Remote sensing small target detection optimization based on YOLOv8

Lujianheng[1] Zhudandan[1] Jiarui[1] Wangning[2]（Corresponding author）

1 School of Artificial Intelligence, Guangzhou Huashang University, Guangzhou, Guangdong, China，511300；

2 School of Artificial Intelligence, Guangzhou Huashang University, Guangzhou, Guangdong, China，511300；

Abstract: To enhance YOLOv8's performance in small remote sensing target detection, this study addresses improvements in network structure, feature fusion, label matching, data augmentation, and loss functions. An optimization model is proposed that integrates a lightweight Backbone, a CBAM attention mechanism, and a BiFPN architecture with multi-scale feature fusion, along with a decoupled detection head, to enhance the ability to detect small target features. K-means++ Anchor clustering and a dynamic label matching strategy are combined to improve positioning accuracy and boost training efficiency.

Keywords : remote sensing image; small target detection; YOLOv8; feature fusion; anchor design; loss function optimization

DOI： 10.69979/3041-0843.25.04.006

1 Introduction

In remote sensing images, small objects are easily affected by factors such as scale, background complexity, and image blur, making detection significantly more difficult. YOLOv8, a newly emerging object detection algorithm, is widely used in remote sensing image processing and optimization due to its balanced approach between accuracy and speed.Wang et al. [1] (2024) developed an improved YOLOv8 method for remote sensing image target detection, which greatly improved the accuracy of detection under complex background conditions; Zhao et al. [2] (2024) developed a lightweight G-YOLO model for infrared remote sensing data to meet the needs of real-time operation of drone platforms; Yi et al. [3] (2023) developed an improved YOLOv8 algorithm for remote sensing small targets, which still performs well in the case of scale imbalance; Liu et al.[4] (2024) developed a YOLO- PDNet model, which uses structural adjustment to improve the perception level of extremely small targets; Bi et al. [5] (2025) proposed SPDC-YOLO, which further optimized the shallow feature channel and expanded the model's expression capability.

Building on the existing foundation, this study constructed a YOLOv8 optimization framework that integrates structural optimization innovation, multi-scale enhancement and expansion, and training mechanism improvement and update, and conducted experimental verification on multiple sets of remote sensing datasets, aiming to improve the accuracy, stability and deployment efficiency of small target detection.

2 Remote Sensing Small Target Detection Optimization Framework Based on YOLOv8

In response to the insufficient features, positioning errors and coupling interference of YOLOv8 in remote sensing small target detection tasks, this paper constructs a multi-module collaborative optimization model (see Figure 2.1). With a lightweight backbone as the cornerstone, it focuses on enhancing the attention to small area responses, adopts multi-scale feature pyramid technology to enhance semantic expression and dense target recognition, and adopts a sample reconstruction module to achieve the decoupling of classification and regression in the detection task, thereby improving accuracy and robustness. Mosaic and Copy-Paste technologies are used to enhance the effect at the data level, and anchor box clustering and dynamic label matching are introduced to optimize the sample allocation strategy. A weighted loss function is implemented to focus on small targets and improve the detection level.

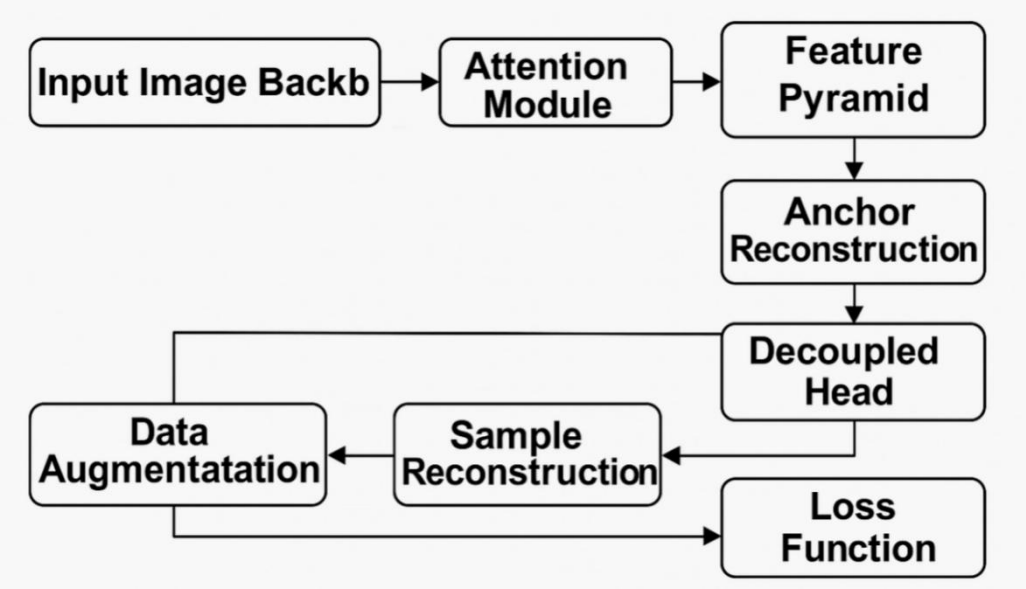

Figure 2.1 Remote sensing small target detection optimization framework based on YOLOv8

2.1 Network structure lightweighting and feature enhancement improvements

The key idea is to use a set of standard convolutions to generate the main feature image, and then generate redundant features by linear transformation, thereby reducing the number of parameters and calculations. Speaking of the Ghost module output, its calculation formula is:

$$Y = Concat\left(y_1, y_2, \ldots, y_m\right), \quad y_i = \phi_i(X), \quad i = 1, 2, \ldots, m \ (3.1)$$

Here, is X defined as the input feature map, ϕ_i i (·) represents the function that achieves the i - th linear transformation, and m is the number of Ghost features. This construction effectively reduces the amount of redundant computation while maintaining the semantic expression capability.The calculation of channel attention is shown below:

$$M_c(X) = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X))) \ (3.2)$$

Among them, $\sigma(\cdot)$ is the Sigmoid function, MLP is the multi-layer perceptron, $AvgPool$ (·) and $MaxPool$ (·) correspond to the global average pooling and maximum pooling operations respectively.

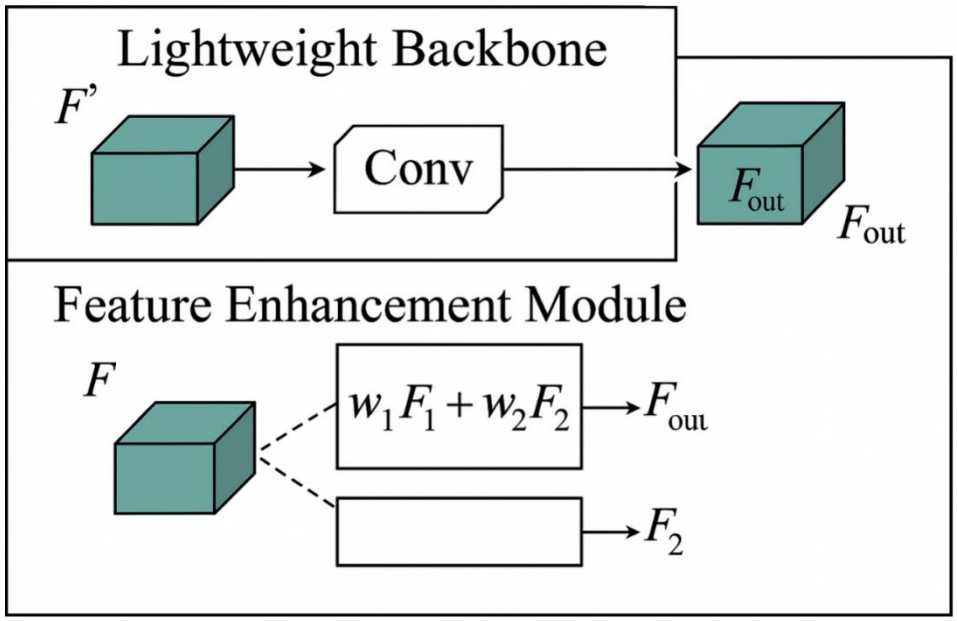

Figure 3.1 Network structure lightweight and feature enhancement module structure diagram

2.2 Multi-scale feature fusion and detection head improvement

The weighted output of the feature fusion node is:

$$O_i = \frac{\sum_{j \in \mathrm{I}(i)} w_j F_j}{\sum_{j \in \mathrm{I}(i)} w_j + \dot{O}} \quad (3.3)$$

Among them, O_i is the feature output by the F_i i- th layer, is the input feature map, is w_i the fusion weighting coefficient that can be clearly learned, and $\mathrm{I}(i)$ can be regarded as the set formed by the input nodes of the i-th layer. $\dot{O}$ In order to avoid the problem of division by zero in the division, the extremely small constant is used. Assuming that the input features are F , the outputs of classification and regression are respectively:

$$P_{cls} = \sigma\left(W_{cls} * F + b_{cls}\right), \quad P_{reg} = W_{reg} * F + b_{reg} \quad (3.4)$$

Among them, W_{cls} , W_{reg} are the weights of the classification and regression branches respectively, b_{cls} , b_{reg} is the bias factor, $*$ which represents the convolution operation, $\sigma(\cdot)$ is the Sigmoid activation function, and adopts the task decomposition approach to greatly improve the accuracy of small target positioning and the classification performance.

3 Experimental verification and result analysis

3.1 Experimental Setup and Dataset Description

The dataset, based on the VisDrone dataset released by Tianjin University's AISKYEYE, covers 14 cities and various scenarios and lighting environments. The samples were grouped into 6471 training samples, 548 validation samples, and 3190 test samples. Small objects accounted for approximately 90% of the detection samples. The images, initially 720×1280 to 1080×1920 resolution, were resized to a uniform 640×640 resolution. Enhancements such as random flipping and brightness and color adjustments were applied, along with bounding box coordinates scaling. A GIoU loss normalization procedure was implemented, and a one-hot encoding technique for the categories was used to filter out small objects smaller than 10 pixels. Table 3.1 presents the statistics of the preprocessed samples.

Table 3.1 Statistics of the first 8 samples of the training set after preprocessing

Image ID	Native resolution WxH	Target quantity	Average target area (px^2)	Small target ratio (<32px)
VisDrone_0001	1280×720	32	724	0.84
VisDrone_0002	1920×1080	47	658	0.91
VisDrone_0003	1280×720	15	412	0.87
VisDrone_0004	1920×1080	59	538	0.89
VisDrone_0005	1280×720	twenty three	689	0.82
VisDrone_0006	1920×1080	41	601	0.88

3.2 Small Object Detection Performance Evaluation

To further verify the detection capability of our model for targets of different sizes, we divided the test set targets into three types according to their area (small targets <32×32, medium targets with an area of 32 to 96, and large targets with an area exceeding 96×96). We calculated the mean average precision (mAP), recall rate, and F1 score for each type of target, demonstrating the model's adaptability and accuracy in the small target scenario. We focused on the small target category, which accounts for more than 60% of the entire dataset and is a major factor affecting remote sensing detection performance.

Table 3.2 Detection performance evaluation indicators divided by target size

Size category	mAP (%)	Recall (%)	F1 Score	Precision (%)
Small	82.5	78.4	0.80	84.7
Medium	88.6	83.2	0.86	89.4
Large	91.3	86.9	0.89	92.6
Overall average	87.5	82.8	0.85	88.9

4 Conclusion

This study addresses the accuracy limitations and model adaptability challenges of small object detection in remote sensing images. Leveraging YOLOv8, this research builds a detection system that integrates structural improvements,

matching strategy adjustments, and enhanced training mechanisms. Future research could further integrate the Transformer architecture, multimodal remote sensing data (such as SAR and optical fusion), and cross-domain transfer mechanisms to enhance the model 's generalizability and task adaptability in complex scenarios.

References

[1]Wang H, Yang H, Chen H, et al. A remote sensing image target detection algorithm based on improved YOLOv8[J]. Applied Sciences, 2024, 14(4): 1557.
[2]Zhao X, Zhang W, Xia Y, et al. G-YOLO: A Lightweight Infrared Aerial Remote Sensing Target Detection Model for UAVs Based on YOLOv8[J]. Drones, 2024, 8(9): 495.
[3]Yi H, Liu B, Zhao B, et al. Small object detection algorithm based on improved YOLOv8 for remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 17: 1734-1747.
[4]Liu X D, Zhang H, Gong W, et al. YOLO-PDNet: Small Target Recognition Improvement for Remote Sensing Image Based on YOLOv8[C]//2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024: 1-9.
[5]Bi J, Li K, Zheng X, et al. SPDC-YOLO: An Efficient Small Target Detection Network Based on Improved YOLOv8 for Drone Aerial Image[J]. Remote Sensing, 2025, 17(4): 685.

This research was funded by the fund project "Research on Object Detection Algorithm Based on Deep Learning" (Project No.: 2022HSXS087).