WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.34
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Improved YOLOv8 Remote Sensing Small Target Detection Method

Peng Wang *, Hanwei Mao, Xiangmeng Ren, Jingwei Yang

School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

*Corresponding Author: Peng Wang (Email: 1743048640@qq.com)

ABSTRACT

Due to the miniaturization, dense arrangement, variable viewing Angle and complexity of background environment in satellite remote sensing technology, the traditional detection methods often encounter the challenge of misidentification and missing detection, thus limiting the detection accuracy. In order to overcome these problems and improve detection efficiency, this paper innovatively proposes an optimized Yolov8 model, which is deeply customized and improved for the detection of tiny objects in remote sensing images. Firstly, BiFormer model is introduced. BiFormer model introduces a two-layer routing attention mechanism, which significantly improves the accuracy and robustness of target detection. In addition, the SPD-Conv module is introduced to realize the conversion from space to depth, and better capture the target features of different dimensions. After rigorous validation of the DOTA-v1.0 dataset, the optimized model achieved a significant improvement in average accuracy (mAP), reaching a level of excellence of 60.3%, which represents a performance leap of about 2.3 percentage points compared to traditional models. It has further promoted the technological progress and application deepening in this field.

KEYWORDS

Target detection; Satellite remote sensing; BiFormer module; SPD-Conv module

1. INTRODUCTION

Satellite remote sensing target detection, as the forefront of deep integration of remote sensing technology and computer vision, focuses on capturing Earth surface images from the perspective of vast space, laying a data highway for accurate identification and positioning of specific targets on the surface. The main task of satellite remote sensing image target detection is to accurately identify and locate the target object with specific shape, color, texture and other features in the image taken by satellite. However, the acquisition environment of remote sensing image is complicated, and the natural factors such as light intensity, shadow change and object masking are like fog, which add many obstacles to the clear presentation of the object. In view of the above problems, researchers are committed to developing an optimization model to open a new chapter of efficient and accurate detection of satellite remote sensing targets [1]. Mei [2] et al. used SPD-RConv to replace the Repconv of the original network to enhance attention to high-value information, and recalculated the candidate boxes of the target using K-means clustering algorithm. Hu et al. [3] proposed a remote sensing target detection model based on YOLOX-Tiny. By improving the multi-scale feature fusion network to make full use of shallow detail information and deep semantic information, the detection ability of small targets is improved. Xie [4] et al. introduced lightweight network Ghost convolution to replace the conventional convolution of the original network, reducing the computational load of

convolutional operations in the network and thus improving the detection efficiency. Liu et al. [5] proposed a model named Sparse R-CNN, in which a self-supervised learning framework and selective query collection were introduced to improve the target detection effect of remote sensing images. Liu et al. [6] proposed a local adaptive threshold algorithm combining non-uniformity and compactness filters for small target detection in infrared remote sensing images. Li et al. [7] introduced spatial pyramid and cross-stage partial channel module, which combined the idea of feature separation and merging. Reduce missed detection and noise effects in small target scenarios. There are still some limitations in the performance of the above detection models, which affect the detection accuracy. In view of the above problems, the improved model is committed to improving the detection accuracy and accuracy of the model on the premise of ensuring the detection speed. This paper takes Yolov8n as the benchmark model. Firstly, BiFormer model is introduced. BiFormer model introduces a double-layer routing attention mechanism, which significantly improves the accuracy and robustness of target detection. In addition, the SPD-Conv module is introduced to realize the conversion from space to depth, and better capture the target features of different dimensions.

2. BASE MODEL

YOLOv8 is the latest iteration of the YOLO family of object detection models. On the basis of inheriting the advantages of the previous generation model, YOLOv8 introduces new network structure and algorithm optimization to further improve the performance of the model. The anchor-free detection head is adopted, and CSPNet is introduced as the basic structure, which makes the model have higher detection accuracy and faster reasoning speed while maintaining light weight. YOLOv8 offers a variety of models with different depths and widths, including n, s, m, l, and x. Among them, Yolov8n is the baseline model chosen in this paper because of its fast, accurate and easy deployment in remote sensing images. YOLOv8's backbone network is responsible for extracting features from input images. It uses an efficient convolution layer structure, combines the advantages of depth-separable convolution and standard convolution, improves the computational efficiency and reduces the number of parameters. At the same time, the introduction of residual connections and novel activation functions, such as SiLU or Mish, further enhanced the performance of the network.

3. MODEL OPTIMISATION

3.1. BiFormer Module

In a traditional vision Transformer, the attention mechanism is the core building block for capturing long-term dependencies in the data. However, this capability often comes with high computing costs and memory footprint. To solve this problem, BiFormer module [8] introduced a two-layer routing attention mechanism, aiming to balance computational efficiency and performance through a dynamic sparse attention mechanism. The two-layer routing attention mechanism is the core of BiFormer module, which realizes the flexible allocation of computation and content awareness through two-layer routing. Specifically, the mechanism first filters out unrelated key-value pairs at the coarse-grained region level, and then applies fine-grained token-to-token attention on the union of the remaining candidate regions. The input feature map is divided into several non-overlapping regions, each containing several feature vectors. The query (Q), key (K) and value (V) tensors are obtained by linear transformations. Construct a region-level directed graph to identify other areas that each region should focus on. Firstly, the interregional affinity matrix (adjacency matrix) is calculated, that is, the correlation between two regions. Then, the most important k connections of each area are reserved through the top-k operation to form the route index matrix. Apply fine-grained token-to-token attention to the remaining routing areas based on the routing index matrix. This step ensures

that attention is focused on the areas most relevant to the query, thereby reducing unnecessary calculations.The structure of BiFormer module is shown in Figure 1.

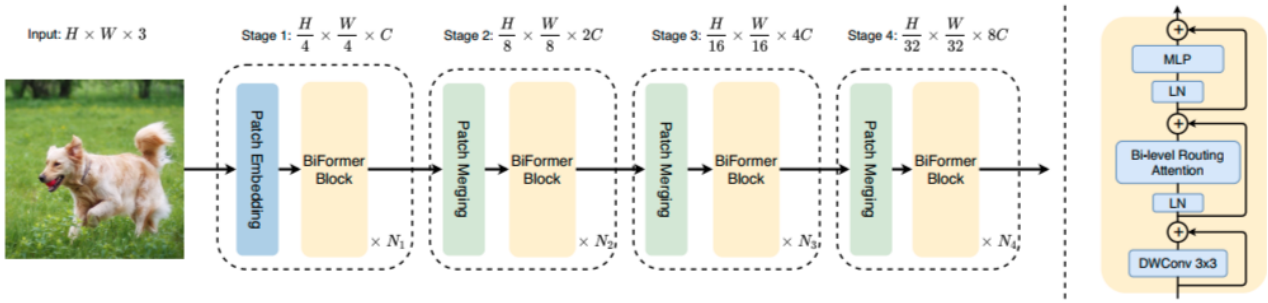


Figure 1. Structural diagram of c3 module and c2f module3.2.SPD-Conv module

In the YOLOv8CNN model, the Head part is usually responsible for the detection task, which integrates multiple layers to fuse multi-scale features in order to better detect objects of different sizes. However, for small targets or complex scenes, these traditional structures sometimes have difficulty adequately capturing detail, resulting in reduced detection accuracy. To improve this situation, the SPD-Conv module can be introduced in the Head section [9]. The core idea of this module is to transform the spatial information of the feature map into richer depth information, which can help the model better "understand" and distinguish the subtle features.

SPD-Conv module is divided into two steps: First, it adopts the spatial to depth conversion, which summarizes the information of each small piece of the feature map into a channel, so that the information that was originally scattered in space is integrated into the depth. In this way, each channel contains information from a different spatial location, thus enhancing the expression of features. These transformed features are then further processed using non-step convolution layers to extract higher-level feature representations. This step helps the model learn more useful features from the transformed depth information, thereby improving detection performance. Consider mapping the intermediate feature of any size S×S×C1 as X, and segment the sequence of sub-feature maps as shown in equation (1).

$$
\begin{aligned}
&f_{0,0} = X[0:s:scale, 0:s:scale],\ f_{1,0} = X[1:S:scale, 0:S:scale], ..., \\
&\qquad\qquad f_{scale-1,0} = X[scale-1:S:scale, 0:S:scale]; \\
&f_{0,1} = X[0:s:scale, 1:s:scale],\ f_{1,1}, ..., \\
&\qquad\qquad f_{scale-1,1} = X[scale-1:S:scale, 1:S:scale]; \\
&\qquad\qquad \vdots \\
&f_{0,scale-1} = X[0:s:scale, scale-1:S:scale],\ f_{1,scale-1}, ..., \\
&\qquad\qquad f_{scale-1,scale-1} = X[scale-1:S:scale, scale-1:S:scale];
\end{aligned}
\tag{1}
$$

4. EXPERIMENTAL RESULT AND ANALYSIS

4.1. Datasets

In order to study object detection in remote sensing images, DOTA-v1.0 [10] dataset released by Wuhan University is selected, which is a huge dataset specially designed for object detection in remote sensing images. In total, the dataset contains 2,806 images tagged with more than 188,000 goals that fall into 15 categories common to everyday life. What makes this dataset special is that it contains images in a variety of weather, color, and brightness conditions, and many of them are dense small targets, which is exactly what this experiment needs for small target detection. In order to train and verify the model, 1893 pictures were randomly selected from the data set as the training set for the model learning. Another 631 images were used as validation sets to evaluate the performance of

the model. This partitioning method not only ensures the training effect of the model, but also facilitates the accurate evaluation of the model.

4.2. Experimental Environment and Parameter Settings

In this paper, YOLOv8n was used as the experimental baseline. The input image size was 640x640, batch_size was set to 8, and epoch was set to 300. All experiments were conducted in the same environment, as shown in Table 1.

Table 1. Experimental environment configuration

Hardware or software	Parameters
GPU	NVIDIA GeForce RTX 2080 Ti
CPU	I7-9700K
Deep learning frameworks	Pytorch 1.12.1
GPU accelerated environment	CUDA 1.17.1
Programming language	Python3.9

4.3. Evaluation Indicators

In this experiment, this field and its important evaluation indexes were used to test the performance of the model, namely, Precision P (Precision), Recall R (Recall), Average Precision AP (Average Precision), and mean average precision mAP (mean Average Precision). The corresponding calculation formula of each index is shown as formula (2) ~ Formula (5).

$$P = \frac{TP}{TP+FP} \times 100\% \tag{2}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{3}$$

$$AP = \int_0^1 P(R)dR \tag{4}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{5}$$

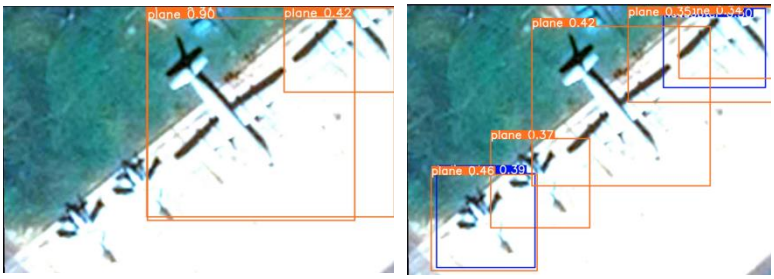
4.4. Ablation Experiments

In order to gain a deeper understanding of the contribution of each module in the YOLOv8n benchmark model to the overall performance, we used a precise control variable method, the ablation experiment, to evaluate the effect of different modules one by one through a combination of them. A total of 4 different model configurations were designed in the experiment, and the detailed results are shown in Table 2. As can be seen from Table 2, the introduction of biformer has increased the accuracy rate by 0.4%, the recall rate by 1.8%, mAP@0.5 by 0.6%, mAP@0.5:0.95 by 0.8%, and the detection speed by 47.8FPS. After the introduction of spd, the accuracy rate increased by 1%, the recall rate increased by 3.5%, mAP@0.5 increased by 1.2%, mAP@0.5 increased by 2%, and the detection speed reached 42FPS. Yolov8n-BS is a detection model with all improved point fusion, and the final average detection accuracy is as high as 60.3, which is 2.3 percentage points higher than the benchmark model. For remote sensing image data set, the improved method proposed in this paper has high accuracy and robustness, and avoids the phenomenon of missing detection and misdetection.

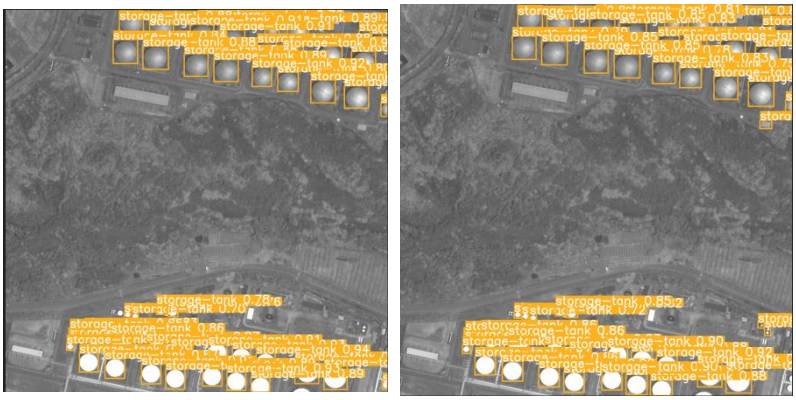
Table 2. Ablation experiments

Algorithm	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	FPS
YOLOv8n	68.9	54.3	58	32.8	45.8
YOLOv8n-biformer	69.3	56.1	58.6	33.6	47.8
YOLOv8n-spd	69.9	57.8	59.2	34.8	42
YOLOv8n-BS	70.4	58.4	60.3	35.4	41.5

(a) Dense small target environment



(b) Strong light environment



(c) Extreme darkness

Figure 2. Comparison of detection results before and after model improvement

4.5. Detection Effect

In view of the challenging factors often encountered in remote sensing image acquisition, such as high target density, image blur and abnormal lighting conditions, in order to intuitively demonstrate the performance improvement of the optimized model in processing complex scenes compared with the traditional model, a comparison diagram of the detection effects of the two models under various

restricted environments was constructed. The comparison focuses on dense small target recognition, target detection under abnormal lighting conditions, and fine edge detection tasks. As shown in Figure 2, the traditional model has poor detection ability in the face of the above complex scenes, and the phenomenon of wrong detection and missing detection is frequent. In contrast, the new and improved model shows excellent adaptability and stability, maintaining high detection accuracy and classification accuracy even under extremely dense targets or extremely abnormal lighting conditions. This significant advantage shows that, therefore, the model designed in this paper is particularly suitable for dealing with the specific needs of remote sensing detection fields such as small target recognition and remote monitoring, in which the characteristics of small target scale and long imaging distance are particularly prominent, so as to effectively overcome the limitations of traditional methods in these complex scenarios.

5. CONCLUSION

By analyzing the characteristics of remote sensing detection, the limitations of detection methods in this field are summarized, and a detection model applied to remote sensing is proposed. BiFormer model is introduced. BiFormer model introduces a two-layer routing attention mechanism, which significantly improves the accuracy and robustness of target detection. In addition, the SPD-Conv module is introduced to realize the conversion from space to depth, and better capture the target features of different dimensions. After rigorous validation of the DOTA-v1.0 dataset, the optimized model achieved a significant improvement in average accuracy (mAP), reaching a level of excellence of 60.3%, which represents a performance leap of about 2.3 percentage points compared to traditional models. The improved model can extract surface information more quickly and accurately to meet the needs of environmental detection and urban planning.

REFERENCES

[1] Chen F K, Li S X. Improved Yolov5 for Target Detection in Unmanned Aerial Vehicle [J]. Journal of Computer Engineering & Applications, 2023, 59(18).

[2] Mei Y L, Cui L K, Geng X J, et al. Remote Sensing Target Detection Based on Improved YOLOv7 [J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2024, 40(04): 38-44.

[3] Hu Z H, Li Y H. Remote Sensing Target Detection Based on YOLOX-Tiny with Biased Feature Fusion Network [J]. Remote Sensing Technology and Application, 2024, 39(03): 590-602.

[4] Xie X. Research on Lightweight Remote Sensing Image Target Detection Algorithm Based on YOLOv8 [D]. Dongguan University of Technology, 2024. DOI: 10.44357/d.cnki.gdgut.2024.000142.

[5] Liu B, Duan R. Research on Remote Sensing Target Detection Based on Sparse R-CNN [J]. Journal of Changchun University of Technology, 2024, 45(02): 147-152. DOI: 10.15923/j.cnki.cn22-1382/t.2024.2.07.

[6] Liu C, Xie F, Dong X, et al. Small target detection from infrared remote sensing images using local adaptive thresholding [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 1941-1952.

[7] Li K, Wang Y, Hu Z. Improved YOLOv7 for small object detection algorithm based on attention and dynamic convolution [J]. Applied Sciences, 2023, 13(16): 9316.

[8] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[9] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459.

[10] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3974-3983.