IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 January 2024, accepted 11 February 2024, date of publication 22 February 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368877

RESEARCH ARTICLE

Research on Optical Remote Sensing Image Target Detection Technique Based on DCH-YOLOv7 Algorithm

CHUNHUI CUI[1], RUGANG WANG[1], YUANYUAN WANG[1], FENG ZHOU[1], XUESHENG BIAN[1], AND JUN CHEN[2]

[1]School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu 224051, China
[2]Yancheng Electric Power Design Institute Ltd., Yancheng, Jiangsu 224006, China

Corresponding author: Rugang Wang (wrg3506@ycit.edu.cn)

This work was supported in part by Jiangsu Graduate Practical Innovation Project under Grant SJCX22_1685, in part by the Natural Science Foundation of China under Grant 62301473, in part by the Major Project of Natural Science Research of Jiangsu Province Colleges and Universities under Grant 19KJA110002, and in part by the Natural Science Research Project of Jiangsu University under Grant 18KJD510010.

ABSTRACT Aiming at the YOLO (You Only Look Once) algorithm's low detection accuracy caused by the complex background environment and large target scale difference in the detection of optical remote sensing images, the Deformable Convolutional Fusion Attention mechanism based DCH-YOLOv7 (Deformable Convolutional Hybrid-YOLOv7) target detection algorithm is proposed in this paper. In this algorithm, deformable convolution is introduced in order to meet the detection of optical remote sensing images with different scale, and at the same time, two modules, PELAN and PMP, are added to effectively improve the network's ability to accurately localize the target features; secondly, a hybrid attention module (ACmix) is used, which effectively enhances the network's sensitivity to the small targets and improves the detection accuracy; lastly, the CIoU loss function is replaced by the WIoU loss function, which, through the adjustment of the weights, improves the detection accuracy of the high-quality anchor frames, and reduces the probability of missed and false detection. Finally, experiments were conducted on publicly available datasets, namely DIOR. Experimental results indicate that the DCH-YOLOv7 algorithm achieved an impressive detection accuracy of 90.6% in mAP@0.5, demonstrating a remarkable improvement of 3.1% over YOLOv7. These results demonstrate that DCH-YOLOv7 algorithm has a certain improvement in the effectiveness of target detection in optical remote sensing imagery, and can better cope with the problems of the dense distribution of small targets, the large differences in target scales, and the complex background.

INDEX TERMS Target detection, deformable convolution, attention mechanism, loss function.

I. INTRODUCTION

Remote sensing image target detection has a wide range of applications [1], including geological exploration, intelligence reconnaissance and urban planning and other fields. Different from traditional natural images, target information in remote sensing images presents fragmented distribution and complex and variable background [2]. This characteristic leads to a large amount of interference information on the feature map, while the dense distribution of individual feature targets further increases the difficulty of target detection in remote sensing images. In comparison, traditional algorithms are less effective in target detection, with low accuracy and easy to miss under complex conditions [3].

With the continuous development of artificial intelligence technology, especially the application of deep learning, researchers have made important breakthroughs in the field of remote sensing image target detection [4], [5]. However, there are still many difficulties in the recognition process of remote sensing images. On the one hand, the background of

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

remote sensing images is complex. Due to the relatively high shooting height of remote sensing images and the complex spatial and geographical environment, the phenomenon of dense occlusion is more common in remote sensing images, which leads to incomplete and discontinuous phenomenon in target detection of remote sensing images. To address this problem, foreground and background segmentation is usually adopted to reduce the background noise and background segmentation to reduce the background noise and highlight the target information, in order to improve the target detection accuracy and generalization ability [6]. On the other hand, remote sensing images have small and clustered targets to be detected. Remote sensing images are taken farther away from the ground, and the targets to be detected are smaller in size and have fewer available features compared to remote sensing images. Therefore, it is easy to miss or misdetect in detection. To address this problem, feature fusion of high-level semantic information and low-level semantic information is usually used to improve the detection ability of small targets [7].

Currently, the commonly used target detection algorithms mainly include one-stage target detection algorithms and two-stage target detection algorithms. Among them, the one-stage target detection algorithms are the detection algorithms represented by YOLO series [8] and SSD [9], which use regression strategy to realize target detection. The two-stage target detection algorithms, on the other hand, are target detection algorithms represented by algorithms such as R-CNN [10], Fast R-CNN [11], and Faster R-CNN [12]. This class of algorithms generates a fitted region based on the original image and utilizes convolutional neural networks for feature extraction to achieve target classification and detection. From the existing research results, the single-stage target detection algorithm does not need to generate a large number of candidate regions. This reduces the detection time and improves the real-time target detection. And it has certain advantages in engineering applications and has achieved better research progress. For example, in 2016, in response to problems such as the slow detection speed of existing target detection, Redmom et al. proposed the YOLOv1 algorithm that allows real-time detection. The core idea of the algorithm is to transform target detection into a regression problem, using the whole image as the input to the network, and then using a convolutional neural network to obtain the location of the bounding box and the category it belongs to. However, it is difficult to predict targets in complex backgrounds and irregular or different sized objects. To further improve the detection accuracy of the algorithm, in 2017, Redmom et al. proposed YOLOv2 [13] based on YOLOv1. The previous YOLOv1 algorithm uses a fully connected layer to obtain the location information of the bounding box, which has poor localization robustness. However, the YOLOv2 network model uses the anchor frame mechanism of the fast RCNN, which can cluster and analyze the target bounding boxes of the training set to find the appropriate size and ratio of the anchor frames and improve the detection accuracy. Subsequently, Redmon et al. proposed the YOLOv3 [14] model in 2018. The model obtains the dimensions of the three prior frames and the ratio of the three prior frames through dimensional clustering, and employs independent logistic regression instead of the original softmax function to support multi-label prediction. Although the YOLOv3 model has improved the detection accuracy compared to previous models, it is still difficult to detect targets in complex backgrounds. In 2020, Bochkovskiy et al. proposed the YOLOv4 real-time detection model by replacing the backbone network with the CSPDarkNet53 network [15] and adding the SPP [16], Feature Pyramid Network (FPN) module [17] and PAN (Path Aggregation Network) module [18] to improve the detection accuracy of the network, but the robustness of the YOLOv4 network model to light changes, occlusion and complex backgrounds is relatively low. In these complex scenarios, misdetections or missed detections may occur. From the research results at this stage, based on the YOLO target detection algorithm, by introducing the feature pyramid structure and multi-scale prediction head, it is able to detect targets of different scales with strong adaptability. However, it is prone to target overlapping and missed detection, so the detection accuracy of the model for dense targets needs to be further improved.

With the continuous development of deep learning algorithms, more and more algorithms have been applied to detect optical remote sensing images. Guo et al. [19] proposed a unified convolutional neural network (CNN) by fusing a multiscale target suggestion network and a multiscale target detection network, which improves the detection accuracy of small targets in high-resolution satellite images. Jiang et al. [20] effectively solved the problem of too narrow bounding boxes for small targets in remote sensing images by combining a dual-lens neural network combined with a staggered localization strategy to effectively solve the problem of overly narrow bounding boxes of small targets in remote sensing images. Yao et al. [21] generated high-quality semantic features by introducing an extended bottleneck structure in the backbone network, thus significantly improving the prediction ability of multi-scale objects. Zhang et al. [22] generated multiple by introducing a new activation function in the backbone network and fusing different layers of features, Zhang et al. generate multiple receptive field features, which effectively improve the detection accuracy of dense small targets in remote sensing images. Yang et al. [23] propose a multi-task rotating region convolutional neural network-based detection model by fusing multilayered features with effective anchor sampling, which improves the accuracy of detecting ships with arbitrary orientations in remote sensing images. Therefore, how to more accurately extract information from remote sensing images under complex backgrounds has become an important research topic in the current field of remote sensing images.

In order to improve the detection accuracy of targets in complex backgrounds in optical remote sensing images, this paper proposes an optimization algorithm based on deformable convolutional fusion attention mechanism for

YOLOv7 [24]. First, a deformable convolution (DCNv2 [25]) mechanism is used in the backbone feature extraction network. This mechanism enables the network model to better adapt to targets with different shapes in optical remote sensing images. And based on this, PELAN and PMP modules are added to make the target localization more accurate. Secondly, the ACmix [26] attention mechanism is introduced into the YOLOv7 network model to suppress the interference of complex background and noise, thus enhancing the network's ability to extract target features under complex background. Finally, the WIoU loss function [27] is utilized as the loss function of bounding box regression as an alternative to the original CIoU loss function [28], which helps to accelerate the convergence speed of the training process, reduces the probability of misdetection and omission, and improves the detection accuracy of the model.

II. BASIC MODEL OF THE YOLO ALGORITHM

YOLOv7 is one of the more advanced target detection algorithms, which integrates strategies such as cascade model-based model scaling, Extended Efficient Long-Range Attention Network (E-ELAN), and convolutional reparameterization [29]. Compared to other known target detection algorithms, YOLOv7 exhibits higher speed and accuracy over a speed range of 5 to 160 frames per second. It succeeds in achieving a good balance between detection speed and accuracy and is therefore widely used in scenarios where small devices are detected in real time. The YOLOv7 network is composed of four parts: the input (Input), the backbone network (Backbone), the neck (Neck), and the detection head (Head), as shown in Figure 1.

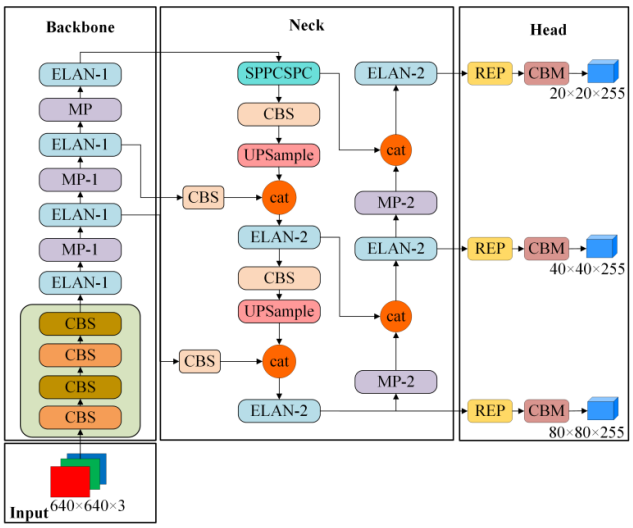


FIGURE 1. Schematic diagram of the YOLO algorithm framework.

After the input module (Input) of YOLOv7 performs a series of, e.g., Mosaic data enhancement algorithms on the input images, the images are uniformly adjusted to the default size (640 × 640 × 3) to meet the input requirements of the backbone network, further enhancing the generalization ability and robustness of the model, so that it can be better adapted to different scenes and changes [30].

The role of the backbone network (Backbone) in YOLOv7 is to extract the feature information in the image and prepare the data for subsequent feature fusion and target detection. It mainly consists of CBS module, ELAN module and MP module. These modules and the way they are connected to each other are designed to efficiently capture semantic information at different scales in an image and provide rich feature representations that can help the YOLOv7 network have better target detection performance and accuracy. Among them, the CBS module consists of a convolutional layer, a batch normalization layer and an activation function layer, which plays the role of feature extraction and channel transformation in YOLOv7. The ELAN module splices the feature map through branches of different depths, controlling the shortest and longest gradient paths, allowing deeper layers of the network to learn and converge efficiently. The MP module is mainly used for downsampling, which fuses the feature maps obtained through the maximum pooling downsampling branch and the convolutional downsampling branch in a way that preserves as much feature information as possible without adding extra computation. These modules play an important role in the backbone network and help to improve the performance and accuracy of target detection.

Neck (Neck) plays the role of feature fusion in target detection and it is mainly composed of Path Aggregation Feature Pyramid Network (PAFPN). As the network extracts and abstracts more image features, the semantic information in the feature map gradually becomes apparent, while the location information may become less precise. The main function of PAFPN is to fully fuse the precise location information at the bottom layer and the abstract semantic information at the top layer, so that the semantic information and the location information in different levels of feature maps can be fully fused. This can further improve the model's accuracy in localizing multi-sized targets, especially in identifying small targets in a large context.

The detection head (Head), on the other hand, uses three different sizes of feature map branches output from the neck for multiscale prediction. In order to speed up the inference of the model, a reparameterization block (RepVGG Block, REP) is used for acceleration. This module reduces the amount of model calculations and improves the efficiency of reasoning. With this structural design, the model can better perform the target detection task and achieve high accuracy prediction at different scales.

III. MODELING ANALYSIS OF THE DCH-YOLOv7 ALGORITHM

Despite the remarkable success of the YOLOv7 algorithm in the field of general-purpose target detection (e.g., vehicle and pedestrian detection) [31], however, there are still some challenges in applying it directly to optical remote sensing image target detection. For example, the problem of the complex background of optical remote sensing images,

in the complex environment, remote sensing image targets and the surrounding environment is difficult to clearly distinguish [32]; the wide range of scale changes of the target in the remote sensing image, so that the small targets in the large background show sparse distribution; compared to the general detection of the image, optical remote sensing images are mostly from the satellite or aircraft overhead view, direct application of the horizontal detection frame of the YOLOv7 algorithm will result in the loss of the orientation information of the remote sensing image target, which reduces the localization accuracy, and even leads to the problem of the omission of the dense target in the post-processing [33]. In order to solve the above difficulties, this paper proposes a DCH-YOLOv7 optimization algorithm that introduces deformable convolution and fused attention mechanism, and the overall structure of the algorithm is shown in Figure 2. From the figure, it can be seen that deformable convolution is embedded in the last MP and ELAN modules in the backbone feature extraction network (Backbone) part and named as PMP module and PELAN module, which improves the model's ability to adapt to the target deformation and spatial variations, and enhances the model's ability to perceive the target boundaries and details, and also introduces the ACmix attentional mechanism in the Neck part, which improves spatial correlations, emphasizes the important regions, suppresses the noises and interferences as well as improves the details of the boundaries, so as to increase the model's accuracy of detecting the optical remote sensing images.

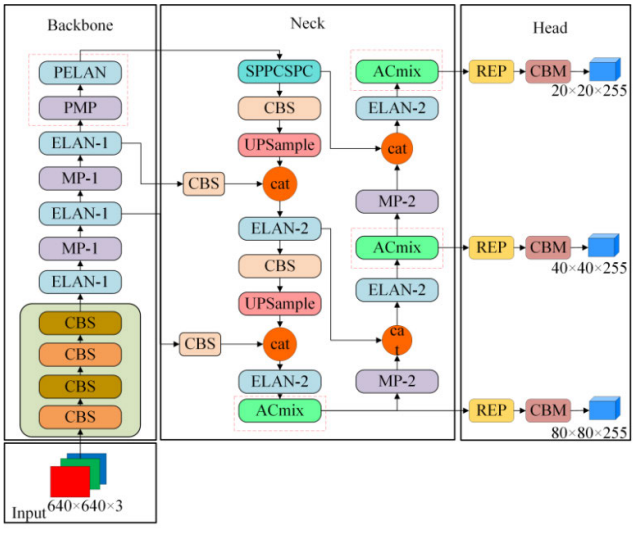


FIGURE 2. Network structure of DCH-YOLOv7 algorithm.

A. ELAN MODULES BASED ON DEFORMABLE CONVOLUTIONS

In order to solve the problem of complex background and difficulty in accurately distinguishing the target from its surroundings in remote sensing images, as well as the challenge of difficulty in accurate recognition due to the wide range of variation in the target scale, this paper proposes an improved method. The method embeds a deformable convolution module with offset learning capability in the ELAN and MP modules of the backbone network by dynamically adjusting the shape and size of the sensory fields sampled by the convolution to adapt to targets of different scales. The structure of the deformable convolution-based DCNS module is shown in Figure 3.

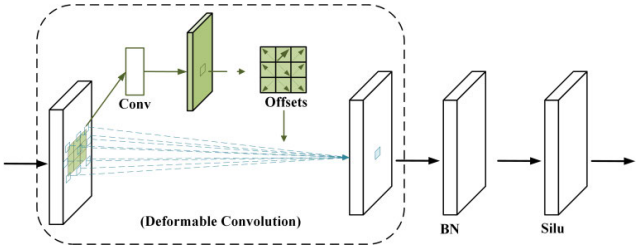


FIGURE 3. Deformable convolution based DCNS module, in which deformable convolution is used to better adapt to targets of different scales, BN is used for normalization and Silu activation function is introduced to further improve the performance of the model.

The traditional convolution operation is a regular convolution, which can only process fixed-size samples. Deformable convolution, on the other hand, has a more flexible receptive field and can adaptively change the size and shape of the receptive field according to the different shapes and sizes of the target objects in the remote sensing image. Deformable convolution interacts with the local or global context by introducing offsets, and has the ability to model long distances to capture a wider range and more complex features. At the same time, deformable convolution has a similar capability of adaptive spatial aggregation. This adaptivity allows deformable convolution to better adapt to different target objects and extract more accurate feature information in different scenarios. In this paper, deformable convolution is used and a multigroup mechanism is introduced to enhance the expressive power of the operator, while the method reduces the complexity of the algorithm by sharing convolutional weights and improves the stability of the training process by normalizing the modulation scalar. In remote sensing image target detection tasks, since the scale and shape of the target may change significantly, this can cause covariate bias problems within the network. As shown in Figure 3, the DCNS module proposed in this paper employs a BN for normalization in order to normalize each small batch of inputs. This approach makes the network more robust to changes in the input data, thus improving the generalization ability of the model. At the same time, the DCNS module also effectively reduces the problem of internal covariate bias during training, making it easier for the network to learn a consistent feature representation of the target. In the DCNS module, an added Silu activation function is introduced to better capture the complex relationships between features. This activation function helps to improve the expressive and fitting ability of the network, which further improves the performance of the model. The improved ELAN module is shown in Figure 4.

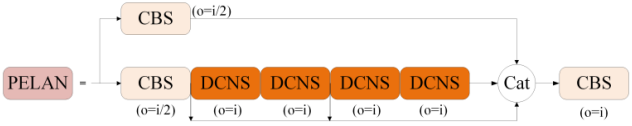


FIGURE 4. Deformable convolution based PELAN module.

B. MP MODULES BASED ON DEFORMABLE CONVOLUTION

In remote sensing image target detection, the use of maximum pooling can lead to some information loss and blurring due to the complex background, thus affecting the detection accuracy [2]. However, by embedding a deformable convolutional layer, the size and position of the sensing field can be adaptively adjusted during the convolution process, allowing the sampling positions around each position to adaptively change as well. In this way, the deformable convolutional layer can better adapt to the shape and size of the target, thus improving the overall detection accuracy, and the improved MP module is shown in Figure 5.

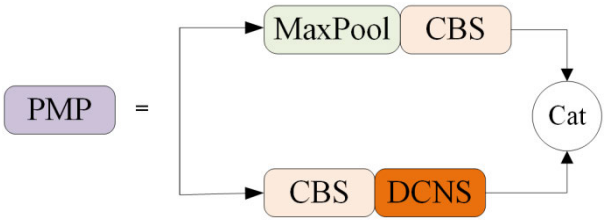


FIGURE 5. Deformable convolution based PMP module.

C. INTEGRATION OF THE ACMIX ATTENTION MECHANISM

In optical remote sensing imagery, especially in complex backgrounds such as nearshore, sea and land boundaries are often mixed together, making it difficult to distinguish ship targets from the background. This leads to the difficulty for the model to extract the target portion of interest from the whole image. To solve this problem, researchers have proposed a variety of attention mechanisms. By introducing the attention mechanism, the model is able to automatically learn which regions are more important for target detection and allocate more attention and computational resources to these regions. In this way, the model is able to better understand complex scenes and accurately extract relevant features of the target from the image.

In this paper, we introduce the ACmix attention mechanism into the YOLOv7 network, which combines the two parts of convolution and self-attention to improve the network's attention to small and medium-sized targets in remote sensing images. Its principle is shown in Figure 6.

First, the input features are projected using three 1×1 convolutions to divide them into N parts, resulting in 3N mapped intermediate features. In the first branch, the convolution operation is used to obtain the feature information of the local receptive field. The intermediate features are shifted and aggregated after passing through the fully connected layer,

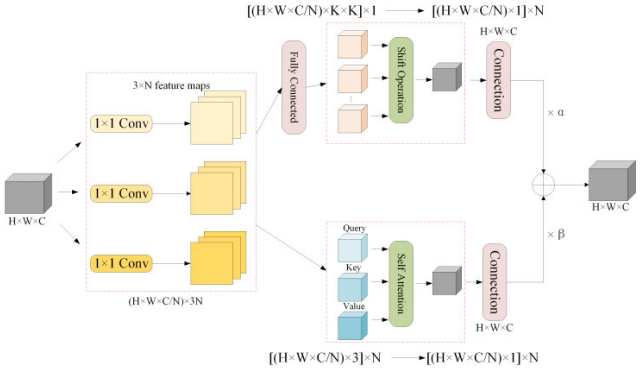


FIGURE 6. ACmix module.

and then, the input features are convolved to obtain features of size $H \times W \times C$. In the second branch, the self-attention mechanism is used to obtain the global receptive field and more attention is paid to the important regions. Here the 3N intermediate features correspond to three feature maps, which are Query, Key and Value. These features follow the principle of multiple self-attention modules, and after convolutional processing the features of size $H \times W \times C$ are obtained. Finally, the outputs of the above two branches are summed and two learnable scalars are used to control the weights between them. As shown in equation (1).

$$F_{out} = \alpha F_{att} + \beta F_{conv} \tag{1}$$

In the ACmix module, the final output is labeled as Fout. Meanwhile,F_{att} and F_{conv} denote the outputs on the self-attention path and the convolutional path, respectively. In this case, the values of both α and β are set to 1. ACmix integrates convolutional and self-attention modules and applies them to the neck (Neck) part of the YOLOv7 network, which utilizes the attention mechanism to weight the prediction results of different bounding boxes. By dynamically adjusting the importance of different bounding boxes based on information such as the feature representation or confidence level of each bounding box, ACmix is able to increase the attention to small targets and reduce the possibility of the network model to produce a missed detection situation when detecting small targets. This in turn improves the overall detection accuracy.

D. IMPROVEMENT OF THE LOSS FUNCTION

The loss function is used to measure the difference between the predicted results and the actual labels, and a good loss function can accelerate the convergence of the network and improve its accuracy. The original YOLOv7 model uses a CIoU loss function, which takes into account the effect of the aspect ratio between the predicted frame and the real frame, the centroid distance, and the overlap area on the loss function. However, for some anchor frames with low labeling quality, when the aspect ratio between the height and width of the predicted and real frames is close to a linear proportion, the relative proportion penalty in the CIoU [22] loss function

degenerates to 0 and no longer works, and the CIoU loss function does not achieve convergence well. The relevant formula for CIoU is as follows:

$$L_{CIoU} = L_{IoU} + \frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2} + \alpha v \quad (2)$$

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{wh + w_{gt}h_{gt} - W_i H_i} \quad (3)$$

$$\alpha = \frac{v}{L_{IoU} + v} \quad (4)$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w}{h} - \arctan\frac{w_{gt}}{h_{gt}})^2 \quad (5)$$

In the CIoU loss function, W_i and H_i denote the dimensions of the overlapping parts of the two frames, W_g and H_g denote the dimensions of the smallest enclosing frame, is the weight coefficient, and is used to compute the similarity of the two frames in terms of aspect ratio metrics.

As can be seen from the above formula, the CIoU loss function mainly considers three factors, namely, the overlap area between the two frames, the distance between the centroids and the aspect ratio, which effectively improves the convergence speed of the loss function. However, when the aspect ratios of the anchor and target frames are the same (i.e. w:h = w_{gt}: h_{gt}), the penalty term for the aspect ratio will lose its effect. Loss functions such as CIoU, GIoU [34], and DIoU [35] all aim to enhance the fitting ability of the bounding box, which implies that they require high labeling quality for the training set. However, in the actual training set, the labeling quality of target objects may be inconsistent, especially for the labeling of small target objects, there are cases of poor labeling. If we blindly emphasize the regression of the bounding box on low-quality targets, we may damage the detection performance of the model. In order to improve the localization ability of the model, this paper introduces the WIoUv1 loss function, which is designed based on the dynamic non-monotonic focusing mechanism. The dynamic non-monotonic focusing mechanism uses "outliers" instead of the traditional IoU to evaluate the quality of the anchor frames, and adopts a more efficient gradient gain allocation strategy, which makes the model pay more attention to the low-quality anchor frames. In this way, the model can localize the target more accurately, which improves the performance of the model. The relevant formula of the WIoUv1 loss function is as follows:

$$L_{IoU} = 1 - IoU \quad (6)$$

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \quad (7)$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (8)$$

where W_g and H_g denote the dimensions of the minimum enclosing box. With superscript $*$ indicates that by separating W_g, H_g from the computational graph and changing them from variables to constants, the R_{WIoU} can be prevented from generating gradients that prevent convergence. According to Eq. (8), we can get the value range of R_{WIoU} as [1, e). This item amplifies the L_{IoU} of normal quality anchor frames by an attention-based approach. Whereas, according to Eq. (3), L_{IoU} takes values in the range of [0, 1]. L_{IoU} reduces the influence of R_{WIoU} on the results when the anchor frame is of high quality and reduces the attention on the centroid distance when the anchor frame overlaps with the target frame better. Compared with the traditional CIoU method, WIoUv1 can effectively avoid the over-penalization of the network by the geometric terms, which enhances the generalization ability of the algorithm. This means that we can reduce the requirement on the quality of dataset labeling and further improve the convergence speed and accuracy of the model.

IV. EXPERIMENTS AND ANALYSIS OF RESULTS

A. DATA SETS AND EVALUATION INDICATORS

To verify the effectiveness of the improved YOLOv7 algorithm in detecting remote sensing images, this paper designs experiments and performs training and evaluation on the publicly available dataset DIOR [36]. The DIOR dataset is a widely used remote sensing target detection dataset provided by Northwestern Polytechnical University. The dataset contains remotely sensed images from different regions and scenes and is intended to provide a more diverse sample of targets for detection. The DIOR dataset contains a total of 23,463 images of 800 pixels × 800 pixels size. Each image is finely labeled with the location and category information of the target object in the image. The dataset covers 20 different target categories including airplanes, airports, baseball fields, basketball courts, bridges, chimneys, dams, highway service areas, highway toll booths, golf courses, athletic fields, harbors, overpasses, boats, stadiums, storage tanks, tennis courts, train stations, vehicles, and windmills. These categories are denoted by C1-C20, respectively. In order to ensure the validity and reliability of the dataset, a strict division of the DIOR dataset is used in this paper. Specifically, the training set contains 5,862 images, the validation set contains 5,863 images, and the test set contains the remaining 11,738 images. Such a division is intended to ensure that the training, validation, and test data are similarly distributed so that the performance of the remote sensing target detection algorithm can be accurately evaluated. Figure 7 presents some images in the DIOR datasets.

B. EXPERIMENTAL SETUP

The optimized algorithmic model in this paper is based on Pytorch 1.8.0 framework, Python version 3.7, the software environment for the experiment is Windows 11, CUDA 11.1; the hardware environment is CPU: Intel Core i7 -12700K, 32 G of RAM; GPU: NVIDIA GeForce RTX 3080, with a video memory of 10 G; the following parameters are used for training, the number of training rounds is set to 300, the batch size is set to 16, and the input image size is 640 × 640 pixels, and the specific parameters are listed in Table 1.

FIGURE 7. Images from the DIOR datasets.

TABLE 1. Experimental parameter setting.

Paramenter	Value
Momentum	0.937
Learning_rate	0.002
Epoch	300
Batch_size	16

C. ANALYSIS OF EXPERIMENTAL RESULTS

In order to verify whether the algorithm in this paper is effective or not, in this section, we mainly focus on the performance of the YOLOv7 optimization algorithm based on deformable convolutional fusion attention mechanism proposed in this paper compared with the performance of target detection algorithms based on convolutional neural networks in recent years, and from the comparison results, we can see that the target detection algorithm proposed in this paper has a better performance on small target detection.

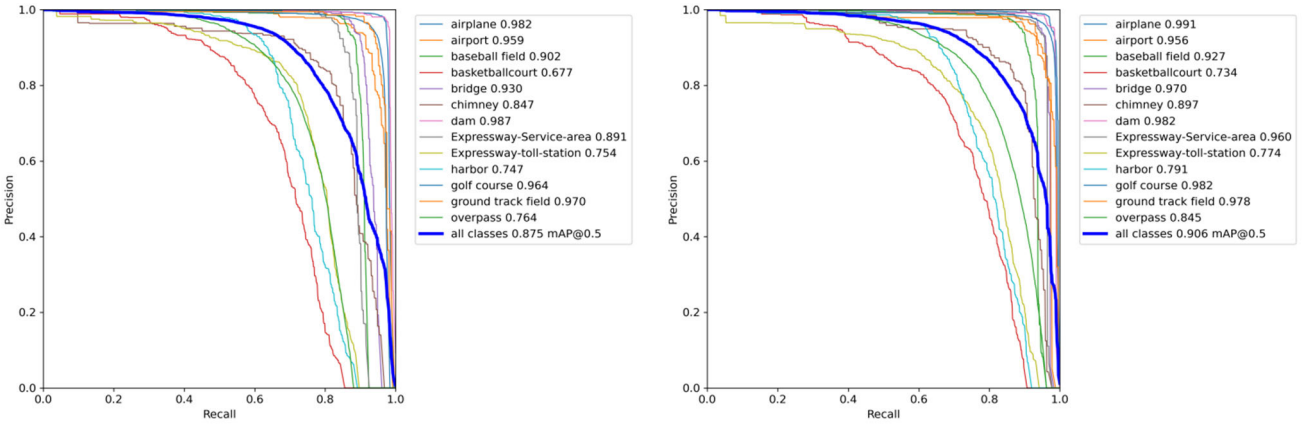
1) PERFORMANCE COMPARISON ON THE DIOR DATASET

In order to evaluate the performance of the proposed model, the previously mentioned training parameters and datasets were used. In the improved YOLOv7 network, the officially provided data enhancement method was used and the image size was effectively resized to 640×640 and 1280×1280 for validation in the input network structure. After validation, the image size of 640×640 was finally selected as more effective and was used as the selected image size for this paper. Then several experiments were conducted on the improved YOLOv7 model and all the experimental results converged and good precision and recall were obtained.

In order to verify the effects of PELAN module, PMP module, ACmix attention mechanism and WIoU loss function on the target detection performance of remote sensing images, this paper conducts ablation experiments on the DIOR dataset to validate the effectiveness of each of the improvement methods proposed in this paper. The results of the ablation experiments are shown in Table 1. Table 1 mAP (mean Average Precision) that is, the mean average precision is the average of all kinds of average precision values detected under the premise of the intersection ratio of positive and negative sample regions is 0.5, Fps indicates the number of images that the algorithm can process per second, which is not only related to the algorithm model, but also the hardware configuration of the experiment. Several experiments were conducted on the DCH-YOLOv7 model and all the experiments converged and good precision and recall were obtained. Compared with the original YOLOv7 model, the mAP (IoU of 0.5) improves from 87.5% to 90.6%, which is a 3.1% improvement. The PR curves before and after the improvement are shown in Figure 8. As can be seen from Table 2, the mAP value of the improved YOLOv7 network model exceeds that of the other combinations when inputting images of the same size, which effectively improves the detection accuracy.

To verify the effectiveness and authenticity of the proposed DCH-YOLOv7 algorithm, ablation experiments were conducted on the DIOR datasets. These experiments evaluated the impact of various modules and techniques on the overall performance of the algorithm. The experimental results are listed in Table 2,where the symbol "$\surd$" indicates the use of a specific moduleor technique.

From Table 2, we can see that the first group is the original Yolov7 algorithm with 87.5% mAP and 37.2M parameters; the second group is replacing part of the ELAN module with the PELAN module, with 87.5% mAP and 32.7M parameters. Although the mAP is not improved, these changes bring about a reduction in the number of parameters and computation, which makes the model lighter; the third group is the replacement of some MP modules with PMP modules, with a mAP of 89.7%, an increase of 2.2%, and a parameter number of 39.8M, an increase of 2.6M; the fourth group is adding the ACmix attention mechanism, with a mAP of 89.8%, an increase of 2. 3%, and a parameter number of 38.3, an uptick of 1.1M; The fifth group is to replace the CIoU loss function with the WIoU loss function, which makes the model pay more attention to the anchor frames of ordinary quality, the mAP is 89.4, which is improved by 1. 9%, and the number of parameters is 33.8M, which is reduced by 3. 4M, which makes the model's detecting speed improved; The sixth group is to replace part of the MP module with the PMP module on the basis of the second group, the mAP is 90. 1%, which is improved by 2. 6%, and the number of parameters is 50.7M, which is increased by 13.5M; The seventh group is based on the sixth group with the addition of the ACmix attention mechanism, the mAP is 90.4%, which is improved by 2. 9%, and the number of parameters is 50.9M, which is increased by 13.7M; the eighth group is the final algorithm proposed in this paper, and the mAP is improved by 3. 1% compared with the original Yolov7, and the number of parameters is 50.7M, which is increased by 13.5M, these results fully proved the effectiveness of the DCH-YOLOv7 algorithm, which is shown to outperform YOLOv7 in various modules and configurations when evaluated on the challenging DIOR dataset.



(a)Original yolov7 model (b)DCH-yolov7 model

FIGURE 8. Comparison of PR curves before and after improvement.

TABLE 2. Ablation experiments for each module.

Group	PELAN	PMP	ACmix	WIoU	mAP@0.5(%)	Params(M)
G1					87.5	37.2
G2	√				87.5	32.7
G3		√			89.7	39.8
G4			√		89.8	38.3
G5				√	89.4	33.8
G6	√	√			90.1	50.7
G7	√	√	√		90.4	50.9
Ours	√	√	√	√	90.6	50.7

TABLE 3. Comparison of object detection performance on different objects between YOLOv7 and DCH-YOLOv7 using the DIOR Dataset.

	IoU	Area	maxDets	YOLOv7	DCH-YOLOv7
Average Precision(AP)	0.50:0.95	all	100	0.534	0.543
	0.50	all	100	0.874	0.887
	0.75	all	100	0.593	0.613
	0.50:0.95	small	100	0.510	0.524
	0.50:0.95	medium	100	0.694	0.745
	0.50:0.95	large	100	0.756	0.789
Average Recall(AR)	0.50:0.95	all	1	0.327	0.386
	0.50:0.95	all	10	0.622	0.665
	0.50:0.95	all	100	0.617	0.624
	0.50:0.95	small	100	0.604	0.674
	0.50:0.95	medium	100	0.742	0.765
	0.50:0.95	large	100	0.814	0.843

In Tables 3, regardless of the consideration of target size, DCH-YOLOv7 consistently outperforms other models, achieving higher AP and AR values across various IoU thresholds. When we factor in the target size, DCH-YOLOv7's superiority becomes even more apparent, with significantly higher AP and AR values compared to YOLOv7. To elaborate further, DCH-YOLOv7 attains remarkable AP values of 52.4% for small object detection, showcasing improvements of 1.4%, respectively, compared to YOLOv7. The AR values for DCH-YOLOv7 in small object detection reach impressive figures of 67.4%, demonstrating substantial improvements of 7.0%, respectively. These experimental results confirm the effectiveness of DCH-YOLOv7 in detecting small objects within optical remote sensing image.

Table 4, on the other hand, shows the comparison of the experimental results between the algorithms in this paper and the recent convolutional neural network-based target detection algorithms on the DIOR dataset. The experimental results of these algorithms are on the same hardware resources using the training and validation sets in DIOR as the training set. As can be seen from the results listed in Table 4, the detection accuracy is improved by 37.3%, 20.6% and 14.2% compared to several classical two-stage target

TABLE 4. Comparative experiments of the DCH-YOLOv7 algorithm on the DIOR dataset.

Model	Backbone	Size	mAP@0.5(%)	Fps(Frames/s)
R-CNN[16]	ResNet18	1000×600	53.3	-
Fast R-CNN[18]	ResNet50	1000×600	70.0	8.1
Faster R-CNN	ResNet50	1000×600	76.4	2.4
Faster R-CNN	VGG16	1000×600	73.2	7.0
SSD[25]	VGG16	512×3512	77.2	38.3
YOLOv3	DarkNet53	300×300	76.7	18.7
YOLOv4	CSP-DarkNet53	300×300	83.4	53.2
YOLOv5s	C3	512×512	82.7	38.8
YOLOX	PA-FPN	512×512	84.9	45.6
YOLOv7	E-ELAN	640×640	87.5	98.5
YOLOv7-tiny	E-ELAN	640×640	78.8	287
YOLOv8s	C2F	300×300	84.2	87
DCH-YOLOv7	**PELAN**	**640×640**	**90.6**	**77**

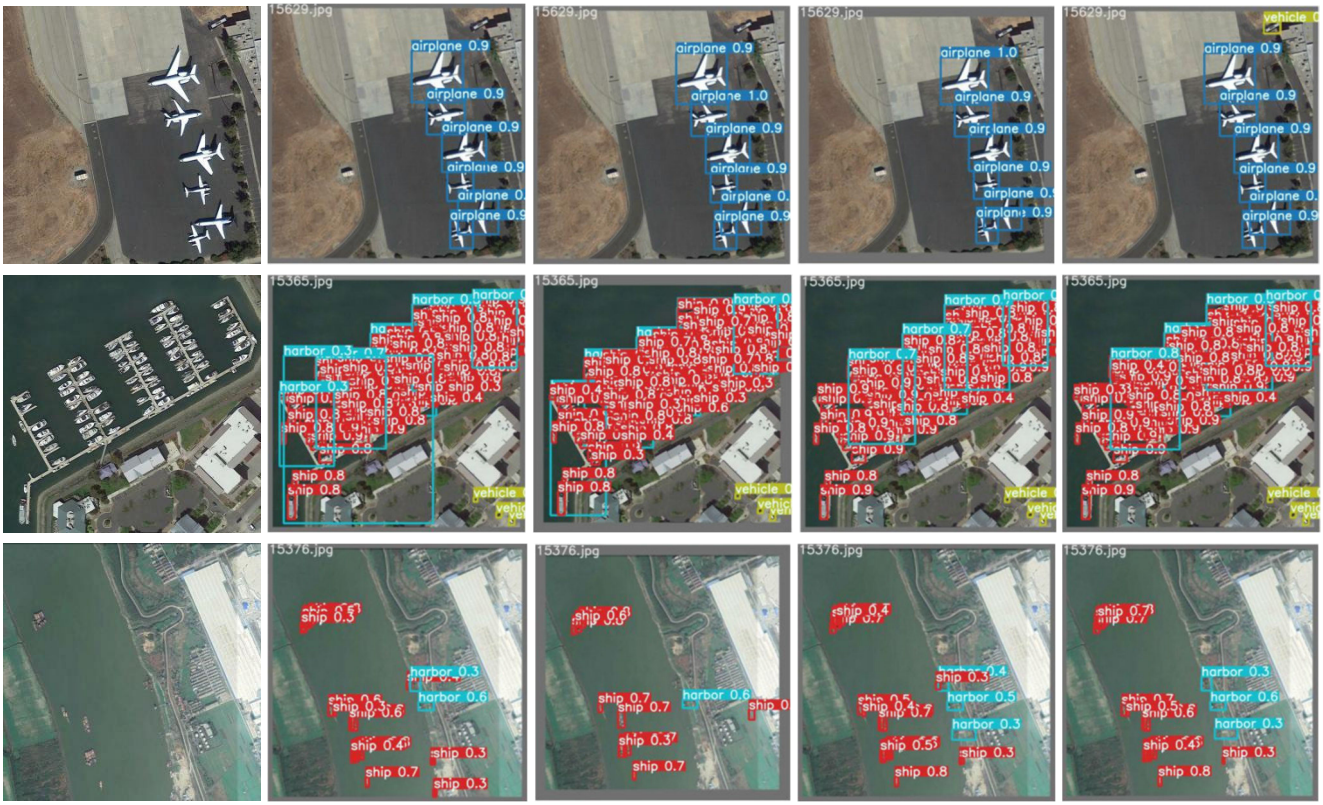


FIGURE 9. Visualization comparison on the DIOR dataset.

detection (R-CNN, Fast R-CNN and Faster R-CNN), and the speed of detection has increased dramatically. There is also a 13.4% improvement in detection accuracy compared to the more popular one-stage algorithm SSD. The detection accuracy is improved by 13.9%, 7.2%, 7.9% and 5.7% compared to YOLOv3, YOLOv4, YOLOv5s and YOLOX of the YOLO family, and by 3.1% from 87.5% to 90.6% compared to the original YOLOv7 model. Although YOLOv7-tiny demonstrates high real-time detection speed, its detection accuracy lags. In contrast, DCH-YOLOv7 achieves an impressive improvement of 11.8% in mAP@0.5. Additionally, when compared to YOLOv8s, one of the latest algorithms in the YOLO series, the proposed DCH-YOLOv7 algorithm achieves a superior balance between detection accuracy and speed, leading to an improvement of 6.4% in mAP@0.5.

The improved YOLOv7 network shows significant accuracy improvement in deep learning detection methods, which makes up for the shortcomings of the original network in

small target detection and further improves the robustness of the network. In conclusion, compared with other mainstream algorithms, the algorithm proposed in this paper has higher accuracy in target detection in remote sensing images.

2) COMPARISON OF THE ACTUAL DETECTION EFFECT OF DIFFERENT DETECTION METHODS

Figure 9 compares the detection effect of YOLOv5s, YOLOv7, YOLOv8s and DCH-YOLOv7 algorithms on DIOR datasets. From the results displayed, the performance of the YOLOv5s is insufficient, and there is an obvious phenomenon of missing detection. the YOLOv7 algorithm and YOLOv8s algorithm are adjusted and optimized based on YOLOv5 algorithm, and although they alleviate the missing detection phenomenon to a certain extent, they are not effective in detecting overlapping targets and incomplete targets. And the YOLOv7 optimization algorithm based on the introduction of deformable convolutional fusion attention mechanism in this paper can effectively improve the detection performance of small targets in optical remote sensing images under complex background.

V. CONCLUSION

Object detection has always been a popular direction in computer vision and digital image processing, widely used in fields such as aviation, transportation, and industry. The DCH-YOLOv7 algorithm proposed in this paper can adapt to image features of different scales and better perceive spatial position information in feature maps. Therefore, the DCH-YOLOv7 algorithm can be used for target detection in military monitoring, traffic planning, pollution control and other fields.

In order to improve the detection accuracy of targets in complex backgrounds during target detection in optical remote sensing images, this paper proposes a YOLOv7 optimization algorithm based on deformable convolution fused attention mechanism. On the basis of the YOLOv7 algorithm, improvements are made in three places: the deformable convolution is introduced and the PELAN and PMP modules are proposed, the fusion attention mechanism ACmix, and the replacement loss function. Firstly, deformable convolution is embedded in the backbone feature extraction network to help the model better adapt to image features with different scales, rotations or distortions. Based on this, PELAN module and PMP module are proposed to improve the model's ability to recognize complex scenes. Secondly, the ACmix attention mechanism is introduced into the YOLOv7 network to suppress the interference of complex background and noise, so that the network can better perceive the spatial location information in the feature map and enhance the ability to extract target features under complex backgrounds for better detection of small targets. Finally, the loss function is replaced with WIoU to further increase the focus on common quality anchor frames, which leads to more accurate anchor frame prediction and effectively reduces the probability of missed and false detections. The experimental results show that the improved YOLOv7 algorithm proposed in this paper has achieved remarkable results in optical remote sensing image target detection.

After experimental verification, the mAP value of the algorithm reaches 90.6% on the DIOR optical remote sensing target detection dataset, which exceeds the current mainstream target detection algorithms. This finding indicates that the algorithm has high accuracy and feasibility, and has some generalization value in practical applications. The DCH-YOLOv7 algorithm proposed in this paper can effectively extract the feature information of optical remote sensing images of different scales, and can suppress the interference of complex background and noise. This helps to improve the detection accuracy of the algorithm. However, the DCH-YOLOv7 algorithm does not fully consider the scale of the algorithm, resulting in slightly poor processing speed. Our future work will investigate how to increase the speed of the algorithm without decreasing the accuracy.

REFERENCES

[1] B. Tu, Z. Wang, H. Ouyang, X. Yang, J. Li, and A. Plaza, "Hyperspectral anomaly detection using the spectral–spatial graph," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542814, doi: 10.1109/TGRS.2022.3217329.

[2] B. Tu, Q. Ren, J. Li, Z. Cao, Y. Chen, and A. Plaza, "NCGLF2: Network combining global and local features for fusion of multisource remote sensing data," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102192.

[3] B. Tu, W. He, Q. Li, Y. Peng, and A. Plaza, "A new context-aware framework for defending against adversarial attacks in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505114, doi: 10.1109/TGRS.2023.3250450.

[4] Y. Qu and W.-H. Li, "Single-stage rotated object detection network based on anchor transformation," *J. Jilin Univ., Eng. Technol. Ed.*, vol. 52, no. 1, pp. 162–173, 2022.

[5] C. Dong, J.-H. Liu, F. Xu, and R.-H. Wang, "Fast ship detection in optical remote sensing images," *J. Jilin Univ. Eng. Technol. Ed.*, vol. 49, no. 4, pp. 1369–1376, 2019.

[6] Z. H. Shi, C. W. Wu, C. J. Li, Z. Z. You, Q. Wang, and C. C. Ma, "Object detection techniques based on deep learning for aerial remote sensing images: A survey," *J. Image Graph.*, vol. 28, no. 9, pp. 2616–2643, 2023.

[7] J. Han, S. Liu, G. Qin, Q. Zhao, H. Zhang, and N. Li, "A local contrast method combined with adaptive background estimation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1442–1446, Sep. 2019.

[8] Z. Lv, R. Wang, Y. Wang, F. Zhou, and N. Guo, "Road scene multi-object detection algorithm based on CMS-YOLO," *IEEE Access*, vol. 11, pp. 121190–121201, 2023.

[9] Q. Lin, R. Wang, Y. Wang, F. Zhou, and N. Guo, "Target detection algorithm incorporating visual expansion mechanism and path syndication," *IEEE Access*, vol. 11, pp. 56973–56982, 2023.

[10] 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.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

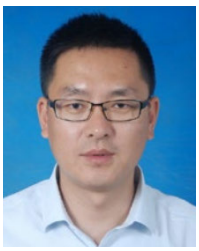
[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[15] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[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.*, 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] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network," *Remote Sens.*, vol. 10, no. 1, p. 131, 2018.
[20] S. Jiang, W. Yao, M. S. Wong, G. Li, Z. Hong, T.-Y. Kuc, and X. Tong, "An optimized deep neural network detecting small and narrow rectangular objects in Google Earth images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1068–1081, 2020, doi: 10.1109/JSTARS.2020.2975606.
[21] Q. Yao, X. Hu, and H. Lei, "Multiscale convolutional neural networks for geospatial object detection in VHR satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 23–27, Jan. 2021, doi: 10.1109/LGRS.2020.2967819.
[22] Y.-Z. Zhang, W. Guo, Z.-Q. Cai, and W.-B. Li, "Remote sensing image target detection combining multi-scale and attention mechanism," *J. Zhejiang Univ. Eng. Sci.*, vol. 56, no. 11, pp. 2215–2223, 2022.
[23] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018, doi: 10.1109/ACCESS.2018.2869884.
[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[25] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9300–9308, doi: 10.1109/CVPR.2019.00953.
[26] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 805–815, doi: 10.1109/CVPR52688.2022.00089.
[27] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.
[28] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022, doi: 10.1109/TCYB.2021.3095305.
[29] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 13728–13737, doi: 10.1109/CVPR46437.2021.01352.
[30] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.
[31] Y. Hou, G. Shi, Y. Zhao, F. Wang, X. Jiang, R. Zhuang, Y. Mei, and X. Ma, "R-YOLO: A YOLO-based method for arbitrary-oriented target detection in high-resolution remote sensing images," *Sensors*, vol. 22, no. 15, p. 5716, Jul. 2022.
[32] Y.-Z. Zhang, W. Guo, and W.-B. Li, "Omnidirectional accurate detection algorithm for dense small objects in remote sensing images," *J. Jilin Univ. Eng. Technol. Ed.*, pp. 1–9, Dec. 2023, doi: 10.13229/j.cnki.jdxbgxb20220715.
[33] J. Yu, S. Liu, and T. Xu, "Research on YOLOv7 remote sensing small target detection algorithm integrating attention mechanism," *Comput. Eng. Appl.*, vol. 59, no. 20, pp. 167–175, 2023.
[34] Y. Yang, Y. Liao, L. Cheng, K. Zhang, H. Wang, and S. Chen, "Remote sensing image aircraft target detection based on GIoU-YOLO v3," in *Proc. 6th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2021, pp. 474–478.
[35] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*.
[36] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

CHUNHUI CUI received the B.S. degree from Shandong Technology and Business University, Shandong, China, in 2020. She is currently pursuing the M.Eng. degree with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. Her current research interests include computer vision technology and image processing technology.

RUGANG WANG received the B.S. degree from Wuhan University of Technology, Wuhan, China, in 1999, the M.S. degree from Jinan University, Guangzhou, China, in 2007, and the Ph.D. degree from Nanjing University, Nanjing, China, in 2012. He is currently a Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His research interests include optical communication networks, novel and key devices for optical communication systems, and image processing technology.

YUANYUAN WANG, photograph and biography not available at the time of publication.



FENG ZHOU received the B.S. and M.S. degrees from Southeast University, Nanjing, China, in 2004 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Army Engineering University of PLA. Since 2017, he has been an Associate Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His research interests include cooperative communication, computer vision technology, and image processing technology.

XUESHENG BIAN, photograph and biography not available at the time of publication.

JUN CHEN, photograph and biography not available at the time of publication.