc. The specific connection mode is shown in Figure. 2.

The module can be added to any convolutional neural network. The backbone network MblieNets consists of five stages, which are denoted as {C1, C2, C3, C4, and C5}. Considering the high spatial resolution of the feature map of the C1 stage and the network model parameters, as well as computational efficiency, starting from the C2 stage, the feature maps obtained through the MblieNets are input to MFCA, and they are defined as  $F_i$ .

#### 1) ATTENTION RESIDUAL DENOISING MODULE

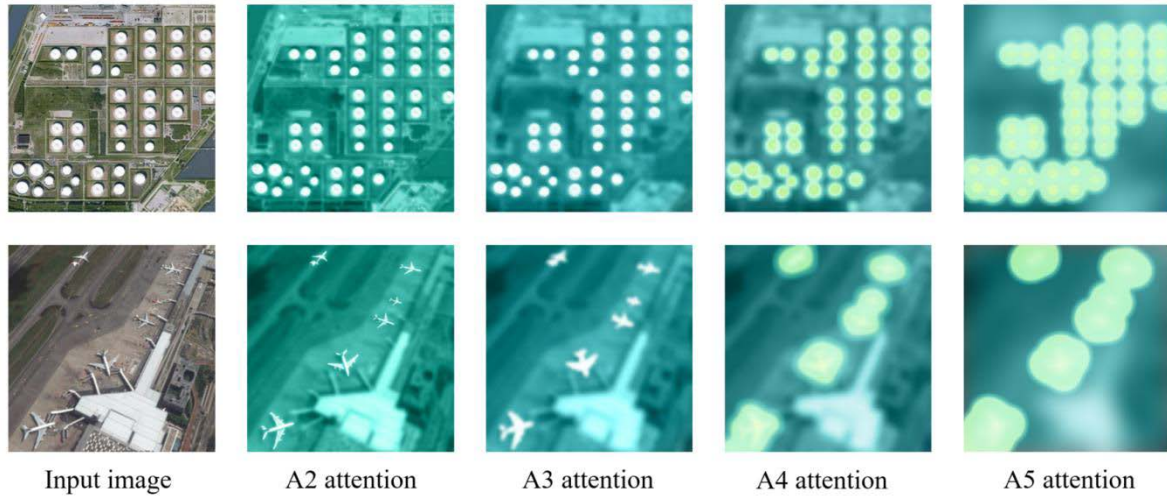
As for the feature map  $F_i \in R^{C \times W \times H}$ , the C, H, and W respectively denote the channel, height, and width of the feature map. Firstly, the ARD uses the attention mechanism to extract the focused attention targets and integrates the global spatial information through the GAP (global maximum pooling). Then, it processes the extracted features with the Sigmoid function and transforms them into the non-linear attention space. The output can be expressed as:

$$S_i = \phi(\theta(F_i)) \tag{1}$$

where attention map calculation  $\theta(\cdot)$  is achieved through GAP. Note that a separate  $\theta(\cdot)$  is implemented to calculate each scale-specific attention map.  $\theta$  is Sigmoid function, and  $S_i$  is the output attention map. The attention map is fused with the output of the original convolution block. The boot output  $F' \in R^{C \times W \times H}$  can be expressed as:

$$F'_i = S_i \otimes F_i \tag{2}$$

where  $i$  is the index of the feature map, and  $\otimes$  denotes element-wise multiplication. ARD performs element-wise multiplication when it is designed as a dot product attention



**FIGURE 3.** Illustration of our proposed multi-size attention responses at different feature scales. The brighter regions indicate higher attention responses. The proposed attention module can focus on informative regions at appropriate feature scales while ignoring irrelevant and noisy areas.

ratio; otherwise, it performs a summation. The nonlinear feature is increased through the  $1 \times 1$  convolutional layers, and then the attention maps are added to the module with the residual connection. The output of ARD is defined as  $Y_i$ , which can be expressed as follows:

$$Y_i = F'_i + F_i \quad (3)$$

## 2) CONVOLUTION OPERATION

As shown in Figure. 3, multi-size processing is performed on the output of ARD to generate more detailed attention information at different scales. In the bottom-up processing, each  $Y_i$  is processed by a corresponding  $3 \times 3$  dilated convolution that is denoted as  $D_i(\cdot)$ . The output of  $D_i(\cdot)$  can be expressed as  $X_i$ . Except for  $Y_2$ , the output of the previous level must be added to each other layers. This process can be expressed as:

$$X_i = \begin{cases} D_i(Y_i) & i = 2 \\ D_i(Y_i + conv(X_{i-1})) & 2 < i \leq 5 \end{cases} \quad (4)$$

Since each  $Y_i$  has a different spatial resolution, a  $3 \times 3$  convolution with a step size of 2 is applied to  $X_i$ , and the convolution result is then merged with  $Y_i$ . Next, the operation in the top-down processing stage is conducted similar to that in the bottom-up stage. Before the addition operation, a deconvolution with a step size of 2 is used to expand the space size. This process can be expressed as:

$$X'_i = \begin{cases} D'_i(Y_i + deconv(Y'_{i-1})) & 2 \leq i < 5 \\ D'_i(Y_i) & i = 5 \end{cases} \quad (5)$$

where  $D'_i(\cdot)$  is also a  $3 \times 3$  dilated convolution, and  $X'_i$  is the output of the top-down processing stage. The dilated convolution can expand the receiving domain, increase the receptive field of the feature map, and obtain richer context information while preserving the global information. Besides, the bottom-up and top-down connections maintain

the flow of attention information between multi-size feature maps. Finally, all the attention weights are generated by  $1 \times 1$  convolution. The attention weight of  $A_i$  is denoted by  $w_i$ . The final output can be expressed as:

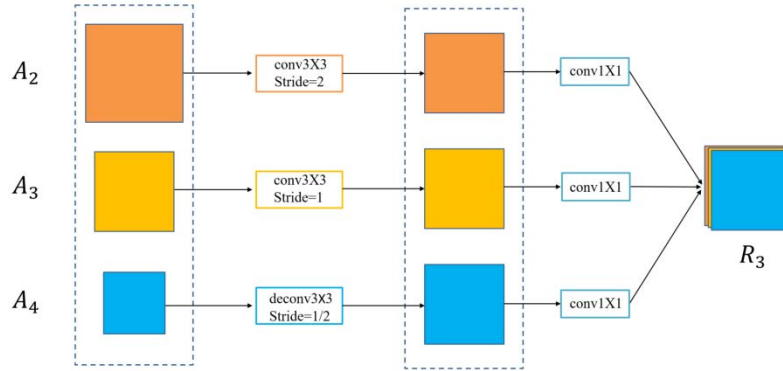
$$A_i = F_i \times w_i \quad (6)$$

As shown in Figure. 3, the proposed MFCA can treat different areas differently at each scale. This enhances the network's feature representation ability for certain important areas so that each area on the feature map has different degrees of importance. For example, smaller airplanes obtain stronger responses at the lower network layers, and the captured information has more detail. Meanwhile, the MFCA helps to weaken the information interference of background and negative sample targets, such as the terminal in the second sample image.

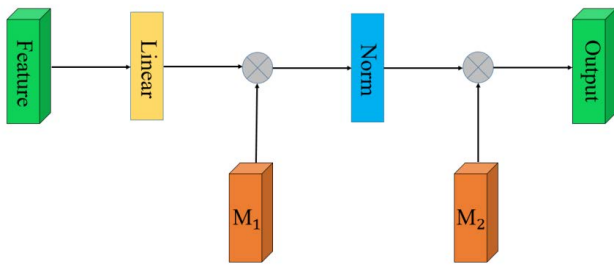
## B. TWO-STAGE DEEP FEATURE FUSION

In feature fusion, features are propagated in a top-down manner, and low-level features can be improved by using strong semantic information of high-level features. However, the features at the highest level lose information due to channel reduction. Since the semantic information has certain inconsistencies, directly fusing these features will reduce the ability of multi-size feature representation. Also, this strategy of fusing feature maps into a single vector may lose spatial relationships and details because multiple targets may appear in an image [30].

Information loss can be greatly reduced by fusing the extracted global context features in two different approaches [31]. The TSDFF module uses two different types of feature fusion, as shown in Figure. 1. Theoretically, the feature maps of adjacent scales have a greater correlation, so fusing these feature maps may reduce the inconsistency between feature targets. The first type of feature fusion



**FIGURE 4.** The first feature fusion process of STDF. After  $3 \times 3$  convolution operations with step sizes of 2, 1, and 1/2, respectively,  $A_2$ ,  $A_3$ , and  $A_4$  form feature maps of the same size. Then, they are fused through  $1 \times 1$  convolution layers.



**FIGURE 5.** The structure of the EA module.  $M_1$  and  $M_2$  are the two outputs of the GLCNet.

independently upsamples, adds patches, and downsamples adjacent features through the  $3 \times 3$  convolutional layer to achieve the same effect. Then, it splices the three adjacent features in dimensions. As shown in Figure. 1, the yellow and blue arrows respectively represent down-sampling and up-sampling, and the green arrow represents the addition of patches. For the convenience of explanation, three adjacent scale feature maps  $A_2$ ,  $A_3$ , and  $A_4$  are taken as examples, and the details of the first fusion process are illustrated in Figure. 4. The output after fusion is:

$$R_3 = W_2 \cdot A_2 \cdot \xi_3^2 + W_3 \cdot A_3 \cdot \xi_3^3 + W_4 \cdot A_4 \cdot \xi_3^4 \quad (7)$$

where  $R_3$  is the output of  $A_3$ .  $W_2$ ,  $W_3$ , and  $W_4$  are the parameter-sharing convolution kernels corresponding to the three feature maps of  $A_2$ ,  $A_3$ , and  $A_4$ . The strides are 2, 1, and 1/2, respectively.  $\xi_3^2$ ,  $\xi_3^3$ , and  $\xi_3^4$  are three spatial weights that respectively represent the importance of  $A_2$ ,  $A_3$ , and  $A_4$  relative to  $A_3$ . The weight generation process is as follows.

After the uniform scale operation, three  $1 \times 1$  convolution layers are used to generate the weight scalar, and they are denoted as  $\gamma_3^2$ ,  $\gamma_3^3$ , and  $\gamma_3^4$ . Taking  $\xi_3^2$  as an example,  $\xi_3^2(i, j)$  represents the spatial weight of  $A_2$  relative to  $A_3$  at point  $(i, j)$ , which can be expressed as:

$$\xi_3^2(i, j) = \frac{\exp(\gamma_3^2(i, j))}{\exp(\gamma_3^2(i, j)) + \exp(\gamma_3^3(i, j)) + \exp(\gamma_3^4(i, j))} \quad (8)$$

From equation (8), it can be seen that the sum of  $\xi_3^2(i, j)$ ,  $\xi_3^3(i, j)$ , and  $\xi_3^4(i, j)$  is 1, and their values are all between 0 and 1.

The first feature fusion can utilize the semantic information of feature maps of different scales better. It achieves higher performance by increasing the channel and further reduces the interference of background noise at the same time.  $1 \times 1$  convolutional layers are used to reduce the feature channels, where the huge semantic gaps between these features are not considered.

The second feature fusion first uses a parallel strategy to perform an element-wise add operation on the feature maps after the first feature fusion. Then, it combines two adjacent feature vectors into a complex vector. The add operation does not increase the dimensionality of the feature maps but increases the amount of information under each dimension, which obviously increases the perception of contextual information.

TSDFF performs a weighted combination on the foreground discrimination of remote sensing images and maximizes the correlation between the feature sets through the two feature fusions. Meanwhile, TSDFF enhances the semantic information of small targets, maximizes the difference between different classes, and further eliminates the influence of noise and complex background.

### C. GLOBAL-LOCAL CONTEXT NETWORK

Considering the correlation between the background and targets in remote sensing images, a global-local context network is designed, which can learn the global scene semantics and use it as a certain prior to better detect the targets in remote sensing images. GLCNet uses the learned correlation as a specific global-local context to compensate for the missing distinguishable target features. The learned correlation can be expressed as follows:

$$G(I) = \psi[\phi_G(A_i)] \quad (9)$$

where  $A_i$  represents the feature mapping from MFAC, and  $\phi_G(\cdot)$  is implemented by the CLSTM module [32] to extract

global features.  $\psi(\cdot)$  represents a pooling operation that compresses the spatial channels of the feature map into a vector, thereby suppressing the scale change problem. There are two sets of CLSTM modules in the network, and positive  $A_i$  and reverse  $A_i$  are respectively input to the modules. The two outputs are input to the two memory modules in SSA.

#### D. SIGNIFICANT SIMPLE ATTENTION MODULE

Self-attention [33] is significant to various visual tasks. Compared with convolution operation, self-attention can acquire more long-range dependency, thereby learning the features that incorporate global information. However, the self-attention mechanism has several obvious defects. First, the large amount of calculation results in a certain amount of calculation redundancy. Also, the self-attention mechanism only uses the information in its own samples but ignores the potential meaningful connection between different samples. To alleviate these problems, external attention [34] is exploited to easily achieve linear complexity by controlling the size of the memory unit. Meanwhile, the useful information of the fused feature map is further screened out so that the features to be detected are more representative. SSA uses four EA modules as attention modules for extracting effective information from the input.

As shown in Figure. 5, external attention can simplify the time complexity of self-attention through two learnable external memory units. Also, the two external memory units are shared for the entire data, so the correlation between different samples can also be implicitly considered. The two units are linear layers, and they can be directly optimized end to end. In the actual operation process, the outputs provided by GLCNet are taken as the two different memory modules that are called  $M_1$  and  $M_2$ . The former stores the key and the latter stores the value. The calculation is as follows:

$$E = \text{Norm}(F_{in}M_1^T) \quad (10)$$

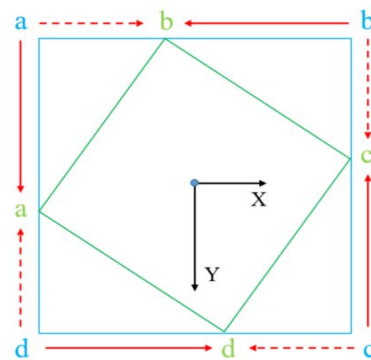
$$F_{out} = EM_2 \quad (11)$$

where  $F_{in}$  and  $F_{out}$  respectively represent the feature maps of input and output;  $\text{Norm}$  represents the normalization operation;  $E$  represents the transition state after normalization operation.

#### E. LOSS FUNCTION DESIGN

The subnet structure includes the classification branch and box branch, and they are respectively responsible for anchor label prediction and location regression.

Due to the existence of multi-posture targets in remote sensing images, the existing area-based rotating object detection methods describe the rotating bounding box with five parameters, including center point coordinate, width, height, and rotation angle, and these methods use smooth L1 as the loss function. However, there are two problems in this method, i.e., the loss discontinuity caused by angle parameters and the influence of different parameter units on network performance. To handle these problems, the 8-parameter ver-



**FIGURE 6.** Regression principle of rotating bounding box. The actual regression process of the green real box is  $\{(a \rightarrow a), (b \rightarrow b), (c \rightarrow c), (d \rightarrow d)\}$ , but obviously the ideal regression process should be  $\{(a \rightarrow b), (b \rightarrow c), (c \rightarrow d), (d \rightarrow a)\}$ .

sion of rotation loss proposed by RSDet [35] is used in this study. It describes the position with four clockwise vertices of the rotation bounding box, suppressing the problem of different parameter units. Figure. 6 shows the regression process from the candidate box to the actual position.

The actual regression process consists of four steps:

- 1) move the four vertices of the prediction frame clockwise;
- 2) keep the vertex order of the prediction frame unchanged;
- 3) move the four vertices of the prediction frame counterclockwise;
- 4) take the minimum value of the above three cases.

The loss function used in this process is expressed as follows, where  $x_i$  and  $y_i$  respectively represent the coordinate offset of the  $i$ -th vertex of the prediction frame and the  $i$ -th vertex of the reference frame.

$$L_{mr} = \min \begin{cases} \sum_{i=0}^3 (|x_{(i+3)\%4} - x_i^*| + |y_{(i+3)\%4} - y_i^*|) \\ \sum_{i=0}^3 (|x_i - x_i^*| + |y_i - y_i^*|) \\ \sum_{i=0}^3 (|x_{(i+1)\%4} - x_i^*| + |y_{(i+1)\%4} - y_i^*|) \end{cases} \quad (12)$$

In the proposed algorithm of this paper, due to the addition of the position offset of the anchor box, the corresponding multi-task loss function should be changed during the end-to-end training. In addition to the basic classification loss and regression loss, it is also necessary to learn the position of the anchor. The complete loss function is expressed as follows:

$$L = \lambda L_{mr} + L_{cls} + L_{reg} \quad (13)$$

where  $L_{cls}$  and  $L_{reg}$  represent the classification loss and regression loss respectively, and  $\lambda$  is a constant.

## IV. EXPERIMENTAL RESULTS

### A. DATASET

Our proposed method is tested on three public datasets, i.e., NWPU VHR-10 [36], DOTA [37], and RSOD [38].

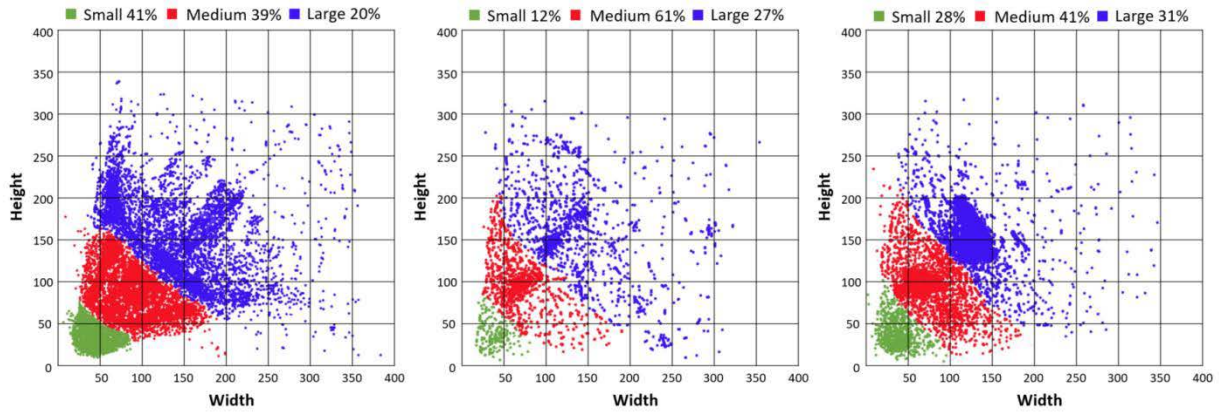


FIGURE 7. The instance size distribution of DOTA, NWPU VHR-10, and RSOD datasets.

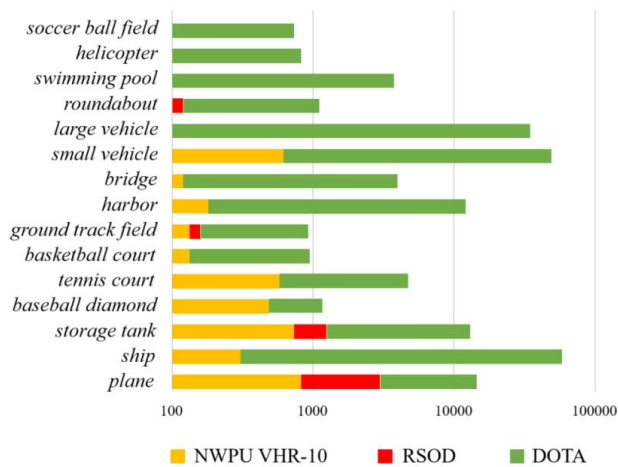


FIGURE 8. The instance number distribution of NWPU VHR-10, DOTA, and RSOD datasets.

Specifically, NWPU VHR-10 contains 800 high-resolution remote sensing image samples, of which 650 are positive samples and the remaining 150 are negative samples. The dataset contains annotations for 10 types of targets such as bridges, harbors, storage tanks, and ground track fields. The labeling method of this dataset is the traditional horizontal bounding box. The DOTA dataset has 2806 aerial images, with sizes ranging from 800 × 800 to 4000 × 4000 pixels. The dataset includes 188,282 targets of 15 categories such as airplanes, vehicles, ships, and football fields. It is the largest and most diverse remote sensing image dataset for object detection recently released. The labeling method of this dataset is the bounding box of any shape and direction determined by four points. RSOD includes four types of targets, including airplanes, playgrounds, overpasses, and oiltanks. In the dataset, there are 4993 aircrafts in 446 images, 191 playgrounds in 189 images, 180 overpasses in 176 images, and 1586 oiltanks in 165 images. The labeling method of this dataset is the traditional horizontal bounding box. The instance size and number distribution of the three

datasets are counted, and the results are shown in Figure 7 and Figure 8, respectively. The detection objects of these datasets are all artificially designed with obvious edge features and strong internal color consistency (e.g., ships, vehicles, and airplanes), while false objects often do not have these characteristics.

**B. EXPERIMENTAL SETTING AND PERFORMANCE EVALUATION INDEX**

Our proposed method was tested with PyTorch and TensorFlow 2.0. The test platform was equipped with Intel Core i7-6700U CPU @ 4.0 GHz, NVIDIA GeForce RTX 4000, and an 8 GB DDR3 memory, and the operating system was Windows 10 64-bit.

As for training parameter settings, the initial learning rate was set to 0.01, and it was attenuated to 1/10 of the original value every 50,000 iterations. The stochastic gradient descent method (SGD) of driving quantity was used to optimize the network. The momentum parameter was set 0.9; the weight attenuation regular term was set to 0.0005; the batch size was set to 32; the confidence threshold was set to 0.5, and the dropout was set to 0.5 to prevent over-fitting. The total training iterations of DOTA, NWPU VHR-10, and RSOD were respectively 200,000, 120,000, and 150,000.

In the experiment, AP and mAP were adopted as evaluation indicators to comprehensively evaluate the network. The ground truth was obtained through manual annotation. TP and FP represent the positive examples that are correctly detected and mistakenly detected respectively. FN represents the positive examples that are mistakenly detected as negative examples. Recall indicates the proportion of correct detection results in the actual targets, and the calculation is shown in equation (14). Precise indicates the accuracy of the detected results, and the calculation is shown in equation (15).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$



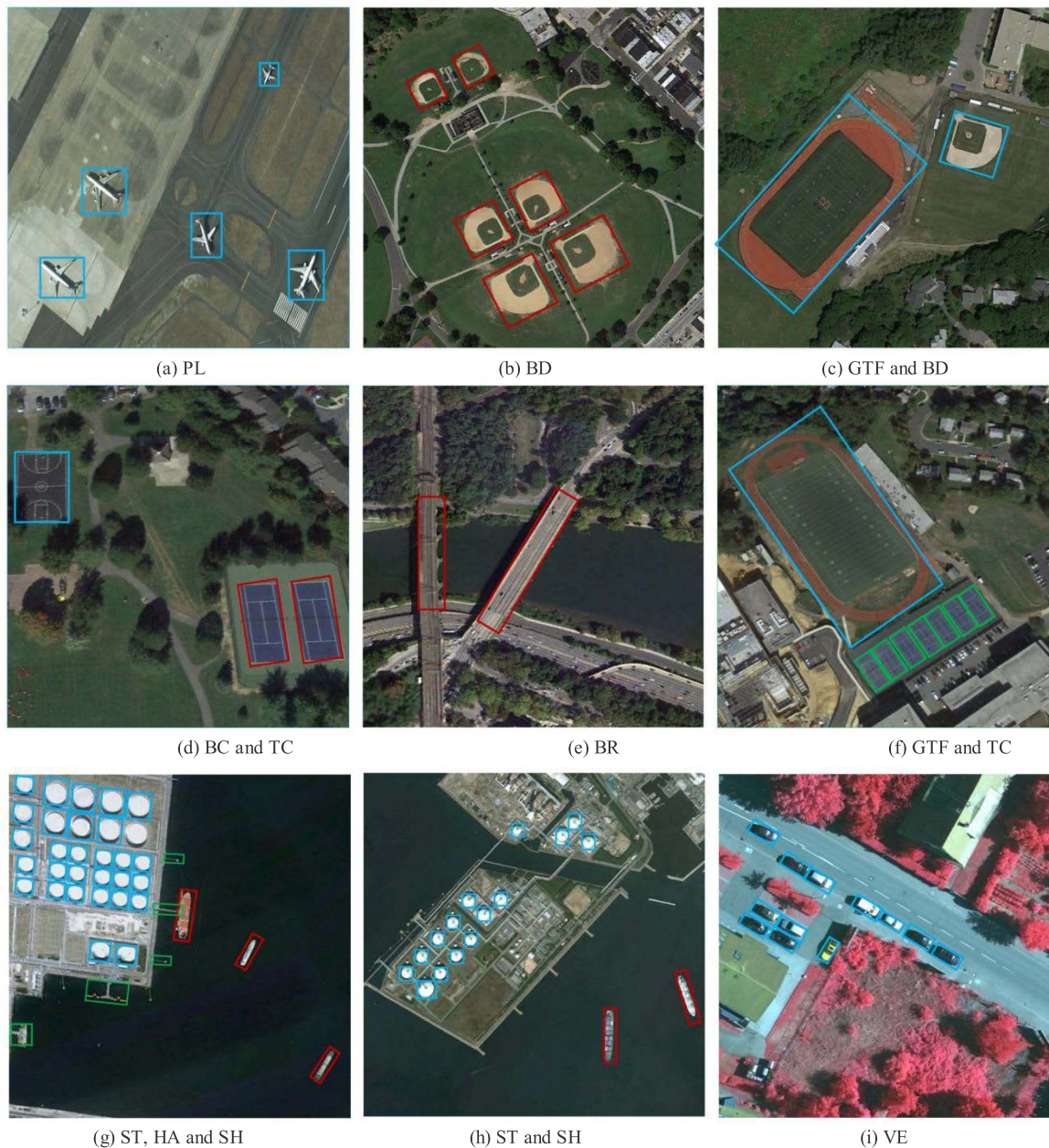


FIGURE 9. The results of our method for detecting different targets in the remote sensing images from the NWPU VHR-10 dataset.

TABLE 1. Experimental results of our proposed method and the state-of-the-art methods on the NWPU VHR-10 dataset.

method	mAP	PL	SH	ST	BD	TC	BC	GTF	HA	BR	VE
CAD-Net	70.83	87.8	81.6	56.1	59.8	63.5	79.8	76.2	70.6	47.3	69.0
R <sup>2</sup> CNN	60.64	80.26	67.82	65.29	59.31	76.45	72.82	62.48	60.52	39.61	66.73
SCRDet	72.51	89.06	73.25	85.68	81.35	<b>90.54</b>	86.99	69.11	65.65	50.92	64.38
SCRDet++	76.32	90.03	87.16	<b>87.89</b>	83.15	89.77	<b>88.01</b>	73.09	72.59	54.36	74.65
YOLT	68.49	88.43	71.95	80.67	74.17	89.88	79.54	73.01	61.58	49.53	72.92
Gliding Vertex	73.16	87.61	<b>88.19</b>	86.23	85.34	89.68	78.99	77.52	73.04	51.69	73.71
RoI Transformer	69.37	88.81	85.02	80.49	77.86	89.16	78.09	75.42	62.94	46.23	71.92
ours	<b>80.31</b>	<b>92.38</b>	87.05	86.37	<b>89.46</b>	90.33	87.94	<b>80.38</b>	<b>74.15</b>	<b>60.72</b>	<b>75.62</b>

In the evaluation results of deep learning, AP represents the average detection accuracy of a certain class of targets, while mAP represents the average accuracy of all classes of targets [39]. The calculation of these two indicators is shown

in equation 16 and equation 17, respectively.

$$AP = \int_0^1 P(R) dR = \sum_{k=0}^n P(k) R(k) \tag{16}$$

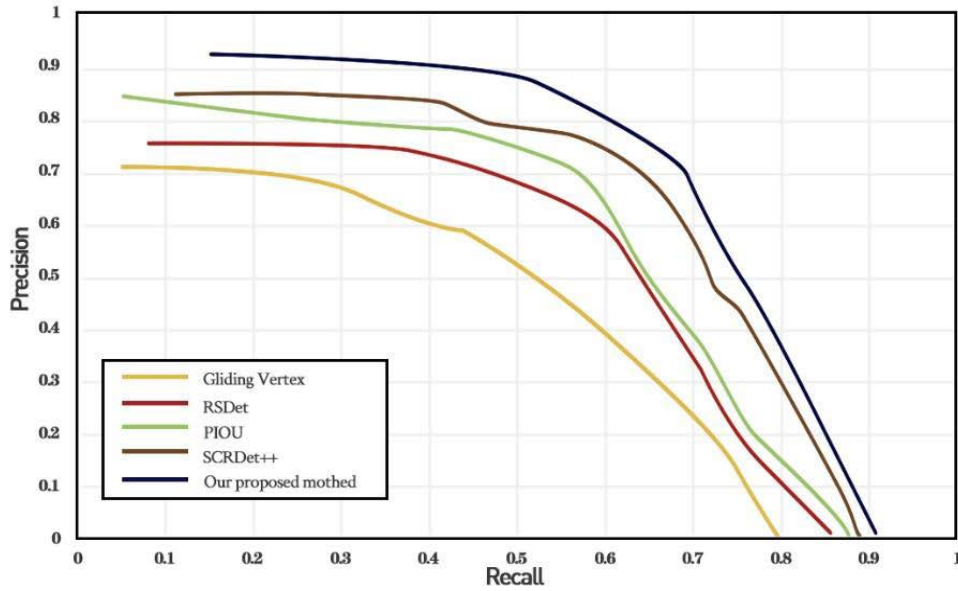


FIGURE 10. Comparison of P-R curves with the above several methods.

TABLE 2. Experimental results of our proposed method and the state-of-the-art methods on the DOTA dataset.

Target category	Gliding Vertex		RSDet		PIOU		SCRDet++		OURS	
	Prec.(%)	Rec.(%)	Prec.(%)	Rec.(%)	Prec.(%)	Rec.(%)	Prec.(%)	Rec.(%)	Prec.(%)	Rec.(%)
PL	<b>93.61</b>	83.96	88.07	91.61	80.85	94.84	81.40	83.13	90.27	<b>94.70</b>
BD	86.47	<b>92.77</b>	<b>90.23</b>	83.04	81.01	84.51	84.52	82.36	89.41	91.49
BR	73.05	83.28	83.18	82.23	82.66	<b>84.89</b>	74.48	81.10	<b>84.91</b>	84.66
GTF	80.56	82.85	85.57	90.14	<b>89.87</b>	86.04	86.84	80.95	88.25	<b>92.53</b>
SV	85.24	83.38	85.64	84.17	80.76	82.64	<b>91.06</b>	82.23	89.13	<b>90.47</b>
LV	89.73	80.97	86.43	85.24	92.74	82.38	87.57	80.61	<b>93.51</b>	<b>91.66</b>
SH	90.81	84.32	84.40	83.28	81.78	<b>85.89</b>	82.27	81.89	<b>94.48</b>	81.81
TC	85.23	83.85	90.29	86.79	91.75	80.09	92.94	<b>90.73</b>	<b>93.98</b>	89.01
BC	77.86	90.09	78.41	84.21	85.83	81.51	<b>86.55</b>	88.45	86.27	<b>91.11</b>
ST	<b>86.71</b>	89.71	82.86	89.01	83.95	80.95	85.96	85.87	85.03	<b>87.46</b>
SBF	82.96	83.21	85.87	88.44	85.85	80.50	89.77	82.26	<b>91.50</b>	<b>91.40</b>
RA	87.73	80.08	92.70	87.87	88.71	<b>88.32</b>	86.91	80.01	<b>93.32</b>	85.39
HA	90.10	81.07	87.14	85.70	88.86	<b>93.01</b>	<b>91.74</b>	93.15	90.14	86.03
SP	89.31	83.85	93.22	81.15	93.22	80.02	86.19	83.65	<b>93.97</b>	<b>90.74</b>
HC	84.65	82.23	84.70	81.80	83.66	86.11	84.47	80.43	<b>87.47</b>	<b>94.86</b>
Average	85.60	84.37	86.58	85.65	86.10	84.78	86.18	83.79	<b>90.11</b>	<b>89.55</b>

TABLE 3. Experimental results of our proposed method and the state-of-the-art methods using the COCO index on the DOTA dataset. The best results are in bold.

method	$AP_s$	$AP_M$	$AP_L$	$AP_{50}$	$AP_{75}$	$AP$
CAD-Net	41.51	70.54	74.51	66.84	40.40	56.18
R <sup>2</sup> CNN	42.83	70.98	76.52	67.79	44.89	57.05
SCRDet	39.38	69.11	72.08	60.44	40.75	56.37
SCRDet++	41.40	71.78	77.79	64.36	48.54	56.86
YOLT	35.66	72.82	78.45	62.29	44.02	56.69
Gliding Vertex	41.09	73.38	78.58	61.09	40.99	57.90
RoI Transformer	37.77	71.68	<b>80.87</b>	62.35	44.81	58.00
RSDet	41.02	66.25	79.41	62.64	49.32	53.54
PIOU	39.73	69.32	76.88	<b>67.89</b>	42.50	59.75
Ours	<b>48.21</b>	<b>74.83</b>	80.83	67.51	<b>51.00</b>	<b>61.14</b>

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \tag{17}$$

where  $P(R)$  represents the precision at the  $R$  point on the recall curve;  $k$  represents the precision cutoff point;  $P(k)$  and  $R(k)$  respectively represent the precision and recall range of the  $k$  point;  $n$  represents the number of precision cutoff points;

$q$  indicates a certain target category, and  $Q$  indicates the total number of target categories.

### C. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENTAL RESULTS ON NWPU VHR-10 DATASET  
Table 1 lists the test results of CAD-Net [22], R<sup>2</sup>CNN [20], SCRDet [23], SCRDet++ [24], YOLT [26], Gliding

Vertex [25], RoI Transformer [21] and our proposed method on NWPU VHR-10 dataset. It can be seen from Table 1 that the mAP of our method for detecting the 10 categories of targets in the NWPU VHR-10 dataset is 80.31%, which is 3.99% higher than that of SCRDet++ and is superior to that of other popular methods at present. Longitudinally, the detection effect of bridges is the worst, and it may be caused by the confidence region setting. If the IOU of other targets is greater than 0.7, the anchor frame is considered as a positive sample; if the IOU is less than 0.3, the anchor frame is regarded as a negative sample. However, for the bridge target, its size is much larger than other targets, so its sensitivity to IOU should be more relaxed in the large aspect ratio rectangle. Horizontally, R<sup>2</sup>CNN performs the worst because it doesn't consider the boundary problem of rotating coordinate frame, which is very unfavorable for object detection in remote sensing images. SCRDet achieves the highest AP in tennis court object detection, and SCRDet++ achieves the highest AP in detecting storage tanks and basketball courts. So, SCRDet series networks perform well for the detection of these neutral targets. Gliding Vertex achieves the highest AP in ship detection, which may be related to its positioning method. Our proposed method achieves the highest AP in detecting other categories of targets. It obtains good detection performance whether the target is the small-sized vehicle, the large-sized bridge, or the medium-sized baseball diamond. This shows that our proposed method has an advantage in multi-size object detection. The detection results are shown in Figure 9.

## 2) EXPERIMENTAL RESULTS ON THE DOTA DATASET

To further evaluate the detection ability of our proposed method for multi-type, multi-size, and multi-posture targets in large-scale databases, experiments are conducted on the DOTA dataset. Our proposed method and other state-of-the-art methods are compared, and the comparison results are listed in Table 2. Our proposed method achieves an average precision of 90.11% without any data enhancement. In terms of precision and recall, our method performs much better than the methods of Gliding Vertex [25], RSDet [35], PIOU [40], and SCRDet++ [24]. It is because our proposed method realizes the scale perception of foreground features and the accurate mining of context information by denoising the complex background.

This study compares the proposed method with the existing saliency detection methods based on deep learning through the P-R (precision & recall) curve, and the result is shown in Figure 10. It can be observed from the figure that our proposed method obtains the best results. When the recall rate is close to 1, the precision of our method is much higher, indicating that its false alarm is much lower than that of the other methods. Also, as for our proposed method, the resulting visual attention map of the target in the remote sensing image of the large scene with a complex background is closer to the ground truth.

**TABLE 4. Detection runtime(second) of our proposed method and the state-of-the-art methods. The best results are in bold.**

method	DOTA	NWPU VHR-10	RSOD
CAD-Net	0.102	0.101	0.096
R <sup>2</sup> CNN	0.146	0.140	0.141
SCRDet	0.130	0.124	0.120
SCRDet++	0.116	0.105	0.107
YOLT	<b>0.085</b>	<b>0.085</b>	<b>0.083</b>
Gliding Vertex	0.112	0.107	0.109
RoI Transformer	0.139	0.128	0.127
RSDet	0.101	0.099	0.094
PIOU	0.141	0.138	0.136
Ours	0.127	0.124	0.119

For a more rigorous evaluation, the COCO metrics is adopted to compare our proposed method to CAD-Net [22], R<sup>2</sup>CNN [20], SCRDet [23], SCRDet++ [24], YOLT [26], Gliding Vertex [25], RoI Transformer [21], RSDet [35], and PIOU [40] on the DOTA dataset. The comparison result is listed in Table 3.  $AP_S$ ,  $AP_M$ , and  $AP_L$  respectively represent the average precision of detecting small, medium, and large targets.  $AP_{50}$  and  $AP_{75}$  represent the average precision under an IOU of 0.5 and 0.75, respectively.

It can be observed that the AP of our proposed method in 15 categories reaches 61.14%, which is better than the AP of other methods. Also, our proposed method achieves the best results in detecting small and medium-sized targets, with an AP of 48.21% and 74.83% respectively. Besides, better results can be obtained under IoU = 0.75 (1.68% higher than RSDet). This indicates that our method can draw a more accurate boundary box, which helps to identify various targets more accurately in remote sensing images with dense targets. Figure. 11 illustrates some detection results of our proposed method for remote sensing images with dense targets.

## 3) EXPERIMENTAL RESULTS ON RSOD DATASET

To further verify the robustness of our proposed method, SCRDet++ [24], RSDet [35], PIOU [40], and our proposed method are exploited to detect all categories of targets on the RSOD dataset. Figure. 12 shows the results of object detection for each category. It can be seen from the figure that our proposed method performs much better than other advanced methods in terms of the correct detection ratio. Specifically, 95.43% of impervious surfaces, 96.67% of aircrafts, 95.27% of playgrounds, 89.92% of overpasses, and 95.62% of oiltanks are correctly detected. Compared with other methods, our proposed method achieves the highest correct detection rate in all categories of targets. Besides, taking GFT and RA as examples, other methods do not perform well in detecting these two targets, leading to a high false detection rate of these two targets. The false detection rate of RSDet is as high as 15.19%. Our proposed method successfully reduces the false detection rate to 9.23%, achieving a great breakthrough. Figure.13 shows the detection performance of our proposed method on the RSOD dataset.

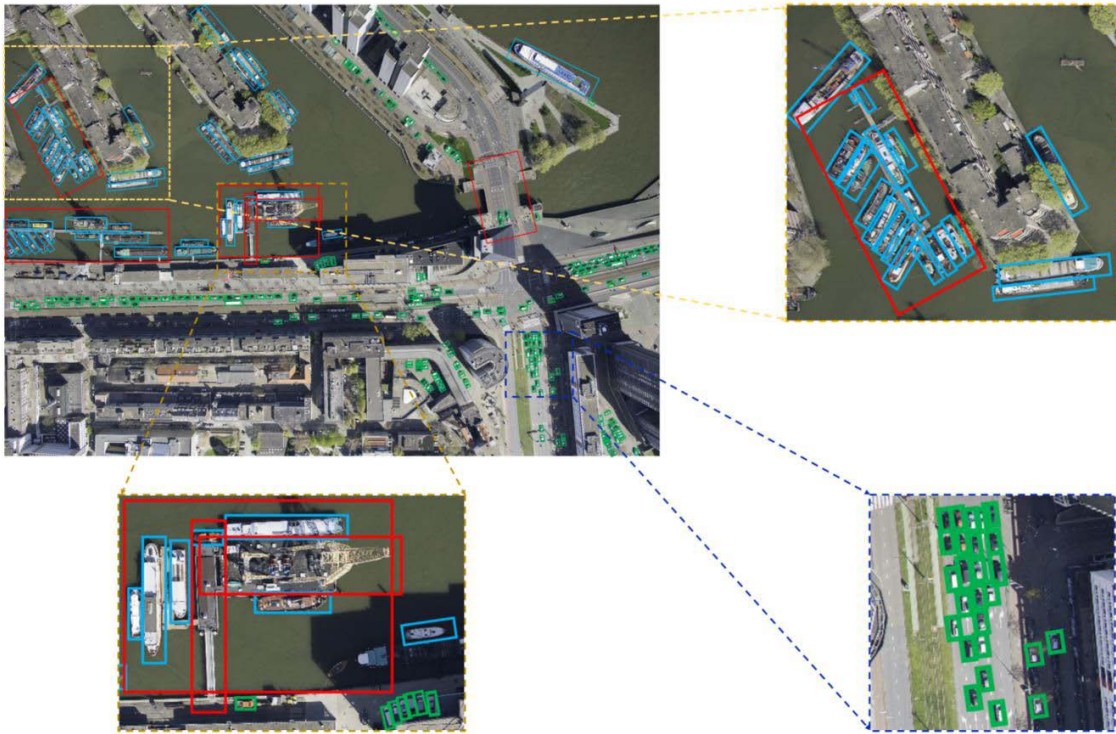


FIGURE 11. Test results of detecting four different targets (ship, harbor, bridge and small vehicle) on the same picture.

	Impervious surface	AC	PG	OP	OT	noise
Impervious surface	93.56%	1.39%	1.21%	0.54%	1.86%	1.44%
AC	0.28%	95.23%	3.04%	0.29%	0.17%	0.99%
PG	0.07%	0.09%	94.11%	4.31%	0.55%	0.87%
OP	4.42%	0.16%	10.64%	83.98%	0.39%	0.41%
OT	1.07%	0.78%	1.29%	1.03%	95.07%	0.76%
noise	36.04%	3.91%	25.35%	2.59%	1.38%	30.73%

(a)

	Impervious surface	AC	PG	OP	OT	noise
Impervious surface	93.72%	1.97%	1.05%	0.33%	1.16%	1.77%
AC	0.45%	95.02%	0.89%	0.38%	0.42%	2.84%
PG	0.10%	0.11%	94.39%	4.51%	0.55%	0.34%
OP	3.15%	0.63%	15.19%	80.13%	0.43%	0.47%
OT	1.53%	0.34%	0.97%	1.65%	94.86%	0.65%
noise	33.48%	3.57%	26.08%	1.36%	1.25%	34.26%

(b)

	Impervious surface	AC	PG	OP	OT	noise
Impervious surface	94.35%	1.12%	1.01%	0.47%	1.09%	1.96%
AC	0.53%	95.54%	0.36%	0.81%	0.62%	2.14%
PG	0.28%	0.49%	94.05%	4.93%	0.18%	0.07%
OP	0.51%	0.19%	13.08%	85.82%	0.13%	0.27%
OT	1.03%	0.44%	0.80%	1.97%	94.31%	1.45%
noise	29.28%	4.06%	29.53%	0.68%	1.83%	34.62%

(c)

	Impervious surface	AC	PG	OP	OT	noise
Impervious surface	95.43%	1.07%	0.95%	0.31%	1.46%	0.78%
AC	0.36%	96.67%	0.34%	0.76%	0.91%	0.96%
PG	0.22%	0.53%	95.27%	3.29%	0.32%	0.37%
OP	0.40%	0.23%	9.23%	89.92%	0.17%	0.05%
OT	0.62%	0.37%	0.18%	1.56%	95.62%	1.65%
noise	26.49%	3.68%	26.79%	0.10%	0.37%	42.57%

(d)

FIGURE 12. The correct detection ratio of SCRDet++, RSDet, PIou, and our proposed method on the RSOD dataset, where (a), (b), (c), and (d) illustrate the detection results of SCRDet++, RSDet, PIou, and our proposed method respectively. The short names for categories are defined as AC-Aircraft, PG-Playground, OP-Overpass and, OT-Oiltank.

#### 4) DETECTION RUNTIME

To compare the detection time of our method with other methods, 200/50/150 images were randomly selected from DOTA, NWPU VHR-10, and RSOD data sets for the experiment of detection runtime, and the average runtime is listed in Table 4. It can be seen that based on the single-stage detection algorithm YOLO, YOLT has the shortest detection

time and the strongest real-time performance. R<sup>2</sup>CNN has the longest detection runtime because it adopts multi-size ROI pooling and oblique frame prediction based on the two-stage detection algorithm Faster RCNN. Our method has a moderate detection runtime among all methods. This is because our method uses two feature fusions, which improves the detection accuracy but leads to slow calculation speed.

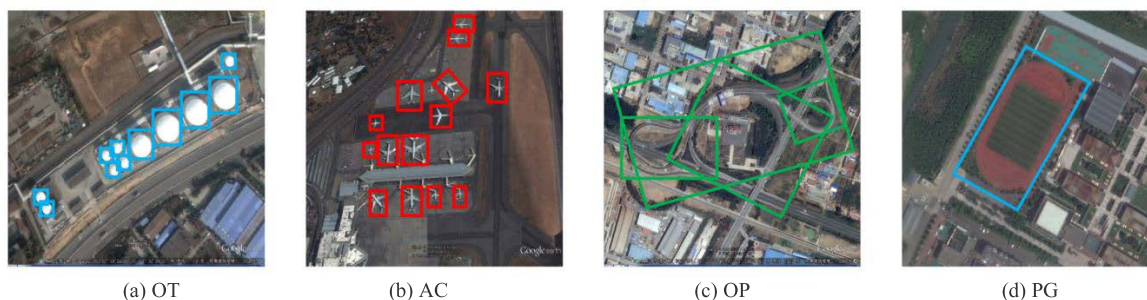


FIGURE 13. The detection performance of our proposed method on the RSOD dataset.

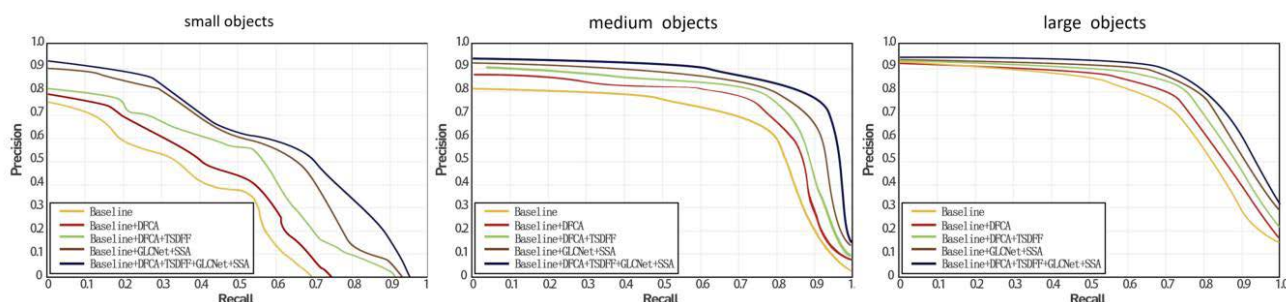


FIGURE 14. Comparison of the P-R curves of ablation experiment. Our proposed method significantly improves the performance of the baseline framework in multi-size object detection.

TABLE 5. The influence of each module (namely DFCA, TSEFF, GLCNet, and SSA) on the object detection performance. The best results are marked in bold.

Baseline	DFCA	TSEFF	GLCNet and SSA	Dataset	$AP_s$	$AP_M$	$AP_L$	$AP_{50}$	$AP_{75}$	$AP$
✓				DOTA	39.21	65.87	69.30	61.09	36.28	54.83
✓	✓			DOTA	41.15	71.23	70.86	64.61	38.74	55.59
✓		✓		DOTA	40.79	70.81	71.51	62.85	38.60	55.23
✓	✓	✓		DOTA	43.62	72.03	75.95	65.74	43.57	56.92
✓	✓		✓	DOTA	44.13	71.76	77.42	63.08	44.30	58.11
✓		✓	✓	DOTA	41.80	70.93	72.05	63.41	39.48	56.03
✓	✓	✓	✓	DOTA	<b>48.21</b>	<b>74.83</b>	<b>80.83</b>	<b>67.51</b>	<b>51.00</b>	<b>61.14</b>

D. ABLATION EXPERIMENT

In this section, the influence of each module in our proposed method on object detection performance is investigated on the DOTA dataset. The ablation results of adding the modules (namely DFCA, TSEFF, GLCNet, and SSA) to the MoblieNets framework are listed in Table 5. The MoblieNets backbone network achieves a detection efficiency of 54.83%. DFCA is conducive to obtaining foreground semantics from large scenes complex backgrounds. It consists of bottom-up and top-down subnets to circulate low-level/intermediate-level and high-level semantic information. It increases the AP of detecting small, medium, and large targets by 1.94%, 5.36%, and 1.56%. Then, for small targets, TSEFF further improves the AP by 2.09% because it can enhance the semantic information of small targets and maximize the differences between the targets of different sizes and categories. Finally, with GLCNet and SSA, the useful information of the fused feature map can be further screened out to make the detected features more characteristic. The final AP is 61.14%.

To show the results of the ablation experiment more intuitively, the P-R curves of detecting small, medium, and large targets are compared. It can be seen from Figure. 14 that the effectiveness of our proposed method in detecting multi-size targets is greatly improved, especially in detecting small targets. When the recall rate is 0.6, the precision of small object detection is about 0.57, which is much higher than that of the backbone network. This improvement indicates that the proposed method can further detect small targets from complex backgrounds, showing that our method is effective for object detection in remote sensing images.

V. CONCLUSION AND FUTURE WORK

In this paper, a model is proposed for multi-size object detection of remote sensing images with large scenes. The model uses the MoblieNets network to extract image features and the MFCA module to pay attention to different regions in the feature map. Then, the feature maps are deeply fused by TSEFF, and the features are characterized by GLCNet and

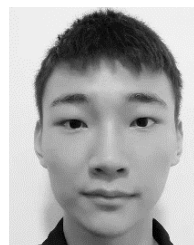
SSA. The experimental results show that our method can be used as an effective target detection method in remote sensing images considering the detection accuracy and detection time. In future work, we will improve the model to realize real-time object detection especially for remote sensing images of large scenes.

## ACKNOWLEDGMENT

(Jinkang Wang and Xiaohui He are co-first authors.)

## REFERENCES

- [1] Z. An, Z. Shi, X. Teng, X. Yu, and W. Tang, "An automated airplane detection system for large panchromatic image with high spatial resolution," *Optik*, vol. 125, no. 12, pp. 2768–2775, Jun. 2014.
- [2] J. Qiu, S. Li, and W. Wang, "A new approach to detect aircrafts in remote sensing images based on corner and edge information fusion," *Microelectron. Comput.*, vol. 28, no. 9, pp. 214–216, Apr. 2011.
- [3] L. I. N. Yu-Dong, H. E. Hong-jie, Y. Zhong-ke, and C. H. E. N. Fan, "Air-plane detection in optical remote sensing image based on sparse-representation," *Acta Photonica Sinica*, vol. 43, no. 9, pp. 1–6, Jul. 2014.
- [4] Y. Yu, H. Guan, D. Zai, and Z. Ji, "Rotation-and-scale-invariant airplane detection in high-resolution satellite images based on deep-Hough-forests," *ISPRS J. Photogramm. Remote Sens.*, vol. 112, pp. 50–64, Feb. 2016.
- [5] 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.
- [6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [7] 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.
- [8] W. Liu, D. Anguelov, D. Erhan, and C. Szeged, *SSD: Single Shot MultiBox Detector*. Cham, Switzerland: Springer, 2016.
- [9] 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.
- [10] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1222–1230.
- [11] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.
- [12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [13] M. Kisantala, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*.
- [14] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.
- [15] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao, "Efficient feature-based image pyramid network for single shot detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7336–7344.
- [16] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.
- [17] 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. (CVPR)*, Jul. 2017, pp. 2117–2125.
- [18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
- [19] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, "ThunderNet: Towards real-time generic object detection on mobile devices," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6718–6727.
- [20] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.
- [21] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for detecting oriented objects in aerial images," 2018, *arXiv:1812.00155*.
- [22] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Aug. 2019.
- [23] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8232–8241.
- [24] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," 2020, *arXiv:2004.13316*.
- [25] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
- [26] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," 2018, *arXiv:1805.09512*.
- [27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [28] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
- [29] X. Hua, X. Wang, T. Rui, D. Wang, and F. Shao, "Real-time object detection in remote sensing images based on visual perception and memory reasoning," *Electronics*, vol. 8, no. 10, p. 1151, Oct. 2019.
- [30] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [31] J. Li, X. Zhao, and H. Li, "Method for detecting road pavement damage based on deep learning," *Proc. SPIE*, vol. 10972, Apr. 2019, Art. no. 109722D.
- [32] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5998–6008.
- [34] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," 2021, *arXiv:2105.02358*.
- [35] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, "Learning modulated loss for rotated object detection," 2019, *arXiv:1911.08299*.
- [36] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
- [37] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," 2017, *arXiv:1711.10398*.
- [38] 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.
- [39] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, Apr. 2019.
- [40] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, and C. Yang, "PloU loss: Towards accurate oriented object detection in complex environments," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 195–211.



**JINKANG WANG** (Student Member, IEEE) received the bachelor's degree in mechanical engineering from the Army Engineering University of PLA, China, in 2020, where he is currently pursuing the master's degree in mechanical engineering. His current research interests include mechanics, machine learning, and computer vision.



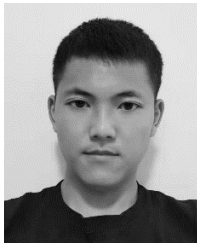
**XIAOHUI HE** (Senior Member, IEEE) was born in 1975. He received the Ph.D. degree from the Army Engineering University of PLA, China. He is an Associate Professor with the Army Engineering University of PLA. His research interests include mechatronics and deep learning.



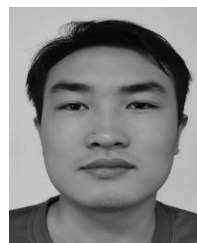
**QUNYAN JIANG** (Student Member, IEEE) is currently pursuing the master's degree with the College of Field Engineering, Army Engineering University of PLA. His research interest includes machine learning.



**SHAO FAMING** (Senior Member, IEEE) was born in 1978. He received the Ph.D. degree from the Army Engineering University of PLA, China. He is an Associate Professor with the Army Engineering University of PLA. His research interests include signal processing, deep learning, and software engineering.



**GUANLIN LU** (Student Member, IEEE) is currently pursuing the master's degree with the College of Field Engineering, Army Engineering University of PLA. His research interest includes machine learning.



**RUIZHE HU** (Student Member, IEEE) is currently pursuing the master's degree with the College of Field Engineering, Army Engineering University of PLA. His research interest includes machine learning.

...