



Article

Toward Efficient UAV-Based Small Object Detection: A Lightweight Network with Enhanced Feature Fusion

Xingyu Di ¹, Kangning Cui ² and Rui-Feng Wang ^{3,4,*}

¹ School of Communication and Artificial Intelligence, School of Integrated Circuits, Nanjing Institute of Technology, Nanjing 211167, China; x00201221511@njit.edu.cn

² Department of Mathematics, City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong; kangnecui2-c@my.cityu.edu.hk

³ College of Engineering, China Agricultural University, 17 Qinghua East Road, Haidian, Beijing 100083, China

⁴ National Innovation Center for Digital Fishery, China Agricultural University, Beijing 100083, China

* Correspondence: sweefongreggiewong@cau.edu.cn

Abstract

UAV-based small target detection is crucial in environmental monitoring, circuit detection, and related applications. However, UAV images often face challenges such as significant scale variation, dense small targets, high inter-class similarity, and intra-class diversity, which can lead to missed detections, thus reducing performance. To solve these problems, this study proposes a lightweight and high-precision model UAV-YOLO based on YOLOv8s. Firstly, a double separation convolution (DSC) module is designed to replace the Bottleneck structure in the C2f module with deep separable convolution and point-by-point convolution fusion, which can reduce the model parameters and calculation complexity while enhancing feature expression. Secondly, a new SPPL module is proposed, which combines spatial pyramid pooling rapid fusion (SPPF) with long-distance dependency modeling (LSKA) to improve the robustness of the model to multi-scale targets through cross-level feature association. Then, DyHead is used to replace the original detector head, and the discrimination ability of small targets in complex background is enhanced by adaptive weight allocation and cross-scale feature optimization fusion. Finally, the WIPIoU loss function is proposed, which integrates the advantages of Wise-IoU, MPDIoU and Inner-IoU, and incorporates the geometric center of bounding box, aspect ratio and overlap degree into a unified measure to improve the localization accuracy of small targets and accelerate the convergence. The experimental results on the VisDrone2019 dataset showed that compared to YOLOv8s, UAV-YOLO achieved an 8.9% improvement in the recall of mAP@0.5 and 6.8%, while the parameters and calculations were reduced by 23.4% and 40.7%, respectively. Additional evaluations of the DIOR, RSOD, and NWPU VHR-10 datasets demonstrate the generalization capability of the model.

Keywords: YOLO; small object detection; UAV aerial photography; lightweight model



Academic Editors: Samsung Lim and Badal Pokharel

Received: 18 April 2025

Revised: 19 June 2025

Accepted: 27 June 2025

Published: 29 June 2025

Citation: Di, X.; Cui, K.; Wang, R.-F. Toward Efficient UAV-Based Small Object Detection: A Lightweight Network with Enhanced Feature Fusion. *Remote Sens.* **2025**, *17*, 2235. <https://doi.org/10.3390/rs17132235>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Unmanned Aerial Vehicles (UAVs) are autonomous platforms controlled via remote commands or preprogrammed instructions [1]. Due to their compact size, ease of operation, and high mobility, UAVs are increasingly deployed across military reconnaissance, urban planning, agriculture, and disaster management [2–12]. However, the aerial perspective of UAV imaging introduces challenges such as small object scale, blurred boundaries, and severe occlusions, leading to high false and missed detection rates [13]. These issues

significantly reduce the performance of object detection algorithms in real-time, high-precision applications. Furthermore, due to limited onboard computing and memory, deploying heavy detection models on UAVs remains impractical [14].

Deep learning has accelerated progress in object detection, driving innovation across computer vision and autonomous systems [15–19]. Existing methods are broadly categorized into two-stage and one-stage detectors. Two-stage algorithms, such as R-CNN and Fast R-CNN [20], first generate region proposals and then classify and regress object locations. Cui et al. [21] improved Faster R-CNN with deformable convolutions, hybrid attention, and Soft-NMS to enhance UAV detection accuracy. Butler et al. [22] extended Mask R-CNN with a non-volume integral branch to address scale variations, while Zhang et al. [23] proposed MS-FRCNN with ResNet50, FPNs (Feature Pyramid Networks), and attention modules for improved forest fire small target detection. Despite their high accuracy, the complexity of two-stage pipelines limits their applicability in real-time UAV operations.

One-stage detectors such as RetinaNet [24], CenterNet [25], SSD [26], and especially the YOLO series [27], provide faster inference and have attracted significant research attention. Cao et al. [28] proposed YOLOv5s-MSES with a small-target layer and multi-scale attention fusion to reduce false detections. Duan et al. [29] enhanced YOLOv8s by introducing a small-object detection head, Inner-WIoU loss, and cross-spatial attention for better feature aggregation. Luo et al. [30] developed YOLOD with HardSwish/Mish activation, EIou loss, and adaptive spatial fusion for UAV images. Zhang et al. [31] introduced HSP-YOLOv8, improving mAP@0.5 by 11% on the Visdrone2019 dataset through structural optimization. Another study by Luo et al. [32] proposed asymmetric convolution modules, enhanced attention, and EIou-NMS, achieving strong performance across multiple datasets.

Although deep learning-based target detection technology has made significant progress, existing methods still face many severe challenges in the scenario of small target detection in UAV aerial images, which severely restrict the improvement of detection performance.

Small target detection is one of the core challenges of UAV aerial image detection. Since drones usually shoot at higher altitudes, ground objects occupy very few pixels in the image, often only a few to a dozen pixels. This makes the object lack sufficient spatial information and texture details in the low-level feature map, which makes it difficult for the detection model to accurately extract and distinguish different categories of small objects. Shang et al. [33] improved the detection performance of small targets in aerial images by applying a small target detection layer to the YOLOv5 model. Zhang et al. [34] improved the YOLOv8n model and enhanced the feature fusion ability by introducing a bidirectional feature pyramid network (BiFPN), achieving more accurate recognition and positioning of small target objects. Qin et al. [35] optimized the FPN structure of the YOLOv7 model and added a small target detection layer to enhance the network's ability to detect small targets.

Fuzzy boundaries and severe occlusion problems are also common in UAV aerial image target detection tasks. On the one hand, the high-altitude overlook angle and weather factors may cause the edges of objects to be blurred, which increases the difficulty of accurately locating the bounding box. On the other hand, in dense scenes, it is very common for objects to block each other, and only a small part of some targets may be exposed, which makes them difficult for the model to reliably identify and locate based on the limited visible area. Wang et al. [36] improved the RTDETR model by combining the HiLo attention mechanism with the in-scale feature interaction module and integrating it into the hybrid encoder to enhance the focus of the model on dense targets and reduce the missed detection rate and false detection rate. Chang et al. [37] improved the YOLOv5s model by adding a coordinated attention mechanism after the convolution operation to improve the model detection accuracy of small targets under image blurring.

Complex background and similarity between classes also make detection difficult. UAV aerial images usually contain broad scenes with rich and complex background elements. At the same time, different categories of targets may have similarities in appearance, shape or color, which makes the model prone to confusion, especially when the target size is small, and this distinction is more difficult. Xiong et al. [38] optimized the spatial attention module in the YOLOv5 model to enhance the representation of small target features while suppressing background interference. Yang et al. [39] used the EMA attention mechanism in YOLOv8 to construct a new C2f-E module in the feature extraction network, grouped channel dimensions into multiple sub-features, and reshaped some channel dimensions into batch dimensions to maximize the feature information of small targets. Zhang et al. [40] adopted two-layer routing attention (BRA) in the feature extraction stage of the YOLOv10 model to effectively reduce background interference.

While these efforts have advanced small object detection from UAV aerial images, challenges remain in terms of efficient cross-scale feature fusion and keeping the model lightweight. Many existing models improve accuracy by integrating complex modules or advanced feature fusion strategies, but this often leads to a large increase in computational complexity and model size, making it incompatible with the limited airborne resources of UAVs. Conversely, attempts at model compression sometimes sacrifice detection accuracy, especially for small objects in complex scenes. Meanwhile, YOLOv8s, while effective in real-time detection, still struggles to handle complex scenarios involving multi-scale variability and small object localization. To address these issues, this paper proposes an improvement of YOLOv8s for UAVs through four main contributions:

- The C2f_DSC design: By introducing deep separable convolution and point-by-point convolution to build a double separation convolution block (DSC), and using DSC to replace the Bottleneck structure of the C2f module, C2f_DSC can effectively reduce the amount of parameters, calculations, and memory accesses, while enhancing feature expression capabilities and improving inference computation efficiency, especially for UAV aerial photography small target detection tasks.
- The SPPL module: By combining the multi-scale feature extraction capability of SPPF with the long-distance dependency capture and self-attention mechanism of LSKA, the diversity and flexibility of feature extraction are improved, and the model's ability to detect targets at various scales is enhanced, making it more suitable for handling multi-scale target detection tasks.
- The integration of DyHead: Dyhead is used to replace the head part of the original model, which aims to solve the shortcomings of the traditional target detection model in feature fusion and dynamic adjustment. By introducing dynamic weight allocation and adaptive feature fusion mechanism, Dyhead can capture and utilize multi-scale feature information more effectively, which improves the accuracy of target detection, especially for small-sized and complex background targets.
- The WIPIoU loss: Three loss functions of Wise-IoU, MPDIoU, and Inner-IoU are combined, and the WIPIoU loss function is proposed, which improves the regression accuracy of bounding box, enables the model to accurately locate the target, and accelerates the convergence speed of the model.

The rest of this paper is arranged as follows: Section 2 describes the technical details of network model improvement; Section 3 shows and analyzes the experimental results; finally, Section 4 summarizes the research work.

2. Methods

2.1. YOLOv8 Model

In 2023, the Ultralytics team released YOLOv8, a major advancement in the YOLO object detection series that demonstrated SOTA performance and received significant attention from both academia and industry. As an improved version based on YOLOv5, YOLOv8 adopts a more flexible and modular architecture, and provides five distinct model variants: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x [41]. These models are designed to meet different requirements for computational resources and detection accuracy by adjusting the depth, width, and channel dimensions of the network. Among them, YOLOv8n is the most lightweight, with approximately 3.0 million parameters and 8.1 GFLOPs, while YOLOv8x is the most powerful, with about 68.1 million parameters and 257.4 GFLOPs. In general, larger models achieve higher accuracy but require more computational resources and have slower inference speeds, whereas smaller models are more suitable for real-time detection and deployment on resource-limited platforms.

The YOLOv8 architecture comprises four main components: the input layer, backbone, neck, and head. As shown in Figure 1, the input layer performs adaptive image scaling and applies Mosaic data augmentation to enhance input diversity and model robustness. The backbone includes CBL, C2f, and SPPF modules. The CBL module integrates convolution, batch normalization, and activation to improve training stability. The C2f module incorporates CSPNet and the E-ELAN structure from YOLOv7, achieving a lightweight design and faster inference [42]. The SPPF module enhances multi-scale feature extraction through pooling operations. The neck adopts a combination of feature pyramid network (FPN) and Path Aggregation Network (PAN), enabling top-down and bottom-up cross-layer feature fusion to strengthen feature representation [43]. The head employs a decoupled design, separating classification and localization into distinct branches before final fusion. This structure reduces computational complexity while improving detection accuracy and generalization. In this study, YOLOv8s is selected as the baseline due to its balance between model size and performance, making it suitable for deployment in UAV-based small target detection scenarios.

2.2. UAV-YOLO Model

In actual UAV small target detection scenarios, due to the huge differences in target sizes, changeable shapes, and complex environmental background interference, the requirements for the multi-scale feature expression ability and anti-interference ability of the target detection model are extremely high. At the same time, the real-time performance of UAV small target detection is crucial for applications such as disaster monitoring and security inspection, so the robustness and detection accuracy of the model cannot be ignored. At present, related research mainly focuses on optimizing multi-scale feature expression, improving target positioning accuracy in complex scenes, and balancing model efficiency and detection performance. To solve these problems, we propose an improved UAV small target detection network UAV-YOLO, which is based on the YOLOv8s architecture and is specially designed for small target detection tasks in UAV images.

As shown in Figure 2, the UAV-YOLO network is optimized on the basis of YOLOv8s. Its depth feature extraction network uses a lightweight C2f_DSC module to replace the traditional convolution structure and uses depth separable convolution to abstract the edges and textures of the input image. Shallow spatial features are abstracted layer by layer, and feature maps with three resolutions of $80 \times 80 \times 64$, $40 \times 40 \times 128$, and $20 \times 20 \times 256$ are generated while greatly reducing the amount of parameters and calculations. Subsequently, these multi-scale features are input into the feature pyramid network integrating SPPL modules. Through the synergy of spatial pyramid pooling and LSKA attention mechanism,

cross-hierarchical long-distance dependencies are established and small target feature responses are strengthened. Furthermore, the WIPIoU loss function is used for regression optimization, and the geometric center, aspect ratio, and overlap degree of the bounding box are included in a unified measure, which effectively improves the positioning accuracy of small targets and the convergence speed of the model. Finally, the DyHead structure is introduced to realize dynamic weight allocation and feature calibration. By adaptively adjusting the fusion weights of different levels of features, the detection performance of small targets in complex backgrounds is significantly enhanced while maintaining scale invariance.

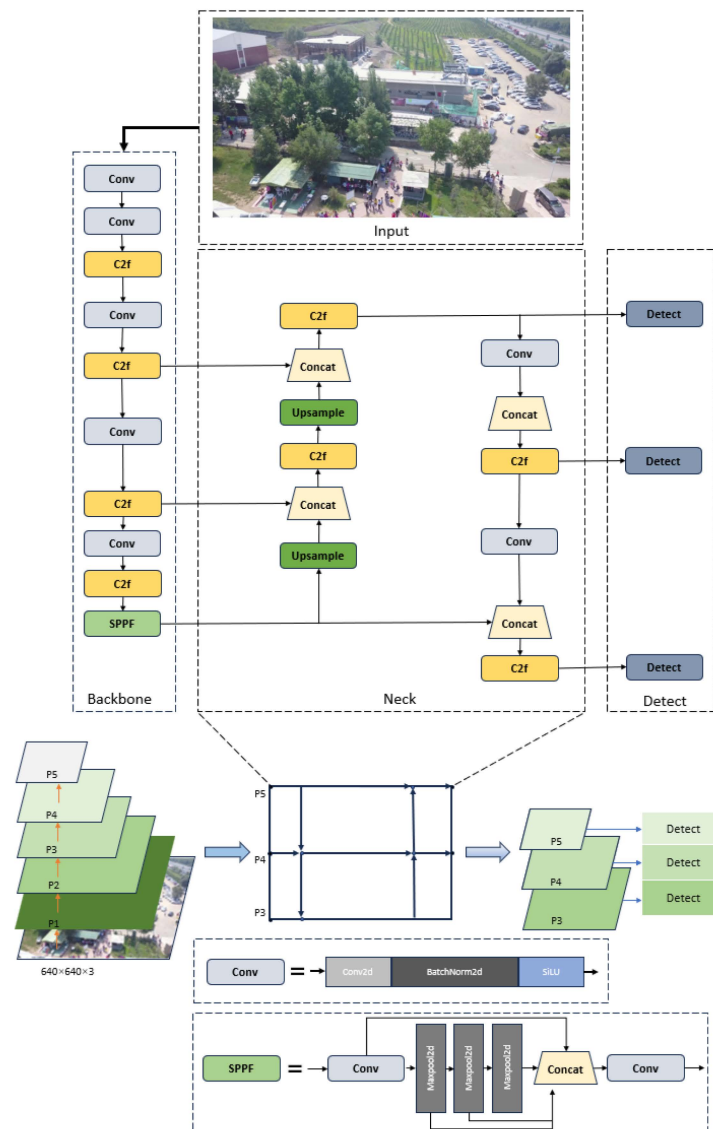


Figure 1. YOLOv8 model structure.

2.2.1. Model Lightweight Improvement

In the small target detection task of UAV aerial images, existing models are often limited due to computational redundancy, a huge amount of parameters and high memory access overhead, and it is difficult to efficiently extract subtle features, which affects the accuracy of small target detection. Therefore, a lightweight backbone network with strong feature extraction capabilities is urgently needed. In response to this challenge, this study proposes a lightweight module called C2f_DSC. Its core design incorporates the idea of Depthwise Separable Convolution into the C2f module. Through the improved Depthwise

convolution operation, only the front or back quarter of the input channel is selectively extracted. The features keep the remaining channels unchanged, greatly reducing the amount of calculation. Then, Pointwise Convolution is applied to integrate channel information and enhance the expressive power of the model. Depthwise convolution and Pointwise Convolution are shown in Figure 3. While inheriting the advantages of C2f parallel connection, the C2f_DSC module significantly reduces calculation and parameter costs and focuses more on key fine-grained features, achieving dual improvements in model lightweight and feature extraction capabilities, allowing the backbone network to generate more discriminant feature maps while keeping the model lightweight. Thus, it significantly improves the performance of the model in small target detection scenarios in UAV aerial images.

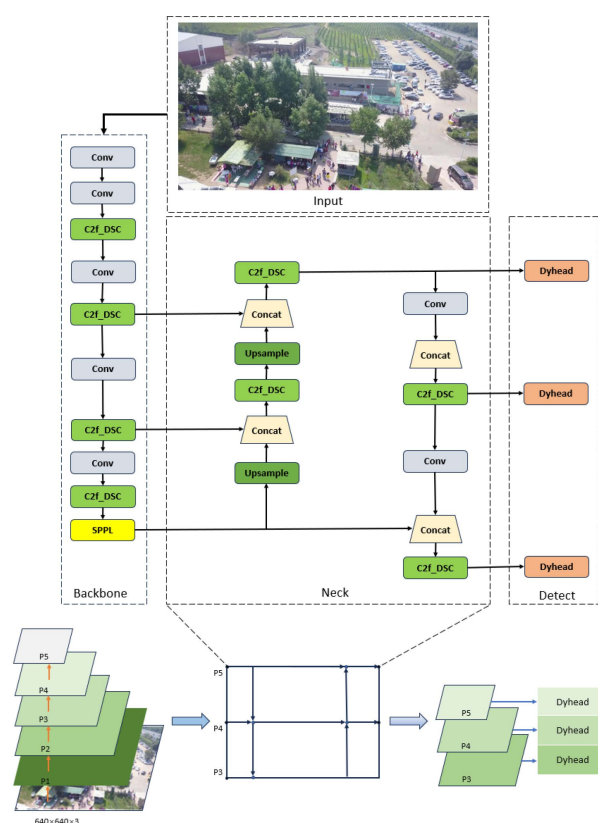


Figure 2. UAV-YOLO model structure.

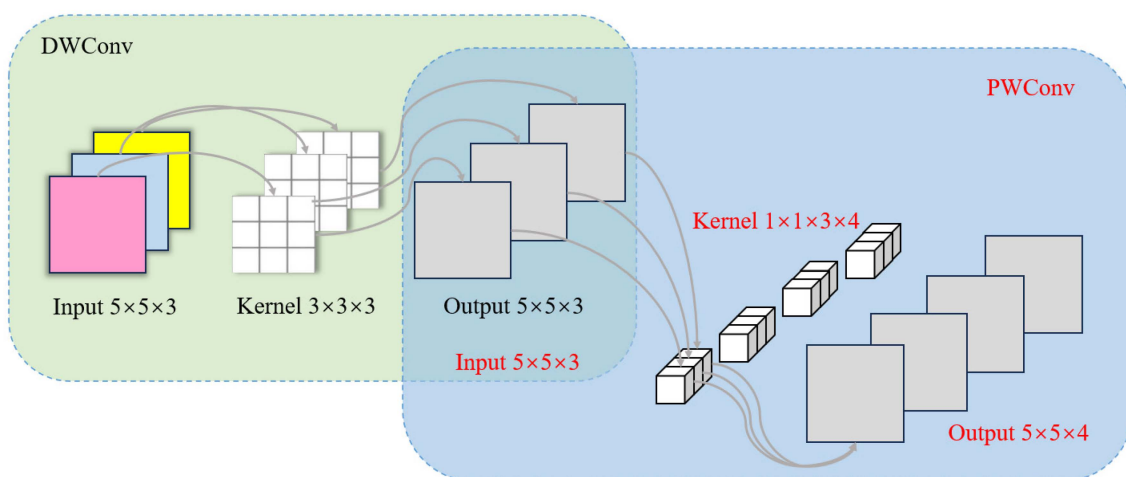


Figure 3. DWConv and PWConv.

Based on DWConv [44] and PWConv, a DSC is designed (Figure 4a), integrating two DWConvs and two PWConvs within an inverted residual structure. The intermediate layer expands channel dimensions to facilitate gradient propagation and feature reuse via shortcut connections. A BatchNorm layer and SiLU activation follow the second PWConv to enhance feature diversity and inference speed. The Feature Extraction Block (FEB) employs this configuration to effectively extract salient features while minimizing parameters and computational cost, thereby improving overall inference efficiency.

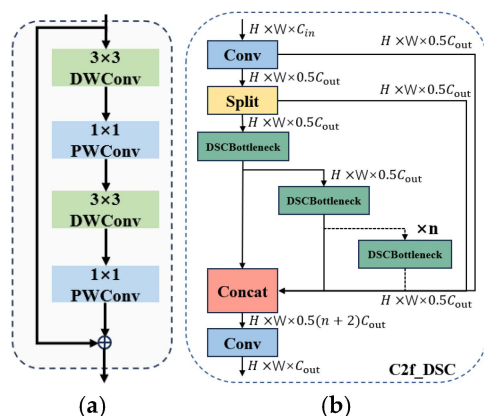


Figure 4. (a) Dual-separable convolution block; (b) C2f_DSC structure diagram.

The C2f_DSC module, illustrated in Figure 4b, replaces the Bottleneck structure in the original C2f module with DSC within the backbone network. A 1×1 convolution is first applied to adjust channel dimensions and enhance feature representation. The input is then split: one part forms a residual connection to retain essential features, while the other undergoes deep feature extraction through n serially connected DSC blocks. This design maintains residual links to support gradient propagation and reduce parameter and computational overhead. Subsequently, multiple parallel branches are fused via a Concat operation, integrating diverse gradient information to enrich feature representations. A final 1×1 convolution yields the output. Compared to the original C2f, the C2f_DSC module significantly lowers parameter count, computational complexity, and memory access, while preserving detection accuracy, thereby enhancing inference speed and supporting efficient small object detection in UAV imagery.

2.2.2. SPPL Module

In the small target detection task of UAV aerial images, effective feature extraction and fusion is the key to achieving high accuracy. However, although existing methods perform well under certain conditions, they often struggle to cope with the inherent scale changes of UAV images, resulting in missed detection or false detection. Traditional single-scale feature extraction cannot fully capture multi-scale targets. At the same time, the discrete spatial distribution characteristics of objects in aerial images are also a big challenge for conventional convolutional neural networks, and it is difficult to establish long-distance dependencies with limited receptive fields. In response to these challenges, this study proposes the SPPL module, which innovatively integrates Spatial Pyramid Pooling Rapid Fusion (SPPF) with the Large Kernel Attention Mechanism (LSKA) [45]. First, a parallel multi-branch architecture is adopted, which incorporates deformable convolution to accurately locate irregularly shaped targets and uses a dual-channel spatial attention mechanism to achieve adaptive weighting of features. Through this design that combines multi-scale feature fusion and dynamic attention allocation, the SPPL module significantly enhances the diversity of feature expression and scene adaptability, thereby

improving the robustness and accuracy of detection. Figure 5 illustrates the architectures of the SPPF and SPPL modules.

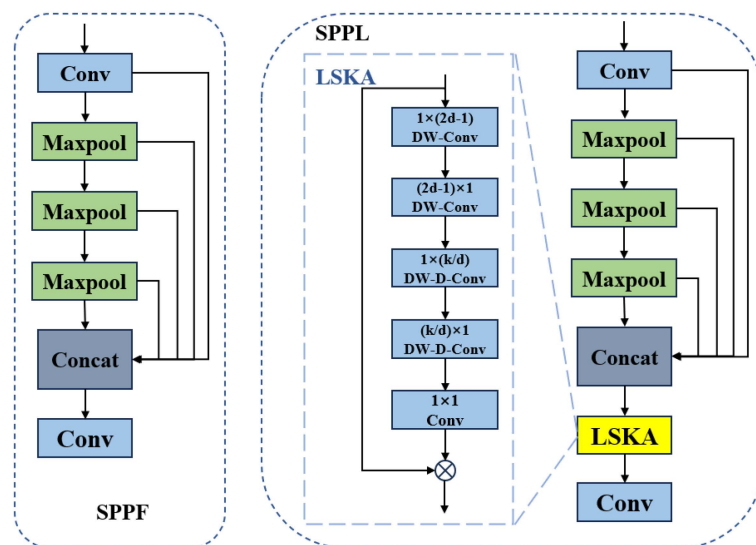


Figure 5. SPPF and SPPL modules.

The SPPF module enhances scale adaptability by applying multi-scale pooling operations to capture features across different resolutions. The core of the LSKA mechanism lies in decomposing a conventional 2D convolution into sequential horizontal and vertical 1D convolutions, enabling efficient directional feature extraction and the generation of an initial attention map that highlights salient regions. LSKA further employs spatially dilated convolutions with varying dilation rates to expand the receptive field and capture broader contextual information without increasing computational costs. The final attention map, produced through additional convolutional layers, is element-wise multiplied with the input feature map to emphasize critical features while suppressing irrelevant ones. By leveraging large kernels, separable convolutions, and dilation, LSKA effectively models long-range dependencies and captures multi-scale target features. This mechanism improves the model's capacity to associate distant features within UAV imagery. Integrating SPPF and LSKA, the SPPL module performs deep feature-level fusion, combining multi-scale representation with long-range attention, thereby significantly enhancing small object detection performance.

2.2.3. Dyhead Attention Detection Head

In the small target detection task of UAV aerial images, although the YOLOv8s model adopts a three-scale prediction structure to deal with targets of different sizes, the lack of effective global context integration among various scales in its prediction stage makes it difficult for the model to make full use of cross-scale semantic association. This limitation is particularly prominent in complex scenarios, which seriously restricts its ability to accurately detect small targets. Aiming at this core problem, this study uses DyHead module to replace the standard detection head. DyHead introduces a dynamic routing mechanism that can adaptively adjust the weights of different feature layers, thus facilitating more efficient multi-scale feature extraction [46]. At the same time, it integrates scale-aware, spatial, and task-aware attention mechanisms by decoupling the attention function into three consecutive operations, each focusing on a specific aspect of feature representation. Through this design, the DyHead module breaks down the barriers of independent work of each scale in traditional multi-scale processing, enhances the model's utilization of global context and multi-scale semantic relationships, and significantly improves the detection of

small targets in complex UAV scenarios. accuracy and robustness. For the feature tensor $F \in R^{L \times S \times C}$, the Dyhead module can be expressed as follows:

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \quad (1)$$

where π_L is the scale attention; π_S is spatial attention; and π_C is perceived attention.

Scale-perceived attention can dynamically fuse features according to the semantic importance of different scales. The calculation formula can be expressed as follows:

$$\pi_L(F) \cdot F = \sigma \left(f \left(\frac{1}{SC} \sum_{S,C} F \right) \right) \cdot F \quad (2)$$

where $f(\cdot)$ denotes a linear function approximated by a 1×1 convolutional layer; $\sigma(x) = \max(0, \min(1, (x + 1)/2))$ is the Hard Sigmoid function, S represents the spatial size of the feature map, and C represents the number of channels of the feature map.

Spatial attention makes attention learning sparse through variability convolution, and then aggregates features across layers at the same spatial position. The calculation formula is

$$\pi_S(F) \cdot F = \frac{1}{L} \sum_{l=1}^L \sum_{k=1}^K \omega_{l,k} \cdot F(l; p_k + \Delta p_k; c) \cdot \Delta m_k \quad (3)$$

where L is the level of the pyramid; K is the number of sparse sampling positions; $p_k + \Delta p_k$ is the position after shift, and Δp_k is the spatial offset of autonomous learning, with the purpose of focusing on a discriminant area; Δm_k is an important measure of autonomous learning at position p_k .

Task-aware attention adapts to a variety of tasks by dynamically switching the switches of feature channels, and can summarize different manifestations of objects and realize joint learning. The expression is

$$\pi_C(F) \cdot F = \max(\alpha^1(F) \cdot F_c + \beta^1(F), \alpha^2(F) \cdot F_c + \beta^2(F)) \quad (4)$$

where F_c is the characteristic slice of the c -th channel; $[\alpha^1, \alpha^2, \beta^1, \beta^2] = \theta(\cdot)$ is a hyperfunction controlling the activation threshold; $\theta(\cdot)$ first performs global average pooling on $L \times S$ dimensions to reduce the dimensionality, then goes through 2 fully connected layers and 1 normalized layer, and finally uses the Shifted Sigmoid function to adjust the output to $[-1, 1]$. The structure of Dyhead module is shown in Figure 6. It integrates scale attention, spatial attention and task-aware attention, and can dynamically adjust the weights of different feature layers. The modules can be stacked in series multiple times, and the output can be used for classification, regression, etc.

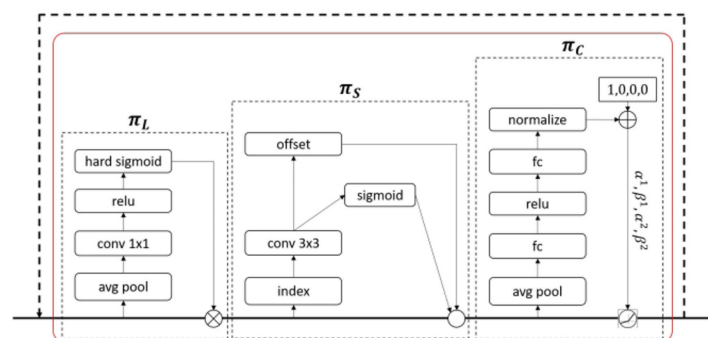


Figure 6. Dyhead module.

The Dyhead module is introduced into the detection head to form the Dyhead dynamic detection head. The three-dimensional tensor deploys an attention mechanism in each dimension, which grants the detection head the ability to dynamically fuse multi-scale features and adaptively fuse multi-level features. It can learn more discriminant representations, and at the same time, it can dynamically fuse pixel-level local features of different FPN output layers to improve the expression and generalization ability of the model.

2.2.4. Loss Function Improvement

In the small target detection task of UAV aerial images, the design of the loss function is crucial to the performance of the algorithm. However, the current YOLOv8s algorithm uses CIOU [47] as the loss function of prediction box regression by default, and its calculation mainly considers three geometric factors: the overlap area between the prediction box and the real box, the distance between the center points, and the aspect ratio. Although CIOU performs well in many scenarios, there are obvious limitations in the CIOU loss function in small target detection tasks of UAV aerial images. First of all, for targets with extremely small scales, small perturbations in their bounding boxes may cause large changes in the overlap area and aspect ratio, making CIOU losses too sensitive to these perturbations, resulting in an unstable training process. Secondly, CIOU's consideration of center point distance may not be able to effectively guide the model in small target scenarios, because the position error of small targets is more significant than its size, and simple center point distance optimization may not be enough to accurately correct its position. Therefore, the CIOU loss function is difficult to fully adapt to its inherent scale challenges and positioning accuracy requirements in processing small target detection tasks in UAV images, which limits the further improvement of model performance. The specific formula of CIOU is as follows:

$$CIOU = IoU - \left(\frac{\rho^2(B^{pred}, B^{gt})}{c^2} + \alpha v \right) \quad (5)$$

$$v = \frac{4}{\pi} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w^{pred}}{h^{pred}} \right) \quad (6)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (7)$$

where α is used as a balance parameter, v is used to measure the length and width, c is the normalized compensation coefficient, and its calculation basis comes from the horizontal span w^{gt} and vertical dimension h^{gt} of the real frame. At the same time, the width and height parameters w^{pred} and h^{pred} of the model prediction frame participate in the spatial alignment calculation, and $\rho^2(B^{pred}, B^{gt})$ represents the Euclidean distance between the prediction box and the center point of the real box. However, CIOU does not consider the quality of the labeled examples of the dataset itself and ignores the impact of low-quality data on detection performance. harm. Therefore, this paper first introduces Wise-IoU [48] as a new bounding box loss function.

β is used to denote the anomaly degree of the prediction box, which is defined as follows:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \quad (8)$$

where L_{IoU}^* is the constant converted from the variable $\overline{L_{IoU}}$, $\overline{L_{IoU}}$ is the running average of the momentum m , and the calculation formula of m is

$$m = \sqrt[n]{0.05} \quad (9)$$

where t is the Epoch value; n is the value of the batchsize.

Wise-IoU is defined as

$$L_{WIoU} = \frac{\beta}{\delta \alpha_1^{\beta-\delta}} R_{WIoU} L_{IoU} \quad (10)$$

$$R_{WIoU} = \exp \left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*} \right) \quad (11)$$

$$L_{IoU} = 1 - IoU \quad (12)$$

where (x, y) and (x_{gt}, y_{gt}) are coordinates of the center points of the prediction frame and the target frame, respectively; W_g and H_g are the dimensions of the smallest bounding box; IoU is the cross-to-merge ratio, which is a quantity commonly used to measure the degree of overlap between the predicted frame and the real frame. α_1 and δ are learning parameters. Wise-IoU reduces the detrimental gradients of low-quality data, thereby improving the overall performance of the model.

But Wise-IoU is not aware of the limitations of IoU itself. Inner-IoU [49] proposes to calculate IoU with an auxiliary border to improve the generalization ability. The specific calculation process is shown in Equations (13) and (14), and the scale factor ratio is used to control the size of the auxiliary bounding box.

$$b_l = x_c - \frac{w \times ratio}{2}, b_r = x_c + \frac{w \times ratio}{2} \quad (13)$$

$$b_t = y_c - \frac{h \times ratio}{2}, b_b = y_c + \frac{h \times ratio}{2} \quad (14)$$

where w and h denote the width and height of the detection frame, b_l , b_r , b_t , b_b represents the left, right, upper and lower boundary coordinates of the auxiliary detection frame respectively, (x_c, y_c) is the center point coordinate of the detection frame.

Through the above two equations, the center point of the detection frame can be transformed to obtain the corner vertices of the auxiliary detection frame. Both the prediction box and the real box output by the model are transformed accordingly, and the calculation results of the real box and the prediction box are represented by b^{gt} and b^{pred} , respectively.

$$S_{inter} = \left(\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l) \right) \times \left(\min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t) \right) \quad (15)$$

$$S_{union} = (w^{gt} \times h^{gt}) \times (ratio)^2 + (w \times h)(ratio)^2 - S_{inter} \quad (16)$$

$$IoU^{inner} = \frac{S_{inter}}{S_{union}} \quad (17)$$

where b_l^{gt} , b_r^{gt} , b_t^{gt} , and b_b^{gt} represent the left, right, upper, and lower boundary coordinates of the auxiliary border corresponding to the ground truth, respectively; w^{gt} and h^{gt} represent the width and height of the real box, respectively; S_{inter} represents the intersection area of the auxiliary border; and S_{union} represents the union area of the auxiliary border.

According to the above three formulas, it can be seen that Inner-IoU actually calculates the IoU between the auxiliary borders. When the ratio = 1, the Inner-IoU loss function is actually the IoU loss function. Since the PCB defect detection task images are all small targets, the IoU will decrease if the labeling box is slightly offset. When the ratio is >1, the auxiliary border is larger than the actual border, which is helpful for sample regression with low IoU. Therefore, for small target detection, the value of the ratio should be greater than 1.

MPDIoU is an improved regression loss function that can minimize the distance between the upper left corner and the lower right corner of the prediction box and the real box [50]. It can successfully deal with the overlap of bounding boxes, has a good effect on complex multi-objective scenes, and can improve the convergence speed.

$$MPDIoU = IoU - \frac{\rho^2(P_1^{pred}, P_1^{gt})}{w^2 + h^2} - \frac{\rho^2(P_2^{pred}, P_2^{gt})}{w^2 + h^2} \quad (18)$$

where P_1^{pred} , P_2^{pred} , P_1^{gt} and P_2^{gt} refer to the points in the upper left corner and the lower right corner of the prediction box and the real box, respectively, and $\rho^2(P_1^{pred}, P_1^{gt})$ calculates the distance between the corresponding points.

Combining inner-IoU and MPDIoU, Inner-MPDIoU is obtained, as shown in Equation (19), which can not only calculate IoU through the auxiliary border, improve the generalization ability, but also solve the overlapping problem of bounding boxes in complex multi-objective scenes.

$$MPDIoU^{inner} = IoU^{inner} - \frac{\rho^2(P_1^{pred}, P_1^{gt})}{w^2 + h^2} - \frac{\rho^2(P_2^{pred}, P_2^{gt})}{w^2 + h^2} \quad (19)$$

On the basis of obtaining inner-MPDIoU, combining Wise-IoU can reduce the harm low-quality data causes with regard to detection performance. At the same time, replacing the IoU calculation part of Wise-IoU with the obtained inner-MPDIoU, IoU can be achieved by calculating it through the auxiliary border, solving the limitation of IoU itself, improving the generalization ability of the model, and finally obtaining the improved loss function WPIIoU; that is, Equation (20), which can effectively improve the model detection effect.

$$L_{WPIIoU} = \frac{\beta}{\delta \alpha_1^{\beta-\delta}} \times \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \times \left(1 - IoU^{inner} + \frac{\rho^2(P_1^{pred}, P_1^{gt})}{w^2 + h^2} + \frac{\rho^2(P_2^{pred}, P_2^{gt})}{w^2 + h^2}\right) \quad (20)$$

3. Results and Analysis

3.1. Experimental Datasets

This study utilizes the VisDrone2019 dataset, curated by the AISKYEYE team from the Machine Learning and Data Mining Laboratory at Tianjin University [51]. Captured via UAVs across diverse scenes, weather conditions, and lighting environments, the dataset comprises over 2.6 million manually annotated bounding boxes, ensuring high annotation accuracy, rich object categories, and comprehensive occlusion scenarios. It includes 10,209 drone-captured images, partitioned into 6471 for training, 548 for validation, and 3190 for testing. The training set contains 353,550 annotated objects, with approximately 50% affected by occlusion—142,873 partially occluded and 33,804 severely occluded. Ten categories of small objects are labeled, making the dataset particularly well-suited for small object detection and recognition research. Representative examples are shown in Figure 7.

3.2. Experiment Environment

Experiments were conducted on a desktop equipped with an Intel® Core™ i5-14400F processor, 16 GB RAM, NVIDIA GeForce RTX 3050 GPU, and Windows 11 (64-bit). The deep learning environment was configured with Python 3.8.19, PyTorch 1.13.1, and CUDA 11.7. Model training employed stochastic gradient descent (SGD) for parameter optimization. Key hyperparameter settings are summarized in Table 1.



Figure 7. VisDrone2019 partial images in the dataset.

Table 1. Model training parameters.

Hyperparameters	Value
Initial learning rate	0.01
Optimizer momentum	0.937
Optimizer weight decay rate	0.0004
batchsize	16
Epoch	200

3.3. Evaluation Metrics

A total of five evaluation metrics are employed in order to provide a comprehensive assessment of the detection model. These include precision, recall, parameter count, computational complexity, and mean average precision (mAP). The equations for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \quad (21)$$

$$R = \frac{TP}{TP + FN} \quad (22)$$

$$AP = \int_0^1 P(R) dR \quad (23)$$

$$mAP = \frac{\sum_0^1 AP_1}{n} \quad (24)$$

TP represents the number of positive samples correctly predicted as positive by the model, FP represents the number of negative samples incorrectly predicted as positive by the model, and FN represents the number of positive samples incorrectly predicted as negative by the model. P refers to the ratio of correctly predicted positive samples among all samples predicted to be positive; R refers to the ratio of correctly predicted positive samples among all actual positive samples. AP represents the area under the Precision–Recall (P–R) curve, while mAP denotes the average of AP for each category.

3.4. Analysis of the Influence of C2f_DSC Module on Model Performance

Although the C2f_DSC module theoretically enhances feature extraction while reducing the parameter count, its practical impact on performance remains uncertain, particularly when applied to deep network structures. To evaluate the optimal integration point within the YOLOv8s architecture, four experimental configurations were tested, as

summarized in Table 2. Replacing only the backbone’s C2f module led to a slight accuracy gain (49.8%→50.3%), a minor decrease in recall (38.4%→38.1%), and notable reductions in parameters (14%) and FLOPs (19%), with a marginal increase in mAP@0.5 (+0.1%). Modifying only the neck module resulted in decreased accuracy (49.8%→48.6%) and recall (38.4%→37.4%), but with a 20% parameter and 28% FLOPs reduction, while mAP@0.5 slightly declined (−0.1%). When both backbone and neck modules were replaced, accuracy and recall further declined (49.8%→47.4%, 38.4%→37.7%), though parameter count and FLOPs dropped significantly by 34% and 47%, respectively, with a mAP@0.5 decrease of 0.2%. These results suggest that while C2f_DSC enhances computational efficiency, performance improvements do not scale linearly with the extent of replacement. Given its highest cost-efficiency, this study adopts C2f_DSC throughout the network.

Table 2. Comparison of C2f_DSC module improvement in different locations.

Model	Precision%	Recall%	Parameters (10 ⁶)	FLOPs (G)	mAP@0.5 (%)
YOLOv8s	49.8	38.4	11.1	28.5	38.4
Backbone	50.3	38.1	9.5	23.1	38.5
Neck	49.6	37.4	8.9	20.6	38.3
Backbone + Neck	47.4	37.7	7.3	15.2	38.2

3.5. Comparison of Loss Functions

In this chapter, we perform quantitative contrast experiments on several loss functions: CIoU, Inner-IoU, GIoU, SIoU, ShapeIoU, DIoU, and WIPIoU. The experimental results are shown in Table 3.

Table 3. Analysis of the impact of loss function on model performance.

Model	Precision%	Recall%	mAP@0.5 (%)
YOLOv8s	49.8	38.4	38.4
YOLOv8s + Inner-IoU	48.1	37.2	37.9
YOLOv8s + GIoU	48.8	37.4	38.0
YOLOv8s + SIoU	48.7	37.9	38.1
YOLOv8s + Shape-IoU	49.2	37.8	38.4
YOLOv8s + DIoU	49.1	38.2	38.5
YOLOv8s + WIPIoU	50.5	38.7	40.4

By comparing the experimental results, we find that the model accuracy, recall, and mAP@0.5 using the WIPIoU loss function are all superior to other loss functions. The accuracy of the model using WIPIoU loss function is significantly improved, which shows that the model reduces false detection and missed detection when identifying small targets, and improves the detection accuracy. The performance of the WIPIoU loss function in terms of recall rate is also very good, indicating that the model can effectively identify more real targets and reduce missed detection, which is particularly important for UAV small target detection in practical applications.

3.6. Comparison of Different Network Detection Heads

To further evaluate the performance of the DyHead detection head, comparative experiments were conducted under identical conditions by integrating DyHead, EfficientHead, SEAMHead, and MultiSEAMHead into the YOLOv8s framework. The results, presented in Table 4, show that the DyHead-enhanced model achieved the highest overall performance, with an accuracy of 51.5%, recall of 44.9%, and mAP@0.5 of 44.7%. While EfficientHead, SEAMHead, and MultiSEAMHead yielded improvements in select metrics, none matched the comprehensive effectiveness of DyHead. Moreover, DyHead introduces

only a marginal increase in model parameters. Therefore, DyHead is identified as the most effective enhancement strategy among those tested.

Table 4. Comparison of different detection heads.

Model	Precision/%	Recall/%	mAP@0.5 (%)	Parameters (10 ⁶)
YOLOv8s	49.8	38.4	38.4	11.1
EfficientHead	53.1	42.6	41.1	11.9
SEAMHead	52.6	39.9	40.5	12.2
MultiSEAMHead	55.3	43.4	42.0	12.2
Dyhead	51.5	44.9	44.7	12.3

3.7. Visual Analysis of the Model Before and After Improvement

We further evaluate the detection ability of the trained model through the confusion matrix. The confusion matrix is shown in Figure 8. It can be seen from the confusion matrix that the improved model can reflect higher classification accuracy and a more balanced classification performance in UAV aerial images.

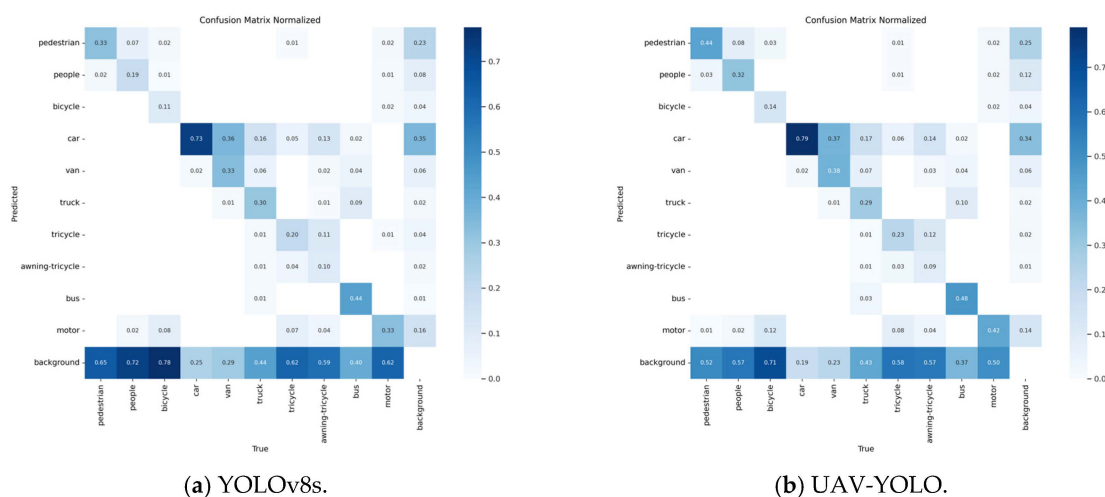


Figure 8. Confusion matrix.

To evaluate the validity and accuracy of the UAV-YOLO model, comparative experiments were performed using the benchmark model YOLOv8s and the UAV-YOLO model, performed using several different scenarios involving the VisDrone2019 dataset. Tables 5 and 6 provide detailed comparisons of performance indicators and detection accuracies for all categories before and after model improvement. The mAP@0.5 of UAV-YOLO was significantly improved by 8.9% compared to YOLOv8s. The parameter amount of UAV-YOLO is 8.5×10^6 , which is 23.4% less than the benchmark model, and the model calculation amount is 16.9 FLOPs, which is 40.7% less than the benchmark model, demonstrating excellent performance. In addition, Figure 9 more intuitively shows the comparison of the recognition accuracy of the YOLOv8s and UAV-YOLO models for all categories on the VisDrone2019 dataset.

Table 5. YOLOv8s and UAV-YOLO models compared on the VisDrone2019 dataset.

Model	FLOPs (G)	Parameters (10 ⁶)	Precision (%)	Recall (%)	mAP@0.5 (%)
YOLOv8s	28.5	11.1	49.8	38.4	38.4
UAV-YOLO	16.9	8.5	52.7	46.9	47.3

Table 6. VisDrone2019 dataset different category target mAP@0.5 comparison.

Model	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning Tricycle	Bus	Motor
YOLOv8s	40.5	31.3	13.0	79.5	44.5	36.3	25.5	15.7	55.4	42.7
UAV-YOLO	55.5	45.0	20.5	88.3	52.2	40.1	32.4	20.3	63.4	55.5

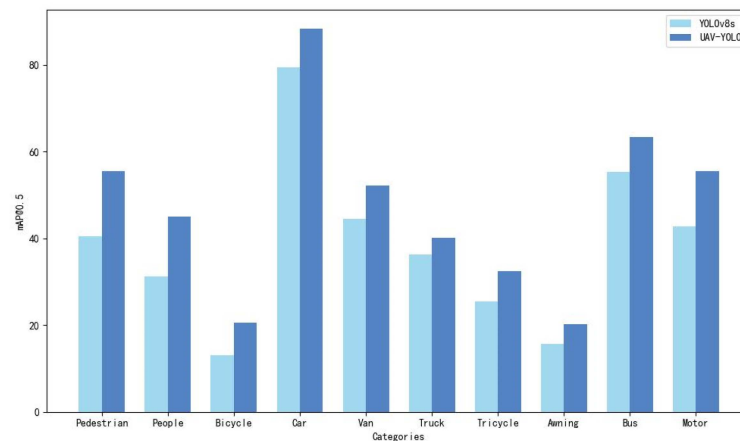
**Figure 9.** Comparison of recognition accuracy of YOLOv8s and UAV-YOLO models on VisDrone2019 dataset.

Figure 10 provides a more intuitive diagram of the test results. In high-exposure scenes, as shown in Figure 10a,b, the UAV-YOLO model shows significant advantages in detecting small objects. Compared with the benchmark model YOLOv8s, UAV-YOLO can detect and locate different targets more accurately, while YOLOv8s may cause missed detection and false detection by identifying signs as motorcycles and not identifying people on the bridge, etc. In the dense target scene, as shown in Figure 10c,d, UAV-YOLO successfully identified motorcycles and pedestrians that the YOLOv8s model failed to identify. In the night scene, as shown in Figure 10e,f, UAV-YOLO maintains excellent detection performance and successfully identifies more cars, pedestrians and motorcycles, while the YOLOv8s model mistakenly identifies motorcycles as bicycles. The ability of the UAV-YOLO model to accurately locate objects in low light conditions further verifies its robustness to illumination changes, providing reliable support for night surveillance and target detection tasks. In multi-scale scenarios, as shown in Figure 10g,h, UAV-YOLO shows excellent performance when detecting distant or incoming and exiting vehicles. It is worth noting that in the lower right corner of the image, UAV-YOLO can more accurately distinguish different vehicle types.

3.8. Ablation Experiment

Given its balance between detection accuracy and speed, YOLOv8s was selected as the baseline model for optimizing small object detection in UAV imagery. To assess the proposed enhancements, nine improved configurations were designed and evaluated under consistent training conditions. Table 7 outlines the modifications, while Table 8 presents the corresponding results on the VisDrone2019 dataset, with optimal values highlighted in bold. The experimental results demonstrate that the best performance is achieved when all four improvements are applied simultaneously, effectively enhancing the extraction and utilization of small object features while simplifying the model architecture.

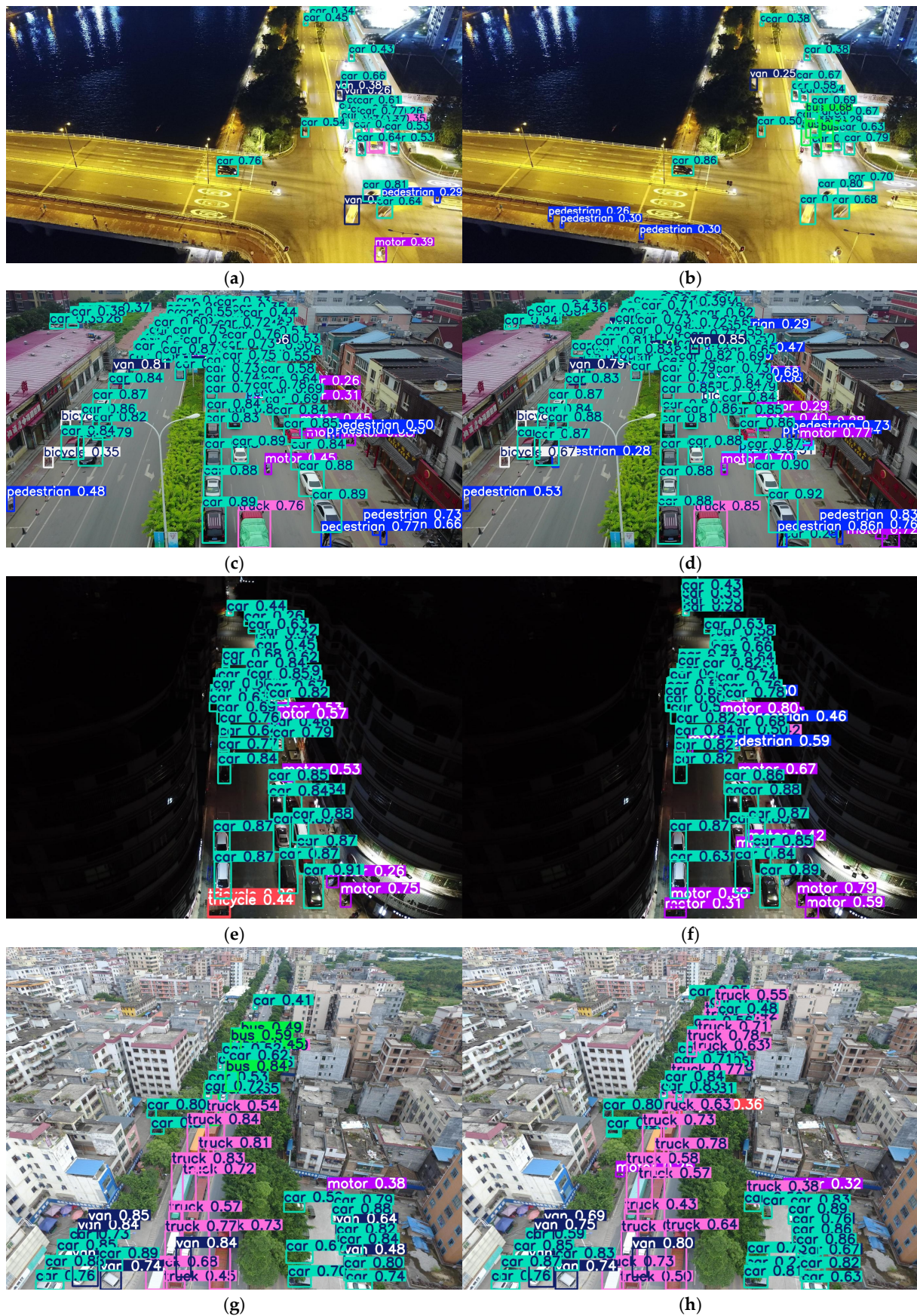


Figure 10. Intuitive diagram of the test results in different scenarios. High exposure scene: (a) YOLOv8s; (b) UAV-YOLO. Dense target scenario: (c) YOLOv8s; (d) UAV-YOLO. Night Scene: (e) YOLOv8s; (f) UAV-YOLO. Multi-scale scenario: (g) YOLOv8s; (h) UAV-YOLO.

Table 7. Different improvement strategy models.

Models	C2f_DSC	SPPL	Dyhead	WIPIoU
YOLOv8s-1				
YOLOv8s-2	✓			
YOLOv8s-3		✓		
YOLOv8s-4			✓	
YOLOv8s-5				✓
YOLOv8s-6	✓	✓	✓	
YOLOv8s-7	✓	✓		✓
YOLOv8s-8		✓	✓	✓
YOLOv8s-9	✓	✓	✓	✓

Table 8. Experimental results of different improvement strategy models.

Models	Precision/%	Recall/%	mAP@0.5 (%)	Parameters (10 ⁶)	FLOPs (G)
YOLOv8s-1	49.8	38.4	38.4	11.1	28.5
YOLOv8s-2	47.4	37.7	38.2	7.3	15.2
YOLOv8s-3	52.5	39.0	40.2	11.2	28.7
YOLOv8s-4	51.5	44.9	44.7	12.3	30.0
YOLOv8s-5	50.5	38.7	40.4	11.1	28.5
YOLOv8s-6	52.3	45.2	45.4	8.5	16.9
YOLOv8s-7	51.7	46.3	46.4	7.5	15.3
YOLOv8s-8	50.7	43.7	42.6	12.4	30.3
YOLOv8s-9	52.7	46.9	47.3	8.5	16.9

In single-module ablation studies, replacing the C2f module with the proposed C2f_DSC module reduced the computational load and parameter count by 40.7% and 23.4%, respectively, with only a marginal mAP@0.5 drop of 0.2%, indicating substantial efficiency gains with minimal accuracy loss. Substituting SPPF with the SPPL module slightly increased complexity but improved mAP@0.5 by 1.8%, validating its effectiveness in feature fusion. Integrating DyHead significantly boosted mAP@0.5 by 6.3% with only minor overhead. Additionally, replacing the CIoU loss function with the proposed WIPIoU yielded a 2.0% mAP@0.5 gain without increasing model complexity, confirming its robustness.

Combined ablation experiments revealed synergistic effects among the modules. Simultaneously incorporating C2f_DSC, SPPL, and DyHead raised mAP@0.5 to 45.4%, with parameter and FLOPs counts of 8.5 M and 16.9, respectively, demonstrating strong complementarity in performance and compactness. Replacing DyHead with WIPIoU under the same conditions further improved mAP@0.5 to 46.4%. In contrast, combining DyHead, SPPL, and WIPIoU yielded a lower mAP@0.5 of 42.6%, with substantially higher parameters (12.4 M) and FLOPs (30.3), indicating the dominant influence of the C2f_DSC module in maintaining lightweight efficiency.

Overall, the optimized UAV-YOLO model achieves a 23.4% reduction in parameters and a 40.7% reduction in computational load while maintaining high detection accuracy. These results confirm the effectiveness of the proposed improvements in achieving a favorable balance between model compactness, computational efficiency, and detection performance for UAV-based small object detection.

3.9. Comparative Analysis with Other Models

To comprehensively assess the performance of UAV-YOLO in UAV-based small object detection, it was compared with several mainstream object detection models, including Faster R-CNN, SSD, RetinaNet, YOLOv5s, YOLOv7-tiny, YOLOX-s, YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, and YOLO12s. Table 9 summarizes the comparative results across

key metrics such as parameter count, computational complexity, accuracy, recall, and mAP@0.5, with optimal values highlighted in bold. A visual comparison is provided in Figures 11 and 12, presenting the dynamic change curve of the mAP@0.5 index of each model on the Visdrone2019 dataset with the growth of training rounds.

As shown in Table 10, SSD and RetinaNet incur high computational costs and parameter counts but yield low mAP@0.5 values (23.9% and 26.5%, respectively), indicating insufficient feature extraction capabilities for small objects. Although Faster R-CNN exhibits the highest resource demands, its detection performance remains limited (mAP@0.5 = 33.2%). YOLOv5s and YOLOX-s are more efficient in terms of resources, but their detection accuracy is similarly constrained, with mAP@0.5 values of 32.8% and 33.1%, respectively. In contrast, advanced YOLO variants (YOLOv7-tiny, YOLOv8s, YOLOv9s [52], YOLOv10s [53], YOLO11s [54], and YOLO12s) achieve better performance, with mAP@0.5 ranging from 36.3% to 38.4%.

Table 9. Results on the DIOR dataset.

Model	YOLOv8s			UAV-YOLO		
	Precision (%)	Recall (%)	mAP@0.5 (%)	Precision (%)	Recall (%)	mAP@0.5 (%)
Airplane	91.5	82.1	88.2	97.2	86.1	94.2
Airport	41.7	52.9	53.3	83.6	65.7	85.8
Baseball field	92.7	91.9	96.4	97.9	94.6	97.2
Basketball court	76.0	88.9	88.0	95.0	89.1	94.4
bridge	59.8	30.8	29.4	72.4	38.8	35.7
chimney	91.2	69.3	81.5	95.2	71.6	91.8
dam	59.9	25.0	32.8	41.9	20.8	41.3
Expressway-service-area	72.1	71.8	77.7	89.1	79.8	90.1
Expressway-toll-station	54.7	77.8	81.2	87.4	77.8	88.9
Golf field	63.9	66.8	63.6	79.4	87.5	81.0
Ground track field	93.3	99.3	98.6	97.4	95.4	96.6
Harbor	74.2	62.2	71.2	86.4	59.8	83.2
Overpass	91.2	69.5	73.9	92.5	58.0	86.7
Ship	90.6	69.7	76.8	96.8	57.7	85.9
Stadium	88.8	99.6	98.2	93.2	87.5	97.1
Storage tank	94.1	37.9	44.6	95.8	44.7	42.8
Tennis court	1	90.8	95.6	98.7	94.7	96.6
Train station	34.0	12.5	13.5	99.6	37.5	45.8
Vehicle	72.3	28.5	36.4	87.7	26.1	44.8
Windmill	82.4	82.9	83.3	97.1	80.9	89.7
mAP@0.5(%)		69.2			78.5	

Table 10. UAV-YOLO compared with other different models.

Model	FLOPs (G)	Parameters (10 ⁶)	Precision (%)	Recall (%)	mAP@0.5 (%)
SSD	87.9	24.5	21.0	25.5	23.9
Faster-RCNN	206.7	41.2	45.3	33.8	33.2
RetinaNet	93.7	19.8	23.5	37.9	26.5
YOLOv5s	15.8	7.0	42.5	34.5	32.8
YOLOv7-tiny	13.1	6.0	47.3	38.6	36.3
YOLOX-s	26.7	8.9	43.2	34.8	33.1
YOLOv8s	28.5	11.1	49.8	38.4	38.4
YOLOv9s	26.7	7.1	48.9	37.2	38.2
YOLOv10s	24.5	8.0	47.4	36.6	37.2
YOLO11s	21.3	9.4	49.3	36.8	37.6
YOLO12s	21.2	9.2	49.1	37.1	37.5
UAV-YOLO	16.9	8.5	58.7	45.2	47.3

UAV-YOLO achieves a mAP@0.5 of 47.3%, with accuracy and recall of 58.7% and 45.2%, respectively. Compared with the YOLOv8s baseline, it reduces parameter count and computational complexity by 23.4% and 40.7%, respectively. These improvements result

from integrating specialized modules for small object detection, enhancing feature extraction in complex UAV scenarios. Although its complexity is slightly higher, this trade-off is justified by the significant accuracy gains. Overall, UAV-YOLO demonstrates an optimal balance between detection accuracy, computational efficiency, and real-time performance, making it well-suited for high-precision small object detection tasks in resource-constrained UAV environments.

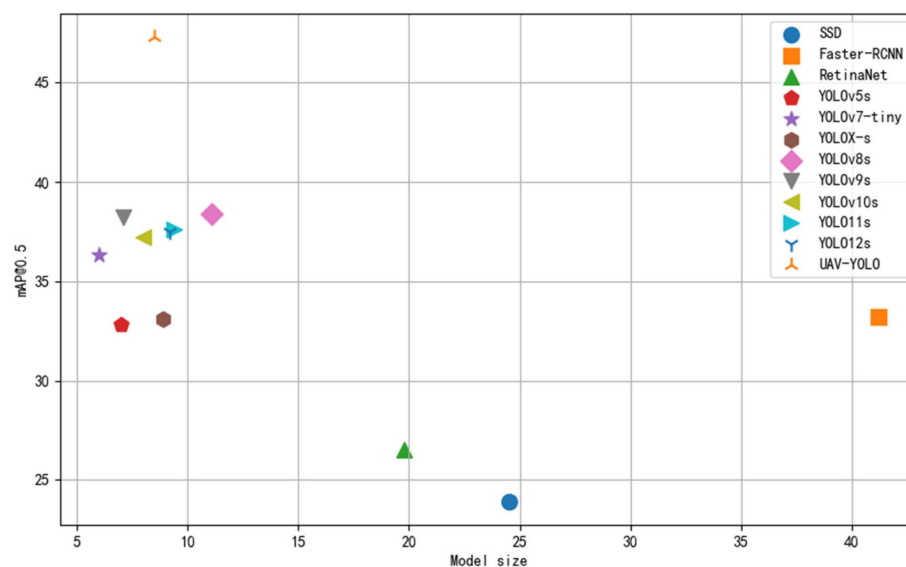


Figure 11. Performance comparison chart of each model on the VisDrone2019 dataset.

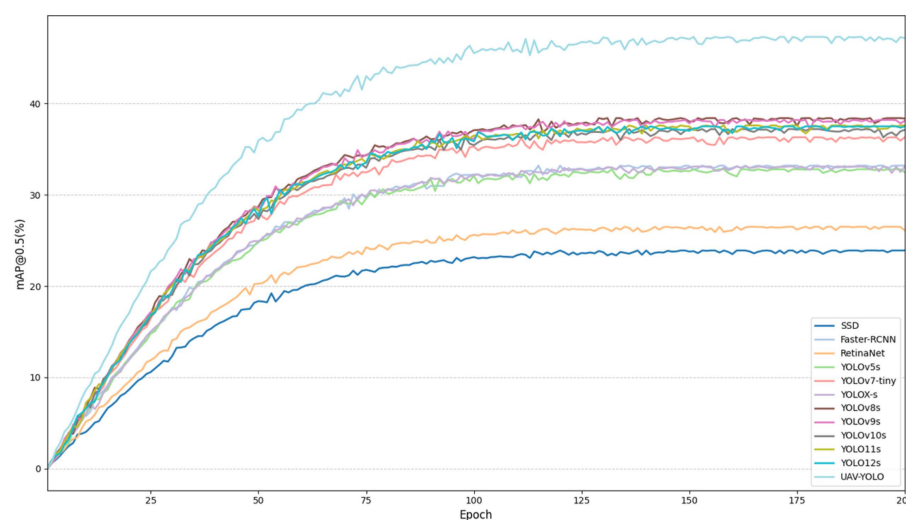


Figure 12. Variation curves of mAP@0.5 with training rounds for different models on the VisDrone2019 dataset.

3.10. Generalization Experiment

In order to further verify the superiority and versatility of the proposed algorithm in this study, this chapter uses DIOR [55], RSOD and NWPU VHR-10 [56] datasets for verification. The experimental results of the YOLOv8s and UAV-YOLO models are shown in Tables 9, 11 and 12, respectively.

The UAV-YOLO model proposed in this study exhibits significant performance improvement when applied to DIOR, RSOD, and NWPU VHR-10. Compared with the benchmark model YOLOv8s, UAV-YOLO achieved 78.5%, 97.5%, and 92.7% on the average accuracy mean (mAP), while YOLOv8s achieved 69.2%, 90.7%, and 89.8%, respectively.

This result shows that UAV-YOLO has better processing capabilities when dealing with challenges such as small targets, occlusion and dense distribution, and it performs better in UAV aerial image detection. In order to demonstrate this advantage more intuitively, Figures 13–15 shows some images detected by the YOLOv8s and UAV-YOLO models on three datasets, respectively, further confirming the superiority of the UAV-YOLO model.

Table 11. Results on the RSOD dataset.

Model	YOLOv8s			UAV-YOLO		
	Precision (%)	Recall (%)	mAP@0.5 (%)	Precision (%)	Recall (%)	mAP@0.5 (%)
Aircraft	95.8	91.0	96.2	96.3	92.1	98.8
Oil tank	93.8	95.5	98.0	95.9	97.2	99.8
Overpass	58.6	73.5	70.6	82.7	94.3	91.5
Playground	94.9	98.6	98.3	95.5	99.1	99.9
mAP@0.5(%)		90.7			97.5	

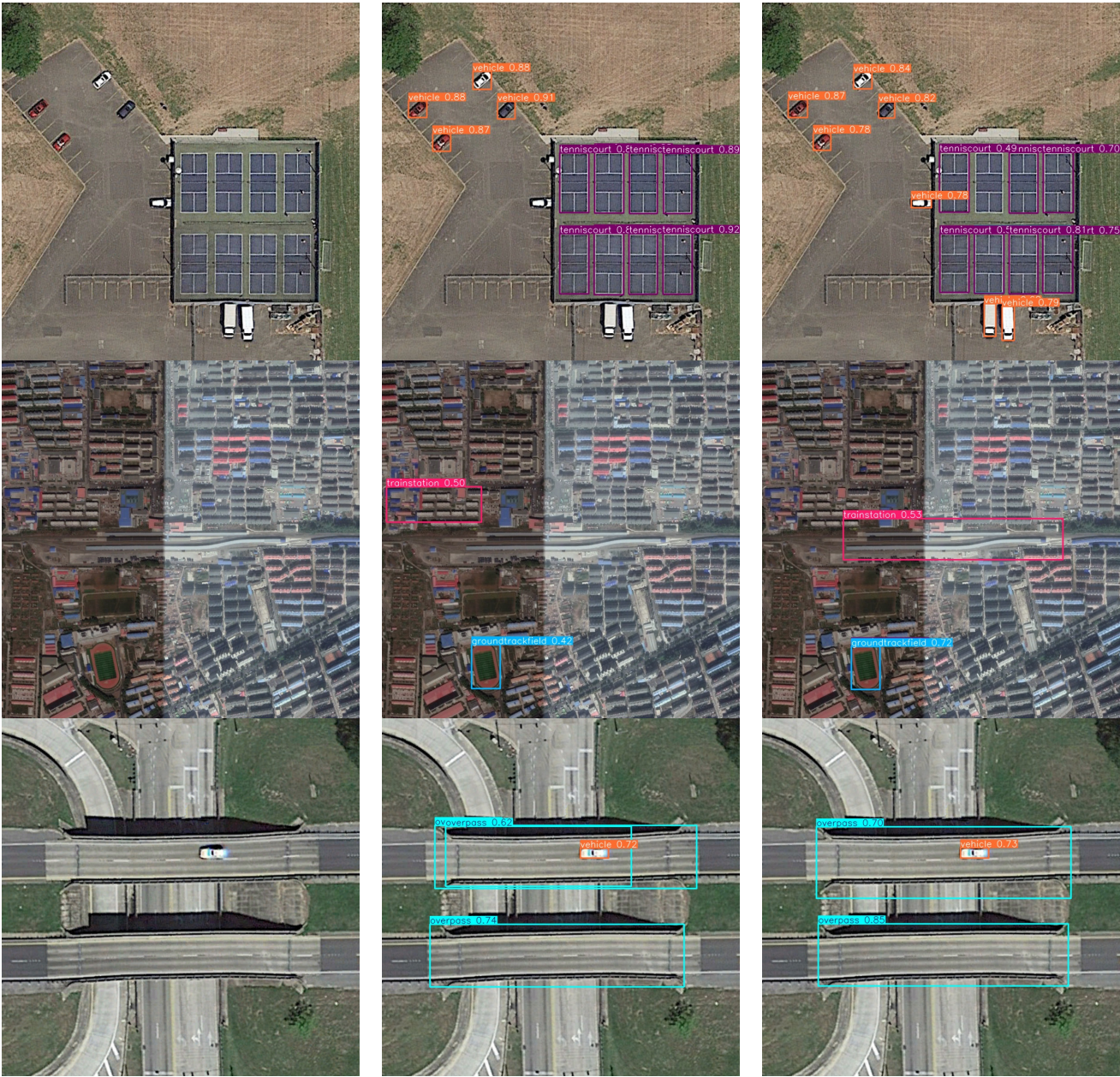


Figure 13. Cont.

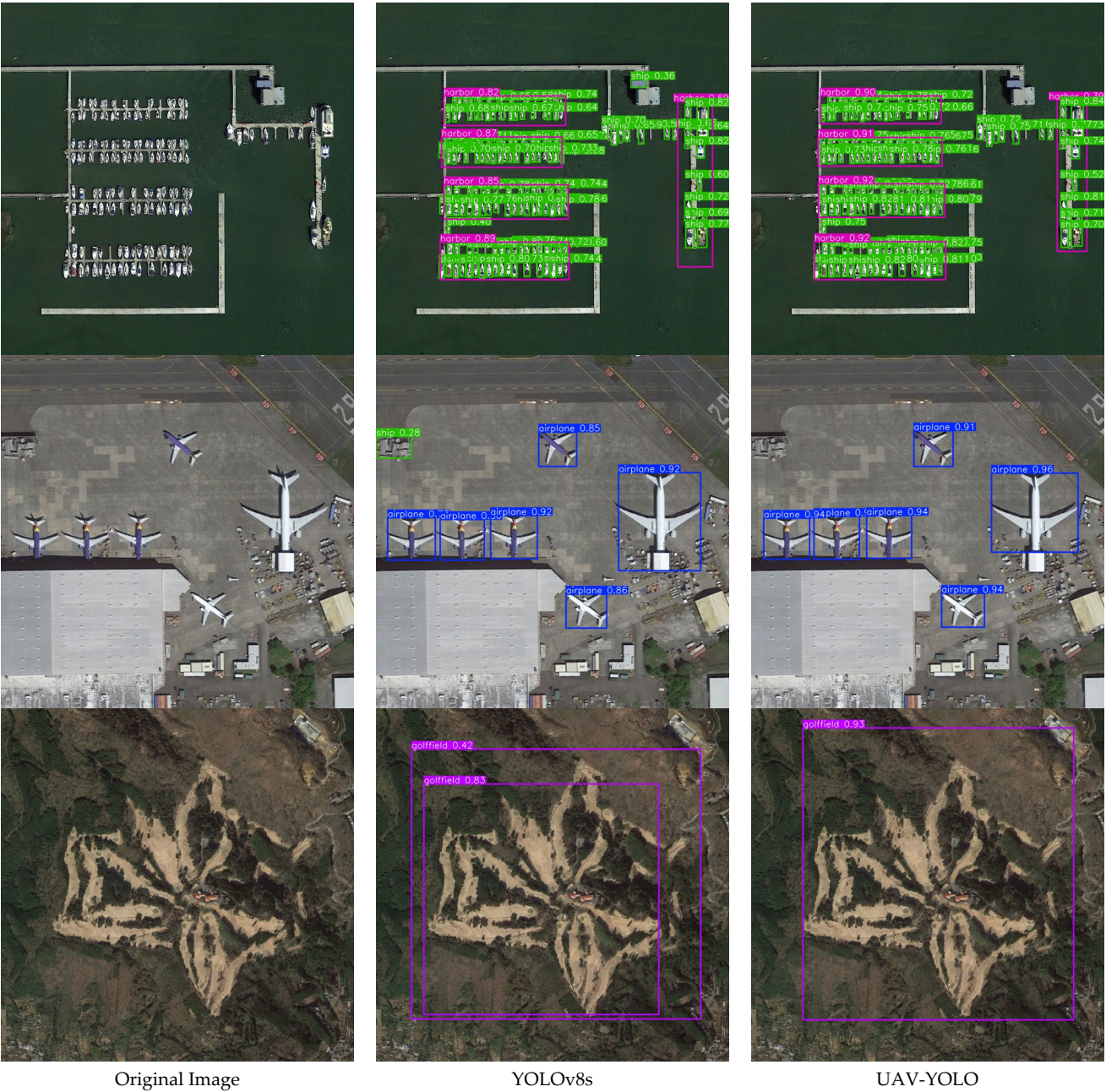


Figure 13. Detection results on the DIOR dataset.

Table 12. Results on the NWPU VHR-10 dataset.

Model	YOLOv8s			UAV-YOLO		
	Precision (%)	Recall (%)	mAP@0.5 (%)	Precision (%)	Recall (%)	mAP@0.5 (%)
Aero plane	97.0	98.6	99.4	96.0	98.6	99.2
Ship	98.7	76.8	93.2	87.1	91.1	94.7
Storage tank	95.8	93.6	94.6	97.7	93.6	96.9
Baseball diamond	94.6	97.3	98.5	92.3	96.5	96.9
Tennis court	95.1	73.0	87.7	96.4	86.2	95.9
Basketball court	92.0	76.4	82.9	93.6	82.2	88.6
Ground Track field	100	90.6	99.3	94.2	92.3	97.5
Harbor	83.5	97.0	95.6	89.4	97.0	96.8
Bridge	72.2	65.0	73.8	85.0	69.2	77.3
Vehicle	89.8	54.8	73.2	89.1	65.6	83.1
mAP@0.5(%)		89.8			92.7	

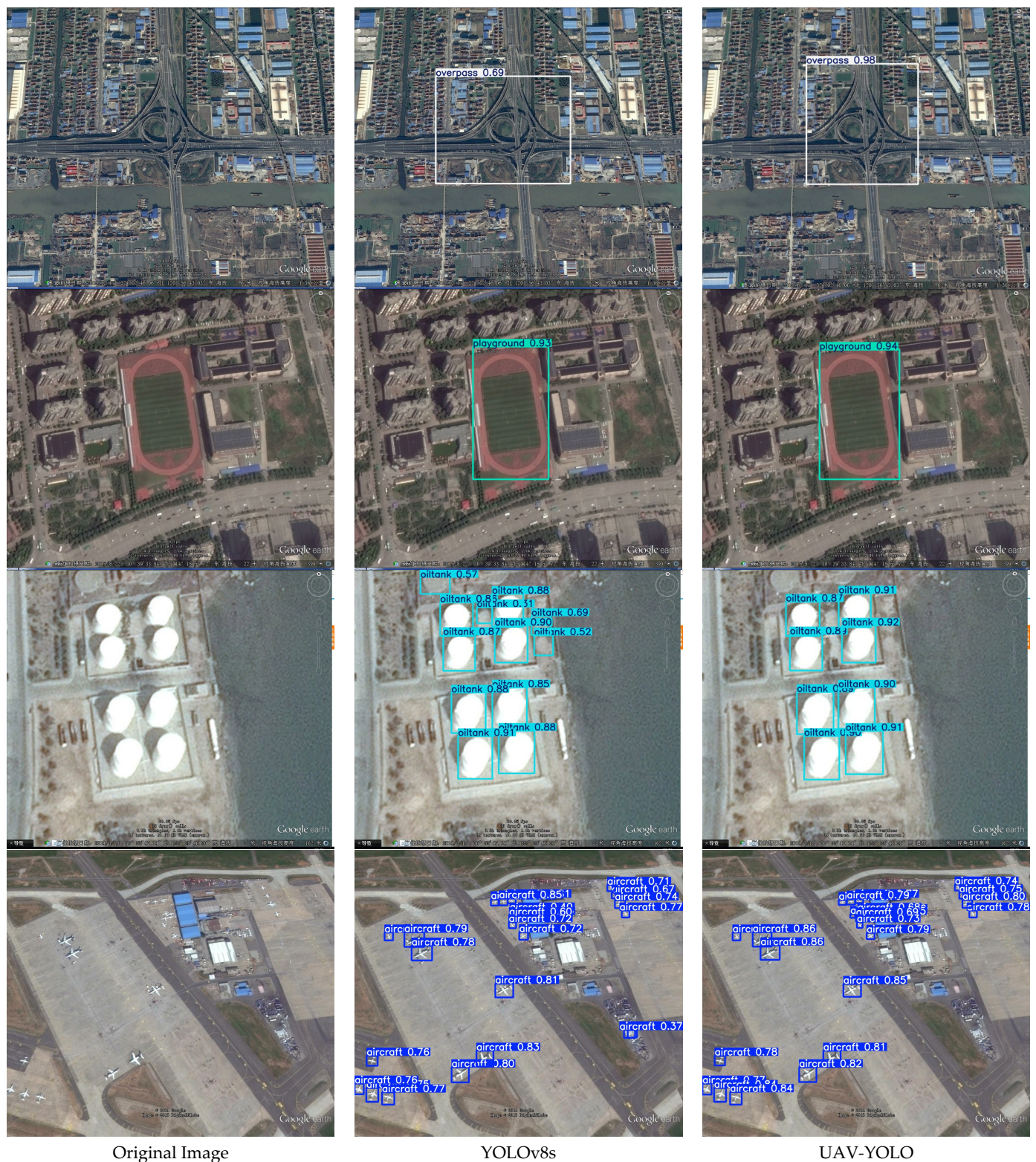


Figure 14. Detection results on the RSOD dataset.



Figure 15. Detection results on the NWPU VHR-10 dataset.

4. Conclusions

In this study, we present UAV-YOLO, an enhanced small object detection model based on YOLOv8s, which is specifically designed for UAV aerial imagery. To reduce model complexity and facilitate the potential deployment on edge devices, the original C2f modules are replaced with the C2f_DSC modules, which adopt depthwise separable and pointwise convolutions. This modification effectively enhances spatial feature extraction and fusion while significantly reducing parameters, computation, and memory access, thereby improving inference efficiency. These effects are particularly beneficial for UAV-based small object detection.

To further improve detection accuracy, we introduce a novel SPPL module that integrates the multi-scale feature extraction capability of SPPF with the long-range dependency modeling and self-attention mechanism of LSKA. This structure enables more effective detection of targets across varying scales and strengthens the model's adaptability to complex multi-scale environments. In addition, the detection head is replaced with DyHead, which employs dynamic weight allocation and adaptive feature fusion to capture multi-scale information more effectively. This enhancement significantly boosts detection performance for small objects in cluttered scenes. Moreover, we adopt the WIPIoU loss function, which integrates several losses to replace the original CIoU loss. WIPIoU introduces auxiliary boundaries to improve robustness against low-quality annotations, while also enhancing generalization and convergence speed through the synergistic effects of the combined losses, ultimately improving small target detection accuracy. Experimental results for the VisDrone2019 dataset show that UAV-YOLO achieves a mAP@0.5 of 47.3%, representing an 8.9% improvement over the baseline YOLOv8s model, and consistently outperforms other mainstream detectors. The model demonstrates strong generalization capability and performs reliably across diverse scenes, including those with complex backgrounds and varying target scales.

Although UAV-YOLO has demonstrated superior performance in a variety of scenarios, its model's ability to extract small target features may decline in extreme environments such as extreme weather or severe lighting changes. Public datasets such as VisDrone and DIOR based on current experiments are not comprehensive enough for labeling small targets in specific fields such as ocean monitoring and forest fires, which may affect the model's generalization ability in professional scenarios.

In future work, Active Disturbance Rejection Control (ADRC) [57] could be incorporated to enhance real-time control robustness and deployed on actual UAV platforms to further validate the model's performance in real-world scenarios. Additionally, the integration of UAV-UGV cooperative navigation could be explored, wherein the UAV provides path planning and navigational guidance for the UGV to execute precision agricultural tasks [58]. Furthermore, the adoption of digital twin technology and synthetic data generation may significantly improve the robustness and accuracy of the UAV-YOLO model under extreme environmental conditions, enabling more reliable performance in diverse field applications [59,60]. Moreover, the integration of multimodal technologies could further enhance the model's adaptability and detection performance across complex environments [61,62].

Author Contributions: Conceptualization, K.C. and R.-F.W.; methodology, X.D., K.C. and R.-F.W.; validation, X.D.; formal analysis, X.D.; investigation, X.D.; data curation, X.D.; writing—original draft, X.D.; software, X.D.; writing—review and editing, K.C. and R.-F.W.; project administration, K.C. and R.-F.W.; supervision, K.C. and R.-F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle.
YOLO	You Only Look Once.
mAP	mean Average Precision.
CIoU	Complete Intersection over Union.
GIoU	Generalized Intersection over Union.
SIoU	Scale Intersection over Union.
FPN	Feature Pyramid Network.
PAN	Path Aggregation Network.
DSC	Dual-Separable Convolution.
SPPF	Spatial Pyramid Pooling Fast.
SPPL	Spatial Pyramid Pooling with Large Kernel Attention.
DWConv	Depthwise Convolution.
PWConv	Pointwise Convolution.
LSKA	Large Kernel Attention.
DyHead	Dynamic Head.
Inner-IoU	Inner Intersection over Union.
MPDIoU	Minimum Predicted Distance Intersection over Union.
WIPIoU	Weighted Perturbation IoU.
UGV	Unmanned Ground Vehicle.

References

1. Li, J.; Liu, A.; Han, G.; Cao, S.; Wang, F.; Wang, X. FedRDR: Federated Reinforcement Distillation-Based Routing Algorithm in UAV-Assisted Networks for Communication Infrastructure Failures. *Drones* **2024**, *8*, 49. [\[CrossRef\]](#)
2. Du, Y.; Wu, T.; Dai, Z.; Xie, H.; Hu, C.; Wei, S. F-Yolov7: Fast and Robust Real-Time UAV Detection. *Computing* **2025**, *107*, 50. [\[CrossRef\]](#)
3. Liu, Y.; He, M.; Hui, B. ESO-DETR: An Improved Real-Time Detection Transformer Model for Enhanced Small Object Detection in UAV Imagery. *Drones* **2025**, *9*, 143. [\[CrossRef\]](#)
4. Wang, R.-F.; Su, W.-H. The Application of Deep Learning in the Whole Potato Production Chain: A Comprehensive Review. *Agriculture* **2024**, *14*, 1225. [\[CrossRef\]](#)
5. Zhang, S.; Xue, X.; Chen, C.; Sun, Z.; Sun, T. Development of a Low-Cost Quadrotor UAV Based on ADRC for Agricultural Remote Sensing. *Int. J. Agric. Biol. Eng.* **2019**, *12*, 82–87. [\[CrossRef\]](#)
6. Cui, K.; Zhu, R.; Wang, M.; Tang, W.; Larsen, G.D.; Pauca, V.P.; Alqahtani, S.; Yang, F.; Segurado, D.; Lutz, D.; et al. Detection and Geographic Localization of Natural Objects in the Wild: A Case Study on Palms. *arXiv* **2025**, arXiv:2502.13023.
7. Niu, Y.; Han, W.; Zhang, H.; Zhang, L.; Chen, H. Estimating Maize Plant Height Using a Crop Surface Model Constructed from UAV RGB Images. *Biosyst. Eng.* **2024**, *241*, 56–67. [\[CrossRef\]](#)
8. Qin, Y.-M.; Tu, Y.-H.; Li, T.; Ni, Y.; Wang, R.-F.; Wang, H. Deep Learning for Sustainable Agriculture: A Systematic Review on Applications in Lettuce Cultivation. *Sustainability* **2025**, *17*, 3190. [\[CrossRef\]](#)
9. Feng, G.; Wang, C.; Wang, A.; Gao, Y.; Zhou, Y.; Huang, S.; Luo, B. Segmentation of Wheat Lodging Areas from UAV Imagery Using an Ultra-Lightweight Network. *Agriculture* **2024**, *14*, 244. [\[CrossRef\]](#)
10. Pei, H.; Sun, Y.; Huang, H.; Zhang, W.; Sheng, J.; Zhang, Z. Weed Detection in Maize Fields by UAV Images Based on Crop Row Preprocessing and Improved YOLOv4. *Agriculture* **2022**, *12*, 975. [\[CrossRef\]](#)
11. Ahmed, S.; Qiu, B.; Kong, C.-W.; Xin, H.; Ahmad, F.; Lin, J. A Data-Driven Dynamic Obstacle Avoidance Method for Liquid-Carrying Plant Protection UAVs. *Agronomy* **2022**, *12*, 873. [\[CrossRef\]](#)

12. Cui, K.; Shao, Z.; Larsen, G.; Pauca, V.; Alqahtani, S.; Segurado, D.; Pinheiro, J.; Wang, M.; Lutz, D.; Plemmons, R.; et al. PalmProbNet: A Probabilistic Approach to Understanding Palm Distributions in Ecuadorian Tropical Forest via Transfer Learning. In Proceedings of the 2024 ACM Southeast Conference, Marietta, GA, USA, 18–20 April 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 272–277.
13. Zhou, S.; Zhou, H.; Qian, L. A Multi-Scale Small Object Detection Algorithm SMA-YOLO for UAV Remote Sensing Images. *Sci. Rep.* **2025**, *15*, 9255. [\[CrossRef\]](#)
14. Nguyen, P.T.; Nguyen, G.L.; Bui, D.D. LW-UAV-YOLOv10: A Lightweight Model for Small UAV Detection on Infrared Data Based on YOLOv10. *Geomatica* **2025**, *77*, 100049. [\[CrossRef\]](#)
15. Wang, Y.; Wang, Z.; Cheng, P.; Tian, P.; Yuan, Z.; Tian, J.; Wang, W.; Zhao, L. Uvcpnnet: A Uav-Vehicle Collaborative Perception Network for 3d Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5615916.
16. Wang, J.; Zhai, Y.; Zhu, L.; Xu, L.; Zhao, Y.; Yuan, H. Sheep-YOLO: A Lightweight Daily Behavior Identification and Counting Method for Housed Sheep. *Meas. Sci. Technol.* **2024**, *36*, 026001. [\[CrossRef\]](#)
17. Wang, Z.; Wang, R.; Wang, M.; Lai, T.; Zhang, M. Self-Supervised Transformer-Based Pre-Training Method with General Plant Infection Dataset. In *Pattern Recognition and Computer Vision, Proceedings of the 7th Chinese Conference, PRCV 2024, Urumqi, China, 18–20 October 2024*; Lin, Z., Cheng, M.-M., He, R., Ubul, K., Silamu, W., Zha, H., Zhou, J., Liu, C.-L., Eds.; Springer Nature: Singapore, 2025; pp. 189–202.
18. Cui, K.; Tang, W.; Zhu, R.; Wang, M.; Larsen, G.D.; Pauca, V.P.; Alqahtani, S.; Yang, F.; Segurado, D.; Fine, P.; et al. Real-Time Localization and Bimodal Point Pattern Analysis of Palms Using UAV Imagery. *arXiv* **2024**, arXiv:2410.11124.
19. Wu, A.-Q.; Li, K.-L.; Song, Z.-Y.; Lou, X.; Hu, P.; Yang, W.; Wang, R.-F. Deep Learning for Sustainable Aquaculture: Opportunities and Challenges. *Sustainability* **2025**, *17*, 5084. [\[CrossRef\]](#)
20. Girshick, R. Fast R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. Cui, G.; Zhang, L. Improved Faster Region Convolutional Neural Network Algorithm for UAV Target Detection in Complex Environment. *Results Eng.* **2024**, *23*, 102487. [\[CrossRef\]](#)
22. Butler, J.; Leung, H. A Novel Keypoint Supplemented R-CNN for UAV Object Detection. *IEEE Sens. J.* **2023**, *23*, 30883–30892. [\[CrossRef\]](#)
23. Zhang, L.; Wang, M.; Ding, Y.; Bu, X. MS-FRCNN: A Multi-Scale Faster RCNN Model for Small Target Forest Fire Detection. *Forests* **2023**, *14*, 616. [\[CrossRef\]](#)
24. 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, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
25. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
27. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
28. Cao, S.; Wang, T.; Li, T.; Mao, Z. UAV Small Target Detection Algorithm Based on an Improved YOLOv5s Model. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103936. [\[CrossRef\]](#)
29. Duan, S.; Wang, T.; Li, T.; Yang, W. M-YOLOv8s: An Improved Small Target Detection Algorithm for UAV Aerial Photography. *J. Vis. Commun. Image Represent.* **2024**, *104*, 104289. [\[CrossRef\]](#)
30. Luo, X.; Wu, Y.; Zhao, L. YOLOD: A Target Detection Method for UAV Aerial Imagery. *Remote Sens.* **2022**, *14*, 3240. [\[CrossRef\]](#)
31. Zhang, H.; Sun, W.; Sun, C.; He, R.; Zhang, Y. Hsp-Yolov8: Uav Aerial Photography Small Target Detection Algorithm. *Drones* **2024**, *8*, 453. [\[CrossRef\]](#)
32. Luo, X.; Wu, Y.; Wang, F. Target Detection Method of UAV Aerial Imagery Based on Improved YOLOv5. *Remote Sens.* **2022**, *14*, 5063. [\[CrossRef\]](#)
33. Shang, J.; Wang, J.; Liu, S.; Wang, C.; Zheng, B. Small Target Detection Algorithm for UAV Aerial Photography Based on Improved YOLOv5s. *Electronics* **2023**, *12*, 2434. [\[CrossRef\]](#)
34. Zhang, X.; Zuo, G. Small Target Detection in UAV View Based on Improved YOLOv8 Algorithm. *Sci. Rep.* **2025**, *15*, 421. [\[CrossRef\]](#)
35. Qin, Z.; Chen, D.; Wang, H. MCA-YOLOv7: An Improved Uav Target Detection Algorithm Based on Yolov7. *IEEE Access* **2024**, *12*, 42642–42650. [\[CrossRef\]](#)
36. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. Phsi-Rtdetr: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography. *Drones* **2024**, *8*, 240. [\[CrossRef\]](#)

37. Chang, Y.; Li, D.; Gao, Y.; Su, Y.; Jia, X. An Improved YOLO Model for UAV Fuzzy Small Target Image Detection. *Appl. Sci.* **2023**, *13*, 5409. [\[CrossRef\]](#)
38. Xiong, X.; He, M.; Li, T.; Zheng, G.; Xu, W.; Fan, X.; Zhang, Y. Adaptive Feature Fusion and Improved Attention Mechanism Based Small Object Detection for UAV Target Tracking. *IEEE Internet Things J.* **2024**, *11*, 21239–21249. [\[CrossRef\]](#)
39. Yang, C.; Li, X.; Zhang, H.; Meng, Y.; Zhang, R.; Yuan, R. Uav Small Target Detection in Complex Scenes Based on Improved Yolov8s. In Proceedings of the 2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Dalian, China, 7–9 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1791–1798.
40. Zhang, Q.; Wang, X.; Shi, H.; Wang, K.; Tian, Y.; Xu, Z.; Zhang, Y.; Jia, G. BRA-YOLOv10: UAV Small Target Detection Based on YOLOv10. *Drones* **2025**, *9*, 159. [\[CrossRef\]](#)
41. Jocher, G.; Chaurasia, A.; Qiu, J. *Ultralytics YOLOv8*; Ultralytics: Frederick, MD, USA, 2023.
42. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
43. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
44. Han, Q.; Fan, Z.; Dai, Q.; Sun, L.; Cheng, M.-M.; Liu, J.; Wang, J. On the Connection between Local Attention and Dynamic Depth-Wise Convolution. *arXiv* **2021**, arXiv:2106.04263.
45. Lau, K.W.; Po, L.-M.; Rehman, Y.A.U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in Cnn. *Expert Syst. Appl.* **2024**, *236*, 121352. [\[CrossRef\]](#)
46. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic Head: Unifying Object Detection Heads with Attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
47. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
48. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
49. Zhang, H.; Xu, C.; Zhang, S. Inner-Iou: More Effective Intersection over Union Loss with Auxiliary Bounding Box. *arXiv* **2023**, arXiv:2311.02877.
50. Ma, S.; Xu, Y. Mpdious: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
51. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226.
52. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. Yolov9: Learning What You Want to Learn Using Programmable Gradient Information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
53. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
54. Khanam, R.; Hussain, M. Yolov11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
55. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [\[CrossRef\]](#)
56. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [\[CrossRef\]](#)
57. Tu, Y.-H.; Wang, R.-F.; Su, W.-H. Active Disturbance Rejection Control—New Trends in Agricultural Cybernetics in the Future: A Comprehensive Review. *Machines* **2025**, *13*, 111. [\[CrossRef\]](#)
58. Wang, R.-F.; Tu, Y.-H.; Chen, Z.-Q.; Zhao, C.-T.; Su, W.-H. A Lettpoint-Yolov11l Based Intelligent Robot for Precision Intra-Row Weeds Control in Lettuce. Available at SSRN 5162748. 2025. Available online: <https://ssrn.com/abstract=5162748> (accessed on 16 April 2025).
59. Ding, H.; Zhao, L.; Yan, J.; Feng, H.-Y. Implementation of Digital Twin in Actual Production: Intelligent Assembly Paradigm for Large-Scale Industrial Equipment. *Machines* **2023**, *11*, 1031. [\[CrossRef\]](#)
60. Liu, D.; Li, Z.; Wu, Z.; Li, C. Digital Twin/MARS-CycleGAN: Enhancing Sim-to-Real Crop/Row Detection for MARS Phenotyping Robot Using Synthetic Images. *J. Field Robot.* **2025**, *42*, 625–640. [\[CrossRef\]](#)

61. Yang, Z.-X.; Li, Y.; Wang, R.-F.; Hu, P.; Su, W.-H. Deep Learning in Multimodal Fusion for Sustainable Plant Care: A Comprehensive Review. *Sustainability* **2025**, *17*, 5255. [[CrossRef](#)]
62. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **2020**, *32*, 829–864. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.