International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 7 (July 2025), PP 66-72

Litter Detection and Recognition for UAV Aerial Photography Based on ImprovedYOLOv8

Peng Xiao[1], Zhan Wen[1, 2*], Meiqin Wu[1], Wenzao Li[1,2]
[1] *School of Communication Engineering, Chengdu University of Information Technology, Chengdu, 610225, Sichuan, China .*
[2] *Meteorological information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology, Chengdu, 610225, Sichuan, China .*
**Corresponding author*

ABSTRACT
Unmanned Aerial Vehicle(UAV) aerial photography has a broad application prospect in environmental monitoring and other fields. Still, its small target detection faces challenges such as geometric deformation, missing features, and dense distribution. Aiming at the shortcomings of YOLOv8 in small target detection, this paper proposes an improvement scheme: introduce the efficient channel attention (ECA) mechanism in the backbone network to enhance the ability to capture fine features, and add a new 160×160 high-resolution detector head to adapt to the high-resolution feature extraction for small targets, meanwhile, removing the original large-target detector head. Experimental results show that the improved model mAP50 increases from 62.2% to 65.4% and mAP50-95 increases from 33.3% to 36.8% of the original network of YOLOv8n, which effectively improves the detection accuracy of small targets in UAV aerial photography scenarios through the synergistic optimization of the attention mechanism and the multi-scale detector head and provides a highly efficient solution to small target-intensive tasks such as remote sensing and aerial photography. It offers an efficient solution for small target-intensive tasks such as remote sensing and aerial photography.
Keywords: *Unmanned Aerial Vehicle; YOLOv8; Small Target Detection; Attention Mechanisms*

Date of Submission: 08-07-2025 Date of acceptance: 20-07-2025

I. INTRODUCTION

In recent years, object detection technology in computer vision [1] has undergone a revolutionary development process, transitioning from traditional methods to deep learning-driven approaches. Early object detection algorithms based on manually designed features, such as the HOG+SVM combination, relied on artificially designed feature extraction rules, resulting in limited detection accuracy and generalization capabilities in complex scenarios. With the rise of deep learning, the R-CNN series of algorithms pioneered a new era of end-to-end object detection by introducing convolutional neural networks (CNNs). Subsequently, Fast R-CNN and Faster R-CNN continuously optimized the region proposal network, significantly improving detection efficiency; The emergence of single-stage detection algorithms such as the YOLO (You Only Look Once) series and SSD (Single Shot MultiBox Detector) has pushed real-time performance and accuracy to new heights, making object detection mature in both theoretical research and engineering applications.

The rise of unmanned aerial vehicle (UAV) technology [2] has injected new vitality into the field of object detection. Compared to traditional ground-based perspectives, the visual systems mounted on UAVs can overcome physical obstacles such as terrain and buildings, providing a unique bird's-eye view to cover large areas and enabling object detection in complex scenarios. At the hardware level, UAVs are equipped with high-resolution optical cameras, infrared thermal imagers, and multispectral sensors, enabling the collection of multimodal data under various lighting and weather conditions; At the algorithm level, by deeply integrating deep learning algorithms with UAV systems and utilizing lightweight network structures and model compression techniques, UAVs can efficiently parse the semantic information of ground objects and accurately locate them on edge devices. For example, in agricultural scenarios, UAVs can quickly identify the growth status of individual crops; in urban traffic monitoring, they can capture the dynamics of vehicles and pedestrians on roads.

This technological integration not only expands the application dimensions of object detection but also demonstrates broad prospects across multiple fields. In environmental monitoring [3], drones can dynamically monitor river pollutants, forest fire hazards, and other issues, promptly identifying minor pollution sources or fire sources; in emergency disaster response, leveraging their mobility advantages, they can swiftly screen for signs of life in earthquake rubble, providing precise coordinates for rescue operations; In the field of intelligent

transportation, drone aerial photography can assist traffic management departments in obtaining a comprehensive view of road conditions and identifying issues such as illegal parking and traffic congestion; in precision agriculture, through multispectral image analysis, drones can accurately detect crop pests and diseases, as well as soil fertility distribution, to support variable-rate fertilization and precision irrigation. The deep integration of drones and object detection technology is driving industries toward greater intelligence and efficiency.

Although UAVs have significant advantages in target detection, such as wide coverage and strong mobility, their technology still faces multiple challenges. UAV target detection technology faces multiple challenges: geometric variation of air view and resolution decay lead to missing target features, which affects model feature extraction; complex weather and lighting interference make it difficult to balance detection accuracy and robustness; multi-target occlusion, cross-scale identification, and multi-task resource allocation conflicts constitute technical pain points, and subtle differences between similar targets put forward higher requirements for algorithmic feature resolution capability. To address the technical bottlenecks, research is shifting from traditional manual feature detection to deep learning methods based on convolutional neural networks (CNN). Deep learning solves the problem of real-time and environment adaptability through end-to-end training and has significant advantages in complex scene feature characterization, cross-scale target detection, and multi-tasking, and its generalization ability and customized architecture become the core solution to balance detection accuracy and real-time performance, which promotes the technology to move forward to engineering applications.

The development of convolution-based deep learning algorithms has gone through key stages. Hinton's Deep Belief Network (DBN) realizes multi-layer feature extraction through unsupervised pre-training, but its complex Boltzmann machine stacking structure leads to high computational costs, limiting its practical application. 2012 saw the beginning of the CNN era with AlexNet, and VGGNet followed by small convolutional kernel stacking, In 2012, AlexNet started the CNN era, followed by VGGNet through small convolutional kernel stacking, GoogleNet introduced multi-scale feature fusion, and ResNet using residual connection to break through the bottleneck of gradient disappearance, which pushed CNN to reach a new height in image classification. CNN's mechanism of local connection and weight sharing reduces the number of parameters, but there are limitations in dealing with sequential data, and although LSTM alleviates the gradient problem of traditional RNNs through the gating mechanism, the computational efficiency of LSTM can hardly meet the real-time demand.

The YOLO (You Only Look Once) algorithm proposed by Joseph Redmon in 2015 created a new paradigm of single-stage target detection, unifying target localization and classification tasks in a single end-to-end network for the first time. YOLO's lightweight network design, highly efficient feature extraction strategy, and end-to-end training mode show its unique advantages in real-time target detection of UAVs and other scenarios that are sensitive to computational resources and promote the rapid advancement of target detection technology from theoretical research to engineering implementation. The YOLO series of target detection algorithms have matured, but there are still significant shortcomings in the field of small target detection, especially in the scene of small size and high-density distribution of objects under the viewpoint of UAVs. To solve this problem, this paper presents a new network structure based on YOLOv8n with modifications. The main contributions of this paper are as follows:

（1）To enhance the capture of subtle features, a layer of improved channel attention mechanism ECA is introduced in the backbone network.

（2）The network adds high-resolution detectors suitable for small targets, enabling more accurate detection of small aerial targets.

II. RELATED WORKS

The YOLO series of object detection algorithms, with their efficient detection speed and excellent overall performance, have matured in the field of general object detection and are widely used in scenarios such as security surveillance and autonomous driving. However, in the field of small object detection, the YOLO series of algorithms still exhibit significant shortcomings, particularly when viewed from a drone's perspective, where their limitations become even more pronounced. During drone aerial photography, due to the high flight altitude, objects in the frame appear smaller in size. In scenarios such as urban buildings and farmland crops, small objects often exhibit dense distribution patterns. On one hand, the downsampling operation in YOLO algorithms compresses image dimensions, leading to further loss of small object features and making them difficult to effectively capture. On the other hand, its grid-based prediction mechanism lacks sufficient localization accuracy for small objects, often resulting in missed detections or false positives. Additionally, the YOLO algorithm does not sufficiently learn the features of small targets during training, resulting in a significant decline in recognition ability when faced with small targets in complex backgrounds.

To address this issue, researchers have explored improvement strategies from multiple dimensions. In terms of network structure optimization, some studies have introduced Feature Pyramid Networks (FPN) or Path Aggregation Networks (PANet) to establish multi-scale feature fusion mechanisms, combining the high-resolution detail information from shallow layers with the semantic information from deep layers to enhance the model's ability to extract small object features. Other studies have attempted to embed attention mechanisms, such as the Spatial Attention Module (SAM) and Channel Attention Module (CAM), into the backbone network to focus the network on small target regions and improve sensitivity to subtle features. In terms of data augmentation techniques, researchers have employed methods such as super-resolution reconstruction and mosaic data augmentation to artificially increase the proportion of small targets in images, enrich the diversity of training data, and help the model learn more small target features. Furthermore, model fusion is another important direction for improvement. By integrating the YOLO algorithm with other algorithms that excel at small object detection, such as integrating lightweight small object detection subnetworks into the YOLO framework, researchers can leverage the strengths of different algorithms to compensate for YOLO's shortcomings in small object detection. Additionally, for drone aerial photography scenarios, some research teams have specifically constructed datasets containing a large number of small object samples, and through targeted training, optimized the model's detection performance in such scenarios.

At the level of feature representation optimization, in the field of feature hierarchical representation, the Dense Similar Object Detector (DSOD) proposed in the literature [4] densely fuses shallow detail features (e.g., edges, textures) with deeper semantic features by constructing cross-layer feature interaction networks. Specifically, the model introduces dense jump connections in the feature pyramid so that the feature maps at each level can receive detailed information from all previous layers, thus constructing a feature pyramid containing multi-granularity semantics. Experiments show that this method improves the detection accuracy by 12.3% on the UAV remote sensing small target dataset, but due to the proliferation of the number of parameters caused by the dense connections, it exposes the tuning complexity and efficiency bottleneck caused by multi-hyper-parameter design. Lin et al. [5] proposed the feature pyramid network (FPN), which achieves multi-scale feature expression optimization with only an 8% increase in computation through the bottom-up path augmentation and horizontal connection mechanism. The core of FPN lies in the use of up-sampling operations to optimize multi-scale feature expression. The core of the FPN is to utilize the up-sampling operation to recover the spatial details of shallow high-resolution features, and at the same time inject the deep semantic information through the lateral connection to form a pyramid structure with both details and semantics. This design improves the insulator defect detection accuracy by 18.7% in a small target detection scenario of UAV power inspection and keeps the inference speed in real time, which provides a classical paradigm for balancing feature characterization and efficiency. Yang et al [6] proposed the MA-ResNet model which innovatively embeds the hybrid attention mechanism into the residual network. Specifically, Spatial Attention and Channel Attention are introduced in the residual block in parallel, with the former capturing the target spatial distribution features through convolutional operation and the latter refining the inter-channel dependencies through global average pooling. Experiments show that this structure improves the detection accuracy of small targets with pixel size < 20×20 by 21.5% in UAV farmland pest detection, and significantly enhances the target-background feature differentiation ability, especially in dense crop backgrounds.

In the direction of feature fusion technology, literature [7] proposes a bi-directional feature fusion module (BFFM), which enhances the feature representation of small targets in complex scenes by constructing a multi-scale feature pyramid, however, this method requires high computational resources and hardware performance. Lu et al. [8] propose an end-to-end feature fusion scheme based on the SSD network, which is focused on the semantic enhancement of shallow features. A feature enhancement module is added to address the shortcoming that the shallow layer of the SSD model contains only low-level visual features, and the potential semantics of the shallow features are refined through convolutional operations. The enhanced shallow features are fused with the deep features by cross-layer splicing: firstly, the deep features are up-sampled to the shallow resolution by bilinear interpolation, then they are spliced with the enhanced shallow features in the channel dimension, and finally, the fused feature map is generated by 1×1 convolution. On the UAV farmland pest detection dataset, this scheme improves the pest target detection accuracy from 55.3% to 68.9% of the original SSD model, but due to the addition of two new convolutional layers and cross-layer splicing operations, the number of model parameters increases from 23.2M to 37.5M, and the amount of computation (GFLOPs) improves by 51%, which results in the frame rate of the UAV detection system with MobileNetv3-SSD decreasing from 45FPS to 1FPS. The frame rate of the UAV detection system equipped with MobileNetv3-SSD decreases from 45FPS to 31FPS, which makes it difficult to meet the real-time detection requirements in dynamic flight scenarios.

In terms of detection structure innovation, literature [9] designed an enhanced sensing field module based on YOLOv7, which effectively improves the detection of small and overlapping targets; Lou et al. [10] proposed dc-YOLOv8, which retains more contextual information by improving the down sampling mechanism;

Zhang et al. [11] introduced a recursive mechanism into the YOLOv5 framework to form a MEU-YOLOv5, which reduces the problem of missed detection of small targets. YOLOv5, which reduces the problem of missed detection of small targets. While these improvements improve the detection accuracy, they generally face the challenge of balancing the computational efficiency and model complexity, which provides diversified technical paths for small target detection in complex UAV scenarios.

III. MATERIAL AND METHODS

3.1 Dataset

The dataset used in this experiment is a publicly available garbage detection and recognition dataset, with data collected from a drone perspective. The dataset contains 772 aerial images taken at different heights and 3,718 COCO-style annotations. The data was collected in a variety of different scenarios.

In the experiments of this paper no data augmentation is performed on the data and the original dataset is directly used for training and evaluation. The dataset is divided into training, validation, and test sets in the ratio of 7:2:1.

3.2 Model Structure

Network architecture diagram

YOLOv8, the most popular algorithm in the YOLO family today, has shown significant advantages in the field of object detection, including excellent performance in detecting small objects from the perspective of UAV aerial photography. In this paper, we improve on the YOLOv8n network. The improved network structure is shown in Fig 1. Improvements are outlined with dotted lines.

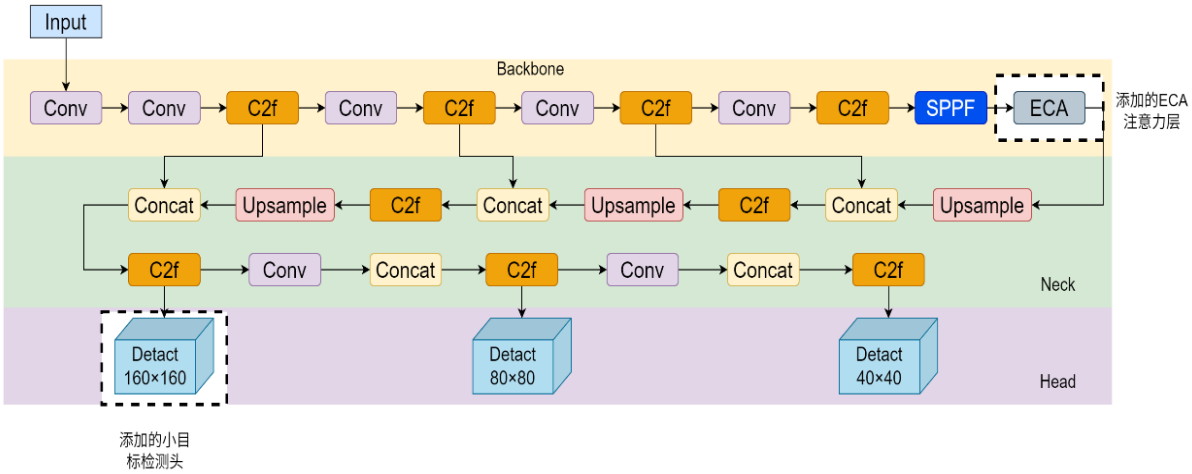

Fig 1. Improved YOLOv8 network structure

Firstly, introduce ECA attention. The core of the ECA mechanism lies in adaptively learning the importance of each channel to precisely focus on the key feature information of the detection target. In the backbone network of traditional YOLO series algorithms, the contribution of each channel to feature extraction is considered equal, making it difficult to highlight effective features and ignore background noise interference when processing small targets. The ECA mechanism avoids complex global information calculations by considering local interaction information between channels, enabling dynamic adjustment of channel weights through a lightweight structure. Specifically, it uses fast one-dimensional convolutions to compute channel attention, allowing the network to quickly capture subtle features that are easily obscured by the background in drone aerial photography scenarios due to their small size, significantly enhancing the model's sensitivity to target features.

Secondly, the 20×20 detection head was removed and replaced with a 160×160 resolution detection head. Since small objects occupy a low proportion of pixels in drone aerial images, traditional detection heads are unable to retain sufficient detail information for small object recognition after multiple layers of downsampling. The introduction of high-resolution detection heads can retain more original image details and support feature extraction at high resolution, thereby accurately capturing the subtle features of small objects.

ECA Attention Mechanism

The ECA attention module avoids dimension reduction and efficiently implements local cross-channel interactions through 1D convolution, thereby extracting cross-channel dependencies. The ECA attention module diagram is shown in Fig 2.

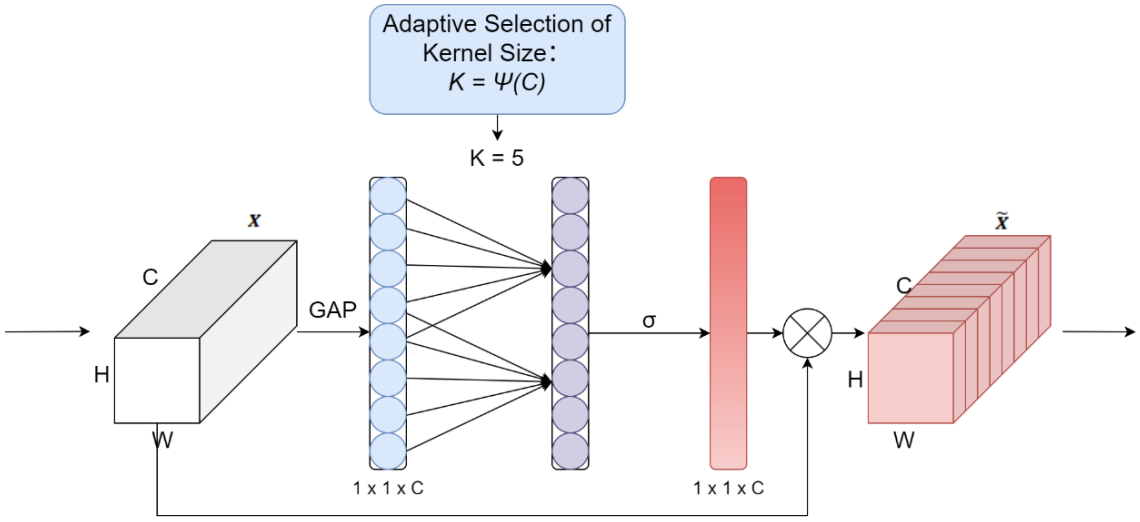

Fig 2. ECA Attention Structure Diagram

The specific steps of the ECA attention mechanism are as follows: (1) Perform global average pooling operation on the input feature map; (2) Perform 1-dimensional convolution operation with a convolution kernel size of k, and obtain the weight w of each channel through the Sigmoid activation function; (3) Multiply the weights with the corresponding elements of the original input feature map, and then adjust the importance of each channel to obtain the final output feature map. It can be seen that the idea and operation of the ECA attention mechanism are extremely simple, and the impact on the processing speed of the network is small.

Improving small target detection

In the structural adjustment of the model neck, the focus is on enhancing the detection capability of small targets. The original YOLOv8 model is equipped with three sizes of detection heads, corresponding to feature map sizes of 20×20, 40×40, and 80×80, respectively, to realize the detection of different sizes of targets through the multi-scale feature mapping mechanism. The feature maps can capture different levels of image semantic information due to the difference in sensory fields. Given the characteristics that the UAV perspective images are mainly small targets and large targets are scarce, a new 160×160 resolution small target detection head is added, and the original 20×20 large target detection head is removed.

The new small target detection head has two major advantages: first, the high-resolution feature map retains richer image detail information, which provides the basis for fine feature extraction of small targets; second, the smaller sensing field allows the network to focus on local detail features, avoiding the dilution of small target features due to the large sensing range. The detection head is deployed in the shallow region of the network, which can rely on the high-resolution feature advantage of the shallow network to realize the rapid localization and accurate identification of small targets. Through the above structural adjustment, the leakage and false detection of small targets can be effectively reduced, significantly improving the target detection performance in complex scenes.

IV. EXPERIMENTSAND DISCUSSION

4.1 Experimental Environment

Table 1 shows the experimental equipment, experimental environment, and model parameters for this paper, using the SGD optimizer to train the network.

Table 1. Experimental Setting

Parameter	Setting
GPU	NVIDIA GeForce RTX 4060 Ti 8 GB
PyTorch	2.5.0
Python	3.11
Epochs	60
Batch Size	16

4.2 Evaluation Metrics

Mean Average Precision (*mAP*) has been chosen to evaluate the network performance.*mAP* can be calculated using the following formula:

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

$$mAP = \frac{1}{N}\sum_{n \in N} AP(n) \tag{2}$$

The *P(R)* curve provides a visual representation of precision performance at different recall rates. Average Precision (AP) provides insight into the overall performance of the model on individual categories. On the other hand, *mAP* is a global performance metric for comprehensively evaluating the model's performance across all categories.

4.3 Experiment

Fig 3 shows the training results of the improved YOLOv8n on the dataset. The rapid climb in early training (first 20 epochs or so) can be seen in the mAP50(B) curve and the mAP50-95(B) curve, indicating that the model learns to "roughly find the target" very quickly and that the base detection capability is built up quickly.

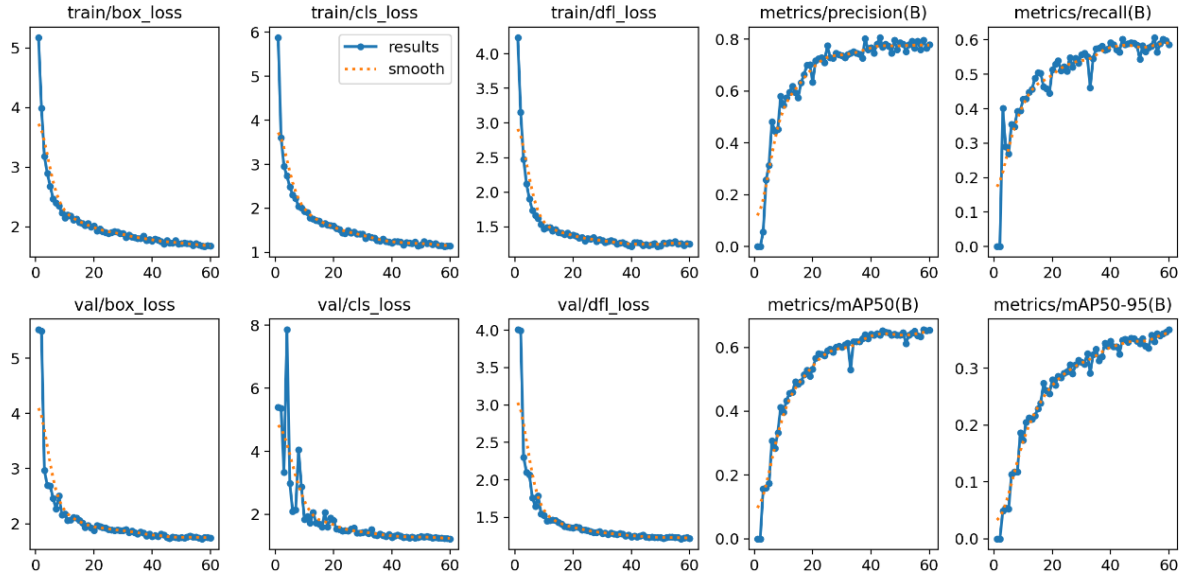

Fig 3. Improve the model training curve

Experiments have shown that the best model achieves a mAP50(B) value of 65.4% on the dataset compared to 62.2% for the original YOLOv8n network, a 3.2% improvement, and a mAP50-95(B) value of 36.8% compared to 33.3% for the original network, a 3.5% improvement. These data illustrate the effectiveness of introducing the ECA attention mechanism and increasing the small-target detection head. Through the synergistic optimization of the ECA attention mechanism and the small-target detection head, the improved version of the YOLOv8n achieves a "two-dimensional performance breakthrough": it strengthens the basic recognition ability of small targets (mAP50 enhancement), and overcomes the high accuracy detection technical bottleneck (mAP50-95 enhancement). This improvement not only verifies the suitability of the attention mechanism and multi-scale detection head for small target tasks but also provides a more efficient and accurate technical solution for target detection in aerial photography, remote sensing, and other "small target intensive" scenes.

Fig 4 shows the P-R plot of the best model, the model is in the low recall stage (Recall ≈ 0~0.6): the Precision maintains a high level of 0.8~1.0 indicating that the model can strictly filter false alarms and almost no false positive prediction at low recall.

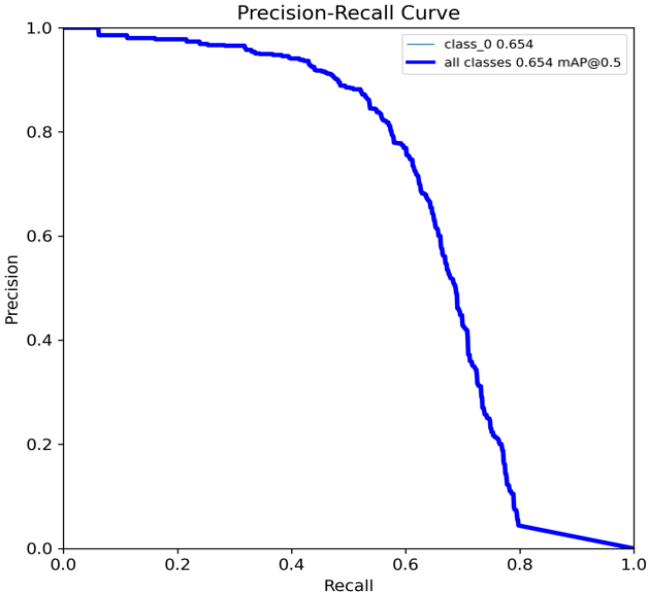

Fig 4. P-R plot of model classification

The above data again verifies that the addition of the new small target detection head can better fuse shallow high-resolution features and make up for the original network's defect of “feature loss” for small targets.

V. CONCLUSION

In this paper, the detection performance of small targets in UAV aerial photography scenarios is significantly improved by introducing the ECA attention mechanism and high-resolution small target detection head in YOLOv8n. The experimental results show that the improved model mAP50 improves from 62.2% to 65.4% of the original network of YOLOv8n, and mAP50-95 improves from 33.3% to 36.8%. This scheme effectively improves the detection accuracy of small targets under UAV aerial photography scenarios through the synergistic optimization of the attention mechanism and the multiscale detection head In the future, we can further explore the lightweight design and multitask fusion to In the future, lightweight design and multi-task fusion can be further explored to enhance the generalization ability in practical applications.

ACKNOWLEDGEMENT

This research was supported by the Sichuan Science and Technology Program, Soft Science Project (No.2022JDR0076) and Sichuan Province Philosophy and Social Science Research Project(No. SC23TJ006). We also would like to thank the sponsors of Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology.

REFERENCES

[1]. Xie, Q. , Li, D. , Yu, Z. , Zhou, J. ,& Wang, J. . (2020). Detecting trees in street images via deep learning with attention module. IEEE Transactions on Instrumentation and Measurement, 69(8), 5395-5406.

[2]. Sun, S. , Yin, Y. , Wang, X. ,& Xu, D. . (2019). Robust visual detection and tracking strategies for autonomous aerial refueling of uavs. IEEE Transactions on Instrumentation and Measurement, 4640-4652.

[3]. Paulin, G. , Sambolek, S. ,&Ivasic-Kos, M. . (2024). Application of raycast method for person geolocalization and distance determination using uav images in real-world land search and rescue scenarios. Expert Systems with Application, 237(Mar. Pt.A), 121495.1-121495.23.

[4]. Wang, X. , Yan, Y. ,& Zhu, S. D. . (2023). Dense-and-similar object detection in aerial images. Pattern recognition letters, 176(Dec.), 153-159.

[5]. Lin, T. Y. , Dollar, P. ,Girshick, R. , He, K. , Hariharan, B. ,& Belongie, S. . (2017). Feature pyramid networks for object detection. IEEE Computer Society.

[6]. Yang, Q. , Ma, S. ,& Guo, M. H. Y. . (2023). A small object detection method for oil leakage defects in substations based on improved faster-rcnn. sensors, 23(17).

[7]. Xiao, J. , Yao, Y. , Zhou, J. , Guo, H. , Yu, Q. ,& Wang, Y. F. . (2023). Fdlr-net: a feature decoupling and localization refinement network for object detection in remote sensing images. Expert Systems with Application.

[8]. Lu, X. , Ji, J. , Xing, Z. ,& Miao, Q. . (2021). Attention and feature fusion ssd for remote sensing object detection. IEEE Transactions on Instrumentation and Measurement, PP(99), 1-1.

[9]. Sun, T. , Chen, H. , Liu, H. , Deng, L. , Liu, L. ,& Li, S. . (2024). Ds-yolov7: dense small object detection algorithm for uav. IEEE Access, 12.

[10]. Lou, H. , Duan, X. , Guo, J. , Liu, H. , Gu, J. ,& Bi, L. , et al. (2023). Dc-yolov8: small-size object detection algorithm based on camera sensor. Electronics (2079-9292), 12(10).

[11]. Zhang, Y. , Dong, D. , Liu, H. , Liu, L. , Deng, L. ,& Gu, J. , et al. (2024). A more efficient algorithm for small target detection in unmanned aerial vehicles. IEEJ Transactions on Electrical and Electronic Engineering, 19(9), 11.