

# Research on small object detection method based on YOLOv8: Application of high-resolution detection layer and Mosaic data augmentation

YiDa Lin \*

School of Henan Polytechnic University, Jiaozuo 454003, China

\* Corresponding Author Email: todaytomorrow.lyd@qq.com.

**Abstract.** For the core problem that the number of small targets in remote sensing images is continually increasing and traditional object detection methods struggle to meet its detection requirements, a small-target detection method based on YOLOv8 is proposed, and its detection performance is enhanced through two key improvements. First, to address the difficulty of detecting small targets, a high-resolution detection layer P2 is added to the Neck part of the YOLOv8 model to enhance the model's feature extraction capability for small targets. Secondly, the Mosaic data augmentation technique is introduced, which stitches multiple images into one image to expand the diversity of training data, thereby improving the model's generalization to small targets. Experiments on the NWPU VHR-10 dataset show that the improved YOLOv8 model significantly outperforms the original model in small-target detection. Specifically, the Recall rate of the improved model reaches 85.6%, which is 9.18% higher than that of the baseline model. At the same time, mAP@0.5 edged up to 0.724, while the more stringent mAP@0.5:0.95 index increased significantly from 0.491 to 0.514. These results validate the effectiveness of the proposed approach in improving small-target detection performance.

**Keywords:** YOLOv8; small object detection; high-resolution detection layer; Mosaic data augmentation; remote sensing imagery.

## 1. Indroduction

With the rapid development of remote sensing technology, remote sensing imagery has been widely used in various fields such as Geographic Information Systems (GIS), environmental monitoring, urban planning, and agricultural management. The high-resolution characteristics of remote sensing imagery make identifying small targets in different scenarios an important task, especially in remote sensing images, where these small targets such as buildings, vehicles, and ships usually occupy only a small number of pixels, posing great challenges to target detection. Traditional object detection methods perform poorly on these small targets and struggle to detect accurately; therefore, improving small target detection capability has become a key research direction in the field of remote sensing image analysis.

Object detection technology, as one of the core tasks in the field of computer vision, has undergone long-term development. From the early region-based methods, such as Selective Search, to the later regression-based end-to-end approaches, such as YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector), object detection methods have continuously improved in accuracy and efficiency. Among them, the YOLO series algorithms have gained broad adoption since their release for their efficiency and real-time performance, especially achieving significant progress in detection speed. However, although models like YOLOv8 perform well in large-object detection, they still face many challenges when handling small objects.

In recent years, with the development of convolutional neural networks (CNN), the YOLO series algorithms have been continually optimized, achieving some improvements in small object detection. YOLOv4 introduces CSPNet (Cross-Stage Partial Network) and SPP (Spatial Pyramid Pooling). YOLOv5 enhances small object detection capabilities through PANet (Path Aggregation Network) and multi-scale feature fusion techniques. YOLOv8 further optimizes its architecture on the basis of previous generations, achieving progress particularly in small object detection. By introducing deeper

network structures and more refined feature fusion methods, YOLOv8 finds a better balance between detection accuracy and speed.

However, despite these advances, existing object detection techniques often face the following challenges when handling small objects: insufficient feature extraction capability: detection models like YOLOv8, while effective for large objects, are weak at capturing fine details of small objects and often overlook them due to lower resolution; poor scale invariance: small objects in remote sensing images are usually set against complex backgrounds and have large scale differences relative to the background, making it difficult for traditional methods to adapt effectively to objects at various scales; insufficient data: although remote sensing datasets are increasing, annotated data remain scarce, causing models to suffer from overfitting and poor generalization during training.

To address shortcomings of existing algorithms, researchers have proposed various improvements: Yang Hongdan et al. [1] proposed an improved small-object detection method based on YOLOv8 to enhance aerial image small-object detection performance. By introducing a lightweight L-MobileViT module to strengthen feature extraction, designing a hierarchical multi-scale fusion module (HF) to combine shallow and deep features, adding a small-object detection head and removing the large-object detection head, they effectively improved detection accuracy and reduced missed detections. Han Songchen et al. [2] proposed an improved Faster-RCNN algorithm to solve low small-object recognition accuracy in airport video surveillance. They use ResNet-101 as the feature extractor, add a sampled “topdown” feature fusion module to generate high-resolution semantic feature maps, and combine an online hard example mining strategy to enhance model robustness and improve detection accuracy. Han Xingbo et al. [3] adapted the YOLOv5 model for remote-sensing small-object detection by adding residual connections and detection heads to improve small-object feature extraction and introducing cross-layer attention to inject semantic information and strengthen suppression of complex backgrounds, addressing the problems of few pixels and difficulty in detecting small objects. Wei Mingjun et al. [4] proposed a multi-scale feature fusion network method to tackle the scarcity of discriminative features and background interference in remote-sensing small objects: they designed a sparse-attention-guided feature fusion module on mid-scale feature maps to increase small-object sensitivity and suppress background; and built a multi-step dilated feature fusion module using dilated convolutions to fuse multi-scale context and improve localization and detection accuracy. Shi Shihao et al. [5], targeting small-object detection in low-quality complex scenes, proposed a “hybrid features + multi-scale fusion” algorithm: they designed an HCTB module combining convolution and self-attention to enhance feature representation; expanded the receptive field with multi-dilation atrous convolutions; and constructed an OKCSM module to fuse multi-scale features while preserving detail, improving detection robustness in complex scenes.

Although the aforementioned studies have made continuous advances in convolutional improvements, attention integration, multi-scale fusion, and loss function optimization, significantly boosting detection performance in general scenarios, they are still constrained in small object detection by dense distributions, ultra-small scales, and complex background interference, manifested as target granularity features being easily lost and higher miss detection rate, mAP/Recall declines in remote sensing scenes. The main objective of this paper is to improve small object detection performance based on YOLOv8, especially in remote sensing images. To this end, this paper proposes two innovative improvements, with the main contributions as follows:

(1) Add High-Resolution Detection Layer (P2): Add a high-resolution detection layer to the Neck of YOLOv8 to enhance the model's feature extraction for small objects. This layer can effectively capture object features at different scales, thereby improving small object detection accuracy.

(2) Introduce Mosaic Data Augmentation: By stitching multiple images into one image using the Mosaic data augmentation method, increase data diversity and improve the model's robustness to small objects. This method effectively expands the training dataset, enabling the model to learn small-object features better during training.

## 2. Experimental Algorithm

### 2.1. Basic Architecture

The object detection model proposed in this paper is based on YOLOv8, with appropriate modifications and improvements to its original architecture to enhance the model's performance in small object detection. As a regression-based object detection model, YOLOv8 adopts an end-to-end architecture consisting of three modules: Backbone, Neck, and Head, as shown in Figure 1. These three modules are responsible for feature extraction, feature fusion, and the final object detection output, respectively.

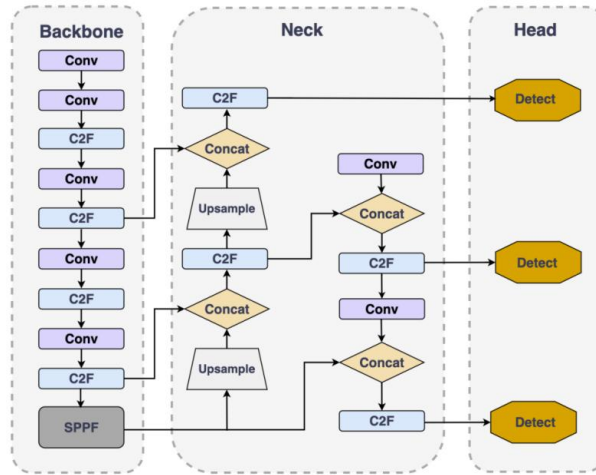


**Figure 1:** YOLOv8 Structure Diagram

Backbone (Backbone Network):

Backbone: The core of YOLOv8, responsible for extracting features at different levels from the input image. YOLOv8 uses improved CSPNet (Cross-Stage Partial Network) and GhostNet, two efficient feature extraction methods. CSPNet optimizes information flow and reduces redundant computation through partial feature sharing, improving the model's computational efficiency. GhostNet further reduces the model's computation burden by introducing lightweight convolution operations, while also enhancing feature extraction capability. The backbone's output feature maps are used as input for subsequent level processing.

Neck:

Neck section is mainly used for multi-scale feature fusion, solving issues in multi-scale object detection, to enhance the model's perception of objects at different sizes. YOLOv8 introduces FPN (Feature Pyramid Network) and PANet (Path Aggregation Network) modules in the Neck section, which can effectively fuse features from different levels, making the model's performance in detecting objects at various scales more balanced.

Head (Head):

Head section is responsible for generating the final object detection results, including object category, bounding box coordinates, and confidence. In YOLOv8, the Head is optimized, adopting the CIoU (Complete Intersection over Union) loss function, and combined with an adaptive feature fusion module to further improve small object detection accuracy. The CIoU loss function not only considers the overlap area of the targets but also introduces optimization of the target center distance and width-height ratio, making the bounding box predictions for small objects more precise. In YOLOv8, the Head module directly outputs the predicted boxes and computes the corresponding loss.

### 2.2. Innovations and Improvements

Small targets are typically small in size, have sparse feature information, and are easily disturbed by the background—although the original YOLOv8 has strong overall performance, it still suffers from missed detections or inaccurate localization when detecting small targets. To address this

challenge, this paper makes three key improvements on the basis of YOLOv8, aiming to enhance the model's practicality and reliability in real scenarios such as remote sensing and high-resolution images.

### **2.2.1. High-Resolution Detection Layer (P2)**

Standard multi-scale object detection networks (such as those using FPN) typically consider multi-scale object detection by feature maps with different downsampling ratios, such as P3, P4, P5. Among them, the high-level P5 has rich semantics but low spatial resolution, suitable for large objects; the middle levels P3/P4 are a compromise for large/medium object detection.

In remote sensing images and other scenarios, small-object detection suffers when deep feature maps are downsampled multiple times, drastically reducing spatial resolution, causing small targets to correspond to very few pixels or be submerged by the background. The characteristic of deep features as “high in semantics, weak in spatial details” makes it difficult for models to accurately locate and classify small targets. Therefore, related studies generally propose preserving or recovering high spatial-resolution feature maps to enhance small-target perception capabilities.

Based on this, introducing the P2 high-resolution detection layer in the YOLOv8 model becomes an effective improvement. P2 is the shallower level in the feature pyramid, with a small downsampling ratio and high spatial resolution, allowing it to retain more fine-grained spatial information and texture details compared to P3-P5.

Implementation level, it is necessary to include P2 in the detection head, so that the model can process large targets using P3-P5 while using P2 to perform small-target classification and bounding box regression, ensuring that small targets can obtain accurate feature representations and detection results even when their pixel proportion in the original image is low, by leveraging high-resolution features. Structurally, the model neck should retain or add a path from the backbone's shallow layers (or via upsampling fusion) to P2, so that P2 simultaneously carries shallow details and semantic information, meeting the needs of the detection head.

### **2.2.2. Mosaic Data Augmentation**

In small object detection tasks, training sets for remote sensing images in dense or complex background scenarios have clear shortcomings: first, the number of small objects is limited and their distribution sparse, which easily causes the model to overfit the background and neglect the objects; second, small object scales are too small and textures easily confuse with the background, making single augmentation methods such as flipping or cropping insufficient to improve the model's generalization and robustness.

To address the above issues, Mosaic augmentation is introduced to enhance the model's ability to adapt to multi-background, multi-scale, and densely packed targets. Its core idea is to randomly select four training images, apply random cropping, scaling, color brightness flipping, and other transformations, then stitch them into a single synthetic image as a new sample, which can fuse multi-scale targets and diverse backgrounds from different original images.

This augmentation is commonly concatenated with  $2 \times 2$ ; the operation process is clear: first randomly select 4 images, then for each image perform scaling, cropping, and transformation, then stitch the four images into the four quadrants of a composite image and finally synchronize the transformed target bounding box coordinates to ensure precise alignment between labels and images. During training, enable it at the data loading stage with a set probability, with probability 1.0 meaning all samples are augmented.

Mosaic enhances small object detection with significant advantages: first, it enriches training data diversity, enabling the model to learn object combinations across multiple backgrounds, scales, and densities, improving adaptability to complex scenes; second, it increases the exposure of small objects, allowing tiny targets to appear at appropriate scales, boosting detection sensitivity and recall; third, it effectively mitigates overfitting, the rich sample combinations help the model avoid dependence on a single scene, enhancing generalization to unknown images.

### 2.2.3. Loss Function

In models already integrated with the P2 high-resolution detection head and Mosaic data augmentation, introducing WIoU to replace CIoU regression loss can further improve mAP\_small, recall, and F1 score for small object detection, while also making training more stable and reducing sensitivity to outliers.

Traditional IoU-series losses homogenize gradients for all positive samples; extremely good samples tend to be over-optimized, extremely poor samples introduce training noise, and the most promising medium-quality samples are often ignored. The core of WIoU is to introduce dynamic non-monotonic focal weight. The core formula is  $WIoU = \text{baseline IoU-type regression term} \times \text{dynamic focal weight}$ , where the baseline term can be CIoU, EIoU, etc. The focal weight gives the highest weight to medium-quality boxes, while reducing weights for extremely good or extremely poor samples. The training process is first using a positive/negative sample allocator to match predicted boxes with ground-truth boxes, computing their geometric quality, then map to a non-monotonic Focal weight, where low samples have small weights to reduce interference, medium samples have high weights to emphasize optimization, and high samples have reduced weights to avoid overfitting. Next, construct the regression loss and accumulate with the original ratios for classification, DFL loss, etc.; the core idea is to tilt the gradient budget toward medium-quality samples.

Compared with CIoU which focuses only on geometric metrics and has homogeneous gradient weights, WIoU introduces dynamic weighting based on sample quality, achieving suppression of the two ends and elevation of the middle, thus improving gradient efficiency and stability. This loss adapts to small object detection, can focus on small targets in the middle-lower IoU range, smooths gradients to reduce training oscillations, and can resist common annotation errors and scale mismatches for remote sensing small objects.

When landing, it is recommended to choose CIoU or EIoU for the base term, use default hyperparameters, maintain the original loss ratio, and form a complement with the P2 detection head and Mosaic augmentation. Note that when data is clean, targets are relatively large, or baseline performance is saturated, WIoU improvements are limited; for extremely sparse tasks, adjust weights to avoid underfitting.

## 3. Experiments and Analysis

### 3.1. Experimental Setup

The experiment is based on the PyTorch 1.9.0 framework, with input image resolution of 640×640 pixels, 100 training epochs, a batch size of 16, using the default YOLOv8 SGD optimizer (paired with cosine annealing learning rate scheduling). Hyperparameters such as learning rate, momentum, and weight decay follow YOLOv8 default settings. The experimental environment configuration is shown in Table 1.

**Table 1:** Experimental Environmental Configuration.

| Setup   | Version parameter             |
|---------|-------------------------------|
| OS      | Window11 Home                 |
| GPU     | NVIDIA GeForce RTX (8GB)      |
| CPU     | 13th Gen Intel Core i7-13700H |
| RAM     | 48 GB RAM                     |
| ROM     | 1 TB SSD                      |
| CUDA    | CUDA 11.2                     |
| PyTorch | PyTorch 1.9.0                 |

### 3.2. Experimental Dataset

This experiment uses the NWPU VHR-10 high-resolution remote sensing object detection public dataset for training and validation. The dataset contains about 800 images, with about 650 positive

images containing targets and about 150 negative images without targets, covering roughly 3.6k target instances in total.

Its scenarios cover complex areas such as ports, airports, urban built-up areas, and sports venues, with real-world interferences including illumination changes, viewpoint differences, and target occlusions. Targets like ships, vehicles, and storage tanks are often densely distributed in small sizes, with a clear long-tail size distribution. The categories retain the original 10-class definition (including airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground trackfield, harbour, bridge, vehicle); in experiments, they are mapped to a continuous index [6].

### 3.3. Evaluation metrics

#### 3.3.1. Basic detection performance metrics

To evaluate the performance of the improved YOLOv8 algorithm, the experiments set metrics from three aspects: detection accuracy, small-object performance, and deployment efficiency.

Basic performance indicators: evaluated using precision, recall, and F1 score, where precision reflects detection accuracy control, recall reflects miss detection control, and F1 score balances the two performances.

Comprehensive detection accuracy indicators: using an average precision framework, including [AP@0.5](#) (area under the PR curve for a single class at IoU=50%), mAP@0.5 (mean AP@0.5 over all classes), mAP@0.5:0.95 (mean AP over multiple IoU thresholds from 0.50 to 0.95 with a step of 0.05);

Small Object Specific Metrics: Defined as small objects by COCO standard (short side of object box in original image < 32px or area < 32), and evaluate improvement with small-object precision, recall, F1-score, and AP.2px2

Deployment Efficiency Metrics: Use FPS (frames per second that the model processes) to evaluate realtime performance, and use parameter count and computational cost (based on input) to evaluate storage and compute overhead.640 × 640

The calculation formula is as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1 - score} = \frac{2 \times P \times R}{P+R} \quad (3)$$

Where TP is the number of correctly predicted positive samples, FP is the number of incorrectly predicted positive samples, and FN is the number of undetected positive samples.

### 3.4. Comparative Experiments

To verify the effectiveness of the proposed improvement method in small target detection for remote sensing images, this paper conducts a comparative experiment with the original YOLOv8 model. Specifically, this paper optimizes the YOLOv8 model through the following three key improvements:

1. Add high-resolution detection layer P2: Introduced an extra high-resolution detection layer P2 in the Neck of YOLOv8, enhancing the model's feature extraction for small objects, especially improving detection accuracy for small objects in high-resolution images.

2. Introduce Mosaic data augmentation: By stitching multiple images into one image, it expands training data diversity, increases robustness to small objects, and reduces missed detections caused by data imbalance or scarcity of small objects.

3. Improved loss function (WIoU): Using Weighted IoU (WIoU) loss instead of the original CIoU loss, it better measures the difference between small object bounding boxes and ground truth boxes, thereby improving small object localization accuracy.

### 3.5. Results analysis:

Accuracy  $P=0.545$ : Although the improved YOLOv8 model introduces high-resolution detection layers and multi-scale data augmentation, the accuracy decreases by 9.17% compared to the original model due to increased complexity of small-object samples.

Recall rate  $R=0.856$ : With Mosaic data augmentation to expand small-object samples and improved loss function for gradient allocation, the model's ability to detect small objects in complex backgrounds is significantly enhanced, with recall increased by 9.18% compared to the original model.

F1-score= $0.666$ : Considering precision and recall together, the F1-score is 1.65% higher than the original model, confirming that the improved model's overall performance on small-object detection tasks is effectively strengthened.

FPS= $33.28$ : Despite adding higher-resolution detection layers and using a complex loss function, the improved model still maintains a high inference speed, with only a small difference in FPS compared to the original model, demonstrating a good balance between detection performance and running efficiency in the improvement scheme.

## 4. Conclusions

Comparative experimental results show that the three proposed improvements in this paper—P2 high-resolution detection layer, Mosaic data augmentation, and WIoU loss function—significantly enhance the YOLOv8 model's overall performance in small target detection tasks for remote sensing images. They not only achieve improvements in core detection metrics such as precision, recall, and F1-score, but also maintain inference speed without a substantial drop, achieving a balance between detection performance and running efficiency, thereby fully validating the feasibility and superiority of the proposed improvements for remote sensing small-target detection tasks.

In this paper, we propose an improved small object detection method based on YOLOv8 to address the difficulties of detecting small targets in remote sensing images, with two key innovative improvements. First, by adding a high-resolution detection layer P2 in the Neck part of YOLOv8, the model's feature extraction capability for small targets is enhanced, improving the detection accuracy for small targets. Second, Mosaic data augmentation technology is introduced, which stitches multiple images to expand the diversity of training data, thereby improving the model's robustness and generalization to small targets.

Experimental results show that the improved YOLOv8 model significantly enhances detection performance on the NWPU VHR-10 dataset. Specifically, the model achieves increases of 9.18% in recall, with also favorable improvements in F1-score and FPS. The above results validate the effectiveness of the proposed method in boosting small object detection capabilities and provide new ideas and approaches for the development of future small object detection technologies.

However, this study still has some limitations, and future work can improve in the following directions. First, although a high-resolution detection layer and Mosaic data augmentation technology are introduced, there is still room for improvement in detecting extremely small targets or in extreme background environments. Second, further optimizing the model's real-time performance and computational efficiency, and achieving more efficient real-time detection by combining edge computing and embedded devices, is another focus for future development.

## References

- [1] Yang Hongdan, Fu Gui, Shao Huichao, et al. Aerial small object detection with multiscale hierarchical features [J]. Online publication date: 2024-11-07.
- [2] Han Songchen, Zhang Bihao, Li Wei, et al. An airport scene small-object detection algorithm based on an improved Faster-RCNN [J]. Online publication date: 2020-02-04. DOI:10.16356/j.1005-2615.2019.06.001.

- [3] Han Xingbo, Li Fan. Remote sensing small-object detection based on cross-layer attention enhancement [J]. Online publication date: 2023-07-26.
- [4] Wei Mingjun, Ge Yihun, Yang Xuan, et al. Remote sensing image small-object detection based on multi-scale feature fusion [J]. Online publication date: 2024-07-00. DOI: 10.16186/j.cnki.1673-9787.2024070040.
- [5] Shi Shehao, Shi Qunshan, Zhou Yang, et al. Optical small target detection algorithm with mixed features and multi-scale fusion [J]. Online publication time: 2025-07-10.
- [6] Author bios: Lin Yida (2005-), male, from Shangqiu, Henan, undergraduate, no title, research interests: machine vision.