RESEARCH ARTICLE

YOLOv7-Bw: A Dense Small Object Efficient Detector Based on Remote Sensing Image

Xuebo Jin[1], **Anshuo Tong**[1], **Xudong Ge**[1], **Huijun Ma**[2,*], **Jiaxi Li**[2], **Heran Fu**[2] **and Longfei Gao**[2]

[1] School of Computer Science and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
[2] National Engineering Laboratory for Agri-product Quality Traceability, BTBU, Beijing, China

Abstract

In recent years, deep learning techniques have been increasingly applied to the detection of remote sensing images. However, the substantial size variation and dense distribution of objects in these images present significant challenges to detection algorithms. Current methods often suffer from low efficiency, missed detections, and inaccurate bounding boxes. To address these issues, this paper presents an improved YOLO algorithm, YOLOv7-bw, designed for efficient remote sensing image detection, thereby advancing object detection applications in the remote sensing industry. YOLOv7-bw enhances the original SPPCSPC pooling pyramid network by incorporating a Bi-level Routing Attention module, which focuses on densely populated target areas to improve the network's feature extraction capabilities. Additionally, it introduces a dynamic non-monotonic WIoUv3 loss function to replace the original CIoU loss function. This substitution ensures that the loss function's gradient allocation strategy aligns more effectively with the current detection scenario, enhancing the network's focus on the detection object. Through comparative experiments on the DIOR remote sensing image dataset, we found that YOLOv7-bw achieved a high mAP@0.5 of 85.63% and a high mAP@0.5:0.95 of 65.93%, surpassing the previous results of 83.7% and 63.9% by approximately 1.93% and 2.03%, respectively. Moreover, compared with commonly used algorithms, YOLOv7-bw demonstrated superior performance, thereby validating the feasibility and enhanced applicability of our proposed algorithm for remote sensing image detection.

Keywords: remote sensing image, YOLO, object detection, mAP.

1 Introduction

Optical remote sensing images [1] are top-down perspective images captured by aerial vehicles or satellites. This unique imaging method results in significant differences compared to conventional daily images. With the continuous advancement of remote sensing technology, the quality of these images has markedly improved, thereby placing higher demands on remote sensing image processing technologies. Remote sensing images have numerous applications and hold significant value in both military and civilian

Academic Editor:
Jing Na

Submitted: 16 February 2024
Accepted: 19 May 2024
Published: 27 May 2024

Vol. 1, **No.** 1, 2024.
10.62762/TIS.2024.137321

***Corresponding author:**
✉ Huijun Ma
mahuijun@th.btbu.edu.cn

Citation
Jin, X., Tong, A., Ge, X., Ma, H., Li, J., Fu, H., & Gao, L. (2024). YOLOv7-Bw: A Dense Small Object Efficient Detector Based on Remote Sensing Image. *ICCK Transactions on Intelligent Systematics*, 1(1), 30–39.

domains. In the military field, remote sensing image data are crucial for processing and analyzing collected intelligence and reconnaissance information. The insights gained can inform various scenarios, such as operational planning and military deployments. In civilian applications, remote sensing images yield valuable results in areas such as land use [2], urban planning, traffic monitoring [3], disaster prevention [4], and ecological protection [5].

Remote sensing images possess characteristics such as large coverage areas, diverse object types, dense objects, and high background complexity, which present significant challenges for detection tasks. Traditional remote sensing image detection methods can be roughly categorized into four types: template matching-based methods [2], shape texture-based methods [3], image segmentation-based methods [4], and visual saliency-based methods [5]. From these methods, it is evident that traditional approaches generally first construct a universal target template and then perform global image matching. Alternatively, potential object areas can be segmented first, followed by using simple feature rules for discrimination. This type of method is prone to a large number of erroneous instances in the detection results, leading to low accuracy and limited applicability. Consequently, these methods can only be effectively used for detecting objects in simple and uniform backgrounds.

Due to the massive growth of data and the improvement of hardware computing power, the theory and technology of deep learning have also developed rapidly, and more and more deep learning methods are being applied in the field of remote sensing image object detection. Deep learning-based object detection algorithms can be divided into region suggestion-based methods (two-stage methods) and regression-based methods (single-stage methods) based on whether region suggestions are generated or not. Two-stage object detectors, such as Faster R-CNN [6], Libra R-CNN [7], and Mask R-CNN [8], first extract regions of interest and then perform further detection and recognition for each region. Although the overall detection accuracy is relatively high, the need to first extract regions of interest and separately classify and regress each region adds additional computational complexity and reduces speed, making it difficult to apply in systems with high real-time requirements. On the other hand, a single-stage object detector does not need to generate individual candidate regions and treats the entire detection process as a whole. It directly regresses and analyzes the bounding boxes and categories of the object from multiple positions in the input image. Typical representative algorithms include the YOLO [9] series, SSD [10], and FCOS [11]. The single-stage algorithm has a fast object detection speed and basically meets the requirements of real-time systems, but the detection accuracy is slightly lower than that of the two-stage object detection method. Overall, deep learning-based methods can automatically obtain deep semantic features of images through training and have stronger expressive power than manually designed features. Additionally, these methods are more sensitive to factors such as the spatial and dense distribution of objects in the image, while their sensitivity to the category of the object is low. Therefore, deep learning-based methods usually do not detect a single type of object but can detect multiple types of objects, which is more in line with the practical application of remote sensing images and has become the mainstream development direction of remote sensing image object detection [15, 16].

In recent years, with the advancement of remote sensing image detection technology, its application scope has gradually expanded, and various improved algorithms have emerged, resulting in significant improvements in detection accuracy and efficiency. Among them, Li et al. [17] proposed a dual-channel feature fusion network that can learn local and contextual attribute features along two independent paths, forming a powerful joint representation to achieve effective detection of remote sensing image objects. Yang et al. [18] proposed an end-to-end rotation detection box object detection algorithm, which improved the detection accuracy of ships. Zhang et al. [19] designed a multi-scale detection network structure based on the YOLOv5s model, which enhanced the detection performance of objects in monitoring scenarios. Jiang et al. [20] combined bijective neural networks and displacement localization strategies to address the problem of narrow bounding boxes for small remote sensing objects. Wang et al. [21] established dense connections between shallow and deep feature maps, solving the problem of significant changes in ship scale. Yang et al. [22] combined multi-layer features with effective anchor sampling to improve sensitivity to small objects. Yao et al. [23] generated high-quality semantic features by introducing an expanded bottleneck structure into the feature pyramid network. Yan et al. [24] improved the performance of small object detection by fusing features across hierarchical channels to retain accurate position information of weak objects. Zhang et al. [25] obtained multiple receptive field features by fusing features from different levels and constructed a new cascaded attention mechanism to enhance the ability to capture features of small remote sensing objects.

This article adopts YOLOv7 as the basic algorithm and proposes YOLOv7-bw. To achieve better performance in precision and dense object prediction tasks, YOLOv7 uses "scaling" and "scaling," addressing the problem of dynamic label allocation and the replacement of reparameterization modules, thereby making the object detector faster and more effective.Secondly, due to the common issues of long shooting distances and blurry imaging in remote sensing images, the WIoUv3 [12] loss function is adopted in the boundary box regression part of the loss function. This approach helps to better focus on and locate the objects to be detected, thereby improving detection accuracy.Finally, for remote sensing images with small objects, clustering of objects can occur easily. To address this, a Bi-level Routing Attention (BRA) module is introduced to better focus on dense object areas and solve the problem of the algorithm's inability to recognize small objects in densely populated areas.

2 Related Work

2.1 Attention mechanism

The attention mechanism in image processing has become one of the popular and important technologies in the field of deep learning [26, 27]. Due to its excellent plug-and-play convenience, it is widely used in various machine learning models [28]. The attention mechanism enhances the model's focus on the most critical regions by weighting the input features, thereby improving the accuracy and performance of image processing tasks. In the early years, the basic idea of the attention mechanism was to first input Query, Key, and Value.

The correlation between Query and Key is calculated to obtain attention scores. After scaling the attention scores (divided by the square root of the dimension), softmax is applied to normalize and obtain the weight coefficients. Finally, the Value values are weighted and summed based on the weight coefficients to obtain the Attention Value, which focuses on the key areas and ignores irrelevant regions. Over time, Vaswani et al. [13] first applied the self-attention mechanism in the field of NLP (Natural Language Processing) and successfully introduced it into the field of computer vision, demonstrating the enormous potential of self-attention models. Unlike ordinary attention mechanisms, self-attention mechanisms reduce dependence on external information and are better at capturing internal correlations of data or features. The key point of the self-attention mechanism is that Q, K, and V are the same variable, or all three originate from the same X, making them homologous. By finding the key points within X, the model can pay more attention to the essential information of X and ignore the unimportant information. This mechanism does not occur between input and output statements but rather between internal elements of input or output statements. However, self-attention in typical global context modeling, such as vanilla attention, calculates the affinity of paired features at all spatial positions, resulting in a high computational burden and heavy memory usage, especially for high-resolution inputs. Therefore, in recent years, research on self-attention modules has been devoted to alleviating this high computational burden, and more work has begun to introduce different manually crafted sparse patterns.

2.2 YOLO

The YOLO series is one of the best-performing algorithms in the current field of object detection. It has significant advantages in recognition accuracy and speed, enabling real-time object detection. Consequently, it is widely used in various industries. Among them, YOLOv7 [14] achieves near-optimal accuracy while maintaining its speed advantage. Its structure is primarily divided into three modules: the input end, the feature extraction backbone network (Backbone), and the detection head output end (Head).

YOLO's real-time detector has been widely recognized and applied in many scenarios by researchers since its inception. It uses a loss function weighted by Bounding Box Regression (BBR) loss, classification loss, and object loss to construct the model. To date, this structure remains the most effective loss function paradigm in object detection tasks, where BBR loss directly affects the localization performance of the model. To further improve the positioning performance of the model, it is essential to design an effective BBR loss.

IoU (Intersection over Union) is used to measure the degree of overlap between predicted and ground truth boxes in object detection tasks. Its calculation is the ratio of the intersection area of the predicted and true boxes to their union area. However, IoU has a significant flaw as a loss function: the gradient of backpropagation disappears when there is no overlap between bounding boxes. This results in the inability to update the overlapping area width between bounding boxes during the training process. To address this issue, existing research has considered many geometric factors related to bounding boxes and constructed penalty terms. Currently, bounding box loss is based on additive loss, using CIoU (Complete IoU). This loss function aligns with the mechanism of target box regression by considering the distance, overlap, scale, and aspect ratio between the target and prediction, thereby making target box regression more stable and avoiding problems such as divergence during training, unlike IoU. However, one of the current drawbacks of YOLOv7 is that CIoU still exhibits certain errors when processing bounding boxes with large aspect ratios, which may lead to poor accuracy of the bounding boxes.

In summary, although YOLOv7 boasts significant advantages in speed and accuracy as an object detection algorithm, it still faces challenges, particularly in handling bounding boxes with large aspect ratios. Future research can focus on refining the bounding box loss function and developing methods to better manage aspect ratios, thereby enhancing the performance of YOLOv7.

3 Methodology

3.1 WIoU

3.1.1 WIoUv1

The block diagram of BBR loss is shown in Figure 1.

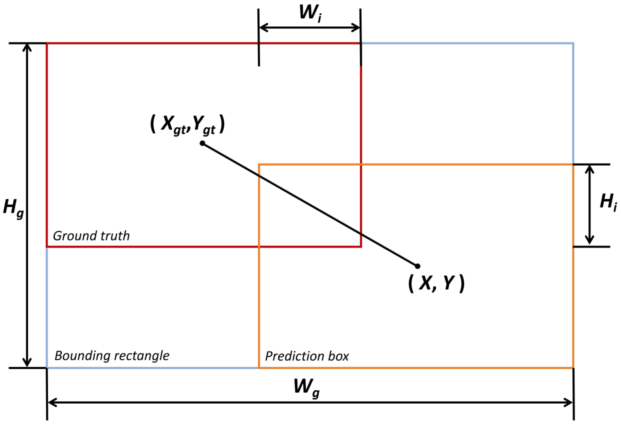


Figure 1. Example of BBR loss.

The loss function used in YOLOv7 is CIoU, which adds consideration for aspect ratio consistency based on the normalized length of the center point connection. Although it addresses the issue where the prediction box cannot be optimized when the negative gradient $\frac{\partial RDIoU}{\partial W_g}$ and $\frac{\partial \mathcal{L}IoU}{\partial W_{\mathrm{g}}}$ cancel out, it inevitably generates many low-quality anchor boxes during the prediction process. Adding geometric metrics such as aspect ratio or distance exacerbates the punishment of these low-quality anchor boxes, thereby reducing the model's generalization ability. An effective loss function should mitigate the penalty of geometric metrics when the anchor box and target box align well, without excessively interfering with training. Based on these principles, distance attention was constructed using distance measurement, breaking away from the traditional additive anchor box loss. The two were multiplied to obtain WIoUv1, incorporating a two-layer attention mechanism:

$$\mathscr{L}_{WIoUv1} = \mathscr{R}_{WIoU}\mathscr{L}_{IoU} \tag{1}$$

$$\mathscr{R}_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{2}$$

where $\mathcal{R}WIoU \in [1, e)$. This significantly amplifies the LIoU of ordinary quality anchor boxes while substantially reducing the RWIoU of high-quality anchor boxes. To prevent RWIoU from generating gradients that hinder convergence, W_g and H_g are separated from the computational graph (indicated by the superscript * to denote this operation).

3.1.2 WIoUv3

Although WIoU v1 can be applied to most scenarios, it still cannot adequately address the problem of small objects in remote sensing images. To better focus on small objects, we opted to use WIoUv3. This version adds a focusing mechanism by constructing a gradient gain (focusing coefficient) calculation method based on v1, replacing the CIoU loss function in YOLOv7. WIoUv3 introduces the concept of "outlier" to describe the quality of anchor boxes, which is specifically defined as:

$$\beta = \frac{\mathscr{L}_{IoU}^*}{\mathscr{L}_{IoU}} \in [0, +\infty) \tag{3}$$

where $\overline{LIoU}$ represents the sliding average with momentum m. A small outlier indicates a high-quality anchor box, so a small gradient gain is allocated to it. Simultaneously, smaller gradient gains are also allocated to anchor boxes with higher outliers, effectively preventing low-quality examples from generating larger harmful gradients. This ultimately focuses the bounding box regression on anchor boxes of ordinary quality. Utilizing β, a non-monotonic focusing coefficient is constructed and applied to WIoU v1:

$$\mathscr{L}_{WIoUv3} = r\mathscr{L}_{WIoUv1}, \quad r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{4}$$

where r represents the gradient gain, and α and δ are artificially set hyperparameters. Since $\overline{LIoU}$ is dynamic, the quality division standard of the anchor box is also dynamic. This enables WIoUv3 to implement the most suitable gradient gain allocation strategy for the current situation at every moment.

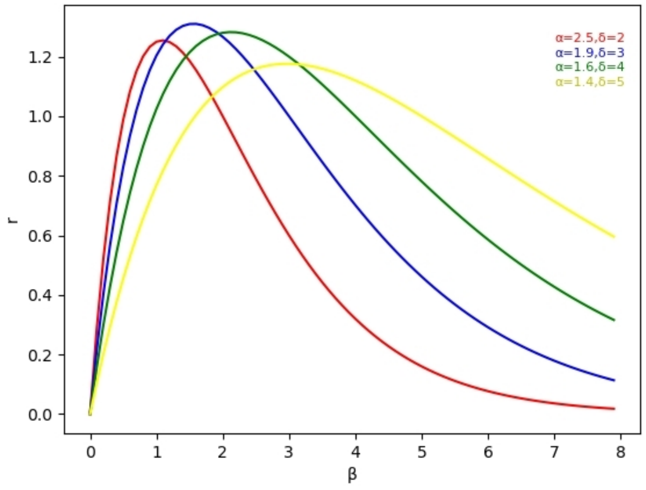
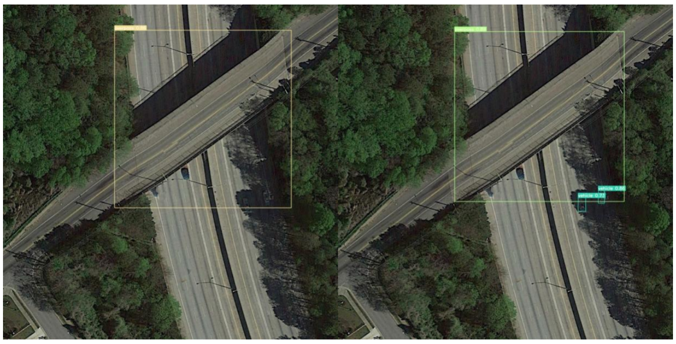


Figure 2. Shows hyperparameters α, δ Controlled β Mapping with gradient gain γ.

After analyzing the influence of gradient gain r on outliers β for several different sets of α and δ, as shown in Figure 2, it can be seen that the blue curve demonstrates better performance. It exhibits smaller gradient gains at both low and high outliers, making the loss function more focused on anchor boxes of ordinary quality. Ultimately, we selected the hyperparameters $\alpha = 1.9$ and $\delta = 3$ for the final experiment (the experiment for determining hyperparameters is provided in Part 3). Additionally, to prevent low-quality anchor boxes from being left behind during early training, we initialized $\overline{LIoU}$ so that when LIoU = 1, it has the maximum gradient gain.

Remote sensing images inherently feature small objects, which are often affected by weather conditions and may also be obscured by shadows, greatly increasing the difficulty of detection. Figure 3(a) shows the detection result using the YOLOv7 source code, which missed three small cars. After applying WIoU v3 to YOLOv7, although one small car was still missed, the unique dynamic non-monotonic focusing mechanism resulted in improvements. As shown in Figure 3(b), both the confidence in object detection and the detection of blurry small objects in shadows have been enhanced to a certain extent.



(a) (b)

Figure 3. Impact of replacing WIoU on YOLO.

3.2 Bi-level Routing Attention(BRA)

BRA is a dynamic, query-aware sparse attention mechanism designed to make each query focus on a small subset of the most semantically relevant key-value pairs. The core idea is to filter out the least relevant key-value pairs at the coarse region level, thus retaining only a small portion of the routing area. Then, fine-grained labels are applied to the attention of the labels in these routing areas. Because BRA involves only dense matrix multiplication, it achieves good performance while maintaining high computational efficiency. The specific steps can be roughly divided into the following three parts:

1) Regional division and input projection:

Input a two-dimensional feature map, $X \in \mathbb{R}^{H\times W\times C}$, and first divide it into $S \times S$ non-overlapping regions, where each region contains $\frac{HW}{S^2}$ feature vectors. This transforms X into $X^r \in \mathbb{R}^{S^2\times\frac{HW}{S^2}\times C}$. Then, derive the linear projection of the query, key, and value as follows:

$$Q = X^r W^{\mathrm{q}}, K = X^r W^{\mathrm{k}}, V = X^r W^{\mathrm{v}} \tag{5}$$

where $W^{\mathrm{q}}, W^{\mathrm{k}}, W^{\mathrm{v}} \in \mathbb{R}^{C\times C}$ are the projection weights for the query, key, and value, respectively.

2) Routing from region to region with a directed graph:

Building on the first step, we establish the participation relationship by constructing a directed graph. First, apply the average value of each region to Q and K separately to obtain $Q^r, K^r \in \mathbb{R}^{S^2\times C}$, and then calculate the adjacency matrix A^r representing the inter-region correlation between Q^r and K^r:

$$A^r = Q^r(K^r)^T \tag{6}$$

The adjacency matrix $A^r \in \mathbb{R}^{S^2 \times S^2}$ measures the degree of semantic correlation between two regions. Next, only the top k connections of each region are retained to trim the correlation graph. Specifically, using the routing index matrix $I^r \in \mathbb{N}^{S^2 \times k}$ and the row-wise top-K operator, the indices of the top k connections are saved row by row:

$$I^r = topkIndex(A^r) \tag{7}$$

where i-th row of I^r contains the index of the first k most relevant regions of the i-th region.

3) Token to token attention:

By utilizing the routing index matrix I^r from region to region, we can apply fine-grained labeling attention. For each query marker in region i, it will focus on all key-value pairs located in the k routing regions and concentrate them together. Specifically, first collect the key and value tensors:

$$K^g = gather(K, I^r), \quad V^g = gather(V, I^r) \tag{8}$$

where K^g and V^g are tensors of the aggregated key and value, and then attention operations are used on the aggregated key-value pairs:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{C}})V \tag{9}$$

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{10}$$

We introduce a context enhancement term $LCE(V)$ here, where the function $LCE(.)$ is parameterized using depthwise separable convolution. The convolution kernel size is set to 5.

In general, BRA collects key-value pairs from the top k relevant windows, utilizes sparsity to skip the calculation of the least relevant regions, and only involves GPU-friendly dense matrix multiplication, as shown in Figure 4.

Specifically, BRA is integrated into the YOLOv7 network. It first performs region-to-region routing on the previously extracted feature maps, and then applies token-to-token attention to obtain new output feature maps. Given that BRA operates on feature maps and the spatial pooling pyramid SPPCSPC aims to avoid image distortion caused by image processing operations and duplicate feature extraction, it was decided to insert the BRA module after SPPCSPC, specifically into the YOLOv7 network.

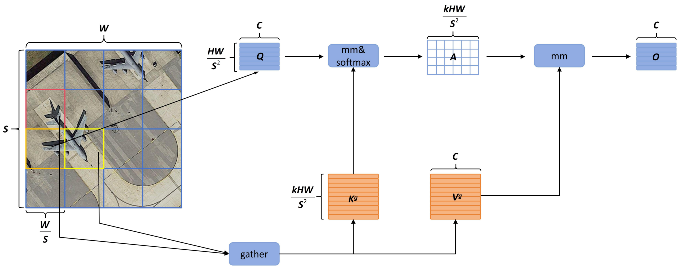


Figure 4. Schematic diagram of BRA principle.

In many remote sensing images, dense small objects are clustered in specific areas. BRA focuses on these areas and applies fine-grained attention mechanisms, which align well with the characteristics of remote sensing images. The actual situation is illustrated in Figure 5. The left Figure 5(a) shows the detection results using the YOLOv7 source code, while the right Figure 5(b) shows the detection results after adding a BRA module on top of the source code. It can be seen that for densely packed small cars, adding the BRA module increases the number of detected cars and reduces the number of falsely detected roofs, thereby improving overall accuracy.

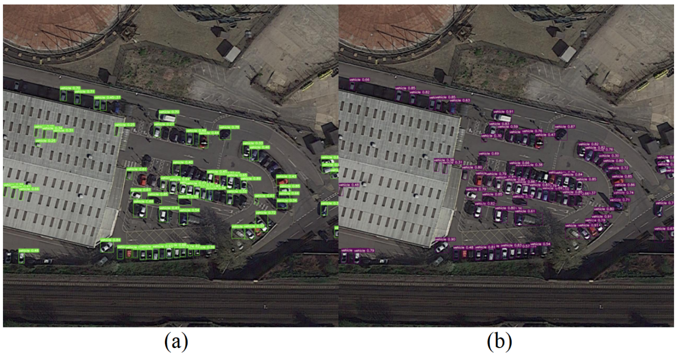
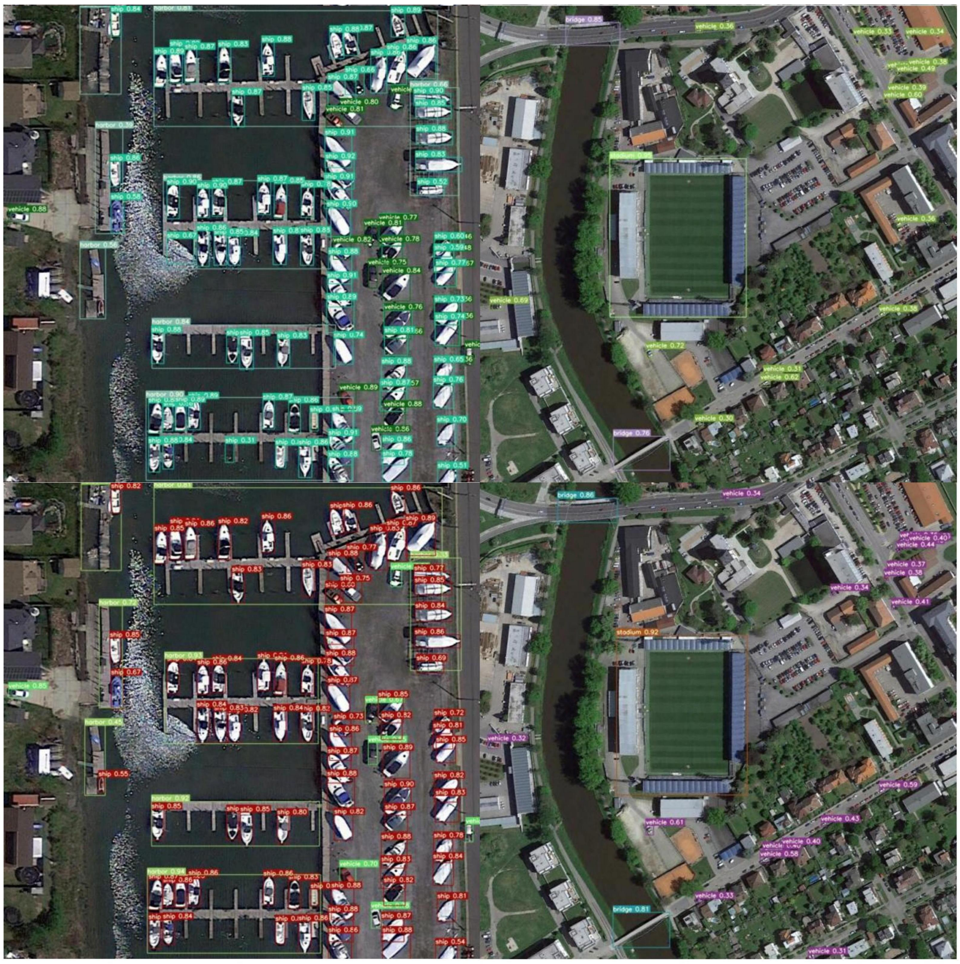


Figure 5. The impact of adding BRA on YOLO.

4 Experiments

4.1 WIoU Hyperparameter experiment

To evaluate the effectiveness of the model, we chose Precision, Recall, and mAP as evaluation metrics. Precision refers to the percentage of predicted true positive samples relative to the total predicted positive



(a) (b)

Figure 6. YOLOv7 and YOLOv7-bw detection results.

samples, and is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

Recall is the percentage of predicted true positive samples relative to the entire true positive sample, and is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

The P-R curve represents the relationship between precision and recall. With Recall on the horizontal axis and Precision on the vertical axis, a curve for a particular category is drawn. The area enclosed under this curve is the AP value for that category.

The mAP (mean Average Precision) is the overall evaluation index of an object detection algorithm for a given dataset, representing the average AP value of all categories. It is currently the most important indicator for evaluating the performance of a model, and is expressed as:

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{13}$$

To determine which hyperparameters are most suitable for application to WIoUv3, we designed a set of comparative experiments based on the PyTorch framework. We selected 20 categories from the MS-COCO dataset, with 28,474 images as training data and 1,219 images as validation data. For the model, we chose YOLOv7-w6 with a layer channel multiplier of 0.75 for training. These models were trained for 120 epochs with batch sizes of 32 and different BBR losses. The experimental results are shown in Table 1.

From Table 1, we can see that WIoUv3 with dynamic non-monotonic focusing performs better than WIoUv1 with only a focusing mechanism. Meanwhile, when $\alpha = 1.9$ and $\delta = 3$, WIoUv3 showed the best performance across different IoU thresholds,

Table 1. Comparison of different WIoU versions and hyperparameters on MS-COCO (20 categories).

Method	mAP@0.5	mAP@0.5:0.95	Recall
WIoU v1 (focusing only)	52.10	61.75	44.20
WIoU v3 (dynamic focusing)	53.80	63.50	45.10
WIoU v3 ($\alpha = 1.9, \delta = 3$)	**54.50**	**64.20**	**45.68**

with values of 54.50, 64.20, and 45.68, respectively. Ultimately, we decided to use $\alpha = 1.9$ and $\delta = 3$ for the final experiment after setting the hyperparameters.

4.2 Ablation experiment

To verify the effectiveness of each module in the model, we conducted a set of ablation experiments. The dataset used is the DIOR remote sensing image dataset, collected from Google Earth by experts in the field of Earth observation interpretation. This dataset includes 23,463 remote sensing images and 190,288 object instances. These object instances are manually marked with axially aligned bounding boxes, covering 20 common object categories, namely airplanes, airports, baseball fields, basketball courts, bridges, chimneys, dams, highway service areas, highway toll stations, ports, golf courses, ground track and field fields, overpasses, ships, sports fields, storage tanks, tennis courts, train stations, vehicles, and windmills.

We selected several images from the test dataset and compared the actual effects of the YOLOv7 source code and our improved YOLOv7-bw. The comparison results are shown in Figure 6. The first image demonstrates the effect of YOLOv7, and the second image demonstrates the effect of YOLOv7-bw. From Figure 6 (a), it can be seen that YOLOv7-bw has fewer false detections of ships as vehicles compared to YOLOv7. From Figure 6 (b), although there are still many vehicles that have not been detected, YOLOv7-bw detected 18 vehicles, which is higher than the 15 vehicles detected by YOLOv7, indicating better performance in vehicle positioning and focusing.

4.3 Comparison with other methods

To further verify the effectiveness of the improved YOLOv7 presented in this article, we selected commonly used algorithms in the field of object detection for comparison, including classic algorithms such as RCNN, SSD, RetinaNet, and CornerNet. The DIOR remote sensing dataset was used for training and testing under the same conditions. The results are shown in Table 2.

Table 2. Comparative experiments.

method	Backbone	mAP
R-CNN	VGG-16	37.7
Faster R-CNN	VGG-16	54.1
SSD	VGG-16	58.6
RetinaNet	ResNet-101	66.1
PANet	ResNet-101	66.1
CornerNet	Hourglass-104	64.9
YOLOv7	ELAN	83.7
YOLOv7-bw(ours)	ELAN	85.6

From Table 2, it can be seen that our proposed YOLOv7-bw algorithm achieves the best mAP@0.5 at an IoU threshold of 0.5. This indicates that the improved YOLOv7-bw algorithm optimizes the detection performance of remote sensing images, enhances overall detection performance, and makes the models competitive. The optimization effect is particularly evident for small and dense objects, meeting the practical detection needs of remote sensing images.

5 Conclusion

To address the problem of dense and blurred objects in remote sensing images, this paper proposes the YOLOv7-bw algorithm based on the YOLOv7 model, introducing the BRA attention mechanism to focus on densely populated object areas. Additionally, we replaced the loss function with WIoUv3 to better focus on and locate the objects to be detected. The algorithm was tested on the DIOR remote sensing image dataset. The experimental results showed that our YOLOv7-bw achieved mAP@0.5 and mAP@0.5:0.95 values of 85.63% and 65.93%, respectively, which were optimal results, demonstrating the feasibility of our algorithm.

However, during the experiment, we also identified shortcomings in the algorithm. Although the overall number of detected vehicles is higher than with YOLOv7, many small vehicles were still not detected and can be optimized. In future research, we will focus more on improving the model's ability to recognize small objects, aiming to achieve better performance in the field of remote sensing images.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Guang-Tao, N., & Hua, H. (2021). A survey of object detection in optical remote sensing images. *Acta Automatica Sinica*, *47*(8), 1749-1768.

[2] Liu, G., Sun, X., Fu, K., & Wang, H. (2012). Aircraft recognition in high-resolution satellite images using coarse-to-fine shape prior. *IEEE Geoscience and Remote Sensing Letters*, *10*(3), 573-577. [CrossRef]

[3] Liu, Q., Xiang, X., Wang, Y., Luo, Z., & Fang, F. (2020). Aircraft detection in remote sensing image based on corner clustering and deep learning. *Engineering Applications of Artificial Intelligence*, *87*, 103333. [CrossRef]

[4] Zhu, C., Zhou, H., Wang, R., & Guo, J. (2010). A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Transactions on geoscience and remote sensing*, *48*(9), 3446-3456. [CrossRef]

[5] Bi, F., Zhu, B., Gao, L., & Bian, M. (2012). A visual search inspired computational model for ship detection in optical satellite images. *IEEE Geoscience and Remote Sensing Letters*, *9*(4), 749-753. [CrossRef]

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, *39*(6), 1137-1149. [CrossRef]

[7] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019, June). Libra R-CNN: Towards Balanced Learning for Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 821-830). IEEE. [CrossRef]

[8] Nie, X., Duan, M., Ding, H., Hu, B., & Wong, E. K. (2020). Attention mask R-CNN for ship detection and segmentation from remote sensing images. *IEEE Access*, *8*, 9325-9334. [CrossRef]

[9] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016, June). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 779-788). IEEE. [CrossRef]

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, September). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Cham: Springer International Publishing. [CrossRef]

[11] Tian, Z., Shen, C., Chen, H., & He, T. (2019, October). FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision* (*ICCV*) (pp. 9626-9635). IEEE. [CrossRef]

[12] Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). Wise-IoU: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[14] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023, June). YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 7464-7475). IEEE. [CrossRef]

[15] Cai, W., Qian, P., Ding, Y., Bi, M., Ning, X., Hong, D., & Bai, X. (2023). Graph-structured convolution-guided continuous context threshold-aware networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1-18. [CrossRef]

[16] Cai, W., Gao, M., Ding, Y., Ning, X., Bai, X., & Qian, P. (2023). Stereo attention cross-decoupling fusion-guided federated neural learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1-16. [CrossRef]

[17] Li, X., Ding, M., & Pižurica, A. (2021). Spectral feature fusion networks with dual attention for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-14. [CrossRef]

[18] Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., & Yan, J. (2021). Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, *34*, 18381-18394. [CrossRef]

[19] Zhang, M., Liu, T., Piao, Y., Yao, S., & Lu, H. (2021, October). Auto-msfnet: Search multi-scale fusion network for salient object detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 667-676). [CrossRef]

[20] Jiang, S., Zhang, J., Wang, W., & Wang, Y. (2023). Automatic inspection of bridge bolts using unmanned aerial vision and adaptive scale unification-based deep learning. *Remote Sensing*, *15*(2), 328. [CrossRef]

[21] Wang, Y., Wang, L., Wang, H., & Li, P. (2019). End-to-end image super-resolution via deep and shallow convolutional networks. *IEEE Access*, *7*, 31959-31970. [CrossRef]

[22] Yang, F., Li, W., Hu, H., Li, W., & Wang, P. (2020). Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images. *Sensors*, *20*(6), 1686. [CrossRef]

[23] Yao, H., Yu, W., Luo, W., Qiang, Z., Luo, D., & Zhang, X. (2023). Learning global-local correspondence with semantic bottleneck for logical anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(5), 3589-3605. [CrossRef]

[24] Yan, R., Yan, L., Cao, Y., Geng, G., & Zhou, P. (2024). One-stop multiscale reconciliation attention network with scribble supervision for salient object detection in optical remote sensing images. *Applied Intelligence*, *54*(5), 3737-3755. [CrossRef]

[25] Zhang, H., & Wu, Y. (2024). CSEF-Net: Cross-Scale SAR Ship Detection Network Based on Efficient Receptive Field and Enhanced Hierarchical Fusion. *Remote Sensing*, *16*(4), 622. [CrossRef]

[26] Roy, A. M., & Bhaduri, J. (2023). DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. *Advanced Engineering Informatics*, *56*, 102007. [CrossRef]

[27] Mahaadevan, V. C., Narayanamoorthi, R., Gono, R., & Moldrik, P. (2023). Automatic identifier of socket for electrical vehicles using SWIN-transformer and SimAM attention mechanism-based EVS YOLO. *IEEE Access*, *11*, 111238-111254. [CrossRef]

[28] Kamilov, U. S., Bouman, C. A., Buzzard, G. T., & Wohlberg, B. (2023). Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, *40*(1), 85-97. [CrossRef]

Xuebo Jin (Fellow, ICCK) received the B.S. and M.S. degreesin control theory and control engineering from Jilin University, Changchun, China, in 1994 and 1997, and the Ph.D. degree in control theory and control engineering from the University of Zhejiang, Zhejiang, China, in 2004. From 2009 to 2012, she was an Assistant Professor with Zhejiang Sci-tech University. Since 2012, she has been a Professor with Beijing Technology and Business University, Beijing, China. Her research includes a variety of areas in information fusion, bigdata analysis, condition estimation, and video tracking. (Email: jinxuebo@btbu.edu.cn)

Anshuo Tong graduated with a Bachelor's degree in Electronic Science and Technology from Beijing Technology and Business University in 2022. Currently, he is a Master's student in Control Engineering at Beijing Technology and Business University. My research focuses on image classification and detection, pattern recognition and information fusion, etc. (Email: tonganshuo@st.btbu.edu.cn)

Xudong Ge graduated from Beijing Technology and Business University with a Master's degree in Control Engineering in 2024. His research focuses on image detection, pattern recognition and information fusion, machine learning, and other related fields. (Email: gexudong@st.btbu.edu.cn)

Huijun Ma graduated from Changchun Institute of Optics and Mechanics with a master's degree in atomic and molecular physics in 2010. She is currently an on-the-job doctoral student in systems science at Beijing Technology and Business University. Research directions include complex system modeling, pattern recognition and information fusion, machine learning, etc. (Email: mahuijun@th.btbu.edu.cn)

Jiaxi Li , a 2022 graduate student, is studying Control Engineering at Beijing University of Business and Technology. My research direction is multi-modal object detection technology and applications. (Email: lijiaxi@st.btbu.edu.cn)

Heran Fu graduated from Beijing Technology and Business University in 2022 with a bachelor's degree in Electronic Science and Technology, and is currently a master's candidate in Control Engineering of Beijing Technology and Business University. His research interests encompass image classification and detection, pattern recognition, deep learning, among others. (Email: fuheran@@st.btbu.edu.cn)

Longfei Gao graduated from the University of Jinan with a bachelor's degree in electrical engineering and automation in 2022. He is now studying in Beijing Technology and Business University, majoring in control engineering. His research interests are image detection and pattern recognition. (Email: gaolongfei@@st.btbu.edu.cn)