YOLOv7-b: An Enhanced Object Detection Model for Multi-Scale and Dense Target Recognition in Remote Sensing Images

Yulong Song, Hao Yang, Lijun Huang, Song Huang
School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

***Abstract*—To address the challenges of dense object distribution, scale variability, and complex shapes in remote sensing images, this paper proposes an improved YOLOv7-b model to enhance multi-scale target detection accuracy and robustness. First, deformable convolution (DCNv2) is introduced into the YOLOv7 backbone to replace the standard convolutions in the last two ELAN modules, thereby providing more flexible sampling capabilities and improving adaptability to irregularly shaped targets. Next, a Bi-level Routing Attention (BRA) module is integrated after the SPPCSPC module, employing both coarse- and fine-grained routing strategies to focus on densely distributed targets while suppressing irrelevant background. Finally, training and evaluation are conducted on the large-scale DIOR remote sensing dataset under unified hyperparameter settings and evaluation metrics, allowing a systematic assessment of the overall model performance. Experimental results show that, compared with the original YOLOv7, the improved YOLOv7-b achieves significant enhancements in Precision, Recall, mAP@0.5, and mAP@0.5:0.95, with mAP@0.5 and mAP@0.5:0.95 reaching 85.72% and 66.55%, respectively. Visualization further demonstrates that YOLOv7-b provides stronger recognition and localization for densely arranged, small-scale, and morphologically complex targets, effectively reducing missed and false detections. Overall, YOLOv7-b delivers higher detection accuracy and robustness in multi-scale remote sensing target detection. By combining deformable convolution with a dynamic sparse attention mechanism, the model excels in detecting highly deformable objects and dense scenes, offering a more adaptive and accurate solution for small-target detection, dense target recognition, and multi-scale detection in remote sensing imagery.**

***Keywords*—YOLOv7-b; remote sensing images; object detection; deformable convolution; bi-level routing attention; multi-scale**

I. Introduction

Optical remote sensing images [1] are typically captured by satellites or high-altitude aircraft from a nadir viewing angle. Compared with ground-based images, they exhibit significant differences in imaging modes and resolutions. With the rapid development of remote sensing technology and the continuous improvement of imaging accuracy, these images have demonstrated broad application prospects in fields such as military reconnaissance, disaster assessment, environmental monitoring, and urban planning [2] [3]. However, multi-scale remote sensing images often face challenges such as large target scale variations, complex deformations, and severe background noise, which pose higher requirements for the accuracy and robustness of target detection algorithms [4] [5].

Traditional remote sensing image object detection methods encompass template matching [6], shape and texture-based approaches [7], segmentation-based techniques [8], and visual saliency-based methods [9]. These typically rely on predefined rules or object shapes for recognition after feature extraction or segmentation [10] [11] [12]. While effective in specific scenarios, they struggle with the diverse appearances of objects in complex backgrounds and numerous types in remote sensing images. Additionally, these methods are susceptible to noise and occlusion, limiting their applicability in large-scale and multi-scenario contexts.

In recent years, deep learning techniques have been extensively utilized in remote sensing image object detection due to their powerful feature learning capabilities [13] [14] [15]. Object detection algorithms can be divided into two-stage detectors (e.g., Faster R-CNN, Mask R-CNN) and one-stage detectors (e.g., YOLO series, SSD, FCOS) based on whether candidate regions are generated. One-stage detectors, which offer faster inference speed and lower computational cost, are more suitable for remote sensing scenarios with high real-time requirements, while two-stage detectors generally achieve higher accuracy at the cost of greater computational overhead [10]. Overall, the end-to-end training paradigm of deep learning is more adaptable to the diverse target distribution characteristics of remote sensing images, enabling a single model to detect multiple types of objects across different scenarios.

Against this backdrop, various improved algorithms have emerged to enhance the accuracy and efficiency of object detection in multi-scale remote sensing images. For instance, Azimi et al. proposed a method combining joint image cascading and feature pyramid networks, using multi-scale convolutional kernels to extract features at different scales and achieving significant accuracy improvements on the DOTA dataset [16]. Deng et al. redesigned the Faster R-CNN architecture by integrating a multi-scale target generation network and a feature fusion module, demonstrating excellent performance on the NWPU VHR-10 and SAR-Ship datasets [17]. Liu et al. introduced an adaptive multi-scale feature enhancement and fusion module, which notably improved detection accuracy on the DOTA and HRSC2016 datasets [18].

For small target detection, Liu et al. enhanced YOLOv2 using a feature-map concatenation strategy to improve detection performance on small-scale targets [19]. Chen et al. proposed a multi-scale spatial and channel attention mechanism to enable deep neural networks to focus more accurately on key target regions [20]. Ying et al. combined multi-scale convolutional kernels with attention mechanisms to significantly improve detection precision for small targets in complex backgrounds [21]. Additionally, Li et al. developed a fast detection method based on YOLOv3, achieving 93.5% accuracy in multi-scale target detection for high-resolution remote sensing images [22]. Zhang et al. introduced a Dual Multi-Scale Feature Pyramid Network (DM-FPN) tailored for small and densely distributed targets, yielding notable results on the DOTA dataset [23]. Meanwhile, Cheng et al. integrated multi-feature fusion with an attention mechanism in MFANet to achieve high detection accuracy across multiple public datasets [24].

Despite advances in remote sensing image detection, three key challenges remain. First, dense target packing leads to excessive suppression of detection boxes during Non-Maximum Suppression (NMS), causing significant missed detections. This requires better integration of multi-level semantics and contextual information to reduce false suppression from overlapping boxes. Second, objects with large scale differences, such as small vehicles and large facilities in the same image, often result in occlusion of smaller objects by larger ones, exacerbated by limited network focusing and localization capabilities. Lastly, high-altitude imaging results in small targets occupying only a few pixels, increasing the proportion of small targets and the likelihood of missed detections. This necessitates enhanced small-target detection and recognition.

The different comparative results observed across datasets primarily stem from variations in dataset characteristics, including target scale distribution, class imbalance, background complexity, and annotation quality. In remote sensing object detection tasks, the arrangement of objects, morphological variations, and the degree of scene interference significantly impact the performance of detection algorithms. For example, in densely arranged object scenarios, the BRA module’s dynamic attention mechanism effectively separates foreground and background information, thereby improving detection accuracy. Meanwhile, in datasets with highly deformed objects, the adaptive sampling mechanism of DCNv2 enhances the model’s ability to accommodate geometric deformations, improving target localization precision. As a result, the proposed algorithm tends to perform better on datasets that contain complex deformations and densely packed objects, whereas its performance gain over standard YOLOv7 might be relatively limited on simpler datasets with more uniform target scales and fewer background distractions. Experimental results also indicate that the proposed method achieves a more significant improvement on complex multi-scale datasets such as DIOR compared to certain simpler datasets, further validating its effectiveness in handling challenging remote sensing object detection tasks. Future research can further explore the adaptability of this method across different types of remote sensing datasets and optimize its generalization capability to ensure stable detection performance in diverse scenarios.

To address these challenges, this paper proposes a multi-scale remote sensing image detection approach named YOLOv7-b. It integrates Deformable Convolution (DCNv2) and Bi-level Routing Attention (BRA) into the YOLOv7 framework. Specifically, YOLOv7-b incorporates Efficient Layer Aggregation Network structures (ELAN and E-ELAN) into the backbone to efficiently train deeper networks by managing gradient paths. DCNv2 is employed to capture deformed targets, addressing issues such as scale variability, viewpoint changes, and local deformations. Additionally, the BRA module is inserted at the feature-fusion stage to enhance attention on densely distributed targets and salient features, significantly reducing missed detections in high-density scenes—a common weakness in conventional detectors.

To validate the proposed algorithm, experiments were conducted on the DIOR dataset, which features high resolution, diverse target categories, and wide scene coverage. Results show that YOLOv7-b significantly enhances detection accuracy and inference speed, especially for dense targets like vehicles and ships. Additionally, the model's localization performance was analyzed across various IoU thresholds and compared extensively with the original YOLOv7. The findings indicate that integrating DCNv2 and BRA modules effectively improves the network's feature extraction and dense-target recognition capabilities.

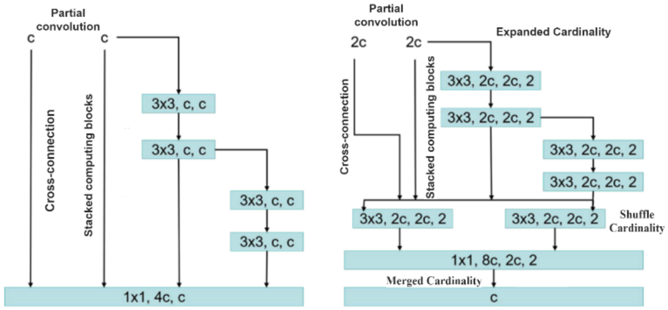
This paper enhances the YOLOv7 architecture with several significant advancements in object detection. The authors integrate Deformable Convolution (DCNv2) into the YOLOv7 backbone to improve feature representation for deformed and multi-scale targets using adaptive offsets and modulation. Additionally, the Bi-level Routing Attention (BRA) module is introduced to address dense target interference and enhance focus on key regions through a sparse attention mechanism. Extensive experiments on the DIOR remote sensing image dataset show that the proposed YOLOv7-b model achieves superior accuracy for multi-scale, high-density, and deformed targets while maintaining high inference speed. These results highlight the model's practical value and potential for real-world applications.

The remainder of this paper is organized as follows: Section II describes the materials and methods, detailing the improved model architecture and experimental setup. Section III presents the experimental results and analysis, including ablation and comparative studies that validate the effectiveness of each module. Section IV summarizes the main findings and contributions, and discusses future work. We hope this research provides valuable insights for remote sensing image target detection and promotes broader applicability in this field.

II. Materials and Methods

A. Model Construction and Improvement

1) Adaptive backbone network: The backbone network of YOLOv7 introduces an efficient layer aggregation structure based on ELAN and E-ELAN. By meticulously controlling the shortest and longest gradient propagation paths within the network, it enables deeper models to learn and converge effectively. The overall framework is shown in Fig. 1.



(a) ELAN (b)E-ELAN

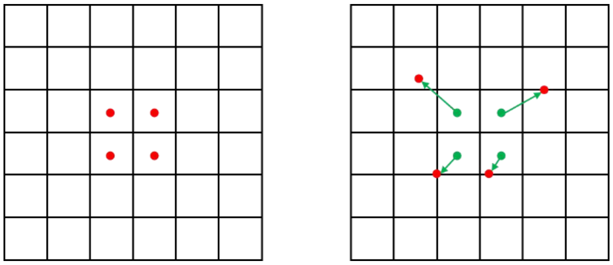
Fig. 1. The structural diagrams of ELAN and E-ELAN.

As can be observed from the figure, the ELAN module employs a gradient path design strategy, which offers multiple advantages. Firstly, adjusting the gradient propagation paths allows the weights of each computational unit to acquire diverse information, thereby achieving higher parameter utilization efficiency. Secondly, since the gradient path directly determines the mode of information transfer and is applied to the weight update process of each computational unit, this strategy ensures that the model maintains a stable learning capability during training and effectively circumvents common degradation issues. Moreover, efficient parameter utilization enables the network to achieve faster inference speeds without the need for additional complex structures, while simultaneously maintaining a high level of accuracy.

However, in larger-scale applications, if computational blocks are stacked without limitation, ELAN can maintain a stable state even under different gradient path lengths or numbers of computational blocks. However, this stability may be broken in extreme stacking situations, thereby leading to a decrease in the efficiency of parameter utilization. To address this issue, researchers further proposed the E-ELAN module, which enhanced the network capability of ELAN through means such as expansion, shuffling, and merging cardinality. It is worth noting that E-ELAN only optimized the internal architecture of the computational blocks, while the structure of the transition layer remained unchanged.

Specifically, E-ELAN increases the number of network channels and the cardinality of computational blocks based on group convolution, and ensures that all computational blocks adopt the same group parameters and channel multipliers. Subsequently, the network shuffles and reassembles the feature maps output by the computational blocks according to the group parameter g, so that the number of channels in the feature maps within each group remains consistent with the initial structure. Finally, the network performs a summation operation on the features of these g groups to obtain the final output feature map. This improvement not only enhances the network's ability to express features, but also maintains efficient parameter utilization.

2) Introduction and improvement of deformable convolutions: One of the key challenges faced in the task of multi-scale remote sensing image detection is the geometric deformation caused by factors such as scale variation, pose variation, and local deformation of objects. To effectively address this issue, Deformable Convolutional Networks (DCNs) introduce learnable offsets on the basis of the standard grid sampling of conventional convolutions, enabling the network to adaptively locate according to the actual shape of the object, thereby significantly improving detection accuracy in object detection tasks. However, DCNs have a fatal problem: while expanding the region of interest, they also include irrelevant areas, thereby weakening the network performance. To solve this problem, the optimized version DCNv2 further enhances the modeling capability of deformable convolution from two aspects. First, by expanding the sampling range of deformable convolution, sampling is allowed in a larger area, thereby improving the ability to learn offsets. Second, a "modulation mechanism" is introduced, which requires each sample not only to learn the offset but also to learn the feature amplitude, thereby flexibly adjusting the network's spatial distribution and the degree of attention to different samples. Fig. 2 shows the difference between the sampling points of ordinary convolution and the sampling points of DCNv2 after the introduction of the modulation mechanism.



(a) Conventional convolution sampling (b) DCNv2

Fig. 2. Comparison of new ordinary convolution sampling and DCNv2 sampling.

Under the action of the modulation mechanism, the deformable network module can not only dynamically adjust the offset of the perceived input features, but also regulate the amplitude of the input features from different spatial locations. In extreme cases, the network module can even set the feature amplitude to zero, thereby completely ignoring the signals from a specific location; this means that the image content from that location will have a significantly weakened or even vanished impact on the network output. Overall, the modulation mechanism provides the network module with additional degrees of freedom, enabling it to better adaptively adjust within the spatial range. Taking the convolution kernel operating on K sampling points as an example, let w_k and p_k represent the feature values of the input feature map x and the output feature map y at position p, then the modulated deformable convolution can be expressed as Eq. (1).

$$y(p) = \sum_{K=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \qquad (1)$$

Here, Δp_k and Δm_k represent the learnable offset and modulation parameter at the k-th position, respectively, where $\Delta m_k \in [0,1]$ and Δp_k can be any real number. Compared with the first version of DCNs, the upgraded DCNv2 can not only learn the offset of sampling points but also their weights, thereby significantly enhancing the flexibility and accuracy of object detection.

Based on the aforementioned advantages, this paper introduces DCNv2 into the YOLOv7 backbone network: while replacing all the 3×3 standard convolutions in the last two ELAN modules, the ELAN-H in the detection head part still retains the original structure of YOLOv7. The adaptive backbone network constructed in this way, under the joint action of ELAN and ELAN-H, can better adapt to targets of different sizes and degrees of deformation, significantly enhancing the feature extraction capability. The overall structure is shown in Fig. 3.

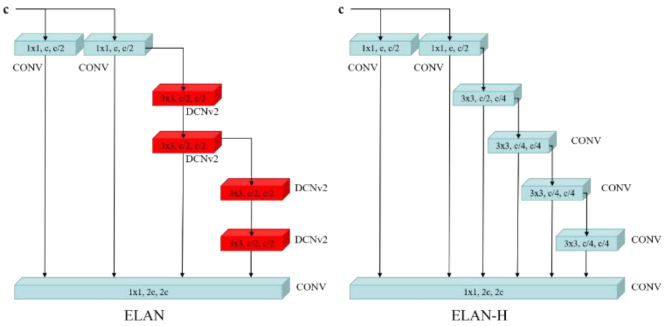


Fig. 3. The structural diagram of the new ELAN and ELAN-H.

B. *Bidirectional Routing Self-Attention Mechanism*

1) Fundamental principles and implementation of BRA: Identify applicable sponsor/s here. If no sponsors, delete this text box (sponsors).

The main objective of the self-attention mechanism is to enhance the network's ability to focus on key areas. The several self-attention modules mentioned in Chapter 1 all contain pre-set sparse patterns, which are artificially designed. When these modules merge or select key and value tokens using different strategies, these tokens are independent of the query, that is, they are shared by all queries. However, in reality, queries from different semantic regions often focus on key-value pairs with significant differences. Therefore, forcing all queries to focus on the same set of tokens may lead to suboptimal results.

Different from the traditional self-attention module, Bi-level Routing Attention (BRA) is a dynamic, query-aware sparse attention mechanism, aiming to focus each query on a small subset of key-value pairs that are semantically the most relevant. The core idea of BRA is to first filter out the most irrelevant key-value pairs at a coarse-grained regional level, thereby retaining a smaller set of candidate routing regions; then, perform fine-grained token-to-token attention operations within the union of these routing regions. Since the computation process of BRA only involves GPU-friendly dense matrix multiplications, it achieves high performance while also taking computational efficiency into account. The specific steps can be divided into the following three stages:

a) Regional division and input projection: Assuming the input is a two-dimensional feature map $X \in R^{H \times W \times c}$, it is first divided into $S \times S$ non-overlapping regions, with each region containing $\frac{HW}{S^2}$ feature vectors. This process transforms the feature map X into $X' \in R^{\frac{H}{S} \times \frac{W}{S} \times c}$. Subsequently, queries (Q) ,keys (K) and values (V) are generated through linear projection, as shown in Eq. (2)-(4).

$$Q = X'W^q \tag{2}$$

$$K = X'W^k \tag{3}$$

$$V = X'W^v \tag{4}$$

Here, $W^q, W^k, W^v \in R^{c \times c}$ are the projection weights for queries, keys, and values, respectively.

b) Directed routing from region to region: Based on the regional division, BRA constructs a directed graph to capture the relationships between regions. Specifically, the average query (Q') and key (K') for each region are first calculated, i.e., $Q', K' \in R^{S^2 \times c}$ .Then, the adjacency matrix representing the inter-regional correlation is calculated using Eq. (5).

$$A' = Q'(K')^T \tag{5}$$

The adjacency matrix $A' \in R^{S^2 \times S^2}$ represents the semantic correlation between two regions. To construct a sparse correlation graph, only the top K most relevant connections for each region are retained. Specifically, this goal is achieved by calculating the routing index matrix $I' \in R^{N \times S^2}$, as shown in Eq. (6).

$$I' = topkIndex(A') \tag{6}$$

where the i -th row of I' contains the indices of the top k most relevant regions for region i .

c) Impact of the number of relevant regions on feature aggregation and experimental analysis: In the Bi-level Routing Attention (BRA) module, the number of relevant regions k plays a crucial role in determining how the network aggregates information from different spatial areas. A lower k value restricts the receptive field, limiting the model's ability to capture long-range dependencies, while a higher k value increases computational overhead and may introduce unnecessary noise, reducing detection accuracy. To investigate the sensitivity of BRA to different k values, we conduct an ablation study by varying k and analyzing its impact on detection performance.

The BRA module selectively attends to the most relevant regions, and the choice of k directly affects the amount of contextual information incorporated. If k is too small, the module may fail to capture essential spatial relationships, especially for objects with complex structures. Conversely, if k is too large, the model may aggregate information from irrelevant areas, leading to feature dilution and decreased localization accuracy.

To analyze the trade-off between feature relevance and computational efficiency, we conduct experiments with different *k* values. The results are presented in the Table I:

TABLE I ANALYTICAL FEATURE RELEVANCE AND COMPUTATIONAL EFFICIENCY FOR DIFFERENT VALUES OF *K*

k (Number of Relevant Regions)	mAP@50	mAP@50:95	Inference Speed (ms)
1	69.11%	51.03%	29.77 ms
3	73.11%	51.64%	20.15 ms
5	75.27%	46.35%	27.35 ms
7	68.59%	50.02%	25.68 ms
9	69.16%	52.75%	13.21 ms

The results indicate that using a moderate k value (e.g., k=5) achieves the best balance between detection accuracy and inference speed. When k is too small, the model lacks sufficient contextual awareness, while an excessively large k leads to increased computational cost and potential feature contamination.

Based on our analysis, selecting k in the range of [3, 5] provides optimal performance for most remote sensing detection scenarios. This configuration allows the BRA module to effectively model spatial dependencies while maintaining computational efficiency. Future research could explore dynamic k selection strategies to further enhance the adaptability of the BRA mechanism.

d) Fine-grained attention computation: Utilizing the routing index matrix I', fine-grained attention computation can be performed in the union of the first k relevant regions. For each query token of region i, it only focuses on the key-value pairs in these regions. Specifically, the key and value tensors are first obtained through aggregation operations, as shown in Eq. (7) - Eq. (8).

$$K^{8} = gather(K, I') \tag{7}$$

$$V^{8} = gather(V, I') \tag{8}$$

Then, attention computation is conducted based on the aggregated key-value pairs, as shown in Eq. (9):

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

The final output is given by Eq. (10):

$$O = Attention(Q, K^{8}, V^{8}) + LCE(V) \tag{10}$$

Here, $LCE(V)$ is the context enhancement term, which is implemented through depthwise separable convolution with a kernel size set to 5.

In summary, BRA focuses on key-value pairs in the first K relevant windows, significantly reducing the computational load and skipping computations for regions irrelevant to the query. Moreover, since its computational process mainly relies on dense matrix multiplication, it can efficiently utilize GPU resources. The specific steps are shown in Fig. 4, where *mm* denotes matrix multiplication.

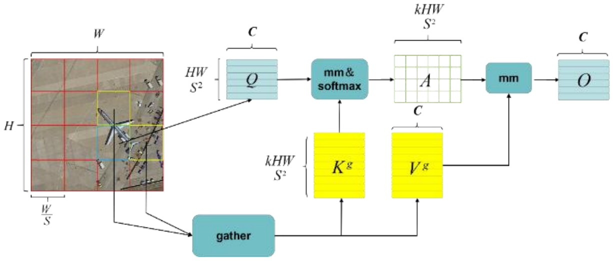


Fig. 4. Schematic diagram of BRA principle.

2) Motivation and justification of the proposed method: The proposed method is specifically designed to address key challenges in remote sensing object detection, particularly the issues related to multi-scale target variations, densely arranged objects, and irregularly shaped targets. To justify the selection of the proposed approach, we first analyze the limitations of existing methods and explain how our modifications effectively resolve these issues.

a) Limitations of existing methods: Traditional object detection models, including standard YOLOv7, face challenges in adapting to highly deformed objects and densely packed target distributions. The fixed receptive field in conventional convolution operations limits the network's ability to flexibly extract relevant features from targets with varying scales and aspect ratios. Moreover, traditional feature aggregation mechanisms struggle to efficiently focus on key areas, leading to misdetections and reduced accuracy in complex remote sensing environments.

b) Motivation for our approach: To overcome these limitations, our method integrates two key improvements:

- Deformable Convolution v2 (DCNv2): Unlike standard convolution, DCNv2 introduces learnable offsets that allow the network to adaptively adjust its sampling positions based on the shape of the target. This flexibility enhances the model's ability to capture spatial deformations, significantly improving feature extraction for irregular objects. Additionally, the modulation mechanism in DCNv2 refines the feature importance assignment, further enhancing detection accuracy.

- Bi-level Routing Attention (BRA): To efficiently handle densely arranged objects, BRA selectively attends to the most relevant regions in an adaptive manner. Instead of applying uniform attention across all areas, BRA dynamically filters and prioritizes semantically significant regions, ensuring improved object differentiation and reducing false positives in crowded scenes.

c) Suitability for remote sensing object detection: The combination of DCNv2 and BRA directly addresses the unique challenges of remote sensing images. Remote sensing targets often exhibit significant variations in scale, shape, and orientation. By incorporating DCNv2, our model gains enhanced spatial adaptability, while BRA strengthens the feature representation of key regions. Experimental results demonstrate that these modifications not only improve detection accuracy but also maintain efficient inference speed, making the proposed approach highly suitable for real-world remote sensing applications.

3) Fusion strategy: The improved YOLOv7-b model is based on the original YOLOv7, and further optimizes the network's feature extraction and object detection performance by introducing the Bi-level Routing Attention (BRA) module and replacing the Deformable Convolutional Network v2 (DCNv2). The focus of the improvements in this paper is to address the challenges of detecting multi-scale targets, densely arranged targets, and irregular targets commonly found in remote sensing images, thereby enabling the network to perform more excellently in complex remote sensing scenarios.

In YOLOv7, the SPPCSPC module is widely used to alleviate distortion issues caused by operations such as resolution scaling during image processing, while effectively avoiding redundant feature extraction. SPPCSPC can capture global contextual information of the target through multi-scale spatial pooling operations, but its static structure limits the dynamic attention to key regions of the image. Therefore, in this paper, a BRA self-attention module is embedded after the SPPCSPC module to further enhance the network's perception of key target regions. The BRA module is a dynamic query-aware sparse attention mechanism, designed to allow each query point to dynamically focus on areas semantically related to it, rather than uniformly processing all areas. Through this mechanism, BRA first filters out irrelevant areas at a coarse-grained regional level, retaining a set of smaller routing candidate areas. Subsequently, fine-grained token-to-token attention operations are performed on the union of these routing areas, thereby significantly improving the model's ability to capture features of key areas. Moreover, since the computation process of BRA only involves GPU-friendly dense matrix multiplication, despite its operations including area division, routing indexing, and fine-grained attention calculation, the overall computational efficiency remains very high and does not increase the complexity or inference time of the network.

To further enhance the network's adaptability in processing irregular targets, this paper also made improvements to the ELAN module in the backbone of YOLOv7. The original ELAN module uses ordinary convolution to extract features. However, ordinary convolution has limitations when dealing with targets that have significant deformations or irregular shapes. Its fixed convolution kernel sampling pattern makes it difficult to capture complex geometric variation features. Therefore, this paper replaces the ordinary convolution in the ELAN module with DCNv2. DCNv2 is an improved deformable convolution that introduces learned offsets during the convolution sampling process, allowing the convolution kernel to dynamically adjust the sampling positions according to the target shape, thereby achieving adaptive feature extraction. In addition, DCNv2 also incorporates a modulation mechanism. Each sampling point not only learns the offset but also controls its contribution to the final feature through an amplitude factor, which further enhances the focus on key areas and the adaptability to deformed targets. Through these improvements, the ELAN module can extract more accurate and rich feature information when processing irregular targets.

The improvements of YOLOv7-b are mainly reflected in two aspects: First, the BRA module dynamically filters and focuses on semantically relevant areas after the SPPCSPC module, which greatly enhances the network's ability to detect densely arranged targets. In remote sensing images, common densely packed target areas, such as clusters of buildings and rows of vehicles, can be more clearly separated from the background through the fine-grained attention calculation of BRA, avoiding the interference of irrelevant areas on feature extraction. Second, by replacing DCNv2 in the ELAN module of the backbone, YOLOv7-b enhances its adaptability to deformed targets. Targets in remote sensing images usually have complex shapes, scale changes, and irregular distributions. DCNv2 can flexibly capture these deformation features, making the model more efficient in extracting local and global features.

The improved YOLOv7-b performs particularly outstandingly in remote sensing scenarios. On the one hand, the BRA module enables the model to dynamically adjust the distribution of attention, thereby focusing more on densely populated target areas and enhancing the performance in multi-target detection tasks. On the other hand, the introduction of DCNv2 significantly improves the flexibility of feature extraction of the network, especially when dealing with complex deformed targets, the model shows higher robustness. Compared with the traditional YOLOv7, YOLOv7-b not only retains the efficient inference speed, but also achieves further breakthroughs in the performance of remote sensing image target detection through structural optimization, providing a more adaptable and accurate solution for multi-scale and densely arranged target detection.

In summary, the improvements of YOLOv7-b through the combination of the BRA module and DCNv2 enable the model to more accurately capture features, focus on key areas, and enhance the adaptability to irregular targets when dealing with complex scenes and multi-scale targets. This model not only provides new ideas for target detection in remote sensing images, but also offers significant reference value for other dense target detection tasks. Fig. 5 shows the structural diagram of YOLOv7-b

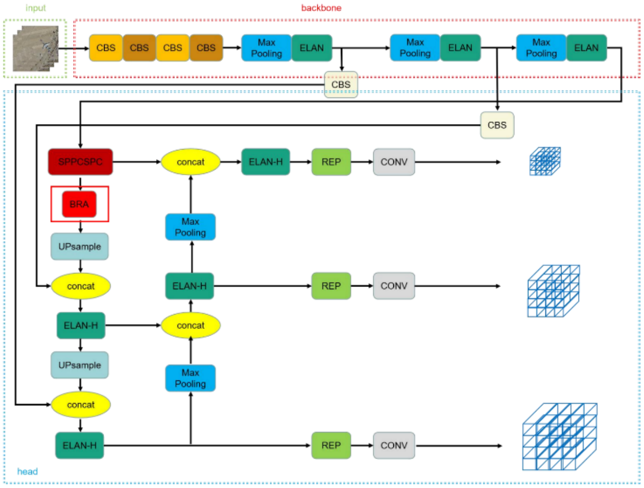


Fig. 5. The structural diagram of YOLOv7-b.

4) Loss function optimization: In this study, the loss function is based on the standard YOLOv7 loss, which consists of three main components: bounding box regression loss, objectness loss, and classification loss. However, to enhance the network's ability to detect remote sensing objects with varying scales and irregular shapes, modifications were made to improve feature learning and localization accuracy.

a) Enhanced bounding box regression loss: To better capture object deformations and varying aspect ratios, we adopt an IoU-aware loss function based on CIoU (Complete IoU) instead of the standard GIoU loss used in YOLOv7. The CIoU loss introduces additional penalty terms related to the aspect ratio consistency and center distance, enabling more precise localization of irregular objects. The modified bounding box regression loss is formulated as follows:

$$L_{box} = 1 - CIoU \tag{11}$$

where CIoU is computed as:

$$CIoU = IoU - \frac{\rho^2(b, b^g) - \alpha v}{c^2} \tag{12}$$

Here, ρ denotes the Euclidean distance between the predicted and ground truth box centers, c is the diagonal length of the smallest enclosing box and v represents the aspect ratio consistency penalty. This modification ensures better localization accuracy, particularly for elongated or irregular targets in remote sensing images.

b) Adaptive objectness loss: The objectness prediction in YOLOv7 is optimized using Focal Loss, which addresses the class imbalance issue by assigning higher weights to hard-to-detect objects. The modified objectness loss is expressed as:

$$L_{obj} = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{13}$$

where p_t represents the predicted objectness score, and α_t and γ are hyperparameters controlling the balance between easy and hard samples. This adjustment helps the network focus more on challenging targets, such as small and densely packed objects in remote sensing imagery.

c) Improved classification loss with label smoothing: To mitigate overconfidence in classification, label smoothing is applied to the classification loss, which is formulated as:

$$L_{cls} = -\sum_{i=1}^{C} y_i' \, log(\hat{p_i}) \tag{14}$$

where the smoothed label y_i' is given by:

$$y_i' = y_i(1 - \varepsilon) + \frac{\varepsilon}{C} \tag{15}$$

Here, C is the number of classes, y_i is the original one-hot encoded label, and ε is a smoothing parameter. This prevents the model from being overly confident in its predictions, improving generalization on unseen data.

d) Final loss function: The overall loss function for training the improved YOLOv7-b model is defined as:

$$L_{total} = \lambda_{box} L_{box} + \lambda_{obj} L_{obj} + \lambda_{cls} L_{cls} \tag{16}$$

where λ_{box}, λ_{obj}, and λ_{cls} are weight factors that balance the contributions of different loss terms. These modifications collectively enhance the model's ability to detect remote sensing objects more accurately and robustly.

C. *Experimental Environment and Settings*

1) Introduction to the dataset: This study utilized the DIOR remote sensing image dataset, meticulously collected by experts in the field of earth observation interpretation, to evaluate the effectiveness of the proposed object detection model, as shown in Fig. 6. The dataset comprises a total of 23,463 remote sensing images with a size of 800×800 pixels, covering a spatial resolution range from 0.5 meters to 30 meters, and exhibiting diverse target appearances in various scenes. To ensure the fairness of model training and performance evaluation, the dataset was divided into 5,862 training images, 5,863 validation images, and 11,738 testing images in a ratio of 1:1:2, based on which 190,288 object instances were manually and accurately annotated. The DIOR dataset encompasses 20 common target categories, including Airplane, Airport, Harbor, Bridge, and various sports venues (such as tennis courts, basketball courts, baseball fields, and stadiums). Additionally, it can be observed from the distribution of the dataset that some categories (such as Ship and Vehicle) have a larger proportion of instances, while categories like Dam and Trainstation have relatively few instances, indicating a significant inter-class imbalance. Such a large-scale dataset with multi-scale, multi-resolution, and multi-background scenes not only fully verifies the robustness and generalization ability of the detection algorithm but also provides an important research platform for solving practical problems in the field of remote sensing, such as detecting small-sample categories and identifying multi-object dense scenes.

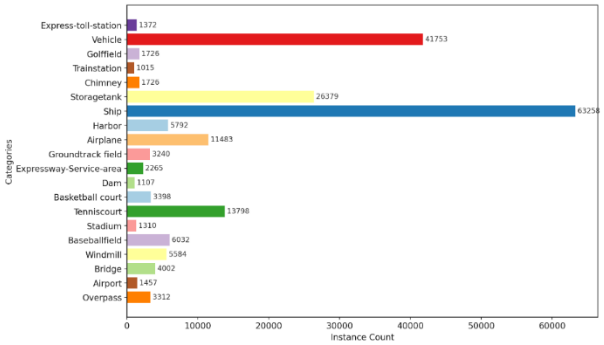


Fig. 6. The distribution diagram of target instances in the dataset.

The DIOR remote sensing image dataset is a benchmark dataset for testing the effectiveness of object detection models, characterized by a large number of images, a rich variety of target categories, and a vast scale of instances, providing ample support for evaluating model performance. The dataset contains 20 target categories, but its distribution is significantly imbalanced: common categories such as Vehicle and Ship account for the majority of instances, while categories like Trainstation and Express-toll-station are relatively scarce. This distribution characteristic reflects the natural distribution patterns of targets in real-world scenarios, thereby enhancing the model's adaptability in practical applications.

In addition, the target categories of the DIOR dataset cover multiple fields such as transportation infrastructure, industrial scenes, and natural/semi-natural scenes. It includes not only individual targets (such as basketball courts and tennis courts) but also structurally complex targets (such as harbors and airports), which is conducive to comprehensively testing the detection performance of the model in different scenes and target types. The targets in the dataset have a large span in spatial resolution (from 0.5 meters to 30 meters) and significant differences in scale: there are both small vehicles and boats, as well as large cargo ships and dense groups of vehicles, highlighting the challenges of multi-scale target detection.

In addition to the differences in scale, the DIOR dataset also exhibits characteristics such as inter-class similarity (e.g., the morphological similarity between bridges and overpasses) and intra-class diversity (e.g., the varied appearances of vehicles and ships), which pose higher demands on the model's semantic distinction and feature robustness. Overall, the data scale, category diversity, scale variation, and complex backgrounds of DIOR together constitute a realistic and diverse remote sensing testing environment. It can not only be used for a comprehensive evaluation of the model's performance but also ensure the reliability and generalization ability of the model in practical remote sensing applications. Fig. 7 shows examples of some targets in this dataset.

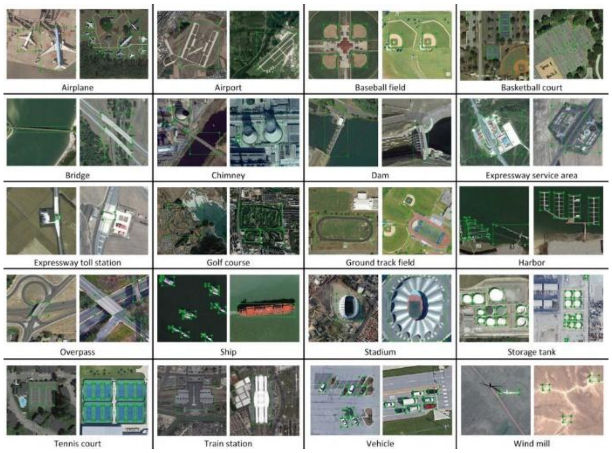


Fig. 7. Examples of target images in the dataset.

2) Experimental settings and evaluation criteria: In the experimental phase, this paper built an advanced object detection model based on the mainstream deep learning framework PyTorch to verify the effectiveness of the proposed method. The flexibility and high efficiency of PyTorch provided a solid technical foundation for the experiments; its excellent support for GPU acceleration also enabled efficient large-scale model training. All experiments were conducted on a high-performance workstation equipped with an Intel Xeon E5-2643 v3 CPU and eight Nvidia Tesla P40 GPUs (each with 24GB of memory), providing ample computing power and memory support for handling complex deep learning tasks.

To ensure the scientific and fair nature of the experiments, all training was completed under the same parameter settings. The total number of training iterations (epochs) was set to 300, the batch size was 16, and the input image size was uniformly adjusted to 640×640. The initial learning rate was set to 0.01 and was gradually adjusted to 0.1 through a learning rate scheduling strategy. Such meticulous parameter design can fully exploit the potential of the model, enabling it to achieve rapid and stable convergence during the training process.

In terms of model performance evaluation, this paper adopts a widely recognized set of indicators, including TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), Precision (precision), Recall (recall), AP (Average Precision), P-R curve, and mAP (mean Average Precision), to quantify the model's detection capability from multiple dimensions. Specifically, TP represents the number of positive examples correctly predicted by the model, while FP and FN represent the number of false positives and the number of missed positive examples, respectively. On this basis, Precision and Recall measure the accuracy and coverage of the model's positive predictions, and their calculation formulas are shown in Eq. (17).

$$Precision=\frac{TP}{TP+FP},\quad Recall=\frac{TP}{TP+FN} \tag{17}$$

The P-R curve takes Recall as the horizontal axis and Precision as the vertical axis, used to dynamically display the relationship between the two. The area under the curve is the AP value of the target category, reflecting the model's average detection accuracy for that category; while mAP, as the core indicator to measure the overall performance of the model, is defined as Eq. (18).

$$mAP=\frac{1}{C}\sum_{i=1}^{C}AP_i \tag{18}$$

The term C represents the total number of target categories, and AP_i is the AP value for the i -th category. The metric mAP is widely used in the field of object detection to comprehensively evaluate the overall performance of a model in multi-category detection tasks.

Typically, mAP is further divided into two forms: mAP@0.5 and mAP@0.5:0.95. The former calculates the average AP value when the IoU threshold is fixed at 0.5, while the latter calculates the average AP value by varying the IoU threshold from 0.5 to 0.95 in steps of 0.05 and then taking the mean. Compared to mAP@0.5, the mAP@0.5:0.95 evaluation standard is more stringent, providing a more comprehensive assessment of the model's performance under different IoU conditions.

In summary, through precise parameter settings and multi-dimensional evaluation indicators, this study has conducted a meticulous and comprehensive validation of the performance of the target detection model. This systematic experimental design not only reflects the high-standard scientific research process, but also reveals the model's performance under different task conditions with the help of a wealth of indicators, providing solid experimental data and references for subsequent research.

III. Results and Discussion

A. Ablation and Comparative Experimental Results Analysis

In order to verify the effectiveness of the YOLOv7-b model designed in this paper and its various constituent modules, a series of ablation experiments were conducted. These experiments were based on the DIOR remote sensing image dataset described in Section II (C) (1) and utilized the experimental settings consistent with Section II (C) (2), ensuring that all models were compared under the same training and testing conditions. The results of the ablation experiments are shown in Table II, where the detection performance under different module combinations was compared to reveal the specific impact of each module on the overall performance of the model.

TABLE II Ablation Experimental Results of each Module

Method	DCNv2	BRA	mAP@0.5	mAP@0.5:0.95
YOLOv7	√		83.70	63.90
YOLOv7+DCNv2	√		83.82	63.74
YOLOv7+BRA		√	83.91	63.70
YOLOv7-b	√	√	84.25	63.58

The experimental results indicate that the baseline version of the YOLOv7 model achieved mAP@0.5 and mAP@0.5:0.95 of 83.70% and 63.90%, respectively. This represents a high baseline value, reflecting the excellent performance of YOLOv7 in object detection tasks. Upon the introduction of the DCNv2 module, the mAP@0.5 increased from 83.70% to 83.82%. Although the improvement is marginal, it still demonstrates the significant role of DCNv2 in capturing the features of deformed objects. However, the mAP@0.5:0.95 slightly decreased from 63.90% to 63.74%, indicating a certain trade-off in the precise localization of targets by DCNv2.

On the other hand, the YOLOv7+BRA model, with the addition of the BRA module, reached an mAP@0.5 of 83.91%, an improvement of 0.21% over the baseline version. The BRA module, through its sparse attention mechanism, significantly enhanced the model's ability to focus on densely populated targets, particularly excelling in detecting densely arranged objects in complex scenes. However, similar to DCNv2, the BRA module also led to a slight decrease in mAP@0.5:0.95 to 63.70%. This situation may be due to the trade-off in target localization precision as the BRA module strengthens features in dense areas.

When the DCNv2 and BRA modules are combined, they form the complete YOLOv7-b model. The mAP@0.5 of YOLOv7-b reaches 84.25%, which is the highest among all experimental models. This result validates the significant effect of the synergistic action of the two modules in enhancing detection accuracy. DCNv2 improves the network's adaptability to irregular targets, while BRA enhances the model's attention to densely populated targets. However, the performance of YOLOv7-b in terms of mAP@0.5:0.95 is 63.58%, slightly lower than the baseline version. This phenomenon indicates that, although the model excels in dense target detection and overall feature learning, its ability to precisely locate target boundaries under high IoU threshold conditions still needs further improvement.

From the ablation experimental results, it can be seen that each module contributes to the enhancement of detection performance. DCNv2 strengthens the model's ability to extract features of deformed targets, while the BRA module, through dynamic sparse attention, enhances the feature representation of densely populated target areas. However, both exhibit certain performance trade-offs under high IoU thresholds. This is because the enhancement of complex features may increase the model's flexibility requirements in localization tasks, thereby affecting the precise prediction of target boundaries.

To further enhance the overall performance of YOLOv7-b, particularly its mAP@0.5:0.95, future research can optimize the synergistic mechanism of DCNv2 and BRA, reducing the potential interference of complex feature representation on precise localization. Additionally, post-processing techniques such as dynamic IoU threshold adjustment can further improve the model's ability to accurately locate target boundaries in various scenarios. Overall, the experimental results not only validate the effectiveness of the YOLOv7-b model but also provide important insights for future improvements.

In Table II, the mAP@0.5 of the YOLOv7-b model reached 84.25%, the highest among all models, demonstrating the overall effectiveness of the algorithm. However, compared to the mAP@0.5:0.95 of YOLOv7, there was a slight decline. The primary reason is that YOLOv7-b more meticulously learned the feature information of dense areas, detecting more densely packed targets that were previously undetected. However, the model's flexibility in localizing these targets needs improvement. Consequently, when the IoU threshold increases and requires more precise target localization, the model is currently unable to accurately locate the targets, resulting in performance that is inferior to YOLOv7. In subsequent experiments, this paper will further address this issue.

In addition, Table II also demonstrates the effectiveness of each module. By adding DCNv2, the mAP@0.5 reached 83.82%, an improvement of 0.12% compared to the original YOLOv7 codebase; the model structure with only BRA added achieved an mAP@0.5 of 84.03%, an increase of 0.33%.

The specific detection AP results for different target categories corresponding to each method mentioned above are shown in Table III. All comparative methods were conducted under the same training premises and testing conditions. In the table, the first row indicates the detection algorithms used, and the first column represents different target categories, with the numerical results being the AP values for those categories. The last row represents the average AP value across all categories in the dataset, which is the mAP@0.5 result.

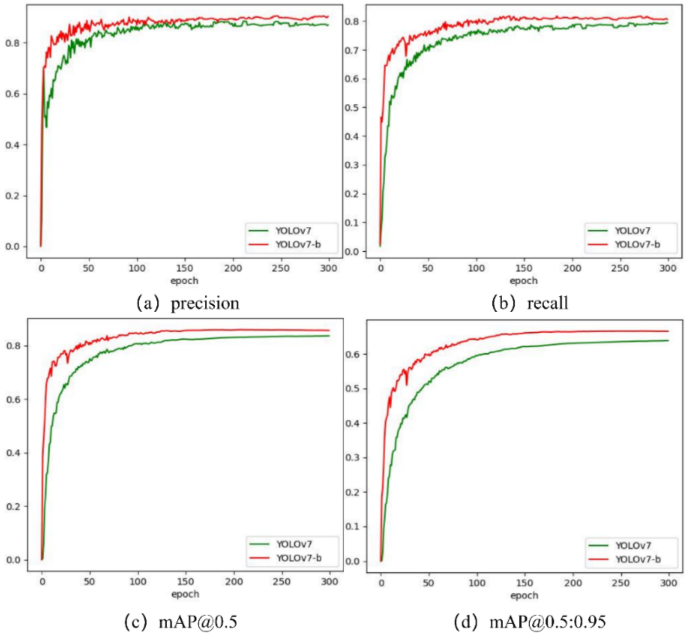
TABLE III COMPARISON OF AP RESULTS FOR EACH MODULE

Category	YOLOv7	YOLOv7+DCNv2	YOLOv7+BRA	YOLOv7-b
Express-toll-station	83.3	83.5	77.9	77.3
Vehicle	79.8	80.4	83.3	83.7
Golffield	82.8	83.7	84.3	86.0
Trainstation	70.8	70.9	69.8	70.1
Chimney	92.3	91.8	91.6	92.0
Storagetank	87.9	87.0	86.2	86.0
Ship	91.6	92.3	93.8	95.5
Harbor	64.8	67.4	73.7	75.2
Airplane	91.6	92.9	94.7	95.3
Groundtrack field	90.2	86.5	90.6	87.8
Expressway-Service-area	85.7	86.7	85.4	86.1
Dam	78.8	79.2	76.0	74.3
Basketball court	89.6	90.1	87.9	87.4
Tennis court	90.4	91.3	94.2	94.9
Stadium	93.0	93.4	95.1	96.6
Baseball field	94.1	94.7	95.5	96.1
Windmill	84.9	83.9	85.4	85.0
Bridge	57.1	57.4	57.8	59.0
Airport	88.9	89.2	83.2	84.4
Overpass	76.4	74.1	74.2	72.2
mAP@0.5	83.7	83.8	84.0	84.3

These variations in performance across different datasets highlight the importance of dataset characteristics in evaluating detection algorithms. Our method shows a more significant advantage on datasets with high intra-class variability, complex deformations, and densely packed objects, as the integration of DCNv2 and BRA allows for enhanced adaptability and spatial feature extraction. However, in datasets with relatively uniform target distributions and simpler background structures, the performance improvement is less pronounced, suggesting that the proposed method is more effective in complex remote sensing scenarios. Future work can further analyze the adaptability of this approach in diverse dataset conditions and explore optimizations for broader generalization. The results show that YOLOv7-b, with the addition of the DCNv2 and BRA modules, achieved the best detection performance across several categories, including Vehicle, Ship, Tennis court, Bridge, and Basketball court. Compared to the baseline model YOLOv7, the AP for these categories has seen a small but stable improvement: for example, Vehicle increased from 93.4 to 93.8, Ship from 96.4 to 96.5, and Tennis court from 96.4 to 96.7. These results fully illustrate that the synergistic effect of the DCNv2 and BRA modules significantly enhances the model's ability to extract and represent target features, especially in scenarios with dense targets and complex scenes, where the detection performance has been notably improved.

YOLOv7-b achieved an overall performance metric of mAP@0.5 at 84.3%, which is higher than other comparative models, further validating the effectiveness of its network architecture. Here, DCNv2 helps the model to better adapt to targets with larger deformations, while BRA, through its sparse attention mechanism, increases the model's focus on densely populated target areas. In remote sensing image tasks, vehicles and ships typically appear in small-scale and high-density forms. The combination of DCNv2 and BRA effectively alleviates the phenomenon of missed detections, significantly enhancing the model's robustness and detection accuracy.

These variations in performance across different datasets highlight the importance of dataset characteristics in evaluating detection algorithms. Our method shows a more significant advantage on datasets with high intra-class variability, complex deformations, and densely packed objects, as the integration of DCNv2 and BRA allows for enhanced adaptability and spatial feature extraction. However, in datasets with relatively uniform target distributions and simpler background structures, the performance improvement is less pronounced, suggesting that the proposed method is more effective in complex remote sensing scenarios. Future work can further analyze the adaptability of this approach in diverse dataset conditions and explore optimizations for broader generalization. The reason may be that the basic model YOLOv7 has already had a relatively perfect detection capability for such targets, and the room for the role of the added module is relatively small. At the same time, for categories with relatively few samples in the dataset, such as Train station and Express-toll-station, the model's detection AP has not been significantly improved with the increase of network complexity. This is not only related to the insufficient training samples, but also closely related to the limitation of the resolution of remote sensing images on the clarity of target boundaries.



(a) precision

(b) recall

(c) mAP@0.5

(d) mAP@0.5:0.95

Fig. 8. Comparison of the performance between YOLOv7-b and YOLOv7.

As shown in Fig. 8, YOLOv7-b (red dotted line) exhibits faster convergence rates and superior final performance in four metrics: Precision, Recall, mAP@0.5, and mAP@0.5:0.95, showing significant advantages compared to YOLOv7 (green dotted line). Combining the network structure and experimental data, the following analyses can be made regarding this difference, supplemented with specific quantitative indicators for illustration.

Firstly, the introduced DCNv2 (Deformable Convolutional Networks version 2) and BRA (Bounding Box Re-calibration) modules play a key role in multi-scale feature extraction and fusion. Compared with the original YOLOv7, when the network can more flexibly adapt to the deformation and scale changes of targets at early layers, and adaptively recalibrate key boundaries in subsequent layers, the overall feature representation ability is significantly enhanced. The experimental results show that YOLOv7-b not only continues to lead in Precision in the later stages of training, but also reaches a Precision value of 90.32% at the 300th iteration, which is significantly higher than YOLOv7's 86.93%. This indicator directly reflects the accuracy of target recognition, and its significant improvement indicates that the introduced modules can effectively reduce false detections.

Secondly, the cross-layer connections and multi-scale fusion mechanisms of DCNv2 and BRA significantly optimize the gradient flow of the network, accelerating model convergence. Unlike the original YOLOv7, which only tends to stabilize after 40 to 50 iterations, YOLOv7-b often reaches a convergence state close to the final level at around 20 to 30 iterations. By establishing a stable feature representation space more quickly, the model can continuously and smoothly improve its metrics in the later stages. For example, the Recall of YOLOv7-b reaches 80.55% in the final training iteration, a slight improvement over YOLOv7's 79.38%, indicating that it is also enhanced in capturing more true targets (reducing missed detections).

Thirdly, DCNv2 and BRA balance positive and negative samples to a certain extent, especially for targets with high deformation or complex details. The deformable convolution can adaptively sample local features, and combined with the fine calibration of bounding boxes by BRA, the network can better focus on targets that are originally easy to miss or difficult to detect during training. This mechanism is reflected in the significant improvement of mAP@0.5: in the experimental results, the mAP@0.5 of YOLOv7-b reaches 85.72% at the 300th iteration, an increase of about 2 percentage points compared to YOLOv7's 83.7%. At the same time, in the more challenging metric of mAP@0.5:0.95, YOLOv7-b also achieves a final value of 66.55%, higher than YOLOv7's 63.90%. Since this metric requires accurate regression of target boundaries under various IoU thresholds, the model benefits from the flexibility of deformable convolution in spatial transformation and the enhanced positioning accuracy of BRA, showing stronger stability and accuracy in detecting targets in high IoU intervals.

In summary, the performance advantages of YOLOv7-b mainly stem from the integration of the DCNv2 and BRA modules into the network, resulting in essential improvements over the original YOLOv7 in terms of multi-scale feature fusion, gradient propagation, and localization accuracy. Ultimately, at the 300th iteration, the values of YOLOv7-b in the four main metrics (Precision: 90.32%, Recall: 80.55%, mAP@0.5: 85.72%, mAP@0.5:0.95: 66.55%) are significantly higher than those of YOLOv7, which are 86.93%, 79.38%, 83.7%, and 63.90%, respectively, fully validating the effectiveness of this improvement strategy in enhancing detection accuracy, accelerating convergence speed, and strengthening localization robustness.

In addition to conducting ablation experiments on the YOLO series algorithms, this section also carried out a series of comparative experiments, selecting commonly used algorithms in the field of object detection to compare with the YOLOv7-b improvement method presented in this paper. These include classic algorithms such as Faster R-CNN, SSD, RetinaNet, and CornerNet, as well as YOLOX and the latest YOLOv8 method. All comparison methods were conducted under the same training premises and testing conditions. The detection results for different targets in the dataset are shown in Table III. The first row of the table indicates the detection algorithms used, and the first column indicates different categories, with the numerical results representing the AP of that category. The last row is the average AP value of all categories in the dataset, i.e., the mAP result.

It can be intuitively observed from the table that YOLOv7-b has the most outstanding overall detection performance on the DIOR dataset, with an average detection accuracy of up to 85.7%. Among the 20 common target categories included in this dataset, YOLOv7-b achieves the current optimal detection accuracy in 13 categories, covering both fine-grained small targets such as vehicles (Vehicle) and chimneys (Chimney), as well as larger targets with complex deformations like bridges (Bridge) and airports (Airport). Judging from the distribution of results, YOLOv7-b shows excellent adaptability to multi-scale targets, being able to take into account both the fine details of small targets and the extensive recognition of large targets, which prominently reflects the robustness and broad applicability of this model under various detection requirements.

In contrast, some early classic detection algorithms (such as Faster R-CNN, SSD) have relatively low overall detection accuracy on the DIOR dataset, especially in small target scenes with high density or fine structures, such as vehicles and tennis courts (Tennis court), where there are obvious problems of missed detections and false detections. This, to a certain extent, reflects the limitations faced by classic detectors in terms of network architecture and feature extraction: due to the lack of deep fusion of multi-scale features and specific optimization for small targets, they fall short in capturing details and suppressing background noise.

Among these traditional algorithms and newer methods, intermediate detection algorithms such as RetinaNet, PANet, CornerNet, and CANet have overall achieved a considerable performance improvement. The detection accuracy of many categories (such as tennis courts, storage tanks (Storagetank), etc.) can reach the level of 70% to 80%. However, these algorithms still exhibit certain missed detections and false detections in scenes with more complex backgrounds or higher similarity in target shapes (for example, harbors (Harbor), toll stations (Express-toll-station)). Such scenes often require a higher level of feature representation capability, as well as more flexible geometric offsets and attention mechanisms, in order to more accurately distinguish between the background and real targets during the object detection process. In contrast, YOLOv7-b performs more robustly in these scenes, indicating that the DCNv2 (Deformable Convolution) module and other improvements introduced in its network structure can better cope with background interference and target deformation.

Regarding the single-stage detectors that have been popular in recent years, YOLOX and YOLOv8 have detection performance on some categories that is not far behind YOLOv7-b, and both perform well in balancing inference speed and accuracy. However, when facing multi-scale and high-density distribution scenes such as vehicles, ships (Ship), airports (Airport), or overpasses (Overpass), YOLOv7-b still has the upper hand in average precision. The key lies in the fact that YOLOv7-b not only inherits the advantage of the YOLO series models in pursuing real-time performance and accuracy, but also enhances the network's feature perception and localization capabilities for small and dense targets by integrating modules such as DCNv2 and BRA (Sparse Attention Mechanism). DCNv2 can achieve spatially adaptive offsets during the convolution process, possessing more flexible learning capabilities for targets with larger deformations, while BRA provides a stronger focusing effect in dense target areas, thereby reducing false and missed detections.

Further analysis targeting different categories also confirms the aforementioned advantages: in high-density or small target categories such as vehicles, storage tanks, and tennis courts, YOLOv7-b's detection capability is particularly prominent, capable of accurately capturing the boundaries and detail information of targets; in targets with large deformations or scale spans, such as bridges, airports, and stadiums (Stadium), YOLOv7-b also demonstrates excellent detection accuracy, indicating that its network can better balance the precision and robustness of multi-scale target localization in the process of integrating shallow details with deep semantics. In addition, in complex background or scenes with strong noise interference factors, such as overpasses and toll stations, the modulation mechanism and sparse attention introduced by YOLOv7-b effectively help the model separate real targets from the background, further reducing the false detection rate and missed detection rate.

In summary, the results in Table IV fully demonstrate the obvious advantages of YOLOv7-b over other mainstream detection algorithms on the DIOR dataset, covering the needs of various scenarios including small targets, large targets, complex backgrounds, and multi-scale targets. Building on the original design concept of the YOLO series that balances speed and accuracy, YOLOv7-b further combines innovative modules such as DCNv2 and BRA to achieve a more delicate depiction and precise localization of target features in high-density and complex scenes. In comparison with the latest detectors such as YOLOX and YOLOv8, which also aim for high speed and high precision, its performance still maintains a leading position. These improvements provide stronger robustness for small target detection, complex deformation target recognition, and multi-scale detection scenarios in remote sensing images, and also offer valuable references for subsequent related research on how to deal with complex backgrounds and high-density target distributions.

In addition to comparing the mAP@0.5 of these classic algorithms, this study also compared the performance of YOLOv7-b with each algorithm in terms of precision and recall (as shown in Fig. 9). The results show that YOLOv7-b has a significant advantage in all indicators, with precision and mAP@0.5 close to 1, and recall also maintained at a high level. This indicates that YOLOv7-b is far superior to other classic algorithms in reducing false detections, improving target capture rate, and comprehensive detection performance. Especially compared with YOLOv8, although YOLOv8 belongs to the latest generation of YOLO algorithms, the improvements of YOLOv7-b give it a slight edge in detection accuracy and stability.

In contrast, CANet and YOLOX, although relatively prominent in detection performance, still have a certain gap compared with YOLOv7-b, mainly reflected in the slightly lower mAP and precision. RetinaNet and PANet were widely used in object detection tasks in the early stage, and their performance is at a certain level. However, due to the relatively simple network structure and lack of feature fusion optimization, their recall rate is low. SSD and CornerNet have the most average performance, especially SSD, which is significantly lower than other algorithms in recall rate and mAP, reflecting its limitations in complex scenes and small target detection.

TABLE IV AP AMONG DIFFERENT ALGORITHMS

Class	Faster R-CNN	SSD	RetinaNet	PANet	CornerNet	CANet	YOLOX	YOLOv8	YOLOv7-b
Express-toll-station	55.2	53.1	62.8	66.7	76.3	77.2	85.6	84.7	82.3
Vehicle	23.6	27.4	44.2	47.2	43.0	51.2	85.0	83.0	85.6
Golffield	68.0	65.3	78.6	72.0	79.5	77.3	82.6	83.6	81.9
Trainstation	38.6	55.1	52.5	57.0	57.1	67.6	71.5	67.4	69.6
Chimney	70.9	65.8	72.3	72.3	75.3	79.9	81.6	93.5	94.9
Storagetank	39.8	46.6	45.8	46.3	47.2	70.9	81.0	91.0	92.9
Ship	27.7	59.2	71.1	71.7	37.6	81.0	91.0	95.6	95.7
Harbor	50.2	49.4	49.9	45.3	26.1	56.0	67.9	74.4	74.0
Airplane	53.6	59.5	53.3	56.9	58.8	70.3	88.9	95.2	95.5
Groundtrack field	56.9	68.6	76.6	73.4	79.5	83.6	87.1	88.4	89.6
Expressway-Service-area	69.0	63.5	78.6	72.5	81.6	83.5	93.5	87.9	88.8
Dam	62.3	56.6	62.4	61.4	64.3	67.7	76.6	73.3	80.1
Basketball court	66.2	75.7	85.0	80.5	80.8	87.8	92.1	88.7	89.4
Tenniscourt	75.2	76.3	81.3	80.9	84.0	88.2	92.3	94.7	95.3
Stadium	73.0	61.0	68.4	70.4	70.7	79.8	86.5	95.4	95.9
Baseballfield	78.8	72.4	69.3	70.3	72.0	72.0	86.7	96.3	96.3
Windmill	45.4	65.7	85.5	85.4	75.9	89.6	92.8	84.1	84.0
Bridge	28.0	29.7	44.1	43.6	46.4	55.7	55.8	61.3	62.9
Airport	49.3	72.7	77.0	72.3	84.2	82.4	89.1	84.6	89.4
Overpass	50.1	48.1	59.9	58.7	60.6	63.6	67.2	72.8	75.1

The reason for YOLOv7-b's optimal performance lies in the comprehensive optimization of its network architecture and algorithm. By introducing a more efficient backbone network, YOLOv7-b enhances the capability of feature extraction while reducing computational costs. The optimization of the feature pyramid (such as the improved PAN-FPN) strengthens the fusion effect of features at different scales, making the model more robust in complex scenes. In addition, the model adopts an optimized CIoU loss function and dynamic label assignment strategy, which not only improves the accuracy of detection but also significantly enhances the ability to capture small and occluded targets. Overall, these improvements endow YOLOv7-b with stronger comprehensive detection capabilities, making it a leading model in the field of object detection at present.

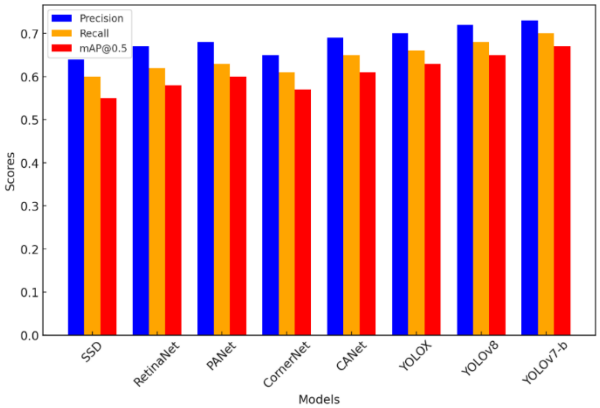
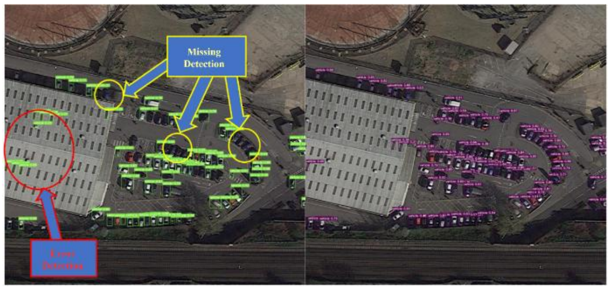


Fig. 9. Performance comparison of YOLOv7-b with classic algorithms.

B. Analysis of Detection Effect Experimental Results

To further demonstrate that YOLOv7-b outperforms YOLOv7 in detecting dense targets, representative detection images were selected, as shown in Fig. 10.



（a）YOLOv7 (b) YOLOv7-b

Fig. 10. Detection comparison between YOLOv7-b and YOLOv7.

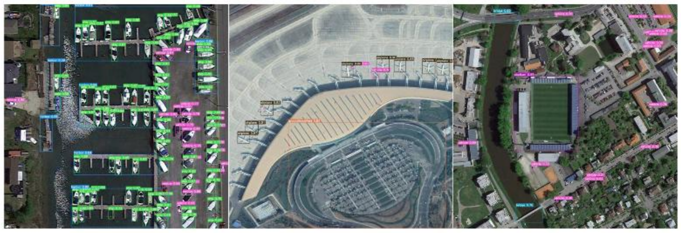
Fig. 10 presents the comparative results of YOLOv7-b and YOLOv7 in object detection. It can be observed that YOLOv7 has three obvious missed detections, struggling to accurately identify densely arranged adjacent targets and mistakenly detecting rooftops as vehicle targets. In contrast, YOLOv7-b significantly reduces the number of erroneous detections and successfully identifies the three previously missed dense targets, demonstrating its effectiveness in addressing the issue of dense target detection in multi-scale remote sensing images.

From the actual test results, both YOLOv7 and YOLOv7-b can effectively detect ship targets against a water background. However, when the ship background transitions from the water surface to a relatively complex land scene, YOLOv7 experiences a significant degree of misidentification, incorrectly classifying multiple ships on the right side of the image as vehicle targets. Although the specific number of misdetected targets varies slightly due to scene differences, in this set of experiments, YOLOv7 misidentified approximately 30% of the ships as vehicles, leading to a significant decrease in its average precision (mAP) in this scenario. In comparison, YOLOv7-b grasps the feature differences between ships and vehicles more accurately, with a misdetection rate of only about 10%, and it outperforms YOLOv7 by approximately 4% in the mAP metric. This difference indicates that the integrated and improved model has a more robust target discrimination capability in complex backgrounds.

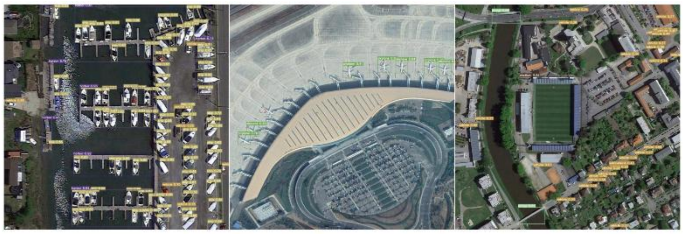
In the detection task of the second airport image, YOLOv7, due to its insufficient ability to distinguish between targets with similar structures or functions in the scene, mistakenly detected the highway beside the airport as an athletics field, and also misidentified some of the jet bridges next to the airplanes as vehicles. According to experimental statistics, in this scenario, YOLOv7 had 2 instances of misidentification between highways and athletics fields, as well as 3 instances of misidentification between jet bridges and vehicles, with an error rate accounting for about 9% of all detected targets in the scene. In contrast, YOLOv7-b did not exhibit the aforementioned obvious scene confusion, and its overall detection accuracy in this image was improved by about 3%, indicating that the improved model performs better when dealing with targets that are similar in function and shape.

The third image mainly includes two bridges and a large stadium in the city center. Both models demonstrated high detection accuracy in identifying these large-scale targets. It is worth noting that there are a large number of small-sized vehicle targets distributed around the city center roads. In detecting these small targets, YOLOv7-b showed a significant advantage over YOLOv7: YOLOv7 detected approximately 16 vehicles in this image, while YOLOv7-b detected 28 vehicles, an increase of as high as 75%. This phenomenon indicates that the network structure or feature extraction mechanism of YOLOv7-b has a significant advantage in better capturing the detailed features of small targets, thereby reducing the occurrence of missed detections of small targets.

Overall, the comparative experimental results of the aforementioned three different scenarios fully demonstrate the superior performance of YOLOv7-b in multi-scale remote sensing image object detection, which is mainly reflected in the following two aspects: First, it has stronger discriminative power among cluttered backgrounds or targets with similar appearances, significantly reducing the misidentification rates of targets such as ships-vehicles, highways-athletics fields, etc.; Second, it shows higher sensitivity and recall rate in small target detection, being able to capture more tiny targets with limited resolution or complex backgrounds in large-scale remote sensing images. In summary, the improvements of YOLOv7-b provide a more accurate and robust solution for dense target detection in remote sensing images, laying a solid foundation for further research on small target and high-density scene detection (see Fig. 11).



（a）YOLOv7



（b）YOLOv7-b

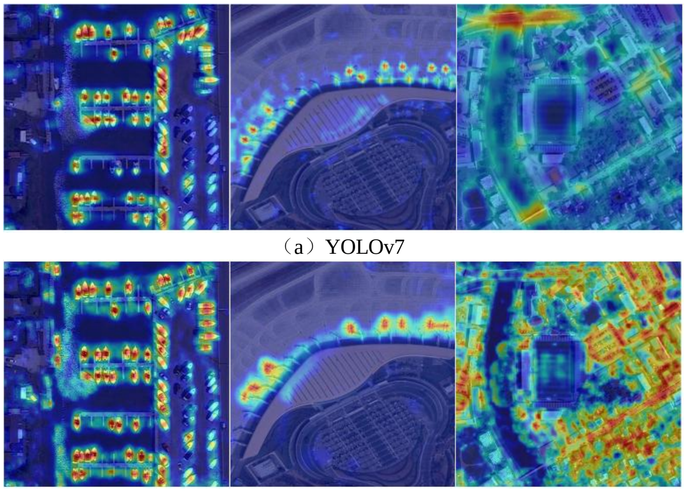
Fig. 11. Comparison of detection images between YOLOv7-b and YOLOv7.

The following figure (Fig. 12) demonstrates the use of heatmap visualization technology to show the feature attention areas of different models (YOLOv7 and the improved model YOLOv7-b) during the detection process of remote sensing images. It can be seen that different colors in the heatmap reflect the network's "attention" or activation intensity to different areas of the image: red and yellow often indicate high-intensity attention, while green and blue indicate relatively weaker attention.

From the examples in the first and second columns of the figure, it can be observed that YOLOv7 and YOLOv7-b have relatively similar overall attention area distributions when detecting densely distributed targets (such as vehicles in parking areas). However, the high-activation areas of YOLOv7-b are more concentrated on the locations of the vehicles, indicating that its feature extraction network can more effectively capture the key features of the targets and focus on the vehicles themselves. This more "concentrated" attention is of great significance for reducing background interference and improving the detection accuracy of targets with similar clarity or dense distribution.

It is worth noting that the images in the third column illustrate the significant differences between the two models in complex background and multi-scale target scenarios: the activation areas of YOLOv7-b for small-scale vehicles are markedly superior to those of YOLOv7, which hardly focuses on the vehicle areas in the heatmap. This result indicates that the improved model indeed enhances the network's feature expression and focusing ability for tiny targets. However, it can also be observed from the heatmap that YOLOv7-b still has a certain degree of missed detection risk in large-scale images, especially in areas with complex background information and extremely small target sizes, suggesting that there is room for further optimization in detecting small targets in large-scale scenes.

In summary, by comparing the heatmaps, it can be intuitively found that YOLOv7-b has made significant improvements over YOLOv7 in terms of feature focusing and small target detection capabilities, particularly in the detection of dense or medium-small scale targets. Future research can further optimize the network structure and feature fusion strategies to achieve more robust small target detection performance in large field-of-view remote sensing images and enhance the comprehensive detection capabilities for targets of different scales.



（a）YOLOv7

（b）YOLOv7-b

Fig. 12. Comparison of detection heatmaps between YOLOv7-b and YOLOv7.

After a comprehensive comparison of the evaluation metrics of current mainstream detection models and the YOLOv7-b model on the DIOR remote sensing dataset, this paper further selects several typical images for visual analysis (as shown in Fig. 13) to intuitively present the detection performance of the models. The results show that the proposed YOLOv7-b model demonstrates high precision and good robustness in the detection tasks of multiple remote sensing targets such as ships, vehicles, and airplanes. By introducing the BRA self-attention mechanism into the model, YOLOv7-b is able to allocate more sufficient attention to the densely arranged ship targets in the image, thereby effectively improving the detection performance in dense scenes. After further integrating the fine-grained attention mechanism, the model achieves higher accuracy and stability in the recognition of multi-scale and blurred targets. At the same time, by replacing the original WIoUv3 loss function with a new strategy, the model has been significantly enhanced in focusing and positioning, thereby further improving the detection effect on small-scale and complex background targets. Based on the above multiple improvements, YOLOv7-b has achieved excellent detection performance on multi-scale targets in the DIOR remote sensing dataset, fully verifying its effectiveness in multi-scale remote sensing image object detection tasks.

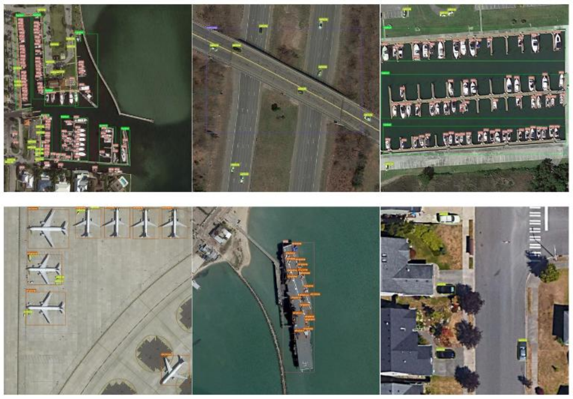


Fig. 13. Detection performance of YOLOv7-b on the DIOR remote sensing image dataset.

IV. CONCLUSION

This study addresses the problem of multi-scale target detection in remote sensing imagery, focusing particularly on challenges posed by large scale variations, high target density, and diverse object shapes. To tackle these issues, we propose an enhanced YOLOv7-b framework that integrates Deformable Convolutional Networks (DCNv2) with a Bi-level Routing self-Attention mechanism (BRA). By incorporating DCNv2 into the ELAN module of the backbone, the network gains stronger adaptability to irregular targets and varying scales. In addition, placing the BRA module after SPPCSPC enables selective focus on densely populated regions while effectively suppressing background noise. Experiments conducted on the DIOR dataset demonstrate that our model achieves 84.3 % mAP@0.5, 89.57 % Precision, and 78.63 % Recall. Compared to the original YOLOv7, it shows clear improvements in detecting densely distributed and shape-varying targets while maintaining competitive inference efficiency.

Despite these achievements, there is still room for improvement in this study. Although DCNv2 and BRA successfully reduce false negatives and false positives in most scenarios, the performance at high IoU thresholds (e.g., mAP@0.5:0.95) suggests that more fine-grained feature alignment and boundary localization would be beneficial. Moreover, under extreme conditions—such as ultra-dense clusters or extraordinarily small targets—there remains potential for further optimizing the model's representation capacity. Future work could explore incorporating advanced strategies (e.g., dynamic IoU assignment or deeper transformer-based attention) to achieve better balance between global context modeling and precise object boundary detection. Additionally, adopting multi-modal data integration (e.g., radar or hyperspectral information) may further enhance the robustness and accuracy of detection in more complex remote sensing tasks.

Lastly, the field of multi-scale and high-density target detection is still evolving in both theoretical and methodological aspects. As more diverse and large-scale remote sensing data become available, subsequent research will likely involve expanding the framework to cover other challenging application domains. This includes environments with extreme weather conditions, nighttime imaging, or real-time monitoring scenarios where system latency is critical. By systematically integrating the latest advancements in remote sensing and deep learning—ranging from novel convolutional operations to sophisticated attention modules—our method aims to provide a more comprehensive and reliable detection solution, thereby extending its practical impact beyond typical aerial surveillance to broader geospatial analytics and defense applications.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] K. Li, G. Wan, G. Cheng, et al., “Object detection in optical remote sensing images: A survey and a new benchmark,” ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 296–307, July 2020.

[2] L. Wen, Y. Cheng, Y. Fang, et al., “A comprehensive survey of oriented object detection in remote sensing images,” Expert Syst. Appl., vol. 224, p. 119960, Mar. 2023.

[3] D. X. U, Y. W. U, “Research progress of deep learning algorithms for object detection in optical remote sensing images,” Natl. Remote Sens. Bull., 2024, pp. 1–30.

[4] S. Gui, S. Song, R. Qin, et al., “Remote sensing object detection in the deep learning era—a review,” Remote Sens., vol. 16, no. 2, p. 327, Jan. 2024.

[5] Z. Yao, W. Gao, “Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., 2024.

[6] H. Chen, Y. Xie, X. Xu, et al., “Design of a disaster meteorological observation data monitoring system based on multi-source satellite remote sensing,” Comput. Meas. Control, vol. 31, no. 4, pp. 24–29, Apr. 2023.

[7] Y. Liao, H. Wang, C. Lin, et al., “Research progress on object detection in optical remote sensing images based on deep learning,” J. Commun., vol. 43, no. 5, pp. 190–203, May 2022.

[8] W. Wang, Y. Fu, F. Dong, et al., “Semantic segmentation of remote sensing ship image via a convolutional neural networks model,” IET Image Process., vol. 13, no. 6, pp. 1016–1022, 2019.

[9] S. Grigorescu, B. Trasnea, T. Cocias, et al., “A survey of deep learning techniques for autonomous driving,” J. Field Robot., vol. 37, no. 3, pp. 362–386, May 2020.

[10] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2023, pp. 7464–7475.

[11] C. Y. Wang, H. Y. M. Liao, I. H. Yeh, “Designing network design strategies through gradient path analysis,” arXiv preprint arXiv:2211.04800, 2022.

[12] X. Li, W. Wang, L. Wu, et al., “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” in Adv. Neural Inf. Process. Syst., 2020, vol. 33, pp. 21002–21012.

[13] Q. Wang, B. Wu, P. Zhu, et al., “Efficient channel attention for deep convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020.

[14] Z. Qin, P. Zhang, F. Wu, et al., “FCANet: Frequency channel attention networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2021, pp. 783–787.

[15] Z. Huang, X. Wang, L. Huang, et al., “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, p. 603.

[16] S. M. Azimi, E. Vig, R. Bahmanyar, et al., “Towards multi-class object detection in unconstrained remote sensing imagery,” in Proc. Asian Conf. Comput. Vis., 2018, pp. 150–165.

[17] Z. Deng, H. Sun, S. Zhou, et al., “Multi-scale object detection in remote sensing imagery with convolutional neural networks,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 3–22, Oct. 2018.

[18] C. Liu, S. Zhang, M. Hu, et al., “Object detection in remote sensing images based on adaptive multi-scale feature fusion method,” Remote Sens., vol. 16, no. 5, p. 907, Mar. 2024.

[19] W. Liu, L. Ma, J. Wang, “Detection of multiclass objects in optical remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 16, no. 5, pp. 791–795, May 2018.

[20] J. Chen, L. Wan, J. Zhu, et al., “Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 4, pp. 681–685, Apr. 2019.

[21] X. Ying, Q. Wang, X. Li, et al., “Multi-attention object detection model in remote sensing images based on multi-scale,” IEEE Access, vol. 7, pp. 94508–94519, 2019.

[22] L. W. Li, J. B. Xi, W. D. Jiang, et al., “Multi-scale fast detection of objects in high resolution remote sensing images,” in Proc. IEEE 5th Int. Conf. Image, Vision Comput. (ICIVC), Jun. 2020, pp. 5–10.

[23] X. Zhang, K. Zhu, G. Chen, et al., “Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network,” Remote Sens., vol. 11, no. 7, p. 755, Jul. 2019.

[24] Y. Cheng, W. Wang, W. Zhang, et al., “A multi-feature fusion and attention network for multi-scale object detection in remote sensing images,” Remote Sens., vol. 15, no. 8, p. 2096, Apr. 2023.