Rice Pest Detection Using Enhanced YOLOv8n with Multilayer Contextual Attention and Deformable Snake Convolution

Shuangyuan Li[1]*, Jianglong Lin[2], Jiaming Liang[3], Tianyu Li[4]
Information Center, Jilin University of Chemical Technology, China[1]
School of Information and Control Engineering, Jilin University of Chemical Technology, China[2, 4]
Beijing Chaitin Technology Co., Ltd,Beijing, 100101, China[3]
No. 45, Chengde Street, Longtan District, Jilin 132022, P. R. China[4]

***Abstract*—Accurate and intelligent detection of rice pests is critical for ensuring food security and advancing precision agriculture. However, due to the small size, irregular morphology, dense distribution, and complex backgrounds of pest targets, traditional lightweight detection models often suffer from low recall and poor localization accuracy in real-world paddy field environments. To address these challenges, this study proposes an enhanced detection model, YOLO-Rice, based on the YOLOv8n framework. First, a Multilayer Contextual Attention (MLCA) module is embedded after the SPPF layer to collaboratively fuse channel, spatial, local, and global contextual information, thereby enhancing the model's sensitivity to subtle pest features. Second, the original C2f structure is redesigned into C2f-DS, which integrates Dynamic Snake Convolution (DSConv) to improve adaptive perception of deformable pest contours and irregular morphological edges. Finally, the conventional CIoU loss is replaced with a WIoU loss to guide the network to focus more effectively on hard-to-fit small and occluded targets. Extensive experiments on a self-constructed dataset of 11 common rice pest species demonstrate that YOLO-Rice achieves 84.8% Precision, 69.9% Recall, 78.7% mAP@0.5, and 63.4% mAP@0.5:0.95, representing significant improvements over the baseline YOLOv8n model. The proposed approach achieves an excellent balance between detection accuracy and computational efficiency, making it highly suitable for real-time deployment on UAVs and edge devices in agricultural pest monitoring applications.**

***Keywords*—Rice pest detection; YOLOv8n; deep learning; real-time detection**

I. Introduction

In recent years, with the continuous advancement of deep learning technology, image-based intelligent detection methods have become an essential pathway to ensure food security and promote agricultural modernization [1]. As one of the world's most important staple crops, rice is highly susceptible to a variety of pest infestations during its growth cycle, including Delphacidae, Cicadellidae, Crambidae, and Curculionidae. These pests not only hinder the healthy growth of rice but can also lead to severe yield reduction or even total crop failure, posing a significant threat to national food security. Consequently, early identification and precise localization of rice pests have become critical research focuses in modern agricultural pest management [2].

Traditional pest monitoring approaches, such as manual inspection and trap-based surveillance, suffer from high labor intensity, low efficiency, and subjective bias, making them unsuitable for large-scale field monitoring. With the rapid development of unmanned aerial vehicle (UAV) remote sensing and the increasing availability of agricultural imagery, computer vision – based automatic pest detection has emerged as a promising solution. Among numerous object detection frameworks, the YOLO (You Only Look Once) series has gained significant attention due to its end-to-end architecture, fast inference speed, and high detection accuracy, making it particularly suitable for agricultural visual tasks[3-4].

As the latest member of the YOLO family, YOLOv8 introduces an anchor-free mechanism and a lightweight C2f structure, achieving a better balance between precision and computational efficiency. However, when applied directly to rice pest detection[5], it still faces several challenges: 1) The pest targets are generally small, densely distributed, and often confused with complex backgrounds such as veins or shadowed regions, leading to false detections and missed targets; 2) Pest bodies exhibit irregular shapes, occlusion, and deformation, making it difficult for standard convolutions to accurately capture boundary features[6]; 3) The conventional bounding box regression loss functions, such as CIoU, show insufficient adaptability to hard samples and slow convergence during training, which limits the final detection accuracy. To overcome these challenges, this study proposes a structure- and loss-enhanced detection model based on YOLOv8n, specifically designed for high-precision detection of multiple rice pest species under complex field environments. The main contributions of this work are summarized as follows:

A Multilayer Contextual Attention (MLCA) mechanism is introduced after the SPPF module to integrate spatial, channel, local, and global contextual features. This enhances the model's focus on pest regions and improves its ability to recognize small and occluded targets. Compared to conventional attention modules such as SE and CBAM, MLCA achieves superior feature discrimination without adding significant computational overhead [7].

A redesigned C2f-DS module based on Dynamic Snake Convolution (DSConv) is proposed. By introducing dynamic offset learning, DSConv adaptively adjusts convolution

*Corresponding author.

sampling positions and shapes, enabling "snake-like" contour perception. This improves the model's capability to capture irregular pest boundaries and deformable structures, particularly for non-rigid shapes such as the feeding edges of Cnaphalocrocis medinalis.

The Weighted IoU (WIoU) loss function is employed to replace the traditional CIoU, enhancing the model's robustness in bounding-box regression. WIoU dynamically assigns gradient weights to samples with varying difficulty, encouraging the network to focus more on hard-to-fit or occluded targets, thereby improving convergence stability and overall mAP performance.

To validate the proposed approach, a large-scale rice pest dataset containing 11 common pest species was constructed. Experimental results demonstrate that the improved YOLOv8n model achieves significantly higher detection accuracy and robustness in complex, cluttered, and deformable pest scenarios while maintaining lightweight deployment capability. Notably, the proposed model outperforms mainstream detectors such as YOLOv5, YOLOv7, and YOLOv11 in key metrics including mAP, Precision, and Recall, verifying its effectiveness and practical potential for real-world rice pest detection tasks.

In summary, this study provides a novel structural optimization approach for object detection in agricultural vision applications and offers a feasible technical foundation for the intelligent deployment of rice pest monitoring systems. Future research will focus on improving the model's cross-regional generalization, multimodal data fusion, and extreme lightweight deployment on agricultural embedded platforms, thereby advancing pest control toward greater efficiency, intelligence, and precision.

II. Related Research

In recent years, with the deep integration of unmanned aerial vehicles (UAVs) and deep learning technologies in precision agriculture, vision-based rice pest detection has become a key research focus. Object detection techniques have evolved from early two-stage frameworks to more efficient single-stage and anchor-free approaches, with researchers exploring model lightweighting, small-object enhancement, deformable feature extraction, and regression loss optimization. Early two-stage detectors such as Faster R-CNN [8] and Mask R-CNN [9] demonstrated strong performance in locating crop diseases and large pest regions. For instance, Ren et al. [10] combined an RPN with a fine classifier for wheat leaf spot detection, achieving 67% mAP, though its inference delay of over 200 ms limited UAV-based real-time deployment. He et al. [11] improved feature alignment through ROI-Align, significantly enhancing lesion boundary segmentation accuracy, but the method remained computationally expensive due to dense proposal generation.

To balance speed and precision, single-stage detectors have gained prominence. Liu et al. [12] introduced SSD, which enables end-to-end detection using multi-scale feature layers and achieves 59 FPS inference. However, it exhibits high miss rates for small pests such as Cnaphalocrocis medinalis and Delphacidae, whose body sizes span only tens of pixels. Lin et al. [13] proposed RetinaNet with Focal Loss to reduce gradient dominance by easy samples, significantly improving small-object recall. Redmon and Farhadi [14] developed YOLOv3, which leverages multi-scale pyramid prediction and efficient convolution to reach 45 FPS, showing robust performance in heavily occluded rice leaf environments. Later, Bochkovskiy et al. [15] released YOLOv4, integrating CSPDarkNet and PANet to better balance speed and accuracy. Jocher et al. [16] further optimized YOLOv5 with automatic anchor learning, depthwise separable convolution, and channel pruning, reducing parameters below 8M. Studies applying YOLOv5 in agricultural pest detection incorporated ECA attention and BiFPN fusion, achieving a 4 – 6% increase in mAP. Li et al. [17] developed a lightweight version of YOLOv7 for Noctuidae detection, reducing model size by 30% and enabling 30 FPS real-time inference on UAV platforms; however, detection of dense Crambidae larvae in rice fields remained unsatisfactory. More recently, Fan et al. [18] embedded SimAM attention and NWD loss into the YOLOv8 backbone and neck for grape leaf disease detection, achieving a 90.4% mAP@0.5 and demonstrating superior fine-grained focus. Nevertheless, their approach did not specifically address pest morphology or leaf occlusion, nor was it validated on diverse rice pest datasets.

Addressing the challenges of small, irregular, and occluded pest targets in paddy fields, Dai et al. [19] proposed Deformable ConvNets, which allow convolutional sampling points to adaptively shift according to object contours, thereby improving feature alignment under occlusion and deformation. Zhu et al. [20] further enhanced this concept through the Modulated DCNv2, incorporating modulation mechanisms to refine sampling precision. In addition, bounding box regression losses have been shown to significantly affect localization accuracy for small objects. While the traditional IoU loss focuses only on overlap, it fails to account for center distance and aspect ratio. To address this, Rezatofighi et al. [21] proposed GIoU, and Zheng et al. [22] extended it to DIoU and CIoU, incorporating geometric penalties that improved convergence speed and localization precision. Building upon this, Bao et al. [23] introduced WIoU v3, which dynamically weights samples according to quality, emphasizing harder instances during training. This approach improved mAP@0.5:0.95 by 4.1% in dense rice pest detection scenarios.

Despite significant progress in lightweight architectures, attention mechanisms, deformable convolutions, and regression loss optimization, most existing studies focus on isolated module improvements. A unified framework that simultaneously addresses small target sensitivity, irregular shape modeling, and robust bounding box regression—while maintaining real-time performance—is still lacking. To bridge this gap, this study integrates Multilayer Contextual Attention (MLCA), Dynamic Snake Convolution (DSConv), and WIoU loss into the anchor-free YOLOv8n architecture. This comprehensive approach aims to achieve superior detection accuracy and robustness for rice pest detection under complex field conditions, enabling efficient deployment on UAV and edge platforms.

III. Proposed Method

This study focuses on the task of rice pest detection based on YOLOv8n, addressing challenges such as small target size,

weak texture features, and irregular pest contours. Based on the original framework, the following improvements are proposed: Firstly, Multilayer Contextual Attention (MLCA) is embedded after the SPPF module to fuse channel, spatial, local, and global contextual features, enhancing the model's responsiveness to weak textures in pest regions [24]. Secondly, the C2f-DS module is designed by replacing some standard convolutions in the original C2f structure with Dynamic Snake Convolution (DSConv), which adaptively learns offsets to capture irregular edges and deformable pest contours. Finally, Weighted IoU (WIoU) loss is adopted to replace CIoU, dynamically assigning gradient weights to samples of varying difficulty, thereby improving the robustness and convergence speed of bounding box regression for small pest targets.

While maintaining the lightweight advantages of YOLOv8n, these improvements jointly optimize the model's perception and localization of "small, dense, and weak" pest targets without significantly increasing computational cost. The overall framework of the proposed YOLO-Rice model is shown in Fig. 1, and the detailed design of each module is described in the following subsections.

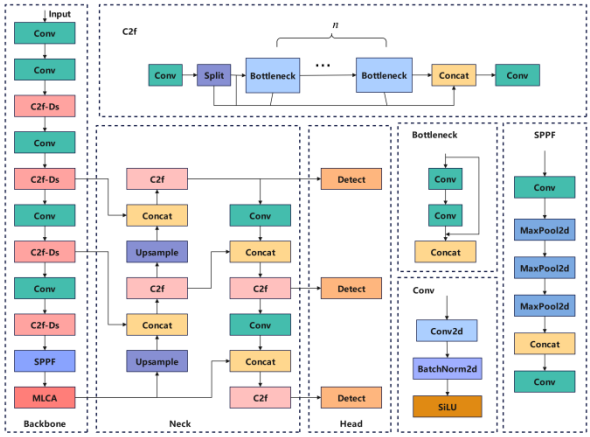

Fig. 1. Overall architecture of the proposed YOLO-Rice model.

A. Reconstruction of the Backbone Network

Rice pest targets are often extremely small and visually similar to the leaf background in both color and texture. Traditional convolutional operations struggle to distinguish pests from background interference, leading to frequent false detections and missed detections. Therefore, it is essential to enhance the model's focus on pest regions through an attention mechanism while maintaining lightweight efficiency. Existing attention modules, such as SE and CBAM, perform one-dimensional weighting along either the channel or spatial axis, which can result in the loss of local information for small targets [25].

To address this limitation, a Multilayer Contextual Attention (MLCA) mechanism is proposed and embedded after the SPPF module. The improved module fuses local pooling, global pooling, and one-dimensional convolution to achieve bidirectional attention across channel and spatial dimensions while integrating both local and global contexts. This structure provides more fine-grained feature enhancement, strengthening the network's feature extraction capability. The structure of MLCA is shown in Fig. 2.

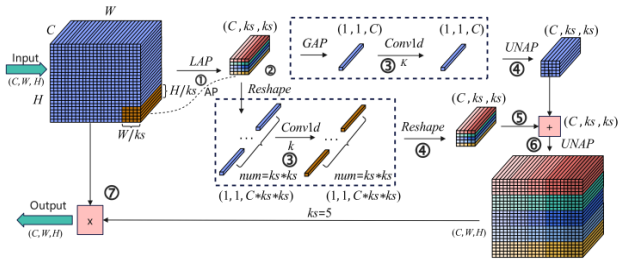

Fig. 2. MLCA structure.

As illustrated in Fig. 2, the MLCA module achieves fine-grained pest feature reinforcement through two collaborative dimensions: local – global complementarity and channel – spatial interaction. Specifically, the input feature map is first processed by Local Average Pooling (LAP) to extract coarse local features from subdivided sub-blocks. It is then divided into two branches: a global branch that applies Global Average Pooling (GAP) to highlight overall field-level context, and a local branch that reshapes the LAP output to preserve channel-level details within each subregion. Both branches are passed through 1D convolution (Conv1d) layers for channel compression and feature transformation. Subsequently, the features are restored to the original spatial resolution through reshaping and upsampling via Unpooled Average Pooling (UNAP), and the two branches are fused along the channel dimension to achieve semantic – detail interaction. Finally, the fused result is expanded to the input resolution and multiplied element-wise with the original feature map to generate a weighted feature map with stronger pest region responses. The functions of the key operations in the MLCA module are explained in Table I.

TABLE I. EXPLANATION OF KEY OPERATIONS IN THE MLCA MODULE

Operation	Role	Output
LAP	Preserve fine-grained spatial details (e.g., pest edges, feeding marks)	Local texture features
GAP	Compresses global context (e.g., leaf-background separation)	Global channel descriptor
UNAP	Aggregates neighborhood information to bridge local and global scales	Medium-scale context
Conv1D	Models inter-channel dependencies (kernel size=5)	Channel attention weights
Feature Fusion	Sums LAP, GAP, UNAP, then multiplies by the original feature map	Refined multi-scale features

To demonstrate the lightweight efficiency of MLCA, it is compared with SE and CBAM in terms of additional parameters, FLOPs, and mAP improvement. The results in Table II show that MLCA introduces only 0.12M parameters and 0.48 GFLOPs, significantly lower than CBAM, while achieving the highest mAP gain (+3.7% mAP@0.5). This confirms that MLCA provides superior feature discrimination with minimal computational overhead, making it ideal for real-time UAV deployment.

TABLE II. COMPARISON OF ATTENTION MODULES ON YOLOV8N BASELINE

Attention Module	Parameters (M)	FLOPs (G)	mAP@0.5 (%)	Δ mAP@0.5 (%)
None (baseline)	5.30	8.1	70.4	-
SE	5.41 (+0.11)	8.2	72.9	+2.5
CBAM	5.45 (+0.15)	8.3	74.2	+3.8
MLCA	5.42 (+0.12)	8.3	74.1	+3.7

B. C2f-DS Module with Dynamic Snake Convolution

Rice pests often exhibit non-rigid, elongated, or curved morphological structures. For instance, Crambidae larvae create irregular chewing marks along leaf edges; Cicadellidae closely resemble leaf textures and are easily overlooked; and Curculionidae often appear partially occluded due to their shrimp-like shapes. Traditional convolutions with fixed sampling points cannot adapt to such boundary variations, leading to decreased detection accuracy.

To overcome these limitations, Dynamic Snake Convolution (DSConv)—a deformable convolution variant capable of adaptively focusing on elongated and curved structures—is integrated into the C2f framework, resulting in a novel module termed C2f-DS. The architecture of DSConv is shown in Fig. 3.

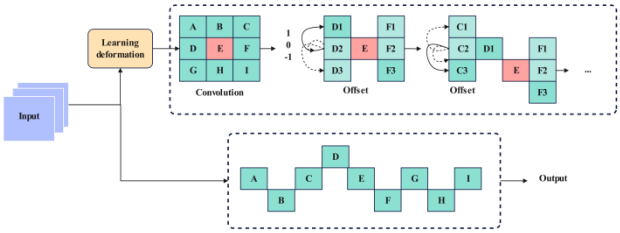

Fig. 3. Architecture of dynamic snake convolution.

The C2f module in YOLOv8n is a lightweight feature extraction unit that balances computational efficiency and representational power. DSConv is embedded into C2f rather than the backbone's basic convolutions or the neck for the following reasons. Firstly, feature hierarchy: C2f operates at multiple scales within the backbone, capturing both fine-grained details and semantic information. Embedding DSConv in C2f allows deformable sampling to be applied across a range of receptive fields, which is essential for detecting pests of varying sizes and deformations. Secondly, Computational efficiency: Replacing all backbone convolutions with DSConv would significantly increase parameters and inference time. C2f already contains a split-and-merge design; replacing only a portion of its standard convolutions with DSConv minimizes additional cost while preserving the benefits of deformable convolution. Finally, Residual integration: The C2f structure inherently includes residual connections, which help stabilize training when introducing dynamic offsets. This mitigates the risk of gradient oscillation caused by excessive deformation [26].

As illustrated in Fig. 4, the proposed C2f-DS module introduces an iterative "snake-like" offset mechanism based on deformable convolution and deeply integrates it within the original C2f structure. This design improves feature extraction while maintaining computational efficiency. Specifically, C2f-DS dynamically samples local pest regions, adaptively capturing irregular leaf-edge contours caused by Crambidae larvae and subtle differences between Cicadellidae nymphs and leaf veins. This significantly enhances the model's ability to fit small pest boundaries. Unlike stacking multiple deformable convolutions, C2f-DS replaces only part of the standard convolutional kernels in C2f, preserving the residual and channel-split design. As a result, the number of parameters and computational cost remain close to those of YOLOv8n, with negligible impact on inference speed.

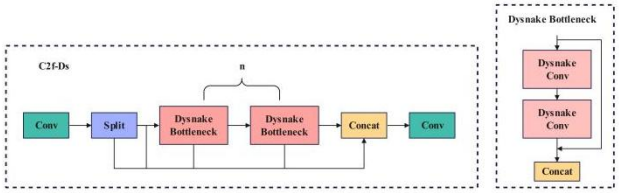

Fig. 4. Structure of the C2f-DS module.

TABLE III. COMPARISON BETWEEN DSCONV (C2F-DS) AND STANDARD CONVOLUTION

Aspect	Standard Convolution	DSConv (C2f-DS)
Sampling positions	Fixed grid, axis-aligned	Adaptive offsets, distributed along target contours
Adaptability to pest boundaries	Low – cannot fit curved or deformed edges	High – fits irregular contours of Crambidae, Curculionidae, etc.
Offset learning	Not applicable	Learned via lightweight auxiliary convolution (~3×3 depthwise)
Parameter increase	Baseline	<5% increase compared to standard C2f
Inference speed impact	Baseline	Comparable (~1–3 FPS drop on RTX 3060)
Best suited for	General feature extraction	Small, elongated, occluded, or clearly deformed pest targets

Unlike standard convolution with a fixed sampling grid, DSConv enhances the perception of irregular pest boundaries by learning adaptive offsets. Specifically, a lightweight auxiliary convolutional layer (typically a 3×3 depthwise convolution) takes the input feature map and predicts a 2D offset field $\triangle p_n$ for each sampling point p_n of the kernel. During training, these offsets are jointly optimized with the main network weights via backpropagation; bilinear interpolation is used to ensure differentiability of fractional sampling positions. At inference, the kernel samples features at shifted positions $p_n + \Delta p_n$, allowing the receptive field to "snake" along elongated or curved pest contours (e.g., the feeding edge of Cnaphalocrocis medialis larvae). Table III compares the key differences between DSConv in the C2f-DS module and standard convolution. It can be seen that DSConv significantly improves boundary fitting for deformable targets while maintaining a lightweight design.

Furthermore, a residual connection is introduced in the iterative offset process to improve training stability and mitigate gradient oscillations caused by excessive deformation. This module significantly enhances detection robustness and precision for small rice pests while maintaining a lightweight structure, making it well-suited for real-time UAV-based agricultural applications.

C. Weighted IoU (WIoU) Loss Function

In small-object pest detection, the precision of bounding box regression plays a crucial role in determining overall detection performance. The original YOLOv8n employs the Complete IoU (CIoU) loss, which simultaneously optimizes IoU, center distance, and aspect ratio. However, for extremely small, irregularly shaped, and background-blended pest targets, CIoU tends to overemphasize high-IoU samples, making it less effective for low-IoU or fuzzy-boundary samples. This often results in inaccurate bounding box regression and higher miss rates [27].

To address this issue, CIoU is replaced with the Weighted IoU (WIoU) loss function. WIoU introduces a dynamic focusing mechanism that adjusts gradient weights according to prediction quality, enabling the model to pay more attention to hard-to-fit small pest boundaries. The relationship between the ground truth and predicted bounding boxes is illustrated in Fig. 5.

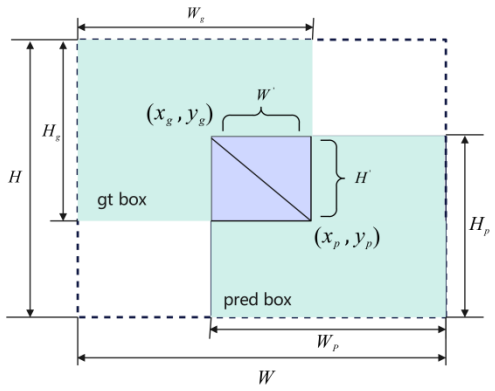

Fig. 5. Illustration of ground truth box and predicted box.

The key mathematical difference between CIoU and WIoU lies in how gradient weights are assigned. CIoU uses a fixed penalty term based on center distance and aspect ratio, treating all samples uniformly. In contrast, WIoU introduces a dynamic focusing coefficient that reduces the contribution of easy samples (high IoU) and amplifies the gradient for hard samples (low IoU). The WIoU loss is calculated as shown in Eq. (1)-(3).

$$L_{WIOU} = r \cdot L_{IOU} \tag{1}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{2}$$

$$\beta = \frac{(W - W_g)^2 + (H - H_g)^2}{{W_g}^2 + {H_g}^2} \tag{3}$$

where, β is the outlier degree, and α, δ are hyperparameters controlling the focusing strength. The weighting factor r dynamically adjusts the gradient magnitude: for samples with small β (easy samples, high IoU), r is small, reducing their contribution; for samples with large β (hard samples, low IoU), r approaches 1, emphasizing the loss. This mechanism forces the network to prioritize poorly localized small targets, resulting in more precise bounding box predictions, especially for occluded or irregularly shaped pests.

IV. Experiments and Analysis

A. Dataset and Experimental Setup

This study targets 11 common rice pest species: Curculionidae, Delphacidae, Cicadellidae, Phlaeothripidae, Cecidomyiidae, Hesperiidae, Crambidae, Chloropidae, Ephydridae, Noctuidae, and Thripidae. The dataset contains 4,559 pest instances manually annotated in YOLO format using LabelImg. The images were split into training, validation, and test sets at a ratio of 7:2:1, resulting in 3,191, 911, and 457 images, respectively, while maintaining class balance across subsets. As shown in Fig. 6, Delphacidae and Cicadellidae dominate the dataset, while Thripidae and Noctuidae are underrepresented, exhibiting a clear long-tail distribution. The bounding box widths and heights are mostly concentrated in small-scale regions (normalized values between 0.0 and 0.4), indicating that most pest targets are small and morphologically diverse.

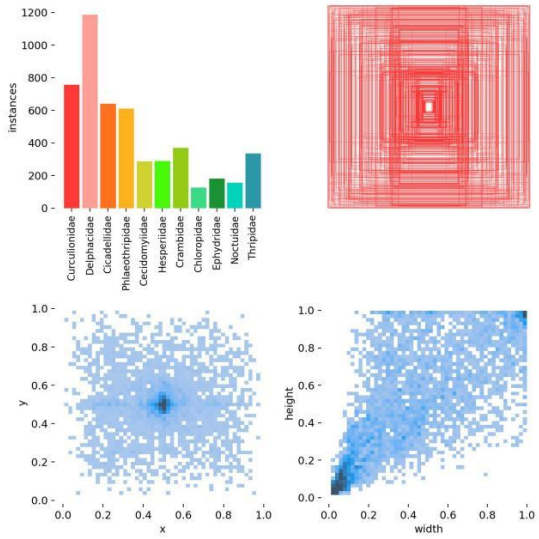

Fig. 6. Distribution of target categories and sizes in the training set.

Partial samples of rice pests are shown in Fig. 7.

Fig. 7. Partial samples of rice pests.

All experiments were conducted on a Windows 11 system with an Intel Core i9-13900H CPU and an NVIDIA GeForce RTX 3060 GPU. Training was accelerated using PyTorch 2.1.0 with CUDA 11.8. All models were trained under identical hyperparameters: the SGD optimizer with 200 epochs, a batch size of 64, and an initial learning rate of 0.01.

B. Evaluation Metrics

To comprehensively evaluate the proposed method, three key metrics are employed: Precision (P), Recall (R), and Mean Average Precision (mAP) to validate the detection accuracy, while Parameter Count and Frame Rate (FPS) assess the model compactness and speed. These metrics are defined as shown in Eq. (4)-(7).

$$P = \frac{TP}{(TP + FP)} \tag{4}$$

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

$$mAP = \frac{\sum_{i=1}^{c} AP_i}{c} \tag{6}$$

$$FPS = \frac{1000ms}{(t1 + t2 + t3)} \tag{7}$$

Here, TP denotes the number of correctly predicted positive samples, FP indicates the number of incorrectly predicted negative samples, FN represents the number of missing targets among the correct ones, c stands for the number of defect categories, t1 is the image preprocessing time, t2 is the image inference time, and t3 is the post-processing time.

C. Ablation Experiments

To verify the effectiveness of the proposed Multilayer Contextual Attention (MLCA), C2f-DS module, and Weighted IoU (WIoU) loss function in the rice pest detection task, a series of ablation experiments were designed and conducted. All experiments were based on the YOLOv8n architecture as the baseline model. Under identical dataset partitioning, data augmentation strategies, and hyperparameter configurations, each improvement module was progressively introduced to quantitatively evaluate the individual contributions and synergistic effects. The experimental results are shown in Table IV.

TABLE IV. ABLATION EXPERIMENT RESULTS

Configuration	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters (M)
Baseline	0.800	0.632	0.704	0.562	5.3
+ MLCA	0.823	0.658	0.741	0.595	5.4
+ C2f-DS	0.816	0.652	0.734	0.588	6.8
+ WIoU	0.812	0.645	0.729	0.580	5.3
MLCA + C2f-DS	0.836	0.678	0.762	0.612	6.9
MLCA + WIoU	0.828	0.665	0.752	0.603	5.4
C2f-DS + WIoU	0.825	0.662	0.748	0.599	6.8
YOLO-Rice	0.848	0.699	0.787	0.634	7.8

From the experimental results in Table IV, it can be observed that each improvement module effectively enhances model performance. When introduced individually, MLCA, C2f-DS, and WIoU improve mAP@0.5 by 3.7%, 3.0%, and 2.5%, respectively, validating their independent effectiveness in feature enhancement, deformable modeling, and regression optimization. In terms of module combinations, the combination of MLCA and C2f-DS achieves the most prominent synergistic effect, with mAP@0.5 reaching 0.762, representing improvements of 2.1% and 2.8% compared to using each module alone, indicating a significant complementary effect between the attention mechanism and deformable convolution. The combinations of MLCA with WIoU and C2f-DS with WIoU also demonstrate favorable synergistic effects. Finally, the YOLO-Rice model, which fully integrates all three modules, achieves optimal performance across all evaluation metrics, with Precision, Recall, mAP@0.5, and mAP@0.5:0.95 reaching 0.848, 0.699, 0.787, and 0.634, respectively, representing improvements of 4.8%, 6.7%, 8.3%, and 7.2% compared to the baseline YOLOv8n. Although the number of parameters increases from 5.3M to 7.8M, the model still maintains a real-time inference speed of 62 FPS, meeting the deployment requirements for UAVs and edge devices. In summary, the ablation experiments fully validate the individual effectiveness and combined synergy of the MLCA, C2f-DS, and WIoU modules in the rice pest detection task. The three modules work together from three dimensions—feature enhancement, deformable modeling, and regression optimization—forming a comprehensive detection optimization system.

D. Comparison of Dysnake Conv Embedding Across Different C2f Layers

To further investigate the influence of DSConv placement within the backbone, embedding DSConv into C2f_1 (shallow), C2f_2 (middle), C2f_3 (deep), and C2f_4, as well as combinations thereof, was compared against the baseline model with standard convolutions. Only the C2f structure at the specified layers was modified; all other configurations remained unchanged. The results are shown in Table V.

TABLE V. EXPERIMENTAL RESULTS OF DSCONV EMBEDDING AT DIFFERENT C2F LAYERS

Dysnake Conv Embed location	Precision	Recall	mAP@0.5	mAP@0.5：0.95	Parameters (M)
No replacement	0.800	0.632	0.704	0.562	5.3
C2f_1	0.810	0.648	0.718	0.576	5.6
C2f_2	0.816	0.652	0.734	0.588	5.8
C2f_3	0.808	0.644	0.725	0.581	6.8
C2f_4	0.803	0.640	0.717	0.573	6.1
C2f_2+C2f_3	0.823	0.662	0.745	0.599	6.5
C2f_1 + C2f_2 + C2f_3	0.829	0.668	0.752	0.605	7.0
C2f_3 + C2f_4	0.819	0.656	0.740	0.593	6.3
Replace All	0.833	0.670	0.760	0.610	7.4

The results indicate that applying DSConv across all four C2f layers significantly enhances multi-scale feature extraction—from fine-grained leaf textures to overall pest semantics—especially for small, deformed, and partially occluded pests such as Crambidae and Delphacidae. Compared to replacing only middle and deep layers, the full-layer configuration increased mAP@0.5 to 0.760 (+1.5%) and mAP@0.5:0.95 to 0.610 (+1.1%), while Precision and Recall reached 0.833 and 0.670, respectively. Although this setup increased parameters by 2.1M, the detection precision and robustness in complex rice-field backgrounds improved significantly, validating the advantage of multi-level collaborative modeling of deformable pest features.

E. Comparison of Attention Mechanisms

This section compares the proposed MLCA with other mainstream attention mechanisms, including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA). All models were built upon YOLOv8n, differing only in the attention module applied. Training configurations were identical across all experiments. The results are presented in Table VI.

TABLE VI. COMPARISON OF ATTENTION MECHANISMS

Model Version	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Baseline_NoAttn	0.800	0.632	0.704	0.562
SE_Attn	0.812	0.645	0.729	0.579
CA_Attn	0.818	0.650	0.735	0.586
CBAM_Attn	0.821	0.654	0.742	0.592
MLCA_Attn	0.823	0.658	0.741	0.595

As shown in Table VI, SE improved mAP@0.5 from 0.704 to 0.729 and mAP@0.5:0.95 from 0.563 to 0.579, demonstrating its effectiveness in suppressing background noise. CA, which captures long-range spatial dependencies, further improved mAP@0.5 to 0.735 and mAP@0.5:0.95 to 0.586. CBAM, integrating both channel and spatial attention, yielded mAP@0.5=0.742 and mAP@0.5:0.95=0.592. The proposed MLCA achieved the highest overall performance, with mAP@0.5:0.95=0.595, Precision=0.823, and Recall=0.658. These results highlight that MLCA provides stronger robustness and localization accuracy by jointly modeling weak pest details and global contextual semantics.

F. Comparison of Loss Functions

This section evaluates the impact of different bounding box regression loss functions. The default CIoU loss in YOLOv8n was replaced with GIoU, DIoU, and WIoU, while maintaining identical training settings. The results are shown in Table VII.

TABLE VII. COMPARISON OF LOSS FUNCTIONS

Model Version	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Baseline_CIoU	0.800	0.632	0.704	0.562
GIoU_Loss	0.801	0.628	0.701	0.553
DIoU_Loss	0.806	0.635	0.710	0.559
WIoU_Loss	0.812	0.645	0.729	0.580

The results show that incorporating center distance and aspect ratio constraints (DIoU, CIoU) improves regression precision and convergence compared to IoU and GIoU. Building upon this, WIoU introduces sample difficulty-aware weighting, allowing the model to focus more on hard-to-detect small or occluded pests. Compared with CIoU, WIoU achieved improvements across all metrics: Precision increased from 0.800 to 0.812, Recall from 0.632 to 0.645, mAP@0.5 by 2.5%, and

mAP@0.5:0.95 by 1.8%. These results confirm that WIoU enhances the alignment between predicted and ground-truth boxes while effectively suppressing training noise from anomalous samples, yielding more stable and generalized detection across multiple IoU thresholds.

G. Comparison with State-of-the-Art Models

To comprehensively evaluate the overall performance of the proposed YOLO-Rice model in the rice pest detection task, this section compares it with several mainstream object detection algorithms. All compared models were trained and tested under identical dataset partitioning, data augmentation strategies, and hyperparameter configurations to ensure fairness. The experimental results are shown in Table VIII.

As shown in Table VIII, YOLO-Rice achieves the best detection performance among all compared models. As traditional detectors, Faster R-CNN and SSD achieve mAP@0.5 of 0.668 and 0.652, respectively, but their large parameter sizes and low inference speeds make them unsuitable for real-time UAV monitoring. Among the lightweight YOLO series models, YOLOv5s, YOLOv7, YOLOv8n, YOLOv9, YOLOv10n, and YOLOv11n all demonstrate a good balance between accuracy and speed. Notably, YOLOv11n achieves a mAP@0.5 of 0.740 and an inference speed of 150 FPS with only 5.5M parameters, exhibiting excellent lightweight performance. Building upon this, the proposed YOLO-Rice model achieves further optimization, with mAP@0.5 reaching 0.787, representing improvements of 8.3% over the baseline YOLOv8n and 4.7% over the latest YOLOv11n. The mAP@0.5:0.95 reaches 0.634, an improvement of 7.2% over YOLOv8n. Precision and Recall reach 0.848 and 0.699, respectively, representing increases of 4.8% and 6.7% compared to YOLOv8n. In terms of inference speed, YOLO-Rice achieves 138 FPS on an RTX 4090, far exceeding the minimum frame rate requirement for real-time detection, fully validating its ability to achieve significant detection accuracy improvements while maintaining lightweight characteristics.

TABLE VIII. PERFORMANCE COMPARISON OF DIFFERENT DETECTION MODELS ON THE RICE PEST DATASET

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS (4090)	Params (M)
Faster R-CNN	0.756	0.582	0.668	0.512	42	41.5
SSD (VGG16)	0.742	0.568	0.652	0.498	85	26.3
YOLOv5s	0.780	0.600	0.680	0.535	156	7.2
YOLOv7	0.790	0.620	0.700	0.552	128	36.9
YOLOv8n	0.800	0.632	0.704	0.562	142	5.3
YOLOv9	0.818	0.655	0.726	0.578	135	8.9
YOLOv10n	0.815	0.655	0.725	0.576	148	5.7
YOLOv11n	0.825	0.670	0.740	0.589	150	5.5
YOLO-Rice	0.848	0.699	0.787	0.634	138	7.8

H. Confusion Matrix Analysis

To further analyze the classification performance of the YOLO-Rice model across different rice pest categories, this section presents the normalized confusion matrices, as shown in Fig. 8. Rows of the confusion matrix represent ground-truth labels, columns denote model predictions, and diagonal elements indicate the correct classification ratios for each category, while off-diagonal elements reflect misclassification cases. The confusion matrix enables intuitive evaluation of the model's recognition capability for individual pest categories and the degree of confusion between classes.

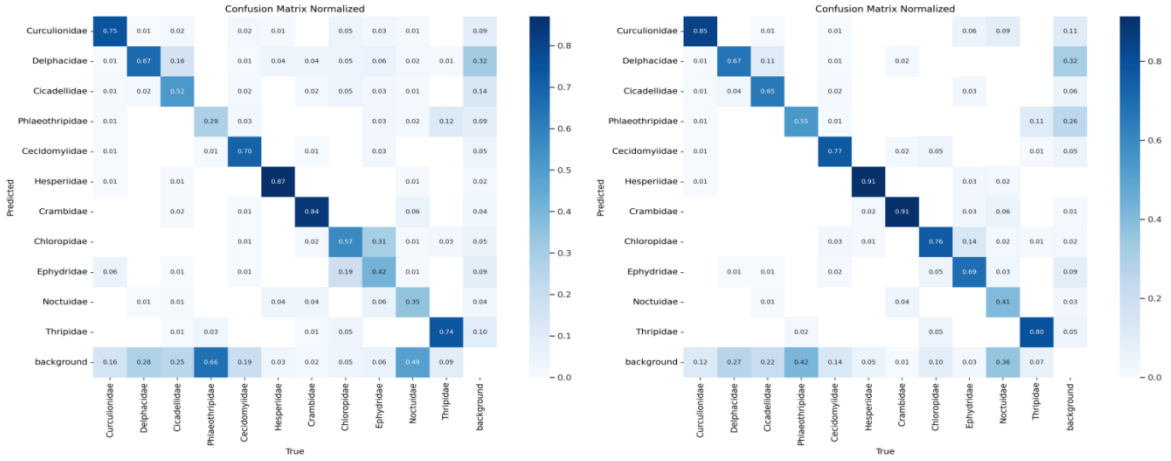

Fig. 8. Confusion matrix: YOLOv8 (left) and YOLO-Rice (right).

As observed from Fig. 8, YOLO-Rice achieves a significant improvement in overall classification accuracy, with substantial increases in the recall rates of most categories. Specifically, the recall for Curculionidae rises from 0.75 to 0.85, for Phlaeothripidae from 0.29 to 0.55, for Ephydridae from 0.42 to 0.69, for Hesperiidae and Crambidae both reaching 0.91, and for Thripidae up to 0.80. Only Delphacidae remains unchanged at 0.67, and Noctuidae sees a slight increase to 0.41. These results demonstrate that the model significantly enhances its ability to identify difficult-to-classify samples and categories with similar morphological features, while also reducing inter-class confusion notably—for instance, the confusion between Ephydridae and Chloropiidae decreases from 0.31 to 0.05. The sole exception is the background class, whose recall drops from 0.66 to 0.42. This decline primarily arises from the model's increased focus on recognizing insect categories, leading to exacerbated misclassification of background regions (e.g., leaf shadows, water reflections, dead spots) as pest taxa such as Phlaeothripidae. While this trade-off is acceptable in some practical applications, it still calls for further optimization to reduce false positives. To address this issue, future work will adopt three strategies: 1) introducing a background suppression loss term that penalizes false positives of the background class; 2) augmenting the training set with diverse negative samples (e.g., empty paddy field images under various lighting conditions); and 3) applying class-balanced sampling or a variant of Focal Loss to down-weight easily confused background patterns. These measures aim to maintain high recall for pest targets while effectively controlling background misclassification, thereby improving overall robustness in real-world paddy field environments. Overall, the proposed YOLO-Rice model, through optimized feature extraction, network architecture, and data augmentation, effectively improves the accuracy and robustness of insect family classification, and the above-mentioned strategies will further enhance its practical deployability.

I. Visualization Analysis of Detection Results

To intuitively demonstrate the practical performance of the proposed YOLO-Rice algorithm in rice pest detection tasks, this study selects detection results from various complex field scenarios for visual comparison and analysis. The Precision–Recall curves before and after improvement are shown in Fig. 9, and the visualization results are presented in Fig. 10.

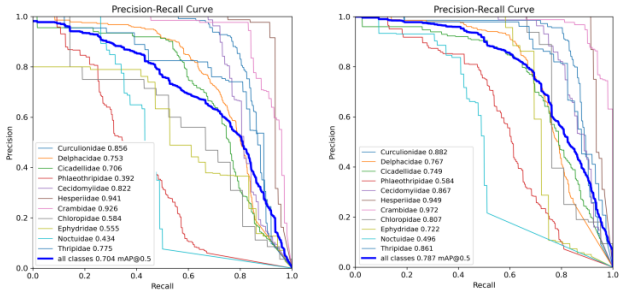

Fig. 9. Precision-recall curve.

Fig. 9 presents the Precision-Recall (PR) curves of the baseline YOLOv8n and the proposed YOLO-Rice model across 11 categories of rice pests. The proposed model achieves a significant overall improvement, with mAP@0.5 increasing from 0.704 to 0.787 (+8.3%). Notably, recognition accuracy for Noctuidae, Delphacidae, and Cicadellidae—categories typically characterized by small size and high occlusion—exhibits remarkable gains of 5.2%, 1.4%, and 4.3%, respectively. This improvement demonstrates that the integration of MLCA and DSConv effectively enhances the model's capacity to distinguish fine-grained pest textures and irregular contours under complex paddy field conditions. The optimized WIoU loss further contributes to improved bounding box regression accuracy, particularly for small and partially occluded targets, resulting in smoother PR curves and higher overall stability across all classes. The before-and-after detection effects are shown in Fig. 10.

As shown in Fig. 10, baseline algorithms such as YOLOv5, YOLOv7, and the lightweight YOLOv8n generally exhibit low detection confidence for Delphacidae, mostly concentrated in the range of 0.28–0.44. These methods suffer from obvious missed detections in dense pest scenarios, poor adaptability to small pests against complex backgrounds such as rice stems and leaves, and insufficient bounding box localization accuracy. Although YOLOv11 improves detection confidence to the range of 0.27–0.68 and slightly increases the number of detected targets, it still suffers from missed detections and exhibits misclassification of pests with similar morphology (e.g., Cicadellidae), indicating limited capability in recognizing densely aggregated targets. In contrast, the proposed algorithm achieves significant improvements in detection completeness, confidence stability, and small-target recognition accuracy. The confidence scores for Delphacidae are stably distributed between 0.57 and 0.91, with most targets exceeding 0.80. In dense scenes, it achieves nearly full coverage detection, with bounding boxes that more closely fit pest contours and no cross-category misclassifications. These results fully validate the robustness and detection advantage of the proposed method in complex rice field scenarios.

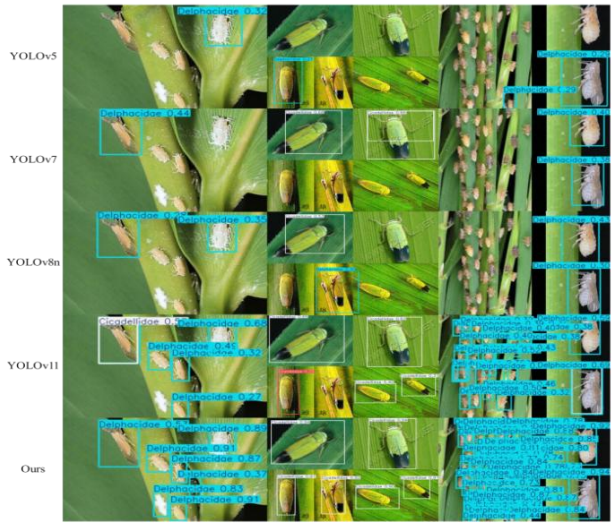

Fig. 10. Visualization comparison of pest detection results by different YOLO algorithms.

Overall, the visualized results substantiate that the proposed YOLO-Rice framework not only achieves quantitative gains in

mAP, Precision, and Recall but also qualitatively improves interpretability and robustness in practical field detection scenarios. These results validate the effectiveness of the proposed attention–deformation–loss joint optimization strategy, highlighting its potential for real-time deployment on UAV-based precision pest monitoring platforms.

V. Conclusion and Future Work

This study proposes YOLO-Rice, a lightweight and high-precision detection model designed for rice pest detection in scenarios characterized by small object size, irregular morphology, dense distribution, and complex backgrounds. Built upon the YOLOv8n framework, the model incorporates three major improvements: 1) the Multilayer Contextual Attention (MLCA) module enhances feature extraction by jointly modeling channel, spatial, and contextual information, thereby improving the model's focus on subtle pest regions; 2) the C2f-DS module integrates Dynamic Snake Convolution (DSConv) to adaptively capture irregular pest contours and deformed structures, significantly enhancing the perception of small and complex targets; and 3) the Weighted IoU (WIoU) loss function replaces CIoU, dynamically adjusting gradient weights to emphasize hard and small samples, thus improving bounding box regression accuracy and robustness.

Experiments on a custom dataset containing 11 common rice pest categories demonstrate that YOLO-Rice achieves a Precision of 84.8%, Recall of 69.9%, mAP@0.5 of 78.7%, and mAP@0.5:0.95 of 63.4%, showing clear superiority over the baseline YOLOv8n. Ablation and comparative experiments verify the effectiveness of each proposed module and the rationality of their integration. Moreover, the model maintains high inference speed (62 FPS on an RTX 3060), satisfying real-time detection requirements for UAV and edge computing platforms.

Notably, the recall of 69.9% indicates that about 30% of pest instances are missed, mainly consisting of extremely small targets (e.g., Thripidae), heavily occluded targets (e.g., Crambidae larvae under leaves), and targets blending into background textures (e.g., Cicadellidae). Future work will address these issues via super-resolution branches, Transformer-based global context modeling, and hard example mining.

Meanwhile, several limitations remain. The dataset was collected from a single growing region, leaving cross-regional generalization unvalidated. The dataset exhibits a long-tail distribution, with Thripidae and Noctuidae underrepresented (AP of 0.52 and 0.48). Test conditions are ideal (daytime, clear weather), excluding low-light, rainy, or motion-blurred scenarios common in UAV deployment. Deployment on edge devices would require further quantization and pruning. Based on the above analysis, future research will focus on: integrating Transformer or self-supervised pre-training for tiny low-texture pests; temporal modeling for UAV video streams; expanding the dataset to cross-regional and degraded-condition samples; model quantization and pruning for embedded devices; and, after augmentation and optimization, making the self-constructed dataset publicly available to facilitate independent replication and follow-up studies.

References

[1] Yuan Y, Chen L, Wu H, et al. Advanced agricultural disease image recognition technologies: A review[J]. Information Processing in Agriculture, 2022, 9(1): 48-59.

[2] Song Y. Application of image recognition technology in agriculture[C]//International Conference on Electrical Engineering and Intelligent Control (EEIC 2024). IET, 2024, 2024: 339-342.

[3] Han F, Guan X, Xu M. Method of intelligent agricultural pest image recognition based on machine vision algorithm[J]. Discover Applied Sciences, 2024, 6(10): 536.

[4] Yang G, Wang J, Nie Z, et al. A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention[J]. Agronomy, 2023, 13(7): 1824.

[5] Xiao B, Nguyen M, Yan W Q. Fruit ripeness identification using YOLOv8 model[J]. Multimedia Tools and Applications, 2024, 83(9): 28039-28056.

[6] Chu Y, Wang J, Ma L, et al. LMSFA-YOLO: A lightweight target detection network in Remote sensing images based on Multiscale feature fusion[J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(4): 63.

[7] Wei M, Chen K, Yan F, et al. YOLO-ESFM: A multi-scale YOLO algorithm for sea surface object detection[J]. International Journal of Naval Architecture and Ocean Engineering, 2025, 17: 100651.

[8] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[10] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137‑1149.

[11] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2961‑2969).

[12] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (ECCV) (pp. 21‑37). Springer.

[13] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2980‑2988).

[14] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[15] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[16] Jocher, G., Chaurasia, A., & Qiu, J. (2022). YOLOv5: Implementation of YOLO family models in PyTorch. GitHub repository, Ultralytics.

[17] Wang, C., Zhang, L., & Zhao, H. (2022). Lightweight pest detection in complex agricultural scenes based on improved YOLOv5 with ECA and BiFPN modules. Computers and Electronics in Agriculture, 200, 107226.

[18] Zhao, Y., Li, X., & Yang, J. (2023). A real-time pest detection method for UAV images using improved YOLOv5 with attention mechanism. Expert Systems with Applications, 216, 119504.

[19] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 764‑773).

[20] Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9308‑9316).

[21] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). 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 (CVPR) (pp. 658‑666).

[22] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 12993-13000.

[23] Bao, Y., Wang, C. Y., & Liao, H. Y. M. (2023). WIoU: Weighted IoU for bounding box regression in object detection. arXiv preprint arXiv:2301.10051.

[24] Du X, Cheng H, Ma Z, et al. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels[J]. Computers and electronics in agriculture, 2023, 214: 108304.

[25] Ai H, Zhu X, Han Y, et al. Extraction of Levees from Paddy Fields Based on the SE-CBAM UNet Model and Remote Sensing Images[J]. Remote Sensing, 2025, 17(11): 1871.

[26] Gao C, Xu J, Liu R. Marine vessel target detection algorithm based on improved YOLOv5[J]. KSII Transactions on Internet and Information Systems (TIIS), 2024, 18(10): 2966-2983.

[27] Tong Z, Chen Y, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint arXiv:2301.10051, 2023.