Lightweight Real-Time Detection Transformer for Tomato Leaf Disease Recognition in Complex Agricultural Scenarios

Jianbo Lu
Nanning Normal University
Yang Liu
Nanning Normal University
Miaomiao Yu
yumm602@163.com
Guangxi Economic and Trade Vocational Institute
Guangliang Wei
Ltd. Nanning

Article

Keywords: disease recognition, RT-DETR, small object detection, lightweight, robustness

Posted Date: June 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9832324/v1

Additional Declarations: No competing interests reported.

Lightweight Real-Time Detection Transformer for Tomato Leaf Disease Recognition in Complex Agricultural Scenarios

Jianbo Lu[1,2], Yang Liu[2], Miaomiao Yu[3,*], Guangliang Wei[4,*]

[1] Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Nanning Normal University, Nanning, 530001, Guangxi, China.
[2] School of Artificial Intelligence, Nanning Normal University, Nanning, 530100, Guangxi, China.
[3] School of Art Design and Architecture, Guangxi Economic and Trade Vocational Institute, Nanning, 530021, Guangxi, China. Email: yumm602@163.com
[4] Guangxi Huiyun Information Technology Co., Ltd. Nanning, 530007, Guangxi, China. Email: brightwei202604@163.com

Abstract: Accurate detection of tomato leaf diseases is essential for sustainable tomato production. To overcome the limitations of existing detection models, such as large parameter sizes, insufficient accuracy, weak robustness, and poor small-target detection performance, this study proposes a lightweight tomato leaf disease detection algorithm based on an improved Real-Time Detection Transformer (RT-DETR), named Tomato Leaf Diseases RT-DETR (TLD-RTDETR). Specifically, a partial convolution-based PConvBlock is introduced into the backbone to enhance feature extraction while reducing model complexity. In addition, a Coordinate Attention-based hierarchical feature pyramid module (CA_HSFPN) is designed to suppress background interference and strengthen small-target feature representation. Furthermore, a learnable positional encoding strategy is integrated into the feature encoding stage to improve the extraction of critical disease features in complex environments. Experimental results show that TLD-RTDETR achieves an mAP of 92.9%, precision of 94.2%, and recall of 87.5% on the tomato leaf disease dataset, outperforming the RT-DETR-R18 baseline by 2.4%, 0.5%, and 3.3%, respectively. Meanwhile, the model size, parameter count, and computational cost are reduced by 38.1%, 38.1%, and 31.2%. Compared with mainstream methods, the proposed model achieves better detection performance with a more lightweight architecture. Additional visualization, anti-interference, and generalization experiments further verify its robustness and cross-scene adaptability, demonstrating its potential for practical deployment in tomato leaf disease detection.

Keywords: disease recognition; RT-DETR; small object detection; lightweight; robustness

1. Introduction

Tomato is one of the most widely cultivated vegetable crops worldwide due to its high nutritional value and strong market demand. However, tomato leaves are highly susceptible to diseases during growth, which can significantly reduce plant health and yield. Studies have shown that crop diseases cause substantial agricultural economic losses worldwide, particularly in Asia and Africa, where loss rates exceed 50% [1]. Therefore, rapid and accurate identification of tomato leaf diseases is essential for effective disease prevention and precision management.

Traditional tomato leaf disease detection mainly relies on manual inspection by experts, which is time-consuming, labor-intensive, and easily affected by environmental conditions. As a result, it cannot satisfy the demands of modern agriculture for efficient and precise disease monitoring. With the rapid development of deep learning, especially convolutional neural networks (CNNs), deep learning-based object detection methods have gradually become the mainstream approach for crop disease detection. Among these approaches, Sun et al. **Error! Unknown switch argument.** introduced a YOLOv8-based tomato disease detection model named YOLO-FMDI. Through architectural refinements, the model lowers computational complexity while improving detection capability, leading to reductions of 18.2% in the number of parameters, 6.1% in billion floating-point operations (GFLOPs), and 15.9% in model size, along with a 1.2% increase in mAP. Yan et al. **Error! Unknown switch argument.** developed the FSM-YOLO model, which enhances the performance of apple leaf disease detection, with mAP, precision, and recall improving by 2.7%, 2.0%, and 4.0%. Liu et al. **Error! Unknown switch argument.** refined YOLOX-Tiny for grape object detection and introduced the YOLOX-RA model. Despite adopting a lightweight configuration, the model achieves a detection speed of 84.88 frames per second (FPS) with a size of 17.53 MB; nevertheless, its accuracy remains limited at 88.75%. Wang et al. **Error! Unknown switch argument.** proposed the BED-YOLO model based on YOLOv10n, which improves tomato disease detection performance in complex field environments, increasing the mAP by 3.9%; however, the number of parameters and billion floating-point operations increase by approximately 60.9% and 46.3%, respectively. Sun et al. **Error! Unknown switch argument.** introduced the Eggplant-DETR model for eggplant disease and pest detection. By incorporating a shallow feature enhancement module based on Context-aware Parallel Squeeze-and-Excitation (CPSE) and a Cross-level Hierarchical Selective Feature Pyramid Network (CHSFPN), the model attains a highly lightweight structure, reducing the number of parameters to 15.77M; nonetheless, its mAP remains relatively low at 77.8%. Wang et al. **Error! Unknown switch argument.** developed the CGM-DETR model for cocoa disease and pest detection. By integrating a context-guided module and a dynamic feature fusion module, the model realizes a lightweight design, decreasing the number of parameters by 33.9%; this reduction is achieved at the cost of performance, with the mAP remaining at only 87%. Zhang et al. **Error! Unknown switch argument.** proposed a lightweight tea disease detection model, WMC-RTDETR. By enhancing feature extraction and attention mechanisms, the model cuts down the number of parameters and computational cost by 35.48% and 40.42%. Wang et al. **Error! Unknown switch argument.** presented the RP-DETR model for rice pest detection. By optimizing the backbone network and feature fusion structure, the model achieves a lightweight architecture, reducing the number of parameters by 25.8%; however, its mAP remains at 76.9%, suggesting significant room for further improvement in detection accuracy. In summary, although existing studies have achieved significant progress in crop disease detection tasks, they still struggle to achieve a satisfactory balance between lightweight model design and detection accuracy. Moreover, in real-world field scenarios, under complex conditions such as drastic illumination variations, background interference, and target occlusion, existing models exhibit limited robustness in detecting small-scale disease targets, which often leads to missed detections and false positives. This paper offers an enhanced lightweight method for detecting tomato leaf diseases, namely TLD-RTDETR, with the objective of achieving an optimal equilibrium between detection efficiency, accuracy, and localization precision. This study's primary contributions are summarized as follows:

(1) To achieve model lightweighting while maintaining feature extraction capability, this study introduces Partial Convolution (PConv) into the BasicBlock to construct a lightweight PConvBlock

module. Specifically, PConv replaces the second standard convolution layer in the residual structure, and an additional nonlinear activation function is incorporated. The reconstructed model reduces parameters, computational cost, and model size by 29.6%, 24.9%, and 29.0%, respectively.

(2) To enhance small-object detection and suppress background interference, a Coordinate Attention (CA) mechanism and a High-level Selection (HS) strategy are integrated into the feature pyramid network of RT-DETR-R18 to construct the CA_HSFPN adaptive feature fusion module. By employing direction-aware spatial encoding and gated weight generation, the proposed module effectively improves feature propagation while reducing computational cost. The improved model achieves a 2.2% increase in mAP50-95, with the parameter size reduced to 12.21 MB.

(3) To improve feature localization under complex backgrounds, a learnable positional encoding strategy is introduced into the feature encoding stage to construct the AIFI_LPE module. Replacing the original fixed positional encoding with an end-to-end learnable representation enhances the positional perception capability of the self-attention mechanism while maintaining computational efficiency. As a result, the model accuracy is improved to 94.2%.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes the proposed TLD-RTDETR algorithm, Section 4 presents experimental results and analysis, and Section 5 concludes the paper.

2 Related Work

2.1 Crop Leaf Disease and Pest Recognition

Agriculture plays a vital role in global food security and economic stability. However, crops are highly vulnerable to diseases and pests during growth, causing significant yield and quality losses. Since early symptoms usually appear on leaves, such as spots, chlorosis, and curling, accurate detection of crop leaf diseases and pests is essential for early warning and effective prevention and control.

Early studies on crop leaf disease and pest recognition primarily relied on traditional computer vision techniques and machine learning algorithms, in which handcrafted image features were designed to achieve automatic identification and classification. Typical machine learning approaches include artificial neural networks **Error! Unknown switch argument.**, support vector machines (SVM) **Error! Unknown switch argument.**, and k-nearest neighbors (KNN) **Error! Unknown switch argument.**. Gurunathan et al. **Error! Unknown switch argument.** proposed a peanut leaf disease recognition method based on a KNN classifier. After extracting image features, the KNN algorithm was employed for classification, achieving a recognition accuracy of 98.46% on the peanut leaf dataset. Sachdeva et al. **Error! Unknown switch argument.** developed a plant leaf disease recognition method incorporating Bayesian learning. By introducing a Bayesian framework to enhance feature representation, the method achieved a recognition accuracy of 98.9% on the PlantVillage dataset. Bhola et al. **Error! Unknown switch argument.** introduced a multi-crop leaf disease recognition method that integrates deep transfer learning with an SVM. By extracting image features from leaf samples and employing the SVM for classification, the method achieved superior performance compared to the baseline model on maize, wheat, and rice leaf datasets. Although the aforementioned traditional methods have achieved satisfactory performance in specific scenarios, the inherent limitations of handcrafted feature selection make it difficult to comprehensively capture the complex and diverse details of disease images, thereby constraining the overall performance and generalization capability of the algorithms.

In recent years, with the continuous advancement of deep learning technologies, new research directions have emerged in the field of crop leaf disease and pest recognition, and related methods have gradually become a major focus of research in this area. Gong et al. **Error! Unknown switch argument.** improved the Faster R-CNN model by incorporating Res2Net and a feature pyramid network to enhance multi-scale feature extraction. In addition, RoIAlign and Soft Non-Maximum Suppression (Soft-NMS) were adopted to improve localization and detection performance, enabling effective recognition of apple leaf diseases under complex background conditions. Alruwaili et al. **Error! Unknown switch argument.** introduced an RTF-RCNN-based tomato leaf disease detection method built upon Faster R-CNN. By integrating static images with real-time video stream data, the method enables efficient recognition of tomato leaf diseases, achieving an accuracy of 97.42%. He et al. **Error! Unknown switch argument.** developed an improved model named MFaster R-CNN based on Faster R-CNN to address the difficulty of maize leaf disease recognition under complex backgrounds and similar lesion conditions. This method effectively improves model performance by introducing batch normalization and a hybrid loss function with a center cost, achieving a mean average precision of 97.23%. Despite the excellent accuracy of the previously discussed methods in recognizing crop leaf diseases, the extensive parameter scale and significant computing complexity of two-stage detection models hinder their deployment in practical field settings. The YOLO series [46][48], utilizing a single-stage detection framework, has emerged as a focal point in crop leaf disease detection owing to its straightforward design, rapid inference speed, and robust real-time capabilities. Meng et al. **Error! Unknown switch argument.** introduced a YOLO-MSM model for the identification of maize leaf diseases. This algorithm adopts multi-scale variable kernel convolution, the C2f-SK module, and the MPDIoU loss function, increasing precision and recall by 0.66% and 1.61%, respectively, while achieving a model size of only 5.4 MB. Xu et al. **Error! Unknown switch argument.** proposed the ALAD-YOLO apple leaf disease detection model based on YOLOv5s. By introducing MobileNet-V3s to compress the backbone network, integrating a coordinate attention mechanism, and replacing ordinary convolutions with group convolutions and depthwise convolutions in the SPPCSPC and C3 modules, the computational cost is reduced to 6.1 G and the mAP reaches 90.2%. To address the challenges of difficult small lesion recognition and limited resources in tomato leaf disease detection, Yao et al. **Error! Unknown switch argument.** proposed the WTAD-YOLO model. This model integrates wavelet transform convolution and an efficient resampling module, achieving an mAP@0.5 of 91.7% while reducing the parameter count by 10.0%. Although the above YOLO-based methods have made some progress in improving detection accuracy and model lightweighting, their reliance on post-processing mechanisms such as Non-Maximum Suppression (NMS) leads to a significant increase in computational overhead as object density rises, thereby limiting the real-time deployment capability of the models.

In contrast, RT-DETR, with its NMS-free end-to-end design, overcomes the deployment limitations of traditional algorithms and demonstrates excellent performance in real-time object detection tasks. Xu et al. **Error! Unknown switch argument.** introduced the DBPC-DETR model for grape leaf disease detection. By constructing a dual-backbone network combining PPHGNetV2 and CSPNet, together with model pruning and distillation techniques, they achieved a very high level of lightweighting. The computational cost was reduced by 71.9%, but the mean average precision (mAP) was only 87.6%. Yao et al. **Error! Unknown switch argument.** proposed a maize leaf disease detection model based on RT-DETR. By introducing the DAttention and SCConv modules to enhance the model's feature extraction and fusion, they improved the detection accuracy by 8.4% compared to the baseline RT-DETR model. Bian et al. **Error! Unknown switch argument.** constructed a multi-scale feature fusion network based on RT-DETR, achieving 97.9% mAP on the tomato leaf disease dataset, but with a parameter increase of 2.43 MB and a computational cost increase of 2.6 G. Ren et al. **Error! Unknown switch argument.** developed the RT-DETR-MC model based on RT-DETR-R18 for lettuce leaf spot detection. By restructuring the cross-scale fusion module and introducing the NWD loss function, the multi-scale feature extraction capability for overlapping lesions is enhanced, achieving 92.6% mAP@0.5, while the parameter count and computational complexity are reduced by only 3.7% and 2.2%, respectively.

Although current crop leaf disease detection models have advanced in enhancing detection accuracy and attaining a lightweight design, they have yet to achieve an acceptable equilibrium between the two. This study proposes a lightweight tomato leaf disease detection algorithm, TLD-RTDETR, based on an enhanced RT-DETR-R18, to tackle the challenges of excessive parameter counts, inadequate detection accuracy, limited robustness, and deficient small-target recognition in current models.

2.2 RT-DETR Model

In the field of object detection, CNN-based detection methods, such as the YOLO series **Error! Unknown switch argument.**, although achieving a good balance between detection speed and accuracy, typically rely on post-processing steps like NMS, which limits the end-to-end optimization capability of the models. In response, researchers have begun exploring transformer-based object detection methods. Among them, the end-to-end detection framework represented by DETR **Error! Unknown switch argument.** achieves an NMS-free object detection pipeline through set prediction and the Hungarian matching strategy.

The original DETR suffers from slow convergence and high training costs, which limits its application in real-time scenarios. Because of this, Baidu proposed the real-time detection transformer model RT-DETR **Error! Unknown switch argument.**. By optimizing the query mechanism and feature encoding approach, it significantly improves the convergence speed and inference efficiency while preserving the advantages of end-to-end detection, achieving a better balance between detection accuracy and real-time performance. Given the requirements of lightweight design and real-time performance for edge deployment in complex field environments, this paper selects RT-DETR-R18, which is based on a lightweight backbone network, as the baseline model for tomato leaf disease detection. On the basis of balancing accuracy and efficiency, the model further improves lightweightness, robustness, and small-target detection capability. The overall architecture is shown in Figure 1.

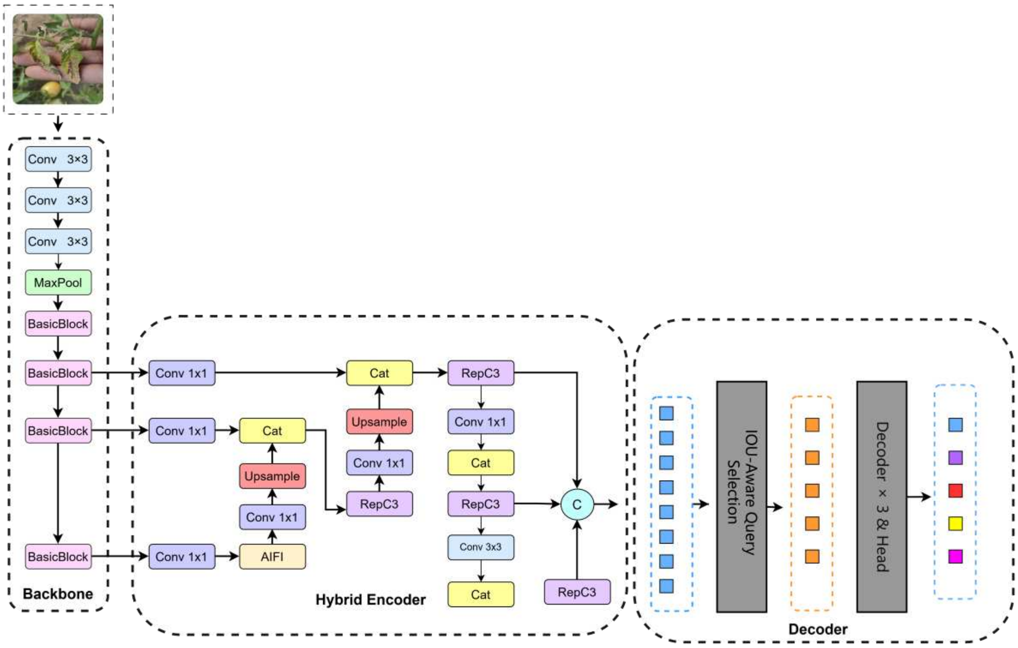


Figure 1 Architecture of the RT-DETR-R18 model

The architecture of RT-DETR-R18 mainly consists of three components: a backbone network, a hybrid encoder, and a decoder. Among them, the backbone network adopts ResNet18, a residual network responsible for extracting multi-scale features from the input image. The efficient hybrid encoder consists of an attention-based intra-scale feature interaction (AIFI) module and a cross-scale feature fusion module (CCFM). The AIFI module operates only on the high-level features (S5) output by the backbone network, using a self-attention mechanism to capture intra-scale feature correlations. The CCFM module, on the other hand, leverages a convolutional neural network (CNN) architecture to effectively fuse features from different levels **Error! Unknown switch argument.**. In the decoding stage, the model employs an IoU-aware query selection mechanism to filter a fixed number of initial object queries from the features. These queries are then iteratively refined by the Transformer decoder, and finally, the model directly outputs the target class scores and bounding box coordinates.

3. Improved RT-DETR

The precise identification of tomato leaf diseases is essential for effective agricultural management and control. Despite RT-DETR-R18 exhibiting superior performance in end-to-end object detection, it encounters the challenge of reconciling accuracy and efficiency in intricate agricultural field contexts: complex backgrounds and minuscule lesions compromise model robustness, resulting in false positives and overlooked detections; concurrently, its substantial parameter count obstructs the streamlined deployment of the model on agricultural edge devices. This paper presents a lightweight detection model, termed Tomato Leaf Diseases RT-DETR (TLD-RTDETR), based on an enhanced RT-DETR-R18. The model seeks to attain an optimal equilibrium among lightweight design, detection precision, and resilience. The whole network architecture is illustrated in Figure 2, with the individual improved modules delineated by red dashed boxes.

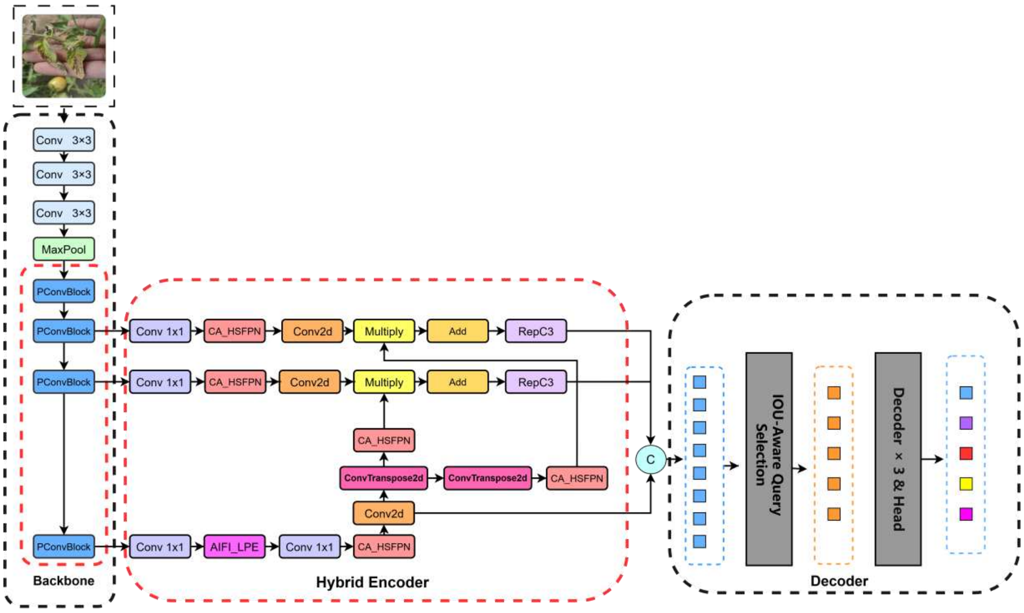


Figure 2 Architecture of the TLD-RTDETR model

In the backbone stage, the TLD-RTDETR model introduces partial convolution to construct a lightweight PConvBlock module, replacing the original basic residual blocks. By reducing spatial computational redundancy, this structure efficiently extracts multi-scale features of tomato leaf diseases while significantly reducing the model's parameter count and computational complexity. In the neck network stage, a hierarchical feature pyramid module incorporating a coordinate attention mechanism (CA_HSFPN) is designed. This module combines a coordinate attention mechanism and a high-level feature selection strategy, performing feature enhancement in both spatial and channel dimensions. While suppressing background noise interference in agricultural fields, it strengthens the feature representation of extremely small-scale diseases at different scales, effectively reducing false alarms and missed detections, thereby improving the model's localization accuracy and robustness in complex environments. In the feature encoding stage, the AIFI_LPE module is innovatively introduced. This module adaptively learns the spatial position representation of feature maps, thereby improving the model's ability to accurately identify key disease features under complex backgrounds.

3.1 Improved PConvBlock Module

In the task of precise recognition of tomato leaf diseases, the original backbone of the RT-DETR-R18 model, due to the extensive use of standard convolution operations, often incurs high spatial computational redundancy, resulting in excessive model parameter count and computational complexity. To achieve lightweight deployment of the model while maintaining efficient feature extraction capability in complex agricultural scenarios, this study introduces partial convolution into the original BasicBlock for reconstruction, proposing an improved lightweight PConvBlock module. Its specific structure is shown in Figure 3.

As shown in Figure 4, partial convolution **Error! Unknown switch argument.** performs regular convolution operations on a subset of channels of the disease feature map. It effectively extracts spatial features while significantly reducing computational redundancy and the number of memory accesses. By reducing the parameter count and computational complexity while enhancing the model's feature representation capability, it meets the requirements of lightweight design and accurate recognition for the task of tomato leaf disease detection. Specifically:

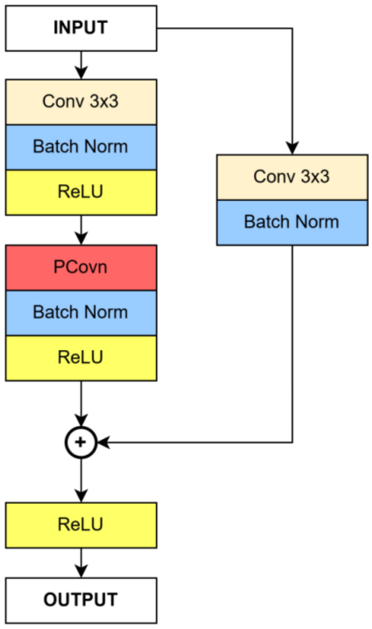


Figure 3 Structure of the PConvBlock module

First, for an input feature map $I \in \mathbb{R}^{c\times \mathrm{h} \times w}$, PConv applies regular convolution only to a subset of its continuous channels I_{c_p}, while leaving the remaining channels I_{c-c_p} unchanged. Here,$c_p = r \cdot c$, where r is the partial ratio, typically set to 1/4. This operation can be expressed as:

$$PConv(I) = Conv_{k\times k}(I_{c_p}) \oplus I_{c-c_p} \tag{1}$$

Where $Conv_{k\times k}$ denotes a regular convolution with a kernel size of $k\times k$ and $\oplus$ represents the concatenation operation along the channel dimension. The floating-point operations (FLOPs) of PConv are:

$$FLOPs = h\times w\times k^2 \times c_p \tag{2}$$

which is only 1/4 of that of regular convolution (when r = 1/4). Meanwhile, its memory access is shown in Equation (3), which is 1/4 of that of regular convolution:

$$Memory_{PConv} \approx h\times w\times 2c_p \tag{3}$$

To further fuse information from all channels, a pointwise convolution (PWConv) is applied after PConv, forming a T-shaped receptive field structure. The overall computation is shown in Equation (4):

$$O = PWConv(PConv(I)) \tag{4}$$

Where $O \in \mathbb{R}^{c\times \mathrm{h} \times w}$ is the output feature map. This structure, while focusing on the central positional features, fully leverages cross-channel information to enhance the completeness and discriminability of feature representation.

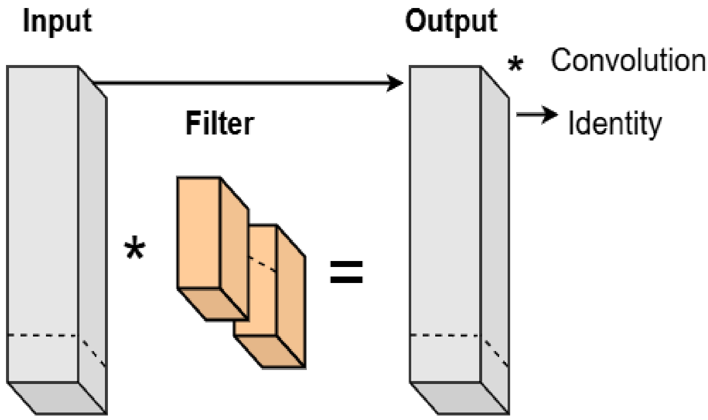


Figure 4 Structure of partial convolution (PConv)

3.2 Learnable Positional Encoding Module

In the task of tomato leaf disease detection, complex field backgrounds often share similarities in color or texture with the lesion regions, which imposes high demands on the model's ability to accurately identify key disease features. In the feature encoding stage, the original RT-DETR-R18 model adopts a positional encoding based on a fixed 2D sinusoidal mechanism **Error! Unknown switch argument.**. Although this predefined, non-trainable periodic encoding possesses translation invariance, it lacks the adaptability to specific data distributions, making it difficult for the model to effectively focus on key disease features under complex background interference, thereby affecting the recognition accuracy of the model.

This study develops the AIFI_LPE module during the feature encoding phase. This module integrates end-to-end learnable positional encoding (Learned Positional Encoding, LPE) **Error! Unknown switch argument.** into the Transformer encoding process, replacing the traditional fixed positional encoding approach. Compared with predefined positional encoding, LPE can adaptively learn the optimal spatial position representation according to the actual spatial distribution patterns of tomato lesions, thereby enhancing the model's ability to perceive and focus on key disease features under complex backgrounds, effectively improving detection performance. Its core computation process is as follows:

First, for the high-level feature map $X \in \mathbb{R}^{N \times C}$ output by the backbone network, this module constructs a trainable $2D$ parameter matrix E_{pos} as a positional prior. The matrix E_{pos} is iteratively optimized together with the loss function of the entire network during backpropagation, thereby dynamically capturing the spatial distribution characteristics of the disease. It is defined as:

$$P_{LPE}=E_{pos} \in \mathbb{R}^{N \times C} \tag{5}$$

where $N \times C$ represents the total sequence length after flattening the feature map along the spatial dimension.

After constructing the positional embedding matrix, the module flattens and transposes the input feature map X into a serialized feature $X_{seq} \in \mathbb{R}^{B \times N \times C}$ to adapt to the subsequent Transformer architecture. It is defined as:

$$X_{seq}=Reshape(X,[B,C,H \times W])^{T}, \quad N=H \times W, \quad X_{seq} \in \mathbb{R}^{B \times N \times C} \tag{6}$$

Then, the learnable positional encoding P_{LPE} is injected into the query $(\mathbf{Q})$ and key $(\mathbf{K})$ vectors via element-wise addition, while the value $(\mathbf{V})$ vector remains unchanged to enhance the model's spatial awareness.

$$Y_{seq}=LayerNorm(MHSA(Q,K,V)+X_{seq}) \tag{7}$$

Finally, the module inversely reshapes the computed 1D serialized output to restore it into a feature vector $\mathbf{Y}$ with a $2D$ spatial structure, so that it can be fed into the subsequent CA_HSFPN neck network for multi-scale fusion. The computation process of this spatial restoration is as follows:

$$\mathbf{Y}=Reshape(Y_{seq}^{T},[B,C,H,W]) \tag{8}$$

3.3 CA_HSFPN Module

For tomato leaf disease detection, different disease types exhibit significant differences in spatial scale and morphological distribution. In particular, early-stage lesions often appear as small-scale local regions with blurred boundaries. Therefore, the model requires a higher capability of multi-scale feature representation and selection. However, the cross-scale feature fusion in the original RT-DETR-R18 model relies primarily on feature concatenation and convolution operations for information aggregation, lacking an attention guidance mechanism for spatial position information and object saliency. As a result, background interferences such as leaf vein textures and light reflections in shallow features are continuously propagated or even cumulatively amplified during the fusion process, thereby weakening the model's discriminative ability for tiny lesions and making it prone to false detections and missed detections.

In light of this, this paper designs a hierarchical feature pyramid module incorporating a coordinate attention mechanism (CA_HSFPN) in the neck network stage. This module embeds Coordinate Attention (CA) **Error! Unknown switch argument.** into the multi-scale feature selection and cross-scale fusion pathways of the HSFPN. The architecture of the CA_HSFPN module is shown in Figure 5.

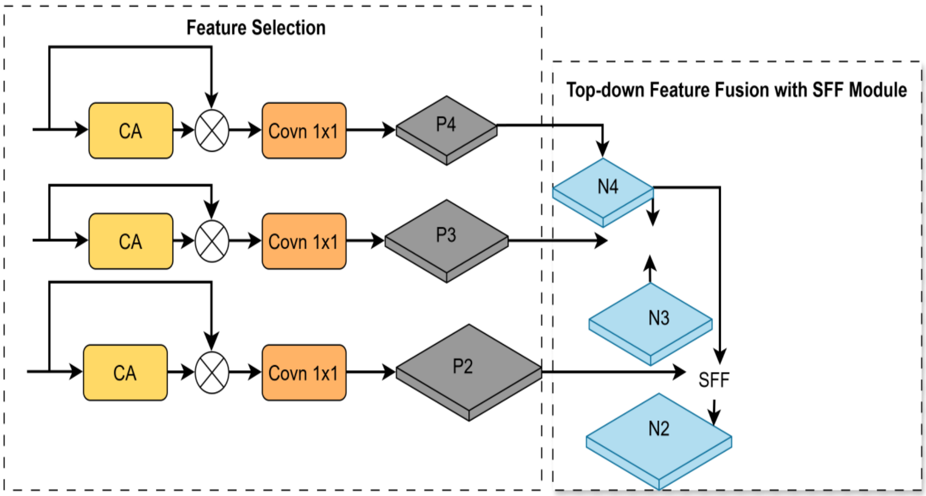


Figure 5 Structure of the CA_HSFPN module

During feature enhancement, based on the coordinate attention mechanism shown in Figure 6, the module performs global average pooling operations on the input features along the horizontal and vertical directions, decomposing the 2D feature map into a pair of directional-aware feature encodings. Subsequently, it generates attention weights for each direction through convolution fusion and nonlinear mapping. Finally, the directional attention is mapped back to the original feature space via element-wise multiplication, thereby effectively modeling the spatial position information of lesion regions under complex backgrounds. In particular:

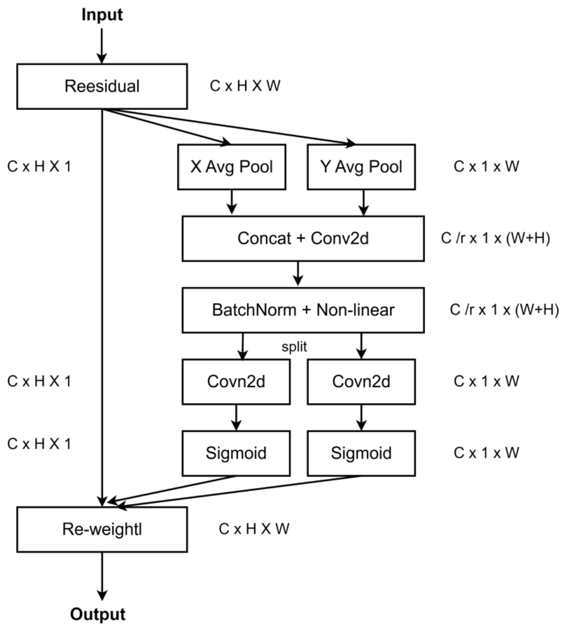


Figure 6 Diagram of coordinate attention mechanism

Given an input feature $X\in\mathbb{R}^{C\times H\times W}$, first, one-dimensional global average pooling is performed along the horizontal and vertical directions, respectively, to obtain directional feature representations:

$$z_{\mathrm{h}}(c,\mathrm{h})=\frac{1}{W}\sum_{i=1}^{W}X(c,\mathrm{h},i),\quad z_{w}(c,w)=\frac{1}{H}\sum_{j=1}^{H}X(c,j,w) \tag{9}$$

Subsequently, the two directional features are fused and then passed through a shared convolutional layer for channel compression and nonlinear mapping:

$$f=\sigma(Conv([z_{\mathrm{h}},z_{w}])) \tag{10}$$

Here,σ(·) denotes a nonlinear activation function. The fused feature is further split into two directional branches along the spatial dimension, and the attention weights are generated separately:

$$g_{\mathrm{h}}=\sigma(Conv_{\mathrm{h}}(f_{\mathrm{h}})),\quad g_{w}=\sigma(Conv_{w}(f_{w})) \tag{11}$$

Finally, the directional attention is mapped back to the original feature space via element-wise multiplication to achieve feature recalibration:

$$\mathrm{Y(c,i,j)=X(c,i,j)\cdot g_{h}(c,i)\cdot g_{w}(c,j)} \tag{12}$$

Thereby adaptively modulating the responses at different positions along the spatial dimension.

Subsequently, in the feature interaction stage, the channel dimension is compressed through a shared convolutional layer with a nonlinear mapping, generating attention weights that are sensitive to spatial positions. Finally, before the top-down feature fusion (SFF), the generated attention weights are used to adaptively recalibrate the feature maps at different scales, thereby enhancing the responses of key regions and suppressing background interference.

Unlike the original RT-DETR-R18 fusion strategy, the proposed method introduces a direction-aware attention mechanism into the feature propagation pathway, enabling spatially guided weighted fusion across different scales. This design effectively suppresses background interference, such as leaf vein textures and light reflections, while enhancing the discriminative representation of small lesion regions. At the same time, it maintains low computational overhead, achieving a balance between detection accuracy and model lightweighting.

4. Experiments

4.1 Experimental Environment and Setup

The experimental parameters are set as follows: the image resolution is 640 × 640, the batch size is set to 4, the number of epochs is set to 300, RT-DETR-R18 is used as the baseline model, and the model is pre-trained using RT-DETR-R18 weights. The experimental environment configuration is shown in Table

Table 1 Experimental Environment

Name	Environment Configuration
Operating System	Win10
Processor	Intel(R)i7-12400
GPU	NVIDIA RTX3090
Deep Learning Framework	PyTorch2.0.0

4.2 Experimental Dataset and Preprocessing

For this experiment, a tomato leaf disease dataset is used to verify the performance of the improved model. Specifically, the tomato leaf disease dataset published by the Roboflow platform **Error! Unknown switch argument.** is adopted. The dataset contains a total of 4,331 images, covering four typical disease states: spotted wilt virus, nitrogen deficiency, potassium deficiency, and healthy leaves. The detailed category division is shown in Table 2. This dataset exhibits strong diversity in lesion morphology, scale distribution, and background complexity, which can well reflect the visual characteristics of leaf diseases in real agricultural scenarios. Some sample examples are shown in Figure 7.

Table 2 Categories of the tomato leaf disease dataset

Category of Disease	Images for Training	Images for Validation	Images for Testing
Spotted Wilt Virus	1010	290	145
Nitrogen Deficiency	303	87	43
Potassium Deficiency	717	205	102
Healthy	1002	284	143
Total	3032	866	433

(a) Healthy (b) Potassium Deficiency (c) Nitrogen Deficiency (d) Spotted Wilt Virus

Figure 7 Images of the tomato leaf disease dataset

In real field environments, factors such as illumination variation, background interference, and target occlusion often reduce the robustness of disease recognition models, leading to missed and false detections. To improve model robustness and generalization in complex scenarios, this study applies diverse data augmentation strategies using Python and Albumentations. Specifically, geometric transformations, including affine transformation, random flipping, perspective transformation, and grid distortion, are employed to simulate viewpoint variations. Gaussian noise, brightness/contrast perturbation, and random shadows are introduced to mimic illumination changes and sensor noise. In addition, a semantically driven occlusion augmentation strategy is designed by randomly overlaying elements such as tomato fruits, stems, water droplets, light spots, and dust to construct more realistic complex backgrounds. Furthermore, a random mixed augmentation strategy is adopted to combine different augmentation operations and parameters, thereby increasing sample diversity and better simulating real field conditions.

After augmentation, the original dataset is expanded to 17,255 images, which helps reduce the risk of overfitting caused by data distribution bias. The training set, validation set, and test set are randomly generated in a ratio of 7:2:1. Some examples of the augmentation effect are shown in Figure 8.

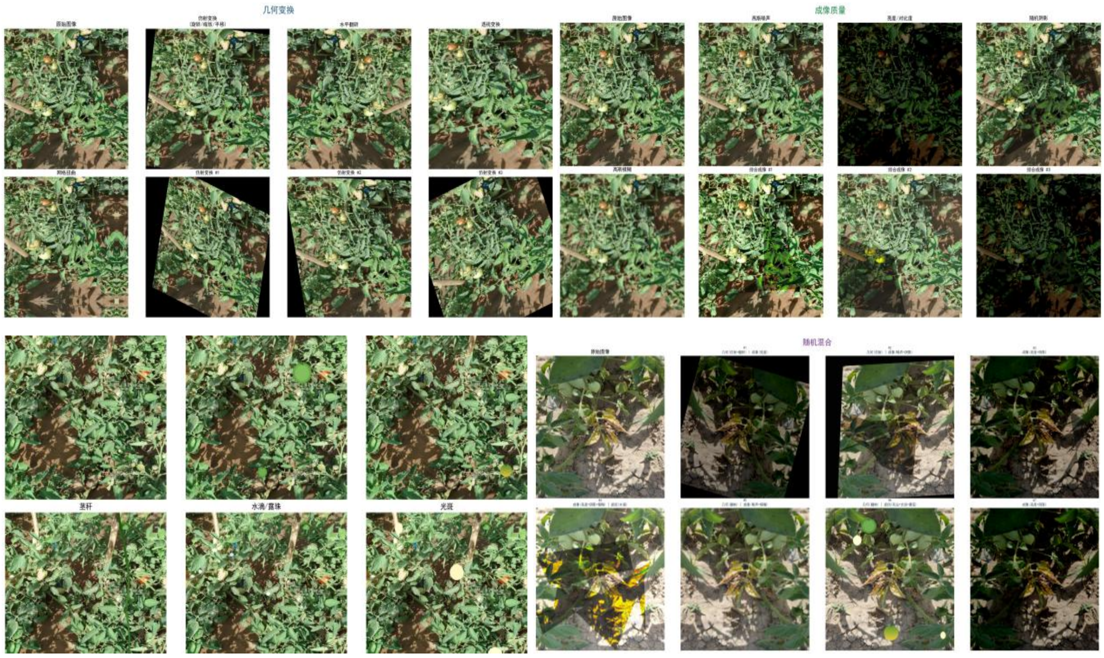


Figure 8 Examples of dataset augmentation

4.3 Evaluation Metrics

For detection performance evaluation, this paper adopts precision (P), recall (R), mAP50, and mAP50:95 as evaluation metrics to comprehensively quantify the model's recognition capability for tomato leaf diseases. Here, P denotes the proportion of predicted disease targets that are actual diseases, reflecting the decision accuracy. Recall measures the model's ability to detect actual disease regions; the higher the recall, the fewer missed detections. mAP50 is the mean of average precision (AP) across all categories when the intersection over union (IoU) threshold is 0.5, reflecting the basic detection performance. mAP50:95 is the mean over IoU thresholds ranging from 0.50 to 0.95, used to evaluate the overall detection capability under different localization accuracy requirements.The specific calculation methods for the above metrics are shown in Equations (12)-(16).

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R = \frac{TP}{TP+FN} \tag{13}$$

$$AP = \int_0^1 P(R)\, dR \tag{14}$$

$$mAP_{50} = \frac{1}{C}\sum_{i=1}^{C} AP_i \quad (IoU=0.5) \tag{15}$$

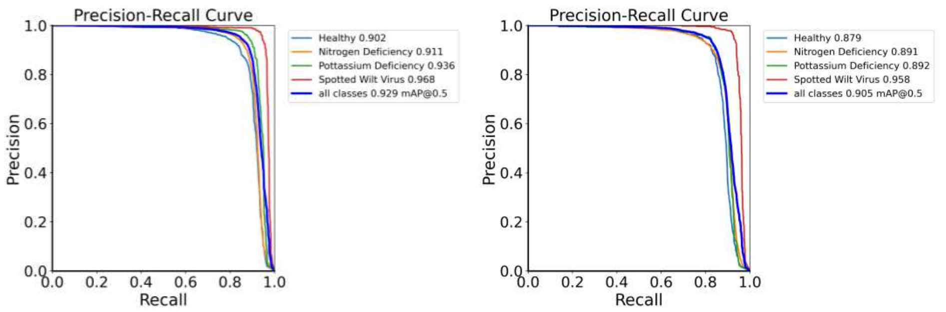
$$mAP_{50-95} = \frac{1}{10}\sum_{t=0.50,0.05}^{0.95} mAP_t \tag{16}$$

To evaluate lightweight and deployment performance, we adopt floating-point operations (GFLOPs), parameter count (Params), and model size as metrics.The lower the GFLOPs, the smaller the inference latency, which is more favorable for real-time detection. The smaller the parameter count and model size, the more they help reduce hardware requirements, facilitating deployment in resource-constrained agricultural scenarios.

4.4 Experimental Results Analysis

4.4.1 Experimental Results

To validate the effectiveness of the improved algorithm proposed in this paper, the baseline model RT-DETR-R18 and the improved model TLD-RTDETR are each trained on the tomato leaf disease dataset under identical parameter configurations. As shown in Figure 9, TLD-RTDETR achieves a mAP50 of 92.9%, which is 2.4 percentage points higher than that of RT-DETR-R18.



(a) Training results of TLD-RTDETR (b) Training results of RT-DETR-R18

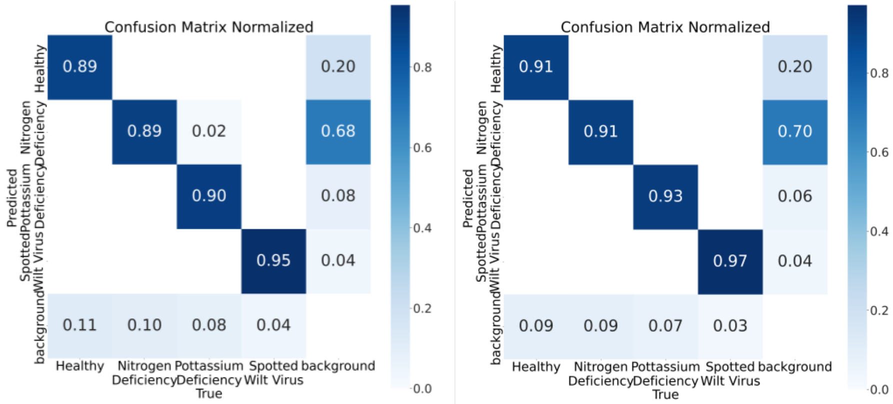
Figure 9 Comparison of training results between RT-DETR-R18 and TLD-RTDETR models

This paper designs comparative experiments from two dimensions, detection accuracy and model lightweighting, to comprehensively evaluate the overall performance of TLD-RTDETR. Detection accuracy is evaluated using recall, precision, mAP50, and mAP50-95, while the degree of lightweighting is measured by parameter count, computational cost (GFLOPs), and model size. The detailed results are shown in Table 3.

Table 3 Model performance comparison on the tomato leaf disease dataset

Model	Recall/%	mAP50/%	mAP50-95/%	Precision/%	FLOPs/G	Params/MB	Size/MB
RT-DETR-R18	84.2	90.5	64.12	93.7	57.0	19.88	77.0
TLD-RTDETR	**87.5**	**92.9**	**66.49**	**94.2**	**39.2**	**12.31**	**47.7**

For a more comprehensive evaluation of the model's recognition performance on each specific disease category, Figure 10 shows the normalized confusion matrices of the two models. The results show that TLD-RTDETR achieves recognition accuracies of 91%, 91%, 93%, and 97% on the four main disease categories, all significantly outperforming the baseline model. Meanwhile, for the "Potassium Deficiency" category, the baseline model misclassifies about 2% of the samples as "Nitrogen Deficiency", indicating that it is prone to feature confusion when handling visually similar diseases. In contrast, TLD-RTDETR effectively eliminates this type of misclassification, demonstrating stronger fine-grained feature discrimination capability. Furthermore, for the common issue in detection tasks where disease targets are falsely detected as background, TLD-RTDETR effectively suppresses the missed detection rate across all categories, further validating its stability and reliability in complex scenarios.



(a) Confusion matrix of RT-DETR-R18 (b) Confusion matrix of TLD-RTDETR

Figure 10 Confusion matrices of the RT-DETR-R18 and TLD-RTDETR models

Overall, compared with RT-DETR-R18, the proposed TLD-RTDETR achieves model lightweighting while further improving detection accuracy. Moreover, it demonstrates significant advantages in discriminating similar diseases and suppressing missed detections.

4.4.2 Ablation Experiments

For the purpose of demonstrating the effectiveness of each improvement of the model on the enhancement of detection performance and lightweight performance, this study conducts ablation experiments on the tomato leaf disease dataset using RT-DETR-R18 as the baseline model. For the three proposed improvement methods, denoted as M 1 (PConvBlock), M2 (CA_HSFPN), and M 3 (AIFI_LP), experimental verification is performed in a stepwise integration manner. A "√" indicates that the corresponding module is adopted. The ablation experiment results are shown in Table 4, where the optimal metrics are marked in bold.

Table 4 Ablation experiments on model detection accuracy metrics

Model	rtdetr-r18	M1	M2	M3	*P* /%	Recall /%	mAP50 /%	mAP 50-95/%	FLOPs/ G	Params/ MB	Size/ MB
Model 1	√	—	—	—	93.7	84.2	90.5	64.1	57.0	19.88	77.0
Model 2	√	√	—	—	93.4	**89.3**	**93.1**	63.5	42.8	14.00	54.6
Model 3	√	√	√	—	93.2	85.9	91.4	65.7	39.2	**12.21**	**47.3**
Model 4	√	√	√	√	**94.2**	87.5	92.9	**66.5**	**39.2**	12.31	47.7

The experimental results show that after introducing the PConvBlock module, RT-DETR-R18 reduces the parameter count, GFLOPs, and model size by 29.6%, 24.9%, and 29%, respectively, while mAP50 increases by 2.6%, respectively, verifying the balance between lightweight design and efficiency achieved by this module. After further integrating the CA_HSFN module, the model achieves a 2.2% increase in mAP50-95 while continuing to optimize lightweight performance. This indicates that the module can significantly enhance the extraction capability of fine lesions and small-target features with low computational overhead, thereby effectively improving the model's robustness. Finally, although introducing the AIFI_LP module brings a slight increase in parameters, the model's accuracy and mAP50-95 jump to 94.2% and 66.5%, respectively. This indicates that by enhancing the global context modeling capability, the module significantly improves the model's discriminability of key disease features under complex backgrounds, verifying its superiority in the tomato leaf disease detection task. The ablation experiment results show that during the gradual introduction of each functional module, although slight fluctuations occur in individual metrics, the overall model performance is not significantly affected. The improved method achieves a good balance between detection accuracy and model lightweighting, taking into account both recognition precision and the practical application requirements of tomato disease and pest detection on edge devices.

4.4.3 Comparison of Lightweight Modules

To evaluate the balance between model lightweighting and recognition performance provided by introducing PConv into the backbone network to build the lightweight PConvBlock convolution module, this paper compares it with several mainstream lightweight modules, i.e., MobileNetV4 **Error! Unknown switch argument.**, GhostNet **Error! Unknown switch argument.**, CGNet **Error! Unknown switch argument.**, DualConv **Error! Unknown switch argument.**, and AKCovn**Error! Unknown switch argument.**. Table 6 presents the experimental results, with optimal metrics highlighted in bold.

Table 5 Comparison experiments of lightweight modules.

Model	Params/M	FLOPs/G	Size/MB	Recall/%	mAP50 /%
RT-DETR-R18	19.88	57.00	77.0	93.7	90.50
RT-DETR-R18+MobileNetV4	**11.31**	**39.50**	**44.4**	80.2	87.50
RT-DETR-R18+DualCovn	15.87	47.30	61.7	87.0	91.59
RT-DETR-R18+CGNet	16.50	47.60	64.3	88.4	92.90
RT-DETR-R18+GhostNet	13.80	44.50	54.7	87.2	92.60
RT-DETR-R18+AKCovn	15.30	47.00	59.5	84.8	91.12
RT-DETR-R18+PCovn	14.00	42.80	54.6	**89.3**	**93.10**

Table 5 reveals that the MobileNetV4 module exhibits the most outstanding performance in model lightweighting, but its detection performance drops significantly, failing to achieve a good balance between lightweighting and accuracy. In contrast, the PCovn module maintains high detection accuracy while ensuring low model complexity. Compared with DualConv, CGNet, and GhostNet, PCovn not only achieves a better level of lightweighting but also delivers the best performance in terms of recall and mAP50, indicating that it effectively enhances feature extraction capability while reducing model redundancy, making it more suitable for resource-constrained agricultural deployment scenarios.

4.4.4 Comparison of Attention Mechanisms

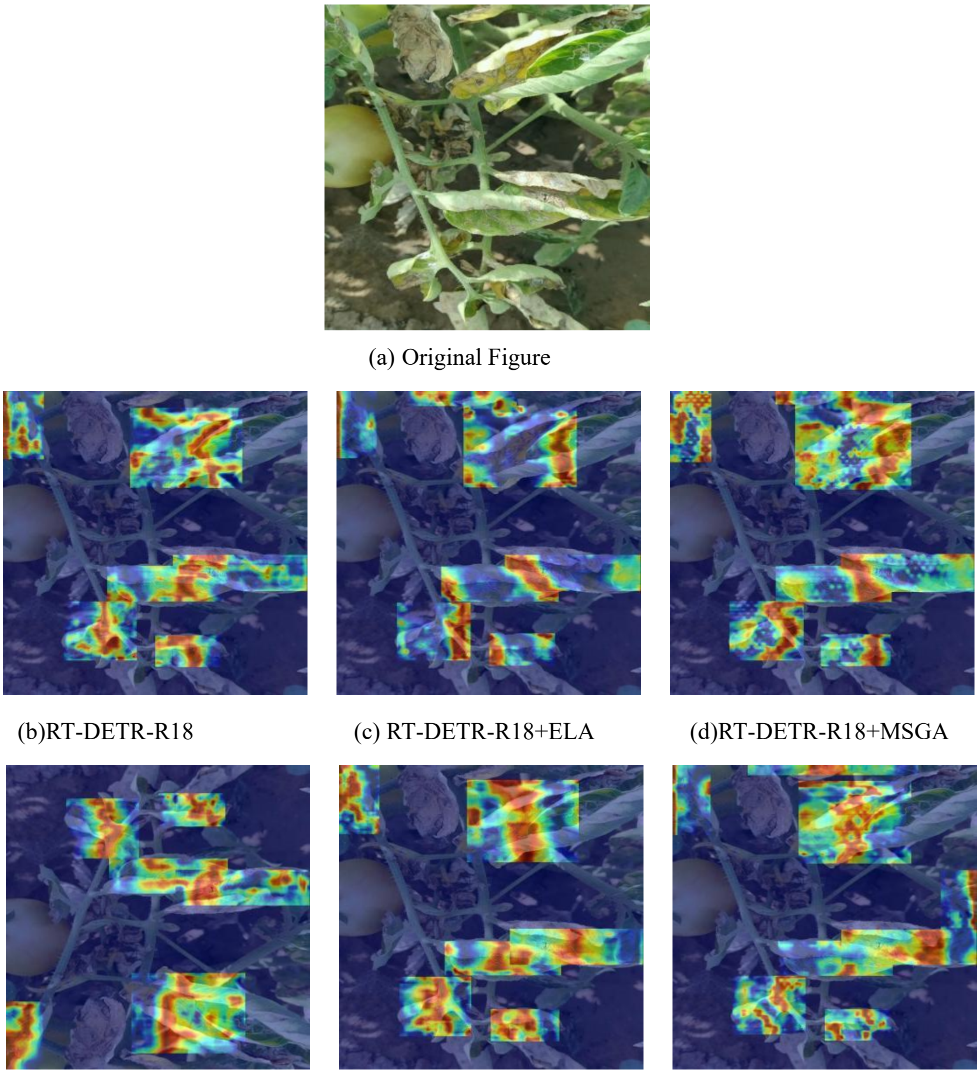
Different attention mechanisms have different sensitivities to the recognition of tomato leaf diseases in different regions. To determine the most suitable attention mechanism for tomato leaf disease recognition and detection, this experiment comparatively evaluates several candidates, including Coordinate Attention (CA) **Error! Unknown switch argument.**, Context Anchor Attention (CAA) **Error! Unknown switch argument.**, Cascaded Group Attention (CGA) **Error! Unknown switch argument.**, Multi-Scale Group Attention (MSGA) **Error! Unknown switch argument.**, and Efficient Local Attention (ELA) **Error! Unknown switch argument.**. Table 7 presents the experimental results on the impact of these mechanisms on model detection performance, with optimal metrics highlighted in bold.

Table 6 Comparison experiments of attention mechanisms

Model	Params/M	FLOPs/G	Size/MB	Precision/%	mAP50 /%
RT-DETR-R18	19.88	57.0	77.0	93.70	90.50
RT-DETR-R18+ELA	19.90	53.7	76.7	93.20	90.70
RT-DETR-R18+MSGA	22.48	71.0	87.0	92.83	91.12
RT-DETR-R18+CGA	19.70	57.0	76.4	93.70	90.60
RT-DETR-R18+CAA	18.60	56.5	71.6	93.80	91.00
RT-DETR-R18+CA	**18.00**	**53.3**	**69.7**	**94.03**	**91.23**

From the comparative analysis in Table 6, it can be seen that after introducing the ELA, CGA, CAA, and CA attention mechanisms into the RT-DETR-R18 model, the parameter count and computational cost of the model are reduced to varying degrees. Among them, the CA attention mechanism performs the best, achieving the highest degree of lightweighting while also delivering the best accuracy and mean average precision. In contrast, introducing the MSGA mechanism, although it improves mAP50, significantly increases the model’s parameter count, GFLOPs, and size, and all accuracy metrics are lower than those of the CA mechanism. This fully demonstrates that the CA attention mechanism can effectively suppress background interferences such as leaf vein textures and light reflections while controlling computational overhead and enhance the discriminative feature representation of small-scale lesion regions. Therefore, this study ultimately chooses to embed the CA attention mechanism into the tomato leaf disease recognition model, aiming to strengthen the multi-scale feature extraction capability and effectively reduce false detections and missed detections under complex backgrounds.

To visually evaluate the effect of different attention mechanisms, heatmaps generated by each method are presented in Figure 11. The highlighted regions indicate the model’s attention to target features, where darker colors represent stronger feature responses. Compared with other mechanisms, Coordinate Attention (CA) enables the model to focus more accurately on tomato leaf disease regions, thereby improving disease recognition performance.



(a) Original Figure

(b)RT-DETR-R18 (c) RT-DETR-R18+ELA (d)RT-DETR-R18+MSGA

(e) RT-DETR-R18+CGA (f) RT-DETR-R18+CAA (g) RT-DETR-R18+CA

Figure 11 Comparison of heatmaps of different attention mechanisms

4.4.5 Model Robustness Test and Comparative Analysis

To systematically evaluate the robustness of the TLD-RTDETR algorithm in complex farmland environments, this study introduces four common types of image degradation into the test set of the tomato leaf disease dataset. In detail, under four severity levels (Severity 1–4, corresponding to intensity coefficients $\eta \in \{0.01, 0.015, 0.02, 0.025\}$), we apply Gaussian noise, salt-and-pepper noise, blur noise, and weather factors, respectively, to simulate real-world interferences such as sensor thermal noise, pixel damage, lens defocus, and adverse weather conditions that may be encountered during actual field collection. By comparing the detection performance of RT-DETR-R18 and TLD-RTDETR, the anti-interference capability of the two algorithms is evaluated both quantitatively and qualitatively. The experiments take mAP50 as the main evaluation metric, and the quantitative comparison results are shown in Tables 7-10, with the best results marked in bold. To visually demonstrate the detection stability of the algorithms under different degradation conditions, Figures 12-20 present the detection visualization results for each severity level.

Table 7 mAP50 Comparison of Two Methods under Different Gaussian Noise Levels

Noise Level	Severity 1	Severity 2	Severity 3	Severity 4
RT-DETR-R18	94.35%	93.99%	93.84%	93.77%
TLD-RTDETR	**94.73%**	**94.73%**	**94.73%**	**94.61%**

Figure 12 Original Image with Ground Truth Bounding Boxes

(a)Severity 1 (b)Severity 2 (c)Severity 3 (d)Severity 4

Figure 13 RT-DETR-R18 Visualization under Different Gaussian Noise Levels

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 14 TLD-RTDETR Visualization under Different Gaussian Noise Levels

Table 8 mAP50 Comparison of Two Methods under Different Levels of Salt-and-Pepper Noise

Noise Level	Severity 1	Severity 2	Severity 3	Severity 4
RT-DETR-R18	90.0%	88.4%	86.35%	84.50%
TLD-RTDETR	**91.5%**	**91.3%**	**90.97%**	**94.61%**

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 15 RT-DETR-R18 Visualization under Different Salt-and-Pepper Noise Levels

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 16 TLD-RTDETR Visualization under Different Salt-and-Pepper Noise Levels

Table 9 mAP50 Comparison of Two Methods under Different Levels of Blur

Noise Level	Severity 1	Severity 2	Severity 3	Severity 4
RT-DETR-R18	94.5%	93.91%	92.17%	91.87%
TLD-RTDETR	**94.9%**	**94.54%**	**92.82%**	**92.48%**

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 17 RT-DETR-R18 Visualization under Different Levels of Blur

（a） Severity 1 (b)Severity 2 (c)Severity 3 (d)Severity 4

Figure 18 TLD-RTDETR Visualization under Different Levels of Blur

Table 10 mAP50 Comparison of Two Methods under Different Weather Conditions

Noise Level	Severity 1	Severity 2	Severity 3	Severity 4
RT-DETR-R18	93.80%	91.58%	90.87%	89.50%
TLD-RTDETR	**94.46%**	**94.46%**	**92.99%**	**92.51%**

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 19 RT-DETR-R18 Visualization under Different Weather Conditions

(a) Severity 1 (b) Severity 2 (c) Severity 3 (d) Severity 4

Figure 20 TLD-RTDETR Visualization under Different Weather Conditions

As shown in Tables 8-11 and Figures 12-20, under image degradation conditions of varying severity, the TLD-RTDETR algorithm consistently outperforms RT-DETR-R18 in terms of mAP50. Further analysis of the visualization results indicates that TLD-RTDETR maintains stable detection performance across all degradation levels, with accurately localized bounding boxes and higher confidence scores than RT-DETR-R18. In contrast, RT-DETR-R18 exhibits varying degrees of false detections and a decline in confidence as the level of degradation increases. These results demonstrate that the TLD-RTDETR algorithm possesses stronger robustness and environmental adaptability in complex field conditions, providing more reliable technical support for the practical deployment of tomato leaf disease detection.

4.4.6 Comparison of Detection Performance across Different Algorithms

To evaluate the generalization ability and robustness of the TLD-RTDETR model, this study compares it with mainstream and state-of-the-art algorithms for tomato disease detection published in recent years. The mainstream models include the two-stage detector Faster R-CNN **Error! Unknown switch argument.** as well as single-stage efficient algorithms such as the YOLO series (v8m **Error! Unknown switch argument.**, v9c **Error! Unknown switch argument.**, v10m **Error! Unknown switch argument.**, v11m **Error! Unknown switch argument.**) and RT-DETR **Error! Unknown switch argument.** (R18, R34). In addition, representative improved algorithms for tomato disease detection published in the past two years, namely LDW-DETR **Error! Unknown switch argument.**, TomatoDet **Error! Unknown switch argument.**, C3ECA-YOLOv5m **Error! Unknown switch argument.**, and DM-YOLO **Error! Unknown switch argument.**, are also selected for comparison. Detailed comparison results are shown in Table 10; optimal metrics are marked in bold.

Table 11 Performance comparison of different algorithms on the tomato disease dataset.

Model	Precision/%	Recall/%	mAP50/%	Params/M	FLOPs/G
Faster R-CNN	63.80	85.7	88.5	41.40	137.20
YOLOv8m	93.80	86.7	92.6	25.80	79.30
YOLO11m	94.10	86.9	92.5	20.11	68.50
YOLOv10m	93.50	86.1	92.5	15.80	59.20
YOLOv9c	93.40	87.3	92.3	25.50	102.10
RT-DETR-R34	93.20	85.6	90.8	31.11	88.80
RT-DETR-R18	93.70	84.2	90.5	19.88	57.00
TomatoDet[**Error! Unknown switch argument.**]	-	-	92.3	13.30	48.98
C3ECA-YOLOv5m	87.76	87.2	90.4	15.94	35.50

DM-YOLO	92.50	86.8	86.4	50.00	238.00
LDW-DETR	84.70	82.9	90.1	16.20	54.70
TLD-RTDETR	**94.20**	**87.5**	**92.9**	**12.31**	**39.20**

From the data in Table 11, it can be seen that the TLD-RTDETR algorithm achieves the most outstanding performance in the tomato leaf disease detection task. In comparison with mainstream object detection algorithms, TLD-RTDETR achieves an mAP50 of 92.94% and a precision of 94.23%, outperforming Faster R-CNN, YOLO11m, YOLOv8m, YOLOv9c, YOLOv10m, RT-DETR-R34, and the baseline model RT-DETR-R18 across multiple key performance metrics while having only 12.31M parameters and 39.2G FLOPs, thus achieving the most lightweight model architecture. Compared with improved algorithms for tomato disease recognition published in the past two years, TLD-RTDETR also demonstrates excellent performance in both detection accuracy and lightweight metrics. Relative to LDW-DETR, TomatoDet, C3ECA-YOLOv5m, and DM-YOLO, our model achieves a significant mAP50 improvement while maintaining the lowest parameter count and relatively low computational cost.

Overall, the experimental results fully validate that the TLD-RTDETR algorithm achieves a favorable balance between model lightweighting and detection accuracy. It is capable of identifying disease targets accurately and efficiently, thereby meeting the practical requirements of the tomato leaf disease recognition task.

4.4.7. Cross-Dataset Generalization Experiment

This paper performs a cross-dataset generalization experiment on a strawberry disease and pest dataset, analogous to the tomato leaf disease recognition test, to further confirm the improved model's generalization performance. This dataset was published by Afzaal et al. **Error! Unknown switch argument.** and contains 2,500 images of strawberry diseases, covering seven different types of diseases and pests. The categories include strawberry angular leaf spot, strawberry anthracnose, blossom blight, gray mold, leaf spot, strawberry fruit powdery mildew, and strawberry leaf powdery mildew. In addition, 4,627 images of leaf diseases from two crop types, bell pepper and potato, covering five disease categories, are selected from the PlantVillage dataset [55] to further verify the scalability of the proposed model. Some example samples are shown in Figure 22.

(b).Angular Leafspot (b).Anthracnose Fruit Rot (c).Blossom Blight

(d).Gray Mold (e).Leaf Spot (f).Powdery Mildew Fruit (g).Powdery Mildew Leaf

Figure 21 Images of the strawberry fruit and leaf disease and pest dataset

(a).Pepper,bell_Bacterial spot (b).Pepper,bell_healthy

(c).Potato Early blight (d).Potato Late blight (e) Potato healthy

Figure 22 Images from the PlantVillage dataset

From Table 12 and Table 13, it can be seen that TLD-RTDETR achieves better overall detection performance than the baseline model RT-DETR-R18 on both the strawberry disease and pest dataset and the PlantVillage dataset. The experimental results demonstrate that TLD-RTDETR not only performs excellently on the tomato leaf disease dataset but also exhibits good generalization ability and robustness in cross-dataset scenarios, effectively adapting to the detection requirements of different crop diseases. This further validates the effectiveness and generality of the improvement strategies proposed in this study.

To validate the transferability and adaptability of the proposed model to a new dataset, this paper conducts a comparative experiment between TLD-RTDETR and the baseline model RT-DETR-R18 on the strawberry disease and pest dataset. The results are shown in Table 13. TLD-RTDETR achieves better overall detection performance than the baseline model, with improvements of 1.3, 1.12, 2.3, and 1.0 percentage points in Precision, mAP50, mAP50-95, and Recall, respectively. This indicates that TLD-RTDETR not only performs excellently on the tomato leaf disease dataset but also exhibits good generalization ability in cross-dataset scenarios, effectively adapting to the detection requirements of different crop diseases.

Table 12 Performance comparison of models on the strawberry disease and pest dataset.

Model	Recall/%	mAP50/%	mAP50-95/%	Precision/%
RT-DETR-R18	81.40	84.50	64.40	90.20

TLD-RTDETR	**82.40**	**85.62**	**66.70**	**91.50**

Table 13 Performance comparison of model detection accuracy on the PlantVillage dataset.

Model	Recall/%	mAP50/%	mAP50-95/%	Precision/%
RT-DETR-R18	96.30	98.50	97.90	96.60
TLD-RTDETR	**97.14**	**98.79**	**98.20**	**99.00**

4.4.8 Comparison of Visualization Effects

In scenarios for recognizing tomato leaf diseases, there exist diseases that are small in size or have inconspicuous features due to factors such as occlusion. Therefore, evaluating the model's accurate localization capability is crucial. To show how the enhanced model reduces missed and false detections, we compare the baseline and TLD-RTDETR on the tomato leaf disease dataset,as shown in Figure 22. The figure also shows the predicted bounding boxes, defect categories, and confidence scores.

(a)RT-DETR-R18

(b)TLD-RTDETR

Figure 23 Comparison of visualization effects between RT-DETR-R18 and TLD-RTDETR

Figure 23(a) and Figure 23(b) show the detection results of the RT-DETR-R18 model on the tomato leaf disease dataset. In the figures, red boxes indicate missed detection targets, and yellow boxes indicate false detection targets. From the comparison, it can be seen that RT-DETR-R18 exhibits varying degrees of missed detections and false detections on the tomato leaf disease dataset. Moreover, on the tomato leaf disease dataset after data augmentation, the baseline model tends to misclassify occluded regions as targets. This is mainly because RT-DETR-R18 adopts the lightweight ResNet-18 as its backbone network. Limited by the network depth, its feature representation capability is relatively restricted, making it difficult to extract discriminative fine-grained features under complex backgrounds and small-target scenarios. Meanwhile, the insufficient multi-scale feature fusion capability leads to a weak perception of small and occluded targets, which in turn results in missed detections and false detections.

In contrast, the improved TLD-RTDETR model exhibits superior localization and recognition capabilities in object detection tasks, enabling it to focus on target regions more accurately. Particularly in complex scenes and small-target detection tasks, the model effectively alleviates missed detections and false detections, leading to more stable overall detection results and demonstrating stronger robustness and accuracy.

5. Conclusion

To address the challenge that existing end-to-end tomato leaf disease detection models struggle to balance between the constraints of embedded deployment and the recognition accuracy in complex environments, this study proposes an improved lightweight tomato leaf disease detection algorithm, named TLD-RTDETR. The PConvBlock module is introduced into the backbone network, leveraging the lightweight mechanism of partial convolution to reduce redundant computation and memory access, thereby significantly decreasing the model parameter size while enhancing feature extraction capability. The AIFI_LPE module introduces learnable positional encoding to adaptively model spatial position information and contextual relationships, further enhancing feature representation capability, thereby improving the model's ability to identify key disease regions under complex backgrounds. In the neck network, the CA_HSFPN structure effectively improves multi-scale feature fusion. The coordinate attention mechanism enhances the model's perception of spatial location information of lesions, while the high-level selection strategy strengthens the representation of small-target disease features. This is verified by visual attention maps, showing that the improved model can focus more accurately on disease regions while suppressing interference from leaf vein textures and complex backgrounds.

Systematic experiments on the tomato leaf disease dataset, strawberry disease and pest dataset, and PlantVillage dataset verify the effectiveness of the proposed method. Compared with the baseline RT-DETR-R18, TLD-RTDETR reduces model size, parameter count, and computational cost by 38.1%, 38.1%, and 31.2%, respectively, while improving mAP50 and recall by 2.4% and 3.3% on the tomato leaf disease dataset. The proposed model also achieves consistent performance improvements on the strawberry disease and pest dataset and the PlantVillage dataset. Visualization and robustness experiments further demonstrate that TLD-RTDETR can more accurately focus on target regions, reduce missed and false detections, and maintain stable detection performance under various degradation conditions. Overall, the proposed method achieves a good balance between detection accuracy, computational efficiency, localization capability, and robustness, demonstrating strong potential for lightweight deployment in real agricultural environments.

Conflict of Interest

The authors declare no competing interests.

Ethical use of research materials

All research materials, including images, graphs, data, and models, were either created by the authors or appropriately obtained from credible sources with explicit permission or licenses. For any copyrighted materials used in this study, permission was obtained from the copyright holders. Where applicable, proper acknowledgment has been provided for all third-party materials, and all data sources are fully referenced in the manuscript. The authors take full responsibility for the ethical use of all research materials.

CRediT authorship contribution statement

Conceptualization: J.L. and Y.L. Formal analysis: J.L. Data collection: J.L. and G.W. Investigation: J.L. and Y.L. Methodology: J.L. and Y.L. Analysis and interpretation of results: J.L. and G.W. Draft manuscript preparation: J.L. Validation: J.L. and M.Y. Writing-original draft: J.L. Writing-review and editing: Y.L. and M.Y. All authors reviewed the results and approved the final version of the manuscript.

Funding

This work was supported by Guangxi Natural Science Foundation(2025GXNSFAA069414) and Guangxi Science and Technology Innovation Platform Program Project(Guike LT2600640001).

This work was supported by Guangxi Vocational College of Economics and Trade College-level Project for Enhancing Basic Research Capacity of Young and Middle-aged Teachers (JMKY202501).

Ackowledgment

We are grateful to all of those who provided useful suggestions for this study.

Data availability

The datasets used in this study are publicly available. Tomato leaf disease images can be accessed via Roboflow Universe. The strawberry disease and pest dataset is provided by Afzaal et al. (2021) https://doi.org/10.3390/s21196565. Additionally, the PlantVillage dataset, an open-access repository of plant health images, is available through Hughes and Salathé (2015) arXiv:1511.08060.

References

[1] BARBEDO J G A. Factors influencing the use of deep learning for plant disease recognition[J]. Biosystems engineering, 2018, 172: 84-91.

[2] SUN H, NICHOLAUS I T, FU R, et al. YOLO-FMDI: a lightweight YOLOv8 focusing on a multi-scale feature diffusion interaction neck for tomato pest and disease detection[J]. Electronics, 2024, 13(15): 2974.

[3] Yan C, Yang K. FSM-YOLO: Apple leaf disease detection network based on adaptive feature capture and spatial context awareness[J]. Digital Signal Processing, 2024, 155: 104770.

[4] LIU B,ZHANG Y,WANG J,et al. An improved lightweight network based on deep learning for grape recognition in unstructured environments[J]. Information Processing in Agriculture，2024，11(2):202-216.

[5] Wang Q, Yan N, Qin Y, et al. BED-YOLO: an enhanced YOLOv10n-Based tomato leaf disease detection algorithm[J]. Sensors, 2025, 25(9): 2882.

[6] Bian S, Su S, Zhou J, et al. Multi-Scale Feature Fusion Based RT-DETR for Tomato Leaf Disease Detection in Complex Backgrounds[J]. Sensors, 2025, 25(23): 7275.

[7] Wang S, Chen S, Dong S, et al. Context-guided and multi-level feature fusion DETR for lightweight detection of cacao pod pests and diseases in complex environments[J]. International Journal of Machine Learning and Cybernetics, 2026, 17(4): 175.

[8] Zhang Y, Song J, Yu X, et al. WMC-RTDETR: a lightweight tea disease detection model[J]. Frontiers in Plant Science, 2025, 16: 1574920.

[9] Wang J, Wang T, Xu Q, et al. RP-DETR: end-to-end rice pests detection using a transformer[J]. Plant Methods, 2025, 21(1): 63.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C].

Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). 2016: 779-788.

[11] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. European conference on computer vision(ECCV). Springer International Publishing, 2020: 213-229.

[12] ZHAO Y A,LV W Y,XU S L,et al.DETRs beat YOLOs on real-time object detection[EB/OL].(2023–04–17).https://arxiv.org/abs/2304.08069.

[13] Zhou Y,Wei Y.UAV-DETR:An enhanced RT-DETR architecture for efficient small object detection in UAV imagery[J].Sensors,2025,25(15):4582.

[14] Chen J, Kao S, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12021-12031.

[15] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[17] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[18] tomato-village-diseases https://universe.roboflow.com/tomatodiseases/tomato-village-diseases

[19] Afzaal, U.; Bhattarai, B.; Pandeya, Y.R.; Lee, J. An Instance Segmentation Model for Strawberry Di seases Based on Mask R-CNN. Sensors 2021, 21, 6565. https://doi.org/10.3390/s21196565

[20] GEETHARAMANI G, PANDIAN A. Identification of plant leaf diseases using a nine-layer deep convolutional neural network[J]. Computers & Electrical Engineering, 2019, 76: 323-338.

[21] Yang, H.; Xue, H.; Lyu, Y.; Mu, M.; Tang, T.; Huang, Z. LDW-DETR: An Efficient Tomato Leaf Disease Detection Algorithm Based on Enhanced RT-DETR. Appl. Sci. 2025, 15, 11620. https://doi.org/10.3390/app152111620

[22] Wang X, Liu J. An efficient deep learning model for tomato disease detection[J]. Plant Methods, 2024, 20(1): 61.

[23] Lee, Y.-S.; Patil, M.P.; Kim, J.G.; Choi, S.S.; Seo, Y.B.; Kim, G.-D. Improved Tomato Leaf Disease Recognition Based on the YOLOv5m with Various Soft Attention Module Combinations. Agriculture 2024, 14, 1472. https://doi.org/10.3390/agriculture14091472

[24] Abulizi A, Ye J, Abudukelimu H, et al. DM-YOLO: Improved YOLOv9 model for tomato leaf disease detection[J]. Frontiers in Plant Science, 2025, 15: 1473928.

[25] JOCHER G,QIU Jing,CHAURASIA A.Ultralytics YOLO11[CP/OL].(2024-09-30).https://github.com/ultralytics/ultralytics.

[26] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[27] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in neural information processing systems, 2024, 37: 107984-108011.

[28] REIS D, KUPEC J, HONG J, et al. Real-time flying object detection with YOLOv8[J]. arXiv preprint arXiv:2305.09972, 2023.

[29] Qin D, Leichner C, Delakis M, et al. MobileNetV4: Universal models for the mobile ecosystem[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 78-96.

[30] Cai X, Lai Q, Wang Y, et al. Poly kernel inception network for remote sensing detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 27706-27716.

[31] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.
[32] Wu T, Tang S, Zhang R, et al. CGNet: A light-weight context guided network for semantic segmentation[J]. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.
[33] Liu X, Peng H, Zheng N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14420-14430.
[34] Xu W, Wan Y, Zhao W. ELA: efficient location attention for deep convolution neural networks[J]. Journal of Real-Time Image Processing, 2025, 22(4): 140.
[35] Zhang X, Zeng H, Guo S, et al. Efficient long-range attention network for image super-resolution[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 649-667.
[36] Zhong J, Chen J, Mian A. DualConv: Dual convolutional kernels for lightweight deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 9528-9535.
[37] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.
[38] Chen Z , Shi G , Wang X .Text Classification Based on Naive Bayes Algorithm with Feature Selection[J].International Journal on Information, 2012, 15(10):4255-4260.
[39] Cover T M , Hart P E .Nearest neighbor pattern classification[J].IEEE Trans.inf.theory, 1953, 13(1):21-27.DOI:10.1109/TIT.1967.1053964.
[40] Gurunathan V, Dhanasekar J, Ishwarya Niranjana M. Plant leaf diseases detection using KNN classifier[C]//2023 9th International conference on advanced computing and communication systems (ICACCS). IEEE, 2023, 1: 2157-2162.
[41] Sachdeva G, Singh P, Kaur P. Plant leaf disease classification using deep Convolutional neural network with Bayesian learning[J]. Materials Today: Proceedings, 2021, 45: 5584-5590.
[42] Bhola A, Kumar P. Deep feature-support vector machine based hybrid model for multi-crop leaf disease identification in Corn, Rice, and Wheat[J]. Multimedia Tools and Applications, 2025, 84(8): 4751-4771.
[43] Gong X, Zhang S. A high-precision detection method of apple leaf diseases using improved faster R-CNN[J]. Agriculture, 2023, 13(2): 240.
[44] Alruwaili M, Siddiqi M H, Khan A, et al. RTF-RCNN: An architecture for real-time tomato plant leaf diseases detection in video streaming using Faster-RCNN[J]. Bioengineering, 2022, 9(10): 565.
[45] He J, Liu T, Li L, et al. MFaster r-CNN for maize leaf diseases detection based on machine vision[J]. Arabian Journal for Science and Engineering, 2023, 48(2): 1437-1449.
[46] Meng Y, Zhan J, Li K, et al. A rapid and precise algorithm for maize leaf disease detection based on YOLO MSM[J]. Scientific Reports, 2025, 15(1): 6016.
[47] Xu W, Wang R. ALAD-YOLO: An lightweight and accurate detector for apple leaf diseases[J]. Frontiers in plant science, 2023, 14: 1204569.
[48] Yao J, Li Y, Xia Z, et al. WTAD-YOLO: a lightweight tomato leaf disease detection model based on YOLO11[J]. Smart Agricultural Technology, 2025: 101349.
[49] Xu D, Mao X, Xia F. DBPC-DETR: a lightweight dual backbone network for real-time identification of multi-scale grape leaf diseases under field conditions[J]. Journal of Real-Time Image Processing, 2026, 23(1): 47.
[50] Yao J, Zhu Z, Yuan M, et al. The Detection of Maize Leaf Disease Based on an Improved Real-Time Detection Transformer Model[J]. Symmetry, 2025, 17(6): 808.
[51] Sun H, Fu R, Zhang Z, et al. Research on eggplant pest and disease detection in complex environments based

on improved RT-DETR[J]. BMC Plant Biology, 2025, 25(1): 1664.

[52] Ren W, Qian T, Li G, et al. RT-DETR-MC: A Novel Fusion of Multi-Scale Features and Gating Mechanism for Lettuce Disease Detection[C]//Proceedings of the 2025 6th International Conference on Computer Information and Big Data Applications. 2025: 196-202.

[53] Ren S , He K , Girshick R ,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.DOI:10.1109/TPAMI.2016.2577031.

[54] Zhang X, Song Y, Song T, et al. AKConv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters[J]. arXiv preprint arXiv:2311.11587, 2023: 2-10.

[55] Hughes D, Salathé M, et al. (2015) An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv:1511.08060