Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MSAL-YOLO: a YOLOv8-based detector for small and densely distributed object detection in UAV aerial imagery

Fei Ding
Fuyang Normal University
Xiufu Du
Fuyang Normal University
Haining Zhang
202507063@fynu.edu.cn
Fuyang Normal University
Haibin Liu
Fuyang Normal University
Liguo Han
Fuyang Normal University

Article

Keywords: UAV aerial imagery, small and dense object detection, YOLOv8-based detector, mixed convolution, coordinate attention, Wise-IoU

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10147699/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 7th, 2026.
See the published version at https://doi.org/10.1038/s41598-026-71102-3.

MSAL-YOLO: a YOLOv8-based detector for small and densely distributed object detection in UAV aerial imagery

Fei Ding[1,2], Xiufu Du[1,2], Haining Zhang[1,2,*], Haibin Liu[1,2], and Liguo Han[1,2]

[1]School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China
[2]Anhui Engineering Research Center for Intelligent Computing and Information Innovation, Fuyang Normal University, Fuyang 236037, China
[*]corresponding:202507063@fynu.edu.cn

ABSTRACT

Object detection in UAV aerial imagery is challenging because of small object scales, dense distributions, occlusions, and complex backgrounds. To address these issues, we propose MSAL-YOLO, an improved YOLOv8-based detector for aerial object detection. Specifically, a Spatial–Channel Mixed Convolution (SCMConv) module is introduced to enhance local feature extraction by combining standard and dilated convolutions. A Multi-Branch Enhanced Coordinate Attention (MBECA) mechanism is further designed to strengthen spatial dependency modeling and contextual interaction. In addition, a Region-Adaptive Wise-IoU (RA-WIoU) loss is developed to alleviate scale imbalance by increasing the contribution of small objects during training. Experiments on the VisDrone2019 dataset show that MSAL-YOLO improves mAP@0.5 and mAP@0.5:0.95 from 30.0% and 17.0% to 35.7% and 20.5%, respectively, while maintaining competitive model complexity. Cross-dataset experiments on DOTAv1 further demonstrate the robustness and generalization capability of the proposed method. These results indicate that MSAL-YOLO provides an effective solution for UAV aerial object detection in complex scenes.

Keywords: UAV aerial imagery; small and dense object detection; YOLOv8-based detector; mixed convolution; coordinate attention; Wise-IoU

Introduction

Deep learning-based object detection has achieved remarkable progress in recent years and has been widely applied in autonomous driving, intelligent surveillance, remote sensing, and unmanned aerial vehicle (UAV) inspection systems[1]. Existing detectors can generally be divided into two categories: two-stage and one-stage methods. Two-stage detectors, represented by Faster R-CNN, achieve high detection accuracy through region proposal mechanisms but often suffer from limited inference efficiency[2]. In contrast, one-stage detectors such as SSD and the YOLO family perform object classification and localization in a unified framework, providing a better balance between accuracy and speed[3–5].

Despite these advances, object detection in UAV aerial imagery remains challenging. Due to high-altitude imaging and large observation ranges, objects often occupy only a small number of pixels and frequently appear with dense distributions, severe occlusions, viewpoint variations, and complex backgrounds[6]. Recent studies have identified insufficient fine-grained feature representation and large scale variation as major obstacles to accurate aerial object detection[7]. Public benchmarks such as VisDrone and DOTA have therefore become important platforms for evaluating UAV-based detection methods[8,9]. Furthermore, recent remote sensing detectors have demonstrated that effective contextual modeling and enlarged receptive fields are critical for robust detection in complex aerial scenes[10].

To address these challenges, extensive research has focused on improving feature representation and multi-scale perception. Feature Pyramid Network (FPN), PANet, and EfficientDet enhance cross-scale information fusion through hierarchical feature aggregation[11–13]. Transformer-based detectors, such as Lite-DETR, further improve feature interaction by exploiting global contextual relationships[14]. Lightweight architectures including FasterNet have also been proposed to improve computational efficiency while maintaining competitive performance[15]. Although these approaches improve multi-scale representation, preserving discriminative features for small objects in complex UAV scenes remains difficult.

Attention mechanisms have also been widely adopted to strengthen feature discrimination and contextual modeling. Representative methods include CBAM, Non-local Networks, and Coordinate Attention[16–18]. More recently, lightweight and dynamic attention architectures, such as SimAM, BiFormer, and EfficientViT, have demonstrated strong capability in modeling long-range dependencies with relatively low computational overhead[19–21]. In parallel, deformable operators,

including Deformable Convolution and Deformable DETR, improve spatial adaptability through dynamic sampling and attention mechanisms[22–24]. Nevertheless, existing approaches often focus on local refinement or single-scale enhancement, which limits their effectiveness for densely distributed small objects in aerial imagery.

Efficient deployment is another important consideration for UAV applications. Lightweight networks such as GhostNet and ShuffleNetV2 reduce computational redundancy through efficient feature generation and channel interaction mechanisms[25,26]. Recent detectors, including Gold-YOLO and DAMO-YOLO, further improve inference efficiency through optimized feature aggregation strategies[27,28]. However, lightweight designs may sacrifice feature representation capability, making it difficult to simultaneously preserve local details and contextual information required for small-object detection.

In addition to feature extraction, bounding-box regression plays an important role in detection performance. A series of IoU-based losses, including GIoU, DIoU, SIoU, Focal-EIoU, α-IoU, MPDIoU, and Wise-IoU, have been proposed to improve localization quality and optimization stability[29–36]. However, most existing losses are not specifically designed for the severe scale imbalance commonly encountered in UAV imagery, where small objects contribute limited optimization signals during training.

Recent detectors, including YOLOv5, YOLOv7, YOLOX, DETR, and YOLOv9, have achieved strong performance across various benchmarks[37–41]. Several UAV-oriented methods, such as TA-YOLO and DEA-YOLOv9, have further improved small-object perception through enhanced attention mechanisms and multi-scale feature learning[42,43]. Nevertheless, three key challenges remain. First, lightweight detectors still struggle to capture discriminative features for small and densely distributed objects. Second, existing attention and feature-fusion strategies often provide insufficient contextual interaction across different scales. Third, current regression losses inadequately address scale imbalance, limiting localization performance for small targets.

To address these issues, this study proposes MSAL-YOLO, an enhanced YOLOv8-based framework for UAV aerial object detection. The proposed framework incorporates three complementary components. First, a Spatial–Channel Mixed Convolution (SCMConv) module is introduced to improve local feature extraction and contextual information integration. Second, a Multi-Branch Enhanced Coordinate Attention (MBECA) mechanism is designed to strengthen spatial awareness and multi-scale contextual modeling. Third, a Region-Adaptive Wise-IoU (RA-WIoU) loss is developed to increase the optimization contribution of small objects through scale-aware dynamic weighting.

The main contributions of this paper are summarized as follows:

(1) A Spatial–Channel Mixed Convolution (SCMConv) module is proposed to improve feature representation by integrating local spatial cues and contextual information, thereby enhancing the perception of small objects while suppressing redundant features.

(2) A Multi-Branch Enhanced Coordinate Attention (MBECA) mechanism is designed to strengthen spatial awareness and cross-branch feature interaction, enabling more effective contextual modeling and multi-scale representation in UAV aerial scenes.

(3) A Region-Adaptive Wise-IoU (RA-WIoU) loss is introduced to address scale imbalance during training by adaptively increasing the optimization contribution of small objects and improving localization performance in dense scenes.

Extensive experiments on the VisDrone2019 and DOTA datasets demonstrate that the proposed method consistently outperforms the baseline detector. Compared with YOLOv8n, MSAL-YOLO achieves clear improvements in detection accuracy while maintaining competitive model complexity, confirming its effectiveness for UAV-based small and dense object detection in complex aerial environments.

Methods

MSAL-YOLO Framework

YOLOv8 provides multiple model variants with different computational complexities. Among them, YOLOv8n offers a favorable balance between detection accuracy, computational efficiency, and real-time deployment capability, making it suitable for UAV platforms. Therefore, YOLOv8n is selected as the baseline architecture in this study. The overall architecture of the baseline model is provided in Supplementary Fig. S1.

YOLOv8n consists of a backbone, a neck, and an anchor-free detection head. The backbone employs C2f modules and a spatial pyramid pooling-fast (SPPF) module for hierarchical feature extraction, while the PAN-FPN neck facilitates multi-scale feature fusion. Although YOLOv8n achieves efficient inference, its capability to represent small and densely distributed objects remains limited in complex UAV aerial scenes.

To address these limitations, an improved framework termed MSAL-YOLO is proposed, as illustrated in Fig. 1. Built upon the lightweight YOLOv8n architecture, the proposed method enhances feature representation, attention modeling, and localization learning while preserving real-time inference capability.

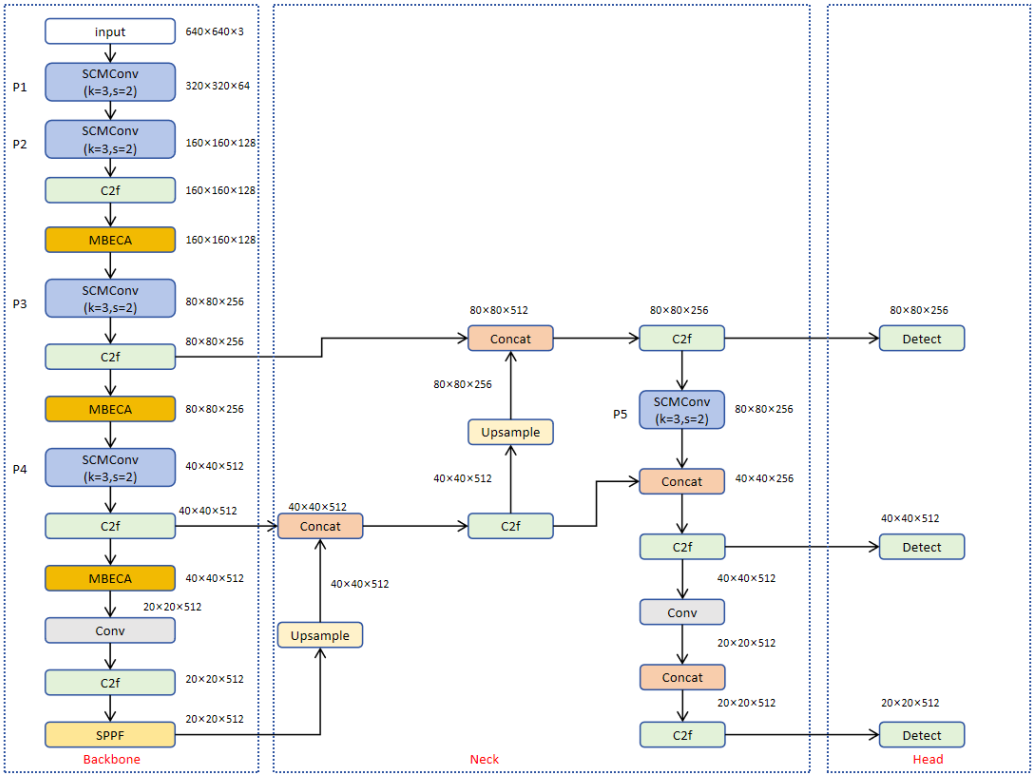


Figure 1. Overall architecture of the proposed MSAL-YOLO framework.

Specifically, the standard convolution layers in the backbone and neck are replaced by the proposed Spatial–Channel Mixed Convolution (SCMConv) modules to improve local feature extraction and contextual interaction. In addition, Multi-Branch Enhanced Coordinate Attention (MBECA) modules are introduced into the backbone to strengthen spatial awareness and suppress background interference in complex aerial environments. Furthermore, the conventional regression loss is replaced with Region-Adaptive Wise-IoU (RA-WIoU), which introduces scale-aware dynamic weighting to improve localization robustness and increase the optimization contribution of small objects.

The resulting framework retains the original multi-scale detection pipeline of YOLOv8 while providing stronger feature discrimination and more robust localization performance for UAV aerial object detection.

Spatial–Channel Mixed Convolution Module

In UAV aerial imagery, object instances often occupy only a limited number of pixels and are frequently distributed in cluttered environments. Accurate feature extraction therefore requires both precise local structural representation and sufficient contextual perception. Standard convolution mainly captures local spatial patterns, whereas dilated convolution enlarges the receptive field but may weaken spatial continuity. To simultaneously preserve fine-grained details and contextual information, a Spatial–Channel Mixed Convolution (SCMConv) module is introduced.

As illustrated in Fig. 2, the proposed SCMConv comprises a local-detail extraction branch, a contextual perception branch, an adaptive fusion mechanism, and a lightweight channel attention module. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, local spatial information is first extracted through a standard convolution operation:

$$X_{\text{local}} = \phi_{3\times 3}(X) \tag{1}$$

where $\phi_{3\times 3}(\cdot)$ denotes a 3×3 convolution followed by Batch Normalization and the SiLU activation function.

To capture broader contextual information, a depthwise separable dilated convolution branch is introduced:

$$X_{\text{context}} = \phi_{pw}\left(\phi_{dw}^{(d=2)}(X)\right) \tag{2}$$

where $\phi_{dw}^{(d=2)}(\cdot)$ represents depthwise dilated convolution with dilation rate $d=2$, and $\phi_{pw}(\cdot)$ denotes a pointwise 1×1 convolution.

The outputs of the two branches are subsequently integrated through learnable fusion coefficients:

$$X_f = \alpha \cdot X_{\text{local}} + \beta \cdot X_{\text{context}} \tag{3}$$

where α and β are trainable parameters initialized to 1.0. During training, the network automatically adjusts their values according to the feature responses of the two branches.

To further enhance channel-wise feature discrimination, a lightweight channel attention mechanism is employed. Global average pooling is first used to aggregate channel-wise statistics, followed by channel compression and reconstruction:

$$A = \sigma\left(W_2 \cdot \delta\left(W_1\left(\text{GAP}(X_f)\right)\right)\right) \tag{4}$$

where $\text{GAP}(\cdot)$ denotes global average pooling, W_1 and W_2 denote two 1×1 convolution layers, and $\delta(\cdot)$ and $\sigma(\cdot)$ represent the SiLU and Sigmoid activation functions, respectively.

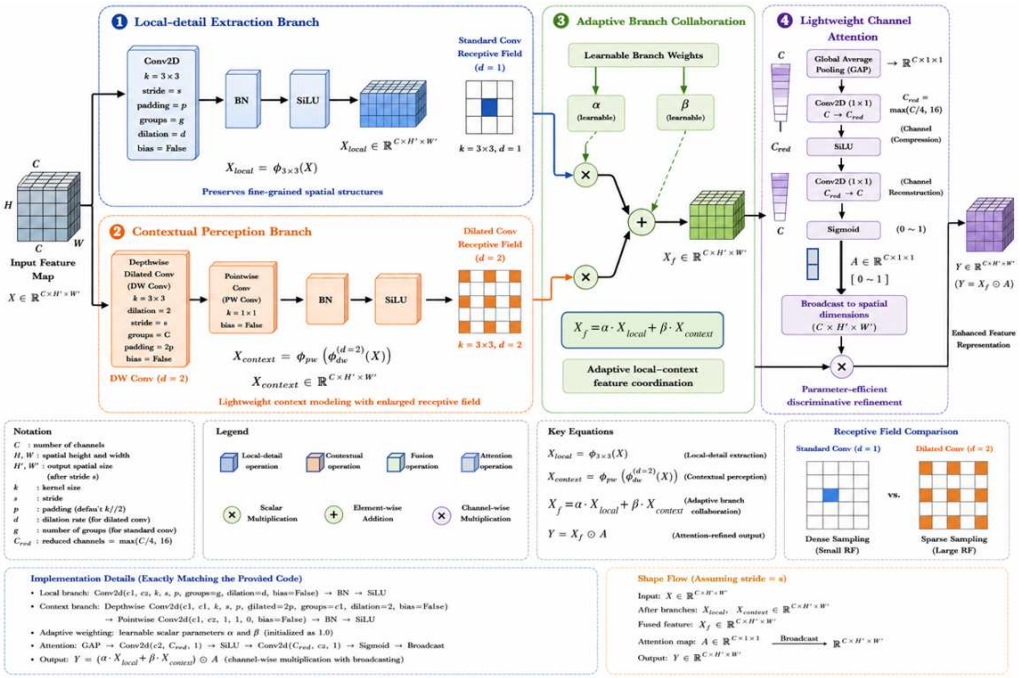


Figure 2. SCMConv module structure diagram.

The resulting attention vector is then used to recalibrate the fused feature representation:

$$Y = X_f \odot A \tag{5}$$

where $\odot$ denotes element-wise multiplication.

By integrating local-detail extraction, contextual perception, adaptive feature fusion, and channel-wise recalibration into a unified framework, SCMConv enhances feature representation capability while maintaining computational efficiency, thereby providing more informative features for UAV object detection.

Multi-Branch Enhanced Coordinate Attention (MBECA) Mechanism

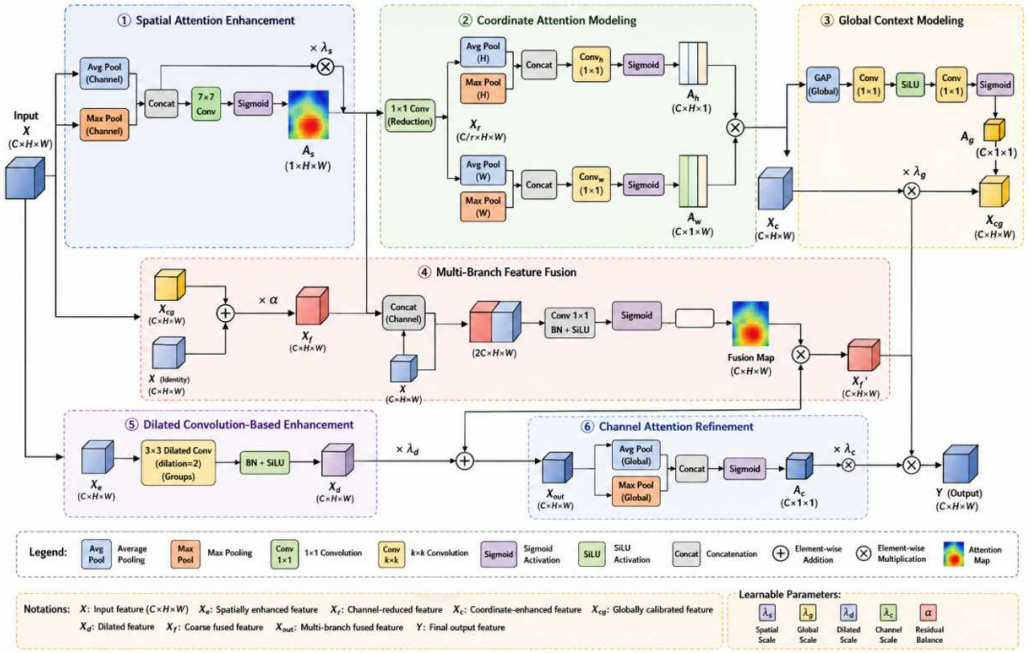


Figure 3. Detailed architecture of the Multi-Branch Enhanced Coordinate Attention (MBECA) module.

To enhance feature representation for small and densely distributed objects in UAV aerial imagery, a Multi-Branch Enhanced Coordinate Attention (MBECA) mechanism is proposed, as illustrated in Fig. 3. Unlike conventional coordinate attention that mainly focuses on directional dependency modeling, MBECA establishes a unified feature interaction framework by jointly exploiting local spatial saliency, directional contextual information, global semantic awareness, and multi-scale contextual aggregation.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, local salient responses are first enhanced through spatial attention:

$$X_e = X \odot (1 + \lambda_s A_s) \tag{6}$$

where A_s denotes the spatial attention map and λ_s is a learnable coefficient.

Subsequently, horizontal and vertical contextual dependencies are jointly modeled to generate coordinate-aware feature representations:

$$X_c = X \odot A_h \odot A_w \tag{7}$$

where A_h and A_w represent directional attention maps along the horizontal and vertical directions, respectively.

To improve semantic consistency under complex UAV scenes, global contextual information is introduced to recalibrate feature responses:

$$X_{cg} = X_c \odot [1 + \lambda_g \sigma(\mathrm{Conv}(\delta(\mathrm{Conv}(\mathrm{GAP}(X_c)))))] \tag{8}$$

where A_g denotes the generated global semantic attention map and λ_g is a learnable scaling coefficient.

Meanwhile, a lightweight dilated convolution branch is employed to capture multi-scale contextual cues:

$$X_{out} = X_{out} + \lambda_d \, \delta\big(\mathrm{BN}\big(\mathrm{Conv}_{3\times3,d=2}(X_e)\big)\big) \tag{9}$$

where $\mathrm{Conv}_{3\times3,d=2}(\cdot)$ denotes a 3×3 dilated convolution with dilation rate $d=2$, $\mathrm{BN}(\cdot)$ denotes Batch Normalization, λ_d is a learnable scaling coefficient, X_e denotes the input feature map, and X_{out} represents the output feature after dilated contextual enhancement.

The complementary information extracted from different feature modeling paths is then integrated through residual-guided fusion:

$$X_f = X_{cg} + \alpha X \tag{10}$$

where α is a learnable balancing coefficient controlling residual information injection.

Finally, channel-adaptive refinement is applied to obtain the output representation:

$$Y = X_{\text{out}} \odot (1 + \lambda_c \sigma(\text{Concat}(\text{GAP}(X_{\text{out}}), \text{GMP}(X_{\text{out}})))) \tag{11}$$

where GMP$(\cdot)$ signify the global max pooling operations, respectively, which are employed to aggregate complementary spatial context. Concat$(\cdot)$ refers to the concatenation operation along the channel dimension. Additionally, λ_c provides final adaptive scaling that adaptively controls the contribution of the generated channel attention map, while the term 1 establishes a residual connection to prevent gradient degradation.

By jointly modeling local saliency, directional dependency, global semantics, and multi-scale contextual information, the proposed MBECA generates more discriminative feature representations for small-object detection in UAV aerial imagery while maintaining computational efficiency.

Region-Adaptive Wise-IoU (RA-WIoU) Loss

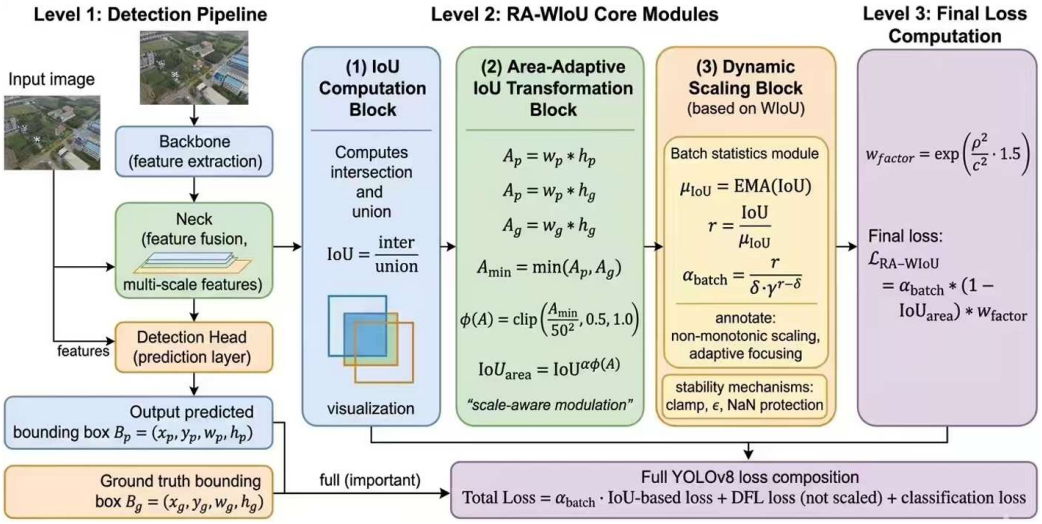


Figure 4. Overview of the Region-Adaptive Wise-IoU (RA-WIoU) Loss Framework.

Accurate localization of small objects remains a challenging problem in UAV aerial imagery. Conventional IoU-based regression losses are inherently scale-invariant and therefore produce similar optimization responses for localization errors occurring in objects of different sizes. However, for tiny objects, even slight localization deviations may cause substantial semantic displacement. Consequently, small targets often receive insufficient optimization during training.

To alleviate this issue, a Region-Adaptive Wise-IoU (RA-WIoU) loss is proposed. Building upon the dynamic focusing mechanism of Wise-IoU, the proposed method introduces an area-adaptive modulation strategy that explicitly incorporates object scale into the regression process. As illustrated in Fig. 4, RA-WIoU enhances localization sensitivity for small objects while maintaining stable optimization behavior for medium- and large-scale targets.

Scale-Adaptive IoU Modulation

Given the predicted bounding box B_p and the ground-truth box B_g, their corresponding areas are denoted as A_p and A_g, respectively. To obtain a stable scale descriptor, the minimum area is adopted:

$$A_{\min} = \min(A_p, A_g) \tag{12}$$

Based on the target scale, an adaptive modulation factor is defined as

$$\phi(A) = \text{clip}\left(\sqrt{\frac{A_{\min}}{A_{\text{ref}}}}, 0.5, 1.0\right) \tag{13}$$

where $\phi(A)$ represents the scale-adaptive modulation factor, $A_{\min}$ denotes the minimum area between the predicted and ground-truth boxes and A_{ref} is the reference object area.

The conventional IoU term is then reformulated as

$$IoU_{\text{area}} = IoU^{\alpha \cdot \phi(A)} \tag{14}$$

where α denotes the modulation coefficient. For small objects, a smaller value of $\phi(A)$ results in a stronger penalty for low-IoU predictions, thereby enhancing localization sensitivity.

Dynamic Focusing and Loss Formulation

To improve robustness against noisy samples and outliers, RA-WIoU inherits the dynamic focusing mechanism of Wise-IoU. Furthermore, a distance-aware weighting term is introduced to improve geometric consistency during optimization:

$$w_{\text{factor}} = \exp\left(\frac{\rho^2}{c^2} \cdot 1.5\right) \tag{15}$$

where ρ denotes the Euclidean distance between the centers of the predicted and ground-truth boxes, and c denotes the diagonal length of the minimum enclosing box.

The final RA-WIoU loss is formulated as:

$$\zeta_{\text{RA-WIoU}} = \alpha_{\text{batch}} \cdot (1 - \text{IoU}_{\text{area}}) \cdot w_{\text{factor}} \tag{16}$$

Within the proposed MSAL-YOLO framework, the dynamic scaling mechanism is applied exclusively to the IoU-based regression component rather than the Distribution Focal Loss (DFL). Therefore, the overall optimization objective is formulated as:

$$\zeta_{\text{Total}} = \zeta_{\text{RA-WIoU}} + \zeta_{\text{DFL}} + \zeta_{\text{cls}} \tag{17}$$

This decoupled design preserves stable distribution learning while enabling adaptive localization optimization for small-object detection.

Experiments and Results

Dataset and Experiment Details

Experiments were conducted using the PyTorch framework on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB). The proposed model was trained for 100 epochs with a batch size of 16 and an input resolution of 640×640. The SGD optimizer was adopted with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 5×10^{-4}. Standard YOLOv8 data augmentation strategies, including Mosaic, MixUp, random horizontal flipping, HSV transformation, scaling, and translation, were applied during training. Unless otherwise specified, all reported results represent the average of three independent runs.

The proposed method was evaluated on two public aerial-image benchmarks, namely VisDrone2019 and DOTAv1. VisDrone2019 contains diverse UAV-captured scenes with significant variations in illumination, weather conditions, object scales, and occlusion levels. Following the official object detection protocol, 6,471 images, 548 images, and 1,610 images were used for training, validation, and testing, respectively. DOTAv1 is a large-scale remote-sensing dataset consisting of 15 object categories and high-resolution aerial images. For the horizontal bounding-box detection task, images were cropped into 1024×1024 patches using a sliding-window strategy with a 200-pixel overlap. The original training and validation sets were merged and re-divided using an 8:2 ratio for evaluation.

Detection performance was assessed using Precision (P), Recall (R), mean Average Precision at IoU thresholds of 0.5 and 0.5:0.95 (mAP@0.5 and mAP@0.5:0.95), model parameters (Params), floating-point operations (GFLOPs), and inference speed measured in frames per second (FPS). These metrics jointly evaluate detection accuracy, localization quality, model complexity, and deployment efficiency.

Ablation study

To evaluate the contribution of each proposed component, ablation experiments were conducted on the YOLOv8n baseline. SCMConv, MBECA, and RA-WIoU were progressively introduced, and the results are summarized in Table 1. All experiments were repeated three times using different random seeds, and the reported values correspond to the average results.

As shown in Table 1, all three components contribute positively to detection performance. Among the individual modules, SCMConv provides the largest gain, improving mAP@0.5 from 30.0% to 33.8% and mAP@0.5:0.95 from 17.0% to 19.5%.

Table 1. Ablation study of the proposed components on VisDrone2019.

SCMConv	MBECA	RA-WIoU	AP_{50}	$AP_{50:95}$	GFLOPs	Params (M)
			0.300	0.170	8.1	3.0
✓			0.338	0.195	10.9	3.9
	✓		0.309	0.177	8.8	3.1
		✓	0.310	0.176	8.2	3.0
✓	✓		0.345	0.201	11.2	4.0
✓		✓	0.355	0.204	11.0	3.9
	✓	✓	0.322	0.182	8.6	3.1
✓	✓	✓	**0.357**	**0.205**	12.4	4.0

MBECA and RA-WIoU also yield consistent improvements, indicating that both feature enhancement and localization optimization are beneficial for UAV object detection.

Combining multiple components leads to further performance gains. In particular, the integration of SCMConv and RA-WIoU achieves mAP@0.5 of 35.5%, suggesting strong complementarity between feature representation and regression optimization. When all components are jointly employed, the proposed model reaches the best performance, with mAP@0.5 and mAP@0.5:0.95 of 35.7% and 20.5%, respectively.

Although the introduction of additional modules slightly increases computational complexity, the final model contains only 4.0M parameters and 12.4 GFLOPs. These results demonstrate that the proposed framework achieves an effective balance between detection accuracy and computational efficiency, making it suitable for real-time UAV deployment scenarios.

Comparative experiments

To assess the effectiveness and generalization capability of the proposed MSAL-YOLO, comparisons were conducted on the VisDrone2019 and DOTAv1 datasets against representative one-stage, two-stage, and transformer-based detectors. All models were trained and evaluated under identical experimental settings.

Table 2. Performance comparison on the VisDrone2019 dataset.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Params (M)	GFLOPs	FPS
YOLOv5s	0.450	0.355	0.305	0.205	9.1	24.1	120
YOLOv5m	0.478	0.366	0.345	0.216	25.0	64.0	84
YOLOv5l	0.501	0.387	0.412	0.242	53.1	134.7	52
YOLOv7-Tiny	0.451	0.346	0.292	0.136	6.0	13.1	106
YOLOv8n	0.415	0.305	0.300	0.170	3.0	8.2	230
YOLOv8s	0.453	0.353	0.356	0.207	11.1	28.5	176
YOLOv8m	0.486	0.367	0.385	0.228	25.8	78.7	116
Faster R-CNN	0.415	0.297	0.300	–	41.0	203.5	20
SSD	0.316	0.290	0.206	–	4.3	3.1	42
SS-YOLO	–	–	0.307	0.173	4.6	13.3	198
RT-DETR	0.469	0.366	0.391	0.223	20.0	57.3	82
Lite-DETR	0.462	0.361	0.372	0.214	15.8	42.6	96
UAV-YOLO	0.422	0.300	0.315	0.182	3.6	11.8	210
RetinaNet	–	–	0.162	0.085	36.2	29.7	–
ATSS[44]	–	–	0.332	0.201	28.5	37.8	–
Gold-YOLO-N	0.451	0.349	0.344	0.201	5.6	14.8	192
DAMO-YOLO-N	0.458	0.356	0.352	0.206	6.9	16.5	168
MSAL-YOLO	**0.461**	**0.359**	**0.357**	**0.205**	**4.0**	**12.4**	**154**

Table 2 summarizes the results on VisDrone2019. MSAL-YOLO achieved 35.7% mAP@0.5 and 20.5% mAP@0.5:0.95 with only 4.0M parameters and 12.4 GFLOPs. Compared with the YOLOv8n baseline, mAP@0.5 improved from 30.0% to 35.7%, corresponding to a gain of 5.7 percentage points.

Although larger models such as YOLOv5l and YOLOv8m obtained higher detection accuracy, they required substantially more parameters and computational cost. Compared with lightweight detectors including Gold-YOLO-N, DAMO-YOLO-N, SS-YOLO[45], and UAV-YOLO, the proposed method achieved competitive or superior detection performance while maintaining relatively low model complexity.

Transformer-based detectors, including RT-DETR[46] and Lite-DETR, demonstrated strong detection capability owing to global feature interaction. However, these methods required considerably higher computational resources than MSAL-YOLO.

The results indicate that the proposed architecture provides an effective trade-off between detection accuracy and computational efficiency for UAV object detection.

To further evaluate cross-dataset generalization, experiments were conducted on the DOTAv1 dataset. As shown in Table 3, MSAL-YOLO achieved 55.5% mAP@0.5 and 31.3% mAP@0.5:0.95, outperforming all compared baseline models. Compared with YOLOv8n, the proposed method improved mAP@0.5 by 1.8 percentage points and mAP@0.5:0.95 by 1.9 percentage points.

Table 3. Performance comparison on the DOTAv1 dataset.

Method	mAP@0.5	mAP@0.5:0.95
YOLOv8n	0.537	0.294
YOLOv7-Tiny	0.458	0.271
YOLOv5s	0.545	0.302
SSD	0.412	0.256
Faster R-CNN	0.368	0.203
R-FCN	0.366	0.204
DETR	0.403	0.246
MSAL-YOLO	**0.555**	**0.313**

The consistent improvements observed on both VisDrone2019 and DOTAv1 suggest that the proposed SCMConv, MBECA, and RA-WIoU components contribute to improved feature representation and localization performance across different aerial imaging scenarios.

Category-wise performance analysis

Category-wise evaluation on VisDrone2019 demonstrated consistent improvements across all object categories (Supplementary Table S1). The largest gains were obtained for *Bus* (+12.7%), *Van* (+7.6%), and *Truck* (+6.8%), while notable improvements were also observed for densely distributed categories such as *Pedestrian* (+6.8%) and *Motor* (+5.6%). In addition, small-object categories, including *Bicycle*, *Tricycle*, and *Awning-tricycle*, benefited from the proposed scale-aware optimization strategy. These results indicate that MSAL-YOLO improves detection performance across a wide range of object scales and scene complexities.

Qualitative analysis and visualization results

Confusion matrix analysis

To further assess category-level recognition performance, the confusion matrices of YOLOv8n and the proposed MSAL-YOLO are provided in Supplementary Fig. S2. Compared with the baseline model, MSAL-YOLO exhibits stronger diagonal responses and fewer off-diagonal errors across most categories, indicating improved classification accuracy and reduced inter-class confusion.

The improvement is particularly evident for categories containing small and densely distributed objects, where missed detections and category ambiguities are common. These results suggest that the proposed framework enhances feature discriminability and localization consistency under challenging UAV imaging conditions. Overall, the confusion matrix analysis supports the quantitative results reported in previous sections and further demonstrates the effectiveness of MSAL-YOLO for small-object detection in aerial imagery.

The improvements are particularly evident for categories characterized by dense distributions and large scale variations, such as pedestrian, motor, and truck. These results suggest that the proposed framework enhances feature discriminability and improves the separation of visually similar categories under complex aerial imaging conditions.

Overall, the confusion matrix analysis demonstrates that MSAL-YOLO effectively reduces both false-positive and false-negative predictions, resulting in more reliable object recognition performance.

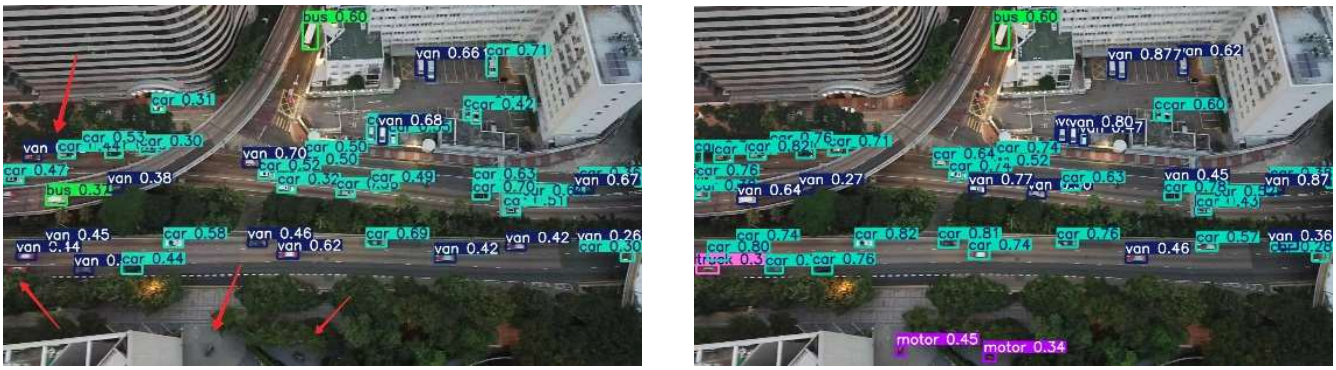
Training dynamics

Fig. S3 illustrates the training and validation curves of YOLOv8n and MSAL-YOLO. The proposed model converges more rapidly and exhibits smoother optimization behavior throughout training.

The regression, classification, and distribution focal losses decrease steadily without noticeable oscillations, while the validation metrics closely follow the training trends. Furthermore, Precision, Recall, mAP$*$50, and mAP$*50-95$ increase consistently and reach stable values earlier than the baseline model.

These observations indicate that the proposed feature enhancement and regression optimization strategies improve both convergence stability and generalization capability.

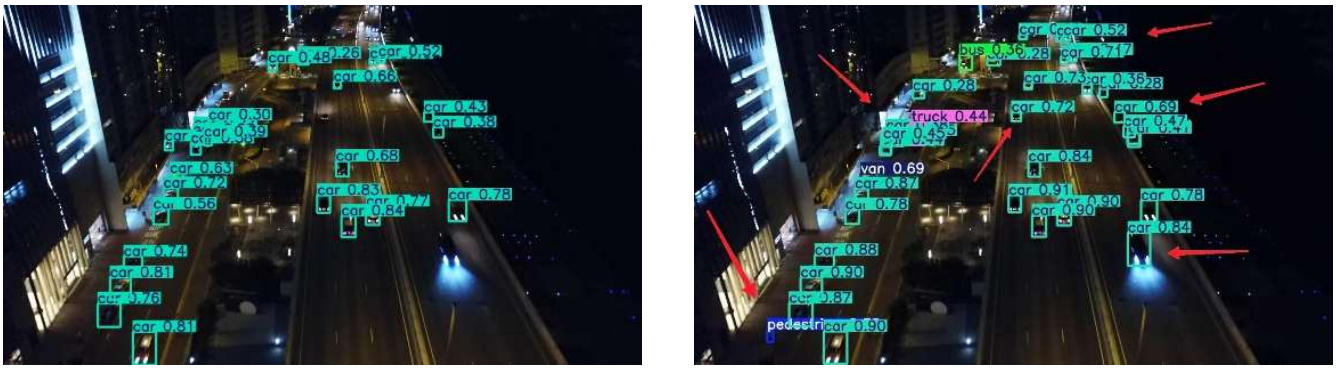
Visual detection results



(a) YOLOv8n **(b)** MSAL-YOLO

Figure 5. Comparison of detection results in dense small-object scenarios.

Fig. 5 compares the detection results of YOLOv8n and MSAL-YOLO in dense small-object scenes. The proposed method detects more targets and exhibits fewer missed detections than the baseline model. The improvement is particularly pronounced in crowded regions containing closely spaced objects. These results suggest that the proposed feature enhancement strategy improves feature discrimination and object separation, thereby facilitating more reliable detection in dense aerial scenarios.



(a) YOLOv8n **(b)** MSAL-YOLO

Figure 6. Comparison of detection results in complex background.

Figure 6 shows detection results in complex-background scenarios. Compared with YOLOv8n, MSAL-YOLO generates fewer false detections and achieves more accurate localization. The performance gain is most apparent in cluttered regions with strong background interference, suggesting improved feature discrimination and robustness in challenging UAV imagery.

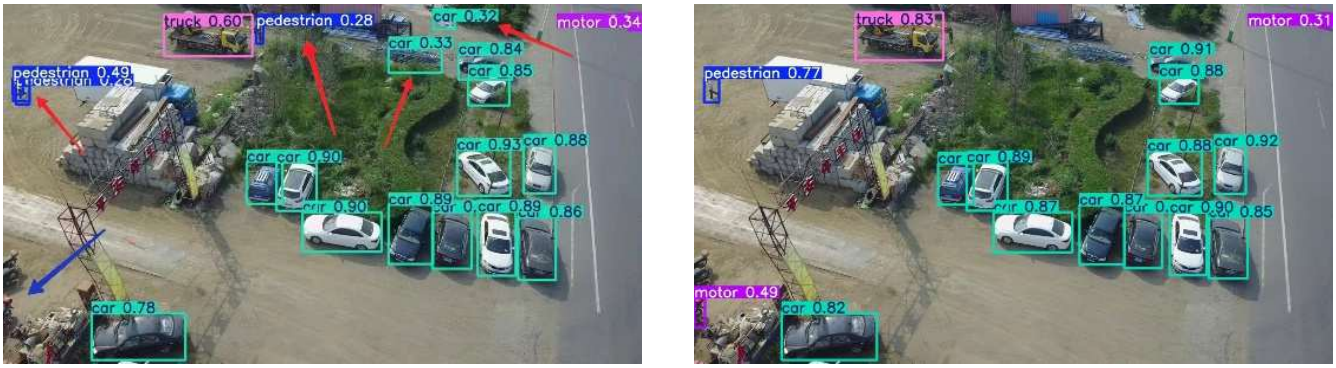
Figure 7 shows representative detection results in occluded and overlapping-object scenarios. Compared with YOLOv8n, MSAL-YOLO detects a larger number of partially visible objects and exhibits fewer missed detections. More accurate localization is also observed in crowded regions, indicating improved robustness under challenging occlusion conditions.

Figure 8 presents detection results for extremely small targets. Compared with YOLOv8n, MSAL-YOLO identifies more small objects and reduces missed detections, particularly for distant targets with limited visual information. These results further demonstrate the effectiveness of the proposed framework for small-object detection in UAV aerial imagery.

Limitations and failure cases

Despite the overall improvements, several challenges remain under extreme conditions. As illustrated in Supplementary Fig. S4, performance degradation may still occur when severe occlusion, extremely dense object distributions, and strong background similarity appear simultaneously. In such cases, object boundaries become ambiguous and the visual characteristics of neighboring targets are difficult to distinguish, leading to occasional missed detections and localization errors.

In addition, the model has not yet been extensively evaluated under severe domain-shift conditions, including variations in sensor characteristics, imaging altitude, adverse weather, motion blur, and image compression artifacts. These factors may affect detection robustness in real-world deployment scenarios.



(a) YOLOv8n **(b)** MSAL-YOLO

Figure 7. Detection efficacy under challenging conditions of occlusion and target overlap.



(a) YOLOv8n **(b)** MSAL-YOLO

Figure 8. Visualization of detection performance for extremely small-scale targets.

Conclusion

This study proposed MSAL-YOLO, a YOLOv8-based framework for object detection in UAV aerial imagery. The proposed method improves the detection of small, densely distributed, and partially occluded objects through enhanced feature representation and localization optimization.

Experiments conducted on the VisDrone2019 and DOTAv1 datasets demonstrated consistent improvements in detection performance. Comparative experiments and ablation studies verified the contribution of the proposed modules, while qualitative analyses further demonstrated the robustness of the method under complex aerial scenes.

Future work will focus on adaptive feature learning and more robust optimization strategies to further improve generalization across diverse imaging conditions. The proposed framework contributes to the development of UAV object detection methods and may facilitate future research in aerial vision applications.

References

1. Liu, L. *et al.* Deep learning for generic object detection: A survey. *Int. journal computer vision* **128**, 261–318 (2020).
2. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. neural information processing systems* **28** (2015).
3. Liu, W. *et al.* Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37 (Springer, 2016).
4. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).
5. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

6. Wang, G. *et al.* Uav-yolov8: A small-object-detection model based on improved yolov8 for uav aerial photography scenarios. *Sensors* **23**, 7190 (2023).

7. Hua, W. & Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **58**, 162 (2025).

8. Du, D. *et al.* Visdrone-sot2019: The vision meets drone single object tracking challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0 (2019).

9. Xia, G.-S. *et al.* Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983 (2018).

10. Cai, X. *et al.* Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27706–27716 (2024).

11. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).

12. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768 (2018).

13. Tan, M., Pang, R. & Le, Q. V. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790 (2020).

14. Li, F. *et al.* Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18558–18567 (2023).

15. Chen, J. *et al.* Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12021–12031 (2023).

16. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).

17. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803 (2018).

18. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13713–13722 (2021).

19. Yang, L., Zhang, R.-Y., Li, L. & Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, 11863–11874 (PMLR, 2021).

20. Zhu, L., Wang, X., Ke, Z., Zhang, W. & Lau, R. W. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10323–10333 (2023).

21. Cai, H., Li, J., Hu, M., Gan, C. & Han, S. Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756* (2022).

22. Dai, J. *et al.* Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773 (2017).

23. Zhu, X. *et al.* Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

24. Xiong, Y. *et al.* Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5652–5661 (2024).

25. Han, K. *et al.* Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1580–1589 (2020).

26. Ma, N., Zhang, X., Zheng, H.-T. & Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131 (2018).

27. Wang, C. *et al.* Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *Adv. neural information processing systems* **36**, 51094–51112 (2023).

28. Xu, X. *et al.* Damo-yolo: A report on real-time object detection design. *arXiv preprint arXiv:2211.15444* (2022).

29. Rezatofighi, H. *et al.* Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666 (2019).

30. Zheng, Z. *et al.* Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 12993–13000 (2020).

31. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).

32. Tong, Z., Chen, Y., Xu, Z. & Yu, R. Wise-iou: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051* (2023).

33. Gevorgyan, Z. Siou loss: More powerful learning for bounding box regression. *arXiv preprint arXiv:2205.12740* (2022).

34. Zhang, Y.-F. *et al.* Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing* **506**, 146–157 (2022).

35. He, J. *et al.* Alpha-iou: A family of power intersection over union losses for bounding box regression. *Adv. Neural Inf. Process. Syst.* **34**, 20230–20242 (2021).

36. Ma, S. & Xu, Y. Mpdiou: a loss for efficient and accurate bounding box regression. *arXiv preprint arXiv:2307.07662* (2023).

37. Jocher, G. *et al.* ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. *Zenodo* (2021).

38. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475 (2023).

39. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021).

40. Carion, N. *et al.* End-to-end object detection with transformers. In *European conference on computer vision*, 213–229 (Springer, 2020).

41. Wang, C.-Y., Yeh, I.-H. & Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, 1–21 (Springer, 2024).

42. Li, M., Chen, Y., Zhang, T. & Huang, W. Ta-yolo: a lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images. *Complex & Intell. Syst.* **10**, 5459–5473 (2024).

43. Li, M. *et al.* Dea-yolov9: A lightweight method for the photovoltaic module defect detection with multi-scale and self-attention mechanism. In *2024 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, 1–6 (IEEE, 2024).

44. Zhang, S., Chi, C., Yao, Y., Lei, Z. & Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768 (2020).

45. Yang, N. *et al.* Ss-yolo: A lightweight deep learning model focused on side-scan sonar target detection. *J. Mar. Sci. Eng.* **13**, 66 (2025).

46. Zhao, Y. *et al.* Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16965–16974 (2024).

Author contributions

F.D., X.D., H.Z. provided materials, laboratory tools and facilities. F.D., X.D. planned and supervised the study.H.Z, H.L. assisted with the model analysis and results. H.L., L.H. carried out the simulation model construction, the implementation of the experiment, experimental data analysis, image graphics, manuscript writing and preparation. L.H. assisted with the literature survey. All authors read and commented on the manuscript.

Funding

This work was supported by Open Project of Anhui Engineering Research Center for Intelligent Computing and Information Innovation Application (ICII202401) and Key Project of Scientific Research in Institutions of Higher Education of Anhui Province (2024AH051456).

Data availability

The datasets used in this study are publicly available. The VisDrone2019 dataset can be obtained from the official VisDrone benchmark website. The DOTA-v1.0 dataset is available from the official DOTA benchmark website.

Code availability

The source code supporting the findings of this study is available from the corresponding author upon reasonable request.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf