Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9007

Target Detection Algorithm for UAV Aerial Images Based on IRE-YOLO

Liu Yang
School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, 300222, China
Email: 2894076249@qq.com

How to cite this paper: Yang, L. (2025). Target detection algorithm for UAV aerial images based on IRE-YOLO. Academic Journal of Emerging Technologies, 1(2), 1-15. ISSN Print: 3104-4417; ISSN Online: 3104-4425.
https://doi.org/10.63313/AJET.9007
Published: 2025-10-01

Abstract

To address challenges in UAV aerial images such as a high proportion of small objects, complex backgrounds, significant scale variations, and limited computing resources, this study proposes an improved object detection algorithm based on YOLOv11n—namely, IRE-YOLO. In the backbone network, the algorithm introduces the inverted residual mobile block (iRMB) attention mechanism to enhance the model's ability to extract features of small objects in complex backgrounds. It replaces some standard convolutions with depthwise separable convolutions (DSConv) to reduce the number of parameters while maintaining feature expression capability. In the neck network, an efficient upsampling block (EUCB) is designed to recover detailed information through hierarchical feature reconstruction, thereby improving the representation quality of small objects' edges and textures. Experimental results on the DOTA dataset show that IRE-YOLO outperforms mainstream lightweight models in key metrics including precision (P), recall (R), mAP@50, and mAP@50:95, achieving 69.4%, 59.0%, 61.0%, and 34.2% respectively. Meanwhile, the model's parameter count is reduced to 8.7MB, realizing an effective balance between accuracy and lightweight design. The research demonstrates that IRE-YOLO has high practical value in small object detection tasks for UAV aerial images and provides a feasible solution for real-time deployment on embedded platforms.

Keywords

UAV aerial image; object detection; YOLOv11; lightweight model

1. Introduction

With the in-depth integration of UAV technology and computer vision, object detection based on UAV aerial images has become a key technical means in fields such as intelligent transportation, forest fire monitoring, military reconnaissance, power transmission line inspection, agricultural monitoring, and urban planning. However, affected by factors including flight altitude, shooting angle, airflow disturbance, and environmental illumination, aerial images often exhibit characteristics such as com-

plex backgrounds, significant differences in target scales, a high proportion of small targets, dense target distribution, severe occlusion, and lack of image details. These characteristics cause general object detection algorithms to face issues such as decreased detection accuracy and increased missed detection rates in this scenario. Traditional methods mostly rely on image stitching and manual feature extraction, which have limited feature expression capabilities in complex backgrounds and struggle to balance the requirements of detection accuracy and real-time performance. Therefore, to meet the practical application needs of UAV platforms, there is an urgent need to develop object detection algorithms that can effectively handle multi-scale small targets while balancing lightweight design and high accuracy.

Currently, existing object detection algorithms can be divided into two main categories: two-stage detection algorithms and one-stage detection algorithms. Represented by Faster R-CNN and Mask R-CNN, two-stage detection algorithms first generate region proposals, and then perform classification and bounding box regression on these regions. Although they achieve high detection accuracy, their detection speed is relatively slow. One-stage detection algorithms, on the other hand, are represented by SSD (Single Shot Multibox Detector) and the YOLO (You Only Look Once) series. These algorithms directly perform convolution operations on images without first generating region proposals; instead, they directly predict the category probability and position coordinates of targets, and obtain the final detection results through a single detection process, thus achieving faster detection speed.

To address problems such as large variations in target size, complex imaging conditions, and uneven distribution of small targets from the UAV perspective, the academic community has proposed a variety of innovative solutions. Yi et al. proposed the LAR-YOLOv8 algorithm based on YOLOv8, which enhances local modules through a dual-branch structure attention mechanism and introduces the RIOU loss function to improve the shape consistency between predicted boxes and ground truth boxes, effectively solving the limitations of small target detection in remote sensing images. Zhu et al. proposed the generalized oversampling Prouhet-Thue-Morse (OGPTM) method, which combines a point-wise multiplication processor (PmuP) to achieve better sidelobe suppression performance in high-speed target detection. Xie et al. proposed FARP-Net for 3D point cloud object detection tasks, innovatively introducing a weighted relational perception proposal module (WRPM), which breaks through the limitations of traditional context information extraction methods. Zhang Huawei et al. improved the activation function of the attention gate and proposed the GUS-YOLO algorithm based on the YOLO series, realizing the improvement of object detection performance on remote sensing datasets.

Despite extensive research dedicated to improving the accuracy of object detection in UAV aerial images, several common challenges remain: high missed detection rates of targets in complex backgrounds, unbalanced detection accuracy for mul-

ti-scale targets, and prominent conflicts between computational efficiency and detection accuracy. To address these issues, this study proposes an improved algorithm based on YOLOv11n, namely IRE-YOLO. Through innovative design, this algorithm not only enhances the detection accuracy of small targets but also significantly reduces model parameters, optimizes the use of computational resources, and accelerates the inference process—providing an efficient and accurate solution for small object detection in UAV aerial images. The main contributions are as follows:

(1) The inverted residual mobile block (iRMB) attention mechanism is introduced into the backbone network. By fusing the inductive bias of CNNs and the dynamic modeling capability of self-attention, the model's ability to extract features of small targets in complex backgrounds is enhanced.

(2) To optimize model efficiency for lightweight deployment, depthwise separable convolution modules (DSConv) are adopted to replace some standard convolutions in the backbone network. This modification significantly reduces the number of parameters while maintaining strong feature expression capability.

(3) An efficient upsampling convolution block (EUCB) is designed and integrated into the neck network. Through a hierarchical feature reconstruction mechanism, spatial details lost during downsampling are recovered, effectively improving the model's ability to represent details such as edges and textures of small targets.

2. Design of the IRE-YOLO Algorithm

YOLOv11, a brand-new state-of-the-art (SOTA) one-stage object detection algorithm, was proposed by numerous researchers in 2024 based on the Ultralytics framework. As the latest member of the YOLO series, this algorithm inherits the design philosophy of treating object detection as a single regression task, enabling real-time prediction of bounding box coordinates and target categories. It maintains high accuracy while achieving excellent detection speed.

According to different model scales, YOLOv11 offers five versions—n, s, m, l, and x—to adapt to application scenarios with varying computational resources and performance requirements. Structurally, the YOLOv11 model mainly consists of three components: Backbone (feature extraction network), Neck (feature fusion network), and Head (detection network). This classic architecture collectively ensures the model's outstanding performance in terms of efficiency and accuracy.

However, when applied to object detection tasks in UAV aerial images, the YOLOv11 algorithm still has certain limitations: For targets with a small pixel proportion, position and detail information are easily lost, leading to a high missed detection rate of small targets; its robustness against complex background interference is insufficient, which affects the purity of feature extraction; in addition, target scales in aerial images vary drastically, and the detailed features of small targets are easily weakened during multiple downsampling processes, thereby impairing detection performance.

To address the aforementioned issues, this study proposes a small object detection algorithm—IRE-YOLOv11—for UAV aerial images, based on the lightweight YOLOv11n. The structure of the algorithm is illustrated in **Figure 1**. Improvements to the algorithm are primarily implemented in three aspects:

First, the inverted residual mobile block (iRMB) attention mechanism is introduced into the backbone network to enhance the model's ability to extract features of small targets against complex backgrounds. Second, depthwise separable convolution (DSConv) modules are adopted to replace some standard convolutions, which not only improves the richness of feature expression and nonlinear fitting capability but also reduces the number of model parameters. Third, the EUCB upsampling module is integrated into the neck network to enhance the ability to recover small target details through refined feature reconstruction.

Specifically, in the improved model, the residual blocks in the C3k2 module of the backbone network are replaced with iRMB modules; some traditional convolutions are substituted with DSConv modules; and the original upsampling operation in the neck network is replaced by EUCB. Through these structural optimizations, IRE-YOLOv11 significantly enhances the detection performance for small targets in UAV images while maintaining a high inference speed.

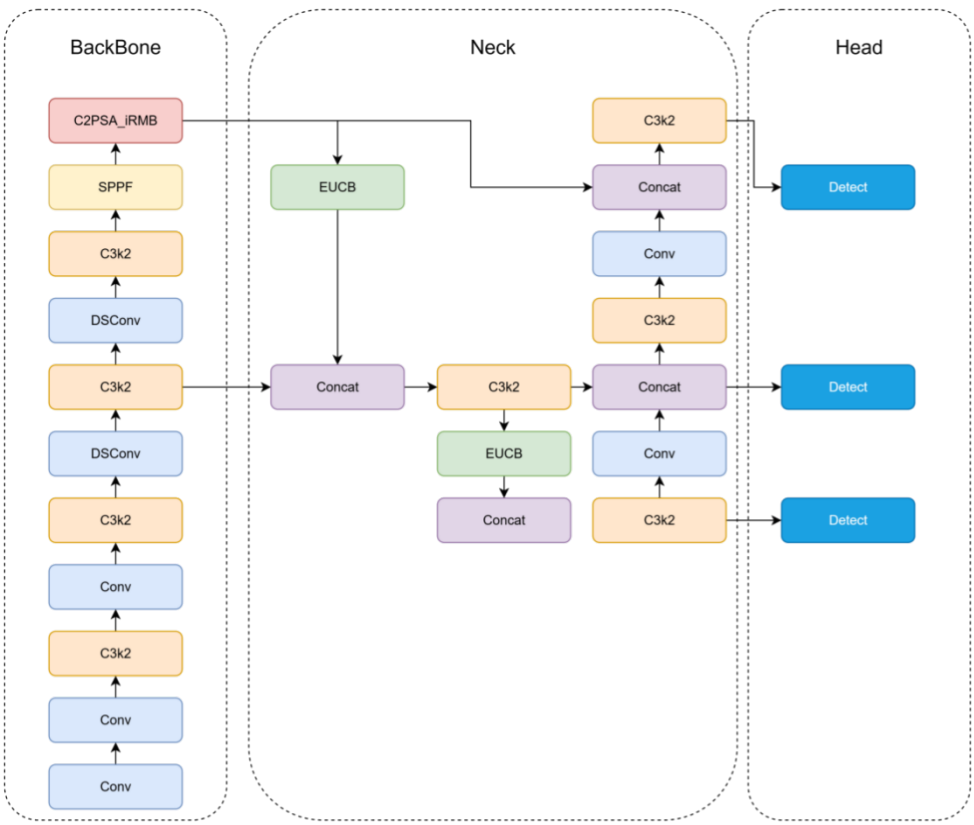

Figure 1. Structure of the IRE-YOLO Model.

2.1. iRMB Attention Mechanism

In UAV aerial images, due to the long shooting distance, the resolution of distant targets is low. Traditional convolutional neural networks, with their limited receptive fields, struggle to effectively capture target details, which easily leads to missed detections. Meanwhile, when the backbone network of YOLOv11 processes multi-scale objects, as it mainly relies on standard convolutions, it cannot fully capture the contextual information of objects of different sizes. This limits its feature expression capability in complex scenes captured by UAVs. Additionally, this design results in high parameter requirements for the model, making it difficult to meet the lightweight requirements of foreign object detection platforms. To address these issues, this paper proposes a multi-scale global attention module, iRMB, to replace the residual blocks in C3k2 of the backbone network. The structure of the iRMB module is shown in **Figure 2**.

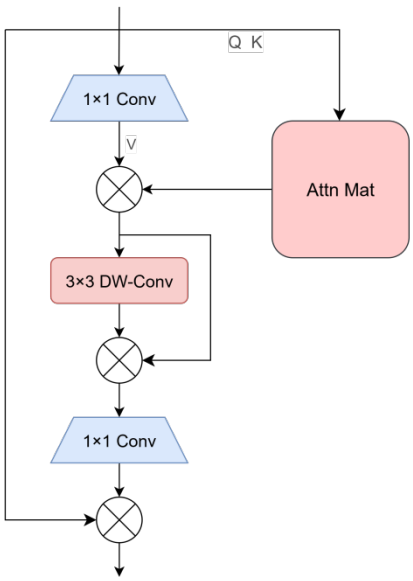

Figure 2. Structure of the iRMB Module.

The core innovation of the iRMB attention mechanism lies in the organic integration of the inductive bias of CNNs and the dynamic expression capability of self-attention. By introducing two adjustable dimensions—"expansion ratio" λ and "efficient operator" $\mathcal{F}$—this mechanism provides flexible structural adaptability. The calculation of its attention weights follows the standard self-attention mechanism, as shown in Equation (1):

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (1)$$

Among them, Q, K, and V are generated by different 1×1 convolutional layers, respectively, and $\sqrt{d_k}$ is the scaling factor. This design enables the model to balance accuracy and efficiency under different computational budgets. From a structural

perspective, iRMB can be regarded as a specific instance of the Meta-Mobile Block. The design supports modular combination through different expansion ratios and operators, forming a dynamic network structure that can adapt to different task requirements.

The iRMB is a hybrid network module that combines depthwise separable convolution (3×3 DW-Conv) and the self-attention mechanism. A 1×1 convolution is used for channel compression and expansion to optimize computational efficiency. Depthwise separable convolution (DW-Conv) is applied to capture spatial features, while the attention mechanism is utilized to capture global dependencies between features. This figure illustrates the structural paradigm of the iRMB.

2.2. DSConv Convolution

In the task of object detection in UAV aerial images, the targets to be detected usually account for a small proportion of pixels. This requires the detection model to extract deep features with a large receptive field and high resolution, which often needs to be achieved by increasing the network depth or improving the input image resolution—thereby significantly increasing the model's parameter count and computational complexity.

However, although the standard convolution operation adopted by traditional object detection models based on convolutional neural networks can simultaneously perform spatial feature extraction and channel information fusion, its integrated design not only pursues accuracy but also introduces a heavy computational burden and parameter redundancy. This contradiction is particularly prominent on embedded UAV platforms with limited computing resources, seriously restricting the real-time deployment of the model.

To improve efficiency while ensuring detection accuracy, this study introduces the depthwise separable convolution module (DSConv) to replace some standard convolution structures in the backbone network, thereby seeking an effective balance between accuracy and efficiency.

As the core operation of convolutional neural networks, standard convolution realizes the filtering and transformation of input feature maps by simultaneously processing spatial correlation and inter-channel information in sliding windows, and it exhibits strong feature expression capability. However, this operation incurs high parameter counts and computational overhead when tackling complex tasks.

To improve model efficiency, this study introduces depthwise separable convolution (DSConv), which decouples standard convolution into two separate steps: depthwise convolution and pointwise convolution. Depthwise convolution performs spatial filtering on each input channel independently without inter-channel information interaction; pointwise convolution, on the other hand, achieves the fusion of channel features and dimensional mapping through 1×1 convolution.

This structure significantly reduces computational complexity and parameter scale

while still maintaining effective feature extraction capability. Its structure is shown in **Figure 3**, and the parameter count and computational load are presented in Equations (2) and (3).

$$Dsconv_{number\ of\ parameters} = D_k^2 \times M + M \times N. \tag{2}$$

$$DSConv_{computational\ complexity} = D_k^2 \times M \times D_{FF}^2 + M \times N \times D_{FF}^2. \tag{3}$$

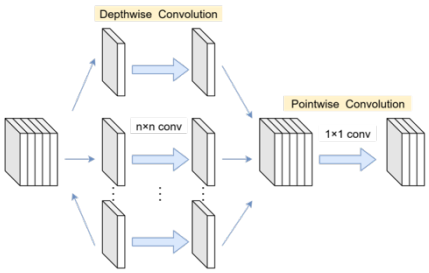

Figure 3. Structure Diagram of DSConv.

2.3. EUCB up - sampling

During the UAV aerial inspection process, due to the varying distances between insulators and the UAV, when the distance is large, the scale of defect targets in the captured images decreases significantly and the number of pixels they occupy is limited—posing great difficulties for defect feature reconstruction.

Traditional upsampling operations use fixed interpolation kernels and lack learnable capabilities. When restoring spatial resolution in the decoding stage, they tend to cause excessive smoothing of detailed features (such as insulator flashover textures or broken edges) during the reconstruction process. In addition, insulator strings are usually densely arranged, and the pixel proportion of defect areas is relatively low; errors in feature reconstruction can easily lead to false detection or missed detection of key targets.

Therefore, inspired by the EMCAD method in the field of medical image segmentation and to address the aforementioned issues, this study introduces an efficient upsampling convolution module to replace the traditional upsampling operation in the neck network. Its structure is shown in **Figure 4**.

Figure 4. Structure of the EUCB Module.

The EUCB is an efficient upsampling module that adopts a hierarchical design to achieve refined feature reconstruction. The specific process is as follows: First, it expands the spatial resolution of the feature map by 2 times through an upsampling operation. Then, it uses 3×3 depthwise separable convolution to enhance spatial feature representation, which significantly reduces computational complexity while expanding the model's receptive field. Next, batch normalization and ReLU activa-

tion function are applied to further improve feature expression capability. Finally, a 1×1 convolution is used to realize channel dimension compression and feature fusion, ensuring that the channels of the output features are accurately aligned with those of the subsequent skip connection layer. This calculation process can be formally expressed as Equation (4).

$$EUCB(X) = Conv_{1\times1}(ReLU(BN(DWC_{3\times3}(Upsample(X))))). \quad (4)$$

Compared with traditional upsampling methods, the key advantage of EUCB lies in its efficient computational design. By replacing standard convolution operations with depthwise separable convolution, the module significantly reduces computational overhead and parameter count while maintaining performance. Although EUCB was originally designed for medical image segmentation, its modular structure exhibits unique adaptability in UAV-based defect detection: the local perception characteristic of depthwise separable convolution can effectively suppress complex background noise, and the channel rearrangement mechanism enhances the cross-channel information flow through implicit feature permutation—enabling the effective fusion of high-level semantics of defect regions and shallow features such as edge details during the upsampling process.

In the IRE-YOLO model proposed in this study, the EUCB module is embedded into the feature fusion part of the neck network to replace the original upsampling operation. This module can effectively recover the spatial detail loss caused by pooling and convolution operations, significantly improve the ability of feature maps to represent sub-pixel-level defects, and at the same time enhance the alignment accuracy and fusion quality between feature maps of different scales—providing a more discriminative multi-scale feature foundation for the detection network.

3. Experiments and Results Analysis

3.1. Dataset

In this study, the DOTA dataset was used for training and validation of the IRE-YOLO algorithm. The DOTA image dataset is a large-scale image dataset specifically designed for detecting small-scale targets that randomly appear in aerial image systems. It includes diverse image sources, covering aerial images collected by different sensors and platforms. As it is applicable for small target detection in aerial images and holds significant advantages over other datasets of the same type in terms of both quantity and quality, the DOTA dataset was selected for the experiments.

The dataset contains 3,156 aerial images, divided into a training set (1,893 images), a validation set (631 images), and a test set (632 images). These images fully retain the annotations for DOTA's 15 target categories, covering typical targets such as aircraft, vehicles, and ships, with a total of over 180,000 instances. The dataset includes various scenarios such as cities, rural areas, and coasts, featur-

ing large target scale differences, dense distribution, and variable orientations. These characteristics enable effective verification of the algorithm's performance in complex aerial environments.

Given that small targets account for more than 60% of the dataset, this study adopted the Mosaic data augmentation technique, which splices four training images into one to significantly increase the occurrence frequency of small targets. Meanwhile, methods such as random rotation, scale transformation, and color space adjustment were used to enhance the model's robustness to changes in target scale and lighting conditions. To adapt to YOLO algorithm training, the original oriented bounding box (OBB) annotations were converted into horizontal detection boxes, and K-means clustering was applied to recalculate the anchor box sizes. All input images were standardized to ensure training stability. The details of the dataset are shown in **Figure 5**.

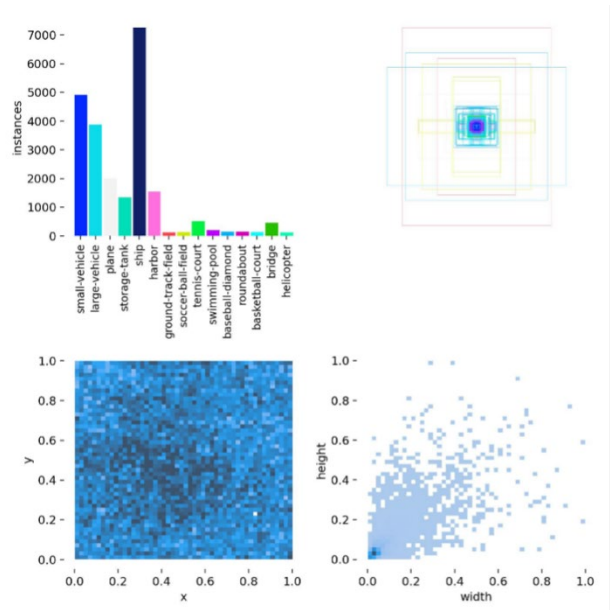

Figure 5. Dataset Visualization Distribution Map.

3.2. Experimental Environment and Evaluation Metrics

All experiments were conducted in the same hardware environment, and the relevant environment configuration is shown in **Table 1**.

Table 1. Table of Relevant Environment Configuration.

software and hardware configuration	version or model
CPU	Intel Xeon Gold 6226R
GPU	NVIDIA GeForce RTX 3080
Python	3.8.19
Pytorch	1.13.1
YOLOv11 version	8.3.24

In this study, the algorithm performance was evaluated using common evaluation metrics for object detection tasks, including: mean average precision (mAP), recall (R), precision (P), model size, and floating-point operations per second (FLOPS).

Precision reflects the proportion of samples predicted as positive by the model that are actually positive, while recall reflects the proportion of actual positive samples that are correctly predicted as positive by the model. These two metrics are defined based on true positives (TP), false positives (FP), and false negatives (FN). Specifically, TP denotes the number of samples predicted as positive and actually positive; FP denotes the number of samples predicted as positive but actually negative; and FN denotes the number of samples actually positive but predicted as negative.

Average precision (AP) is the average of precision values across different categories, used to measure the model's comprehensive performance on all categories. mAP is the average value of AP across all categories.

The corresponding calculation formulas are shown in Equations (5), (6), (7), and (8):

$$R = \frac{TP}{TP+FN}. \tag{5}$$

$$P = \frac{TP}{TP+FP}. \tag{6}$$

$$AP = \int_0^1 P(R)dR. \tag{7}$$

$$mAP = \frac{1}{n}\sum_1^n AP_i. \tag{8}$$

3.3. Ablation Experiment

To further evaluate the comprehensive performance of the model, this study plotted Precision-Recall (PR) curves for comparative analysis. As shown in **Figure 6**, the PR curve of IRE-YOLO is closer to the upper-right corner of the coordinate axis compared to the baseline model, and it has the largest Area Under the Curve (AUC). This indicates that IRE-YOLO can maintain high precision and recall across different confidence thresholds.

Notably, in the high-recall region, the precision of IRE-YOLO decreases more gently, which demonstrates that the model exhibits better robustness in detecting hard samples. By comparing the PR curves of other ablation experiment models, it can be observed that as each component in the IRE-YOLO module is incorporated, the curve gradually shifts toward the upper-right direction. This finding mutually corroborates the quantitative analysis results and intuitively illustrates the contribution of each module to the model's performance.

To verify the improvement of detection accuracy by each improvement strategy adopted in this study, ablation experiments experiments were conducted using YOLOv11n as the baseline model. The evaluation metrics included precision (P), recall (R), mAP@0.5, mAP@50:95, FPS, and model size. Each group of experiments

adopted the same training strategy, and ablation experiments were performed through combinations of multiple improvements. The experimental results are shown in **Table 2**.

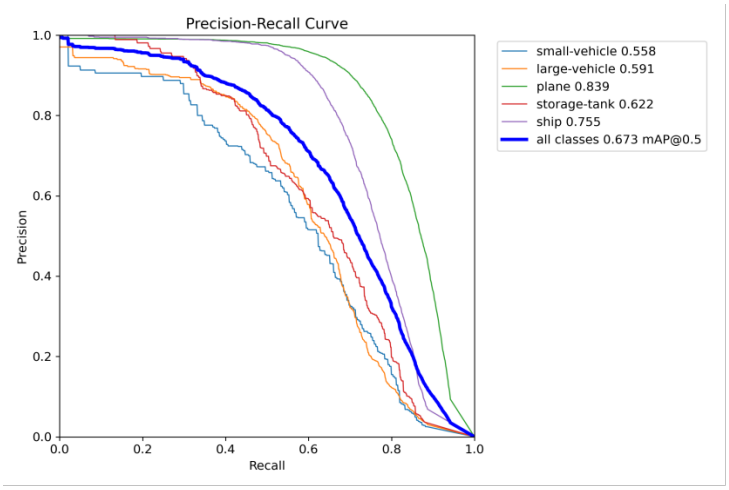

Figure 6. P-R Curve.

Table 2. Ablation Experiment Results.

Baseline	iRMB	DSConv	EUCB	P/%	R/%	mAP @50/%	mAP @50:95/%	Model Size/MB	FPS
Yolov11n				65.6	56.6	56.7	32.6	9.3	26.6
	✓			67.4	56.9	56.4	33.0	10.1	27.4
		✓		66.6	58.1	56.4	32.7	7.9	25.5
			✓	68.1	57.4	58.3	33.4	11.0	24.6
	✓	✓	✓	69.4	59.0	61.0	34.2	8.7	23.0

Analysis of the ablation experiments shows that the introduction of each module has brought targeted improvements to the model performance. When the iRMBattention mechanism was added alone, the model's precision (P) increased to 67.4%, enhancing its ability to extract features of small targets against complex backgrounds. The adoption of DSConv depthwise separable convolution reduced the model size to 7.9MB, achieving effective lightweighting, while increasing the recall (R) to 58.1%. The introduction of the EUCB upsampling module significantly improved the quality of detail reconstruction, raising mAP@50 to 58.3%.

When the three modules were integrated to form IRE-YOLO, the model exhibited the best comprehensive performance: precision (P) and recall (R) increased to 69.4% and 59.0% respectively; mAP@50 and mAP@50:95 reached 61.0% and 34.2% respectively; and the model size (8.7MB) was better than that of the baseline, achieving an effective balance between accuracy and lightweighting. This indicates that through the synergistic effect of multiple modules, IRE-YOLO effectively enhances the ability to capture features of multi-scale small targets while maintaining a lightweight model, improving the recall rate of positive samples and reducing missed detections and false detections.

In summary, the ablation experiments verify the effectiveness and complementarity

of each module in IRE-YOLO, and the overall model achieves a coordinated improvement in accuracy and efficiency in the task of small target detection in UAV aerial photography.

3.4. Comparative Experiments

To verify the effectiveness of the proposed model, IRE-YOLO was first trained and validated on the dataset alongside mainstream object detection models (such as YOLOv5, YOLOv8, YOLOv10, and YOLOv11). The models were compared based on metrics including mAP@50, mAP@50:95, precision (P), recall (R), FPS, and model size. The results of the comparative experiments are shown in **Table 3**.

Table 3. Comparative Experiment Results.

Model	P/%	R/%	mAP @50/%	mAP @50:95/%	Model Size/MB	FPS
YOLOv5n	62.0	46.8	39.6	24.3	3.6	35.7
YOLOv8n	63.5	44.9	39.2	24.8	6.0	33.2
YOLOv10n	61.0	44.2	37.5	23.3	5.5	30.9
YOLOv11n	65.6	56.6	56.7	32.6	9.3	26.6
IRE-YOLO	69.4	59.0	61.0	34.2	8.7	23.0

The comparative experiment results in Table 3 show that the IRE-YOLO model proposed in this study outperforms current mainstream lightweight detection models in multiple performance metrics.

In terms of detection accuracy, the mAP@50 and mAP@50:95 of IRE-YOLO reach 61.0% and 34.2% respectively, which are 4.3 percentage points and 1.6 percentage points higher than those of the baseline model YOLOv11n. Meanwhile, the precision (P) and recall (R) are increased to 69.4% and 59.0% respectively, reflecting that the model significantly enhances its ability to identify positive samples while maintaining high classification accuracy.

In terms of model complexity, the parameter count of IRE-YOLO is 8.7MB, which is lower than the 9.3MB of YOLOv11n, demonstrating excellent lightweight characteristics. Although the inference speed (FPS) is 23.0, slightly lower than that of the baseline model, IRE-YOLO achieves the optimal balance between accuracy and model complexity. This improved architecture is therefore more suitable for lightweight devices such as UAVs.

In summary, IRE-YOLO achieves improved detection performance while reducing the number of parameters, and has strong engineering application value.

3.5. Visualization Results

To intuitively compare the detection performance of the improved algorithm proposed in this study and YOLOv11n in UAV aerial photography scenarios, aerial images covering several complex scenarios from the DOTA dataset—including dense small targets, varying lighting conditions, multi-scale targets, and low-light environments—were selected for a detection result comparison experiment. Among

these, clustered small targets and complex lighting are common challenging scenarios in aerial image detection. The multi-scale target scenario refers to images where targets with significant size differences (e.g., pedestrians, cars, and trucks) coexist, requiring the model to balance both shallow detail features and deep semantic features. The experimental results are shown in **Figure 7** and **Figure 8**.

Figure 7. Comparison of YOLOv11 and IRE-TOLO Detection Results.

Figure 8. Comparison of YOLOv11 and IRE-TOLO Detection Result Diagrams.

In the figures below: the left column shows the original images, the middle column presents the detection results of the YOLOv11n model under different scenarios, and the right column displays the detection results of the improved IRE-YOLO algorithm proposed in this study. Differences in detection performance between IRE-YOLO and YOLOv11n are marked with red boxes in the figures.

In the scenario with dense small targets, compared with the baseline algorithm, IRE-YOLO has a lower missed detection rate. It can accurately identify targets in highly overlapping environments, achieving better detection performance than the baseline. In scenarios with varying natural lighting, IRE-YOLO can suppress interference from background noise while preserving feature information that is more

critical for target decision-making, resulting in better detection performance than the baseline. In multi-scale scenarios, IRE-YOLO can effectively adapt to changes in targets of different scales, accurately detect objects of various sizes, and achieve a higher recognition rate for pedestrians and cars. In low-light nighttime scenarios, IRE-YOLO can overcome the impact of insufficient light on target recognition, outperforming the baseline algorithm in detection results.

In summary, the IRE-YOLO algorithm demonstrates better detection performance across all complex scenarios, and exhibits stronger robustness especially under unstable lighting conditions.

4. Conclusion

To address challenges in UAV aerial images—such as a high proportion of small targets, complex backgrounds, significant target scale differences, and limited computing resources—this study proposes an efficient small target detection algorithm, IRE-YOLO, by making targeted improvements to the YOLOv11n model from multiple dimensions, including feature extraction, model lightweighting, and detail reconstruction. This algorithm significantly enhances detection accuracy and reduces model complexity while maintaining a high inference speed.

First, to strengthen the model's ability to capture features of small targets in complex backgrounds, an inverted residual mobile block (iRMB) attention mechanism is integrated into the backbone network. This effectively fuses local details with global contextual information. Second, considering lightweight deployment requirements, depthwise separable convolution modules (DSConv) are used to replace some standard convolutions, which drastically reduces the number of parameters while preserving the model's representational capacity. Finally, to solve the problem of detail loss during upsampling, an efficient upsampling convolution module (EUCB) is designed in the neck network. Through hierarchical feature reconstruction, this module improves the ability to recover key details of small targets, such as edges and textures.

Experimental results on the public DOTA dataset show that the IRE-YOLO algorithm outperforms mainstream lightweight models in key metrics including mAP@50 and recall. Meanwhile, the model's parameter count is further reduced, achieving a good balance between accuracy and efficiency. However, the algorithm still faces a certain risk of missed detections in scenarios with extremely high target density or severe occlusion. Future research will focus on optimizing the model's robustness in extreme scenarios and exploring its application in more challenging real-time inspection tasks.

References

[1] TIAN M ,CUI M ,CHEN Z , et al. MFR-YOLOv10: Object detection in UAV-taken images based on multilayer feature reconstruction network[J].Chinese Journal of Aeronautics,2025,38(11):103456-103456.DOI:10.1016/J.CJA.2025.103456.

[2] Zhang J ,Gao M ,Song L , et al. REA-YOLO for small object detection in UAV aerial images[J].The Journal of Supercomputing,2025,81(14):1332-1332.DOI:10.1007/S11227-025-07836-0.

[3] Yan Y ,Zhang L . A Review of YOLO Series Algorithms in Object Detection for UAV Aerial Images[J].International Core Journal of Engineering,2025,11(9):74-92.DOI:10.6919/ICJE.202509_11(9).0009.

[4] Zhou H ,Yin W ,Sun K , et al. FO-YOLO for small object detection in drone aerial imagery[J].The Journal of Supercomputing,2025,81(12):1208-1208.DOI:10.1007/S11227-025-07688-8.

[5] Bian Y ,Teng H ,Lv Q , et al. YOLO-AAHU: An adaptive multi-scale small target and occluded target detection algorithm for drone aerial photography scenarios[J].Engineering Research Express,2025,7(3):035226-035226.DOI:10.1088/2631-8695/ADEEFE.

[6] Ying Z ,Yisheng H .A Survey of SAR Image Target Detection Based on Convolutional Neural Networks[J].Remote Sensing,2022,14(24):6240-6240.

[7] Yihunie S A ,E. J B .A Survey on Deep-Learning-Based LiDAR 3D Object Detection for Autonomous Driving[J].Sensors,2022,22(24):9577-9577.

[8] Bao D ,Hao L .Survey of Target Detection Based on Neural Network[J].Journal of Physics: Conference Series,2021,1952(2):

[9] Kang T ,Yiquan W ,Fei Z .Recent advances in small object detection based on deep learning: A review[J].Image and Vision Computing,2020,97(prepublish):103910-103910.

[10] Zhao Z ,Zheng P ,Xu S , et al.Object Detection With Deep Learning: A Review.[J].IEEE Trans. Neural Netw. Learning Syst.,2019,30(11):3212-3232.