agronomy

MDPI

Article

Lightweight Structure and Attention Fusion for In-Field Crop Pest and Disease Detection

Zijing Luo [1], Yunsen Liang [2], Naimin Kong [1,3], Lirui Liang [1], Wenjun Peng [1], Yujie Yao [2], Chi Qin [1], Xiaohan Lu [1], Mingman Xu [3], Yining Zhang [1], Chenyang Lin [4], Chengyao Jiang [1], Mengyao Li [1], Yangxia Zheng [1], Yameng Jiang [5,6,]* and Wei Lu [1,]*

[1] College of Horticulture, Sichuan Agricultural University, Chengdu 611130, China
[2] College of Information Engineering, Sichuan Agricultural University, Ya'an 625014, China
[3] College of Electrical and Mechanical Engineering, Sichuan Agricultural University, Ya'an 625014, China
[4] College of Resouces, Sichuan Agricultural University, Chengdu 611130, China
[5] Institute of Urban Agriculture, Chinese Academy of Agricultural Sciences, Chengdu 610213, China
[6] Information Technology Group, Wageningen University & Research, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands
* Correspondence: yameng.jiang@wur.nl (Y.J.); weilu@sicau.edu.cn (W.L.)

Abstract

In agricultural production, plant diseases and pests are among the major threats to crop yield and quality. Existing agricultural pest and disease identification methods have problems such as small target scales, complex background environments, and unbalanced sample distributions. This paper proposes a lightweight improved target detection model, YOLOv5s-LiteAttn. Based on YOLOv5s, the model introduces GhostConv and Depthwise Conv to reduce the number of parameters and computational complexity, and it combines CBAM and Coordinate Attention mechanisms to enhance the network's feature representation capability. Experimental results show that, compared with the basic YOLOv5s model, the number of parameters of the improved model is reduced by 22.75%, and the computational load is reduced by 16.77%. At the same time, mAP@0.5–0.95 is increased by 3.3 percentage points, and recall is improved by 1.1 percentage points. In addition, the inference speed increases from 121 FPS to 142 FPS at an input resolution of 640 × 640, further confirming that the proposed model achieves a favorable trade-off between accuracy and efficiency. The average precision of YOLOv5s-LiteAttn is 97.1%, which outperforms the existing mainstream lightweight detection models. Moreover, an independent test set containing 4328 newly collected field images was established to evaluate generalization and practical applicability. Despite a slight performance decrease compared with the validation results, the model maintained an mAP@0.5–0.95 of 95.8%, significantly outperforming the baseline model, thereby confirming its robustness and cross-domain adaptability. These results confirm that the model has high precision and is lightweight, making it effective for the detection of agricultural diseases and pests.

Keywords: agricultural pest and disease detection; lightweight model; YOLOv5; attention mechanism; target detection; mobile deployment; deep learning

check for updates

Academic Editor: Paul Kwan

Received: 25 October 2025
Revised: 26 November 2025
Accepted: 9 December 2025
Published: 15 December 2025

Citation: Luo, Z.; Liang, Y.; Kong, N.; Liang, L.; Peng, W.; Yao, Y.; Qin, C.; Lu, X.; Xu, M.; Zhang, Y.; et al. Lightweight Structure and Attention Fusion for In-Field Crop Pest and Disease Detection. *Agronomy* **2025**, *15*, 2879. https://doi.org/10.3390/agronomy15122879

1. Introduction

In the process of agricultural production, plant diseases and pests have always been one of the main factors threatening the stable yield and income increase in crops. Pests, by feeding on crop tissues, transmitting pathogens, or causing abiotic stress, not only hinder crop growth and reduce quality but also lead to the premature death of plants [1,2]. At the same time, crop diseases are also widespread and often occur in conjunction with pests, further exacerbating crop losses [3]. Therefore, how to achieve accurate identification and efficient control of agricultural diseases and pests has always been a core issue in agricultural scientific research and production practice.

Traditional pest and disease identification methods mainly rely on manual surveys and expert experience. Although this approach is intuitive, it has problems such as high labor intensity, low efficiency, strong subjectivity, and proneness to errors. Particularly in large-area farmlands, manual identification can hardly cover the entire area in a timely manner, which often leads to the miss of the optimal prevention and control period and causes greater losses [4]. With the development of agricultural informatization and intelligence, automatic detection methods based on image processing and pattern recognition have gradually emerged [5]. Early Machine learning methods, such as Support Vector Machines (SVM) [6], Artificial Neural Networks (ANN) [7], and Genetic Algorithms (GA) [8], have been applied to pest identification tasks. These methods can achieve automatic classification and recognition of crop images to a certain extent, but they usually rely on manual feature extraction, have poor adaptability to complex backgrounds and variable pest morphologies, and have limited generalization ability, making it difficult to meet the requirements for real-time performance and robustness in actual production [9].

In recent years, the rapid development of Deep Learning (DL) has provided new opportunities for the intelligent identification of agricultural pests and diseases [10]. Models such as Convolutional Neural Networks (CNNs) can automatically extract deep features from images and have shown significant advantages in target detection and image classification tasks. In the field of pest and disease detection, deep learning technology has not only improved the recognition accuracy but also significantly enhanced the detection efficiency. For example, Amrani et al. proposed a pest detection model based on Bayesian multi-task learning, which uses ResNet18 as the backbone network to achieve aphid recognition, with an accuracy rate ranging from 59% to 75.77% [11]. Ye et al. proposed the PestNAS model, which optimizes the network structure through adaptive feature fusion and evolutionary neural architecture search, and is superior to traditional models in terms of accuracy [12]. Domestically, Zhang Huan et al. improved the fruit tree pest recognition model based on MobileViT, introducing Pania convolution and atrous spatial pyramid pooling modules. While increasing the accuracy by 7.5%, the model parameters were reduced by 33.86%, demonstrating the potential of balancing lightweight and high precision [13]. In recent years, knowledge graphs [14] have been introduced into agricultural pest and disease research. By integrating heterogeneous data such as pest morphological characteristics, damage symptoms, and control methods, semantic relationships between entities are established to provide support for intelligent diagnosis and question-answering systems [15]. This direction has provided new ideas for the comprehensive identification and intelligent prevention and control of agricultural pests and diseases, but its combination with deep learning models is still in the exploratory stage.

How to balance recognition accuracy with real-time performance and model lightweightness has become the core challenge in agricultural pest and disease detection. Although traditional convolutional structures can enhance feature expression, they often come at the cost of linear growth in FLOPs and model parameters; while pure attention mechanisms are prone to overfitting or redundancy under lightweight constraints, resulting in poor application effects of the model in actual agricultural scenarios [16–18]. In this study, YOLOv5s [19] was first selected as the base model because it has shown good adaptability in agricultural pest and disease detection, which is specifically reflected in three aspects: first, it has real-time processing capability, suitable for dynamic recognition needs; second, it has strong robustness to complex backgrounds (such as occlusion by crop leaves); third, it has demonstrated superior performance in similar biological detection tasks. To address these challenges, this study proposes YOLOv5s-LiteAttn, an improved lightweight detection model based on YOLOv5s. Our work makes the following key contributions: (1) Constructed a computationally efficient backbone by integrating GhostConv and Depthwise Convolution modules, significantly reducing the model's parameter count and computational load. (2) Introduced a hybrid attention mechanism, combining CBAM and Coordinate Attention, to enhance the network's feature representation capability, leading to improved accuracy in identifying small and occluded pests. (3) Systematically elucidated the performance trade-offs through comprehensive ablation studies and benchmark comparisons, demonstrating that our model achieves a superior balance between high accuracy and lightweight design, making it highly suitable for real-world agricultural deployment.

2. Materials and Methods

2.1. Overview of the Model Improvements

The structure of YOLOv5 generally consists of four parts: the input end, the backbone network, the neck (feature fusion network), and the head (prediction layer). According to different application requirements, YOLOv5 consists of multiple versions such as YOLOv5s, YOLOv5n, YOLOv5m, YOLOv5l, and YOLOv5x. Among them, YOLOv5s is mainly characterized as lightweight and high efficiency, with significantly reduced parameters and computation time, making it suitable for resource-constrained environments and mobile deployment. While ensuring detection accuracy, YOLOv5s also has real-time performance and portability, which can better meet the dual needs of high efficiency and being lightweight for agricultural pest and disease identification. Therefore, this study selects YOLOv5s as the basic detection model for the detection and identification of agricultural crop pests and diseases.

This paper proposes the YOLOv5s-LiteAttn model based on YOLOv5s, optimizing it from two aspects: lightweight design and attention mechanism. In terms of lightweighting, the GhostBottleneck [20] module, as shown in Figure 1, generates intrinsic features through a small number of convolutions, then expands them into high-dimensional representations using inexpensive linear transformations, and combines with Depthwise [21] convolution to provide a local receptive field. This not only significantly reduces FLOPs and parameters but also retains lesion textures.

In terms of the attention mechanism, CBAM [22] acts in the Backbone stage, suppressing background noise through step-by-step screening of channels and spaces to extract cleaner features of crops and pests; Coordinate Attention [23] introduces direction awareness and long-range dependence in the Head stage, performing geometric and positional refinement and re-calibration on the features after multi-scale fusion.

This "first compression then focusing" hierarchical design forms a clear information processing path: GhostBottleneck ensures high information density input under the premise of lightweight, CBAM filters out interference and highlights significant regions in the Backbone, CoordAtt refines the structure and positioning in the Head, and C3 [24] completes cross-scale residual fusion. Compared with the original YOLOv5s, this design reduces both the number of parameters and FLOPs, with the incremental overhead of the attention module being low and controllable. The "cheap-to-rich" generation of GhostBottleneck improves the effective dimension under unit computing power, CBAM increases the feature signal-to-noise ratio, CoordAtt integrates spatial direction priors, and the residual path ensures the stability of gradient flow. Finally, this structure, as shown in Figure 2, significantly enhances the ability to distinguish and locate pest and disease targets while maintaining lightweight, achieving a dual improvement in detection accuracy and efficiency under limited computing power conditions.

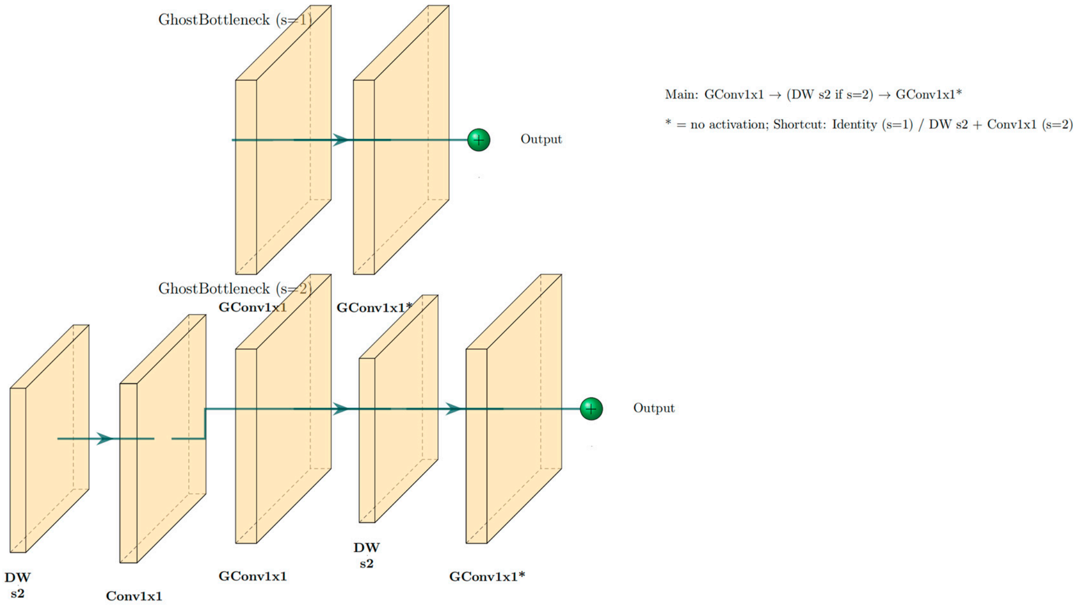

GhostBottleneck (s=1)
Output
Main: GConv1x1 → (DW s2 if s=2) → GConv1x1*
* = no activation; Shortcut: Identity (s=1) / DW s2 + Conv1x1 (s=2)
GhostBottleneck (s=2)
GConv1x1
GConv1x1*
Output
DW
s2
Conv1x1
GConv1x1
DW
s2
GConv1x1*

Figure 1. Architecture of the GhostBottleneck Module.

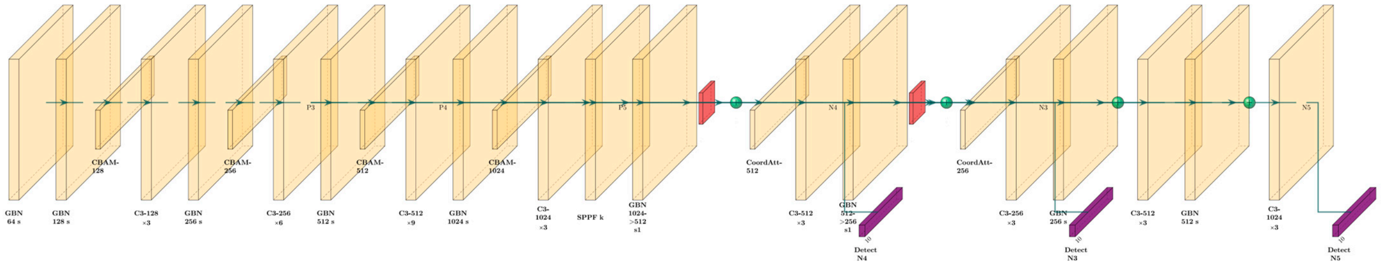


Figure 2. Architecture of the YOLOv5s-LiteAttn Module.

2.2. GhostConv

GhostConv [20] is a lightweight convolution module, as shown in Figure 3, whose design goal is to improve the computational efficiency of convolutional neural networks. Compared with traditional convolution layers, GhostConv reduces redundant computations in convolution operations by generating "ghost features", thereby significantly reducing the number of parameters and computational complexity while ensuring the feature expression ability. Its core idea is to first extract some intrinsic feature maps through standard convolution, and then generate additional ghost feature maps using cheap linear transformation operations, so as to expand the feature representation ability and improve computational efficiency. The structure of GhostConv consists of two stages: the first stage uses standard convolution to obtain original features, and the second stage generates more pseudo-features through lightweight operations. These pseudo-features and the original features together form a complete feature map. Compared with traditional convolution, this method can significantly reduce the amount of computation while maintaining similar feature expression capabilities.

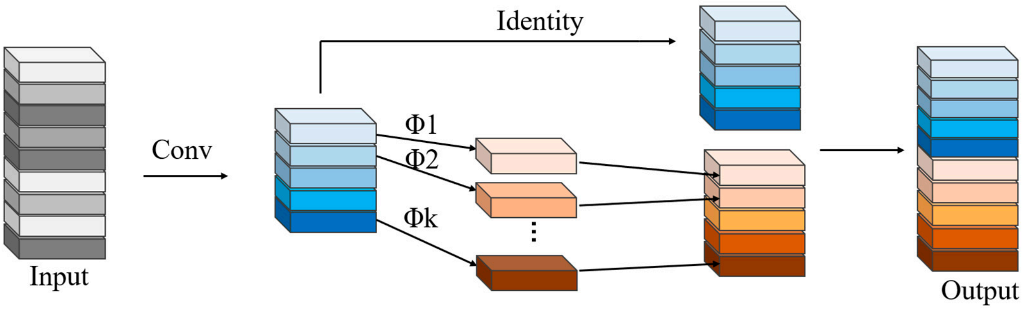


Figure 3. Architecture of the GhostConv.

2.3. Depthwise Convolution

Depthwise Convolution [21] (DWConv) is a lightweight convolution operation method aimed at significantly reducing the number of parameters and computational load while ensuring the feature extraction capability. Unlike traditional convolution, which performs fully connected calculations between input channels and output channels, DWConv decomposes convolution into two steps: first, it independently performs spatial convolution (Depthwise) on each input channel, and then completes information fusion between channels through pointwise convolution (Pointwise, 1 × 1 convolution). This can effectively reduce the redundancy of convolution operations and significantly lower FLOPs and the number of parameters, making it particularly suitable for deployment in mobile and computing power-constrained scenarios.

Compared with standard convolution, DWConv has significant advantages in terms of model size and latency, and shows a good trade-off ability in maintaining model accuracy. DWConv is widely used in lightweight networks such as MobileNetV1 and MobileNetV2 [25], which verifies its core role in efficient model design. The structure of DWConv is shown in Figure 4.

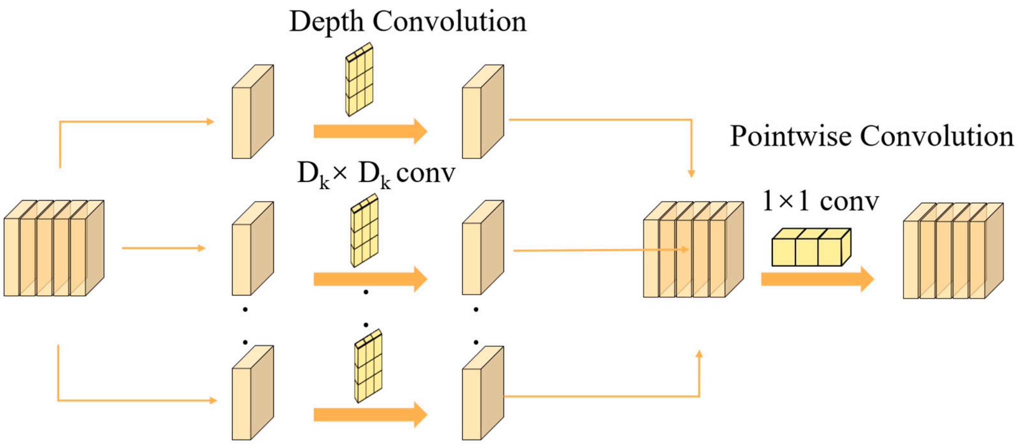


Figure 4. Architecture of the DWCon.

The calculation process of DWConv can be formally expressed as follows: the input feature map is grouped by channels, and each channel is convolved only with its corresponding convolution kernel to capture local spatial features; then, cross-channel linear combination is performed through pointwise convolution to complete feature fusion. This design maintains the representational capability of convolutional neural networks while significantly reducing the consumption of computing resources, making it one of the fundamental operators in lightweight neural networks.

2.4. CBAM

The attention mechanism CBAM [22] considers two important aspects, namely the spatial and channel dimensions, when processing features. It simultaneously employs global average pooling and global max pooling to ensure that key information is not lost. As shown in Figure 5, the CBAM module includes a Channel Attention Module (CAM) and a Spatial Attention Module (SAM).

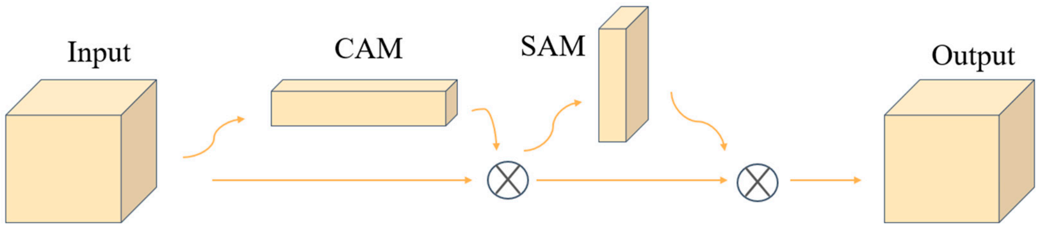


Figure 5. Architecture of the CBAM.

The channel dimension is mainly used to capture high-level abstract information of features, while the spatial dimension focuses more on preserving the position information of objects. Attention mechanisms in both the channel and spatial dimensions are introduced, respectively, to more comprehensively process the information of feature maps while saving parameters and computing power. The input feature map F obtains the channel attention map through the channel attention module. $M_c(F)$ and $M_c(F)$ are multiplied element-wise to obtain the feature map F, as shown in Equation (1).

$$F^{'} = M_c(F) \otimes F \tag{1}$$

Apply the spatial attention module $M_s(F')$ to F' and multiply it element-wise with F' to obtain the final feature map F, as shown in Equation (2)

$$F'' = M_S(F^{'}) \otimes F^{'} \tag{2}$$

$M_c(F)$ the definition includes a multi-layer perceptron (MLP) processing the feature maps that have undergone average pooling (AvgPool)and max pooling (MaxPool), ($F^G{}_{avg}$ and $F^G{}_{max}$) as shown in Equation (3).

$$\begin{aligned} M_\tau(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \\ &= \sigma\Big(W_1\Big(W_0\Big(F^G_{\mathrm{avg}}\Big)\Big) + W_1\big(W_0\big(F^G_{max}\big)\big)\Big) \end{aligned} \tag{3}$$

where σ—sigmoid operation; W_0—weight matrix of the 1st layer; W_1—weight matrix of the 2nd layer.

$M_c(F)$ after processing the feature map with channel attention through global average pooling ($F^G{}_{avg}$) and global max pooling ($F^G{}_{max}$),a 7 × 7 convolution operation is used to generate a spatial attention map, as shown in Equation (4)

$$\begin{aligned} M_*(F) &= \sigma\big(f^{7*7}([\mathrm{AvgPool}(F), \mathrm{MaxPool}(F)])\big) \\ &= \sigma\Big(7^{T\times 7}\Big(\Big[F^*_{\mathrm{avg}}\,; F^*_{\mathrm{avg}}\Big]\Big)\Big) \end{aligned} \tag{4}$$

2.5. Coordinate Attention

Coordinate Attention [23] (CA) is an attention mechanism that balances fine positioning and efficient representation, aiming to introduce long-range dependencies while maintaining lightweight feature modeling. Unlike traditional channel attention modules that only capture global statistics, CA explicitly encodes spatial information into attention weights, and realizes the fusion of position perception and channel selection through decomposed direction-sensitive modeling. Its structure is shown in Figure 6.

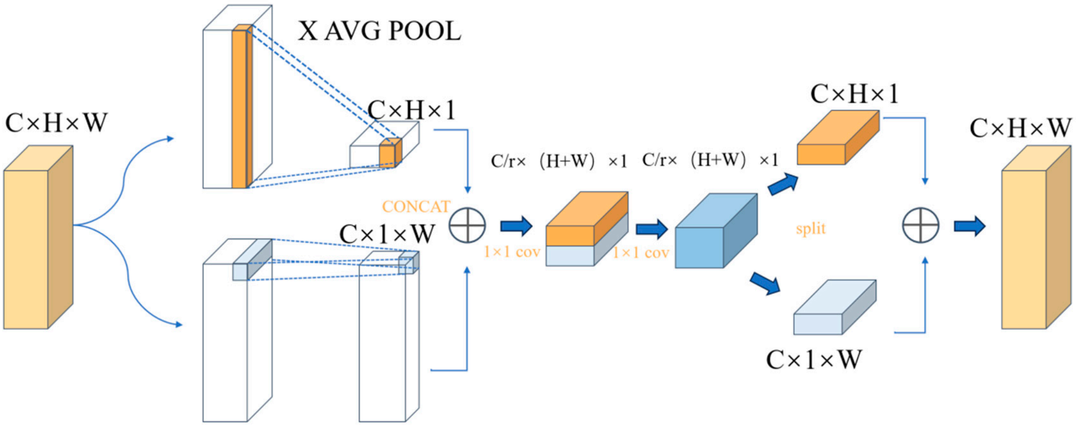


Figure 6. Architecture of the CA.

In terms of design, CA first performs global average pooling on the input features in the horizontal and vertical directions, respectively, to obtain two sets of one-dimensional descriptors containing directional position information. Subsequently, these descriptors are compressed into a low-dimensional space through a shared transformation network to reduce computational overhead and capture cross-channel correlations. Then, the original number of channels is restored through branch mapping, and position-sensitive attention weights are generated by broadcasting in the horizontal and vertical directions, respectively. Finally, it is multiplied point by point with the original features to complete the channel-position joint recalibration.

The advantages of CA are reflected in two aspects: first, it uses decomposed direction modeling to explicitly introduce spatial position information into channel attention, thereby enhancing the network's ability to express target geometric structures and long-range dependencies while maintaining lightweight; second, the module design is concise and can be seamlessly embedded into existing convolutional networks, bringing stable accuracy improvements in tasks such as image classification, object detection, and semantic segmentation.

2.6. Dataset

The dataset used in this study is constructed by combining the PlantDoc [26] dataset with additional images collected through Google and Bing. Only images from webpages that explicitly permit academic or non-commercial research use, or that do not include restrictive copyright statements, were downloaded; for webpages with unclear licensing, only URLs were recorded and the raw images were not included in the dataset. Images captured through search engines, covering 9 major categories, including 9 types of healthy conditions and 20 types of disease and pest conditions, involving common cash crops such as apples, tomatoes, strawberries, and potatoes. To ensure data quality and avoid redundancy, perceptual hashing (pHash) was applied to detect duplicate or near-duplicate images between PlantDoc and web-collected samples. All automatically detected cases were manually verified by two researchers. Images with low resolution, severe blur, overexposure, large occlusion, or prominent watermarking were removed. The dataset contains 28,721 images, each with a resolution of 640 × 640 pixels. Use the Labellmg annotation tool to annotate the target positions and categories of diseases and pests in the filtered images. Approximately 15% of all annotations were independently reviewed by two plant protection experts. For images with annotation inconsistencies, the final labels were determined through consensus discussion to ensure biological correctness and minimize systematic labeling errors. Refer to PASCALVOC2007 to generate annotation files in xml format, which record the image name, image size, types of diseases and pests, and location information of diseases and pests, as shown in Table 1.

Table 1. Dataset Category Details.

Species	Class	Num	Category
Apple	Apple Healthy	949	Healthy Plant
	Apple Black Rot	943	Fungal Disease
	Apple Scab	945	
	Cedar Apple Rust	948	
Bell-pepper	Bell-pepper Healthy	945	Healthy Plant
	Bell pepper Bacterial Spot	944	Bacterial Disease
Cherry	Cherry Healthy	940	Healthy Plant
	Cherry Powdery Mildew	941	Fungal Disease
Corn	Corn Healthy	940	Healthy Plant
	Corn Cercospora Leaf Spot	953	Fungal Disease
	Corn Common Rust	964	
	Northern Leaf Blight	1003	
Grape	Grape Healthy	946	Healthy Plant
	Grape Black Rot	945	Fungal Disease
	Grape Leaf Blight	966	
	Grape Esca	942	Other Diseases/Conditions
Peach	Peach Healthy	941	Healthy Plant
	Peach Bacterial Spot	946	Bacterial Disease
Potato	Potato Healthy	934	Healthy Plant
	Potato Early Blight	947	Fungal Disease
	Potato Late Blight	947	
Strawberry	Strawberry Healthy	946	Healthy Plant
	Strawberry Leaf Scorch	1048	Other Diseases/Conditions
Tomato	Tomato Leaf Healthy	946	Healthy Plant
	Tomato Bacterial Spot	956	Bacterial Disease
	Tomato Leaf Mould	944	Fungal Disease
	Tomato Early Blight	967	
	Tomato Late Blight	1067	
	Tomato Septoria Leaf Spot	952	

2.7. Implementation Details

The number of training epochs is set to 100 and the batch size is 64. To enhance the model's generalization capability under complex field conditions, multiple data augmentation strategies were employed during training, including random horizontal flipping, random rotation (±10°), random scaling (0.8–1.2), brightness and contrast perturbation, HSV jittering, and Gaussian noise injection. Mosaic and MixUp augmentation were additionally applied during early training epochs to further increase sample diversity. The data is divided into a training set and a valid set in an approximate ratio of 7:3. The experiment was conducted in an environment built on Ubuntu 22.04 operating system with Python 3.9 and Pytorch 2.6.0. The CPU model is Intel Xeon Platinum 8352V @ 2.10 GHz, and the GPU model is NVIDIA RTX 4090 (24 GB). Meanwhile, CUDA 12.4 was used to accelerate the computations. The optimizer was SGD with a momentum of 0.937, an initial learning rate of 0.01, a cosine annealing learning rate scheduler with a three-epoch warm-up, and a weight decay of 0.0005. No pretrained weights were used to ensure full reproducibility.

2.8. Evaluation Metrics

This study mainly uses precision P, R, F_1 scores, and mean average precision (mAP) to measure model accuracy, and it uses parameter count (Params) and computational complexity (FLOPs) to evaluate model performance (Equations (5)–(8)).

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

$$\text{mAP} = \frac{\sum_{i=1}^{N} \int_0^1 P_i(R)\mathrm{d}R}{N} \times 100\% \tag{8}$$

where TP is the number of positive samples correctly predicted as positive samples; FP is the number of negative samples incorrectly predicted as positive samples; FN is the number of positive samples incorrectly predicted as negative samples; N is the number of categories.

3. Results and Discussion

3.1. Ablation Study

To systematically validate the independent effects of individual lightweight modules and attention mechanisms within the YOLOv5s framework, as well as the advantages of multi-module collaborative fusion, this section designs and implements a set of ablation experiments. YOLOv5s was adopted as the baseline model. First, single lightweight modules (including GhostConv and Depthwise Convolution) and single attention mechanisms (including the CBAM channel-spatial attention mechanism and Coordinate Attention) were separately embedded into either the Backbone or Neck components of the baseline model. Under the premise of preserving the overall topological structure of the network, a series of improved single-module models were constructed via module replacement or insertion. Subsequently, the performance of each single-module model was compared and analyzed against that of the target model (YOLOv5s-LiteAttn), which integrates multiple modules through collaborative fusion.

The evaluation process centered on both the core performance metrics for object detection tasks and the lightweight properties of the model, specifically encompassing precision, recall, mean average precision (mAP@0.5–0.95), floating-point operations (FLOPs), and parameter count. By quantitatively assessing the contribution of each individual module to the model's detection accuracy and efficiency, the functional roles of different modules were clearly defined. Concurrently, this study verified the necessity and superiority of multi-module collaborative fusion in enhancing the model's comprehensive performance—particularly in balancing detection accuracy and computational efficiency. These findings provide experimental evidence to support the rationality of the structural design of the YOLOv5s-LiteAttn model. It can be seen from the training loss curve shown in Figure 7. It illustrates the variation trends of different loss components during the training process of the model. The horizontal axis represents the number of training epochs, while the vertical axis denotes the corresponding loss values. As shown in the figure, the Box Loss, Object Loss, and Class Loss all exhibit a rapid decline in the early training stages, followed by a gradual stabilization as the number of epochs increases.

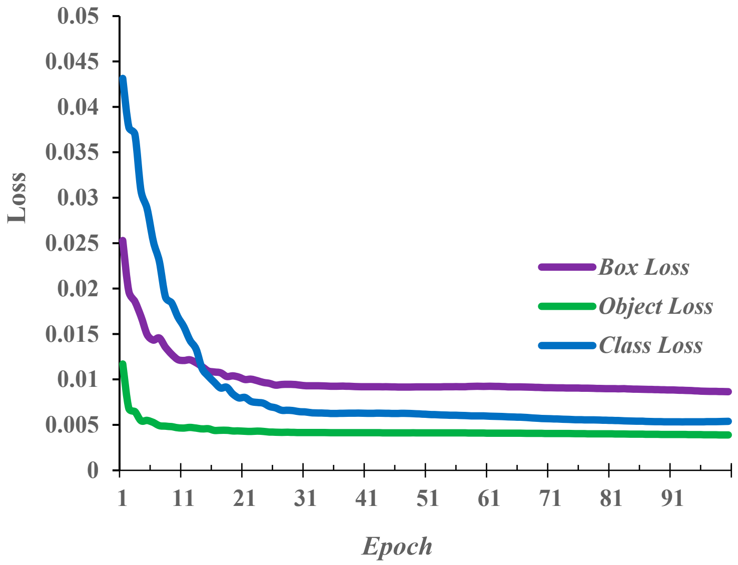


Figure 7. Training loss curve.

Specifically, the Box Loss (purple curve) decreases sharply within the first few epochs and then converges to a stable level around 0.01, indicating that the model quickly learns accurate bounding box regression. The Object Loss (green curve) shows the lowest magnitude among the three, demonstrating that the model effectively captures the confidence in objects and achieves good localization capability. Meanwhile, the Class Loss (blue curve) initially presents the highest value but undergoes the most significant reduction, eventually stabilizing at a low level, reflecting the model's enhanced ability to distinguish different categories as training progresses.

Overall, the declining and stable trends of all loss components suggest that the model training process converges effectively, and the proposed network achieves a balance between classification and localization performance.

YOLOv5s-Ghost replaces the standard convolution in the Backbone with GhostConv to test its effect in reducing redundant computations and the number of parameters; YOLOv5s-DW uses Depthwise Convolution to replace some standard convolutions to evaluate its performance in reducing FLOPs while maintaining feature extraction capability; YOLOv5s-CBAM embeds the CBAM module in the Backbone to verify the contribution of channel and spatial attention in feature screening and suppressing background interference; YOLOv5s-CoordAtt introduces Coordinate Attention in the fusion stage of Neck and Head to test the improvement effect of position perception and direction-sensitive modeling on fine-grained target detection. Through the above experimental design, the impact of a single module on model performance can be comprehensively evaluated, and a comparison basis can be provided for the finally proposed "dual lightweight + dual attention" YOLOv5s-LiteAttn, thereby proving the necessity and superiority of multi-module collaborative fusion.

Compared to the YOLOv5-CBAM-C3TR [27] model, which achieved mAP@0.5:0.95 of 40.9%, precision of 70.9%, and recall of 69.5% in apple leaf disease detection, the proposed YOLOv5s-LiteAttn model demonstrates significantly higher performance, with a mAP@0.5–0.95 of 97.1% and a recall rate of 97.9%. Furthermore, YOLOv5s-LiteAttn reduces the model parameters to 5.50 M, a 23% reduction compared to YOLOv5-CBAM-C3TR (7.12 M), and decreases the computational load (FLOPs) by 13.40%. These results indicate that YOLOv5s-LiteAttn not only outperforms YOLOv5-CBAM-C3TR in terms of detection accuracy but also effectively reduces the computational cost, making it more suitable for practical deployment in resource-constrained environments.

As shown in Table 2, the baseline YOLOv5s achieves a precision of 95.1%, a recall of 96.8%, and an mAP@0.5–0.95 of 93.8%, with 7.12 M parameters and 16.10 GFLOPs. Replacing standard convolution with GhostConv substantially reduces the parameters to 5.89 M and lowers FLOPs to 13.30 G, while improving mAP to 96.1%, demonstrating that GhostConv effectively removes redundant computation without degrading detection quality. Depthwise Convolution further reduces computation to 11.46 GFLOPs and maintains competitive accuracy, confirming its advantage for lightweight model design. Regarding attention mechanisms, CBAM increases recall but slightly decreases overall mAP, whereas Coordinate Attention achieves a more notable gain (95.7% mAP), indicating better performance in fine-grained spatial feature refinement.

Table 2. Ablation results.

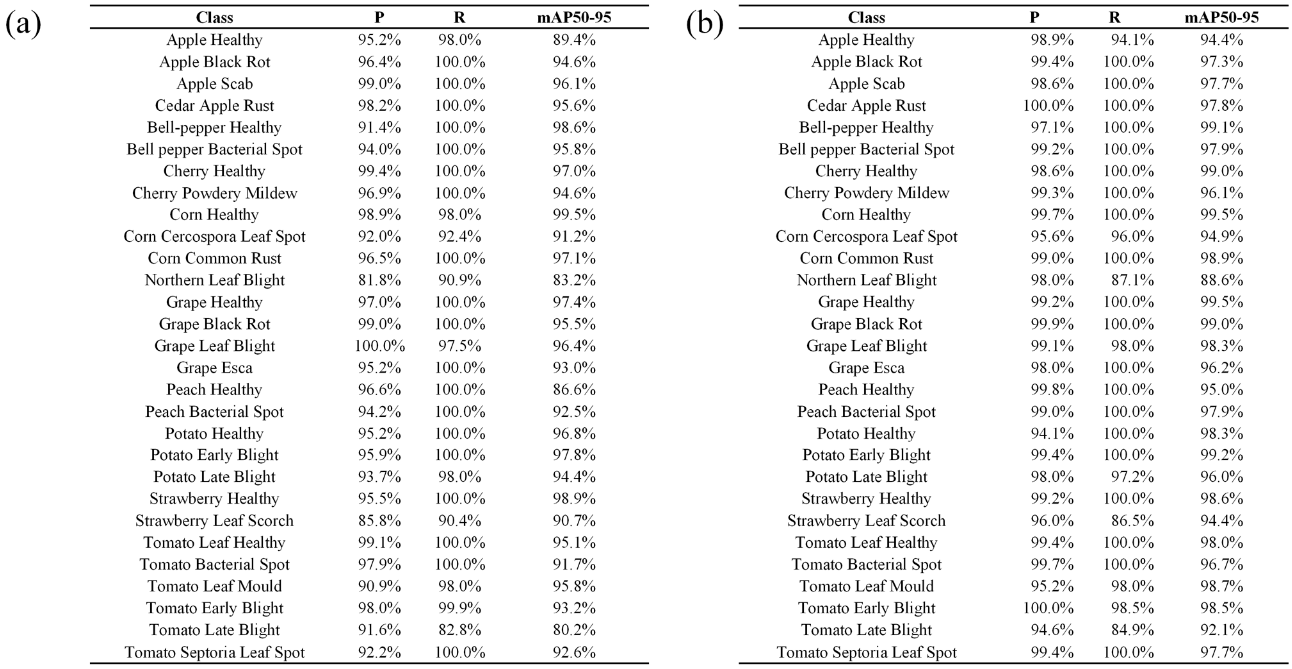
Model	Layers	Params (M)	FLOPs (GF)	P (%)	R (%)	mAP@0.5–0.95 (%)	FPS	Size (MB)
YOLOv5s Baseline	157	7.12	16.10	95.1	96.8	93.8	121.00	13.90
YOLOv5s-Ghost	184	5.89	13.30	96.1	96.2	96.1	147.00	11.58
YOLOv5s-CBAM	185	6.15	14.10	94.1	97.2	91.3	133.00	12.13
YOLOv5s-CoordAtt	245	7.15	16.20	95.0	94.0	95.7	112.00	14.02
YOLOv5s-DW	223	5.09	11.46	96.0	96.0	92.6	167.00	10.00
YOLOv5s-LiteAtten	394	5.50	13.40	98.4	97.9	97.1	142.00	11.00

Among all variants, YOLOv5s-LiteAttn—combining GhostConv, Depthwise Convolution, CBAM, and Coordinate Attention—achieves the best overall performance. It attains 97.1% mAP@0.5–0.95 and 97.9% recall while reducing parameters to 5.50 M and FLOPs to 13.40 G. Compared with the baseline, this corresponds to a 3.3-point increase in mAP and a 1.1-point improvement in recall, along with 22.75% fewer parameters and 16.77% fewer FLOPs. These results clearly demonstrate the effectiveness of multi-module collaborative fusion and highlight the superior balance between accuracy and efficiency achieved by the proposed YOLOv5s-LiteAttn model.

3.2. Comparison Between Baseline YOLOv5s and YOLOv5s-LiteAttn

In complex crop pest and disease identification scenarios, by comparing the detection performance of YOLOv5s and YOLOv5s-LiteAttn in Figure 8, the differences between the models can be intuitively observed.

The original YOLOv5s is prone to interference from the background when multiple types of pests and diseases are densely distributed, and some small targets have positioning deviations or category confusion. However, YOLOv5s-LiteAttn achieves efficient feature extraction by introducing GhostBottleneck and Depthwise convolution into the backbone network, and combines CBAM and Coordinate Attention to enhance the expressive ability of significant regions, making the detection results more stable and accurate. Especially when there is mutual interference between disease spots and leaf textures, the improved model can more clearly distinguish the boundaries of pest and disease targets and reduce misidentifications. At the same time, under conditions of light changes or strong image noise, the attention mechanism effectively suppresses irrelevant features and improves the robustness of the model in complex backgrounds. Overall, YOLOv5s-LiteAttn outperforms YOLOv5s in both small target detection accuracy and positioning accuracy, showing stronger adaptability and practical value. Combined with the quantitative indicators in Table 2, the improved model increases mAP@0.5–0.95 by 3.3 percentage points and recall by 1.1 percentage points compared with YOLOv5s, further confirming its advantages in detection performance. This is generally consistent with the findings reported in Vegetable Disease Detection Using an Improved YOLOv8 Algorithm in the Greenhouse Plant Environment [28]. However, the independently developed YOLOv5s-LiteAttn model in this study demonstrates superior detection performance under a more stringent mAP@0.5–0.95 evaluation criterion and with a greater number of detection categories.



(a)

Class	P	R	mAP50-95
Apple Healthy	95.2%	98.0%	89.4%
Apple Black Rot	96.4%	100.0%	94.6%
Apple Scab	99.0%	100.0%	96.1%
Cedar Apple Rust	98.2%	100.0%	95.6%
Bell-pepper Healthy	91.4%	100.0%	98.6%
Bell pepper Bacterial Spot	94.0%	100.0%	95.8%
Cherry Healthy	99.4%	100.0%	97.0%
Cherry Powdery Mildew	96.9%	100.0%	94.6%
Corn Healthy	98.9%	98.0%	99.5%
Corn Cercospora Leaf Spot	92.0%	92.4%	91.2%
Corn Common Rust	96.5%	100.0%	97.1%
Northern Leaf Blight	81.8%	90.9%	83.2%
Grape Healthy	97.0%	100.0%	97.4%
Grape Black Rot	99.0%	100.0%	95.5%
Grape Leaf Blight	100.0%	97.5%	96.4%
Grape Esca	95.2%	100.0%	93.0%
Peach Healthy	96.6%	100.0%	86.6%
Peach Bacterial Spot	94.2%	100.0%	92.5%
Potato Healthy	95.2%	100.0%	96.8%
Potato Early Blight	95.9%	100.0%	97.8%
Potato Late Blight	93.7%	98.0%	94.4%
Strawberry Healthy	95.5%	100.0%	98.9%
Strawberry Leaf Scorch	85.8%	90.4%	90.7%
Tomato Leaf Healthy	99.1%	100.0%	95.1%
Tomato Bacterial Spot	97.9%	100.0%	91.7%
Tomato Leaf Mould	90.9%	98.0%	95.8%
Tomato Early Blight	98.0%	99.9%	93.2%
Tomato Late Blight	91.6%	82.8%	80.2%
Tomato Septoria Leaf Spot	92.2%	100.0%	92.6%

(b)

Class	P	R	mAP50-95
Apple Healthy	98.9%	94.1%	94.4%
Apple Black Rot	99.4%	100.0%	97.3%
Apple Scab	98.6%	100.0%	97.7%
Cedar Apple Rust	100.0%	100.0%	97.8%
Bell-pepper Healthy	97.1%	100.0%	99.1%
Bell pepper Bacterial Spot	99.2%	100.0%	97.9%
Cherry Healthy	98.6%	100.0%	99.0%
Cherry Powdery Mildew	99.3%	100.0%	96.1%
Corn Healthy	99.7%	100.0%	99.5%
Corn Cercospora Leaf Spot	95.6%	96.0%	94.9%
Corn Common Rust	99.0%	100.0%	98.9%
Northern Leaf Blight	98.0%	87.1%	88.6%
Grape Healthy	99.2%	100.0%	99.5%
Grape Black Rot	99.9%	100.0%	99.0%
Grape Leaf Blight	99.1%	98.0%	98.3%
Grape Esca	98.0%	100.0%	96.2%
Peach Healthy	99.8%	100.0%	95.0%
Peach Bacterial Spot	99.0%	100.0%	97.9%
Potato Healthy	94.1%	100.0%	98.3%
Potato Early Blight	99.4%	100.0%	99.2%
Potato Late Blight	98.0%	97.2%	96.0%
Strawberry Healthy	99.2%	100.0%	98.6%
Strawberry Leaf Scorch	96.0%	86.5%	94.4%
Tomato Leaf Healthy	99.4%	100.0%	98.0%
Tomato Bacterial Spot	99.7%	100.0%	96.7%
Tomato Leaf Mould	95.2%	98.0%	98.7%
Tomato Early Blight	100.0%	98.5%	98.5%
Tomato Late Blight	94.6%	84.9%	92.1%
Tomato Septoria Leaf Spot	99.4%	100.0%	97.7%

Figure 8. Comparison chart of mAP@0.5–0.95; (**a**) YOLOv5s; (**b**) YOLOv5—LIteAttn.

3.3. Comparative Experiment

To further verify the performance advantages of the proposed YOLOv5s-LiteAttn model, this study compares it with various classic and mainstream object detection algorithms in Table 3. For a fair comparison, all models in Table 3 were trained from scratch under identical training settings for 100 epochs, without using any COCO or externally pre-trained weights.

Table 3. Comparison of results of different target detection models.

Model	Params/M	FLOPs/G	mAP@0.5–0.95/%	R/%	Size/MB	FPS
YOLOv5s	7.12	16.10	93.8	96.8	13.9	121.0
YOLOv7-tiny	6.12	13.40	89.2	91.2	47.1	105.0
YOLOX-s	8.95	26.84	91.6	91.6	68.6	70.0
YOLOv11-s	9.42	21.40	98.3	98.4	18.8	88.0
SSD-MobileNetV3 large	2.76	2.09	82.1	92.5	10.8	235.0
Faster R-CNN(R50-FPN)	41.55	182.54	73.8	85.3	158.0	12.0
EfficientDet-D0	3.85	7.79	70.4	86.2	15.1	195.0
YOLOv5s-LiteAttn	5.50	13.40	97.1	97.9	11.0	142.0

To ensure the scientificity and fairness of the comparison, lightweight detection models that are widely used and representative in agricultural scenarios, as well as some mainstream models with high accuracy, are selected, including YOLOv5s, YOLOv7-tiny, YOLOX-s, YOLOv11-s, SSD-MobileNetV3 [29,30] large, Faster RCNN (R50-FPN) [31], and EfficientDet-D0 [32]. These models have been widely applied and verified in crop pest and disease detection tasks, and can well reflect the actual effects of lightweight and high-precision models in the agricultural field, as shown in Figure 9.

As summarized in Table 3, YOLOv5s-LiteAttn achieves a mAP@0.5–0.95 of 97.1% and a recall of 97.9%, outperforming all lightweight competitors. Compared with YOLOv7-tiny (89.2% mAP) and YOLOX-s (91.6% mAP), the proposed model demonstrates substantially stronger discriminative ability for small and subtle lesion targets. Although YOLOv11-s attains the highest mAP (98.3%), it requires more parameters (9.42 M), higher computational cost (21.40 GFLOPs), and exhibits a slower inference speed (88 FPS). This indicates that YOLOv11-s prioritizes accuracy at the expense of deployment efficiency.

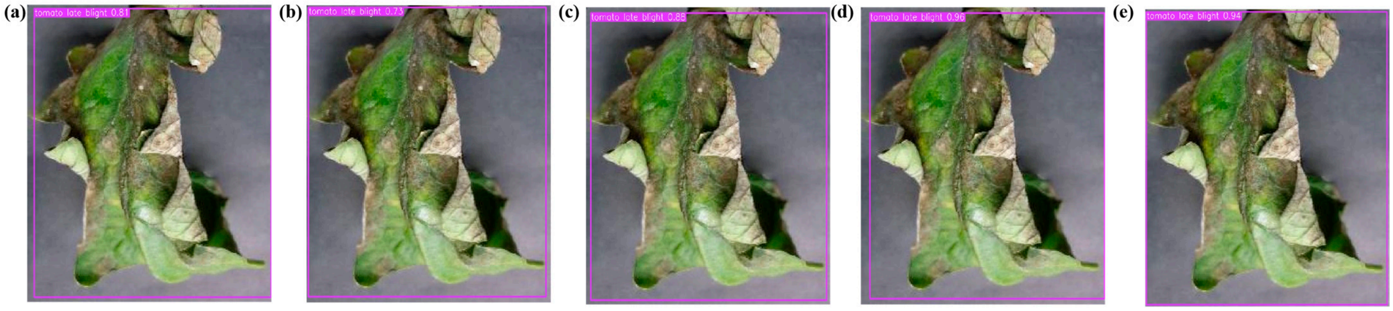


Figure 9. Comparison of visualization results from different models. (**a**) YOLOv5s; (**b**) YOLOv7-tiny; (**c**) YOLOX-s; (**d**) YOLOv11-s; (**e**) YOLOv5s-LiteAttn.

In contrast, YOLOv5s-LiteAttn achieves the best overall balance between accuracy and efficiency. With only 5.50 M parameters and 13.40 GFLOPs, the model delivers the fastest inference speed among all tested models (142 FPS). Notably, it improves detection accuracy over YOLOv5s by 3.3 percentage points while simultaneously reducing parameters by 22.75% and FLOPs by 16.77%. This demonstrates that the combination of lightweight convolution modules and the hybrid attention mechanism enhances feature representation without increasing computation burden.

Overall, the comparative experiments clearly show that YOLOv5s-LiteAttn offers the highest comprehensive performance across detection accuracy, computational efficiency, and real-time capability. Its ability to deliver near-YOLOv11-s accuracy at substantially lower computational cost underscores its strong deployment potential for field-based agricultural monitoring systems where both precision and speed are critical.

3.4. Per-Class Performance Analysis

To further assess class-level performance regarding per-category behavior, the confusion matrix and precision–recall (PR) curves of YOLOv5s-LiteAttn are presented in Figures 10 and 11.

The confusion matrix shows that most categories achieve very high true-positive rates (generally above 0.95), indicating strong class separability across diverse crops and diseases. Only a few visually similar categories—such as Apple Scab vs. Apple Black Rot and Potato Healthy vs. Potato Early Blight-exhibit minor confusion, which is expected due to overlapping lesion patterns or subtle early-stage symptoms.

The PR curves further confirm the model's robustness. The macro-average curve remains close to the upper boundary, and most categories maintain high precision even at large recall values. Lesion-rich categories such as Northern Leaf Blight, Strawberry Leaf Scorch, and Tomato Late Blight demonstrate particularly strong PR profiles, consistent with their notable per-class improvements. A few healthy-class categories display slightly lower recall, likely due to their greater similarity to early disease symptoms.

Overall, the per-class results demonstrate that YOLOv5s-LiteAttn achieves stable and reliable detection across most categories, with limited confusion among difficult classes. These findings support the effectiveness of the proposed lightweight and attention-enhanced structure in improving fine-grained feature discrimination and maintaining strong generalization in practical agricultural environments.

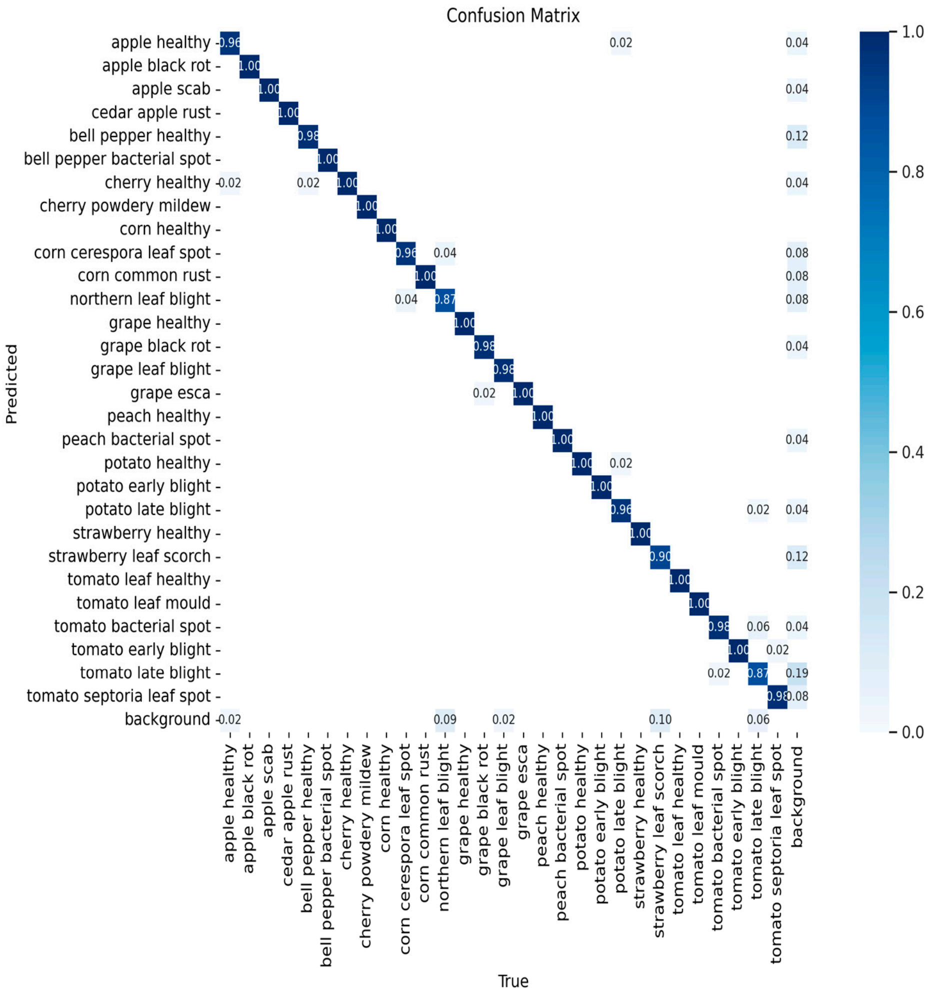


Figure 10. The confusion matrix.

3.5. Application Evaluation Based on an Independent Test Set

To further assess the cross-domain robustness and field applicability of the proposed model, an independent test set containing 4328 in-field images was constructed. All images were collected at the Modern Agricultural R&D Base of Sichuan Agricultural University (Chongzhou, Chengdu, Sichuan Province) between March and June 2025 using an iPhone 13 under natural lighting, occlusion, and background conditions. All images in the independent test set were annotated following the same protocol as the main dataset. Initial bounding boxes and category labels were produced by three trained annotators. Subsequently, a stratified sample of 600 images (approximately 14% of the set) was independently reviewed by two plant pathology experts. For cases with inconsistent labels, the annotators and experts jointly discussed and revised the annotations until consensus was reached, ensuring biological correctness and reducing systematic labeling bias. This independent dataset covers a wide range of real field scenarios and thus provides a rigorous evaluation of generalization performance.

As shown in Table 4, both YOLOv5s and YOLOv5s-LiteAttn experience performance drops compared with the validation results, which is expected due to the clear distribution shift between controlled dataset conditions and real in-field images. Nevertheless, YOLOv5s-LiteAttn maintains strong detection capability, achieving 95.8% mAP@0.5–0.95, 97.1% recall, and 140 FPS, all of which remain consistently superior to the baseline YOLOv5s (91.2% mAP@0.5–0.95, 95.9% recall, 118 FPS). The improved robustness demonstrates that the proposed lightweight structure and hybrid attention mechanism effectively mitigate background noise, illumination variation, and fine-grained lesion ambiguity present in actual field sampling environments.

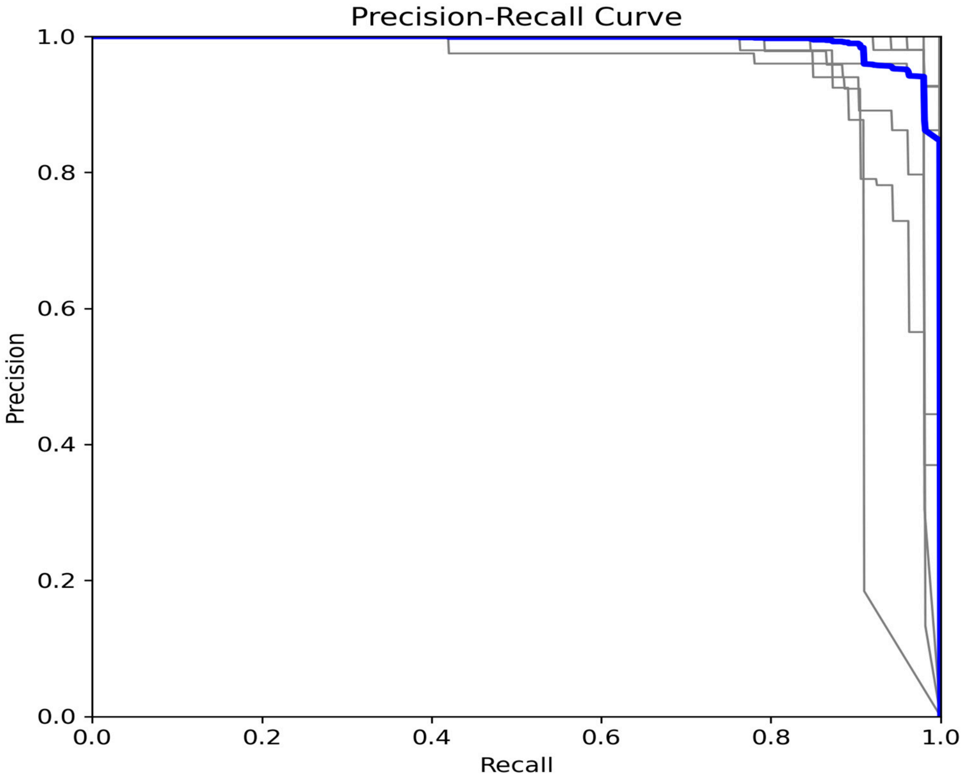


Figure 11. PR.

Table 4. Comparison of results of the baseline model.

Model	Params/M	FLOPs/G	mAP@0.5–0.95/%	R/%	FPS	Size/MB
YOLOv5s	7.12	16.10	91.2	95.9	118.0	13.9
YOLOv5s-LiteAttn	5.50	13.40	95.8	97.1	140.0	11.0

These results further confirm that YOLOv5s-LiteAttn is not only accurate and efficient during validation but also exhibits stable transferability when exposed to previously unseen real-world distributions. Therefore, the model has significant potential for practical deployment in intelligent agricultural monitoring systems and field-based pest and disease diagnosis.

Notably, despite the minor performance degradation, the YOLOv5s-LiteAttn model consistently outperformed the baseline YOLOv5s, highlighting the effectiveness of the proposed lightweight structure and hybrid attention mechanism in maintaining robustness and transferability under complex backgrounds and variable acquisition conditions. Therefore, the independent test results further substantiate the stability, generalization ability, and deployment potential of the YOLOv5s-LiteAttn model for real-world agricultural pest and disease detection scenarios.

4. Conclusions

The accurate identification of pests and diseases remains a critical requirement for modern precision agriculture. This study addressed the challenge of developing a detection model that balances performance with computational efficiency for practical deployment. An improved lightweight model, YOLOv5s-LiteAttn, was introduced based on the YOLOv5s architecture. The main contributions of this work were fourfold:

(1) A lightweight model architecture was constructed through the integration of the GhostConv and Depthwise Convolution modules. This design effectively reduced the model's computational complexity and parameter count.

(2) The model's superior trade-off between detection accuracy and operational efficiency was demonstrated through comparative analyses. The proposed model achieved enhanced performance in key metrics compared to the baseline and other mainstream lightweight models.

(3) The critical role of a hybrid attention mechanism in improving feature representation was elucidated. This integration was shown to enhance the model's robustness.

(4) To further verify generalization performance, an independent test set of 4328 newly collected field images was employed. Although the model exhibited a slight performance decline, its mAP@0.5–0.95 remained at 95.8%, still markedly outperforming the baseline, thus confirming the model's robustness and cross-domain adaptability.

Nevertheless, this study still has certain limitations. On the one hand, the dataset is mainly derived from a limited number of crop categories, so the cross-regional generalization ability of the model remains to be verified; on the other hand, this paper focuses on the optimization of algorithm performance and lacks in-depth research on the biological connections between different diseases and pests and their dynamic evolution in agricultural ecosystems.

Therefore, future research will mainly proceed in three directions: first, further expanding the dataset by incorporating more crop types and samples from complex environments to enhance the robustness and adaptability of the model in different scenarios [33]; second, exploring multimodal information fusion, such as integrating hyperspectral data, infrared imaging, and meteorological data, to improve the comprehensiveness and accuracy of pest and disease identification [34]; third, focusing on time-series behavior modeling to potentially introduce LSTM, 3D convolution, or Transformer modules to predict the dynamic changes and development trends of pests and diseases [35]. Through these expansions, the YOLOv5s-LiteAttn model proposed in this study is expected to become a core tool in agricultural intelligent monitoring and precise prevention and control in the future, promoting the development of pest and disease management towards automation, intelligence, and sustainability.

Author Contributions: Z.L.: conceptualization, data curation, formal analysis, methodology, software, writing—original draft, resources, and writing—review and editing. Y.L.: conceptualization, writing (review and editing), data curation, software, validation, and formal analysis. N.K.: conceptualization, writing (review and editing), investigation, validation, and software. L.L.: data curation, software, and validation. W.P.: methodology, resources, validation, and writing—review and editing. Y.Y.: investigation, software, and methodology. C.Q.: resources, formal analysis, and validation. X.L.: formal analysis, methodology, visualization, and validation. M.X.: investigation, software, and visualization. Y.Z. (Yining Zhang): writing (original draft), data curation, and resources. C.L.: investigation, resources, and validation. C.J.: data curation, visualization, validation, and writing—review and editing. M.L.: validation, writing (original draft), and writing (review and editing). Y.Z. (Yangxia Zheng): project administration, data curation, and methodology. Y.J.: conceptualization, funding acquisition, writing (review and editing), visualization, software, project administration, investigation, validation, resources, and data curation. W.L.: conceptualization, formal analysis, writing—review and editing, writing (original draft), validation, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Foreign Experts Program (H20240506); the Sichuan Haiju High-Level Talents Introduction Project (2025HJRC0048); the National Agricultural Science and Technology Innovation System Sichuan Characteristic Vegetable Innovation Team

Project (SCCXTD-2024-22); and the Key R&D Program Project of Xinjiang Province (grant number: 2023B02020).

Data Availability Statement: Data is contained within the article. The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

References

1. Liu, Y.-S.; Liu, Y.; Guo, L.-Y. Impact of climatic change on agricultural production and response strategies in China. *Chin. J. Eco-Agric.* **2010**, *18*, 905–910. [CrossRef]
2. Li, S.; Feng, Z.; Yang, B.; Li, H.; Liao, F.; Gao, Y.; Liu, S.; Tang, J.; Yao, Q. An intelligent monitoring system of diseases and pests on rice canopy. *Front. Plant Sci.* **2022**, *13*, 972286. [CrossRef] [PubMed]
3. Oerke, E.-C. Crop losses to pests. *J. Agric. Sci.* **2006**, *144*, 31–43. [CrossRef]
4. Li, R.; He, Y.; Li, Y.; Qin, W.; Abbas, A.; Ji, R.; Li, S.; Wu, Y.; Sun, X.; Yang, J. Identification of cotton pest and disease based on CFNet-VoV-GCSP-LSKNet-YOLOv8s: A new era of precision agriculture. *Front. Plant Sci.* **2024**, *15*, 1348402. [CrossRef]
5. Wang, X.; Zhang, S.; Zhang, T. Crop insect pest detection based on dilated multi-scale attention U-Net. *Plant Methods* **2024**, *20*, 34. [CrossRef]
6. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
7. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
8. Holland, J.H. Adaptation in natural and artificial systems. *Univ. Mich. Press Google Sch.* **1975**, *2*, 29–41. [CrossRef]
9. Chen, M.; Wang, J.; Chen, Y.; Guo, M.; Zheng, N. Weight-based ensemble method for crop pest identification. *Ecol. Inform.* **2024**, *82*, 102693. [CrossRef]
10. Guan, H.; Fu, C.; Zhang, G.; Li, K.; Wang, P.; Zhu, Z. A lightweight model for efficient identification of plant diseases and pests based on deep learning. *Front. Plant Sci.* **2023**, *14*, 1227011. [CrossRef]
11. Amrani, A.; Diepeveen, D.; Murray, D.; Jones, M.G.; Sohel, F. Multi-task learning model for agricultural pest detection from crop-plant imagery: A Bayesian approach. *Comput. Electron. Agric.* **2024**, *218*, 108719. [CrossRef]
12. Ye, Y.; Chen, Y.; Xiong, S. Field detection of pests based on adaptive feature fusion and evolutionary neural architecture search. *Comput. Electron. Agric.* **2024**, *221*, 108936. [CrossRef]
13. Zhang, H.; Zhou, Y.; Wang, K.; Wang, C.; Li, H. Fruit Tree Pest Identification Method Based on MobileViT-PC-ASPP and Transfer Learning. *Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach.* **2024**, *55*, 1–14.
14. Liu, X.; Mao, T.; Shi, Y.; Ren, Y. Overview of knowledge reasoning for knowledge graph. *Neurocomputing* **2024**, *585*, 127571. [CrossRef]
15. Wang, Z.Y.; Yu, Q.; Wang, N.; Wang, Y.G. A review of intelligent question answering research based on knowledge graphs. *J. Comput. Eng. Appl.* **2020**, *56*, 1–11. [CrossRef]
16. Wang, X.; Huang, Z.; Zhang, S.; Zhu, J.; Gamba, P.; Feng, L. GMSR: Gradient-Integrated Mamba for Spectral Reconstruction from RGB Images. *Neural Netw.* **2025**, *193*, 108020. [CrossRef]
17. Li, X.; Li, S. Transformer help CNN see better: A lightweight hybrid apple disease identification model based on transformers. *Agriculture* **2022**, *12*, 884. [CrossRef]
18. Ma, L.; Hu, Y.; Meng, Y.; Li, Z.; Chen, G. Multi-plant disease identification based on lightweight ResNet18 model. *Agronomy* **2023**, *13*, 2702. [CrossRef]
19. Jocher, G. *Ultralytics/Yolov5: V3.1—Bug Fixes and Performance Improvements, v3.1*; Zenodo: Geneva, Switzerland, 2020.
20. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features From Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589. [CrossRef]
21. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
22. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
24. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.

25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
26. Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad, India, 5–7 January 2020; pp. 249–253.
27. Lv, M.; Su, W.-H. YOLOV5-CBAM-C3TR: An optimized model based on transformer module and attention mechanism for apple leaf disease detection. *Front. Plant Sci.* **2024**, *14*, 1323301. [CrossRef]
28. Wang, X.; Liu, J. Vegetable disease detection using an improved YOLOv8 algorithm in the greenhouse plant environment. *Sci. Rep.* **2024**, *14*, 4261. [CrossRef]
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
32. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
33. Li, W.; Zhou, B.; Zhou, Y.; Jiang, C.; Ruan, M.; Ke, T.; Wang, H.; Lv, C. Grape Disease Detection Using Transformer-Based Integration of Vision and Environmental Sensing. *Agronomy* **2025**, *15*, 831. [CrossRef]
34. Cao, Z.; Sun, S.; Bao, X. A Review of Computer Vision and Deep Learning Applications in Crop Growth Management. *Appl. Sci.* **2025**, *15*, 8438. [CrossRef]
35. Taha, M.F.; Mao, H.; Zhang, Z.; Elmasry, G.; Awad, M.A.; Abdalla, A.; Mousa, S.; Elwakeel, A.E.; Elsherbiny, O. Emerging technologies for precision crop management towards agriculture 5.0: A comprehensive overview. *Agriculture* **2025**, *15*, 582. [CrossRef]

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