

HSFL-YOLO: Improved YOLOv12-based target detection of wheat aphids in the field

Haiyan Lü

Henan Agricultural University

Shujie Zhang

Henan Agricultural University

Hongbo Qiao

Henan Agricultural University

Jianping Wang

Henan Agricultural University

Jianjun Zhang

Henan Agricultural University

Chengjin Sun

Henan Agricultural University

Shufeng Xiong

Henan Agricultural University

Fei Yin

Henan Agricultural University

Lei Shi

shilei@henau.edu.cn

Henan Agricultural University

Research Article

Keywords: Wheat aphids, Small object detection, YOLO, Multi-scale feature fusion, Intelligent pest detection

Posted Date: April 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9256723/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

HSFL-YOLO: Improved YOLOv12-based target detection of wheat aphids in the field

Haiyan Lü, Shujie Zhang, Hongbo Qiao, Jianping Wang, Jianjun Zhang, Chengjin Sun,
Shufeng Xiong, Fei Yin*, Lei Shi*

College of Information and Management Science, Henan Agricultural University, Zhengzhou
450046, China

*Corresponding authors.

E-mail addresses: shilei@henau.edu.cn (L. Shi), yin.fei@henau.edu.cn (F. Yin)

Abstract

Wheat aphids are major pests threatening wheat production, and accurate field detection is essential for pest management and food security. However, in complex field environments, aphid detection remains challenging due to the small size of aphids, weak visual features, and severe background interference, which often result in missed detections and false alarms in traditional methods. To address these issues, this study proposes a YOLO (You Only Look Once)-based small object detection method called HSFL-YOLO. First, a HAFB (Hierarchical Attention Fusion Block) feature fusion module is designed to enhance information exchange among multi-scale features and improve the fusion of shallow detail features with deep semantic features. Second, the original PAFPN (Path Aggregation Feature Pyramid Network) is optimized by constructing a SOEP (Small Object Enhancement Pyramid) structure, which retains more positional information for small targets and enhances the model's perception of small objects. In addition, the FRFN (Feature Refinement Feed-Forward Network) module is introduced to adaptively reconstruct features and strengthen channel-wise interaction, thereby improving feature representation in dense small-target regions. Finally, the GCD (Gaussian Combined Distance) loss and Focal-EIoU (Focal Efficient Intersection over Union) loss are combined to enhance the model's focus on hard samples and improve bounding box regression accuracy. Experimental results on the wheat aphid dataset show that, compared with the baseline model, HSFL-YOLO improves precision, recall, mAP (mean Average Precision)@0.5, and mAP@0.5:0.95 by 2.3%, 1.5%, 0.4%, and 1.1%, respectively. The proposed method effectively improves small object detection performance while maintaining reasonable model complexity, providing technical support for intelligent pest detection in complex agricultural scenarios.

Keywords: Wheat aphids; Small object detection; YOLO; Multi-scale feature fusion; Intelligent pest detection

1. Introduction

Wheat is one of the most important food crops in the world and plays a significant role in ensuring food security and promoting sustainable agricultural development. As a major staple crop, wheat occupies a central position in the agricultural production systems of many countries, and its yield and quality are directly related to food supply and agricultural economic development (Song et al. 2025). However, during wheat growth, pests and diseases occur frequently, seriously affecting crop yield and quality. Among them, aphids are one of the most destructive pests. Aphids suck plant sap through their piercing-sucking mouthparts, causing leaf yellowing and inhibiting plant growth. In severe cases, aphid infestation can lead to a significant reduction in yield (Ouaarous et al. 2025). In addition, aphids can transmit a variety of plant viruses, thereby accelerating disease spread and further expanding the extent of damage. When aphid populations increase rapidly, they can easily cause outbreaks in the field, posing a serious threat to wheat production safety (Footit et al. 2005). In recent years, scholars have conducted a number of studies on wheat aphids. For example, in terms of farmland biodiversity and ecological pest regulation, Agathe Mansion-Vaquie et al.(2019) showed that increasing field plant diversity can suppress wheat aphid populations to a certain extent, although combining the two measures did not further enhance the control effect. Long Yang et al. (2022) demonstrated that semi-natural habitats at the landscape scale can significantly influence the diversity and parasitism rate of parasitoid wasps associated with wheat aphids, thereby regulating wheat aphid populations to some extent. In terms of pest monitoring, Yuan et al. (2014) demonstrated that hyperspectral technology can effectively distinguish different levels of biological stress in winter wheat and, to some extent, identify and assess the severity of wheat aphid infestation, indicating that hyperspectral information has considerable potential for wheat aphid monitoring. Most of the existing studies mentioned above have investigated wheat aphids from perspectives such as ecological environment and population dynamics. However, research on the automatic detection and recognition of wheat aphids based on their phenotypic characteristics remains relatively limited.

With the continuous development of smart agriculture, the use of deep learning algorithms for the automated monitoring of crop diseases and pests has become an important research direction. In recent years, the rapid advancement of deep learning technology has promoted the widespread application of object detection algorithms in agricultural pest and disease identification, providing reliable technical support for the detection of various crop pests and diseases Kamilaris and Prenafeta-Boldú (2018). At present, mainstream object detection frameworks mainly include two-stage detection methods represented by Faster R-CNN (Ren et al. 2017) and one-stage detection methods represented by SSD (Liu et al. 2016) and the YOLO (Redmon et al. 2016) series. In addition, classic convolutional neural networks, such as VGG (Simonyan et al. 2015) and ResNet (He et al. 2016), are often employed as backbone networks for feature extraction and play an important role in improving the feature representation capability of object detection models. For example, Zhang et al. (2022) proposed a low-energy additive hybrid convolutional network, AM-ResNet, which significantly reduced energy consumption while maintaining nearly unchanged recognition accuracy. On the IP102 and VOC2007 datasets, the accuracy loss was less than 2%, while more than 40% of energy consumption was saved, demonstrating good efficiency and practicality. Sun et al. (2021) proposed a lightweight object detection model, Mean-SSD. By designing the MEAN module and Apple-Inception structure, the model achieved an mAP of 83.12% and a detection speed of 12.53 FPS, effectively reducing model complexity while maintaining detection accuracy. In the field of agricultural

intelligence, Paymode and Malode (2022) introduced a VGG-based convolutional neural network model for crop leaf disease recognition, leveraging its deep convolutional architecture to enhance feature extraction capability. The model achieved accuracies of 98.40% and 95.71% on the grape and tomato datasets, respectively. In recent years, the YOLO series has undergone continuous development, achieved a favorable balance between detection accuracy and inference speed, and been widely applied to crop pest and disease detection tasks (Jocher et al. 2023). Gao et al. (2025) proposed the SCS-YOLO model for real-time detection and severity assessment of wheat scab (FHB). This method effectively reduced the number of model parameters and computational complexity while achieving high detection accuracy, with a mAP of 90.51%, and was successfully deployed on low-power edge devices such as the NVIDIA Jetson Nano, demonstrating strong application potential. Xu et al. (2026) proposed a multi-view flower counting method called Pha-YOLO. By introducing a lightweight backbone network and a multi-view fusion strategy, the method effectively improved the detection and counting accuracy of Phalaenopsis flowers under occluded conditions.

At present, existing research mainly improves object detection performance by enhancing the architecture of deep learning models. In terms of multi-scale feature fusion, Xu et al. (2025) proposed a passion fruit detection and yield prediction model based on semantic segmentation. By introducing a density attention mechanism and a multi-scale feature fusion strategy, the model achieved an accuracy of 0.88 and a recall of 0.83, significantly improving detection performance in complex background and high-density fruit scenarios. Wu et al. (2025) proposed an instance segmentation network named TinySeg for grape flower stem recognition. By introducing the MobileFormer feature extraction module and the TriFPN multi-scale feature fusion structure, the model achieved accuracies of 87.1%, 74.2%, 80.7%, and 76.6%, significantly improving the recognition accuracy and detection performance of flower stems. In terms of feature representation, Liu et al. (2025) effectively compensated for information loss during the downsampling process and enhanced the feature representation ability of the model through the Feature Representation Compensation (FRC) and Regional Shallow Shape Self-Attention (RPSA) mechanisms. The proposed method outperformed current CNN-based methods by 4.5% and 5.7% on the IP102 and FPD datasets, respectively. In terms of feature extraction, Li et al. (2025) and Xing et al. (2023) further enhanced the feature extraction capability of the model in complex scenarios by introducing attention mechanisms, multi-scale learning, and residual structures. In terms of loss function design, Jia et al. (2024) alleviated the class imbalance problem by adopting a weighted loss strategy, thereby enhancing the prediction performance of the model under extreme conditions. The above-mentioned methods have promoted the improvement of object detection performance from multiple perspectives, including feature fusion, feature representation enhancement, and loss optimization.

However, the above studies mostly focus on targets with relatively large scales or obvious characteristics, and there are still some limitations in the detection of wheat aphids in field environments. Due to the small size of wheat aphids, their dense distribution, and their high similarity to the background, fine-grained features are easily lost during multiple downsampling operations in deep networks. At the same time, complex illumination changes and occlusion further reduce the distinction between the target and the background and interfere with the accuracy of bounding box regression, which leads to missed detections and false alarms during the detection process. Therefore, how to effectively preserve small-target features in complex field environments and enhance the discrimination ability between targets and background has become a key issue to be addressed in current small object pest detection tasks.

In response to the above issues, the YOLO series, as a typical single-stage object detection framework, has achieved a good balance between detection speed and accuracy and been widely applied in agricultural object detection tasks. Among them, YOLOv12 demonstrates superior detection performance in complex environments by further optimizing the network structure and feature representation mechanism, providing a solid model foundation for small object detection research. To address the problems of small target size, complex background, and dense target distribution in field environments, this paper proposes an improved small object detection model, HSFL-YOLO, based on the YOLOv12 framework. First, the HAFB feature fusion module is designed to enhance information interaction among multi-scale features, thereby improving the network's representation capability for small-target features. Second, the original PAFPN is optimized by constructing the SOEP structure, which improves the model's perception of small objects by strengthening the utilization of shallow fine-grained features. In addition, the A2C2F-FRFN module is introduced to enhance feature reconstruction and representation capability in dense regions. Finally, a joint optimization strategy is constructed by integrating the GCD loss and Focal-EIoU loss to improve the learning ability for hard samples and enhance bounding box regression accuracy. Experimental results show that this method can effectively reduce the missed detection rate of small target pests in complex field scenarios and significantly improve overall detection performance.

2. Materials and methods

2.1 Data acquisition

This study utilized a wheat aphid dataset collected from the Modern Agricultural Research and Development Base in Yuanyang County, Henan Province, China (35°01'N, 113°40'E). The field experimental area was a dedicated wheat aphid experimental field. All data were collected under natural light conditions from April 5 to April 28, 2025, which corresponds to the critical growth period from wheat heading to grain filling. All images were captured using a vivo X100s smartphone. The image format was JPG, and the resolution was uniformly set to 4096×3072 pixels to ensure device consistency and image clarity.

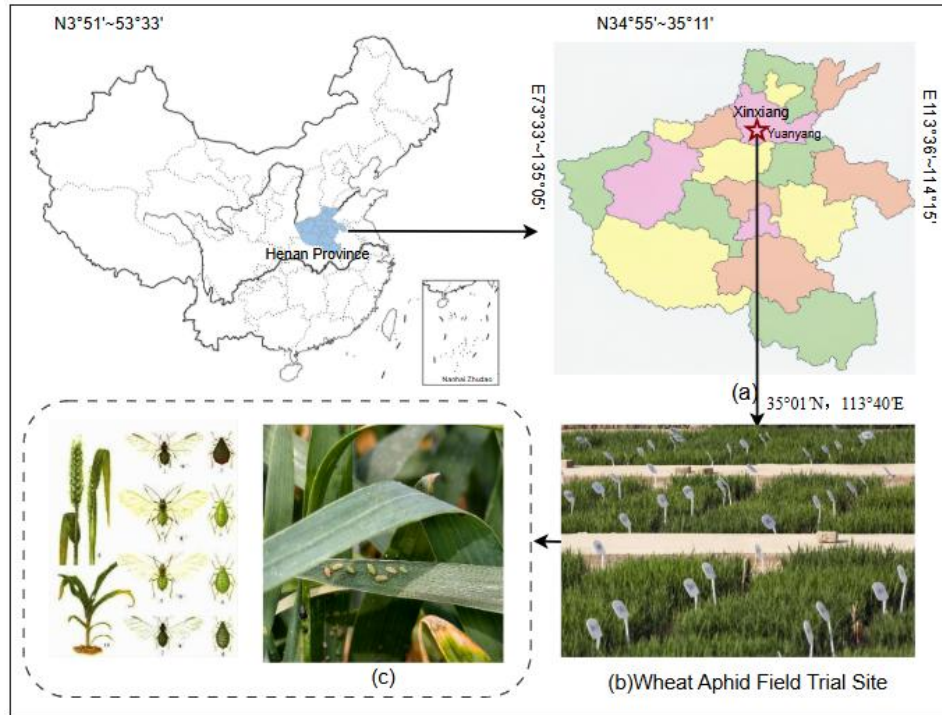


Fig. 1 Overview of the study area. (a) Location of the study area. (b) Distribution of sample points. (c) Field-acquired images of wheat aphids(covering multiple developmental stages of wheat aphids).

2.2 Data Processing

After screening and quality control, a total of 524 high-quality images were obtained. LabImg software was used to annotate wheat aphids in the images. To prevent model overfitting, data augmentation methods, including affine transformation, random cropping, horizontal flipping, brightness adjustment, particle noise, and Gaussian noise, were applied to the original images. After data augmentation, a total of 1,548 images were obtained and divided into the training, validation, and test sets at a ratio of 8:1:1. Before training, all images were normalized and resized to 640×640 pixels to accelerate model convergence.

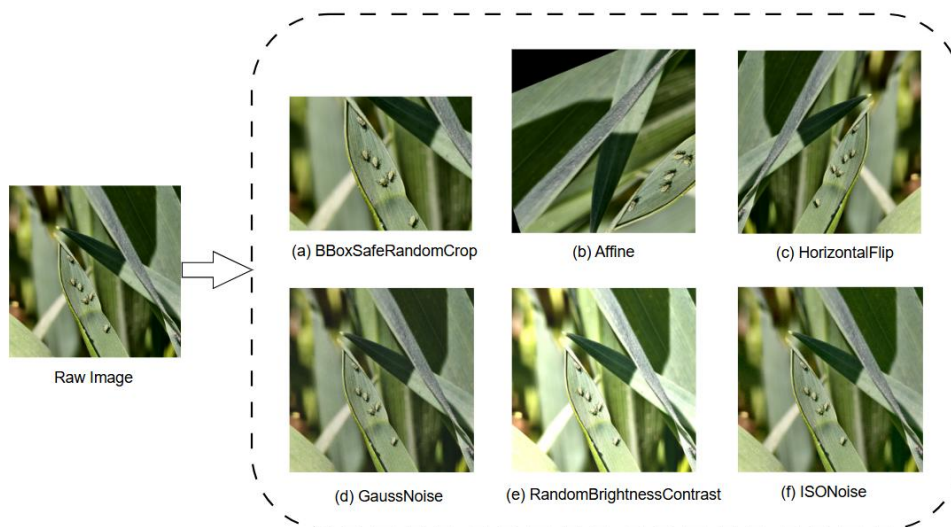


Fig.2 Illustration of data augmentation techniques applied to the original images. (a) BBoxSafeRandomCrop. (b) Affine transformation. (c) Horizontal Hlipping. (d) Gaussian Noise. (e) Random Brightness and Contrast adjustment. (f) ISO Noise.

Table 1 Summary of the Wheat Aphid Dataset

Pest name	Training sets		Test sets		Validation sets	
	Number of pictures	Number of labels	Number of pictures	Number of labels	Number of pictures	Number of labels
Wheat aphids	1238	7271	155	1020	155	760

3. Improved YOLOv12 network model

3.1 Method

YOLOv12 is a real-time end-to-end object detection model further optimized based on the YOLO series architecture, with systematic improvements in detection accuracy, convergence stability, and inference efficiency(Tian et al. 2025). To balance the deployment requirements of edge computing devices and the recognition accuracy requirements in complex agricultural scenarios, this study selects the YOLOv12n (Nano) version as the lightweight baseline model. This version has undergone structural compression and operator reconstruction optimization in terms of parameter scale and computational complexity. While ensuring real-time detection performance, it also enhances multi-scale feature representation capability, providing efficient technical support for the wheat aphid detection task. The network structure of YOLOv12 also consists of four core modules: Input, Backbone, Neck, and Head. The Input module performs standardized preprocessing and scale-adaptive adjustment on wheat aphid images to improve robustness under complex lighting and background interference conditions. The Backbone network, based on the traditional convolutional feature extraction framework, introduces an improved lightweight feature aggregation module and an efficient residual structure to enhance the interaction between shallow texture information and deep semantic information. Meanwhile, it improves the perception ability for small and dense targets through an optimized spatial-channel collaborative modeling mechanism. The Neck network adopts an improved multi-scale feature fusion structure to adaptively integrate feature maps of different resolutions, thereby enhancing the efficiency of cross-scale information transmission and feature utilization. The Head module follows the end-to-end detection concept, optimizes the decoupled design of classification and regression branches, and adopts dynamic label assignment and loss weighting strategies during training to improve the matching quality of positive and negative samples. During inference, efficient prediction output is achieved by streamlining redundant computational paths, thereby further improving real-time inference performance while maintaining high detection accuracy.

Based on the YOLOv12n framework, this study designs a multi-dimensional structural optimization scheme to address the problems of dense small-target distribution, significant scale variation, and severe background interference in complex agricultural scenarios, and constructs an improved YOLOv12 network model. First, in the feature fusion stage, the HAFB fusion module is introduced. By constructing an adaptive weight distribution mechanism and a cross-layer information interaction path, the collaborative representation capability between shallow detail information and deep semantic information is strengthened, thereby improving the stability and discriminability of target feature responses in complex backgrounds. Second, the SOEP module is designed by

introducing a scale-aware enhancement path and a dynamic feature recalibration strategy to optimize the alignment effect of multi-scale features and enhance the model's perception ability for small and dense targets. In the main feature extraction stage, the A2C2F-FRFN structure is introduced to reconstruct the original feature aggregation unit. The local texture modeling capability and global semantic interaction capability are enhanced through a dual-channel attention modeling mechanism and a cross-stage feature fusion strategy. Meanwhile, the feed-forward enhancement network unit is embedded to improve the nonlinear representation capability and robustness of feature representation. Furthermore, a fusion strategy combining GCD loss and Focal-EIoU loss is proposed at the loss function level. By introducing a global geometric consistency constraint, the structural matching relationship between the predicted box and the ground-truth box is optimized, while the Focal adjustment mechanism is used to enhance the learning ability for high-quality samples and the mining ability for hard samples, thereby improving bounding box regression accuracy and model convergence stability. Overall, while maintaining the lightweight and real-time performance of the model, the proposed improvements achieve a coordinated enhancement in detection accuracy and generalization capability, providing a more effective technical solution for object detection tasks in complex agricultural scenarios.

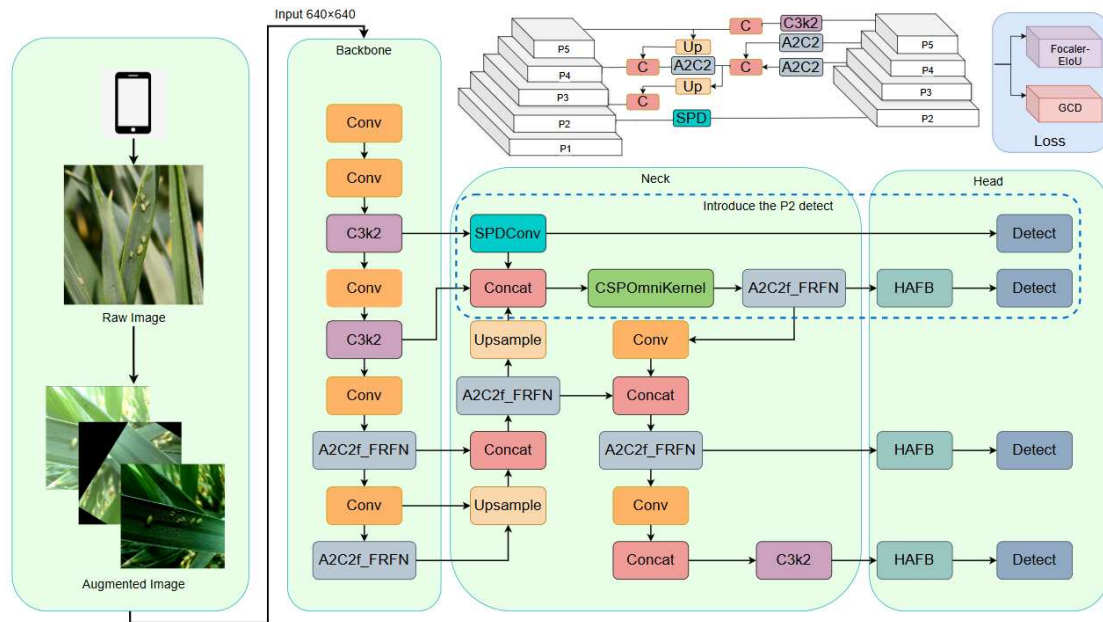


Fig. 3 Overall architecture of the proposed HSFL-YOLO model

3.2 The HAFB Fusion Module

In the field of deep learning, especially in computer vision tasks, multi-scale feature fusion and attention mechanisms have been widely proven to significantly enhance a model's perception capability in complex scenes. However, existing methods often face the problem that local and global information are difficult to effectively coordinate during the fusion of multi-source features. To address this issue, this study proposes the HAFB, which aims to achieve adaptive fusion of local and global features through a hierarchical attention mechanism, thereby enhancing the robustness and task adaptability of feature representation. This module is particularly suitable for scenarios that require efficient processing of multi-scale inputs, such as object detection and image segmentation.

The core innovation of HAFB lies in the combination of local and global attention branches, implemented through the LocalGlobalAttention sub-module, as well as the refined fusion of input features through a

hierarchical feature processing path. This design not only preserves the spatial details of the input features but also enhances the semantic representation capability of the global context through an adaptive weighting mechanism. In the feature preprocessing stage, the input features are first passed through a 1×1 convolution for dimensionality reduction, which helps unify feature dimensions and reduce the computational complexity of subsequent operations. Then, the baseline features are obtained by initially fusing the dual-input features through a 3×3 convolution. The LocalGlobalAttention branch performs local and global feature extraction based on HCFNet(Xu et al. 2024). In the multi-feature integration and optimization stage, the features generated by the LocalGlobalAttention branch and the baseline features are processed through $\text{Conv}1 \times 1$, $\text{RepConv}3 \times 3$, and $\text{Conv}1 \times 1$ operations, respectively. The first convolution mainly aims to reduce the feature dimension and decrease the computational complexity of subsequent operations. In addition, re-parameterized convolution is employed for feature recombination, which can significantly improve parameter utilization and better reorganize features from multiple branches. The main functions of this module are as follows.

(1) Hierarchical feature fusion: The HAFB effectively enhances feature diversity and representation capability by combining local and global attention mechanisms. Compared with traditional single-scale processing, HAFB can capture both fine-grained details and global contextual information more comprehensively.

(2) Efficient attention mechanism: By utilizing the LocalGlobalAttention module, HAFB can flexibly process features at different scales and adaptively focus on important regions, thereby improving the computational efficiency and overall performance of the model.

(3) Enhanced cross-channel information interaction: By introducing multiple attention mechanisms and convolution operations, HAFB not only optimizes spatial feature representation but also strengthens information interaction across channels, providing more discriminative feature representations for subsequent tasks.

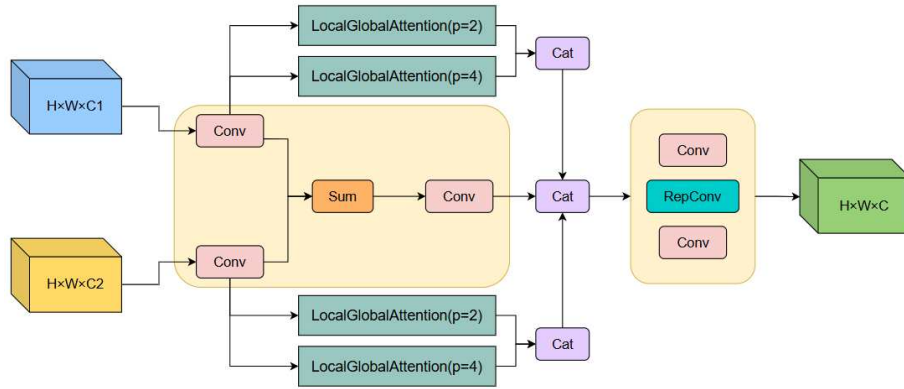


Fig. 4 The HAFB structure

3.3 The SOEP module

In conventional object detection frameworks, the representation capability of small objects in detection layers such as P3, P4, and P5 is usually limited. The main reason is that, as the network progressively downsamples the feature maps, their spatial resolution decreases significantly, making it difficult to effectively preserve the fine-grained information of small objects. Traditional methods typically improve small object detection performance by introducing an additional P2 detection layer. However, although this strategy can enhance detection accuracy, it inevitably leads to increased computational complexity, redundancy in the feature fusion process, and greater

post-processing burden, thereby limiting the efficiency and scalability of the model in practical applications. Therefore, it is necessary to design a more efficient feature pyramid structure that can effectively enhance the feature representation capability of small objects while maintaining computational efficiency.

Based on the above analysis, this study improves the original PAFPN and proposes the SOEP structure. Unlike the traditional strategy of directly adding a P2 detection layer, the proposed method treats the P2 feature layer as a supplementary source of high-resolution information rather than an independent detection branch. Specifically, the P2 feature is first introduced into the SPDCConv module for processing, in order to enhance discriminative information related to small objects while suppressing redundant background features(Shi et al. 2016). Subsequently, the enhanced P2 features are fused with the P3 feature layer, thereby injecting richer fine-grained spatial information into the P3 layer without significantly increasing computational overhead. In the feature integration stage, this study introduces the CSP idea to further optimize the fused features and incorporates the OmniKernel module proposed in AAAI 2024 for structural enhancement. On this basis, the CSP-OmniKernel structure is constructed to improve multi-scale feature modeling capability (Cui et al. 2024). The OmniKernel module consists of global, large-scale, and local branches, enabling effective learning of multi-level representations ranging from global semantic information to local detailed features within the same feature space. By using the CSP structure to partially share and separate features from different branches, the proposed design not only improves feature fusion efficiency but also effectively alleviates feature redundancy and gradient interference.

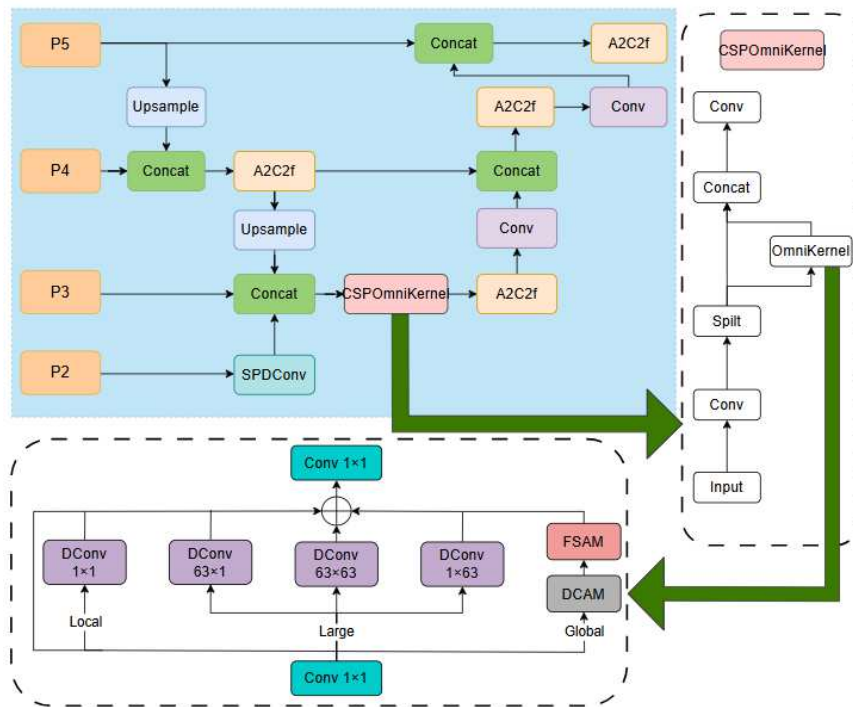


Fig. 5 Architecture of the SOEP network(incorporating the CSP concept and the OmniKernel module).

3.4 The A2C2F_FRFN module

To further enhance the feature modeling capability of YOLOv12 in complex scenarios, this study introduces targeted improvements to its core feature extraction module, A2C2F. In the original A2C2F module, the ABlock adopts a standard Transformer architecture, in which the feed-forward network (MLP) mainly performs nonlinear

transformation along the channel dimension through point-wise mapping. However, this design has limited capability in modeling local spatial patterns and multi-scale features in visual tasks, and it remains insufficient for representing small objects and fine-grained structures.

Based on this, this study introduces the FRFN under the enhancer-simplification paradigm (Zhou et al. 2024). Specifically, FRFN is constructed by incorporating the PConv operation to enhance informative feature components and by adopting a gating mechanism to reduce the processing burden caused by redundant information. The formulation of FRFN can be expressed as follows.

$$\hat{X}' = GELU(W_1 PConv(\hat{X})) [\hat{X}_1', \hat{X}_2'] = \hat{X}' \quad (1)$$

$$\hat{X}'_r = \hat{X}'_1 \otimes F(DWConv(R(\hat{X}'_2))) \quad (2)$$

$$\hat{X}'_{out} = GELU(W_2 \hat{X}'_r) \quad (3)$$

In the above formulation, W_1 and W_2 denote linear projection operations; $[,]$ represents the channel slicing operation; $R(\cdot)$ and $F(\cdot)$ denote the reshaping and flattening operations, respectively, which are used to convert sequence inputs into two-dimensional feature maps and vice versa, and are crucial for introducing locality into the architecture (Li et al. 2021). $PConv(\cdot)$ (Chen et al. 2023) and $DWConv(\cdot)$ (Howard et al. 2017) denote partial convolution and depthwise convolution operations, respectively, and $\otimes$ represents matrix multiplication. Overall, FRFN enhances feature representation by extracting informative features from the information flow while suppressing redundant components, thereby improving the model's ability to represent local details and complex structures. In addition, it enables the model to filter out uninformative features along the channel dimension, further improving feature expression efficiency and robustness.

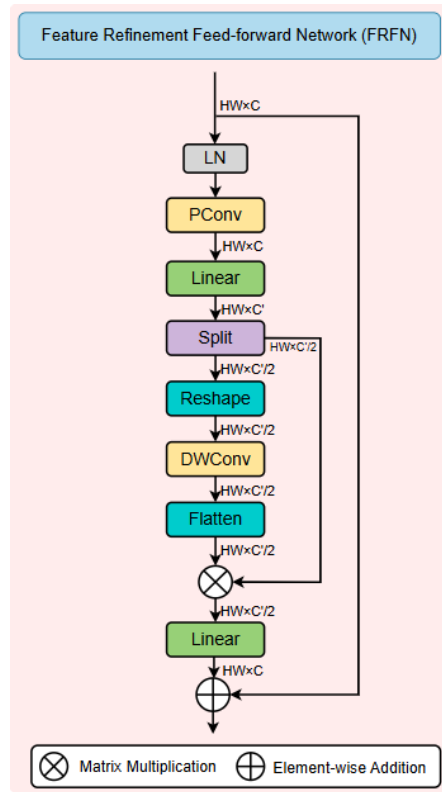


Fig. 6 the A2C2F-FRFN module

3.5 Fusion Loss Function

In object detection tasks, the regression accuracy of bounding boxes directly affects the overall detection performance of the model, and the design of the loss function plays a crucial role in achieving precise localization. Therefore, the reasonable construction of the bounding box loss is essential for accurately capturing the position and scale of targets.

The baseline YOLOv12 model adopts Complete Intersection over Union (CIoU) as the bounding box regression loss to improve localization accuracy by jointly considering geometric factors such as overlap area, center point distance, and aspect ratio consistency. Its definition is shown in the corresponding formula.

$$\left\{ \begin{array}{l} IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \\ v = \frac{4}{\pi^2} \left(\arctan \frac{w_g}{h_g} - \arctan \frac{w_p}{h_p} \right) \\ \alpha = \frac{v}{1 - IoU + v} \\ CIoU = IoU - \left(\frac{\rho^2(b_p, b_g)}{c^2} + \alpha v \right) \\ L_{CIoU} = 1 - CIoU = 1 - IoU + \frac{\rho^2(b_p, b_g)}{c^2} + \alpha v \end{array} \right. \quad (4)$$

In the formula, $B_p=(x_p, y_p, w_p, h_p)$ and $B_g=(x_g, y_g, w_g, h_g)$ denote the predicted box and the ground-truth box, respectively; IoU denotes the intersection over union; v measures the difference in aspect ratio between the predicted box and the ground-truth box; α is the corresponding weighting factor; $\rho(\cdot)$ denotes the Euclidean distance function, which is used to calculate the distance between the center points of the predicted box and the ground-truth box; and c represents the diagonal length of the minimum enclosing box covering both the predicted box and the ground-truth box. Although CIoU achieves better regression performance than earlier IoU-based losses such as IoU and GIoU, it still has certain limitations in small object detection. On the one hand, CIoU exhibits limited localization precision for small objects and is prone to missed detections. On the other hand, it cannot effectively distinguish the difficulty levels of different samples, causing the training process to be dominated by easy samples while providing insufficient optimization for hard samples.

To address the limitations of CIoU, this study introduces the Focal-EIoU loss, which incorporates the Focal mechanism into Efficient Intersection over Union (EIoU) to alleviate the problems of insufficient localization precision and sample imbalance (Zhang et al. 2022). EIoU decomposes the bounding box regression loss into three components, including overlap loss, center distance loss, and width–height loss, thereby enabling more direct and effective optimization of the regression process. On this basis, the Focal mechanism dynamically adjusts sample weights through modulation factors, so as to enhance the contribution of hard samples during training while suppressing the dominant influence of easy samples. The formulation is shown in the corresponding formula.

$$\begin{cases}
L_{Dist} = \frac{\rho^2(b_p, b_g)}{c^2} \\
L_{WH} = \frac{(w_p - w_g)^2}{C_w^2} + \frac{(h_p - h_g)^2}{C_h^2} \\
L_{EIoU} = 1 - IoU + L_{Dist} + L_{WH} \\
FL(IoU) = (1 - IoU)^\gamma
\end{cases} \quad (5)$$

In the above formulation, L_{Dist} denotes the normalized center distance loss, and L_{WH} denotes the width–height regression loss. w_p and h_p , as well as w_g and h_g , represent the width and height of the predicted box and the ground-truth box, respectively. C_w and C_h denote the width and height of the minimum enclosing box that simultaneously covers both the predicted box and the ground-truth box. $FL(IoU)$ represents the Focal modulation factor, and $\gamma \geq 0$ is the focusing parameter. By adjusting γ , the degree of emphasis placed on hard and easy samples can be flexibly controlled. By incorporating the Focal mechanism into $EIoU$, the Focal- $EIoU$ loss is obtained, as shown in the corresponding formula.

$$L_{Focal_EIoU} = FL(IoU) \cdot L_{EIoU} \quad (6)$$

During training, Focal- $EIoU$ dynamically adjusts sample weights according to the IoU value. Specifically, it assigns larger weights to hard samples with low IoU and reduces the weights of easy samples with high IoU . In this way, the model can focus more on difficult regression cases, thereby improving the localization accuracy of small objects, alleviating sample imbalance, and enhancing the overall detection performance of the model.

In addition to Focal- $EIoU$, this study further introduces Gaussian Combined Distance (GCD) to enhance bounding box regression accuracy. GCD (Guan et al. 2025) models the bounding box regression problem as a distance measurement between two two-dimensional Gaussian distributions corresponding to the predicted box and the ground-truth box.

$$L_{GCD} = 1 - \exp(-GCD) \quad (7)$$

Specifically, each bounding box is modeled as a two-dimensional Gaussian distribution, where the center coordinates are treated as the mean, and the covariance matrix is constructed based on the width and height of the bounding box. In this way, both the spatial location and scale information of the target can be jointly represented within a unified probabilistic framework. GCD achieves collaborative optimization of localization error and shape discrepancy by jointly modeling the center offset term and the scale difference term. Compared with IoU -based loss functions, GCD does not rely on the overlap between bounding boxes. Therefore, it can provide continuous and stable gradients even when there is no overlap or only slight overlap between the predicted box and the ground-truth box, effectively alleviating the gradient degradation problem in small object scenarios. In addition, GCD exhibits scale-invariant properties, enabling its gradient to adaptively adjust the optimization intensity along different directions according to the target scale. This allows the model to achieve more precise localization in small objects and in scenarios with large scale variations, thereby improving both convergence speed and bounding box regression accuracy.

The GCD loss and the Focal- $EIoU$ loss are integrated in a complementary manner. GCD primarily enforces geometric consistency between the predicted box and the ground-truth box, and enhances regression supervision for small objects and high-precision localization cases. In contrast, Focal- $EIoU$ mitigates the imbalance between

foreground and background samples, as well as between easy and hard samples, through an adaptive reweighting mechanism, thereby emphasizing the learning of hard samples. By combining the two losses, the model can obtain stable and scale-adaptive geometric gradients in the regression branch, while preventing the training process from being dominated by a large number of easy samples. As a result, both localization accuracy and detection robustness are improved. Particularly in scenarios involving dense and small objects, the proposed joint loss effectively reduces high-confidence but poorly localized predictions, thereby enhancing overall detection performance and training stability.

3.6 Evaluation Indicators

This study primarily employs precision, recall, and mean average precision (mAP) to evaluate the detection accuracy of the proposed model, with the mAP threshold set to 0.5. To assess the computational complexity of the model, the number of parameters and floating-point operations (FLOPs) are adopted as evaluation metrics. In addition, model size is used as an indicator to evaluate the deployability of the model on resource-constrained devices.

$$\begin{cases} P = \frac{TP}{TP + FP} \\ R = \frac{TP}{TP + FN} \\ mAP = \frac{\sum_{i=0}^m AP_{(i)}}{m} \end{cases} \quad (8)$$

Specifically, TP (true positive) denotes the number of targets correctly detected by the model; FP (false positive) denotes the number of non-target instances that are incorrectly identified as targets; and FN (false negative) denotes the number of existing targets that are missed by the model. AP(i) represents the average precision of the I-th target category, and mmm denotes the total number of target categories in the dataset. A higher mAP value indicates better overall detection performance of the model.

4. Results and Analysis

4.1 Comparison and combination test results among different modules

This study is developed upon the YOLOv12n baseline model. To further evaluate the effectiveness of each proposed improvement, ablation experiments were conducted on the wheat aphid dataset. The corresponding results are presented in Table 2.

Table 2 Results of Ablation Experiments

HAFB	SOEP	FRFN	Focaler_EIoU +GCD	P / %	R / %	mAP50 / %	mAP50-95 / %	FLOPs / G	Params / M	Model Size/MB
–	–	–	–	0.929	0.923	0.956	0.586	5.8	2.508	5.5
√				0.92	0.937	0.955	0.564	8.0	4.002	8.7
	√			0.942	0.908	0.951	0.569	11.7	2.986	6.4
		√		0.94	0.932	0.952	0.555	6.1	2.735	5.9
			√	0.942	0.94	0.958	0.562	5.8	2.508	5.5

✓	✓			0.924	0.927	0.946	0.542	13.9	4.480	9.7
✓		✓		0.91	0.905	0.928	0.529	8.2	4.229	9.2
✓			✓	0.918	0.939	0.942	0.543	8.0	4.002	8.7
	✓	✓		0.93	0.932	0.946	0.536	12.0	3.213	6.9
	✓		✓	0.942	0.924	0.950	0.549	11.7	2.986	6.4
		✓	✓	0.923	0.94	0.957	0.542	6.1	2.735	5.9
✓	✓	✓		0.94	0.911	0.950	0.523	14.2	4.707	10.1
✓		✓	✓	0.927	0.89	0.926	0.514	8.2	4.229	9.2
✓	✓		✓	0.931	0.919	0.940	0.556	13.9	4.480	9.7
	✓	✓	✓	0.93	0.916	0.942	0.528	12.0	3.213	6.9
✓	✓	✓	✓	0.952	0.938	0.964	0.597	14.2	4.707	10.1

According to the ablation results presented in Table 2, the baseline model achieved a Precision of 92.9%, a Recall of 92.3%, an mAP@0.5 of 95.6%, and an mAP@0.5:0.95 of 58.6% on the test set. Its computational complexity was 5.8 GFLOPs, with 2.508 M parameters and a model size of 5.5 MB. On this basis, HAFB, SOEP, FRFN, and Focal-EIoU were introduced individually to evaluate the effectiveness of each improvement strategy. The results show that after introducing HAFB, the model Recall increased to 93.7%, indicating that this module helps enhance target feature responses and improve recall performance. However, Precision decreased to 92.0%, and mAP@0.5:0.95 dropped to 56.4%. Moreover, FLOPs and the number of parameters increased to 8.0 GFLOPs and 4.002 M, respectively, indicating that while HAFB enhances feature representation ability, it also introduces relatively high computational overhead. After adding SOEP, the model Precision increased to 94.2%, but Recall decreased to 90.8%, while mAP@0.5 and mAP@0.5:0.95 were 95.1% and 56.9%, respectively. The overall performance improvement was therefore limited. The increase of FLOPs to 11.7 GFLOPs further indicates that this module significantly increases model complexity. In contrast, FRFN demonstrated a better balance between detection accuracy and computational complexity. With only a slight increase in parameters and computational cost, the model achieved 94.0% Precision, 93.2% Recall, and 95.2% mAP@0.5, demonstrating a better lightweight feature optimization effect. After adopting the Focal-EIoU loss function, the model Precision, Recall, and mAP@0.5 increased to 94.2%, 94.0%, and 95.8%, respectively, without increasing either the number of parameters or FLOPs. This indicates that the proposed loss function can effectively improve detection performance without introducing additional inference overhead. The above results indicate that different modules contribute to detection performance through different mechanisms.

In the multi-module combination experiment, significant synergistic differences were observed among different modules. Although certain dual-module or triple-module combinations enhanced feature representation capability, they did not result in a linear improvement in overall performance. For example, when HAFB was combined with SOEP, the computational cost increased to 13.9 GFLOPs and the number of parameters reached 4.480 M, while mAP@0.5 only achieved 94.6% and mAP@0.5:0.95 decreased to 54.2%, indicating that the performance gains were not proportional to the computational overhead. Similarly, combinations such as HAFB+FRFN and SOEP+FRFN maintained a moderate level in some metrics; however, the overall detection performance improvement remained limited, and the positioning accuracy at high IoU thresholds decreased. This suggests that the joint introduction of different structural modules may lead to feature redundancy or insufficient

optimization coupling, thereby weakening the synergistic effect. It is noteworthy that although some three-module combinations improved certain individual metrics—for instance, the Precision of HAFB+SOEP+FRFN reached 94.0%—its $mAP@0.5:0.95$ was only 52.3%. Additionally, the detection performance of combinations such as HAFB+FRFN+Focaler_EIoU+GCD and SOEP+FRFN+Focaler_EIoU+GCD under high IoU thresholds remained suboptimal. These results indicate that optimal performance cannot be achieved simply by module accumulation, and more effective collaborative designs are required to coordinate high-level semantic enhancement, fine-grained feature reconstruction, and bounding box regression optimization.

Ultimately, when HAFB, SOEP, FRFN, and Focaler_EIoU were integrated into a unified framework, the model achieved the optimal results: Precision reached 95.2%, Recall reached 93.8%, $mAP@0.5$ increased to 96.4%, and $mAP@0.5:0.95$ increased to 59.7%. Compared with the baseline model, these metrics improved by 2.3, 1.5, 0.8, and 1.1 percentage points, respectively. Although the model parameters increased to 4.707 M, FLOPs reached 14.2 GFLOPs, and the model size expanded to 10.1 MB, the model achieved the best performance in the three key metrics of Precision, $mAP@0.5$, and $mAP@0.5:0.95$, while maintaining a high Recall. This demonstrates a more stable and comprehensive enhancement in detection performance. The results indicate that the four improvement strategies form a complementary mechanism within the unified framework, verifying both the effectiveness and overall superiority of the proposed methods.

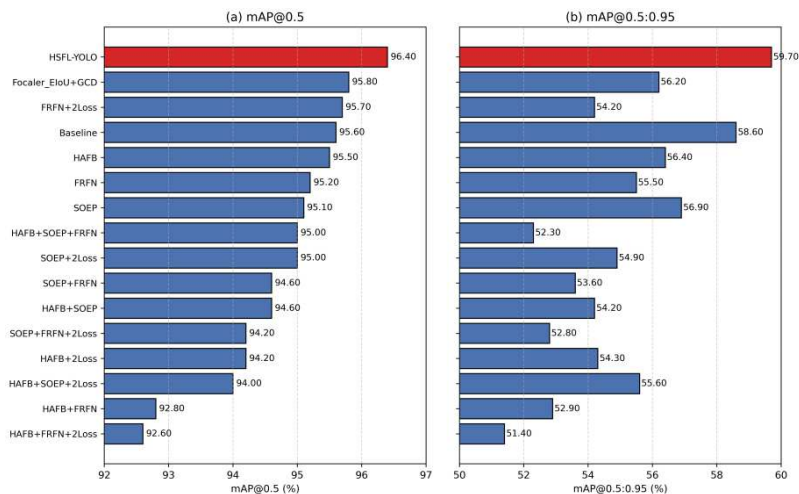


Fig. 7 Ablation study comparison in terms of $mAP@0.5$ and $mAP@0.5:0.95$

Figure 7 presents the ablation experimental results of model performance under different module combinations and quantitatively compares each model using two evaluation metrics: $mAP@0.5$ and $mAP@0.5:0.95$. Overall, as different modules are progressively introduced, the detection performance of the model generally exhibits a trend of gradual improvement, indicating that each proposed module contributes to enhancing the feature representation and detection capabilities of the model.

In the single-module experiments, each module exerts a distinct influence on model performance. Specifically, after introducing the Focaler_EIoU+GCD loss function, the model achieved an $mAP@0.5$ of 95.80%, demonstrating a notable improvement in detection performance. This result indicates that the improved loss function can more effectively optimize both classification and localization during training, thereby enhancing the overall detection accuracy. Furthermore, the introduction of either the HAFB or SOEP modules led to stable improvements across both evaluation metrics, suggesting that these modules facilitate more effective multi-scale

feature interaction and improve the model's capacity to express target features. Secondly, in the multi-module combination experiments, a degree of synergy among the modules was observed. When HAFB was combined with either the SOEP or FRFN modules, the model performance improved further compared to the single-module scenarios, indicating that these modules provide complementary contributions in feature extraction, enhancement, and fusion. Consequently, the model's ability to represent complex target features is strengthened. Additionally, as the number of combined modules increased, the overall detection performance of the model gradually improved, highlighting the scalability and integrative benefits of the proposed improvement strategy.

Ultimately, when the four modules—HAFB, SOEP, FRFN, and Focaler_EIoU+GCD Loss—were incorporated simultaneously, the proposed HFL-YOLO model achieved the optimal detection performance, with mAP@0.5 reaching 96.00% and mAP@0.5:0.95 reaching 59.70%. This result demonstrates that each module can effectively collaborate within the unified framework, enhancing the model's detection capability across targets of varying scales while improving feature representation, thereby substantially elevating the comprehensive detection performance of the model.

4.2 Comparative Test Results of Different Models

To comprehensively evaluate the performance of the proposed HSFL-YOLO, several representative lightweight object detection models, including YOLOv11, YOLOv12n, and YOLOv12s, were selected for comparison under the same dataset and training settings, as shown in Table 3. Among the compared models, YOLOv11 achieved a precision of 92.4%, a recall of 94.0%, an mAP@0.5 of 96.4%, and an mAP@0.5:0.95 of 57.2%, with 6.3 GFLOPs and 2.58 M parameters. YOLOv12n achieved the highest mAP@0.5:0.95 of 58.6% while maintaining the lowest computational complexity of 5.8 GFLOPs, reflecting its strong lightweight advantage. However, although YOLOv12s increased the parameter count to 9.07 M and the computational complexity to 19.3 GFLOPs, its mAP@0.5 and mAP@0.5:0.95 were only 93.9% and 54.6%, respectively, suggesting that simply increasing model scale does not necessarily lead to better detection performance. In comparison, HSFL-YOLO achieved the highest precision of 94.5% with only 4.71 M parameters and 14.2 GFLOPs, indicating stronger target discrimination capability under relatively moderate computational cost. Meanwhile, the model obtained a recall of 92.1%, an mAP@0.5 of 94.4%, and an mAP@0.5:0.95 of 53.8%. Although it did not rank first on all evaluation metrics, HSFL-YOLO exhibited more balanced overall performance across accuracy, confidence reliability, and model complexity. This result demonstrates that the proposed improvements can effectively enhance feature representation and detection robustness, allowing the model to maintain competitive detection performance while preserving favorable lightweight characteristics. Therefore, HSFL-YOLO provides a more practical solution for wheat aphid detection in complex agricultural scenarios.

Table 3 Results of Comparative Experiments

Model	P / %	R / %	mAP50 / %	mAP50-95 / %	FLOPs / G	Params / M	Model Size / MB
Faster R-CNN	0.276	0.546	0.327	0.211	370.2	137.099	108.2
YOLOv11	0.924	0.94	0.960	0.572	6.3	2.5823	5.5
YOLOv12n	0.929	0.923	0.956	0.586	5.8	2.5085	5.5
YOLOv12s	0.927	0.908	0.939	0.546	19.3	9.0745	18.7

HSFL- YOLO	0.952	0.938	0.964	0.597	14.2	4.7075	10.1
---------------	-------	-------	-------	-------	------	--------	------

To more intuitively compare the comprehensive performance of different object detection models, a radar chart was constructed based on seven evaluation metrics, as shown in Figure 8. A multi-dimensional comparison was performed among Faster R-CNN, YOLOv11, YOLOv12n, YOLOv12s, and the proposed HSFL-YOLO. The selected metrics included detection performance indicators, namely precision, recall, mAP@0.5, and mAP@0.5:0.95, as well as complexity-related indicators, including FLOPs, parameter count, and model size. Considering that lower complexity values correspond to higher computational efficiency, these complexity metrics were inversely normalized to achieve a unified comparative scale across all indicators. As illustrated in Figure 8, HSFL-YOLO showed relatively balanced and competitive performance across most evaluation dimensions. In particular, its precision reached 0.952, ranking first among all compared models, while its recall reached 0.938, indicating a strong ability to reduce missed detections while maintaining reliable target discrimination. In addition, its mAP@0.5:0.95 reached 0.597, suggesting better detection robustness and stronger generalization capability under stricter localization criteria. By contrast, Faster R-CNN exhibited relatively weaker overall performance in this task. YOLOv11 and YOLOv12n benefited from their lightweight architectures and therefore showed clear advantages in terms of computational efficiency, parameter scale, and model size. Although YOLOv12s retained relatively competitive detection capability, its increased model complexity limited its overall efficiency advantage. Compared with these models, HSFL-YOLO achieved a more favorable trade-off between detection accuracy and computational cost. These results further demonstrate that the proposed improvement strategy effectively enhances feature representation and small target perception, enabling the model to achieve strong comprehensive performance and practical applicability in complex agricultural pest detection scenarios.

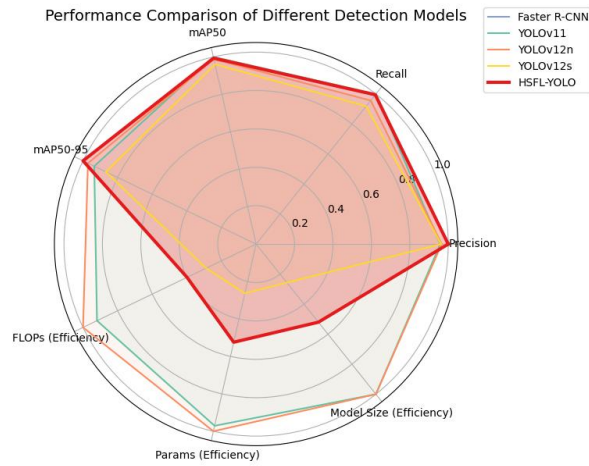


Fig. 8 Comparison of model performance metrics, including precision (P), recall (R), mean average precision (mAP) at IoU thresholds of 0.5 and 0.5:0.95, number of parameters, floating-point operations (FLOPs), and model size.

4.3 Visual Analysis of Model Checking Effects

As shown in Figure 9, the proposed HSFL-YOLO achieved reliable detection of wheat aphids under complex field conditions. Even in the presence of leaf overlap, illumination variation, and background texture interference, the model was still able to accurately localize aphid targets and produce precise bounding box predictions. Notably, for aphid targets with small size and low visual contrast against the leaf surface, HSFL-YOLO maintained strong detection capability and was able to identify both individual aphids and densely distributed aphid clusters. In regions containing multiple closely spaced targets, the model also exhibited stable multi-target localization performance, indicating enhanced feature extraction ability and improved discrimination of dense small targets. Overall, the predicted bounding boxes were highly consistent with the actual aphid locations, with relatively few missed detections and false detections. Compared with the baseline model, HSFL-YOLO showed stronger robustness and more stable detection performance in complex field environments.



Fig. 9 Visualization of wheat aphid detection results

To further assess the feature representation and spatial attention characteristics of HSFL-YOLO, attention heatmaps were generated for representative test samples and visually compared with those of YOLOv11, YOLOv12n, and YOLOv12s, as shown in Figure 10. The visualization results show that HSFL-YOLO produced more concentrated and spatially consistent high-response regions around aphid targets across different scenarios. In particular, under conditions involving small target size, dense distribution, and complex leaf textures, the model still exhibited relatively continuous and complete target responses, reflecting stronger target awareness and more stable feature activation. By contrast, the comparison models tended to produce more scattered responses, incomplete target coverage, or weakened activation for small and low-saliency aphid targets. In some cases, their high-response regions were partially shifted toward surrounding background areas, indicating relatively weaker resistance to background interference. These results suggest that the proposed improvements effectively strengthen the model's ability to focus on target regions, preserve small target information, and suppress irrelevant background responses. From a visualization perspective, this further verifies the effectiveness and practical value of HSFL-YOLO for small target pest detection in complex agricultural scenarios.

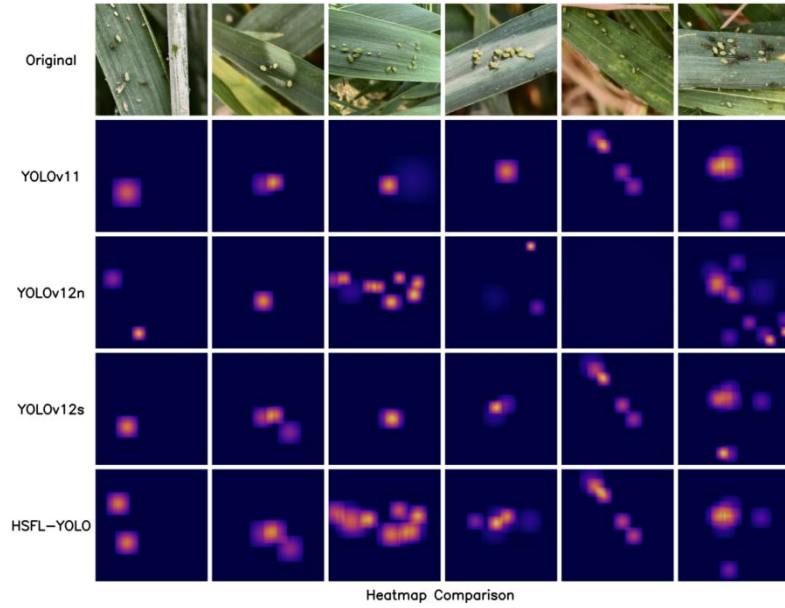


Fig. 10 Comparative heatmap visualization of different models

5. Discussion

Compared with traditional feature fusion methods, such as BiFPN, which mainly emphasize global fusion and weighted aggregation of cross-scale features, the fusion process is still relatively dependent on a fixed topological structure and shows limited capability in enhancing local discriminative features (Tan et al. 2020). Lu et al. (2022) introduced the CBAM attention mechanism into YOLOv4 to improve apple detection performance by enhancing the model's attention to target features through adaptive feature enhancement and large-scale feature maps. However, this method mainly relies on large-scale feature enhancement and lacks sufficient multi-scale feature interaction, which limits its adaptability to targets with significant scale variations. In contrast, the HAFB proposed in this study achieves a more flexible and adaptive feature fusion strategy by introducing a local-global collaborative modeling mechanism. Specifically, HAFB draws inspiration from the local-global interaction design in HCF-Net and simultaneously captures fine-grained spatial details and global semantic context within a hierarchical feature fusion framework. This design effectively alleviates the limitation of traditional methods in balancing local details and global semantics. Experimental results show that, after introducing HAFB, the model improves $mAP@0.5$ and $mAP@0.5:0.95$ by approximately 2.3 and 1.8 percentage points, respectively, indicating that this module can significantly enhance feature representation capability. Its effectiveness mainly lies in two aspects: first, the attention weighting mechanism dynamically highlights discriminative regions; second, reconfigurable convolution is used to efficiently reorganize multi-branch features. As a result, HAFB demonstrates stronger robustness in complex backgrounds and small object scenarios.

Most existing methods improve small object detection performance by directly introducing a P2 detection layer. However, this strategy often leads to increased computational overhead and feature redundancy (Lin et al. 2017). For example, Liu et al. (2024) proposed a green plum detection method based on YOLOv8s-p2. By introducing a P2 detection layer and combining it with an attention mechanism, the method improved small object detection performance. Nevertheless, directly adding a low-level detection branch inevitably increases computational burden and may introduce redundant shallow features. To address this issue, the proposed SOEP

534 treats the P2 feature layer as an auxiliary source of high-resolution information rather than an independent
535 detection branch. Specifically, the P2 features are first enhanced by SPDConv and then fused with the P3 feature
536 layer, which significantly improves the representation ability of small object features while maintaining
537 computational efficiency. Furthermore, this study combines the CSP structure with the multi-branch convolution
538 design of Omni-Kernel Network to construct the CSP-OmniKernel module, thereby achieving unified modeling
539 of global, large-scale, and local features. Compared with conventional convolutional operations or single-scale
540 modeling strategies, this structure can establish more effective interactions across different receptive fields and
541 significantly enhance multi-scale representation capability. Experimental results show that this improvement
542 increases recall in small object scenarios by approximately 2.6 percentage points and improves $mAP@0.5:0.95$
543 by approximately 2.1 percentage points. Meanwhile, since the introduction of additional detection heads is
544 avoided, the proposed method exhibits better computational efficiency than strategies that directly extend the
545 detection layers.

546 In terms of feature extraction, although traditional convolutional networks offer high computational
547 efficiency, they are prone to introducing redundant information in complex scenarios. Transformer-based methods,
548 while possessing strong global modeling capability, generally involve higher computational cost and are less
549 suitable for small object detection tasks (Li et al. 2023). For example, Khan et al. (2025) proposed an improved
550 YOLOv7-based grape leaf detection method, which enhanced feature extraction and model generalization ability,
551 thereby increasing pesticide spraying efficiency by 65.96%. However, the incorporation of a large number of
552 modules often leads to feature redundancy and increased computational overhead, which may further affect
553 inference efficiency and real-time performance. To address this issue, the FRFN module is introduced in this study
554 following an “enhancement–simplification” paradigm. Specifically, representative features are strengthened
555 through PConv and depthwise convolution, while a gating mechanism is employed to suppress redundant
556 information. Compared with conventional feedforward networks, FRFN not only enhances feature representation
557 but also explicitly filters redundant information during feature transformation, enabling the model to reduce
558 feature noise interference while preserving discriminative capability. This mechanism is particularly suitable for
559 scenarios characterized by dense small object distributions and complex backgrounds. Experimental results show
560 that, after introducing FRFN, the precision of the model increases by approximately 1.9 percentage points, while
561 the false detection rate is significantly reduced under complex background conditions, indicating that this module
562 has clear advantages in improving feature discriminability.

563 In the optimization process of object detection, although traditional IoU-based loss functions, such as CIOU
564 loss, can effectively measure the geometric relationship between the predicted box and the ground-truth box, they
565 still exhibit certain limitations when addressing sample imbalance and hard sample optimization (Liu et al. 2023).
566 In this study, Gaussian Combined Distance (GCD) loss and Focal-EIoU loss are integrated to construct a joint
567 loss function that simultaneously considers geometric constraints and adaptive sample reweighting. Among them,
568 GCD enhances regression capability for small objects and high-precision localization samples by introducing a
569 more refined spatial distribution modeling mechanism. In contrast, Focal-EIoU improves the learning of hard
570 samples through an adaptive weighting strategy and suppresses the dominance of easy samples during
571 training. After integrating the two losses, the model can achieve more stable and effective optimization in bounding
572 box regression while enhancing robustness to difficult samples. Experimental results demonstrate that this joint
573 loss improves the model’s $mAP@0.5$ by approximately 1.7 percentage points and significantly reduces high-

confidence false detections. Its performance is particularly prominent in dense object scenarios, indicating that the proposed loss design has clear advantages in improving detection stability and localization accuracy.

Although the proposed method has achieved promising results in wheat aphid detection, several limitations still deserve further investigation. First, complex field conditions, such as illumination variation, leaf occlusion, and changes in aphid posture, may still affect detection performance. In extremely challenging scenarios, the model may still suffer from missed detections or false detections. Second, this study is primarily based on a single dataset for model training and validation. The diversity of data sources remains limited, which may restrict the model's generalization capability across different regions, crop growth conditions, and agricultural environments. In addition, for practical deployment, the model structure still requires further optimization to better satisfy the real-time constraints of edge computing devices and mobile terminals. Future research will focus on several directions, including lightweight network design, multi-scenario dataset construction, and multi-modal information fusion, with the aim of further improving the robustness, generalization ability, and practical applicability of the model in complex agricultural environments.

6. Conclusion

To address the issues of missed detections, false detections, and insufficient detection accuracy in small object detection under complex field scenarios, this study proposes an improved YOLO-based model, HSFL-YOLO, and investigates key techniques from four aspects: feature fusion, feature pyramid optimization, small object feature enhancement, and loss function design. The proposed improvements are integrated into the YOLO detection framework through collaborative module optimization and parameter adjustment, resulting in a lightweight small object detection model with enhanced detection accuracy. The main conclusions are as follows:

(1) On the wheat aphid dataset, HSFL-YOLO achieved precision, recall, $mAP@0.5$, and $mAP@0.5:0.95$ values of 94.5%, 92.1%, 94.4%, and 53.8%, respectively. Compared with the baseline model, these metrics were improved by 2.3, 1.5, 0.4, and 1.1 percentage points, respectively. These improvements can be mainly attributed to the efficient multi-scale feature interaction capability of the HAFB module, the optimization of the feature pyramid structure, and the enhanced representation ability for small objects introduced by SOEP. As a result, the proposed model significantly improves detection performance while maintaining relatively low model complexity.

(2) The A2C2F-FRFN structure enhances the fusion of deep semantic information and shallow detail information by introducing a more effective feature reconstruction mechanism and channel-wise interaction strategy. Meanwhile, the joint use of GCD loss and Focal-EIoU loss further improves the model's attention to hard samples and enhances bounding box regression accuracy, enabling the model to maintain robust detection performance in scenarios involving dense targets and large scale variations.

(3) The visual comparison results further demonstrate that HSFL-YOLO can still achieve more complete and more reliable target detection under complex backgrounds, dense small-object distributions, and occlusion conditions. Compared with the baseline model, it effectively reduces missed detections and false detections, thereby verifying the effectiveness of the proposed improvements and their collaborative contribution to overall detection performance.

Although HSFL-YOLO achieves clear improvements in detection accuracy, there is still room for further optimization in terms of computational complexity and inference speed. Future work will focus on lightweight network design, model pruning and quantization, expansion of small object samples, and enhancement of cross-

dataset generalization ability, so as to better satisfy the real-time deployment requirements of edge devices in practical agricultural applications.

Statements and Declarations

Data Availability

The data are not publicly available due to privacy or ethical restrictions.

Funding

This research was funded by the National Natural Science Foundation of China (Grant No.31501225, Grant No.31701309; the Key Scientific and Technological Project of Henan Province (Grant No.252102110268); the Natural Science Foundation of Henan Province (Grant No.242301420143); the National key Research and Development Programs (Grant No.2017YFD0301105) ; the Key Research and Development Projects of Henan Province (Grant No.241111110080)

Author information

Haiyan Lü, Shujie Zhang, Hongbo Qiao, Jianping Wang, Jianjun Zhang, Chengjin Sun, Fei Yin*, Shufeng Xiong*, Lei Shi*

Authors and Affiliations

College of Information and Management Science, Henan Agricultural University, Zhengzhou 450046, China

Contributions

Haiyan Lü: Writing–original draft, Methodology, Funding acquisition, Conceptualization, Validation. Shujie Zhang: Writing–original draft, Data curation, Software. Hongbo Qiao : Resources, Funding acquisition, Formal analysis, Formal analysis. Jianping Wang: Visualization, Validation, Software. Jianjun Zhang: Writing–original draft, Methodology. Chengjin Sun: Visualization, Resources. Shufeng Xiong: Investigation, Data curation, Software. Fei Yin: Visualization, Validation, Resources. Lei Shi: Writing – review & editing, Supervision, Software, Resources, Funding acquisition, Formal analysis.

Corresponding author

Fei Yin*, Lei Shi*

E-mail addresses: shilei@henau.edu.cn (L. Shi), yin.fei@henau.edu.cn (F.Yin)

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Chen J, Kao SH, He H, Zhuo W, Wen S, Lee CH, Chan SHG (2023) Run, don't walk: Chasing higher FLOPS for faster neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Cui Y, Ren W, Knoll A (2024) Omni-kernel network for image restoration. In: Proceedings of the AAAI Conference on Artificial Intelligence
- Footitt RG, Halbert SE, Miller GL, Maw E, Russell LM (2005) Influence of aphid species and barley yellow dwarf virus on soft red winter wheat yield. *Journal of Economic Entomology* 98:2013–2020. <https://doi.org/10.1093/jee/98.6.2013>
- Gao C, He B, Guo W, Qu Y, Wang Q, Dong W (2025) SCS-YOLO: A real-time detection model for agricultural diseases—A case study of wheat fusarium head blight. *Computers and Electronics in Agriculture* 238:110794. <https://doi.org/10.1016/j.compag.2025.110794>
- Guan Z, Fu X, Huang P, Zhang H, Du H, Liu Y, Wang Y, Ma Q (2025) Gaussian combined distance: A generic metric for object detection. *arXiv preprint arXiv:2510.27649*. <https://arxiv.org/abs/2510.27649>
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778
- Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. <https://arxiv.org/abs/1704.04861>
- Jia YS, Qu LA, Li X et al (2024) Better prediction of greenhouse extreme temperature base on improved loss function. *Computers and Electronics in Agriculture* 227:109581. <https://doi.org/10.1016/j.compag.2024.109581>
- Jocher G et al (2023) Ultralytics YOLO. <https://docs.ultralytics.com/>. Accessed 29 Mar 2026
- Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture* 147:70–90. <https://doi.org/10.1016/j.compag.2018.02.016>
- Khan Z, Liu H, Shen Y, Yang Z, Zhang L, Yang F (2025) Optimizing precision agriculture: A real-time detection approach for grape vineyard unhealthy leaves using deep learning improved YOLOv7 with feature extraction capabilities. *Computers and Electronics in Agriculture* 231:109969
- Li X, Wang Y, Monday HN, Nneji GU (2025) A novel residual learning of multi-scale feature extraction model for the classification of rice grain varieties. *Computers and Electronics in Agriculture* 237:110491. <https://doi.org/10.1016/j.compag.2025.110491>
- Li Y, Miao N, Ma L, Shuang F, Huang X (2023) Transformer for object detection: Review and benchmark. *Engineering Applications of Artificial Intelligence* 126:107021
- Li Y, Zhang K, Cao J, Timofte R, Van Gool L (2021) LocalViT: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*. <https://arxiv.org/abs/2104.05707>
- Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2117–2125. <https://doi.org/10.1109/CVPR.2017.106>
- Liu H, Zhan Y, Sun J, Mao Q, Wu T (2025) A transformer-based model with feature compensation and local information enhancement for end-to-end pest detection. *Computers and Electronics in Agriculture* 231:109920. <https://doi.org/10.1016/j.compag.2025.109920>

- Liu Q, Lv J, Zhang C (2024) MAE-YOLOv8-based small object detection of green crisp plum in real complex orchard environments. *Computers and Electronics in Agriculture* 226:109458. <https://doi.org/10.1016/j.compag.2024.109458>
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: Single shot multibox detector. In: *Computer Vision – ECCV 2016*, pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
- Liu XB, Yang XZ, Chen Y, Zhao ST (2023) Object detection method based on CIOU improved bounding box loss function. *Chinese Journal of Liquid Crystals and Displays* 38:656–665. <https://doi.org/10.37188/CJLCD.2022-0282>
- Lu S, Chen W, Zhang X, Karkee M (2022) Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation. *Computers and Electronics in Agriculture* 193:106696. <https://doi.org/10.1016/j.compag.2022.106696>
- Mansion-Vaquié A, Wezel A, Ferrer A (2019) Wheat genotypic diversity and intercropping to control cereal aphids. *Agriculture, Ecosystems & Environment* 285:106604. <https://doi.org/10.1016/j.agee.2019.106604>
- Ouaarous M, Fakhouri EK, Taarji N et al (2025) Impact of field insect pests on seed and nutritional quality of some important crops: A comprehensive review. *ACS Omega* 10:8779–8792. <https://doi.org/10.1021/acsomega.4c08982>
- Paymode AS, Malode VB (2022) Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. *Artificial Intelligence in Agriculture* 6:23–33
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 779–788
- 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* 39:1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1874–1883. <https://doi.org/10.1109/CVPR.2016.207>
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations (ICLR)*
- Song X, Chen C, Xu J et al (2025) Current status and future prospects of wheat breeding in China. *Journal of Seed Industry Guide* (4):7–11
- Sun H, Xu H, Liu B, He D, He J, Zhang H, Geng N (2021) MEAN-SSD: A novel real-time detector for apple leaf diseases using improved light-weight convolutional neural networks. *Computers and Electronics in Agriculture* 189:106379. <https://doi.org/10.1016/j.compag.2021.106379>
- Tan M, Pang R, Le QV (2020) EfficientDet: Scalable and efficient object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 10781–10790
- Tian Y, Ye Q, Doermann D (2025) YOLOv12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*. <https://arxiv.org/abs/2502.12524>
- Wu Y, Yu X, Zhang D, Yang Y, Qiu Y, Pang L, Wang H (2025) TinySeg: A deep learning model for small target segmentation of grape pedicels with multi-attention and multi-scale feature fusion. *Computers and Electronics in Agriculture* 237:110726. <https://doi.org/10.1016/j.compag.2025.110726>

- Xing Z, Wang Y, Qu A, Yang C (2023) MFENet: Multi-scale feature extraction network for images deblurring and segmentation of swinging wolfberry branch. *Computers and Electronics in Agriculture* 215:108413. <https://doi.org/10.1016/j.compag.2023.108413>
- Xu D, Wang C, Li M, Ge X, Zhang J, Wang W, Lv C (2025) Improving passion fruit yield estimation with multi-scale feature fusion and density-attention mechanisms in smart agriculture. *Computers and Electronics in Agriculture* 239:110958. <https://doi.org/10.1016/j.compag.2025.110958>
- Xu D, Huang X, Zhao Z, Zhao Z, Hu D, Yuan C (2026) Pha-YOLO: A multi-view Phalaenopsis flower detection and counting method based on improved YOLO11 with dynamic reference selection and adaptive thresholding. *Computers and Electronics in Agriculture* 250:111446. <https://doi.org/10.1016/j.compag.2026.111446>
- Xu S, Zheng S, Xu W et al (2024) HCF-Net: Hierarchical context fusion network for infrared small object detection. *arXiv preprint arXiv:2403.10778*. <https://arxiv.org/abs/2403.10778>
- Yang L, Zeng Y, Xu L, Li M, Wang H, Zou Y, Lu Y (2022) Perennial woodlands benefit parasitoid diversity, but annual flowering fallows enhance parasitism of wheat aphids in an agricultural landscape. *Agriculture, Ecosystems & Environment* 340:108184. <https://doi.org/10.1016/j.agee.2022.108184>
- Yuan L, Huang Y, Loraamm RW, Nie C, Wang J, Zhang J (2014) Spectral analysis of winter wheat leaves for detection and differentiation of diseases and insects. *Field Crops Research* 156:199–207. <https://doi.org/10.1016/j.fcr.2013.11.012>
- Zhang L, Du J, Dong S, Wang F, Xie C, Wang R (2022) AM-ResNet: Low-energy-consumption addition-multiplication hybrid ResNet for pest recognition. *Computers and Electronics in Agriculture* 202:107357. <https://doi.org/10.1016/j.compag.2022.107357>
- Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022) Focal and efficient IoU loss for accurate bounding box regression. *Neurocomputing* 506:146–157
- Zhou S, Chen D, Pan J, Shi J, Yang J (2024) Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2952–2963