BCS-YOLOv8s: A Detecting Method for Dense Small Targets in Remote Sensing Images Based on Improved YOLOv8s

Wutao Du, Xinyu Ouyang, Nannan Zhao, Yifan Ouyang

***Abstract*—The remote sensing image contains a lot of dense small targets, which increases the difficulty of object detection. The loss of small target feature information in feature fusion is rarely taken into account by the target detection algorithms used in remote sensing images today. A dense small object detection method based on improved YOLOv8s, namely BCS-YOLOv8s, is proposed to address this issue. The innovation of the proposed method is mainly reflected in three aspects. First, the backbone network was modified to incorporate Bi-Level Routing Attention (BRA), a dynamic sparse attention mechanism, which increased the model's concentration on tiny targets without appreciably changing its parameters. Second, CSNeck is employed as the neck of the model. Content-Aware ReAssembly of Features (CARAFE) is implemented as the upsampling module to minimize information loss of tiny targets in the feature combination. To enhance the model's capacity to identify tiny targets, a detection head and small target detection layer are also incorporated. Thirdly, pre-selected frame regression may be made faster and more accurate by employing structured IoU (SIoU). The ultimate experimental findings demonstrate that BCS-YOLOv8s lowers the missed rate of tiny targets and enhances the model's detection ability. The average detection accuracy (mAP) of this model on the DIOR dataset is 89.5%, which is 3.2% higher than the base model. Compared with other mainstream models such as YOLOv5 and YOLOv7, this model has better performance in all aspects. By using this strategy, the model's capacity to recognize dense and tiny targets is successfully improved.**

***Index Terms*—small object detection, remote sensing, YOLOv8, DIOR, SIoU, Bi-Level Routing Attention, CARAFE.**

I. Introduction

IMAGES from remote sensing have become more and more significant in people's lives and careers in the past few decades. Additionally, as target recognition technology advances, the use of target recognition in images collected by remote sensing is becoming progressively more crucial. However, the task of target detection is made more difficult by the complex background and the abundance of small, dense targets in these images. Thus, the focus of study has shifted to how to get around these challenges and enhance the efficiency and accuracy of remote sensing picture target recognition techniques [1].

In remote sensing image identification, deep learning powered object recognition approaches have shown impressive results, surpassing conventional object detection methods in terms of accuracy and efficiency. The two types of object detection methods are one-stage approaches and two-stage approaches, which are determined by the detection strategy. Where the two-stage approaches divide the object detection task into: generating region suggestions and classifying and correcting the location of the region suggestions in a classifier, such as R-CNN [2], SPP-Net [3], Fast R-CNN [4], Mast R-CNN [5]and Faster R-CNN [6]. The one-stage approaches generates bounding boxes directly by regression prediction objects, such as the You Only Look Once (YOLO) series [7]–[10], SSD [11]and EfficientDet [12].

In the subject of remote sensing picture target recognition, there are a lot of study findings available right now. Zhou et al. [13]proposed a data enhancement approache to solve the problem of monotonous image background. An attention-based feature combination SSD approach was presented by Lu et al. [14] to enhance the effectiveness of model detection for tiny targets. Liu et al. [15] introduced ResNet [16] in YOLOv3 to optimize the backbone network. An end-to-end pyramid network for multiple sizes target identification was proposed by Wang et al [17]. Wu et al. [18] proposed a SEF module in YOLOv8 based on lightweight convolution (SEConv), which speeds up the detection process. And the multi-scale attention mechanism is added to the method, thus increasing the feature extraction capability. Wang et al. [19] introduced a bidirectional feature pyramid network (BiFPN) into the YOLOv8 model, while multi-head self-attention is integrated into the network and data enhancement techniques are used to improve model robustness. Li et al. [20] integrating the Conv2Former module into the YOLOv7+ model, which enhances the extraction of spatial information features with more precision. At the same time, there are a number of studies that can be informative. Zhang et al. [21] proposes a double-layer semicomposite backbone network structure to enhance the ability of the backbone network to extract target features. Zhang et al. [22] propose a bidirectional partial dynamic fusion module to facilitates cross-level interactive fusion of feature information.

High altitude above the ground and a single imaging viewpoint characterize photos from remote sensing. As a result,

Manuscript received August 14, 2024; revised December 12, 2024. This work is supported by the Fundamental Research Funds for the Liaoning Universities of China (Grant No. LJ212410146005).

Wutao Du is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, 114051, China (e-mail: 18713912310@163.com).

Xinyu Ouyang is a professor of the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, 114051, China (Corresponding author, e-mail: 13392862@qq.com).

Nan-Nan Zhao is a professor of the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan,Liaoning, 114051, China (Corresponding author, e-mail: 723306003@qq.com).

Yifan Ouyang is an undergraduate student of School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Sepang, Selangor Darul Ehsan, 43900, Malaysia (e-mail: AIT2209084@xmu.edu.my).

there are many different types and quantities of objects in photographs from remote sensing, as well as a huge number of small, compact objects, and the object's orientation might change. Photos from remote sensing will be modified by various shooting platforms, weather and lighting conditions, and other factors.

The present paper proposes a BCS-YOLOv8s remote sensing picture target recognition approach based on YOLOv8 to address the issues mentioned above. This model exhibits strong performance in enhancing the detection accuracy for remote sensing images, and the main contributions are as follows:

(1) A novel dynamic sparse attention mechanism is introduced in the backbone network to improve the model's capture and mapping of key features, optimizing the detection performance without overburdening the parameter count.

(2) A new neck structure, CSNeck, is proposed. A content-aware upsampling operator is introduced in the neck, and the detection layer for small targets is extended so that the model better preserves small target feature information during feature fusion.

(3) To increase the algorithm's preselection box regression's speed and accuracy, SIoU is used as the loss function.

II. Related Work

The You Only Look Once (YOLO) family of models is currently the dominant object detection model and has accomplished remarkable success in the domain of computer vision.

The YOLOv8 model combines the advantages of many excellent models, and it has more powerful performance than previous versions. Figure 1 depicts the network topology of the YOLOv8 model, which is comprised of three sections: the head, neck, and backbone.

A. Backbone

The backbone network of YOLOv8 is borrowed from the structure of CSPDarknet53, and improvements are made on it. To create five feature maps at various sizes for feature combination in the neck, the input features are downsampled five times using additional feature extraction methods. A convolutional layer, a BN normalization layer, and a SiLU activation function constitute the CBS module. It is used for downsampling operations; To increase the network's capacity for extracting features, the C2f module is capable of expanding the receptive field. The C2f module adds extra layer hopping links and split operations compared to the C3 module and eliminates the convolutional operations in the branches. It greatly increases the effectiveness of the algorithm by ensuring lightweighting while also obtaining greater gradient flow data and adjusting the total amount of channels in accordance with the algorithm's size; the SPPF module achieves multi-scale fusion by three consecutive maximal pooling with residual structure and finally fusing the pooling front with the result of each pooling. Compared with the SPP structure, the SPPF structure reduces the computational volume and further increases the receptive field while ensuring multi-scale fusion.

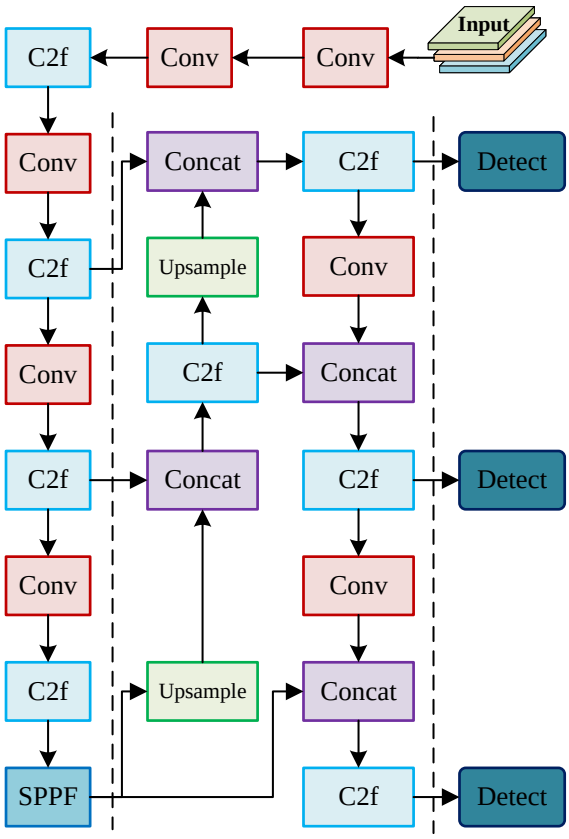

Fig. 1. The YOLOv8 structure diagram

B. Neck

Feature fusion is the primary task of the neck. By adopting the PAN-FPN construction, which is split into two halves for the feature fusion process from top to bottom and from bottom to top and fully using the features gained from the backbone network, the neck may improve the semantic expression and localization ability on different scales. Firstly, the deep feature maps are fused with the coarse-grained feature maps from top to bottom by up-sampling operation to get richer feature representations. Secondly, the fused shallow feature map is then subjected to a down-sampling operation to fuse it again with the feature map in the first step.

In the up-sampling operation, the YOLOv8 model uses bilinear interpolation, which only considers the neighborhood of the interpolated pixels and ignores the overall continuity, which will seriously affect the overall quality of the feature map. Another commonly used upsampling method is the deconvolution method, which uses the same convolution kernel globally, which severely limits the model's ability to cope with local variations. At the same time, both approaches invariably include a lot of variables, which slows down the model. This is particularly noticeable when attempting to identify objects in photos from remote sensing.

C. Head

The head is used to predict the object position and category. The YOLOv8 model uses the Decoupled Head design. The Coupled Head uses a sequence of convolutional and fully connected layers at the network's conclusion to forecast the bounding box's position, size, and category at several scales

concurrently. Decoupled Head is to separate the classification and detection heads, two parallel branches are taken to extract the category features and location features respectively. The two parallel branches complete the classification and localization tasks with one layer of $1 * 1$ convolution respectively to enhance target detection performance.

The decoupled head not only improves model accuracy, but also enhances the network's convergence performance. The decoupled head also has a better expressive ability than the coupled head, which enhances the robustness of the model, allows for better modeling of the relationship between position and category, and improves object detection performance.

D. Loss function

The loss function in the YOLOv8 model consists of two parts: classification loss and regression loss, which uses a cross entropy loss (BCE Loss) for classification loss and a combination of DFL + CIoU Loss for regression loss.

CIoU calculates the loss by calculating the distance and the overlapping area between the center of the preselected box and the real box. CIoU includes an aspect ratio term to ensure the regression quality. Even yet, there are still some issues with CIoU. For instance, the aspect ratio term does not appropriately react to a situation in which the dimensions of the real and preselected box differ, but their aspect ratios are consistent. Meanwhile, due to the application of inverse trigonometric functions, the speed of the model receives limitations.

III. Method

This study presents the BCS-YOLOv8s model, which is an improvement on YOLOv8, figure 2 shows its structure. The improvement enhances the algorithm's incapacity to identify tiny, dense objects. The next section provides a detailed description of it.

A. Improvements in Backbone

Figure 3 illustrates the introduction of Bi-Level Routing Attention (BRA) [23] into the YOLOv8 architecture to enhance the feature extraction capability for tiny items in remote sensing pictures. The main concept is to filter the region with low correlation first, instead of directly calculating the correlation between two elements. This approach can successfully avoid the model size from being too big while boosting the algorithm's accuracy by comparison to the usual attention mechanism. The benefit is particularly noticeable on remote sensing photos.

Firstly, the feature map $X \in R^{H*W*C}$ is divided into $S*S$ non-overlapping regions, and then each region has HW/S^2 feature vectors. Then we are able to obtain Q, K,and V, i.e., query, key, and value for each region. Next, the region-to-region routing of the directed graph is used. Calculate the average values Q^r and K^r of Q and K for each region, and then use matrix operation to acquire the adjacency matrix $A^r \in R^{S^2*S^2}$ of the area friendship network. Next, determine the routing index matrix I^r via row-by-row topk operation, as shown in Equations (1) and (2).

$$A^r = Q^r \left(K^r\right)^T \tag{1}$$

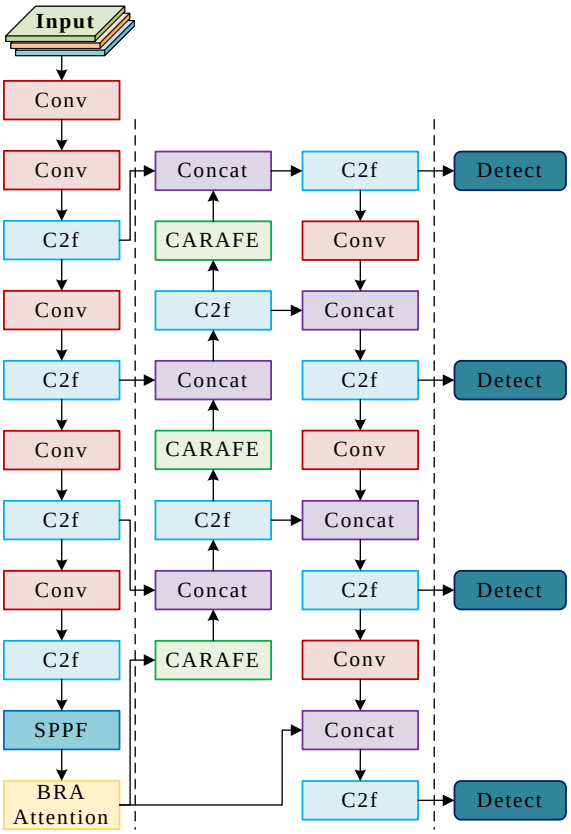

Fig. 2. BCS-YOLOv8s structure diagram

$$I^r = topkIndex\left(A^r\right) \tag{2}$$

Finally, the anchor box attention mechanism is used. For each query in area i, BRA will pay attention to all key-value pairs in the concatenation of k routing areas indexed as $I^r(i,1)$, $I^r(i,2)$, ..., $I^r(i,k)$, as shown in Equation (3).

$$K^g = gather\left(K, I^r\right), V^g = gather\left(V, I^r\right) \tag{3}$$

The gathered combinations of key-value are then the center of attention, as illustrated by Equation (4), where K^g, V^g is a collection of key-value tensors and $LCE(V)$ is a local context augmentation term.

$$O = Attention\left(Q, K^g, V^g\right) + LCE\left(V\right) \tag{4}$$

B. Improvements in Neck

Remote sensing images contain a multitude of densely distributed small-scale targets, and the general model only focuses on the fusion of deep feature maps, and some of the small targets appear to lose semantic information after repeated deep feature extraction and fusion, which reduces the reliability of the algorithm recognition. Furthermore, the YOLOv8 algorithm's detecting head is unable to recognize small-scale objects accurately, which is necessary to finish the target identification job of photos from remote sensing. Therefore, a new neck structure, CSNeck, is proposed to address the above problems, which can meaningfully ameliorate the situation of small target semantic loss and insufficient performance. Its structure is displayed in Figure 4.

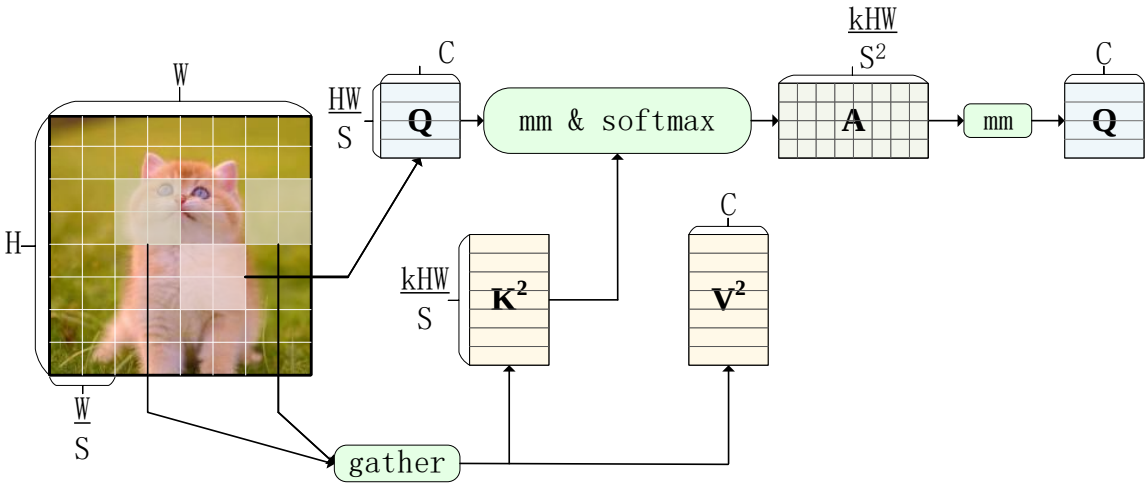

Fig. 3. BRA Schematic diagram

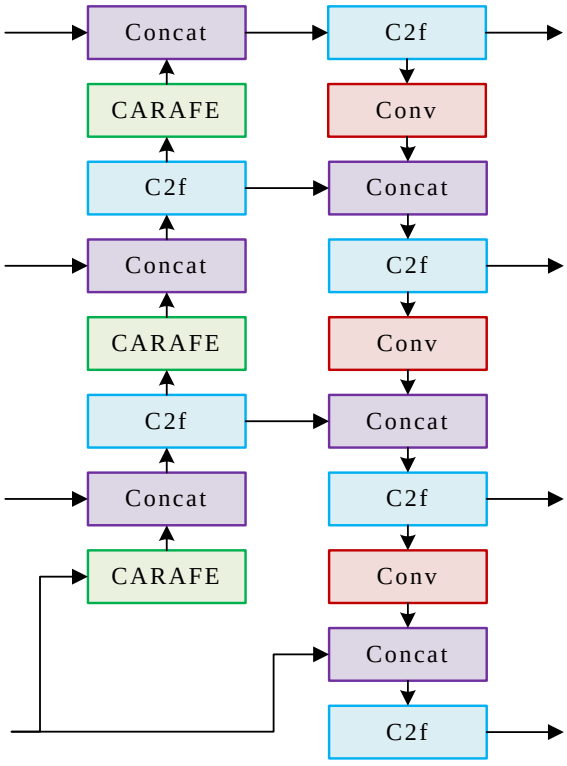

Fig. 4. CSNeck structure diagram

Content-Aware Reassembly of Features(CARAFE) [24] module is introduced in CSNeck to replace the original up-sampling module. Compared to the original up-sampling module, the CARAFE module has a larger receptive field and can better aggregate contextual information to form coherent features. Secondly, it can perform content-aware processing on different samples to form an adaptive kernel applicable to the current sample, which solves the problem of insufficient ability to cope with local changes of, for example, deconvolution methods. In addition, the CARAFE module is able to introduce a small number of covariates to enhance the model's effectiveness, solving the problem of large computational volume. The specific flow of CARAFE is as follows.

CARAFE will reorganize feature map X of size $C*H*W$ into feature map X' of size $C*\sigma H*\sigma W$ in two steps, and for each feature position $l' = [\sigma i, \sigma j]$ in feature map X' there is a position $l = [i, j]$ located in the original feature map corresponding to it, where σ is the up-sampling rate. Firstly, the kernel prediction module Ψ generates a recombination kernel $W_{l'}$ of size $C_{up}*H*W$ based on the neighborhood l' of location l in the original feature map X, and this process can be expressed as Equation (5). where $N(X_l, k)$ stands for the neighborhood of location l in the original feature map X of size $k*k$. C_{up} is the number of channels of the recombination kernel, which is used to specify the kernel dimension, and can be expressed as Equation (6).

$$W_{l'} = \Psi\left(N\left(X_l, k_{encoder}\right)\right) \tag{5}$$

$$C_{up} = \sigma^2 k_{up}^2 \tag{6}$$

Secondly, the content-aware recombination module recombines the features with the kernel to generate a new feature map X' , the process that can be represented as Equation (7). Φ is the content-aware recombination module, which is

responsible for combining each of the local feature maps with the corresponding recombination kernel $W_{l'}$, as shown in Equation (8), where $r = k_{up}/2$.

$$X'_{l'} = \Phi\left(N\left(X_l, k_{up}\right), W_{l'}\right) \tag{7}$$

$$\Phi\left(X'_{l'}\right) = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W_{l'(n,m)} * X_{(j+n,j+m)} \tag{8}$$

There are two main parameters $k_{encoder}$ and k_{up} that determine the final result in the CARAFE process, representing the size of the context area used to generate the recombination kernel and the size of the context area used for feature recombination, respectively. Typically, the relationship is shown in Equation (9). Through experiments, we determined the optimal parameter as $k_{encoder} = 1, k_{up} = 3$.

$$k_{encoder} = k_{up} - 2 \tag{9}$$

Meanwhile, we also introduce the shallow feature maps from the backbone network into CSNeck, add new small-target feature fusion structures and small-target detection heads, and further improve the model's capacity to learn multi-size target feature information.

C. Improvements in Loss function

SIoU is utilized as the loss function to both accelerate the speed and enhance the accuracy of anchor regression. Compared with CIoU, SIoU solves the defect of aspect ratio by separately computing the length and width of anchor boxes and real boxes. Also SIoU takes into account the angle between the two boxes by defining an angular penalty metric, which enables the anchor box to quickly drift to the nearest axis first, and the subsequent regression process only needs to regress to one coordinate (X or Y). SIoU can be divided into four parts: angular loss, distance loss, shape loss, and IoU loss.

Angle loss is used to describe the minimum angle between the line segment connecting the centres of the preselector box and the real box and the x-y axis (shown in Fig. 5), $\Lambda = 0$ when the two centrally connected lines overlap the x or y axes, and $\Lambda = 1$ when $\alpha = 45^o$. This penalty directs the preselector box to move to the nearest axis and reduces the total number of degrees of freedom in the BBR. The equations are shown in Equation (10) to Equation (13).

$$\Lambda = 1 - 2 * sin^2\left(arcsin\left(x\right) - \frac{\pi}{4}\right) \tag{10}$$

$$x = \frac{c_h}{\sigma} = sin\left(\alpha\right) \tag{11}$$

$$\sigma = \sqrt{\left(b^{gt}_{c_x} - b_{c_x}\right)^2 + \left(b^{gt}_{c_y} - b_{c_y}\right)^2} \tag{12}$$

$$c_h = max\left(b^{gt}_{c_y}, b_{c_y}\right) - min\left(b^{gt}_{c_y}, b_{c_y}\right) \tag{13}$$

Distance loss is used to describe the gap between the anchor box's centroid and the real box's centroid (as shown in Fig. 6), and this penalty cost is positively correlated with the angle loss. When $\alpha \to 0$, the contribution of distance loss will gradually decrease; when $\alpha = 45^o$, its contribution is the largest. The formulas are shown in Equation (14) to Equation (16).

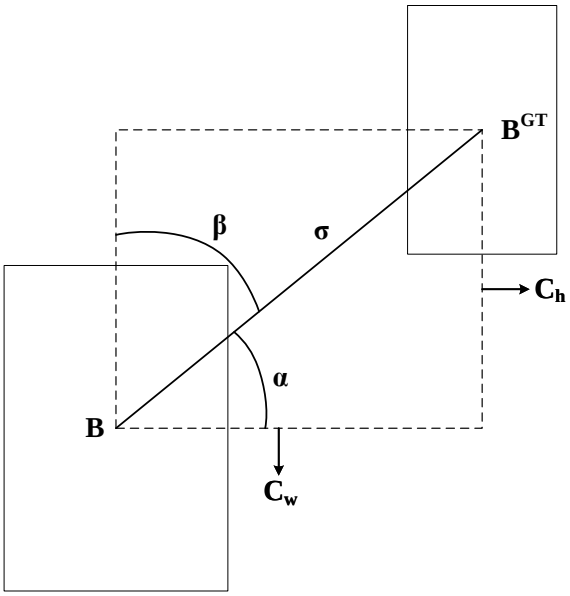

Fig. 5. Angular loss schematic

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma\rho t}\right) \tag{14}$$

$$\gamma = 2 - \Lambda \tag{15}$$

$$\rho_w = \left(\frac{b^{gt}_{c_x} - b_{c_x}}{c_w}\right)^2, \rho_h = \left(\frac{b^{gt}_{c_y} - b_{c_y}}{c_h}\right)^2 \tag{16}$$

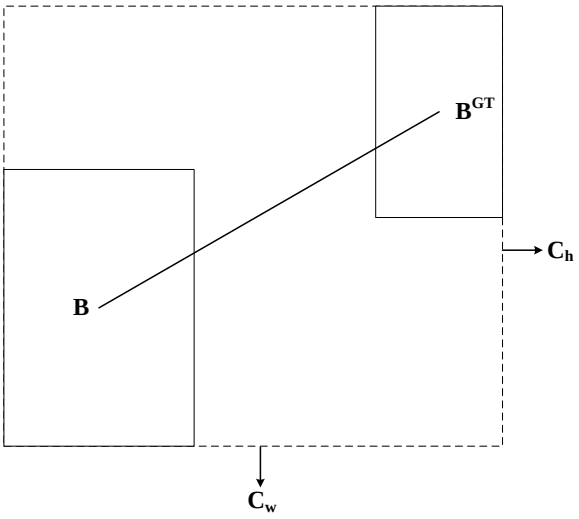

Fig. 6. Distance loss schematic

In shape loss, SIoU factors in the proportional relationship between length and width of the anchor boxes and real boxes, nonetheless, in contrast to CIoU, it is determined by computing the difference between the lengths of the two boxes and the ratio of the maximum lengths of the two boxes (same for the widths). SIoU achieves the effect of the overall shape convergence by the convergence of the two

edges separately. The formulas are shown in Equation (17) and Equation (18).

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega t}\right)^{\theta} \tag{17}$$

$$\omega_w = \frac{|w - w^{gt}|}{max\left(w, w^{gt}\right)}, \omega_h = \frac{|h - h^{gt}|}{max\left(h, h^{gt}\right)} \tag{18}$$

Among them, θ is an important parameter. For different datasets, the value of θ is different. When $\theta = 1$, SIoU immediately corrects the shape of the anchor boxes, which seriously affects the free motion of the regression. It is derived from the experiments that the value of θ should be between 2 and 6.

At this point, we obtain the overall composition of SIoU, which is shown in Equation (19).

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{19}$$

IV. Experiment

A. Experimental Setting

The datasets, related valuation indicators, training methodology, and setup of the experiment are all covered in this section.

Table 1 displays the ambient conditions as well as the hardware platform utilized throughout the experiment.

TABLE I
Comparison of Feature Extraction Modules

Parameters	Configuration
CPU	i5-12400F 2.50GHz
GPU	NVIDIA GeForce RTX 3060
RAM	32G
GPU memory size	12G
Operating systems	Windows 11
Deep learning architecture	Pytorch2.0.1+Cuda11.8

The YOLOv8 model is classified into five models according to the width and depth. In order to have better research and improvement, YOLOv8s is used as the base model. Table 2 displays some key parameters in the algorithm training.

TABLE II
Key Training Parameters

Parameters	Setup
Epoch numbers	500
Momentum setting	0.937
Initial learning rate	0.01
Final learning rate	0.01
Weight decay	0.0005
Batch size	8
Input image size	640
Optimizer	Auto

The DIOR dataset [25], a well-known object recognition dataset for photographs from remote sensing, contains 20 types of targets, with a total of 23463 images and 192472 targets. The DIOR dataset contains a rich variety of types of detected objects, a large number of objects, an irregular distribution, and a large span of sizes, which can better validate the advantages of the improved model.

Figures 7 and 8 show the distribution of the number of targets of various classes in the dataset as well as the distribution of target bounding box sizes as well as locations. The target size of this dataset spans a wide range, the distribution is dense, and the target distribution location covers the whole picture, which can be a good test of the model performance. In the present study, the ratio of the training, validation, and test sets is 6:2:2.

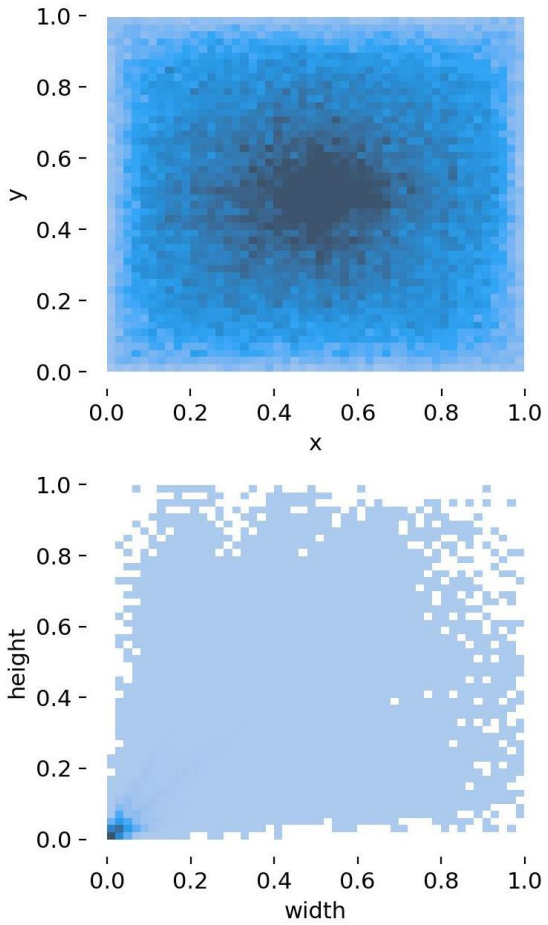

Fig. 7. distribution of target bounding box

The NWPU VHR-10 dataset [26] is a geographic remote sensing dataset produced by the Northwestern Polytechnical University (NWPU) in 2014, which has 650 images containing targets and 150 background images, totaling 800 images, with a total of 10 target categories. This dataset is used in this paper for assisted validation and generalizability validation experiments.

Precision (P), recall (R), total average precision (mAP0.5, mAP0.5:0.95), number of parameters, model size, and detection speed were utilized as assessment criteria to precisely analyze the enhanced algorithm's detection ability. In the definitions of each performance metric below, samples classified as TP are those that both the model and the data really anticipated to be positive; FP denotes samples that were really negative but were predicted by the model to be positive; FN denotes samples that were positive even though the model had projected them to be negative.

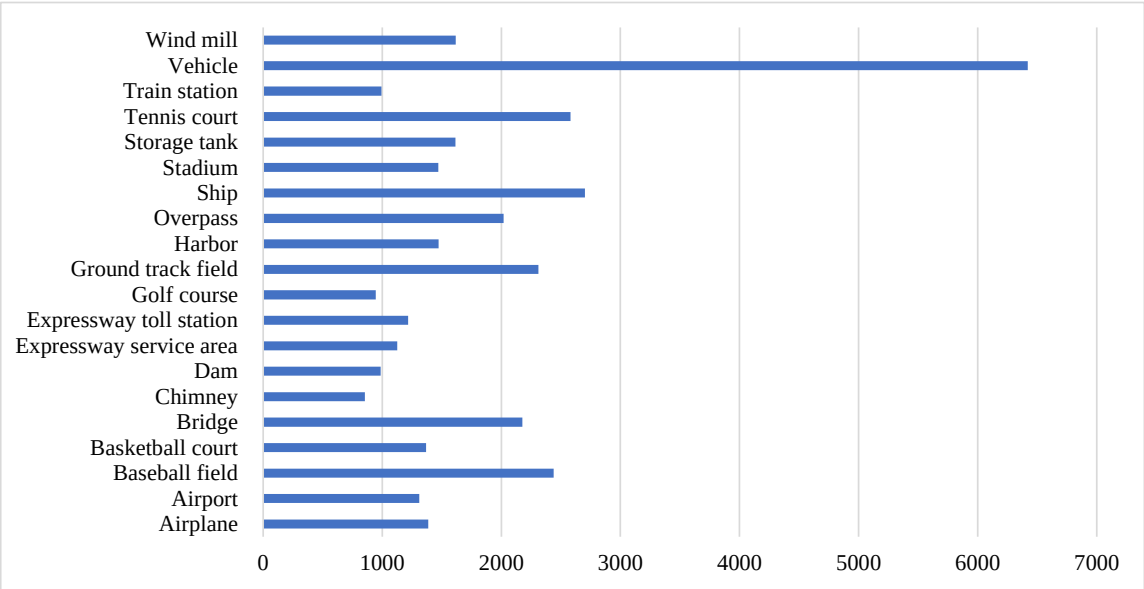

Fig. 8. Distribution of classes

As demonstrated by Equation 20, precision (P) is the ratio of the number of positive data samples that the model predicts to the total number of samples that are projected to be positive.

$$P = \frac{TP}{TP + FP} \tag{20}$$

Equation 21 illustrates recall (R), which is the ratio of the number of positive data samples that the model correctly predicted to the actual number of positive samples.

$$R = \frac{TP}{TP + FN} \tag{21}$$

Average Precision (AP) is the area enclosed by the P-R curve and represents the average prediction precision of the samples in a given category, as shown in Equation 22.

$$AP = \int_0^1 P\left(r\right) \mathrm{d}r \tag{22}$$

Recall is plotted on the x-axis of the P-R curve, while precision is plotted on the y-axis. The average accuracy is shown by the region that is bounded by the P-R curve, the x- and y-axes.

The total average precision (mAP) is a parameter obtained by the weighted average of the AP values across all sample categories, which serves as a comprehensive metric to evaluate the model's detection performance under all sample categories, as shown in Equation 23.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{23}$$

where N represents the number of sample types in the dataset. Furthermore, when the algorithm IoU is configured to 0.5, the overall average accuracy is represented by mAP0.5, and when the algorithm IoU is configured between 0.5 and 0.95, it is represented by mAP0.5:0.95. IoU represents the ratio between the intersection and concatenation of the predicted target bounding box and the true box.

B. Experiment Details

Comparative studies between the enhanced model and the original YOLOv8s model on the DIOR dataset were conducted to validate the improvement. Table 3 shows the values of AP and mAP for each category for both models. The improved model has a great improvement compared to the original model, and the mAP value is increased from 0.863 to 0.895, which indicates that the improved method can effectively improve the model's detection accuracy and performance of dense small targets on remote sensing images.

TABLE III
COMPARISON OF AP (MAP) BEFORE AND AFTER IMPROVEMENT ON DIOR DATASET

Classification	AP(before)	AP(after)
all	0.863	0.895
Airplane	0.955	0.978
Airport	0.923	0.922
Baseballfield	0.948	0.979
Baskballcourt	0.912	0.935
Bridge	0.630	0.685
Chimney	0.931	0.935
Dam	0.825	0.829
Expressway service area	0.971	0.983
Expressway toll station	0.853	0.936
Golf course	0.864	0.900
Ground track field	0.911	0.938
Harbor	0.743	0.793
Overpass	0.703	0.769
Ship	0.954	0.954
Stadium	0.972	0.969
Storage tank	0.872	0.926
Tennis court	0.971	0.988
Train station	0.670	0.768
Vehicle	0.685	0.750
Wind mill	0.962	0.963

Figures 9, 10, and 11 show the difference between the

improved model and the original model in terms of mAP0.5, mAP0.5:0.95, and recall during training. The original model starts to converge near 80 rounds, while the improved model already converges near 30 rounds. Both the detection accuracy and convergence speed of the modified algorithm are significantly higher than the original model, proving the effectiveness of the improvement.

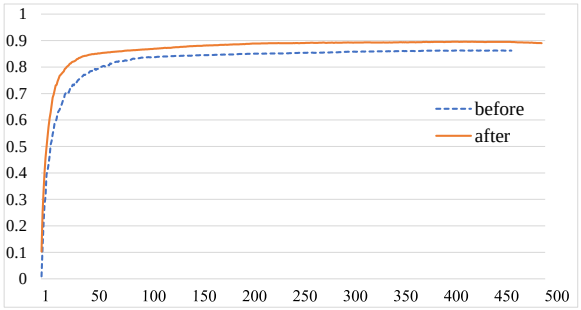

Fig. 9. Comparison of mAP0.5 before and after improvement

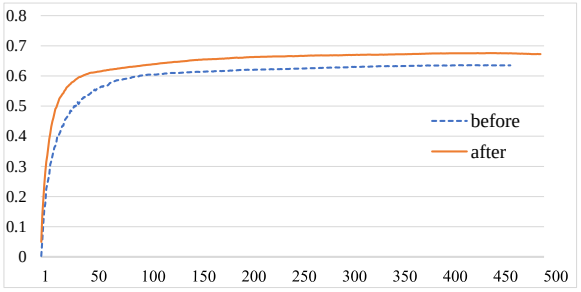

Fig. 10. Comparison of mAP0.5-0.95 before and after improvement

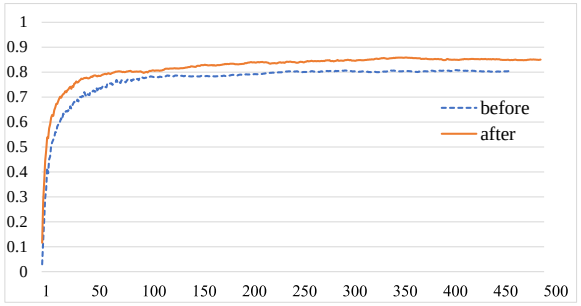

Fig. 11. Comparison of Recall before and after improvement

In order to better validate the effectiveness and generalizability of the improvement, the NWPU VHR-10 dataset was used to perform the auxiliary validation, and Table 4 displays the experimental outcomes. This suggests that the improvement approach may be used on many datasets and is generally applicable.

TABLE IV
COMPARISON OF MAP BEFORE AND AFTER IMPROVEMENT ON NWPU VHR-10 DATASET

Models	mAP0.5	mAP0.5:0.95
YOLOv8s	0.88849	0.52938
BCS-YOLOv8s	0.92724	0.59142

A comparison was conducted with mainstream methods to further validate the efficacy and superiority of the improved method. Firstly, in terms of the attention, the BRA attention adopted is compared with the current mainstream CA, SE, and MHSA attention, and the results are shown in Table 4.

TABLE V
COMPARISON OF ATTENTION MECHANISMS

Model	Attention	mAP0.5	mAP0.5:0.95
YOLOv8s		0.863	0.636
	SE	0.86	0.632
	MHSA	0.868	0.641
	BRA	0.876	0.652

From the above table, the BRA attention mechanism improves the model mAP by 1.3%, which has a better effect compared to the current mainstream attention mechanism on remote sensing images with a multitude of densely packed small objects. The critical content is also better enabled to be emphasized by the model, while the feature of region-to-region routing gives the model a higher efficiency.

Secondly, in terms of the neck, CSNeck is compared with Bifpn, a current popular neck improvement solution, and Table 6 displays the outcomes.

TABLE VI
COMPARISON OF NECK

Model	Neck	mAP0.5	mAP0.5:0.95
YOLOv8s		0.863	0.636
	Bifpn	0.858	0.628
	CSNeck	0.879	0.654

From the above table, it becomes apparent that CSNeck improves the model mAP by 1.6%, compared with bifpn, has better results on remote sensing images, improves the information loss in feature fusion by adjusting the up-sampling, and the introduction of the small target detection layer is also more targeted at the identification of tiny targets.

A blation experiments have been conducted to validate that each of the improvement strategies presented in this study has enhanced the detection capability of the model, and the results are shown in Table 7.

TABLE VII
RESULTS OF ABLATION EXPERIMENTS

Model	YOLOv8s			
BRA		Y	Y	Y
CSNeck			Y	Y
SIoU				Y
mAP0.5	0.863	0.876	0.890	0.895
mAP0.5:0.95	0.636	0.652	0.666	0.676
Recall	0.807	0.816	0.843	0.852
Precision	0.888	0.885	0.891	0.889
Parameter	11.1	11.4	11.9	11.9

The information in the table above shows that the model's detection performance has been enhanced to varying degrees by each enhancement technique suggested in this research. By adding the BRA attention mechanism to the backbone network, the mAP is improved by 1.3% and the attention to main elements is improved. CSNeck, which replaced the previous neck, reduces the features lost during the feature

fusion process, while the tiny target detection layer enhances the capacity to recognize tiny objects, which improves the mAP by 1.4%. SIoU is utilized in the loss function section to increase the accuracy and efficiency of regression of anchor box, resulting in a 0.5% improvement in the mAP.

The majority of the detection measures are substantially increased, and the revised model has a 3.2% overall improvement in mAP.

The paper carries out a comparative experiment to confirm the enhanced model's efficacy even further. We contrast the enhanced algorithm with a few popular ones. The results are shown in Table 8.

TABLE VIII
COMPARISON EXPERIMENTAL RESULTS

Models	mAP50	mAP50:95	Precision	Recall	Parameter
YOLOv5s	0.859	0.625	0.881	0.803	7.01
YOLOv6	0.846	0.628	0.885	0.780	16.3
YOLOv7	0.865	0.639	0.888	0.811	64
YOLOv8s	0.863	0.636	0.888	0.807	11.1
CSB-YOLOv8s	0.895	0.676	0.889	0.852	11.9

It can be obtained that the majority of the enhanced algorithm's performances, which are suggested in this study, are found to be more effective than those of the mainstream algorithms, demonstrating the approach's feasibility and excellence.

C. Visualization and Analysis

In order to visualize the prediction effect of BCS-YOLOv8s, inference experiments are conducted in this paper. Representative images are predicted using the improved model versus the original model. These pictures contain cross-scale targets and an abundance of little targets, which are objective and suitable for tests of prediction. Figure 12 and Figure 13 displays a comparison of the forecast outcomes.

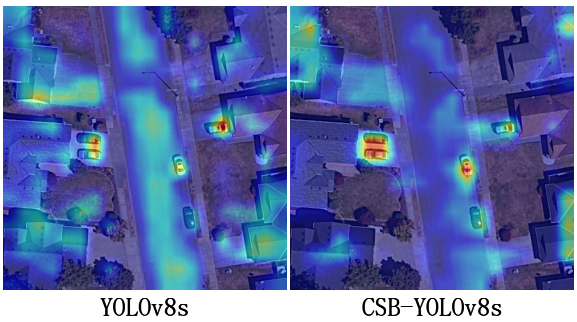

Fig. 12. Comparison of Heat map

In the research, heat maps were produced using Grade CAM. The areas of the feature map that the algorithm concentrates on are shown visually via the heat maps. Regions in the feature map with high confidence gradient values tend to be more dark red, and regions with low gradient values tend to be more dark blue shaded. Figure 12 displays a heat map contrast of the upgraded model with the original

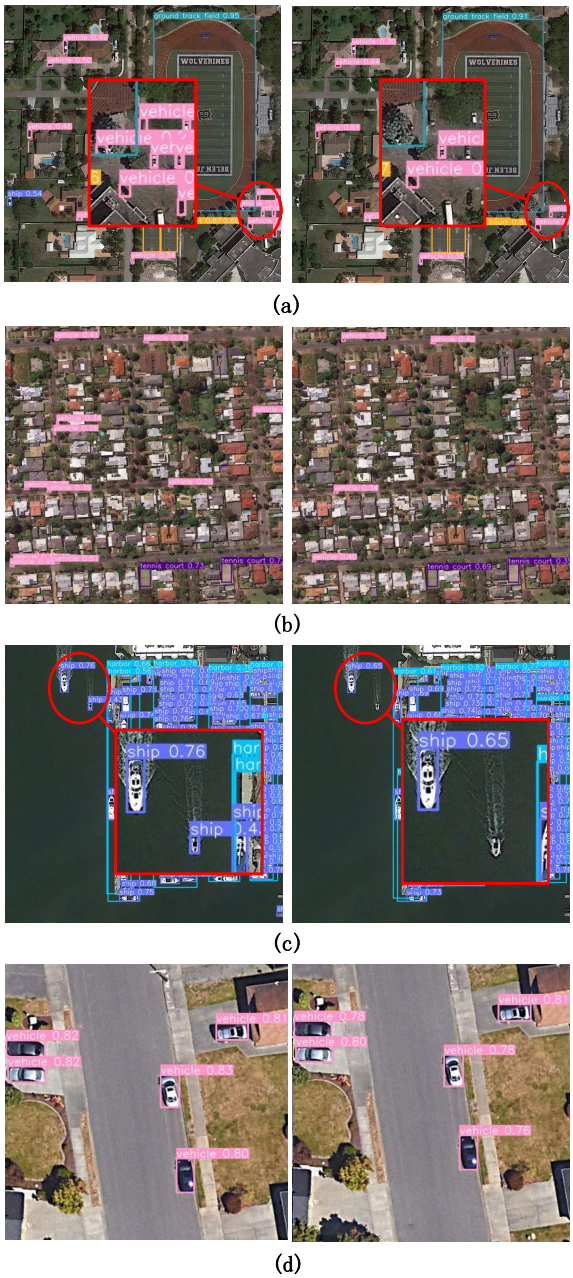

Fig. 13. Comparison of Projected results

model. The graphic illustrates how the YOLOv8s model is readily disrupted by background noise and pays little attention to tiny objects. While the BCS-YOLOv8s model pays more attention to small targets and more effectively muffles background noise, which makes the model's attention more focused on the target and improves the overall model performance.

In the Figure 13, the left side shows the prediction results of the improved model BCS-YOLOv8s, and the right side shows the prediction results of the original model YOLOv8s. 13(a) the original model omits the small target car in the lower right corner, and the improved model predicts it correctly; 13(b) the original model also has a large number of omissions, and the correctness of the prediction of the improved model is significantly improved; 13(c) the original model omits the boat in the upper left corner, and the improved model predicts correctly, and the accuracy of other

targets is improved; the enhanced model's capacity for prediction in 13(d) is noticeably better than the original model's. This proves that the improvement proposed in this paper substantially diminishes the omission rate and enhances the accuracy of predicting small and dense targets.

V. Conclusion

In this paper, the initial backbone portion is supplemented with the BRA attention to enhance the model's attention to crucial information, particularly the small targets; second, in order to address the issue with semantic information loss of the tiny objects during the feature fusion process, CARAFE is utilized to substitute the up-sampling module in the neck; meanwhile, a new small-target detection layer is added to enhance the model's ability of small target detection by introducing a shallow feature map in the backbone into the feature fusion; and finally, the SIoU is used to improve the pre-selected frame regression performance to progressively optimize the model efficiency. The model's detection accuracy for dense, tiny objects in remote sensing pictures can be enhanced by the aforementioned changes. Without a significant increase in model parameters, the mAP of the algorithm is 89.5% on the DIOR dataset. The improved average detection accuracy is 3.2% higher compared to the original model, which significantly improves the target detection performance and validates the generalization of the improvement on different datasets. In addition, the improved model outperforms the classical methods in the same class in terms of detection accuracy.

Due to the addition of a new detection layer after the improvement, the complexity of the model structure increases, resulting in an increase in the model FLOPs, and there is still potential for enhancement of the computational resource consumption. Therefore, the next step of the research focuses on reducing the model resource consumption and the number of parameters to make the model lightweight without affecting the model detection effectiveness.

References

[1] G. Cheng and J. Han, "A Survey on Object Detection in Optical Remote Sensing Images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[4] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[9] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2778–2788.

[10] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot Multibox Detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[12] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and Efficient Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790.

[13] H. Zhou, A. Ma, Y. Niu, and Z. Ma, "Small-Object Detection for UAV-Based Images Using a Distance Metric Method," *Drones*, vol. 6, no. 10, p. 308, 2022.

[14] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and Feature Fusion SSD for Remote Sensing Object Detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[15] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "Uav-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective," *Sensors*, vol. 20, no. 8, p. 2238, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3377–3390, 2019.

[18] T. Wu and Y. Dong, "YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition," *Applied Sciences*, vol. 13, no. 24, p. 12977, 2023.

[19] K. Wang and Z. Liu, "BA-YOLO for Object Detection in Satellite Remote Sensing Images," *Applied Sciences*, vol. 13, no. 24, p. 13122, 2023.

[20] S. Li and W. Liu, "Small Target Detection Model in Aerial Images Based on YOLOv7X+," *Engineering Letters*, vol. 32, no. 2, pp. 436–443, 2024.

[21] X. Zhang and Y. Tian, "Traffic Sign Detection Algorithm Based on Improved YOLOv8s," *Engineering Letters*, vol. 32, no. 1, pp. 168–178, 2024.

[22] Y. Zhang, M. Ma, Z. Wang, J. Li, and Y. Sun, "POD-YOLO Object Detection Model Based on Bi-directional Dynamic Cross-level Pyramid Network," *Engineering Letters*, vol. 32, no. 5, pp. 995–1003, 2024.

[23] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: Vision Transformer with Bi-Level Routing Attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 323–10 333.

[24] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "Carafe: Content-Aware Reassembly of Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3007–3016.

[25] 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[26] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.