ENGINEERING Agriculture

Engineering Journals

RESEARCH ARTICLE

2026 / Volume 13 / Issue 5 / 26682

A lightweight tomato leaf disease detection method integrating multiscale perception and detail enhancement

Qian YUAN, Hui LIU (✉)

Key Laboratory of Traffic Safety on Track of Ministry of Education, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China.

KEYWORDS

Agricultural image analysis, dynamically refined intersection over union loss function, lightweight object detection, multiscale feature fusion, tomato leaf disease detection

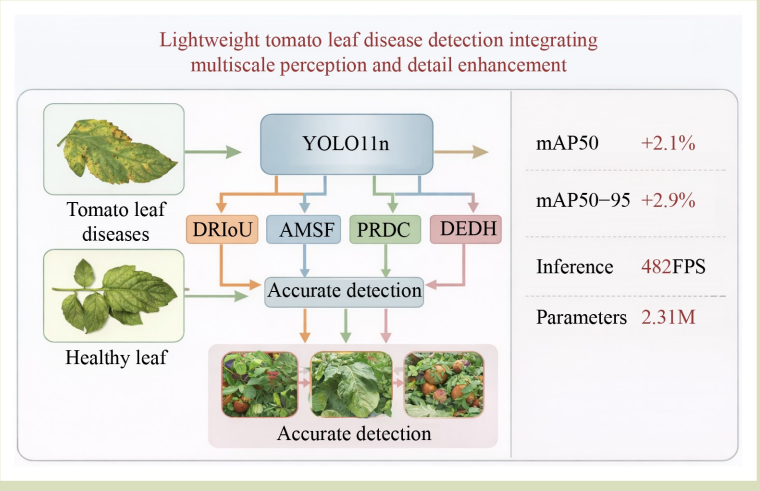
HIGHLIGHTS

- A lightweight tomato leaf disease detection method is developed based on an improved You Only Look Once version 11 nano (YOLO11n) framework.
- A dynamically refined intersection over union loss is introduced to improve box regression across training stages.
- An adaptive multiscale fusion module enhances feature representation for lesions with diverse sizes.
- A progressive receptive field module based on dilated convolutions improves contextual perception in complex backgrounds.
- A detail enhanced detection head strengthens boundary and texture perception while keeping the model compact.

Received July 8, 2025;
Accepted December 15, 2025.

Correspondence: csuliuhui@csu.edu.cn

GRAPHICAL ABSTRACT



ABSTRACT

To address the challenges of large model size, limited detection accuracy and poor adaptability to complex environments in tomato leaf disease detection tasks, this paper proposes a lightweight and efficient detection method based on an improved YOLO11n. First, a dynamically refined intersection over union loss function is introduced to optimize bounding box regression quality across different training stages. Subsequently, an adaptive multiscale fusion module is designed to enhance feature extraction adaptability to varying scales. To further strengthen spatial perception across lesions of different sizes, a progressive receptive field via dilated convolutions module is proposed. Finally, a detail enhanced detection head is incorporated to improve detection performance on small-scale and blurred-boundary disease regions. Extensive experiments validate the effectiveness of the proposed approach, achieving a 2.1% improvement in mean Average Precision at an Intersection-over-Union (IoU) threshold of 0.5 (mAP50) and a 2.9% improvement in mean Average Precision (mAP) averaged over Intersection-over-Union thresholds from 0.5 to

0.95 (mAP50–95) compared with the YOLO11n baseline, while boosting inference speed to 482 frames/second. The proposed method demonstrates excellent accuracy, real-time performance and lightweight deployment capability, providing a novel technical solution and practical support for intelligent agricultural disease detection.

© The Author(s) 2026. Published by Higher Education Press. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0)

1 Introduction

Agricultural production is directly linked to the food security of hundreds of millions of people worldwide, and the health status of crops has a significant impact on agricultural yield and quality. Consequently, research on efficient and accurate technologies for monitoring crop growth and detecting diseases has attracted increasing attention in the agricultural sector[1]. The application of object detection technology in agriculture, by accurately identifying crop pests, diseases and abnormal conditions, can significantly improve production efficiency, reduce labor and resource waste, and ultimately enhance crop quality and yield[2].

Tomato is one of the world's most important economic crops, prized for its rich nutritional value and high yield[3]. However, during cultivation, tomatoes are often threatened by a range of fungal, bacterial and viral diseases. These diseases not only inhibit normal plant growth but may also cause fruit deformation, rot or discoloration, severely reducing market value and resulting in substantial economic losses for growers[4,5]. Among them, tomato leaf diseases are particularly common and diverse. As shown in Fig. 1, our dataset included eight typical disease categories: early blight, late blight, leaf miner, leaf mold, tomato yellow leaf curl virus, Septoria leaf spot, red spider mite and yellow curl leaf virus, along with a healthy category for comparison. Therefore, the development of efficient and accurate tomato disease detection technologies has become an urgent need[6]. Early methods for crop disease detection primarily relied on manually designed features and threshold-based segmentation techniques, including edge detection[7,8], color analysis and texture recognition[9]. Although these methods achieved reasonable results under controlled conditions, they performed poorly in the complex

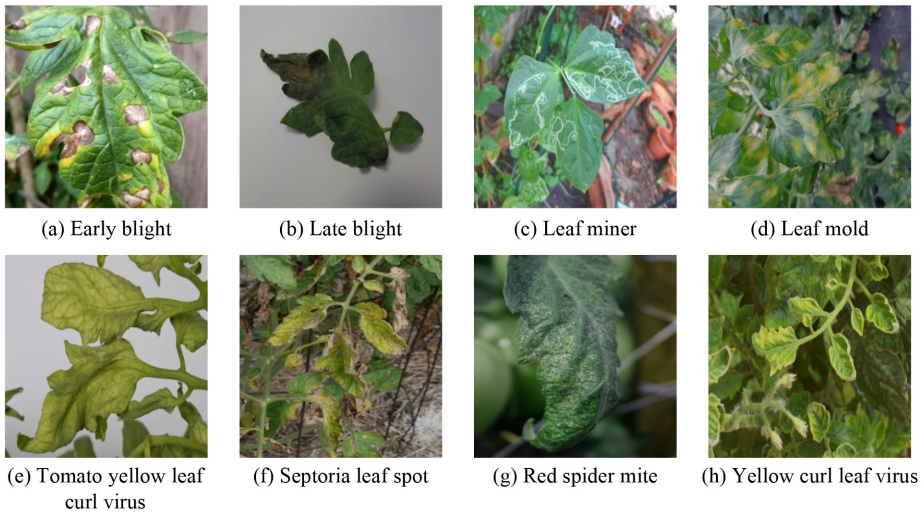


Fig. 1 Representative images of eight tomato leaf disease categories from the tomato leaf disease dataset used in this study, consisting of selected Plant Village images and real-world field samples collected under diverse natural conditions.

environments typical of real-world agricultural production, exhibiting limited generalization capability[10]. Also, traditional machine learning approaches require a high degree of expertise in feature selection and struggle to handle high-dimensional, large-scale data analysis, restricting their practical applications[11].

In contrast to previous methods, deep learning models have demonstrated powerful capabilities in automatic feature extraction and image learning, driving significant progress in agricultural recognition and detection tasks[12,13]. With the rapid advancement of artificial intelligence, especially deep learning, object detection algorithms such as the You Only Look Once (YOLO)[14] series, single shot multiBox detector (SSD)[15] series, and faster regions with Convolutional neural network features (R-CNN)[16] have been widely adopted in agricultural applications, showcasing notable advantages[17]. Nagamani and Sarojadevi[18] conducted a comparative study between support vector machines (SVM)[19] and R-CNN for tomato disease detection. The results demonstrated that R-CNN outperformed SVM in terms of detection accuracy and feature representation capability. Yang et al.[20] proposed a multisource data fusion method based on ShuffleNet V2 for grape leaf disease detection, achieving high detection accuracy. Chu et al.[21] improved the YOLOv4 model to develop a lightweight tomato pest and disease detection framework, achieving reduced parameters and computational cost with minimal loss in accuracy. Jing et al.[22] introduced an enhanced YOLOv5 model that incorporates an automatic labeling algorithm, a weighted bidirectional feature pyramid network and an attention mechanism, resulting in efficient and balanced detection of nine types of tomato leaf diseases. Xue et al.[23] developed a real-time fine-grained tomato disease detection model based on YOLOv8[24], achieving an accuracy of 92.6% on a tomato disease dataset.

Although deep learning methods have significantly advanced tomato disease detection, notable challenges remain. Many current models are limited to detecting specific lesions or single targets, making it difficult to handle the wide range and complexity of tomato images. In addition, their performance often drops under changing natural conditions such as varying light or blocked leaves. Real-time detection also falls short of practical needs. Therefore, it is essential to further improve detection models with more targeted and efficient designs that meet the demands of real-world agricultural environments [25,26].

(1) Among current object detection models, the YOLO series is widely used in agricultural disease detection due to its strong real-time performance and computational efficiency. Based on the YOLO11n[27] architecture, this study proposes an improved tomato disease detection model that incorporates a multiscale feature fusion strategy through architectural enhancements and loss function optimization. To address the challenges posed by indistinct lesions, blurred boundaries, scale variability and environmental complexity such as diverse lighting and weather conditions, this study aims to develop a lightweight, accurate and robust tomato leaf disease detection framework suitable for real-world agricultural scenarios. Accordingly, the objectives of this work were: to improve localization accuracy for small, indistinct, and boundary-blurred lesions, by adopting a more flexible and stage-aware bounding-box regression strategy that remains reliable under complex field conditions;

(2) To enhance multiscale feature representation while maintaining a compact and efficient model, enabling robust detection of lesions with diverse sizes and shapes without imposing excessive parameter or computational overhead;

(3) To strengthen cross-scale spatial context perception in complex backgrounds, so that the model can better distinguish lesion regions from background clutter and varying illumination by more effectively integrating local and global contextual information; and

(4) To improve fine-grained recognition of subtle and low-contrast disease patterns, with particular emphasis on edge and texture details, thereby reducing missed detections and misclassifications of early-stage or weakly expressed lesions.

2 Materials and methods

YOLO11n developed by Ultralytics (Frederick, MD, USA) and released in September 2024, represents the latest generation of object detection algorithms. As the newest addition to the YOLO series, it delivers notable improvements in accuracy, speed, and computational efficiency, further reinforcing the' leading position of the series in real-time object detection. Staying true to the end-to-end, single-stage detection philosophy of YOLO, YOLO11n achieves substantial performance gains through the optimization of several core components. As shown in Fig. 2, its architecture integrates a range of enhanced modules.

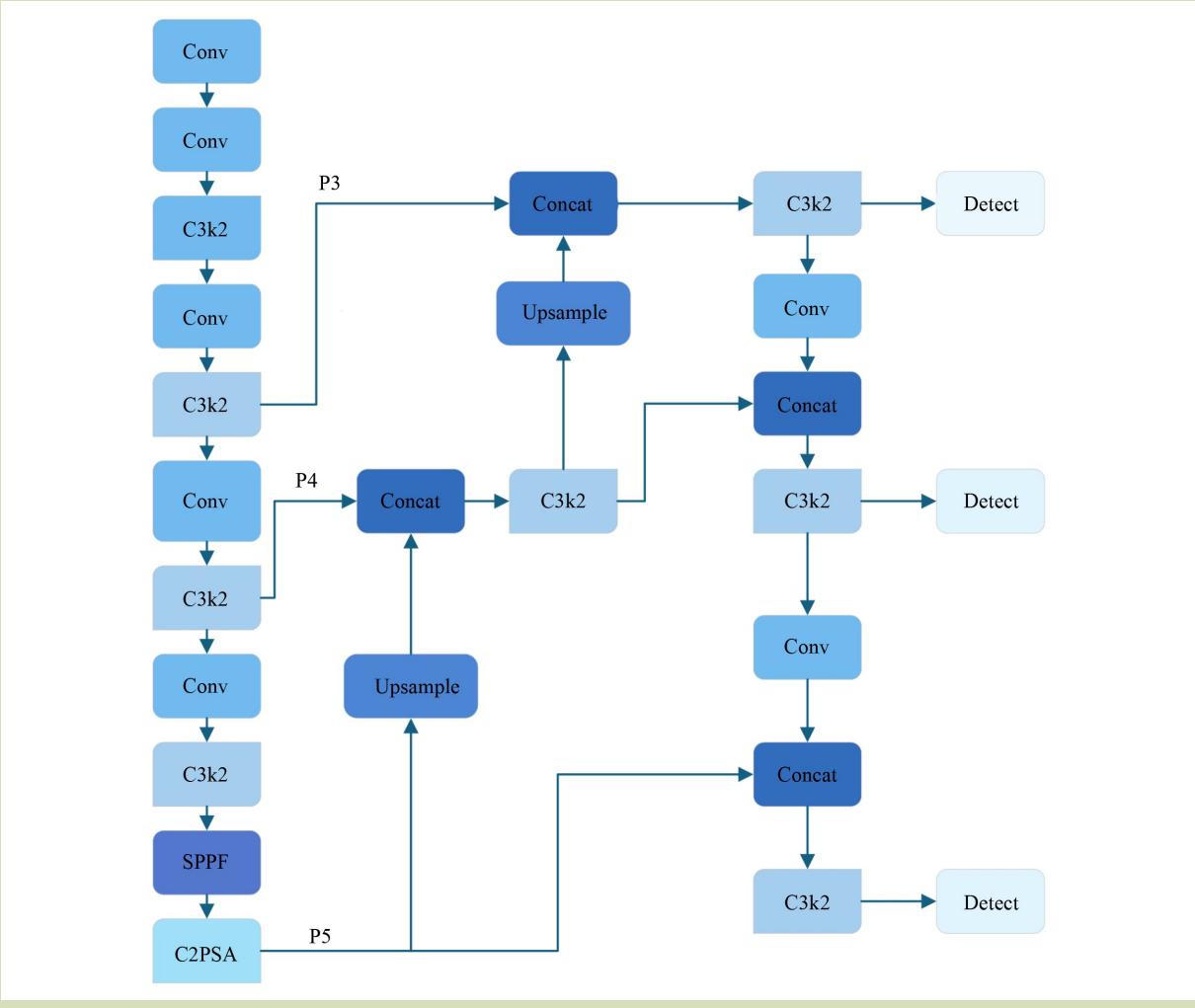


Fig. 2 Overview of the YOLO11n architecture. The model consists of an input module, backbone, neck and detection head. Key components include the C3k2 blocks, the SPPF module and the C2PSA attention-enhanced module.

Structurally, YOLO11n comprises four main components: the input module, backbone, neck and detection head. The backbone use stacked convolutional layers and introduces the C3k2 module, which improves computational efficiency over the previous C2f structure while maintaining strong feature extraction capabilities. The model also retains the spatial pyramid pooling-fast (SPPF) module and augments it with a C2PSA module incorporating a spatial attention mechanism, enabling more precise focus on key regions of the input image. The neck adopts the C3k2 module in conjunction with upsampling and multiscale feature concatenation to further enhance adaptability to objects of various sizes. The detection head handles final classification and bounding box regression to generate output predictions. Overall, YOLO11n achieves a well-balanced improvement in both speed and accuracy through its structural refinements and attention-based mechanisms.

To address persistent challenges in tomato leaf disease detection, including small and indistinct lesions, blurred boundaries and poor adaptability to varying lesion scales, this study evaluated four targeted improvements built upon the YOLO11n architecture. Each enhancement was designed to correspond to a specific component within the model framework, forming a cohesive and task-oriented optimization strategy. At the loss function level, a dynamically refined

intersect over union (DRIoU) loss was introduced to adaptively balance the localization gradient throughout training, thereby improving bounding-box precision and training stability. Structurally, several targeted module substitutions were implemented to enhance feature representation and detection efficiency. The original C3k2 blocks in both the backbone and neck were replaced with a lightweight C3k2-A variant integrated with an adaptive multiscale fusion (AMSF) mechanism, which enables adaptive cross-scale feature interaction and strengthens the extraction of discriminative lesion information. Within the backbone, the original SPPF module was replaced by a progressive receptive field via dilated convolutions (PRDC) module, which progressively expands the receptive field and improves contextual perception under complex environmental conditions. In the detection head, the conventional detection head is substituted with a detail-enhanced detection head (DEDH) that emphasizes fine-grained texture and low-contrast regions, improving recognition of subtle disease manifestations. Collectively, these replacements created a coherent lightweight framework that enhances multiscale adaptability, spatial context awareness, and detailed feature perception for agricultural disease detection.

2.1 Dynamically refined intersect over union

In tomato leaf disease detection tasks, target regions often present challenges such as blurred boundaries, varied lesion shapes and significant overlap or adhesion between instances. These complexities require the model to not only perform rapid localization of lesion areas during the early stages of training, but also to gradually refine the prediction boundaries for accurate detection of subtle or early-stage symptoms. To meet these demands, the bounding box regression loss function must adapt its focus and gradient behavior dynamically throughout the different phases of training.

The core idea of unified intersect over union (UIoU) lies in applying a scaling factor, referred to as the ratio, to transform the predicted bounding boxes. This involves linearly scaling the width and height of the boxes according to the ratio, which is dynamically adjusted as training progresses. By gradually reducing the ratio from a larger initial value to a smaller one, the model is able to tolerate lower-quality predictions during the early stages of training and progressively shift its focus toward high-quality targets with greater overlap in the later stages[28]. The core computation is:

$$\mathrm{UIoU}(B_p, B_{gt}; \mathrm{ratio}) = \mathrm{IoU}(B_p \times \mathrm{ratio}, B_{gt} \times \mathrm{ratio}) \quad (1)$$

(1) where IoU denotes the standard intersection-over-union metric defined earlier, and ratio is a dynamic factor that varies with training epochs. It scales the width and height of the bounding boxes proportionally while keeping the center position unchanged. While UIoU demonstrates effectiveness by dynamically scaling predictions, the key lies in the design of the ratio function. In the original method, strategies such as linear decay, cosine decay or fractional decay were used to control the ratio. However, these approaches present notable limitations: linear decay, although simple, maintains a constant decay rate, lacking adaptability to different phases of training and therefore cannot realize early-stage tolerance, mid-stage stabilization and late-stage fine-tuning;

(2) cosine decay offers a smooth transition but its curvature is inherently dictated by the cosine function, limiting flexibility in adjusting the turning point and gradient variation; and

(3) fractional decay provides some degree of smoothness but its overall variation is too small, causing the ratio to decline too slowly, which fails to sufficiently distinguish between training stages and allows low-quality predictions to interfere in later stages.

In summary, these established decay strategies lack the necessary flexibility, tunability and stage-awareness, making it difficult to meet the evolving demands of deep detection models throughout training.

To overcome these limitations, we developed DRIoU, a more flexible and precisely controllable ratio function for improved IoU computation. The proposed ratio design is based on a sigmoid-like function, allowing precise control over the ratio value at mid-epoch and achieving a smooth transition throughout the entire training process. The mathematical definition is:

$$\mathrm{ratio}(t) = a \cdot \frac{1}{1 + \exp^{k(e-c)}} + b \quad (2)$$

where e denotes the training epoch index, a represents the total decay amplitude, b is the final convergence value, k is a hyperparameter controlling the decay speed of the ratio (with a default value of 0.05, adjustable as needed) and c is the turning point of the curve. The latter can be derived based on the target epoch as:

$$c = e^* - \frac{1}{k} \ln\left(\frac{a}{r^* - b} - 1\right) \quad (3)$$

where $e^* \in [0,E]$ is the specified target epoch and r^* is the target

ratio value. For example, assuming the total number of training epochs is set to 200, the initial ratio is set to 2.0, and the final convergence value is 0.5, resulting in $a = 1.5$ and $b = 0.5$. When the target epoch is set to 100 and the target ratio set to 1.25, c is 100. The visualization of the result is shown in the following Fig. 3, which illustrates how different ratio functions evolve with the number of training epochs.

For completeness, we additionally included a simple back-loaded polynomial schedule to visualize a different decay trend during training:

$$\text{ratio}_{\text{poly}}(e) = 2.0 - 1.5\left(\frac{e}{200}\right)^2 \tag{4}$$

This curve decreases slowly at the beginning and accelerates toward the end, yielding a monotone, back-loaded trajectory. We used this polynomial schedule only as a supplementary baseline to illustrate an alternative decay trend; in contrast, our proposed ratio design achieves a more controlled and flexible evolution over training via its tunable parameters.

2.2 Adaptive multiscale fusion module

To address the complex lesion morphology and significant scale variations observed in tomato disease images, this study proposes a lightweight and efficient AMSF. In real-world tomato leaf disease datasets, different types of diseases typically appear as spots, stripes or large leaf blotches of varying sizes. Traditional convolutional operations with a single kernel scale struggle to effectively capture such diverse features simultaneously.

To overcome this limitation, the AMSF module divides the input feature map along the channel dimension into three groups. Of these, 50% of the channels are retained without modification, aiming to preserve low-level detail information. The remaining two channel groups are passed through standard 3 × 3 and 5 × 5 convolutional kernels, respectively, which are responsible for extracting features from medium-scale and large-scale lesion areas. Through this collaborative multi-kernel design, AMSF effectively captures the hierarchical spatial structure of disease lesions across different scales. The structure is shown in Fig. 4.

The feature maps generated by the multiscale branches are concatenated along the channel dimension and then passed through a 1 × 1 convolution to compress the channel number and fuse information. This design ensures that the output features possess both strong spatial representation and computational efficiency. The structure not only improves the sensitivity of the model to various types of tomato leaf diseases but also achieves an acceptable balance between boundary preservation for small lesions and perception of large infected areas, making it well suited for fine-grained disease detection in high-resolution images.

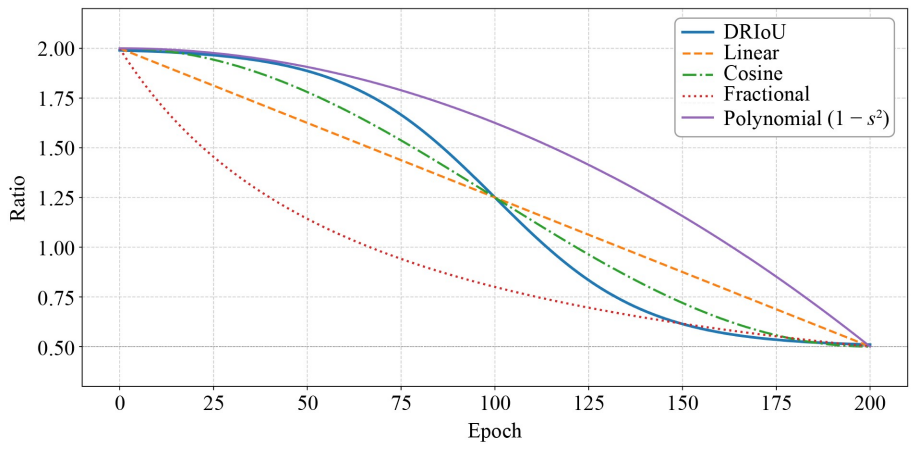


Fig. 3 Comparison of ratio scheduling functions used in the intersect over union computation. The curves illustrate the decay patterns of linear, cosine, fractional, polynomial and the proposed DRIoU schedule over training epochs. The DRIoU curve provides more flexible mid-epoch control and smoother transitions, improving stage-aware regression behavior.

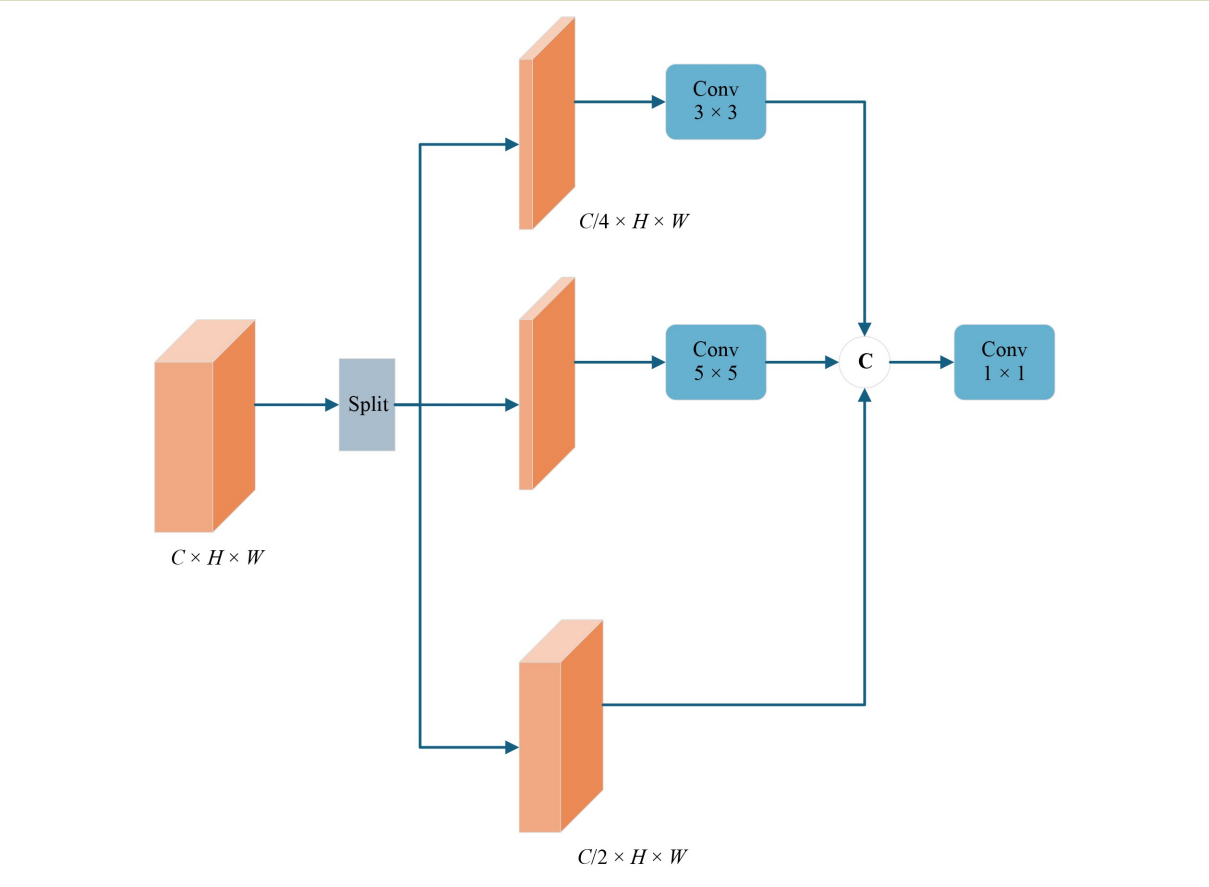


Fig. 4 Structure of the adaptive multiscale fusion module. The input features are split into three channel groups: a direct-pass branch, a 3 × 3 convolution branch and a 5 × 5 convolution branch. C denotes channel-wise concatenation, where feature maps from different branches are concatenated along the channel dimension before being fused by a 1 × 1 convolution.

In the original YOLO11n architecture, the C3k2 module has certain limitations in feature extraction, and its parameter count and computational complexity remain relatively demanding for resource-constrained embedded devices. To enhance the feature representation capability while reducing the computational burden, we propose replacing the original C3k2 with an improved C3k2-A module. Specifically, the standard convolution units within the bottleneck structure of C3k2 are replaced with the proposed AMSF module, which significantly strengthens the ability of the network to extract multiscale features. Meanwhile, this design expands the receptive field while retaining critical features, thereby improving the capacity of the network for spatial perception and semantic expression. As a result, the proposed modification achieves a better trade-off between feature extraction performance and computational efficiency. The integrated design of the improved module is shown in Fig. 5.

2.3 Progressive receptive field via dilated convolutions module

In tomato leaf disease detection tasks, the traditional SPPF achieves effective multiscale feature fusion through rapid pooling operations. However, when the diseased regions exhibit drastic scale variations and rich fine-grained details, pooling operations often lead to the loss of local details, making it difficult for the model to accurately capture the diverse and intricate features of the diseased areas.

To address the diversity of lesion morphology and the significant scale differences in tomato leaf disease images, we propose a PRDC module. The goal of this design is to significantly enhance the feature perception and representation capabilities of the model across different scales of diseased regions.

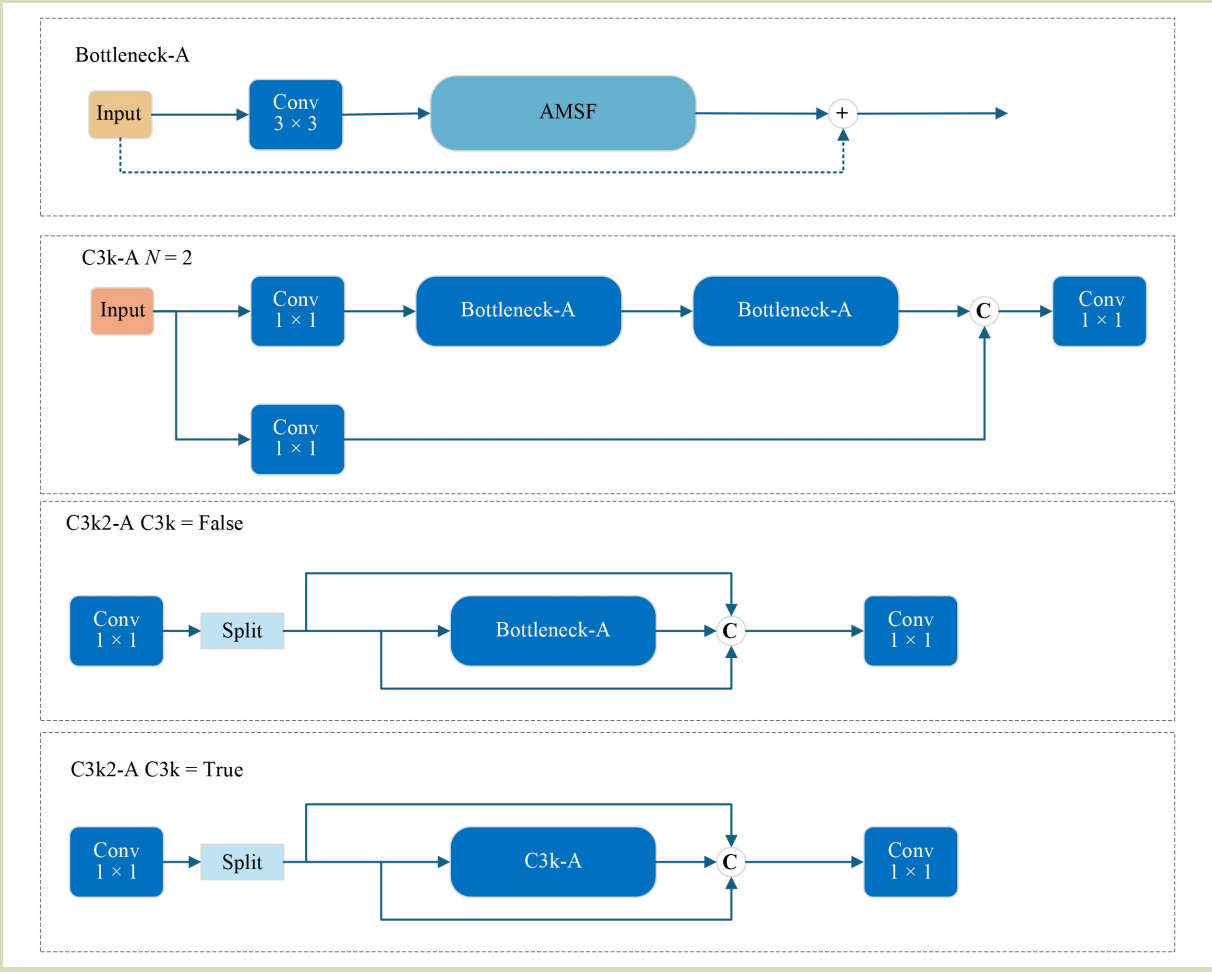


Fig. 5 Integration of the adaptive multiscale fusion (AMSF) module within the lightweight C3k2-A structure. The original convolution units inside C3k2 are replaced by AMSF to achieve multiscale feature extraction.

The PRDC module progressively expands the receptive field by stacking multiple dilated convolutional layers, effectively capturing rich spatial context information while maintaining a relatively low computational cost.

As shown in Fig. 6, the PRDC module first applies a 1 × 1 convolution to compress the input feature channels to half of their original number, thereby reducing computational cost and improving efficiency. After channel compression, the resulting feature map is denoted as F_0. This feature serves as the starting point for the progressive dilated convolution process.

Subsequently, three 3 × 3 dilated convolutional layers are applied in sequence, with dilation rates 1, 2 and 5. The three dilated layers share a single convolution kernel W_s. Only the dilation rate varies among them, while the stride remains 1 and the padding equals the dilation to keep the feature map size unchanged. The input to each dilated layer is represented by F_{prev}, indicating the feature map generated in the previous step. At the beginning of the process, F_{prev} is initialized as F_0. After each convolution, the output feature map F_d becomes the new F_{prev}, which is then passed into the next dilated layer. In this way, the receptive field of the feature map is progressively expanded while maintaining continuity between stages. The algorithm that summarizes this forward process is given in Table 1.

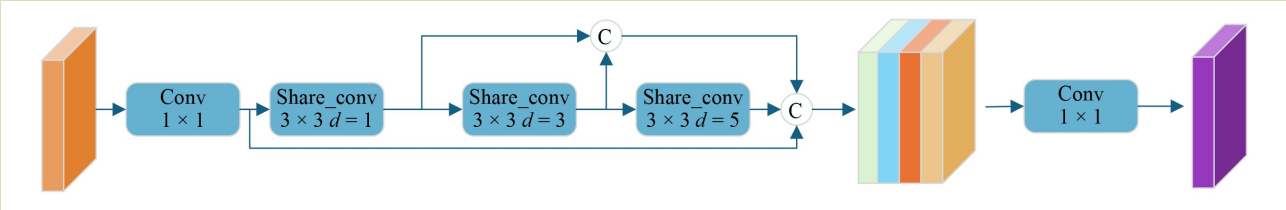


Fig. 6 Structure of the PRDC (progressive receptive field via dilated convolutions) module. The module compresses input channels using a 1 × 1 convolution, followed by three sequential dilated convolution layers with increasing dilation (*d*) rates, where d denotes the dilation factor of the convolution kernel. Shared weights and skip connections help preserve detailed textures while expanding contextual perception.

Table 1 Algorithm of the PRDC (progressive receptive field via dilated convolutions) module

Step	Operation
1	Input feature map $X \in R^{B \times C \times H \times W}$
2	$F_0 \leftarrow$ Conv 1×1(X) channel compression; $F_{prev} \leftarrow F_0$
3	for each $d \in \{1, 3, 5\}$ do
4	$F_d \leftarrow$ Conv 3 × 3(F_{prev}, W; dilation = d, padding = d, stride = 1)
5	Append F_d to feature list; $F_{prev} \leftarrow F_d$
6	$F_{out} \leftarrow$ Concat(F_0, F_1, F_2, F_3)
7	$Y \leftarrow$ Conv 1 × 1(F_{out})
8	**Return** Y

The layers with smaller dilation rates emphasize local and fine-grained details, which help capture small lesion textures, while the layers with larger dilation rates expand the receptive field to incorporate broader contextual information of larger diseased areas. Through this progressive connection mechanism, PRDC ensures that both local detail and global context are effectively represented throughout the forward propagation process.

Also, to further reduce parameter count and computational overhead, the PRDC module introduces a shared-weight mechanism, where the three dilated convolutional layers with different dilation rates share a single set of convolutional kernel weights. This design significantly cuts down the number of parameters and lowers model complexity, making the PRDC module more suitable for deployment on resource-constrained embedded devices.

Through this design, PRDC efficiently and progressively captures multiscale features, ranging from small local spots to large-scale leaf blight or moldy areas, allowing precise and comprehensive modeling of diseased regions in images. After the progressive dilated convolutions, the outputs from all stages are fused and concatenated, and the initial 1 × 1 compressed feature is added via a skip connection. This strategy effectively integrates shallow edge features with deep semantic information, which is particularly beneficial for tomato leaf disease detection tasks. It enables the model to represent spatial features from fine-grained lesion boundaries to large-scale diseased regions comprehensively.

Finally, a subsequent 1 × 1 convolution is applied to the fused feature map to further integrate channel information, resulting in a more compact, efficient, and semantically rich feature representation, thereby enhancing the overall detection performance of the model for tomato leaf diseases.

2.4 Detail enhanced detection head

From this study, we propose a novel DEDH structure tailored for tomato leaf disease detection tasks, based on a deeply optimized and improved version of the classical YOLO11n detection head. The overall structure is shown in Fig. 7.

In the detection stage, a lightweight DEDH is designed to replace the conventional decoupled detection heads at all feature pyramid levels (P3–P5). At the input stage, multiscale features are first projected through a unified 1 × 1 convolution with group normalization, which aligns channel dimensions and stabilizes optimization under small-batch conditions. The channel dimension is compressed to 256, effectively reducing computational overhead while maintaining representative capacity. Subsequently, these aligned features are passed through two cascaded detail-enhancement convolution (DEConv) modules, which are shared across all scales to reduce redundant parameters. Each DEConv module contains five parallel convolutional branches, center-difference, horizontal-

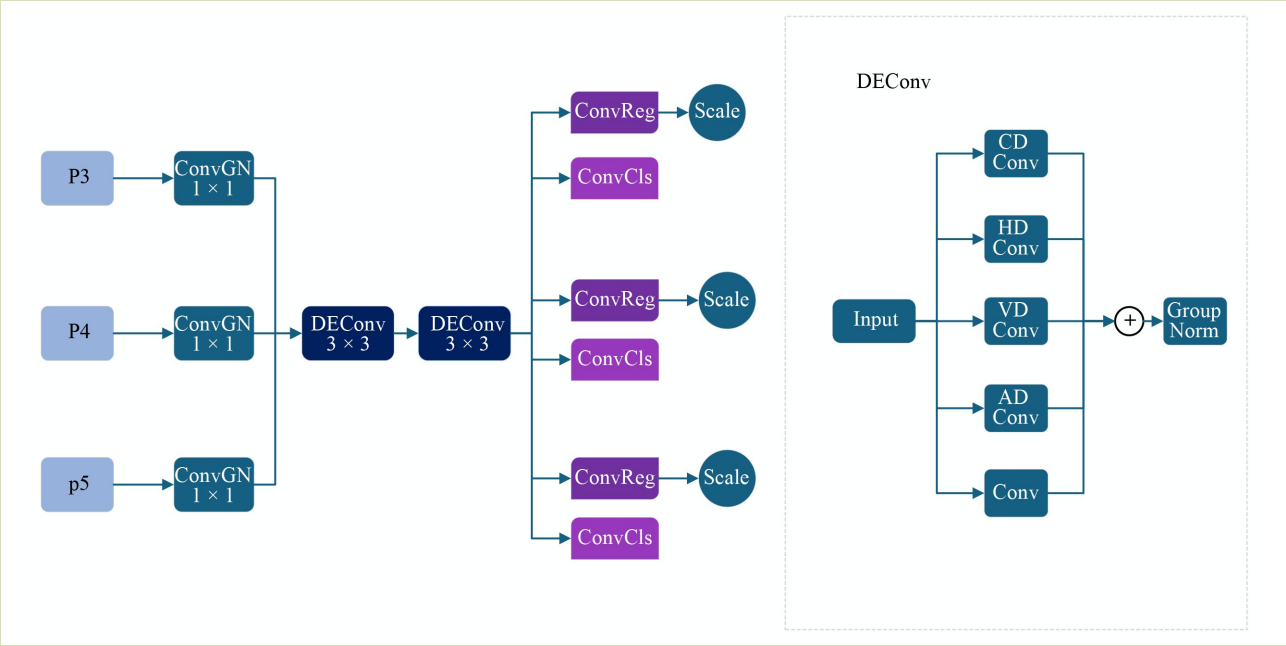


Fig. 7 Structure of the detail-enhanced detection head and the detail-enhancement convolution (DEConv). Each DEConv block contains five parallel convolutional branches, including center-difference convolution (CD Conv), which models local intensity variations relative to the kernel center; horizontal-difference convolution (HD Conv) and vertical-difference convolution (VD Conv), which enhance horizontal and vertical gradient responses, respectively; asymmetric-difference convolution (AD Conv), which captures directional edge features using asymmetric kernels; and a standard convolution branch for complementary feature extraction. Shared weights and differential filtering strengthen fine-grained texture representation and boundary perception.

difference, vertical-difference, asymmetric-difference and a standard convolution, whose outputs are summed and normalized. This design introduces multidirectional gradient cues and fine-grained texture enhancement, allowing the head to more effectively capture lesion boundaries and subtle structural variations in diseased leaves.

After feature enhancement, the refined outputs are processed by two lightweight 1 × 1 convolutional branches for classification and regression, respectively. A learnable Scale parameter is introduced in the regression branch to automatically calibrate the prediction magnitude across different feature levels, improving localization consistency.

Overall, the multilevel feature refinement and differential convolution mechanism in DEDH notably enhance the capability of the model to detect small-scale lesions and areas with blurred boundaries, substantially improving detection accuracy and generalization under complex background conditions.

From this study, we propose four novel differential convolution modules to effectively enhance the ability of convolutional neural networks to capture feature details, edges and directional changes in images. These modules include CD Conv, AD Conv, HD Conv and VD Conv.

To clarify the working principle of differential convolution, we first define each position within a standard 3 × 3 convolution kernel using the letters A to I, arranged from the top-left to the bottom-right, with each letter corresponding to the weight at that specific position in the kernel:

$$\text{Kernel} = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix} \tag{5}$$

Differential convolution modifies the kernel weights to emphasize the differences between local pixels, rather than relying solely on traditional weighted summation. It focuses on capturing relative changes among neighboring pixels, which enhances the extraction of edge information and directional

features. This technique is particularly effective for highlighting local structural details, especially in the presence of complex boundaries and varied texture patterns within an image.

As shown in Fig. 8, these modules model pixel differences within the standard 3 × 3 convolutional kernel through different strategies, effectively enhancing edge sensitivity and directional information expression in the feature maps.

2.4.1 CD Conv

The CD Conv module highlights the intensity differences between the center position and its surrounding pixels by modifying the convolutional kernel weights. Specifically, it retains the original weights of the eight neighboring positions in a 3 × 3 kernel and assigns the negative sum of these weights to the center position. This adjustment enhances the contrast between the center and its surroundings, thereby improving the representation of edges and lesion boundaries.

2.4.2 AD Conv

The AD Conv module focuses on modeling asymmetric local gradient changes by calculating differences between adjacent pixels, particularly between left and upper neighbors. Unlike traditional symmetric center-based differencing, AD Conv emphasizes direction-specific variations, such as the difference between A and D (A–D), B and A (B–A) and C and B (C–B). This asymmetric approach effectively captures rotation, skewness and subtle deformation patterns, thereby improving the ability of the model to represent complex lesion morphologies.

2.4.3 HD Conv

The HD Conv module specializes in extracting horizontal features by jointly modeling the left-side pixels (e.g., A, D and G) and the negated right-side pixels (e.g., −A, −D and −G) within the kernel. This operation strengthens the horizontal gradient responses, enhancing the sensitivity of the model to horizontally aligned cracks, streaks and other fine-grained structures. It is crucial for in detecting horizontally distributed lesions or stripe-like disease patterns on tomato leaves.

2.4.4 VD Conv

The VD Conv focuses on extracting vertical edges by combining the topside pixels (A, B and C) with the negated bottom-side pixels (−A, −B and −C). Since it measures changes along the vertical axis, it enhances responses to horizontally extended lesions, streaks and vein crossings, helping the model to detect horizontally distributed disease patterns in tomato leaves.

Figure 9 shows differential convolution on tomato leaf lesions using the VD and CD branches. It includes a diseased region with horizontally extended mining patterns, the VD-branch heatmap from the second DEConv block, showing strong activation along the internal streaks and texture transitions of the lesion. Also, it shows another lesion with a contrast-dominated fuzzy boundary and the CD-branch heatmap from the second DEConv block, emphasizing intensity transitions between the central area and its surrounding tissue. All feature maps are bilinearly upsampled to the input resolution and normalized with a shared color scale. The corresponding quantitative analysis is described in a later section.

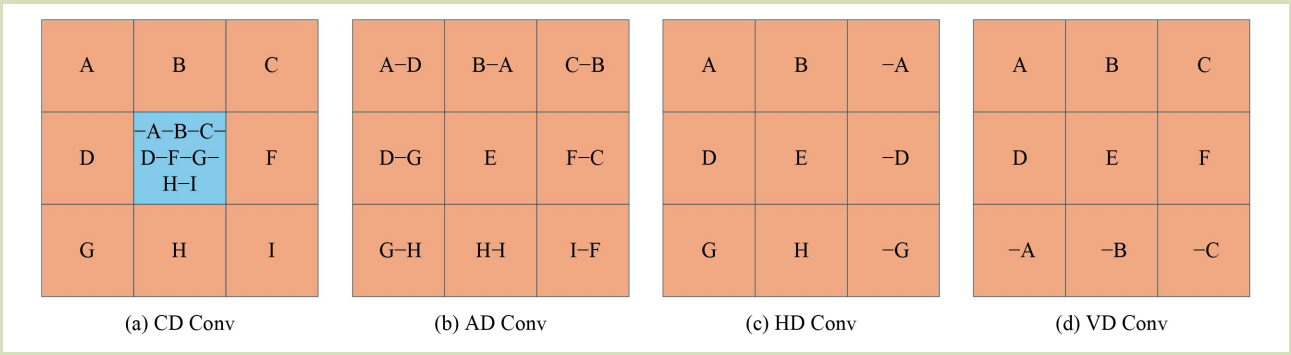


Fig. 8 Illustration of differential convolution modules (CD, AD, HD and VD Conv). The figure shows how different local gradient patterns are emphasized by modifying the 3 × 3 kernel weights, thereby enhancing edge detection and directional feature extraction.

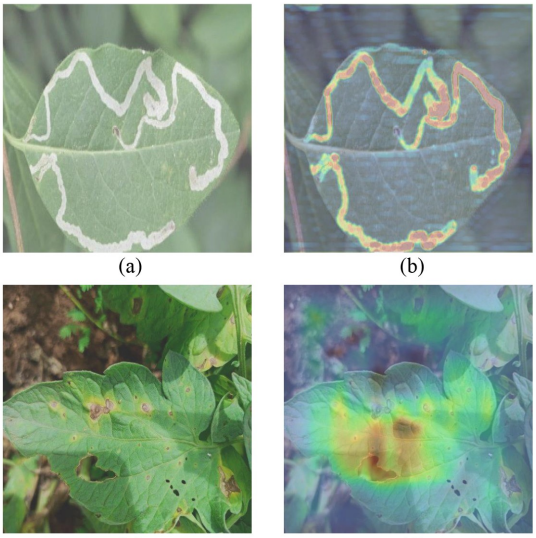


Fig. 9 VD- and CD-branch visualizations on tomato leaf lesions: (a) input, (b) VD heatmap, (c) input, and (d) CD heatmap, all taken from the second detail-enhancement convolution block.

2.5 Experimental environment

The experiments were conducted on a Windows 11 operating system with an NVIDIA GeForce RTX 4060 Ti GPU. The CPU used is an Intel Core i7-14700F, with 32 GB of RAM. The adopted deep learning framework was PyTorch 2.4.1.

During training, the initial learning rate was set to 0.01, the momentum to 0.937 and the weight decay to 0.0005. Considering the dataset size and hardware capabilities, the batch size was set to 16 and the model was trained for 200 epochs. All input images were resized to 640 × 640 pixels before training and inference to maintain uniform feature resolution and balance between detection accuracy and computational efficiency.

2.6 Dataset and preprocessing

The Plant Village dataset provides a reliable foundation for tomato leaf disease detection research but primarily contains images collected under controlled experimental conditions. To improve adaptability in real agricultural environments, we reconstructed and expanded the dataset by selecting about 61.3% of high-quality Plant Village images and supplementing them with 38.7% of real-world tomato leaf disease samples collected through web crawling from open agricultural image repositories and field photography. All images were manually annotated using LabelImg, ensuring precise bounding boxes for lesion regions and category labels.

To enhance data diversity and model robustness, multiple augmentation strategies were used, including mosaic fusion, random scaling, brightness adjustment and fog simulation, simulating environmental variations such as uneven lighting, shadows and background interference. The final dataset consists of 7000 images (5416 for training, 801 for validation and 783 for testing), containing 25337 annotated instances across nine categories. As shown in Fig. 10, the dataset included simulated tomato leaf images under diverse lighting conditions. Detailed annotation statistics for each category are listed in Table 2. This balanced composition ensures that the model evaluation better reflects performance in practical agricultural scenarios.

2.7 Evaluation metrics

In visual object detection tasks, model performance is typically evaluated using a combination of metrics. The most commonly used indicators include precision, recall, mean average precision, number of parameters and giga floating-point operations (GFLOPs).

In classification-related tasks, predictions are generally categorized into four outcomes: true positive (TP), true negative, false positive (FP) and false negative (FN). Based on these, two key evaluation metrics are defined:

Precision: the proportion of correctly predicted positive samples out of all predicted positives,

$$P = \frac{TP}{TP + FP} \tag{6}$$

Recall: the proportion of correctly predicted positive samples out of all actual positives,

$$P = \frac{TP}{TP + FN} \tag{7}$$

To further quantify the performance of object detection models, average precision (AP) is commonly used to evaluate the trade-off between precision and recall for a single class. It is calculated based on the precision-recall (P-R) curve, where prediction results are sorted by confidence scores, and

Fig. 10 Examples of tomato leaf images after augmentation under different lighting conditions. (a–e) The original images; (b–f) the corresponding augmented samples with simulated low-light, foggy and strong-light conditions, used to improve model robustness

corresponding precision and recall values are computed by five steps.

First, the area under the P-R curve is computed to represent the AP value:

Table 2 Annotation statistics for each tomato leaf disease category.

Category	Annotation
Early blight	2486
Healthy	2703
Late blight	3053
Leaf miner	2451
Leaf mold	3063
Tomato yellow leaf curl virus	3020
Septoria leaf spot	2953
Red spider mite	2331
Yellow curl leaf virus	3277

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\,\mathrm{dR} \tag{8}$$

Second, the mAP, the mean of AP values, is calculated across all categories, and is used to evaluate the overall performance of multiclass object detection models:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i \tag{9}$$

Third, parameters refer to the total number of learnable weights in the model and serve as a key indicator of the size and structural complexity of the model. It is typically calculated by summing the weights across all layers. Models with fewer parameters are generally more lightweight and suitable for deployment in resource-constrained environments, while those with more parameters often have stronger feature representation capabilities at the cost of higher storage and computational demands.

Fourth, GFLOPs indicate the number of floating-point operations (in billions) required during the forward pass of the model. This metric reflects the computational cost and inference complexity of the model, and it is often used to assess performance efficiency on hardware platforms.

Finally, frames per second (FPS) measures the number of images the model can process per second. It is a critical indicator of real-time performance in object detection tasks.

To quantitatively evaluate how effectively the model learns and aligns with image structures, we used the relative directional alignment score (RDAS). For each test image, we tapped the detect-input P3 fused tensor $\mathrm{T} \in \mathrm{R}^{C\times H'\times W'}$ (Detet-P3)and form a single-channel map by channel-wise mean:

$$F(u,v) = \frac{1}{C}\sum_{c=1}^{C} I_c(u,v) \tag{10}$$

Then bilinearly upsample F to the image size and apply min-max normalization. From the grayscale image we computed Scharr gradients G_x and G_y, and orientation energies:

$$E_H = |G_x|,\ E_V = |G_y|,\ E_{D+} = \left|\frac{G_x+G_y}{\sqrt{2}}\right|,\ E_{D-} = \left|\frac{G_x-G_y}{\sqrt{2}}\right|,$$
$$E_{\mathrm{edge}} = \sqrt{G_x^2+G_y^2} \tag{11}$$

RDAS is the Spearman rank correlation between vec(F) and each energy vector:

$$\mathrm{RDAS}_H = \rho(F,E_H),\ \mathrm{RDAS}_V = \rho(F,E_V) \tag{12}$$

$$\mathrm{RDAS}_D = \max\{\rho(F,E_{D+}),\rho(F,E_{D-})\},\mathrm{RDAS}_{\mathrm{edge}} = \rho\left(F,E_{\mathrm{edge}}\right) \tag{13}$$

For each test image, four RDAS values were calculated and summarized as mean, standard deviation and improvement rate.

3 Results and discussion

The UIoU framework suffers from a relatively fixed ratio function that lacks flexibility and fine-grained control. To overcome this limitation, we developed DRIoU, a ratio function that provides greater flexibility and precise controllability. To examine the influence of different ratio scheduling strategies on overall model performance, we conducted comparative experiments using the linear, cosine, fractional, polynomial and DRIoU strategies.

As shown in Table 3, DRIoU consistently achieved the best performance across all evaluation metrics. It delivered higher precision and recall compared with the other strategies, indicating improved localization accuracy and stronger target-capture capability. In terms of overall performance, DRIoU attained the highest scores on both mean Average Precision at an Intersection-over-Union threshold of 0.5 (mAP50) and mean Average Precision averaged over Intersection-over-Union thresholds from 0.5 to 0.95 (mAP50–95), surpassing the linear, cosine, fractional and polynomial options. The polynomial schedule was included as a supplementary comparison to represent a distinct slow-to-fast decay pattern that had not been covered by previous functions. While it provides an additional reference for understanding different decay responses, its curvature and turning characteristics remain fixed and lack tunability.

Table 3 Performance comparison of different ratio functions for IoU

Method	Precision	Recall	mAP50	mAP50–95
YOLO11n	0.9381	0.873	0.935	0.848
YOLO11n + UIoU (linear)	0.9377	0.883	0.938	0.854
YOLO11n + UIoU (cosine)	0.9381	0.887	0.940	0.857
YOLO11n + UIoU (fractional)	0.9379	0.884	0.939	0.855
YOLO11n + polynomial	0.9375	0.879	0.937	0.853
YOLO11n + DRIoU	0.9383	0.892	0.941	0.860

Note: YOLO11n + UIoU (linear), YOLO11n + UIoU (cosine), and YOLO11n + UIoU (fractional) denote YOLO11n models trained with the unified intersection over union (UIoU) loss, where the ratio factor is scheduled using linear decay, cosine decay, and fractional decay strategies, respectively. YOLO11n + polynomial adopts a polynomial decay function to control the ratio factor during training. YOLO11n + DRIoU represents YOLO11n equipped with the proposed dynamically refined intersection over union (DRIoU) loss.

These findings demonstrate that DRIoU not only enhances the focus of the model on high-quality bounding boxes but also provides greater stability across varying IoU thresholds, thereby improving robustness in complex detection scenarios. The strength of DRIoU lies in its stage-aware ratio scheduling mechanism: during early training, a larger ratio increases tolerance for coarse predictions and facilitates rapid structure learning, while in later stages, the ratio decays more sharply and gradually stabilizes, guiding the model toward refined high-IoU regression, particularly beneficial for detecting lesions with blurred boundaries or subtle early-stage symptoms.

In summary, by enabling flexible control over both the decay rate and the transition dynamics of the ratio function, DRIoU achieves a more balanced trade-off between precision and recall, leading to substantial improvements in overall detection performance for tomato leaf disease identification.

To comprehensively evaluate the performance advantages of the proposed DRIoU method in bounding box regression tasks, we conducted comparative experiments with several classical and enhanced IoU-based loss functions, including distance intersection over union[29], complete intersection over union, efficient intersection over union[30] and SCYLLA intersection over union (SIoU)[31]. All experiments were done using the same YOLO11n backbone network.

DRIoU achieves the highest overall performance among all compared methods (Table 4). It gave superior precision and recall, reflecting both enhanced localization accuracy and stronger target-capture capability. In terms of the comprehensive evaluation metrics, DRIoU gave the best results on both mAP50 and mAP50–95, highlighting its robustness and stability across varying IoU thresholds. Compared with Eliot and SIoU, DRIoU provided greater consistency in detecting objects of different scales and effectively suppresses regression noise in overlapping or occluded regions.

Table 4 Comparison of intersect over union-based loss functions on YOLO11n

Method	Precision	Recall	mAP50	mAP50–95
YOLO11n	0.938	0.873	0.935	0.848
YOLO11n + DIoU	0.935	0.887	0.940	0.854
YOLO11n + CIoU	0.938	0.881	0.939	0.853
YOLO11n + EIoU	0.934	0.882	0.935	0.852
YOLO11n + SIoU	0.936	0.883	0.938	0.857
YOLO11n + DRIoU	0.938	0.892	0.941	0.860

Note: YOLO11n + DIoU, YOLO11n + CIoU, YOLO11n + EIoU, and YOLO11n + SIoU denote YOLO11n models trained with distance intersection over union (DIoU), complete intersection over union (CIoU), efficient intersection over union (EIoU) and SCYLLA intersection over union (SIoU), respectively. YOLO11n + DRIoU represents YOLO11n equipped with the proposed dynamically refined intersection over union loss.

The superior performance of DRIoU can be attributed to its adaptive ratio scheduling mechanism. In the early stages of training, the ratio remains larger, allowing the model to tolerate low-quality predictions and rapidly learn general object structures. As training progresses, the ratio decreases sharply and then stabilizes, guiding the model to concentrate on refining high-IoU predictions and improving boundary precision. This adaptive behavior enhances both the reliability and quality of the final predictions.

In summary, by dynamically controlling the decay pattern of the ratio function, DRIoU achieves a more balanced trade-off between precision and recall and consistently improves the regression accuracy and generalization capability of the model in tomato leaf disease detection.

3.1 Ablation evaluation

To systematically evaluate the contribution of each proposed module to tomato disease detection performance, we conducted a series of stepwise ablation experiments. The results show that each module contributes incremental improvements in both detection accuracy and computational efficiency.

Starting from the YOLOv11n baseline, the model achieved an mAP50 of 0.935 and an mAP50–95 of 0.848, with 2.58 million parameters, 6.3 GFLOPs, and an inference speed of 407 FPS. The corresponding precision and recall were 0.938 and 0.873, respectively. After integrating the proposed DRIoU loss function, mAP50 increased to 0.941 and mAP50–95 to 0.860, without adding any parameters or computational complexity. This demonstrates that DRIoU enhanced bounding-box regression stability and accelerates convergence while maintaining the same model scale.

Replacing the original C3k2 block with the lightweight C3k2-A module further improves efficiency. This module incorporated a 50% direct-pass branch combined with 3 × 3 and 5 × 5 convolutional sub-branches, followed by a 1 × 1 fusion layer for channel recombination. Such a design reduces redundant convolution operations on half of the feature channels and leverages multiscale perception within a smaller computational scope. As a result, the total parameters dropped from 2.58 to 2.48 million and GFLOPs slightly decreased from 6.3 to 6.2, while inference speed rises from 406 FPS to 427 FPS. The detection accuracy also improved, with mAP50 rising to 0.945 and mAP50–95 to 0.863, indicating that refined multiscale feature aggregation achieves a better balance between model compactness and precision.

Introducing the PRDC module yielded another used improvement. The mAP50 increased from 0.945 to 0.953, and mAP50–95 from 0.863 to 0.870, while precision and recall rise to 0.945 and 0.901, respectively. The added computational cost was modest: parameters increased from 2.48 to 2.63 million, GFLOPs from 6.2 to 6.4, and inference speed decreased slightly from 427 to 398 FPS. The parameter growth primarily came from the newly introduced 3 × 3 dilated convolution kernels that replace the parameter-free max-pooling operations in the original SPPF module. Each PRDC branch applies the same 3 × 3 convolution weights under different dilation rates, thereby expanding the receptive field without a proportional increase in parameters. The remaining 1 × 1 compression and fusion layers are retained from SPPF for channel alignment and feature integration, contributing only marginally to the total increase. Overall, the PRDC design converts the static pooling pyramid into a learnable multiscale receptive field extractor, enhancing contextual feature capture and improving detection accuracy with minimal computational overhead.

To further examine the sensitivity of dilation configurations, several combinations were compared under identical settings (Table 5). The (1, 3, 5) configuration gave the highest accuracy, with mAP50 of 0.953 and mAP50–95 of 0.870, outperforming the narrower (1, 2, 3) and wider (1, 5, 7) settings. This indicates that (1, 3, 5) provides a balanced receptive-field distribution, large enough to capture coarse lesion context without degrading local details, offering the most stable trade-off between accuracy and efficiency.

Finally, adding the DEDH module to the detection head provided the best overall performance. The model reaches an mAP50 of 0.956 and an mAP50–95 of 0.877, while the parameters decrease to 2.31 million and the inference speed rises to 482 FPS. Precision and recall also improve to 0.953 and 0.904, respectively. The reduction in parameters was mainly because DEDH uses a unified 1 × 1 projection to 256 channels at each scale, which narrows the feature width before detection. Also, the detail-enhancement weights were uniformly shared across detection scales rather than redundantly learned for each, and the original, heavier decoupled heads were replaced with lightweight 1 × 1 convolutional layers to further reduce redundancy while maintaining detection precision.

Table 5 Performance comparison of the progressive receptive field via dilated convolutions module with different dilation configurations

Dilation	Parameters (millions)	GFLOPs	Precision	Recall	mAP50	mAP50–95
(1,2,3)	2.63	6.4	0.942	0.896	0.949	0.867
(1,3,5)	2.63	6.4	0.945	0.901	0.953	0.870
(1,2,5)	2.63	6.4	0.943	0.899	0.951	0.869
(1,5,7)	2.63	6.4	0.941	0.897	0.950	0.866

Note: The dilation configuration (d1, d2, d3) denotes the dilation rates of the three sequential dilated convolution layers in the PRDC module, where d1, d2, and d3 correspond to the first, second and third dilated convolution layers, respectively. For example, (1,2,3) means the first layer uses dilation 1, the second uses dilation 2, and the third uses dilation 3. GFLOPs denotes the number of floating-point operations, measured in giga floating-point operations. Precision and Recall represent the detection precision and recall, respectively.

To verify that the gains arise from more direction- and edge-consistent representations rather than from box scoring alone, we evaluated image-wide RDAS on the Detect-P3 fused features over 783 images, where Detect-P3 denotes the fused P3-scale feature map output by the neck and used as input to the detection head. As shown in Table 6, replacing the original detection head with DEDH provided consistent increases across all four metrics: RDAS-H rises from 0.024 to 0.057, RDAS-V from 0.014 to 0.060, RDAS-D from 0.053 to 0.087 and RDAS-Edge from 0.007 to 0.059, with comparable or reduced variability. Statistical tests confirmed that all these improvements were significant, with the largest gains for V and Edge, indicating tighter boundary alignment and stronger sensitivity to horizontally distributed lesion textures. Given that both models were evaluated at the same Detect-P3 feature level, the comparison remains controlled; although this stage precedes the differential branches, end-to-end optimization shapes upstream features, so Detect-P3 RDAS remains indicative of the enhancements introduced by DEDH.

Overall, the ablation results demonstrate that each proposed module provides consistent and complementary performance gains at different stages of the model (Table 7). Collectively, these enhancements lead to a 2.1% increase in mAP50 and a 2.9% increase in mAP50–95. These findings confirm the comprehensive advantages of the proposed method in accuracy, efficiency, and real-time performance, and further validate its practical applicability and scalability for complex agricultural disease detection scenarios.

Table 6 Relative directional alignment score (RDAS) on Detect-P3 for 783 test images

Metric	Mean *A*	Std *A*	Mean *B*	Std *B*
$RDAS_H$	0.024	0.256	0.057	0.216
$RDAS_V$	0.014	0.239	0.060	0.191
$RDAS_D$	0.053	0.242	0.087	0.200
$RDAS_{Edge}$	0.007	0.190	0.059	0.171

Note: Detect-P3 refers to the fused P3-scale feature map fed into the detection head. *A* and *B* denote the baseline model with the original detection head and the model equipped with the proposed detail-enhanced detection head, respectively. Mean and std indicate the dataset average and standard deviation across images.

3.2 Comparative experiments

To further validate the comprehensive advantages of the proposed model in terms of lightweight design and detection performance, we conducted comparative experiments against several mainstream lightweight detectors, including YOLOv5s[32], YOLOv8s, YOLOv10n, YOLOv10s[33] and YOLOv12[34,35], as well as classical detection models such as Faster R-CNN and SSD. Table 8 summarizes the comparative performance across different detection models.

The older two-stage models such as Faster R-CNN delivered relatively high accuracy, with an mAP50 of 0.922, but their excessive parameter count (63.2 million) and high computational cost (370 GFLOPs) lead to limited inference speed and poor deployment flexibility. The SSD model has a smaller size (12.3 million parameters) but yields lower detection accuracy (mAP50 of 0.920 and mAP50–95 of 0.837).

Among the more recent one-stage lightweight detectors, YOLOv5s and YOLOv8s provided an acceptable balance between compactness and accuracy, achieving mAP50 values of 0.937 and 0.936, respectively. However, their parameter counts (9.1 and 11.2 million) and computational costs (23.8 and 28.5 GFLOPs) were relatively high compared with newer ultra-lightweight models.

Table 7 Ablation study evaluating the contributions of individual proposed modules to detection accuracy, computational complexity and inference speed

Baseline	DRIoU	C3k2-A	PRDC	DEDH	Parameter (million)	GFLOPs	FPS	Precision	Recall	mAP50	Map50–95
√					2.58	6.3	406.76	0.938	0.873	0.935	0.848
√	√				2.58	6.3	406.76	0.938	0.892	0.941	0.860
√	√	√			2.48	6.2	427.06	0.939	0.892	0.945	0.863
√	√	√	√		2.63	6.4	397.56	0.945	0.901	0.953	0.870
√	√	√	√	√	2.31	5.9	482.36	0.953	0.904	0.956	0.877

Table 8 Comparison of the proposed method with lightweight and mainstream object detectors in terms of model complexity and detection performance

Method	Parameter (million)	GFLOPs	Precision	Recall	mAP50	mAP50–95
Faster-R-CNN	63.2	370	0.923	0.853	0.922	0.845
SSD	12.3	63.2	0.925	0.856	0.920	0.837
YOLOv5s	9.1	23.8	0.940	0.876	0.937	0.851
YOLOv8s	11.2	28.5	0.941	0.877	0.936	0.852
YOLOv10n	2.26	6.5	0.939	0.870	0.932	0.845
YOLOv10s	7.22	21.4	0.945	0.879	0.943	0.862
YOLO11n (baseline)	2.58	6.3	0.938	0.873	0.935	0.848
YOLO12	2.51	5.8	0.947	0.874	0.939	0.849
Ours	2.31	5.9	0.953	0.904	0.956	0.877

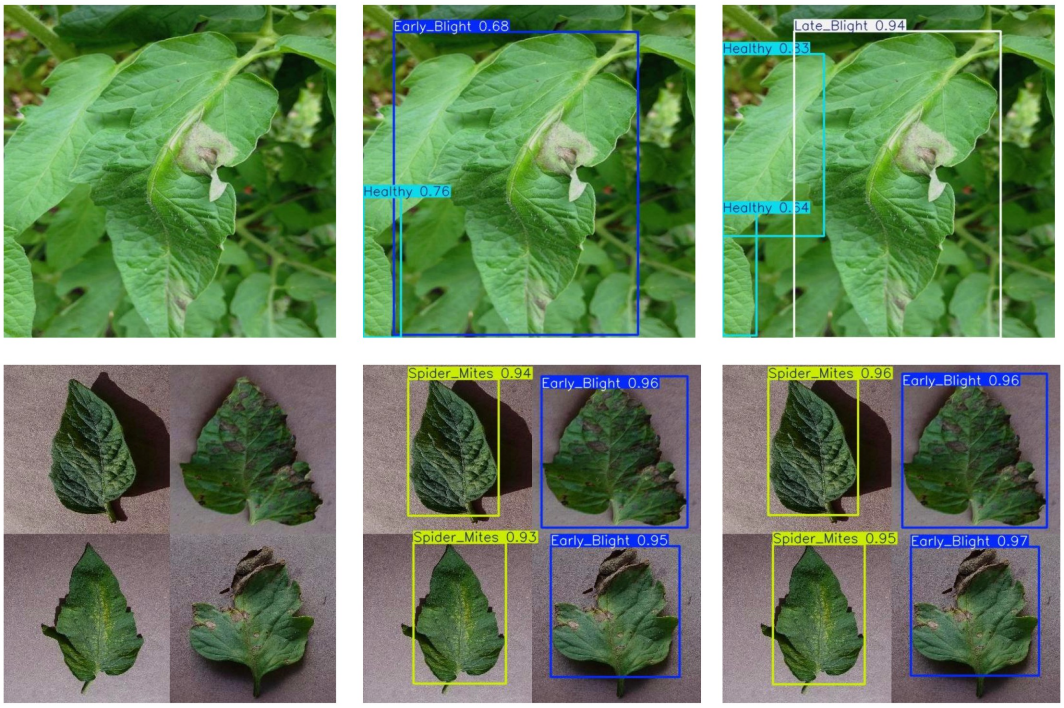
YOLOv10n and YOLOv10s further reduce complexity and achieved comparable results, but they still fell short in comprehensive accuracy-efficiency trade-offs. YOLOv12[36,37] uses an attention-centered backbone and achieves an mAP50 of 0.939 and an mAP50–95 of 0.849 with 2.51 million parameters and 5.8 GFLOPs. Its weaker spatial priors make it more data-dependent and less effective for small or boundary-blurred lesions. This structural limitation reduces its overall generalization compared with the proposed convolution-based design.

In contrast, the proposed model achieves an mAP50 of 0.956 and an mAP50–95 of 0.877 with only 2.31 million parameters and 5.9 GFLOPs. Precision and recall reached 0.953 and 0.904, respectively, and a peak inference speed of 482 FPS, substantially higher than the other methods evaluated.

These results demonstrate that the proposed method delivers high detection accuracy with extremely low computational overhead, combining high precision, low latency, and strong scalability. Such characteristics make it well suited for real-time agricultural disease detection and deployment on resource-constrained edge devices.

To visually demonstrate the practical effectiveness of the proposed method in tomato leaf disease detection tasks, Fig. 11 provides a comparison of detection results on representative samples, including the original image, results from the YOLO11n baseline model and those from the proposed approach.

In the first sample, the YOLO11n model gave misclassification and false detection, mistakenly identifying a late blight lesion as early blight, and failure to detect a healthy region. In contrast, the proposed method accurately detected the late blight lesion with a confidence score of 0.94 and its bounding box aligns closely with the actual diseased area, effectively reducing both false positives and missed detections. In this second sample, the



(a) Original Image (b) Yolov11n (c) Proposed Model

Fig. 11 Visual comparison of detection results between the YOLO11n baseline and the proposed method on representative tomato leaf disease images. From left to right, the columns show the original input images, detection results of the YOLO11n baseline and detection results of the proposed method. The proposed method demonstrates more accurate localization and improved recognition of small-scale and blurred-boundary disease regions.

proposed method gave higher confidence scores in detecting both spider mites and late blight compared to YOLO11n, further enhancing the precision of disease localization.

These comparisons indicate that the proposed multiscale feature fusion and detail-enhanced detection strategies significantly improve the ability of the model to perceive small lesions and regions with blurred boundaries, substantially reducing false detection and omission rates. The method provides superior accuracy and robustness under complex background conditions.

4 Conclusions

To address the challenges of excessive model parameters, limited detection accuracy and poor adaptability to complex agricultural environments in tomato leaf disease detection, this study developed an efficient and lightweight detection method based on improvements to the YOLO11n architecture.

This proposed method introduces several key innovations: the C3k2-A module integrates an AMSF mechanism to enhance feature richness and scale adaptability; the PRDC module replaces the traditional SPPF structure to significantly improve spatial modeling of lesions at different scales; the DEDH detection head strengthens the ability of the model to detect small lesions and blurred boundaries; and the improved loss function, DRIoU, is proposed to dynamically adjust regression gradients across different training stages.

Extensive experiments on the improved dataset demonstrated the effectiveness of the proposed approach. With only 2.31 million parameters and 5.9 GFLOPs, the model achieves 0.956 mAP50, 0.877 mAP50–95, 0.953 precision and 0.904 recall,

while maintaining an inference speed of 482 FPS. Compared with mainstream lightweight detectors such as YOLOv5s, YOLOv8s and YOLOv10n, the proposed model delivers superior performance across accuracy, speed, and model compactness. Ablation evaluation further confirmed both the independent contributions and the synergistic benefits of each introduced module.

In summary, the method presented in this paper achieves a strong balance between detection accuracy, real-time inference and deployment efficiency. It offers a practical and scalable solution for robust tomato leaf disease detection under real-world conditions, and provides valuable guidance and technical support for future developments in intelligent agricultural disease monitoring.

Acknowledgements

This work was fully supported by the National Natural Science Foundation of China (52072412).

Compliance with ethics guidelines

Qian Yuan and Hui Liu declare that declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Kibriya H, Rafique R, Ahmad W, Adnan S M. Tomato leaf disease detection using convolution neural network. In: Proceedings of the 2021 International Bhurban Conference on Applied Sciences and Technology (IBCAST). Islamabad, Pakistan: *IEEE*, 2021, 346–351
2. Wei W B, Rui X T. Study on edge detection method. *Computer Engineering and Applications*, 2006, **42**(30): 88–91 (in Chinese)
3. Sanida M V, Sanida T, Sideris A, Dasygenis M. An efficient hybrid CNN classification model for tomato crop disease. *Technologies*, 2023, **11**(1): 10
4. Jia J Z, Zhang T Z. Typical diseases and control measures of tomato in sunlight greenhouses. *Contemporary Horticulture*, 2020, **43**(1): 173–174 (in Chinese)
5. Guo W J, Feng Q, Li X Z. Research progress of convolutional neural network model based on crop disease detection and recognition. *Journal of Chinese Agricultural Mechanization*, 2022, **43**(10): 157–166 (in Chinese)
6. Patil M A, Manohar M. Enhanced radial basis function neural network for tomato plant disease leaf image segmentation. *Ecological Informatics*, 2022, **70**: 101752
7. Shao M Y, Zhang J H, Feng Q, Chai X J, Zhang N, Zhang W R. Research progress of deep learning in detection and recognition of plant leaf diseases. *Smart Agriculture*, 2022, **4**(1): 29–46 (in Chinese)
8. Khalid M M, Karan O. Deep learning for plant disease detection. *International Journal of Mathematics, Statistics, and Computer Science*, 2024, **2**: 75–84
9. Cubero S, Aleixos N, Moltó E, Gómez-Sanchis J, Blasco J. Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food and Bioprocess Technology*, 2011, **4**(4): 487–504
10. Adnan M A, Ahammed F, Rahman M F, Alam N, Prince I A, Ahamed M T. Raspberry Pi-powered IoT device for blind navigation using MobileNetV3-SSD image processing. In: Proceedings of International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). Dhaka, Bangladesh: *IEEE*, 2023, 274–279
11. Fei L K, Lu G M, Jia W, Teng S H, Zhang D. Feature extraction methods for palmprint recognition: a survey and evaluation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(2): 346–363
12. Durga Bhavani K, Ferni Ukrit M. Design of inception with deep convolutional neural network based fall detection and classification model. *Multimedia Tools and Applications*, 2024, **83**(8): 23799–23817
13. Attallah O. Tomato leaf disease classification via compact convolutional neural networks with transfer learning and feature selection. *Horticulturae*, 2023, **9**(2): 149
14. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: *IEEE*, 2016, 779–788
15. 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 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: *Springer*, 2016, 21–37

16. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: *IEEE*, 2014, 580–587
17. Shoaib M, Shah B, Ei-Sappagh S, Ali A, Ullah A, Alenezi F, Gechev T, Hussain T, Ali F. An advanced deep learning models-based plant disease detection: a review of recent research. *Frontiers in Plant Science*, 2023, **14**: 1158933
18. Nagamani H S, Sarojadevi H. Tomato leaf disease detection using deep learning techniques. *International Journal of Advanced Computer Science and Applications*, 2022, **13**(1): 305–311
19. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, **20**(3): 273–297
20. Yang R, Lu X Y, Huang J, Zhou J, Jiao J, Liu Y F, Liu F, Su B F, Gu P W. A multi-source data fusion decision-making method for disease and pest detection of grape foliage based on ShuffleNet V2. *Remote Sensing*, 2021, **13**(24): 5102
21. Chu X, Li X, Luo B, Wang X D, Huang S. Identification method of tomato leaf diseases based on improved YOLOv4 algorithm. *Jiangsu Journal of Agricultural Sciences*, 2023, **39**(5): 1199−1208 (in Chinese)
22. Jing J P, Li S F, Qiao C, Li K Y, Zhu X Y, Zhang L X. A tomato disease identification method based on leaf image automatic labeling algorithm and improved YOLOv5 model. *Journal of the Science of Food and Agriculture*, 2023, **103**(14): 7070–7082
23. Xue X, Liu P, Zhou W. Detection of real-time fine-grained plant disease based on improved YOLOv8 algorithm. *Journal of Chinese Agricultural Mechanization*, 2024, **45**(5): 188−194 (in Chinese)
24. Varghese R, Sambath M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness. In: Proceedings of the International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Chennai, India: *IEEE*, 2024, 1–6
25. Andrushia A D, Patricia A T. Artificial bee colony optimization (ABC) for grape leaves disease detection. *Evolving Systems*, 2020, **11**(1): 105–117
26. Alessandrini M, Rivera R C F, Falaschetti L, Pau D, Tomaselli V, Turchetti C. A grapevine leaves dataset for early detection and classification of esca disease in vineyards through machine learning. *Data in Brief*, 2021, **35**: 106809
27. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv, 2024: 2410.17725
28. Luo X J, Cai Z H, Shao B, Wang Y X. Unified-IoU: for high-quality object detection. arXiv, 2024: 2408.06636
29. Zheng Z H, Wang P, Liu W, Li J Z, Ye R G, Ren D W. Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, NY, USA: *AAAI*, 2020, 12993–13000
30. Zhang Y F, Ren W Q, Zhang Z, Jia Z, Wang L, Tan T N. Focal and efficient IoU loss for accurate bounding box regression. *Neurocomputing*, 2022, **506**: 146–157
31. Gevorgyan Z. SIoU loss: more powerful learning for bounding box regression. arXiv, 2022: 2205.12740
32. Khanam R, Hussain M. What is YOLOv5: a deep look into the internal features of the popular object detector. arXiv, 2024: 2407.20892
33. Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G, Ding G G. YOLOv10: real-time end-to-end object detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: *Curran Associates Inc*., 2024, 3429
34. Tian Y J, Ye Q X, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv, 2025: 2502.12524
35. Tian Y J, Xie L X, Qiu J H, Jiao J B, Wang Y W, Tian Q, Ye Q X. Fast-iTPN: integrally pre-trained transformer pyramid network with token migration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 9766–9779
36. Dao T. FlashAttention-2: faster attention with better parallelism and work partitioning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: *ICLR*, 2024
37. Fang Y X, Wang W, Xie B H, Sun Q, Wu L, Wang X G, Huang T J, Wang X L, Cao Y. EVA: exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: *IEEE*, 2023, 19358–19369