Optical Remote Sensing Object Detection Based on Background Separation and Small Object Compensation Strategy

Yan Dong [1,2], Haotian Yang [1], Shanliang Liu [1*], Guangshuai Gao [1], Chunlei Li [1]

[1] School of Electronic and Information Engineering, Zhongyuan University of Technology, ZhengZhou, China

[2] School of Automation Engineering, University of Electronic Science and Technology of China, ChengDu, China

[*] Correspondence:shliucauc@163.com

***Abstract*—Solid and accurate object detection in optical remote sensing images still remains significant challenges such as complex background and weak object information. To alleviate above problems, we propose a revolutionary one-stage object detection network. Specifically, the proposed effective localization attention is embedded in deep feature maps with more channels, and is used to locate channels that are effective for detection tasks through one dimensional convolution operations. Following that, a small object compensation strategy is proposed to use compensation fusion operation to enable the reuse and compensation of weak object information. Additionally, a background separation strategy is designed to separate foreground and background, highlighting the features of interest, suppressing background noise. To ascertain our model, extensive experiments are conducted in three public datasets, it can simply achieve mAP of 94.2%, 70.7% and 80.5% in NWPU VHR-10, DIOR and DOTA datasets.**

***Index Terms*—Optical remote sensing images, object detection, effective localization attention module, background separation strategy, small object compensation strategy.**

I. Introduction

IT is easy to obtain overwhelming high-resolution remote sensing images, and object detection in remote sensing images has become a research priority presently [1]. The purpose of remote sensing object detection is to determine the position and category of the object in an image, which has important significance and widely application in military, navigation, agriculture, [2], [3], [4], [5], [6], [7]. Inversely, achieving higher detection accuracy while keeping faster reference speed is still an urgent and pressing problem to be solved.

Recently, deep learning methods have been widely used, because of the strong feature extraction capability and advanced performance, and many object detectors based on CNN have been proposed, such as mainstream one-stage detection algorithms [8], [9], [10]. Paradoxically, in contrast to images in natural scenes, remote sensing images are usually taken at different altitudes, contain many small objects, and the image background is relatively complex. Relatively small objects in an image contain less information, and as the CNN progresses deeper, the loss of information about small objects becomes more severe, making it difficult to exactly detect objects. Consequentially, due to the complex background and suffer from insufficient information on objects, the features which they are extracted usually have noise, which caused false detection easily. Some samples are shown in Fig. 1.

Fig. 1. These samples explain the two challenges in remote sensing object detection. The top two graphs represent complex backgrounds and the bottom two graphs represent weak object information

For the difficulty in detecting small objects, many researchers have generated characteristics of high quality by constructing and aggregating multi-scale features. For instance, the improved feature fusion network used a weighted fusion method to fuse different feature layers, achieving higher accuracy and superior real-time performance [11]. To reduce

This work was supported by NSFC (No.62301623, No. 62072489), IRTSTHN (21IRTSTHN013), Leading talents of science and technology in the Central Plain of China (234200510009), Henan province key science and technology research projects (222102210008, 232102211002, 232102211030). (Corresponding author: Shanliang Liu.)

The author are with the School of Electronic Information, Zhongyuan University of Technology, Zhengzhou 450007, China(e-mail: dy@zut.edu.cn;1148406478@qq.com;shliucauc@163.com; 6911@zut.edu.cn; lichunlei1979@zut.edu.cn).

information loss on small objects due to continuous downsampling operations, [12], [13], [14] used parallel unfolding convolutional layers with different rates to reconstruct different levels in the feature pyramid and enhance the contextual information of small objects. Though these methods had improved detection capabilities, they still have some drawbacks. The direct feature fusion of different feature layers extracted from the backbone network introduces a large amount of noise while enriching the contextual information. Simultaneously, they did not take into account the variability between different feature layers, which can affect the detection results. For small objects, improving detection accuracy is still limited by using only feature fusion enhancement techniques. While these methods treated attention and convolution as distinct parts, the relations between them were not fully exploited. Analogously, they did not take into account the decoupling of background and foreground.

For the complex background, Dong et al. [15] integrated local and global contextual information into FPN. Ma et al. [16] improved the YOLOv5 algorithm for small objects detection, integrating the CBAM to make target information to be enhanced. Cheng et al. [17] introduced a multi-scale feature aggregation module, which can achieve information fusion and interaction by aggregating feature maps at different scales.

To mitigate above problems, we propose an optical remote sensing object detection algorithm based on background separation and small object compensation strategy. First, an effective channel attention module is proposed to focus more on channel features that are effective for detection task and achieve more powerful feature extraction during the fusion process. Following that, a small object compensation strategy is used to reuse the small object information to avoid flooding the network with small object information. Subsequently, the background separation strategy is to separate background and foreground objects to enhance the object features. Additionally, the enhanced features are sent to prediction, that consist of three independent regression branching, and are integrated at prediction time. And, the public DIOR [18], NWPU VHR-10 [12], and DOTA [19] datasets are used to demonstrate that our proposed algorithm can solve these problems as described above, our approach is better than other state-of-the-art detectors based CNN. Our method can achieve a mAP of 70.7%, 94.2% and 80.5% on DIOR, NWPU VHR-10 and DOTA datasets, respectively.

In nutshell, our works mainly contributes to the following:

1) An effective localization attention module (ELAM) is proposed to focus more on channel features that are effective for detection task, provide effective feature descriptions and reduce redundant information.

2) A small object compensation strategy (SOCS) is proposed to effectively extract contextual information while also compensating for discriminative details and small objects to enhance feature representation.

3) Be aimed at the complex background, a background separation strategy (BSS) is proposed to separate the foreground and background, enhance the features of the objects.

4) Extensive experiments are carried out on the widely used NWPU VHR-10, DIOR and DOTA datasets, the results have proven the effectiveness of our model.

II. Related Work

A. Remote Sensing Object Detection

Since complex background information in remote sensing images can interfere with target detection, a network with hybrid attention was proposed to get contextual information [20]. A new multiscale variability attention module is designed and added to the top of the feature pyramid to highlight features [21]. Li et al. [22] proposed a parameter free masking module that can detach instance related foreground and instance independent background in multiscale features. To mitigate the interference of complex backgrounds, Yang et al. [23] achieved accurate positioning and classification of objects by introducing the Coordinate Attention Module(CoAM). Hu et al. [24] proposed an attention-guided multi-scale detection network structure, that can successfully detect ships even in complex scenarios. Ma et al. [25] proposed a one stage scale aware network for remote sensing target detection, a target saliency enhancement strategy was to enhance the features of interest through the proposed affiliation function and suppresses the background information, consequently preventing false and missed detections.

For the large scale variation of targets in remote sensing images, a novel model with density map and attention mechanism (DA2FNet) was proposed [26]. Shen et al. [27] improved the Adaptively Spatial Feature Fusion (ASFF) to fuse multiple branches, improving accuracy through fusion. Wan et al. [28] introduced a MobileNetV2S as the backbone of YOLOX, and designed a detection head that is conducive to multi-scale detection. Teng et al. [29] proposed a model that integrates global contextual cues extracted and local contextual cues encoding local spatial contextual correlation, and designed adaptive anchor blocks using rich semantic features to effectively mitigate scale variations. Wang et al. [30] proposed a new multiscale enhancement network (MSE-Net) that integrates Laplacian kernels with fewer parallel multiscale convolutional layers to provide multiscale description enhancement. Li et al. [31] proposed a network (SGFTHR) with hybrid residual operations to extract structural information with significant differences for easy identification.

Remote sensing image imaging is characterized by arbitrary orientation. For the rotating targets, Shi et al. [32] proposed a method for detecting aircraft in arbitrary orientation in high-resolution aerial images. Hu et al. [33] designed a local and nonlocal attention model to obtain local and nonlocal features separately. In order to smooth L1 loss, Yang et al. [34] introduced the IoU constant, for more accurate rotation estimation to solve the boundary problem of rotating bounding box. A rotate bounding box is used for ship detection [35]. Dong et al. [36] designed a vector field filter and a neural network for remote sensing. Wang et al. [37] used transformation to convert regression to classification to eliminate the confounding information caused by angular discontinuities. Jiang et al. [38] proposed a field edge decomposition rotate bounding box based on centernet to avoid boundary discontinuity.

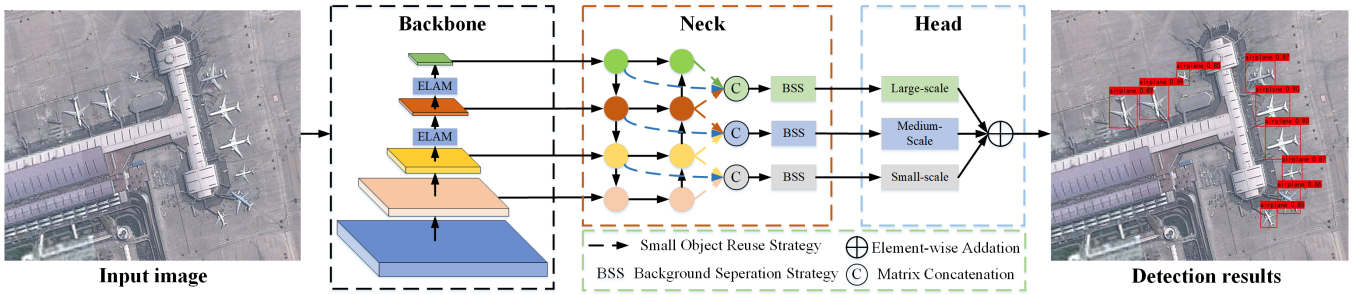

Fig. 2. The structure diagram of our method. From left to right are the backbone, neck, and detection head in sequence. An effective localization attention is proposed to perform feature refinement on small objects. Then, the small object compensation strategy can be reused by obtaining contextual feature information at different scales and for small objects. The background separation strategy decouples the instance related foreground and the instance independent background. Finally, the prediction maps are generated using the prediction head.

B. Attention Mechanism

Attention mechanism is applied to machine translation tasks originally, it mainly has two aspects, one is determining which part of the input needs to be focused, another is allocating limited information processing resources for important parts. Attention mechanism is used in various areas of deep learning, whether in image segmentation, speech processing, or in computer vision and natural language processing. Hou et al. [39] designed a new attention in order to emphasize positional information, called "coordinated attention", which can encode feature maps that enhance the target of interest. It is easy to be implemented and is a lightweight structure. A new normalization based attention module (NAM) was proposed to achieve better performance while also ensuring high computational efficiency by suppresses less significant weights [40].

C. Feature Fusion

Feature fusion refers to the combination of feature information from different sources or layers to generate richer and more representative feature representations.

Zhao et al. [41] had added a multi-scale attention feature fusion (MSAFF) structure based on YOLOX, which can expand the receptive field to capture larger contextual information. Song et al. [42] designed adaptive instance normalization (AdaIN) blocks during the fusion process to improve the performance of the object detection model by improving it's adaptability to different types of images, and used attention modules (AM) to help the model better handle small domain changes and local information weights. Wang et al. [43] had introduced a feature fusion module to effectively aggregate features at different levels. Zhao et al. [44] proposed an Attention Feature Fusion Module (AFFM) that fully integrates and refines texture features and semantic features of aircraft. A multi-scale feature fusion module was proposed to obtain local details and global contextual features [45]. Chen et al. [46] had designed a feature fusion module that facilitates the full transmission of spatial and semantic information through effective bidirectional cross connections and weighted fusion.

Inspired by these works, attention mechanisms and various strategies are aggregated in our work, such as small object compensation strategy and background separation strategy.

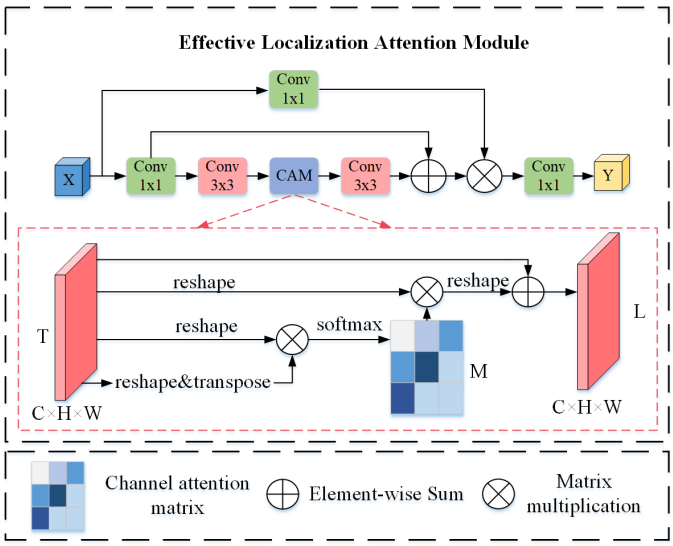

Fig. 3. The details of ELAM.

III. Proposed Method

A. Overview

The network model, as illustrated in Fig. 2. We use YOLOX as the baseline, and the detection algorithm consists of three parts: feature extraction, enhanced feature extraction, and prediction. Firstly, in the feature extraction stage, an effective localization attention module is designed and embedded into the backbone network to achieve feature refinement of small objects by locating key channels. Simultaneously, to enhance detection ability of small objects a small object compensation strategy is proposed, by obtaining contextual feature information at different scales and reusing small objects. Additionally, the background separation strategy is used to separate the object related foreground and the object independent background to further promote the ability to detect objects in complex contexts. Finally, different feature maps are send to the detection head for prediction.

B. Effective Localization Attention Module

We design an effective localization attention module based on channel attention module (CAM), the details are shown

in Fig. 3. Due to the shallow feature maps mainly contain local details and edge information about the image, they have relatively rich information on localization. Paradoxically, shallow feature maps have relatively less semantic information compared to deep feature maps, which make it difficult to express higher-level semantic information such as relationships and categories between various objects in the image. Equally, deep feature maps can capture richer semantic information and identify various objects and their categories in the image through multi-level abstraction and induction. Whereas, deep feature maps have relatively less positioning information, and difficult to provide precise positioning information compared to shallow feature maps. Consequently, in computer vision tasks, it is usually necessary to combine the advantages of shallow and deep feature maps to achieve more accurate and refined image analysis and understanding. Effective localization attention module can localize channel features that are practical for the detection task. As consequence, we embed it before C4 and C5 of the backbone network.

As Fig. 3, we split the input X into horizontal and vertical branches. The vertical branch undergoes 1×1 convolution for channel number adjustment. The horizontal branch is captured information between channels through different convolutional operations based on CAM. The detailed diagram is shown in Fig. 3. After Element-wise Sum, we obtain the output in the horizontal direction. After that, we perform matrix multiplication on two outputs in different directions. In the end, a 1×1 convolution is used for dimensionality reduction to obtain the output Y.

For CAM, the original input feature map $T \in R^{C \times H \times W}$ is reshaped into $C \times K$, K = $H \times W$, we multiply the reshaped result with its transposition to obtain the attention feature map M. Subsequently, we multiply M with its reshaped result and add the multiplied result to the original feature map to achieve $L(C \times H \times W)$. Where the equation is expressed as follows:

$$m_{ij} = \frac{\exp(A_i A_j)}{\sum_{i=1}^{C} \exp(A_i A_j)} \tag{1}$$

$$L_j = \beta \sum_{i=1}^{C} (m_{ij} A_i) + A_j \tag{2}$$

Where m_{ij} measures the i_{th} channel's impact on the j_{th} channel. β is a scale parameter, we set it as 1.

In short, an effective localization attention module (ELAM) is proposed to focus more on channel features that are effective for the detection task and provide validity feature descriptions.

C. Small Object Compensation Strategy

Small targets contain relatively little information, but when extracting features, we often use operations such as convolution, down-sampling, and pooling, which may lead to information loss. As in Fig. 4, we visualize the adjacent feature maps extracted from the backbone network. Observe that as the depth of the network increases, the small object information is gradually lost. As the number of iterations increases, the model gradually focuses more on large objects.

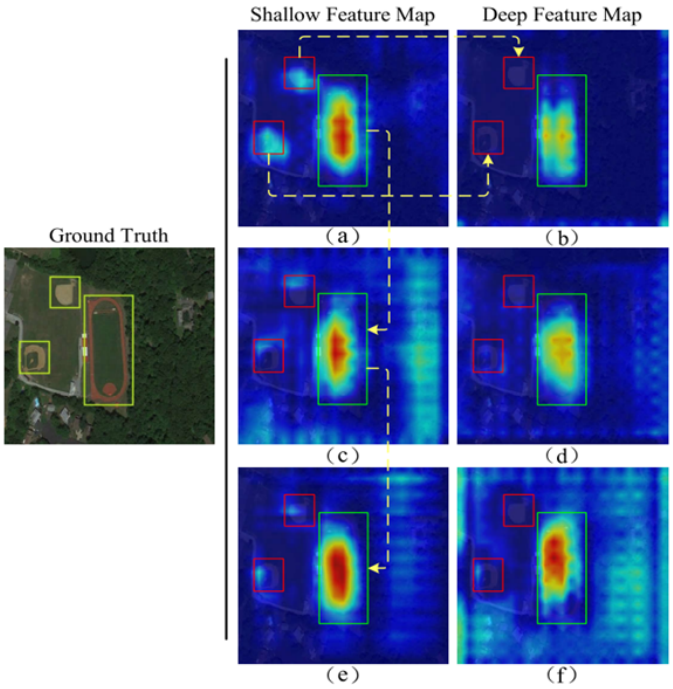

Fig. 4. Feature map visualization. By comparing each row, we find that the information of small objects gradually become lost with the development of network; by comparing each column, we find that as the number of training repetitions increases, it gradually focuses more on large objects.

As consequence, convolution, pooling and other operations are also an influential factor in the unsatisfactory detection of small targets. The information reuse capability of the model for small objects needs to be improved.for enhancing the detection accuracy of small objects.

Follow that, we propose a small object compensation strategy, the details are shown in Fig. 5. After feature extraction, we obtained different effective feature layers, where F1 represents the shallow feature map, F2 represents the middle layer, and F3 represents the deep feature map. A 3×3 convolutional operation is performed on F3 for channel number reduction and model parameter reduction, and then W is obtained after the Sigmoid activation function, which is to enhance the nonlinear representation capability of our model.

In order to acquire the composite semantic feature map, we perform matrix concatenation operations on F4 and F1. It can not only retain the preservation of spatial and semantic information about the feature map as much as possible, but also effectively extract multi-scale contextual information, which is efficacious in detecting remote sensing small objects. Consequently, channel shuffle is performed to exchange the features between groups (shuffle), so that each group contains the features of other groups to achieve the purpose of enhanced feature extraction. As a result, a 3×3 convolution operation is performed to obtain the output feature map F.

In brief, we propose a small object compensation strategy. It enables the fusion and interaction of contextual information by compensating information operations, which can efficiently moderate the sluggish performance results of small object detection.

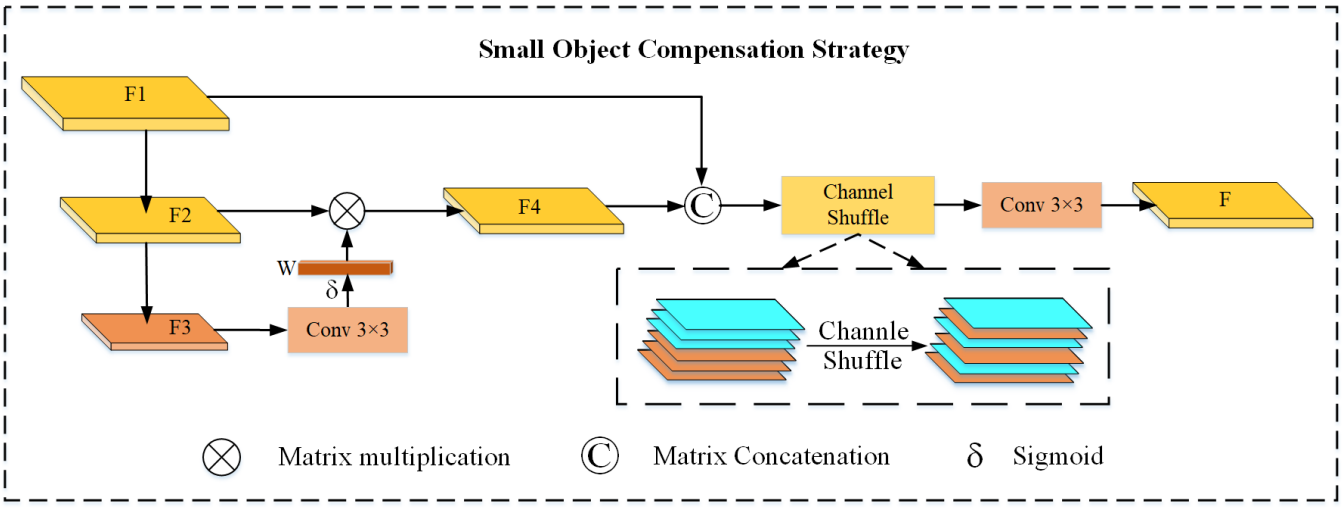

Fig. 5. Structure diagram of the small object compensation strategy. F1 indicates shallow feature map, F2 indicates deep feature map. The fusion and interaction of contextual information through compensation information operations can effectively alleviate the problem of poor performance results of small object detection.

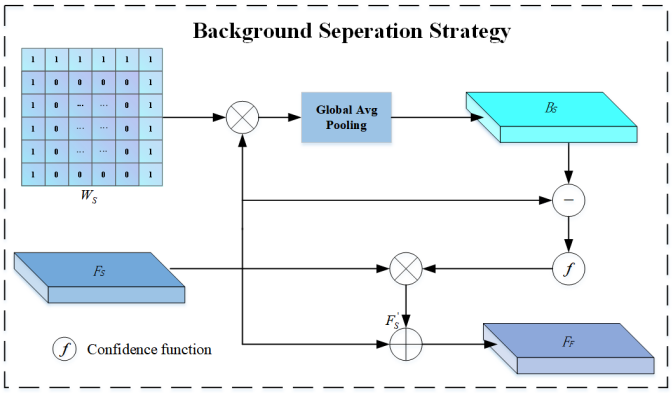

Fig. 6. Background Separation Strategy. Enhancing features of interest, suppresses background noise.

D. Background Separation Strategy

As the consequence of the special characteristics of remote sensing imaging technology, the viewing angle of remote sensing images is mostly overhead, which inevitably obscures the target, causing incomplete features of the object, consequently leading to the missed detection of some objects. Additionally, the remote sensing image field of view is relatively large and contains a variety of backgrounds, which will introduce a lot of noise in the feature extraction process and produce stronger interference to the object detection, further weakening the feature characterization ability of the object.

In a nutshell, a background separation strategy is proposed in this paper, and the detail diagram is shown in Fig. 6. We consider the edge pixels as background. Our formula for calculating F_s and $B_s \in R^{C\times H\times W}$ is as follows:

$$B_s = AvgPool(W_s \times F_s) \tag{3}$$

Where $W_s \in R^{C\times H\times W}$ is a tensor with learning ability, the edge value is 1, other areas value is 0. We use the learnable tensor to obtain edge pixels of F_s and use the global maximum pooling to obtain the background B_s, which preserves the spatial structure information of F_s. During the process of training, we continuously optimize the element values of W_s to better represent the background. Then, we consider B_s as the difference between the feature maps F_s and B_s. The calculation formula is shown as follows: We use the learnable tensor to obtain all edge pixels of W_s .The calculation formula is shown as follows:

$$\Delta W_s = F_s - B_s \tag{4}$$

Usually, there are differences between the element values in correspondence with the object area and the background. $\triangle W_s(i,j)$ is obtained by subtracting between F_s and B_s , the greater of it, the more likely a specific location in the feature map will contain the object information, and therefore the model should focus more on that object region. We define the confidence function f to give an account of correlation, determined as follows:

$$f = \begin{cases} 1 - e^{-\Delta W_s} & \text{if } \Delta W_s > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We can get the feature map after background separation by putting the input feature map through the correlation function calculation. Whereas, in this process, we should understand that the value of f is really between [0,1], which reflects the confidence magnitude of each element belonging to the object. It is mainly used to make our model focus more on the object region and suppress the background region. Consequently, we can the expression of the final feature map used for detection, determined as follows:

$$F_F = F_s \oplus (F_s^{'} \times f) \tag{6}$$

The background separation strategy can separate the object from the complex background in accordance with f , it boosts the features of the object, reduces the interference of noise, the details of which will be developed in the Section of Experiments.

TABLE I
ABALATION EXPERIMENTS ON NWPU VHR-10 (%)

Group	Baseline	ELAM	SOCS	BSS	mAP
Group1	✓				92.3
Group2	✓	✓			93.4
Group3	✓		✓		93.5
Group4	✓			✓	93.3
Group5	✓	✓	✓		93.6
Group6	✓	✓		✓	93.7
Group7	✓		✓	✓	93.9
Group8	✓	✓	✓	✓	94.2

TABLE II
ABALATION EXPERIMENTS ON DIOR (%)

Group	Baseline	ELAM	SOCS	BSS	mAP
Group1	✓				66.9
Group2	✓	✓			68.3
Group3	✓		✓		68.5
Group4	✓			✓	68.3
Group5	✓	✓	✓		68.6
Group6	✓	✓		✓	68.9
Group7	✓		✓	✓	69.3
Group8	✓	✓	✓	✓	70.7

TABLE III
ABALATION EXPERIMENTS ON DOTA (%)

Group	Baseline	ELAM	SOCS	BSS	mAP
Group1	✓				79.0
Group2	✓	✓			79.8
Group3	✓		✓		79.9
Group4	✓			✓	79.8
Group5	✓	✓	✓		80.1
Group6	✓	✓		✓	80.1
Group7	✓		✓	✓	80.2
Group8	✓	✓	✓	✓	80.5

IV. EXPERIMENTS

We first describe the dataset and the evaluation metrics utilized in the experiment. To demonstrate the effectiveness of our approach, we conduct extensive experiments with three sets of data separately. Accordingly, we conduct extensive ablation experiments to estimate the performance of each proposed module.

A. Datasets

In our experiment, we evaluated and analyzed the achievements of our method on three different datasets, DIOR, NWPU VHR-10 and DOTA.

1) DIOR: It's a large-scale, publicly available object detection dataset released by Northwestern Polytechnical University. It consists of 23463 images, totaling 20 categories. 11725 images are used for training and validation , 11738 images are used for testing.

2) NWPU VHR-10: It's also a public remote sensing image object detection dataset that includes 650 high-resolution optical remote sensing images in a total of 10 categories.

3) DOTA: DOTA contains 15 types of objects of different scales, directions, and shapes. We split it into 640×640 pixel blocks during training and validation.

B. Evaluation Metrics and Parameter Settings

The achievements of the network are measured using the mean Average Precision (mAP) and Average Precision (AP). AP measures the performance of the learned model in each category; the higher the AP value, the better the classification. mAP measures the performance of the model in all categories, and its range is [0, 1].

The mAP is the average of the AP across all categories. It evaluates the performance of the model more comprehensively compared to accuracy and recall. Therefore, the above two evaluation metrics are important indicators of the achievement of the object detection algorithm. Precision indicates how many of the predicted positive samples are truly positive, and mAP is the average precision for multiple objects. Precision, recall, AP and mAP are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \frac{1}{N} AP_i \tag{10}$$

Where TP, FP, and FN separately represent the sum total of true positives, false positives, and false negatives.

The training process is divided into freeze training and unfreeze training. We choose YOLOX as our benchmark model and trained a total of 150 epochs. First, the pre-trained model is loaded and the freeze training is executed in the first 50 epochs, we set the learning rate to 0.001, and use the cosine annealing method to train and adjust only the later parts of the backbone network. Unfreeze training is executed in the last 100 epochs, we set learning rate to 0.0001. At this stage, the cosine annealing algorithm is still used. We use SGD to optimize the model throughout the entire training process. The same parameter settings are used during the training process of all experiments. We use Pytorch on DGX-Station-A100 GPU to train our model, and use a framework of Python 1.7.1 and CUDA version 11.0.

C. Experimental Results

We first presented the contributions of three modules in our model through ablation experiments, including effective attention mechanism modules, small object reuse strategies, and background separation strategies. Then we compared our model with other state-of-the-art detectors on three widely used datasets, NWPU VHR-10, DIOR and DOTA dataset.

1) Ablation Experiment

To evaluate the effectiveness of different modules, we conducted ablation experiments on the NWPU VHR-10 and DIOR dataset, all experiments used the same settings, we can see the results in Table I, Table II and Table III.

It can be seen that the method proposed in this article has achieved significant improvements in terms of effective

TABLE IV
COMPARISON EXPERIMENTS ON DIOR DATASET (%)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Expressway service area	Expressway toll station	Goff course
C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Ground track field	Harbor	Overpass	Ship	Stadium	Storage tank	Tennis court	Train station	Vehicle	Windmill

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	mAP
R-CNN [47]	35.6	43.0	53.8	62.3	15.6	53.7	33.7	50.2	33.5	50.1	49.3	39.5	30.9	9.1	60.8	18.0	54.0	36.1	9.1	16.4	37.7
SSD [9]	59.5	72.7	72.4	75.7	29.7	65.8	56.6	63.5	53.1	65.3	68.6	49.4	48.1	59.2	61.0	46.6	76.3	55.1	27.4	65.7	58.6
F-RCNN [48]	53.6	49.3	**78.8**	66.2	28.0	70.9	62.3	69.0	56.0	68.9	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1
CornerNet [49]	58.8	**84.2**	72.0	80.8	**46.4**	75.3	**64.3**	**81.6**	**76.3**	**79.5**	**79.5**	26.1	**60.6**	37.6	70.7	45.2	84.0	57.1	43.0	75.9	64.9
RetinaNet [10]	53.7	77.3	69.0	81.3	44.1	72.3	62.5	76.2	66.0	77.7	74.2	50.7	59.6	71.2	69.3	44.8	81.3	54.2	45.1	83.4	65.7
YOLOX [50]	86.7	70.5	74.0	89.0	40.8	74.8	46.9	56.5	57.2	72.7	70.9	60.9	57.1	88.4	61.9	71.7	88.6	35.8	52.0	82.2	66.9
SCFNet [51]	82.1	78.0	77.7	88.7	42.8	76.0	57.8	60.9	59.7	77.3	41.8	61.9	57.3	89.3	70.8	74.9	86.9	55.6	52.7	76.0	69.9
RAST-YOLO [52]	84.3	76.4	78.7	85.9	40.2	76.8	50.2	62.6	56.5	77.1	73.7	61.1	56.6	**91.1**	**74.3**	**77.9**	**89.3**	53.3	54.0	76.2	69.8
Ours	**87.4**	78.4	74.5	**89.1**	44.3	**77.6**	59.8	58.5	60.2	78.2	71.8	**62.4**	59.2	89.7	65.7	74.0	88.5	**55.6**	**54.4**	**84.8**	**70.7**

TABLE V
COMPARISON EXPERIMENTS ON NWPU VHR-10 DATASET (%)

Airplane-APL, Baseball diamond-BD, Basketball court-BC, Bridge-BR, Ground track field-GTF, Harbor-HA, Ship-SH, Storage tank-STO, Tennis court-TC, Vehicle-VE.

Methods	APL	BD	BC	BR	GTF	HA	SH	STO	TC	VE	mAP
F-RCNN [48]	82.8	96.3	68.8	78.8	98.4	82.5	77.5	52.5	62.9	63.8	76.4
M-RCNN [53]	93.2	90.4	91.2	60.6	95.2	75.2	75.5	92.9	90.3	74.2	83.9
RFBNet300 [54]	97.2	97.7	93.8	**97.6**	96.5	**98.5**	77.4	59.8	81.6	55.2	85.5
YOLOv4 [55]	94.9	98.3	67.5	95.9	99.3	80.7	78.6	95.4	88.2	67.7	86.7
DNN [56]	93.0	92.8	89.0	81.0	78.0	76.0	84.5	87.1	82.0	84.5	84.8
RICAOD [57]	99.7	92.9	80.3	68.5	90.8	80.3	90.8	90.6	90.3	87.1	87.1
MEDNet [58]	99.2	98.5	95.2	75.1	98.3	88.1	94.4	82.2	95.4	89.3	91.6
EGAT-LSTM [59]	97.3	96.5	94.5	80.1	94.2	86.2	**96.7**	**97.2**	86.6	90.8	92.0
YOLOX [50]	99.4	**99.9**	95.8	72.2	**100**	94.8	84.3	92.9	94.8	88.9	92.3
ours	**99.9**	97.3	**98.8**	85.4	**100**	97.7	85.9	89.7	**96.6**	**90.8**	**94.2**

TABLE VI
COMPARISON EXPERIMENTS ON DOTA DATASET (%)

Plane-PL, Baseball diamond-BD, Bridge-BR, Ground track field-GTF, Small vehicle-SV, Large vehicle-LV, Ship-SH, Tennis court-TC, Basketball court-BC, Storage tank-ST, Soccer ball field-SBF, Roundabout-RA, Tennis court-TC, Harbor-HA, Swimming pool-SP, helicopter-HP.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SCRDet [60]	89.9	80.6	52.1	68.3	68.3	60.3	72.4	90.8	87.9	86.8	65.0	66.6	66.2	68.2	65.2	72.6
FAOD [61]	90.2	79.6	45.5	76.4	73.1	68.2	79.6	90.8	83.4	84.7	53.4	65.4	74.2	69.7	64.9	73.2
RSDet [62]	90.1	82.0	53.8	68.5	70.2	78.7	73.6	91.2	87.1	84.7	64.3	68.2	66.1	69.3	63.7	74.1
S2ANet [63]	89.1	82.8	48.4	71.1	78.1	78.4	87.3	90.8	84.9	85.6	60.4	62.6	65.2	69.3	57.9	74.1
CSL [64]	90.2	85.5	54.6	75.3	70.4	73.5	77.6	90.8	86.1	86.7	69.6	68.0	73.8	71.1	68.9	76.1
CFA [65]	89.1	83.2	54.3	66.8	**81.2**	80.9	87.2	90.2	84.3	86.1	52.3	69.9	75.5	**80.8**	67.9	76.6
SASM [66]	89.5	**85.9**	57.7	78.4	79.8	84.2	89.3	90.9	58.8	**87.3**	63.8	67.8	78.7	79.4	69.4	77.3
YOLOX [50]	88.9	81.1	53.2	79.0	76.6	77.3	87.3	90.1	85.2	86.1	66.2	76.9	76.6	75.4	85.8	79.0
APOG [67]	89.9	85.5	**60.9**	**81.5**	78.7	85.3	88.8	90.9	**87.6**	70.5	**71.5**	**82.0**	77.4	74.5	80.6	80.3
DODet [68]	89.9	85.5	58	81.2	78.7	**85.5**	88.5	90.9	87.1	87.8	70.5	71.5	**82.0**	77.4	74.4	80.3
Ours	**91.3**	83.9	54.7	80.6	79.3	79.7	**90.7**	**91.2**	87.5	85.1	66.5	76.8	79.3	75.2	**87.0**	**80.5**

localization attention module, small object compensation strategy, and background separation strategy. Compared with the baseline, the mAPs are 1.1%, 1.2%, and 1.0% higher in NWPU VHR-10 dataset, the mAPs are 1.4%, 1.6%, and 1.4% higher in DIOR dataset, the mAPs are 0.8%, 0.9%, and 0.8% higher in DOTA dataset. ELAM is proposed to focus more on channel features that are effective for the detection task, it provides validity feature descriptions. SOCS is proposed to compensate in formation, which is effective for small object detection. BSS can separate the object from the complex background, reduce the interference of noise.

2) Comparison Results

We compared our model with other detectors on three datasets, NWPU VHR-10, DIOR and DOTA. The achievement of each category is evaluated by AP, and the overall performance on the dataset is measured by mAP over the entire dataset.

The experimental results of DIOR. We evaluated our model on DIOR dataset and compared it with other outstanding detectors, including RCNN [47], SSD [9], RAST-YOLO [52], Faster RCNN [48], YOLOX [50], SCFNet [51] and Retinanet [10], we can see the results in Table IV. Our model has a map of 70.7% across all categories of DIOR, which is significantly superior to other methods. Especially in small

TABLE VII
PERFORMANCE ANALYSIS OF THE MODELS ON NWPU VHR-10 DATASET

Method	Training time	Detection speed	GFLOPs	Params	mAP
F-RCNN [48]	4.9h	5.6fps	402.18	137.10M	76.4%
M-RCNN [53]	5.0h	5.1fps	354.16	63.73M	83.9%
Ours	1.1h	13.0fps	26.06	19.39M	94.2%

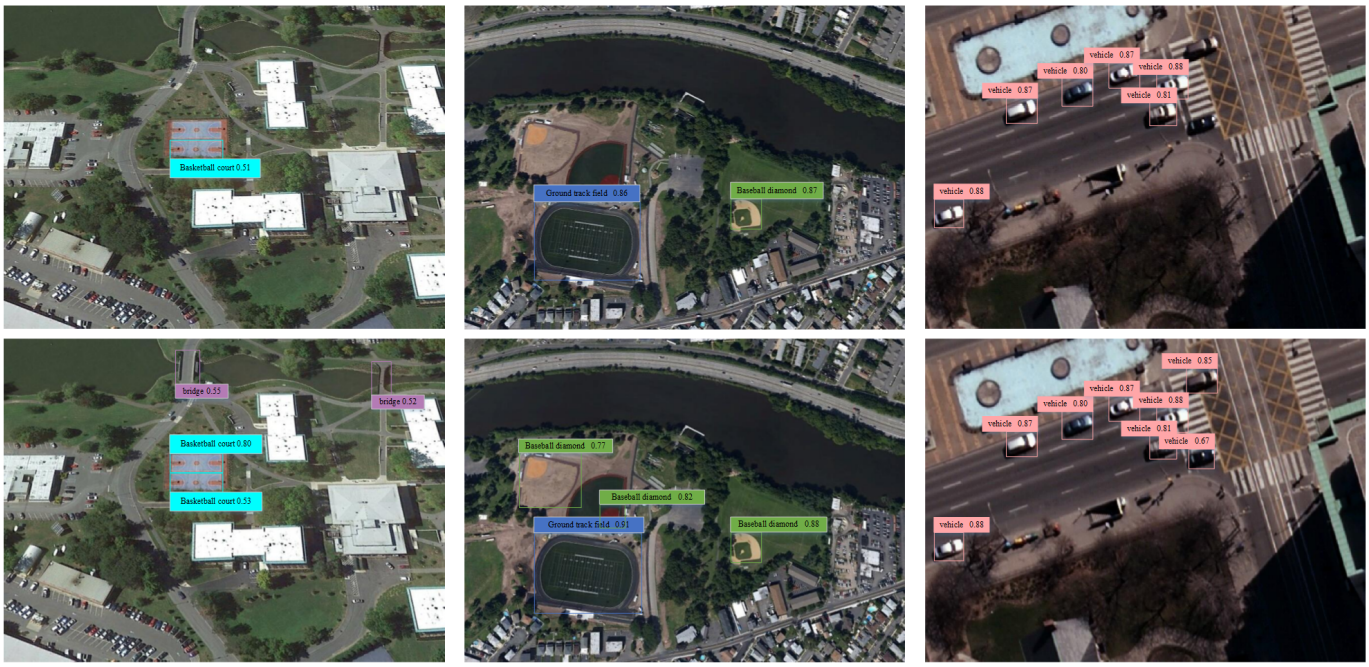

Fig. 7. Detection results by comparing our method against YOLOX on NWPU VHR-10. The top row of photos show the detection results of YOLOX, and the following are our detection results.

object categories, our results are the best. For example, the APs for airplane and vehicle categories are 87.4%, and 54.4%, respectively, indicating that our model can accurately detect small objects under limited information. It demonstrates the effectiveness of small object compensation strategy. Similarly, for large objects such as basketball court and chimney, the map values of the model are 89.1% and 77.6%. It is higher than all other methods. Our model can correctly locate small objects and some objects under complex background interference.

The experimental results of NWPU VHR-10. We also evaluated our method on NWPU VHR-10 and compared it with else methods like Faster R-CNN [48], Mask R-CNN [53], RFBNet300 [54], DNN [56] and EGAT-LSTM [59]. We can see the results on NWPU VHR-10 in Table V. The mAP of our model is 94.2%, followed by YOLOX and EGAT-LSTM. Especially on the airplane, we get the AP of 99.9% , which is difficult to achieve. Additionally, our detector has also made significant improvements on other types of objects, such as Basketball court, Bridge and Vehicle, which have reached 98.8%, 96.6% and 90.8% AP respectively. It can be seen that our model has a substantial performance improvement in the field of remote sensing object detection. In summary, this method has good detection performance among all detectors.

The experimental results of DOTA. Table VI reports the mAP of the method and other latest object detection models, including SCRDet [60], FAOD [61], RSDet [62], S2ANet [63], CSL [64], CFA [65], SASM [66], DODet [68], AOPG [67]. Bold values indicate that the corresponding model achieves the best detection performance in this type of object. It shows our method can effectively detect the objects and achieve the mAP of 80.5%. Analogously, our model could achieve optimal performance in 4/15 specific object categories.

We compared our training time, detection speed, GFLOPs, and parameter count with other models on NWPU VHR-10 datasets. The results are shown in Table VII. It can be seen that our method obtains the best accuracy and the fastest inference compared to other important models.And our algorithm has the optimal complexity.

V. DISCUSSION

In this article, we propose a new one stage detection network in complex remote sensing scenes. In response to the small objects, an effective localization attention module is proposed, which focuses more on channel features that are effective for detection tasks, providing effective feature descriptions for object detection. Its mAP achieved a 1.1% improvement on NWPU VHR-10 dataset. Homoplastically, a strategy of compensating for small objects is proposed, which effectively extracts contextual information while also compensating for identifying details and small object information, enhancing feature representation capabilities. Its mAP achieved a 1.2% improvement on NWPU VHR-10 dataset. Simultaneously, for

remote sensing objects in complex backgrounds, a background separation strategy is proposed, which can separate the foreground and background, enhance the features of the objects and reduce the noise. Its mAP achieved a 1.0% improvement on NWPU VHR-10 dataset.

The visualization result is shown in Fig. 7, which proves that our method is more friendly to small objects. More importantly, our model can detect objects that cannot be detected by other detection methods. It can be seen that our method can successfully detect targets similar to the background, such as bridge, vehicle, and baseball diamond, as well as small targets such as basketball court and vehicle. Our method has significantly improved the detection results on remote sensing object datasets, with mAPs of 94.2%, 70.7%, and 80.5% on the NWPU VHR-10, DIOR, and DOTA datasets, respectively.

VI. Conclusion

We propose a one-stage remote sensing object detection algorithm for small objects and complex backgrounds. This method uses the anchor-free detection algorithm YOLOX as the framework. We design an effective localization attention module, and embeds it into the backbone network to locate channel features. This operation is effective for inspection tasks and is well adaptable to multi-scale object sizes. Follow that, a small object compensation strategy is proposed to achieve the reuse and compensation of small object information, to reduce the loss of small object information. On this basis, a background separation strategy is also proposed to enhance the features of interest to suppress background noise. Under the same dataset setting, in comparison else methods, our method has achieved superior accuracy of remote sensing objects to a certain extent, which verifies the superiority of this method in complex background and small objects.

In the following research, we plan to apply the existing network to practical applications and model pruning and quantization of the model without compromising current performance.

References

[1] P. Wang, X. Sun, and W. Diao, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*,vol.58,no.5,pp.3377–3390, May.2020.

[2] G. Mattyus, "Near real-time automatic marine vessel detection on optical satellite images," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci., vol.XL-1/W1, no.1, pp.233–237,May.2008.*

[3] J. Leitloff, D.Rosenbaum,and F. Kurz, "An operational system for estimating road traffic information from aerial images," *Remote Sens.*,vol.6, no.11, pp.11315–11341, 2014.

[4] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*,vol.53, no.7, pp.3639–3657, Jul. 2015.

[5] 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.

[6] Y. Li, X. Huang, and H. Liu,"Unsupervised deep feature learning for urban village detection from high-resolution remote sensing images" *Photogramm. Eng. Remote Sens.*,vol.83, pp.567–579, Aug.2017.

[7] Y. Zhuang, L. Li, and H. Chen, "Small sample set inshore ship detection from VHR optical remote sensing images based on structured sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*,vol.13, pp.2145–2160, 2020.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[9] W. Liu, D. Anguelov, D.Erhan, C.Szegedy, Scott, C.Fu, and A.Berg, "SSD: Single Shot MultiBox Detector," in *Proc. Comput.Vis.(ECCV)*, Oct.2016, pp. 21-27.

[10] TY. Lin, P. Goyal, R.Girshick, and K.He, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[11] RJ. Chen, H. Mai, L. Luo, X. Chen, and K. Wu, "Effective Feature Fusion Network in BIFPN for Small Object Detection," in *International Conference on Image Processing (ICIP)*,Anchorage, AK, USA, 2021, pp. 699-703.

[12] 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, 2016.

[13] HV. Habi, H. Messer, and J. Han, "Recurrent neural network for rain estimation using commercial microwave links," *IEEE Trans. Geosci. Remote Sens.*,vol.59, no.5, pp.3672–3681, May,2021.

[14] D. Comite, and N. Pierdicca, "Decorrelation of the near-specular land scattering in bistatic radar systems," *IEEE Trans. Geosci. Remote Sens.*, vol.60, pp.1–13, 2022.

[15] X. Dong, Y. Qin, R. Fu, Y. Gao, S. Liu and Y. Ye, "Remote Sensing Object Detection Based on Gated Context-Aware Module," *IEEE Trans. Geosci. Remote Sens. Lett.*,vol.19, pp.1–5, 2022.

[16] P. Ma, and J.Che,"Remote Sensing Image Detection Based on Attention Mechanism and YOLOv5," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*,2022, pp.376-388.

[17] Y. Cheng, W. Wang, anf W. Zhang, "A Multi-Feature Fusion and Attention Network for Multi-Scale Object Detection in Remote Sensing Images," *Remote Sens.*, vol.15, no.8, pp.2096, 2023.

[18] K. Li, G.Wan, and G. Chang, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol.159, pp.296-307, 2020.

[19] G. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang,"DOTA: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,2018, pp.3974-3983.

[20] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang and Q. Li, "CDD-Net: A Context-Driven Detection Network for Multiclass Object Detection," *IEEE Geosci. Remote Sens. Lett.*,vol.19, pp.1–5, 2022.

[21] X. Dong, Y. Qin, R. Fu, and S. Liu, "Multiscale Deformable Attention and Multilevel Features Aggregation for Remote Sensing Object Detection," *IEEE Geosci. Remote Sens. Lett.*,vol.19, pp.1–5, 2022.

[22] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-Aware Distillation for Efficient Object Detection in Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol.61, pp.1–11, Jan.2023.

[23] X. Yang, X. Zhang, N. Wang, and X. Gao, "A Robust One-Stage Detector for Multiscale Ship Detection With Complex Background in Massive SAR Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–12,Nov.2023.

[24] J. Hu, X. Zhi, S. Jiang, H. Tang, W. Zhang, and L. Bruzzone, "Supervised Multi-Scale Attention-Guided Ship Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–14,Sep.2022.

[25] W. Ma, N. Li, H. Zhu, L. Jiao, and X. Tang, "Feature Split–Merge–Enhancement Network for Remote Sensing Object Detection," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–17,Jan.2022.

[26] Y. Guo, X. Tong, X. Xu, S. Liu, Y. Feng, and H. Xie, "An Anchor-Free Network With Density Map and Attention Mechanism for Multiscale Object Detection in Aerial Images," *IEEE Geosci. Remote Sens. Lett.*,vol.19, pp.1–5, 2022.

[27] C. Zhan, X. Duan, S. Xu, and Z. Song, "An improved UAV object detection algorithm based on ASFF-YOLOv5s," in *International Conference on Image and Graphics (ICIG)*, 2007, pp. 519-523.

[28] H. Wan, J. Chen, and Z. Huang, "AFSar: An Anchor-Free SAR Object Detection Algorithm Based on Multiscale Enhancement Representation Learning," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–14, 2022.

[29] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to Local: Clip-LSTM-Based Object Detection From Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–13, Mar.2021.

[30] T. Zhang, Y. Zhuang, G. Wang, and S. Dong, "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–18, 2021.

[31] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-Guided Feature Transform Hybrid Residual Network for Remote Sensing Object Detection," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–13, 2022.

[32] F. Shi, T. Zhang, and T. Zhang, "Orientation-Aware Vehicle Detection in Aerial Images via an Anchor-Free Object Detection Approach," *IEEE Trans. Geosci. Remote Sens.*,vol.59,no.6 pp.5221–5233, Jun. 2021.

[33] Q. Hu, S. Hu, and S. Liu, "BANet: A Balance Attention Network for Anchor-Free Ship Detection in SAR Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–12, 2022.

[34] X. Yang, J. Yang, J. Yan, and Y. Zhang, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8232–8241.

[35] Y. Yang, Z. Pan, Y. Hu, and C. Ding, "CPS-Det: An Anchor-Free Based Rotation Detector for Ship Detection," *Remote Sens.*, vol.13, no.11, 2021.

[36] Y. Dong, F. Chen, S. Han, and H. Liu, "Ship Object Detection of Remote Sensing Image Based on Visual Attention," *Remote Sens.*, vol.13, no.16, 2021.

[37] P. Wang, Y. Niu, J. Wang, F. Ma, and C. Zhang, "Arbitrarily Oriented Dense Object Detection Based on Center Point Network in Remote Sensing Images," *Remote Sens.*, vol.14, no.7, 2022.

[38] X. Jiang, H. Xie, and J. Chen, "Arbitrarily-Oriented Dense Object Detection Based on Center Point Network in Remote Sensing Images," *Remote Sens.*, vol.15, no.3,pp.673, 2023.

[39] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 13713-13722.

[40] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," 2021, arXiv:2111.12419.

[41] W. Zhao, Y. Kang, H. Chen, and Z. Zhao, "Adaptively Attentional Feature Fusion Oriented to Multiscale Object Detection in Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*,vol.72, pp.1–11, 2023.

[42] B. Song, P. Liu, J. Li, L. Wang, and L. Zhang, "MLFF-GAN: A Multilevel Feature Fusion With GAN for Spatiotemporal Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–16, 2022.

[43] X. Wang, S. Wang, C. Ning, and H. Zhou, "Enhanced Feature Pyramid Network With Deep Semantic Embedding for Remote Sensing Scene Classification," *IEEE Trans. Geosci. Remote Sens.*,vol.59,no.9. pp.7918–7932, Sep. 2021.

[44] Y. Zhao, L. Zhao, Z. Liu, D. Hu, G. Kuang, and L. Liu, "Attentional feature refinement and alignment network for aircraft detection in SAR imagery," 2022, arXiv:2201.07124.

[45] Z. Wang, J. Guo, L. Zeng, and C. Zhang, "MLFFNet: Multilevel Feature Fusion Network for Object Detection in Sonar Images," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–19, 2022.

[46] R. Chen, Z. Cai, and W. Cao, "MFFN: An Underwater Sensing Scene Image Enhancement Method Based on Multiscale Feature Fusion Network," *IEEE Trans. Geosci. Remote Sens.*,vol.60, pp.1–12, 2022.

[47] R. Girshick, J. Donahue, and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580-587.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Faster r-cnn: Towards real-time object detection with region proposal networks*, vol.28, 2015.

[49] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *IEEE Int. Conf.Comput. Vis.(ICCV)*, 2019, pp. 6569-6578.

[50] Z. Ge, S. Liu, and F. Wang, "YOLOX: Exceeding yolo series in 2021," 2021, arXiv:2107.08430.

[51] C. Yue, J. Yan, Y. Zhang, Z. Luo, Y. Liu, and P. Guo, "Multi-level learning counting via pyramid vision transformer and CNN," *Expert Systems with Applications*, vol.224, pp.119980.2023.

[52] X. Jiang, and Y. Wu, "Remote Sensing Object Detection Based on Convolution and Swin Transformer," *IEEE Access*, 2023.

[53] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE Int. Conf.Comput. Vis.(ICCV)*, Sep.2017, pp. 2980-2988.

[54] S. Liu, and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Comput.Vis.(ECCV)*, Oct.2018, pp. 385-400.

[55] A. Bochkovskiy, CY. Wang, HYM. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, arXiv:2004.10934.

[56] S. Jiang, W. Yao, MS. Wong, G. Li, and Z. Hong, "An optimized deep neural network detecting small and narrow rectangular objects in Google Earth Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*,vol.13,pp.1068-1081. 2023.

[57] 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.

[58] Q. Lin, J. Zhao, B. Du, G. Fu,and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples, " *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[59] S. Tian, L. Kang, X. Xing, J. Tian, C. Fan, and Y. Zhang, "A relation-augmented embedded graph attention network for remote sensing object detection, " *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[60] X. Yang, J. Yang, J. Yan, and Y. Zhang, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *IEEE Int. Conf.Comput. Vis.(ICCV)*, 2019, pp. 8232-8241.

[61] C. Li, C. Xu, Z. Cui, D. Wang, and T. Zhang, "Feature-attentioned object detection in remote sensing imagery," in *International Conference on Image Processing (ICIP)*,Taipei, Taiwan. 2021, pp. 3886-3890.

[62] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.35, no.3. 2021, pp.2458-2466.

[63] J. Han, J. Ding, J. Li, and G. -S. Xia, "Align Deep Features for Oriented Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-11, 2022.

[64] X. Yang, and J. Yan, "Arbitrary-Oriented Object Detection with Circular Smooth Label," in *Proc. Comput.Vis.(ECCV)*, Glasgow, UK, Aug. 2020, pp.677-694.

[65] Z. Guo, C. Liu, X. Zhang, J. Jiao, and X. Ji, "Beyond Bounding-Box: Convex-hull Feature Adaptation for Oriented and Densely Packed Object Detection," in *Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp.8792-8801.

[66] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.36, no.1. 2022, pp.923-932.

[67] G. Cheng, J. Wang, K. Li, X. Xie, and C. Lang, "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-11, 2022.

[68] G. Cheng, Y. Yao, S. Li, and K. Li, "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-11, 2022.

Yan Dong received the B.S. and M.S. degrees from the School of electrical engineering, Zhengzhou University, China. She is currently an Associate Professor in the School of Electric and Information Engineering, Zhongyuan University of Technology, China. In recent years, she has published more than 30 academic papers, authorized three patents and published one book.

Her research interests include artificial intelligence, pattern recognition and surface defect detection based on machine vision.

Haotian Yang received the B.S. degree from Heilongjiang University of Science and Technology, Harbin, China. He is currently pursuing the M.S. degree with Zhongyuan University of Technology, Zhengzhou, China.

His research interests include remote sensing, computer vision, and machine learning.

Shanliang Liu received the B.S. and M.S. degrees in signal and information processing from the School of Electronic and Information Engineering, Zhongyuan University of Technology, Zhengzhou, China, in 2016 and 2019, respectively. He received the Ph.D. degree at the Civil Aviation University of China.

He is currently teaching at Zhongyuan University of Technology. He focuses on deep learning, image processing, and object detection.

Guangshuai Gao received the B.S. and M.S. degree in applied physics and signal and information processing from the Zhongyuan University of Technology, Zhengzhou, China, in 2014 and 2017, respectively, and the Ph.D. degree in computer science from School of Computer Science and Engineering, Beihang University, Beijing, China, in 2022.

He is currently a Lecturer with the School of Electronics and Information, Zhongyuan University of Technology. His research interests include image processing, digital machine learning, and remote sensing imagery interpretation.

Chunlei Li received the M.S. degree from Hohai University, Nanjing, China, in 2004, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012. He is currently a Professor with the School of Electronics and Information, Zhongyuan University of Technology, Zhengzhou. In recent years, he has authored or coauthored more than 50 technical articles and has authored 2 books. His research interests include computer vision and pattern recognition.