Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065152

ARTICLE

SPD-YOLO: A Method for Detecting Maize Disease Pests Using Improved YOLOv7

Zhunruo Feng[1], Ruomeng Shi[2], Yuhan Jiang[3], Yiming Han[1], Zeyang Ma[1] and Yuheng Ren[4,*]

[1]School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China
[2]International Business School Suzhou, Xi'an Jiaotong Liverpool University, Suzhou, 215123, China
[3]School of the Arts, Universiti Sains Malaysia, Penang, 11700, Malaysia
[4]School of Digital Industry, Jimei University, Xiamen, 361021, China

*Corresponding Author: Yuheng Ren. Email: szcyxy@jmu.edu.cn

Received: 05 March 2025; Accepted: 19 May 2025; Published: 03 July 2025

ABSTRACT: In this study, we propose Space-to-Depth and You Only Look Once Version 7 (SPD-YOLOv7), an accurate and efficient method for detecting pests in maize crops, addressing challenges such as small pest sizes, blurred images, low resolution, and significant species variation across different growth stages. To improve the model's ability to generalize and its robustness, we incorporate target background analysis, data augmentation, and processing techniques like Gaussian noise and brightness adjustment. In target detection, increasing the depth of the neural network can lead to the loss of small target information. To overcome this, we introduce the Space-to-Depth Convolution (SPD-Conv) module into the SPD-YOLOv7 framework, replacing certain convolutional layers in the traditional system backbone and head network. This modification helps retain small target features and location information. Additionally, the Efficient Layer Aggregation Network-Wide (ELAN-W) module is combined with the Convolutional Block Attention Module (CBAM) attention mechanism to extract more efficient features. Experimental results show that the enhanced YOLOv7 model achieves an accuracy of 98.38%, with an average accuracy of 99.4%, outperforming the original YOLOv7 model. These improvements represent an increase of 2.46% in accuracy and 3.19% in average accuracy. The results indicate that the enhanced YOLOv7 model is more efficient and real-time, offering valuable insights for maize pest control.

KEYWORDS: Deep learning; improved YOLOv7; attention mechanism; SPD-Conv module; insect pest detection

1 Introduction

As a leading agricultural country, China regards maize as one of its most crucial food crops, with its production having a direct impact on both the national economy and public welfare [1]. Pests, including maize aphids, grass armyworms, pelagic bugs, rice locusts, and Lucifer bifasciata, pose significant threats to maize food security. Research indicates that, without effective pest control measures, corn yield losses due to armyworms can reach as high as 48.35%, with ear damage rates reaching up to 98.91%. Infestations by Laminaria biculata can result in yield reductions of 10% to 30% [2]. Traditional pest monitoring for maize heavily relies on the agricultural expertise and experience of professionals and growers. While this method can be effective on small-scale farms, it is inefficient, inaccurate, and overly dependent on the experience of personnel.

The swift progress in AI, machine learning, and deep learning has notably enhanced the ability to detect plant pests. Research has demonstrated the potential of these technologies for more accurate pest identification. Xiang et al. [3] utilized image feature extraction, segmentation, and classifier design.while Venkateswara et al. [4] proposed an algorithm that integrates deep learning for pest monitoring and classification, effectively addressing the issue of data imbalance. Deep learning techniques have notably improved model accuracy and generalization in pest detection [5], which extract multi-scale features from large datasets. However, challenges, such as small pest sizes, high feature similarity, and low image resolution, still complicate detection in real-world agricultural settings [6,7].

Despite these advancements, maize pest detection remains challenging due to issues such as small pest sizes, high feature similarity, and low image resolution. To overcome these challenges, this study proposes an enhanced YOLOv7 algorithm. A comprehensive dataset of common maize pests was constructed based on data collected from the Xian Intelligent Agriculture Industrial Park in Northern Wilderness. To enhance detection accuracy, we integrated the SPD-Conv module in place of conventional convolution layers, minimizing the loss of essential details and boosting the identification of small pests, we integrated the hybrid attention mechanism CBAM into the ELAN-W module, allowing the enhanced model to more effectively capture and refine feature information across both channel and spatial domains. These optimizations significantly improve detection accuracy, even in complex and cluttered environments. The proposed model provides a robust and efficient solution for detecting maize pests in challenging natural settings. It offers a valuable technical contribution to pest monitoring and control, ultimately supporting more effective pest management practices in agriculture.

The paper is structured as follows: Section 2 provides a review of related studies in this area. In Section 3, the proposed algorithm improvements are discussed in detail, highlighting key advances. Section 4 introduces experimental parameters and experimental environment. In Section 5, experiments are carried out and the results are comprehensively analyzed. In Section 6, we briefly summarize the algorithms introduced in this study.

2 Related Works

The You Only Look Once (YOLO) framework was developed by Redmon et al. [8]. This approach made detection faster than methods like R-CNN [9]. YOLO divides the input image into a grid, predicting bounding boxes and class probabilities for each cell. YOLOv2 [10] and YOLOv3 [11] enhanced accuracy with multi-scale detection, anchor boxes, and a stronger backbone network for better small object detection. YOLOv4 [12] introduced the Cross Stage Partial Darknet-53 (CSP-Darknet53) backbone and advanced training techniques, improving speed and accuracy for real-time applications. YOLOv5 [13] further optimized speed and accuracy with a PyTorch implementation, gaining rapid industry adoption, which is a community-driven release. The YOLOv7 [14] iteration offers even faster inference and outstanding performance, making it highly effective for real-time tasks like pest detection in agriculture, where both speed and accuracy are essential.

The YOLOv7 architecture is composed of three core components, illustrated in Fig. 1. The input layer handles images with dimensions of $640 \times 640 \times 3$, which are first preprocessed before being passed through the backbone network. This backbone network is built on the foundation of the YOLOv5 architecture and is enhanced with several key elements, including the Efficient Layer Aggregation Networks (ELAN) structure, the Max Pooling 1 (MP1) structure, and the CBS module. These components collaborate to efficiently extract vital features from the input images, ensuring that important visual information is captured for the subsequent stages of the model.

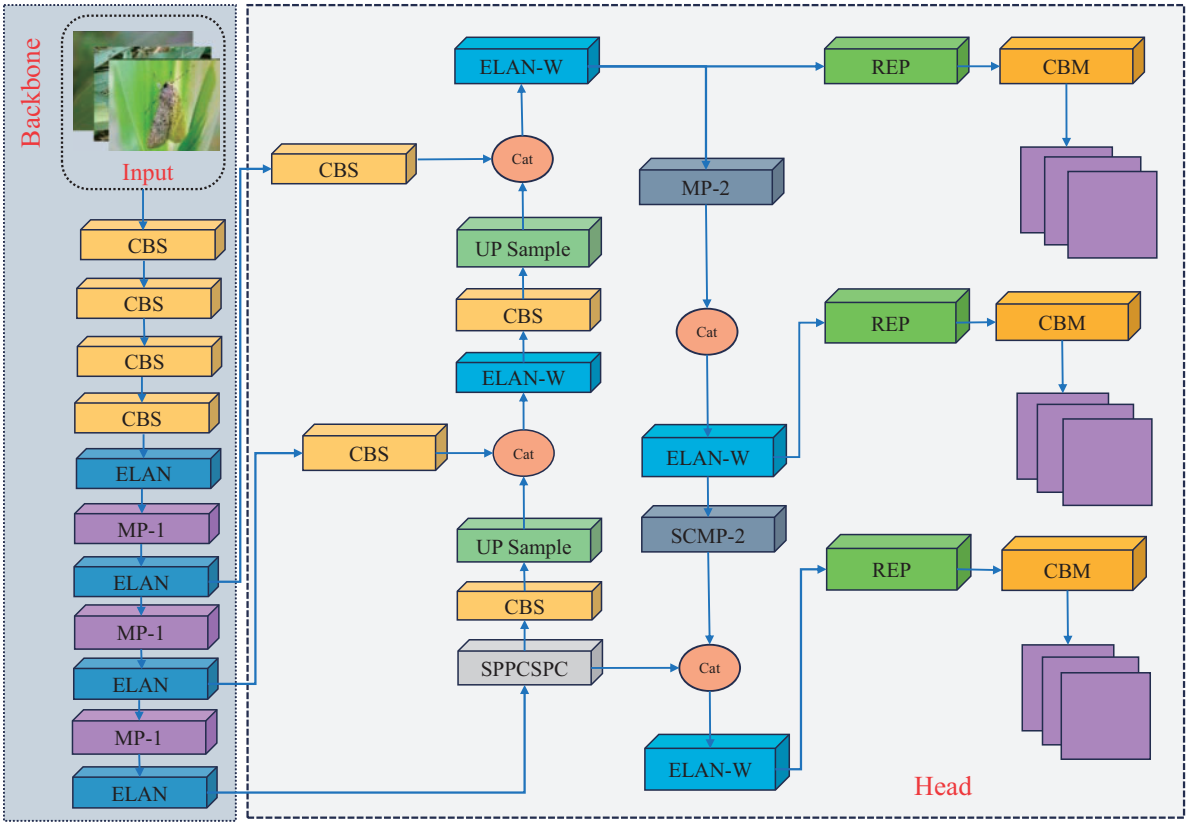


Figure 1: The YOLOv7 network structure

The detection head, which generates the final predictions, incorporates advanced structures like Spatial Pyramid Pooling and Cross Stage Partial Channel. These advanced structures enhance the model's ability to capture multi-scale features, contributing to better detection performance across various object sizes. In addition, the detection head incorporates further feature extraction components like Efficient Layer Aggregation Network-High (ELAN-H), Max Pooling ×2 (MP2), and Re-parameterizable Convolution (RepConv) layers, all of which support the refinement and fusion of feature maps generated by the backbone. This fusion process enables multi-scale target detection, allowing the model to make predictions at three distinct scales. As a result, YOLOv7 significantly improves detection accuracy, particularly for objects of different sizes, in complex visual environments.

The outputs from the various backbone layers are refined and fused in the detection head, allowing for multi-scale target detection. This fusion process enables the model to generate predictions at three distinct scales, optimizing detection accuracy for different sizes of objects [15].

3 Methods

This section mainly introduces the relevant improvement algorithm of the paper, including the Space to Depth module, Convolutional Block Attention Module (CBAM) attention mechanism, and improved YOLOv7 model.

3.1 Space to Depth Module

The SPD-Conv module consists of two main convolution operations: space-to-depth and non-strided convolution layers [16]. It serves as a replacement for traditional step convolutions, mitigating the loss of detailed information typically encountered in small object detection due to the limited pixel representation. By utilizing SPD-Conv, the accuracy of small object detection is significantly enhanced, and more detailed information is preserved. Fig. 2 shows the operation flowchart of SPD-Conv.

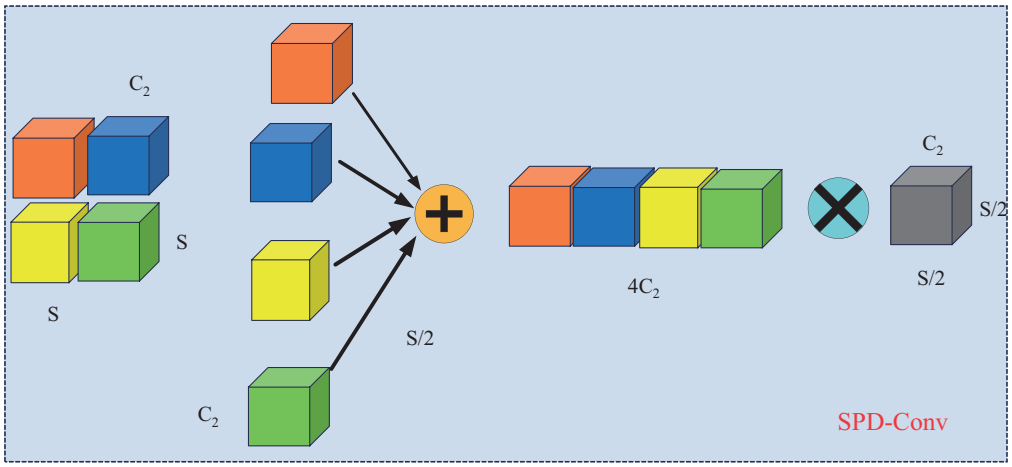


Figure 2: The SPD-Conv structure model. Strided convolution and pooling layers are commonly used in CNN architectures. However, they often lead to the loss of fine-grained details and inefficient feature learning. To overcome this limitation, SPDGConv (Space to Depth) has been proposed as a novel CNN block that replaces the traditional strided convolution and pooling layers. This approach significantly improves the network's performance on low-resolution images and enhances the detection of small targets, as illustrated in Eq. (1).

$$X\left(S \times S \times C_1\right) \rightarrow X_1\left(\frac{S}{N} \times \frac{S}{N} N^2 C_1\right) \tag{1}$$

where X denotes any intermediate feature map of size $(S \times S \times C_1)$. Typically, a feature map X can be partitioned into sub-feature maps, each divisible by N. Consequently, each sub-feature map is downsampled by a factor of N, as illustrated in Fig. 2. The four sub-feature maps are generated. They are then concatenated along the channel dimension to produce the feature map X_1, which exhibits a reduced spatial dimension and an increased number of channels. After the SPD feature transformation layer, if $N^2 C_1 > C_2$, the next layer feature map X_2 in Eq. (2) is further derived.

$$X_1\left(\frac{S}{N} \times \frac{S}{N} \times N^2 C_1\right) \rightarrow X_2\left(\frac{S}{N} \times \frac{S}{N} \times C_2\right) \tag{2}$$

where C_2 represents the non-strided convolutional layer of the filter. Among them. This operation has the advantage of down sampling the feature map while retaining the distinguishing feature information.

In this study, to tackle the issue of detail loss caused by step convolution, modifications are made to the MP-1 and MP-2 modules in both the Backbone and Detection Head. Specifically, the CBS convolution following the max pooling layer is replaced with the SPD-Conv module, as shown in Fig. 3. This change aims to preserve more intricate details by avoiding the information loss typically associated with step convolutions. The third CBS convolution at the input side of the Backbone is also substituted with the SPD-Conv module, further enhancing the model's ability to retain crucial features.

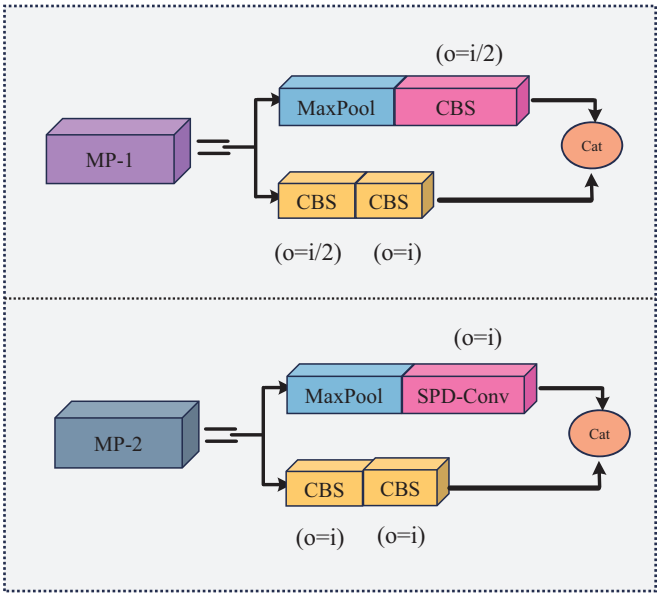


Figure 3: Improved front MP1 and MP2 modules

These adjustments significantly improve the model's performance, particularly in scenarios involving small object detection, where retaining fine-grained details is essential for accurate predictions. The SPD-Conv module is designed to reduce detail loss. It also improves the network's overall efficacy in various tasks. This ensures that important features are preserved during feature extraction. Fig. 4 shows the structure of the improved module and the enhancements made to the model architecture.

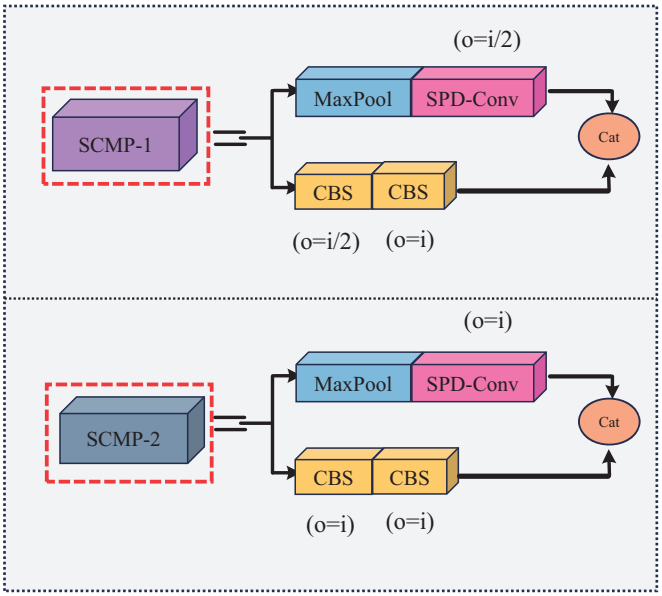


Figure 4: Improved SCMP-1 and SCMP-2 modules

3.2 CBAM Module

Fig. 5a shows a lightweight and powerful attention mechanism for efficient feature refinement. It includes two main components. Upon receiving the intermediate feature map, CBAM independently processes both the channel and spatial dimensions. It improves the overall feature representation. After processing, the resulting attention maps are fused with the original feature map, enabling the model to adaptively refine the features based on both channel-wise and spatial-wise importance. This fusion process allows for more effective feature optimization, leading to better performance in various tasks.

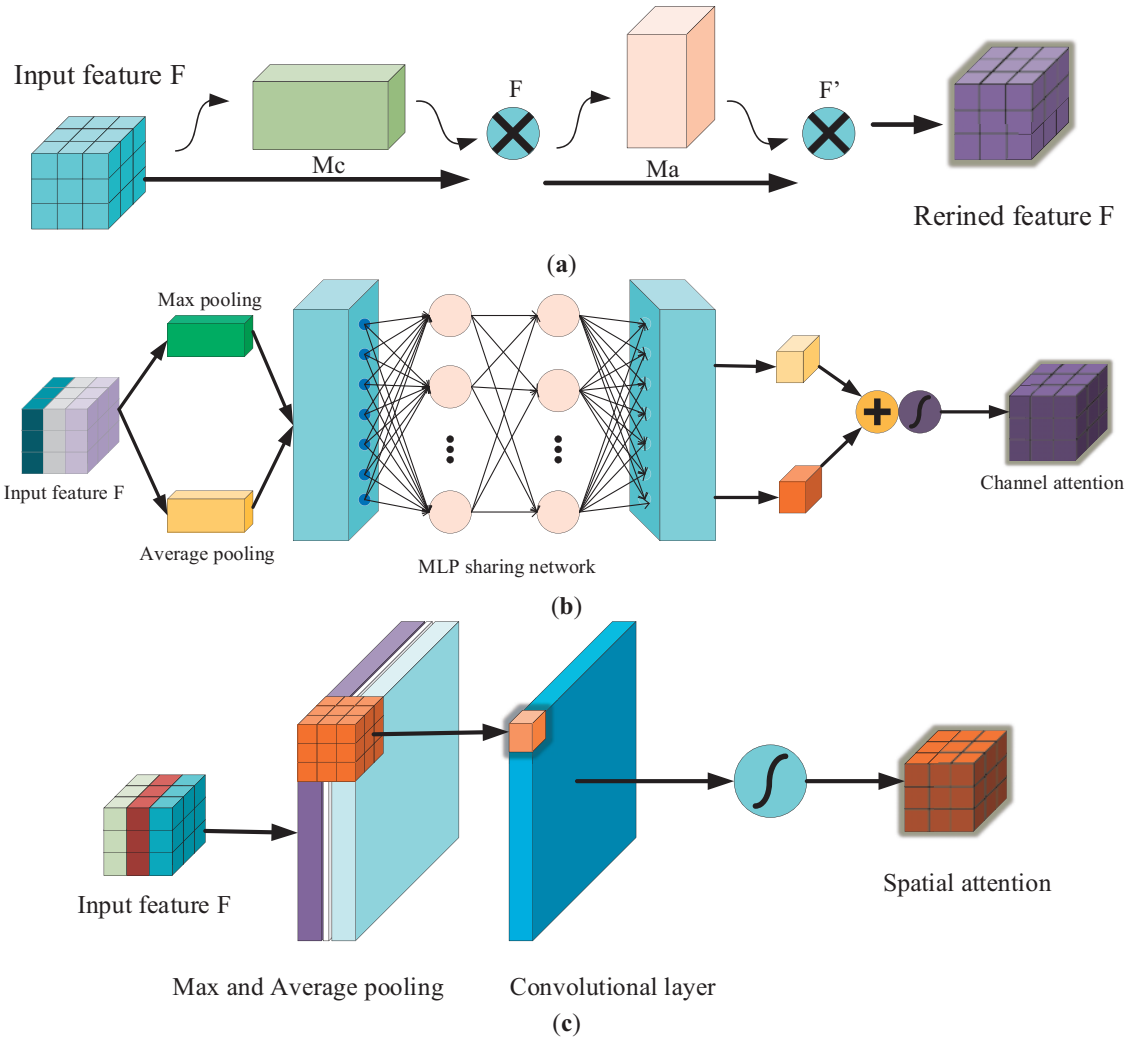


Figure 5: The structural framework of CBAM. (**a**) CBMM overall structure. (**b**) Channel attention mechanism overview. (**c**) Overview of the spatial attention mechanism

Fig. 5b shows the channel attention mechanism. It works by applying global average pooling (GAP) and GMP to the input feature map. These pooled feature maps are then merged and processed using a shared multi-layer perception (MLP) for ascending and descending operations. The final channel attention map is generated through an activation function. This approach allows CBAM to prioritize important features.

Fig. 5c shows the spatial attention mechanism. The input feature map undergoes GAP and GMP, creating two separate channel feature representations that capture both global and local contextual information. These two feature maps are then concatenated, combining the complementary information from each pooling method. The concatenated feature map is passed through a 7 × 7 convolutional layer to refine the features. This helps the model focus on the most relevant features, improving its ability to prioritize key information and boosting overall performance.

This study modifies the ELAN-W module to improve the detection of small target pests in complex environments. Specifically, the final convolutional layer is replaced with a CBAM attention mechanism, as illustrated in Fig. 6. This modification enables more efficient extraction of key local detail features of pests, significantly improving the learning and feature representation capabilities of the neural network. As a result, the model's accuracy in detecting small target pests is enhanced, thereby increasing its robustness and performance in practical applications.

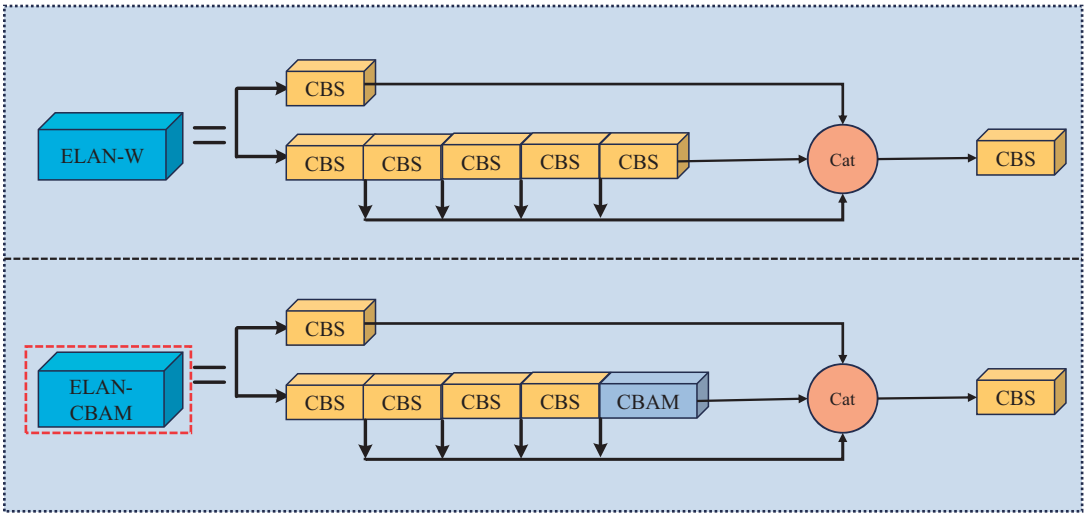


Figure 6: Before and after the improvement of ELAN-W3.3 SPD-YOLOV7 model

As the convolutional neural network's depth grows, the probability of losing fine details about small target pests increases, primarily due to the convolutional stride and pooling operations. To solve this, we enhance the MP-1 and MP-2 in the Backbone and Head of YOLOv7 by replacing the CBS module with the SPD-Conv module. This enhancement significantly boosts the detection accuracy of small target pests. Furthermore, to enhance the feature extraction and learning capabilities for maize crop pests, the final convolutional layer of the ELAN-W module is substituted with a CBAM attention mechanism. The model adjusts attention in both the channel and spatial dimensions. This highlights key features like color, shape, and texture of the pests while suppressing redundant features. It ensures precise extraction of relevant information. Fig. 7 shows the improved SPD-YOLOv7 model. The modified parts are highlighted in the green dotted box. These improvements make the network more suitable for target detection in complex environments.

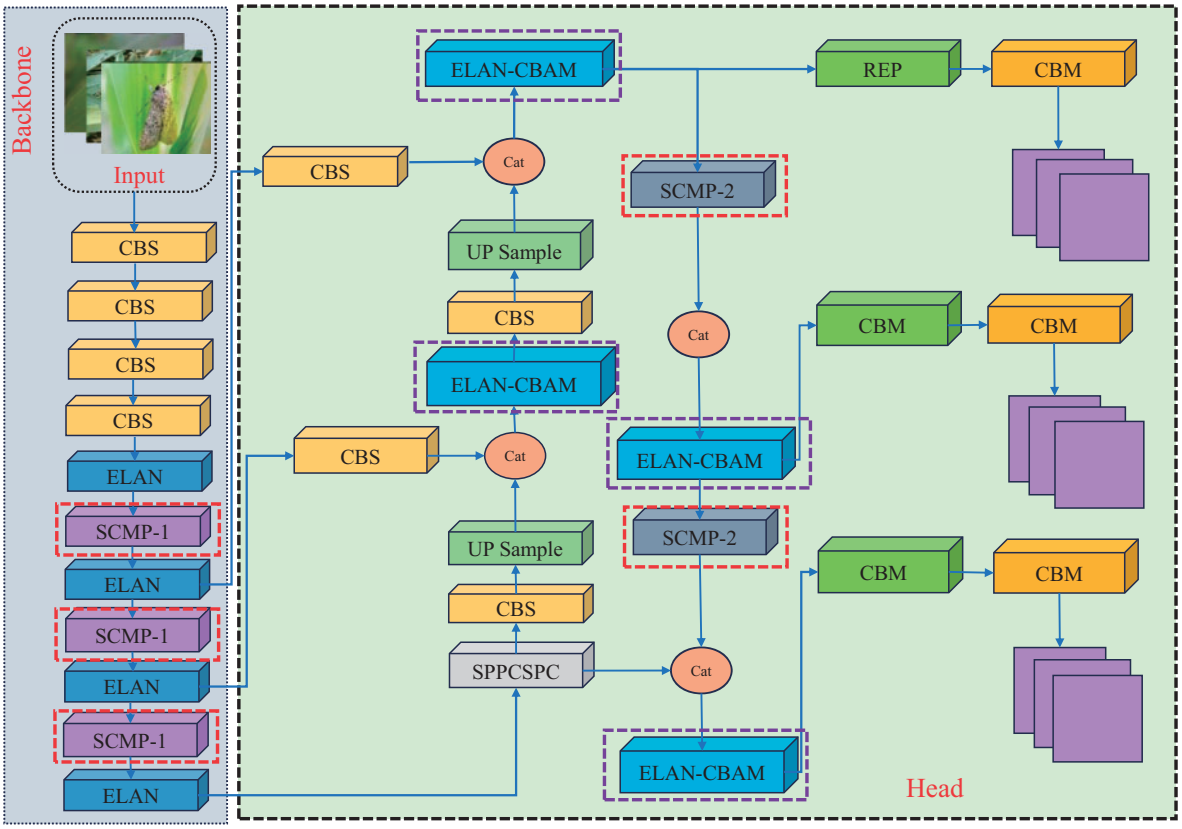


Figure 7: The SPD-YOLOv7 structure diagram was improved

4 Experimental Setup

This section describes the small object datasets utilized in the study, providing details on their implementation and key evaluation metrics. Additionally, it covers the experimental parameters and the setup of the experimental environment.

4.1 Experimental Data Set

The experimental data for this study were collected from the Xian Intelligent Agriculture Industrial Park, located in the northern wilderness of Heilongjiang Province [17]. To ensure the authenticity and generalizability of the data, approximately 1 mu (0.0667 hectares) of experimental field was selected for data collection. The imaging equipment used included a Canon D600 digital camera, equipped with a 35–135 mm medium telephoto lens and a 100 mm macro lens for capturing close-up images of corn crop pests. Data collection took place between mid-June and early August 2023, focusing on three primary maize pests: locusts, armyworms, and Lucifer bimaculate. During the data screening process, images of poor quality—due to issues such as blurring or overexposure—were excluded, resulting in a dataset of 1340 images, with approximately 400 samples per pest species. Fig. 8 displays sample images, with rows 1 to 3 illustrating examples of locusts, armyworms, and firefly beetles. These carefully collected and curated data enable accurate identification and analysis of maize pests.

Figure 8: Maize pest dataset

4.2 Data Processing

To reduce training time and improve efficiency, this study applies image compression and normalization techniques to ensure compatibility with the video memory and computational resources required by the YOLOv7 network model. The original resolution of the maize pest dataset was 5472 × 3648 pixels, which was uniformly compressed to 1080 × 720 pixels. For data annotation, the Labeling tool was used to label the three pest species, with annotations in the VOC format, saved in XML files. To ensure compatibility with the YOLO model, Python scripts were employed to convert the annotations into the YOLO format and save them in TXT files.

Efforts were made to balance the sample sizes for each maize pest species. However, discrepancies remained across different growth stages. For instance, only about 100 adult armyworm samples were collected, while 380 larvae samples were available. To simulate the real-world growing conditions of maize pests and address the issue of imbalanced data that could degrade model performance, additional data processing was performed. These included adding Gaussian noise and applying dark and light adjustments to simulate varying lighting conditions. These augmentations increased the total number of pest images to 3000. Finally, it was divided into training, validation, and test sets with an 8:1:1 ratio. Fig. 9 shows sample images after augmentation.

Figure 9: Data enhance example

4.3 Experimental Parameter Setting

4.3.1 Experimental Environment and Hyperparameter Setting

The experiments were carried out on a system running Windows 11. The hardware setup includes an AMD Ryzen 7 7735H processor with Radeon Graphics and an RTX 4060 GPU with 8 GB of memory. PyTorch was used as the deep learning framework, and Python 3.8.18 was the programming language. The development environment was PyCharm 2025. Table 1 provides the hyperparameter configurations for the experimental training.

Table 1: Experimental hyperparameter settings

Argument	Parameter value
Learning rate	0.01
Momentum	0.937
Weight	0.0005
Iteration cycle	300
Lot size	16
Image size	640 × 640
Learning rate attenuation method	Cosine annealing algorithm

4.4 Evaluation Index of Model Training

This study uses standard evaluation metrics for target detection. These include precision (P), recall (R), mean average precision (mAP), detection speed (FPS), and the loss function. These metrics provide a comprehensive assessment of the model's performance, evaluating accuracy, recall, multi-class average precision, and real-time detection speed in target detection tasks. While additional cross-validation was not performed in this work, we recognize that further validation through cross-validation or other robustness measures would enhance the confidence in the reported performance across different contexts and datasets. We acknowledge the limitations of the current approach in the manuscript and believe the selected metrics offer strong evidence of the model's generalizability. Future work will aim to include these additional validation measures.

Precision (P) is the ratio of correctly identified objects to the total number of identified objects, both correct and incorrect. The mathematical expression is shown in Eq. (3).

$$P = \frac{TP}{TP + FP} \tag{3}$$

here, TP is the number of corn pest images correctly detected by the model. FP (False Positive) is the number of pest images incorrectly identified by the model. Recall (R) is the ratio of correctly detected targets to the total number of actual targets, including the missed ones. The mathematical expression for recall is given in Eq. (4).

$$R = \frac{TP}{TP + FN} \tag{4}$$

here, FN denotes the number of pest images on corn crops that the model failed to detect. Mean Average Precision (mAP) is computed by evaluating the accuracy at different recall thresholds and averaging these accuracies to assess the model's performance across various recall levels. It is a widely used metric for

objectively assessing the performance of target detection models. The mathematical formula for mAP is provided in Eq. (5).

$$mAP = \frac{1}{N}\sum_{k=i}^{N} AP\,(k) \tag{5}$$

here, N represents the total number of maize pest classes, while K is the threshold value. $AP\ (k)$ refers to the Average Precision (AP) for the detected pest class k. FPS measures the number of frames the model processes or detects per second, reflecting the model's detection speed. To evaluate the stability of the model's performance, classification loss is used in this study. This loss represents binary classification, typically calculated using the sigmoid or SoftMax functions. The formula for this loss is provided in Eq. (6).

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c \log s_i^c \tag{6}$$

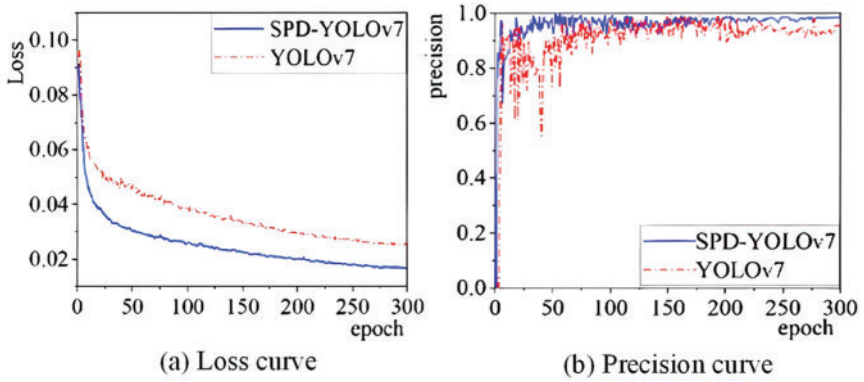
where N represents the total number of image samples collected. g denotes the number of pest individuals correctly detected. C indicates the number of corn crop pest images that were incorrectly detected.

5 Results and Analysis of the Experiment

This section presents the experiment results, including visual and data analysis. It also provides a comparison of the experimental outcomes, focusing on both subjective and objective evaluations.

5.1 Comparative Analysis of Experiments before and after Improvement

The training results before and after the improvements are shown in Fig. 10. In Fig. 10a, with training parameters held constant for both the YOLOv7 and SPD-YOLOv7 models, the loss values for both models decrease steadily as the number of training epochs increases, indicating the absence of overfitting. Both models reached optimal performance after 300 epochs. At this stage, the loss value of the SPD-YOLOv7 model was 0.008. Moreover, the convergence of confidence loss for the SPD-YOLOv7 model was faster and consistently lower compared to YOLOv7.



(a) Loss curve (b) Precision curve

Figure 10: (Continued)

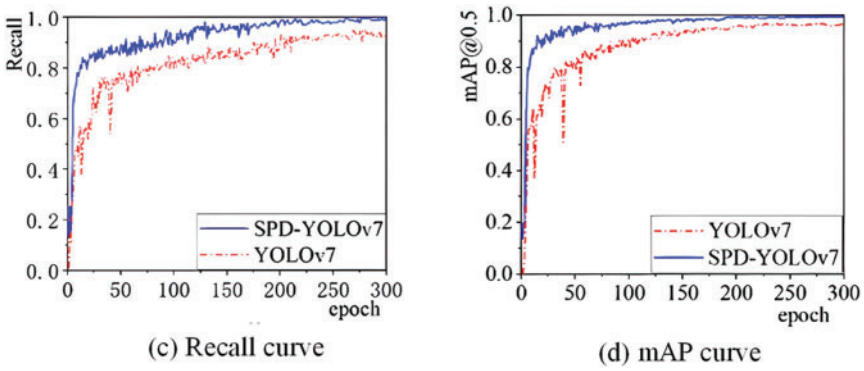


Figure 10: Experimental results graph

The improved SPD-YOLOv7 model achieved a 2.9 percentage point increase in accuracy, a 6.8 percentage point increase in recall, and a 3.19 percentage point increase in mAP. The training curve became more stable as well. The specific details are shown in Fig. 10b–d. The loss value serves as a key indicator of training quality; a lower loss indicates a smaller gap between predicted and actual bounding boxes, reflecting improved detection performance. While these results highlight the effectiveness of the proposed improvements, further validation, such as cross-validation or other robustness measures, would be beneficial to confirm the generalizability of the reported performance across different datasets and real-world scenarios.

Fig. 11 displays the confusion matrix, which allows for a visual examination of the classification outcomes for each category. In this matrix, each row corresponds to the predicted category, each column represents the actual category, and the diagonal elements show the proportion of correct classifications. An analysis of the results in the matrix shows that the majority of the targets are accurately predicted, suggesting that the model performs effectively.

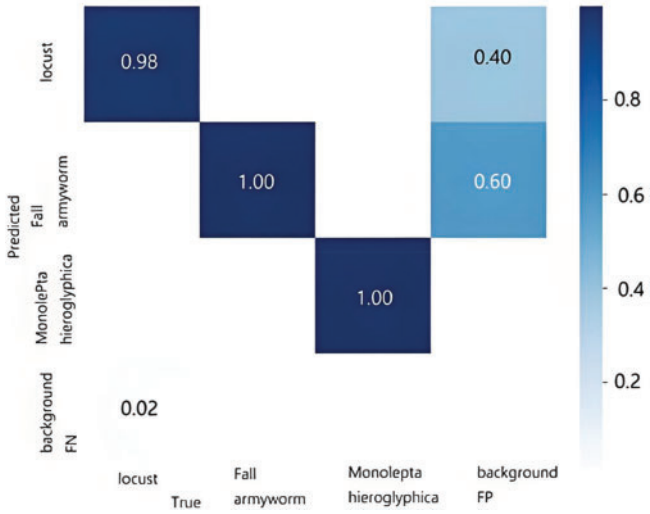


Figure 11: Confusion matrix diagram

5.2 Comparison of Test Results

YOLOv7 and SPD-YOLOv7 were tested on the same set of pest images to visually evaluate their detection performance. Fig. 12 illustrates that the improved SPD-YOLOv7 model greatly reduced false and missed detections compared to the original YOLOv7 model. The SPD-YOLOv7 model achieved higher detection accuracy than the original model. Experimental results confirm that the enhanced SPD-YOLOv7 model offers superior detection performance and improved detection capabilities.

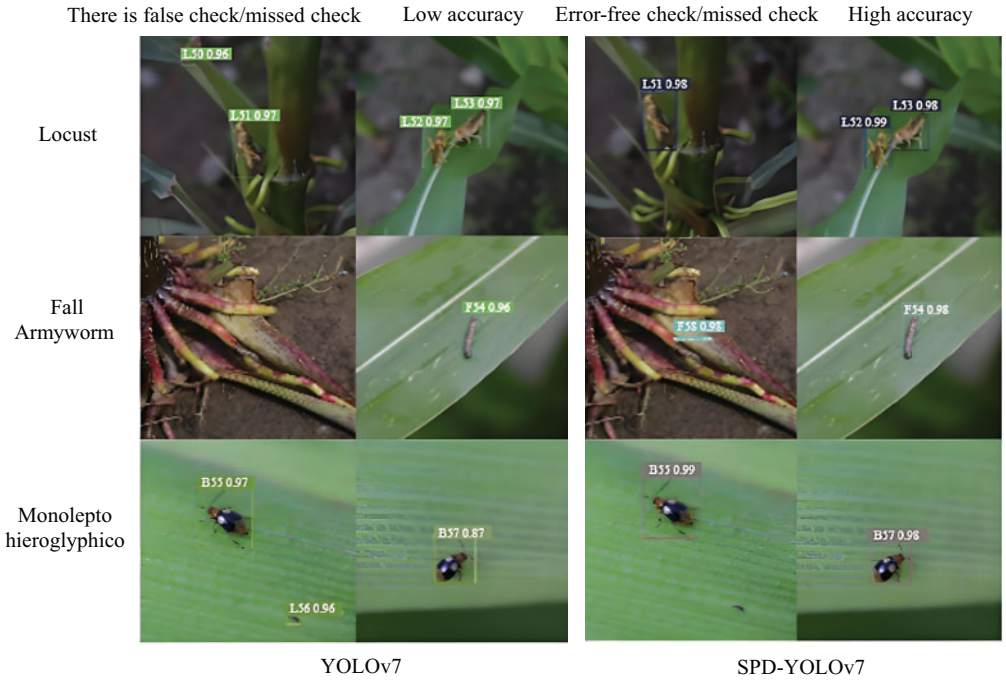


Figure 12: Test effect picture

5.3 Analysis of Ablation Results

Ablation experiments were conducted to evaluate the impact of each improvement on the maize pest detection model and validate the effectiveness of the enhancements. This study designed four sets of ablation experiments with different model configurations, and the results are presented in Table 2.

Table 2: Comparison of ablation performance

Test No.	Model	P	R	mAP	FPS
1	YOLOv7	95.92	92.27	96.21	71.7
2	YOLOv7-SPD	98.50	94.32	97.89	68.3
3	YOLOv7-CBAM	96.41	98.23	99.20	72.0
4	TOLOv7-SPD-CBAM	98.38	99.51	99.40	69.0

As presented in the data in Table 2. Experiment 1 used the original YOLOv7 model. It performed well in maize pest detection with an accuracy of 95.92%, a recall rate of 92.27%, a mAP of 96.21%, and an FPS of 71.7. In Experiment 2, the SPD-Conv modules were added to both the Backbone and Head of the

original model. The results indicated that, compared to Experiment 1, the mAP increased by 1.68 percentage points, reaching 98.50%, and accuracy improved by 2.58 percentage points. These enhancements suggest that the SPD-Conv module effectively reduces detail loss and improves feature extraction for small target objects. While these results underscore the effectiveness of the proposed improvements, further validation through cross-validation or other robustness measures would be beneficial to confirm the consistency and generalizability of the reported performance across varied datasets and scenarios.

In Experiment 3, the original model was enhanced by incorporating the CBAM attention mechanism.The results showed further improvement, with the mAP increasing by 2.99 percentage points to 99.20%, and the recall rate increasing by 3.91 percentage points compared to Experiment 1. This shows that the CBAM attention mechanism effectively focuses attention on both channel and spatial dimensions, enhancing recall and detection accuracy for small target pests.

Experiment 4 combined the improvements from Experiments 2 and 3. While the detection speed was slightly reduced, the accuracy, recall rate, and mAP increased by 2.46, 7.24, and 3.19 percentage points, respectively. The SPD-YOLOv7 model proposed in this study achieved the best overall performance, with an accuracy of 98.38%, a recall rate of 99.51%, a mAP of 99.40%, and an FPS of 69.5. The results from the ablation experiment confirm that the proposed improvements contribute significantly to the performance enhancement of the improvement in the performance of the YOLOv7 target detection model.

5.4 Comparison Experiment of Mainstream Target Detection Algorithms

The improved YOLOv7 algorithm exceeds Faster R-CNN, YOLOv3, YOLOv4, and YOLOv5 in accuracy, recall rate, and average precision in Table 3. Although the improved YOLOv7 exhibits a slightly lower detection speed compared to YOLOv5, it still demonstrates superior overall performance. Specifically, the improved YOLOv7 achieves a 3.78–9.41 percentage point increase in accuracy, a 9.32–17.67 percentage point improvement in recall rate, and a 4.15–8.56 percentage point gain in average precision compared to other mainstream models. These results further highlight the enhanced target detection capabilities of the improved YOLOv7 algorithm.

Table 3: Comparison of results of mainstream target detection algorithms

Model	P	R	mAP	FPS
Faster-RCNN	93.48	83.89	91.07	46
YOLOv3	88.97	81.84	92.33	59
YOLOv4	92.06	84.14	90.84	65
YOLOv5	94.60	90.19	95.25	74
SPD-YOLOv7	98.38	99.51	99.40	69.0

5.5 Model Practicality Analysis

To assess the practicality of the improved YOLOv7 algorithm, we first compared the number of parameters of the improved YOLOv7 with that of other algorithms. This comparison helps evaluate the model's efficiency based on computational complexity and resource needs.

The comparative analysis in Table 4 shows that the improved YOLOv7 model has fewer parameters and is more compact than the other models. This not only results in a smaller model size but also enhances efficiency in terms of storage and transfer. Despite the reduced number of parameters, the improved YOLOv7

model still requires higher hardware specifications. To ensure the model's portability, pre-training on the dataset can be performed, followed by deployment of the pre-trained model to the application for execution.

Table 4: Comparison of parameters of mainstream target detection algorithms

Model	Parameter quantity/M	Model size/Mb
Faster-RCNN	3.8	54.3
YOLOv3	6.0	25.6
YOLOv4	5.6	28.3
YOLOv5	5.8	14.4
SPD-YOLOv7	3.6	12.8

Based on the comparative analysis of the parameter count in Table 4, the enhanced YOLOv7 model used in this study demonstrates a smaller number of parameters and is more compact than other models. Additionally, the reduced model size leads to lower storage and transfer requirements, improving overall efficiency. Although the improved YOLOv7 model has fewer parameters, it still demands higher hardware specifications. To enhance model portability, pre-training can be conducted using the target dataset, and the pre-trained model can then be deployed on mobile applications. This approach effectively reduces hardware demands while ensuring the model's compatibility across various devices. In Fig. 13, the model was successfully deployed on different mobile devices, utilizing an embedded system for precise pest detection in maize crops.

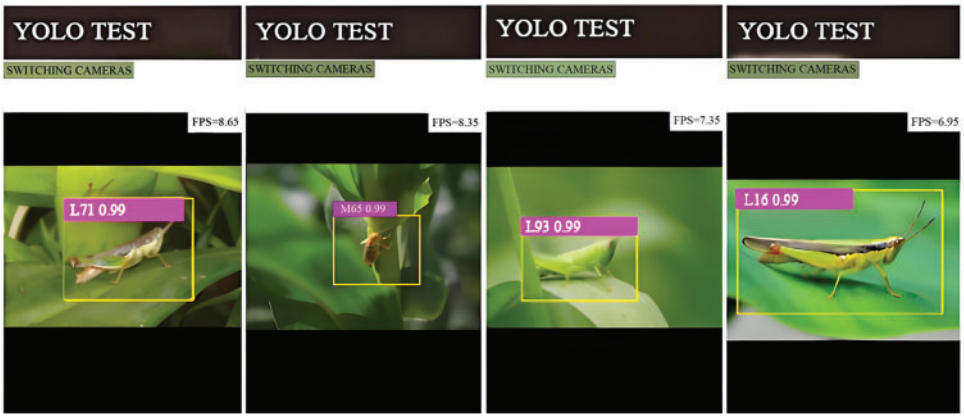


Figure 13: Effect diagram of mobile app detection

The image on the right in Fig. 13 shows the output from the high-end mobile phone client, demonstrating how the system leverages the pre-trained model within the mobile app for target detection. While detection accuracy is slightly lower compared to more powerful hardware configurations, the model's lightweight design allows it to operate efficiently on mobile devices. Future improvements could focus on further reducing the model size through techniques such as pruning, quantization, and the integration of ROI pooling layers to accelerate computation. Additionally, to enhance accuracy, the model could be deployed on a cloud server with more robust processing power. Image data could then be uploaded to the cloud for analysis, with the final classification results sent back to the mobile device. This approach enables the

system to work smoothly on mobile devices with different hardware configurations. It is ideal for real-world applications like precision agriculture, where real-time detection and analysis of crops, pests, or diseases on mobile devices is crucial. Real-time detection and analysis of crops, pests, or diseases via mobile devices is critical. By incorporating the proposed YOLO model into these practical applications, the system could significantly streamline agricultural workflows, improving both efficiency and accuracy in field operations.

6 Conclusion

To address challenges in maize pest detection, including small size, low resolution, and variability across growth stages, we propose an enhanced YOLOv7-based algorithm. By incorporating the lightweight CBAM hybrid attention mechanism, the model improves accuracy and real-time performance, particularly in extracting small target features in complex environments. Field-captured dataset brightness adjustments simulate real-world conditions and increase training difficulty, while the SPD-Conv module reduces information loss, enriches feature extraction, and boosts detection accuracy for small pests. Experimental results demonstrate notable improvements in detection accuracy while maintaining speed.

Furthermore, while the proposed method shows clear benefits, there are limitations, such as the lack of cross validation and other robustness measures for validation. Future work will address these considerations. The scalability of the model will be explored. Its applicability to a wider range of real-world scenarios will also be examined. In addition, future research will focus on adding more validation measures. These will help assess the model's generalization ability and reliability.

Acknowledgement: We thank the editors and anonymous reviewers for their suggestions. At the same time, thank the corresponding author for their support and help.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization: Zhunruo Feng, Ruomeng Shi; Methodology: Zhunruo Feng, Ruomeng Shi, Yuhan Jiang, Yiming Han; Software: Zhunruo Feng, Ruomeng Shi; Validation: Zhunruo Feng, Ruomeng Shi, Yuheng Ren; Resources: Yuheng Ren; Data curation: Zeyang Ma, Yuheng Ren; Writing—original draft preparation: Zhunruo Feng, Yuhan Jiang, Yuheng Ren; Writing—review and editing: Zhunruo Feng, Zeyang Ma, Yiming Han; Supervision: Yuheng Ren; Project administration: Zhunruo Feng, Yuhan Jiang, Yuheng Ren. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang N, Fu S, Rao Q, Zhang G, Ding M. Insect-YOLO: a new method of crop insect detection. Comput Electron Agric. 2025;232(1):110085. doi:10.1016/j.compag.2025.110085.
2. Meng Y, Zhan J, Li K, Yan F, Zhang L. A rapid and precise algorithm for maize leaf disease detection based on YOLO MSM. Sci Rep. 2025;15(1):6016. doi:10.1038/s41598-025-88399-1.
3. Xiang W, Wu D, Wang J. Enhancing stem localization in precision agriculture: a two-stage approach combining YOLOv5 with EffiStemNet. Comput Electron Agric. 2025;231(11):109914. doi:10.1016/j.compag.2025.109914.
4. Venkateswara S, Padmanabhan J. Deep learning based agricultural pest monitoring and classification. Sci Rep. 2025;15(1):8684. doi:10.1038/s41598-025-92659-5.
5. Bi X. Research and application of pest identification technology using geographic information system and computer image recognition. J Phys. 2021;2033(1):12064. doi:10.1088/1742-6596/2033/1/012064.

6. Suzauddola M, Zhang D, Zeb A, Chen J, Wei L, Rayhan A. Advanced deep learn model crop-specif cross-crop pest identif. Expert Syst Appl. 2025;274(3):126896. doi:10.1016/j.eswa.2025.126896.
7. Raj G, Prabadevi B. Enhancing surface detection: a comprehensive analysis of various YOLO models. Heliyon. 2025;11(3):e42433. doi:10.1016/j.heliyon.2025.e42433.
8. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA.
9. Zhao X, Jia H, Ni Y. A novel three-dimensional object detection with the modified you only look once method. Int J Adv Robot Syst. 2018;3(2):116. doi:10.1177/1729881418765507.
10. Yue X, Qi K, Na X, Yang F, Wang W. Deep learning for recognition and detection of plant diseases and pests. Neural Comput Applic. 2025;3(7553):12. doi:10.1007/s00521-025-11125-5.
11. Tang X, Ruan C, Li X, Li B, Fu C. MSC-YOLO: improved YOLOv7 based on multi-scale spatial context for small object detection in UAV-view. Comput Mater Contin. 2024;79(1):983–1003. doi:10.32604/cmc.2024.047541.
12. Dhoundiyal P, Sharma V, Vats S, Rawat P. Progressive hierarchical model for plant disease diagnosis. SN Comput Sci. 2025;6(2):102. doi:10.1007/s42979-024-03582-x.
13. Huang Z, Yang S, Zhou MC, Gong Z, Abusorrah A, Lin C, et al. Making accurate object detection at the edge: review and new approach. Artif Intell Rev. 2022;55(3):2245–74. doi:10.1007/s10462-021-10059-3.
14. Alhwaiti Y, Khan M, Asim M, Siddiqi M, Ishaq M, Alruwaili M. Leveraging YOLO deep learning models to enhance plant disease identification. Sci Rep. 2025;15(1):7969. doi:10.1038/s41598-025-92143-0.
15. Yan Z. A YOLO-NL object detector for real-time detection. Expert Syst Appl. 2024;238(9):122256. doi:10.1016/j.eswa.2023.122256.
16. Cui H, Wei Z. Multi-scale receptive field detection network. IEEE Access. 2019;7:138825–32. doi:10.1109/ACCESS.2019.2942077.
17. Tong K, Wu Y. Rethinking PASCAL-VOC and MS-COCO dataset for small object detection. J Vis Commun Image Represent. 2023;93(4):103830. doi:10.1016/j.jvcir.2023.103830.