applied sciences

MDPI

Article

LDW-DETR: An Efficient Tomato Leaf Disease Detection Algorithm Based on Enhanced RT-DETR

Hua Yang *, Hao Xue, Yanjie Lyu, Mingzhi Mu, Tianwei Tang and Zhongke Huang

School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430040, China; ksilverbell@outlook.com (H.X.); lvyj8863@163.com (Y.L.); xiaomumu1203@gmail.com (M.M.); 15337172258@163.com (T.T.); hyy520hyy666@outlook.com (Z.H.)

* Correspondence: huayang@whpu.edu.cn

Abstract

Tomato is one of the most important economic crops in the world, but it is prone to diseases during the growth process, so the detection of tomato diseases is very important. However, when detecting tomato diseases in natural environments, existing models are easily affected by environmental factors such as occlusion and illumination, as well as the small size of lesions. In response to these challenges, this paper proposes a tomato leaf disease detection framework LDW-DETR based on multi-scale fusion. First, the local-global feature fusion (LGFF) module is designed by referring to the idea of the PPA module, which can effectively capture local and global features, thereby enhancing the detection ability of small lesions in complex backgrounds. Second, the CSPDarknet architecture is introduced as the backbone network of LDW-DETR to improve the efficiency of feature extraction. In addition, the bottleneck layer of the C2f component is improved by integrating Strip Block and Contextualized Gated Linear Unit (CGLU) to enhance the perception ability of lesion edges and textures. Finally, the WIoU v3 loss function is used to optimize the bounding box regression process. The experimental results show that compared with RT-DETR, the LDW-DETR model improves mAP@0.5 and mAP@0.5–0.95 by 2.6% and 3.7%, respectively, while the number of parameters is reduced by 17.9%. In addition, it still maintains high robustness and generalization ability in cross-dataset experiments. These results show that LDW-DETR has good detection performance and generalization ability in the tomato leaf disease detection task.

Keywords: tomato leaf disease; RT-DETR; disease detection; deep learning

check for updates

Academic Editor: Pedro Couto

Received: 10 October 2025
Revised: 28 October 2025
Accepted: 29 October 2025
Published: 30 October 2025

Citation: 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

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Tomatoes originated in South America and are one of the most important vegetables in the world. They rank second globally in terms of cultivation area and have high economic value, serving as an important source of income for many countries [1]. Beyond its economic importance, tomato is recognized for its substantial nutritional value, as it serves as a rich source of vitamin C, potassium, folate, and carotenoids, which are essential for human health maintenance and disease prevention [2]. However, in practical cultivation, tomato yield is constrained by multiple factors. First, due to intensive selection during evolution and domestication, as well as a severe genetic bottleneck, cultivated tomatoes exhibit relatively low genetic diversity. Consequently, tomatoes are more susceptible to pathogen invasion at various stages of growth and harvest, with more than 200 related diseases having been documented to date [3]. Second, natural environmental factors such as climate

change, together with anthropogenic factors including poor drainage and inadequate fertilization, can further increase the incidence of bacterial infections in tomatoes [4]. These diseases can severely impact tomato yields and economic value [5].

Second, natural environmental factors such as climate change, together with anthropogenic factors including poor drainage and inadequate fertilization, can further increase the incidence of bacterial infections in tomatoes, thereby leading to substantial reductions in both yield and economic value. Therefore, the early detection and identification of tomato diseases, along with the adoption of appropriate control measures, are crucial for ensuring crop yield and economic value. As the initial symptoms of most diseases usually appear on the leaves, the accurate recognition of tomato leaf diseases is particularly important [6].

Traditional disease detection methods are usually performed through visual inspection of leaves by observers. However, this manual identification approach is not only inefficient but also limited in accuracy, as it is easily affected by factors such as the observer's theoretical knowledge, fatigue level, and environmental variations [7]. When symptoms are poorly understood or assessed by inexperienced observers, simple visual comparison is often inadequate for reliable diagnosis and may lead to misclassification [8]. With the rapid progress of convolutional neural networks in recent years, deep learning has been extensively utilized in a variety of areas, such as wind power forecasting [9], traffic prediction [10], Interval prediction [11], text classification [12], and agricultural pest and disease detection [13]. Object detection, as a crucial subfield of deep learning, has made significant advances in recent years and has been increasingly applied in the agricultural field, offering new opportunities for tomato leaf disease identification [14,15]. Compared with traditional detection methods, object detection algorithms offer significant advantages in feature extraction, detection accuracy, and detection speed [16].

Currently, mainstream object detection algorithms are generally divided into single-stage methods, two-stage methods, and transformer-based methods [17]. Mainstream two-stage methods are primarily represented by models such as Faster R-CNN [18] and Mask R-CNN [19]. While these algorithms typically attain a high level of detection accuracy, their efficiency is relatively low due to the generation of many candidate bounding boxes. Gong et al. [20] incorporated Res2Net and a feature pyramid network architecture as the feature extraction backbone into the Faster R-CNN model for apple leaf disease detection and achieved 63.1% mAP@0.5. Mainstream one-stage object detection algorithms mainly include the YOLO series [21–25], SSD [26], and EfficientDet [27]. Owing to their higher efficiency, these methods have been widely applied in the field of crop leaf disease identification. For example, Wang et al. [28] proposed an apple leaf disease detection method based on YOLOv5, in which a mobile inverted residual bottleneck convolution integrated with the CBAM was designed to enhance feature extraction capability. In addition, the Ghost module was employed to replace standard convolutions, making the model more lightweight, and the method achieved 94% mAP@0.5. Abudukelimu Abulizi et al. [29] proposed DM-YOLO for tomato leaf disease detection. This method integrates DySample and MPDIoU, enabling the model to effectively identify early subtle lesion areas, thereby achieving precise detection of tomato diseases and accurate localization of subtle edge features of lesions at different scales. Yang et al. [30] improved the YOLOv8 model for corn leaf disease detection by introducing slim-neck and Global Attention Mechanism (GAM) attention, enhancing the model's ability to identify corn leaf spot disease. Compared with the original YOLOv8, their approach achieved a 3.79% improvement in mAP@0.5.

Initially designed for NLP applications, Transformers have in recent years been applied to object detection, owing to their strong capacity for capturing global contextual information [31]. Detection transformer (DETR) [32] is the first model to apply transformer to the field of object detection. It breaks the paradigm of the traditional object detec-

tion framework and directly regards object detection as a sequence prediction problem, simplifying the detection process. Yang et al. [33] proposed a Dense Higher-Level Composition Detection Transformer (DHLC-DETR) methodology based on the DETR model. This method is mainly used for rice pest and disease detection and can effectively identify three types of diseases in rice: sheath blight, rice blast, and flax spot disease. Ultimately, compared with the original DETR model, mAP@0.5 improved by 17.3%. However, such models have many parameters, high computational costs, and slow inference speeds, making them unsuitable for the real-time detection demands of modern agriculture. Subsequently, Zhao et al. introduced RT-DETR [34], the first real-time end-to-end object detector based on Transformer. By adopting an efficient decoder architecture and a multi-scale feature fusion strategy, RT-DETR preserves the global modeling capability of Transformer while significantly reducing inference costs and improving detection speed. These characteristics make it particularly suitable for agricultural pest and disease detection scenarios where real-time performance is essential.

Although significant progress has been made in applying object detection models to agriculture, challenges still exist in the detection of tomato leaf diseases. These include insufficient detection capability for small-scale lesion areas, inadequate differentiation between diseases with similar features, and high computational costs. To address these issues, this paper proposes the LDW-DETR(LGFF-CSPDarknet-WIoU-DETR) model for tomato leaf disease detection, based on the RT-DETR architecture. This approach not only reduces computational overhead but also enhances detection accuracy for small lesions on tomato leaves in complex backgrounds. The main contributions of this work are as follows:

1. Based on the patch-aware mechanism in the PPA module, a local-global feature fusion (LGFF) module is proposed. The module combines local details and global context information, and designs a multi-branch structure by drawing on the Patch-aware mechanism. The local branch focuses on the extraction of fine-grained features, and the global branch provides more extensive background information. This design significantly improves the model's ability to identify small lesions of tomato leaf disease in the natural environment, especially in complex backgrounds.
2. The CSPDarknet architecture is introduced as the backbone network of LDW-DETR to improve the efficiency of feature extraction. In addition, the C2f-SCG (CSCG) module is proposed to improve the CSPDarknet architecture, which mainly replaces the bottleneck layer in C2f by integrating Strip Block and Contextualized Gated Linear Unit (CGLU). The backbone network not only reduces the number of parameters, but also enhances the perception ability of lesion edges and textures.
3. Replace the GIOU loss function with the WIOU V3 loss function. Optimize the bounding box regression process to enhance model accuracy.

2. Materials and Methods

2.1. Materials

2.1.1. Data Collection

The data used in this study were obtained from three datasets: PlantVillage [35], FieldPlant [36], and PlantDoc [37]. The PlantVillage dataset is the first publicly available plant disease dataset, containing images of 13 plant species and 26 types of diseases. Specifically, it includes 18,160 images of tomato leaf diseases, all of which were taken in a laboratory environment. The PlantDoc dataset mainly consists of images downloaded from the Internet. It contains 2598 images covering 13 plant species and 27 categories, and represents the first publicly available plant disease dataset collected from uncontrolled environments. The Fieldplant dataset contains 5170 annotated field leaf images collected from the Kalonmai plantation, which can be used to train efficient plant disease detection

models based on field image correlation and object detection models. More transparent dataset samples are shown in Figure 1.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 1. Overview of the dataset. (**a**) healthy; (**b**) late blight; (**c**) brown spot; (**d**) yellow curl virus; (**e**) bacterial.

Among the three datasets, the PlantVillage dataset was captured in a laboratory environment and lacks real-world background contexts. The Fieldplant dataset was captured in a botanical garden and includes backgrounds under various lighting and weather conditions. The PlantDoc dataset consists of images sourced from the internet, with varying image quality. Therefore, we selected and integrated tomato leaf disease images from the above three datasets based on image quality, background diversity, and the number of disease categories. Ultimately, we obtained 1917 tomato leaf disease images, including five categories: healthy, brown spot, bacterial, late blight, and yellow curl virus. The specific distribution of quantities is shown in Table 1.

Table 1. Category distribution.

Class Name	Number
healthy	358
late blight	382
brown spot	414
yellow curl virus	389
bacterial	374
total	1917

As shown in Table 1, the dataset contains sample sizes across five categories as follows: 358 healthy leaves, 414 with brown spot, 374 with bacterial disease, 382 with late blight, and 389 with yellow leaf curl virus, totaling 1917 images. This distribution demonstrates a high degree of class balance. Specifically, the ratio between the most frequent category (brown spot) and the least frequent (healthy leaves) is only 1.16:1, with all categories representing between 18.7% and 21.6% of the dataset, forming an approximately uniform distribution.

2.1.2. Data Preprocessing

The constructed dataset was annotated using LabelImg v1.8.1, and the annotation files were stored as txt files in YOLO format. Subsequently, the 1917 annotated image files were divided into training, validation, and test sets in a ratio of 7:2:1. To enhance the model's robustness and generalization capability, this study incorporated various data augmentation techniques. To ensure objectivity, augmentation was performed only on the training set. The data augmentation methods used included horizontal and vertical flipping, rotation, grayscale adjustment, saturation adjustment, and exposure adjustment. Ultimately, the number of images in the training set increased from 1347 to 4041, while the number of images in the test set and training set was 377 and 193, respectively.

2.2. RT-DETR Model

RT-DETR is the first real-time end-to-end object detector based on transformers, jointly proposed by teams from Baidu and the School of Electronic and Computer Engineering at Peking University. It primarily consists of a backbone, a Transformer decoder with an auxiliary pre-detection head, and an efficient hybrid encoder. RT-DETR offers multiple model configurations to accommodate different application scenarios. In this study, RT-DETR-R18 was adopted as the baseline model, in which ResNet-18 is employed as the backbone of RT-DETR. The structure of RT-DETR-R18 is illustrated in Figure 2.

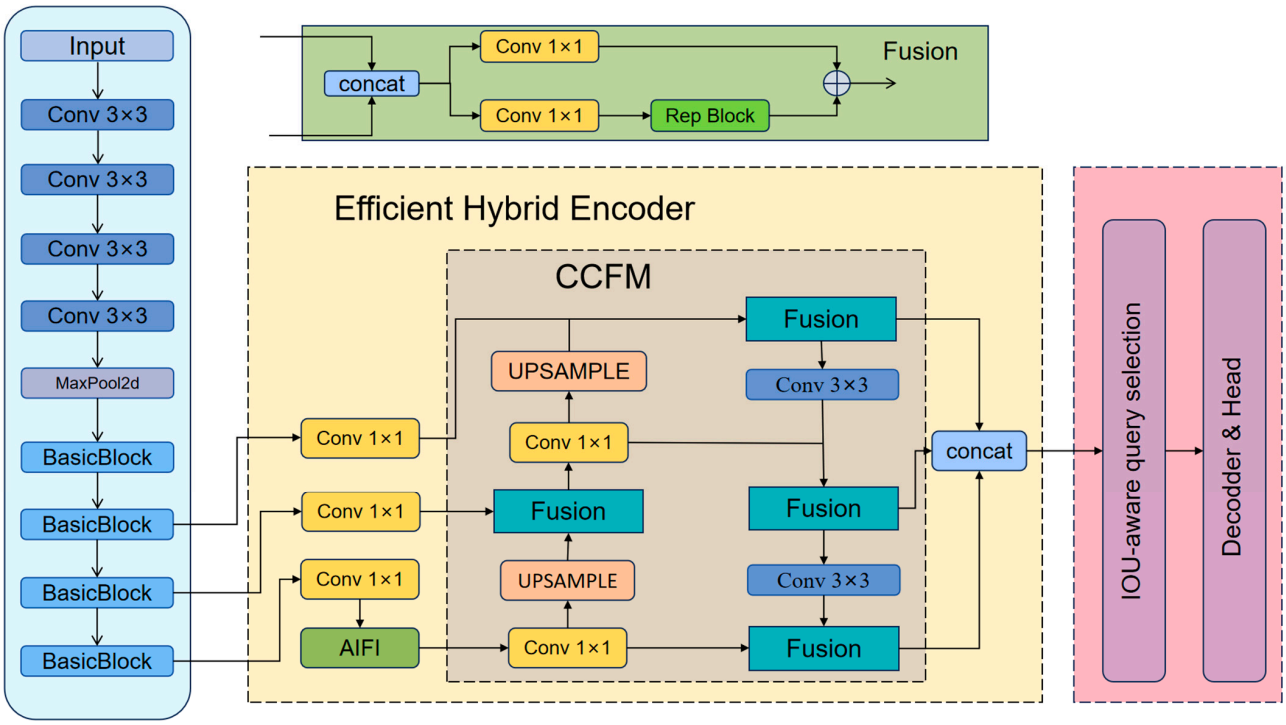


Figure 2. RT-DETR network architecture diagram.

The main function of the RT-DETR backbone network is to extract multi-scale features from images and use the last three layers of features as input for the efficient hybrid encoder. The efficient hybrid encoder mainly consists of AIFI and CCFM. Among them, AIFI is a feature interaction module based on the attention mechanism, which mainly processes deep features. It performs feature interaction operations on high-level features s5 through a single-scale transform encoder, facilitating subsequent module localization and target recognition. The multi-scale feature information obtained after interaction will be sent to CCFM for processing. The structure of CCFM is optimized through the integration of multiple fusion blocks into its fusion path, where each block is composed of convolutional layers. Their function is to amalgamate features from two adjacent scales, thereby producing a new feature representation. It then fuses multiple outputs by pointwise addition to generate a new feature sequence. Finally, A fixed number of features are sampled from the encoder's output as initial queries. These queries are then passed to a decoder that incorporates an auxiliary prediction head. This decoder undergoes iterative optimization, yielding the final prediction boxes and their confidence scores.

2.3. Improved LDW-DETR

In order to overcome the problems of insufficient feature discrimination in small lesion detection and complex background of tomato leaves in RT-DETR, this paper proposes an

improved detection framework, LDW-DETR. The model introduces the following three improvements on the basis of maintaining the efficient end-to-end structure of RT-DETR.

First, a novel LGFF module is designed. This module draws on the patch-aware mechanism and realizes the adaptive fusion of local fine-grained features and global context features through a multi-branch structure, thereby enhancing the model's ability to distinguish small-scale lesions and similar lesion patterns, and effectively improving the detection accuracy.

Second, the CSPDarknet architecture is introduced as the backbone network, and the CSCG module is proposed to improve the architecture. CSCG replaces the bottleneck layer in C2f by combining Strip Block and CGLU, which can better capture the strip-shaped and marginal lesion features in the leaves. The backbone network can significantly reduce the number of model parameters while ensuring the feature extraction capability.

Finally, the WIoU v3 loss function is introduced in the bounding box regression part to replace the traditional GIoU. WIoU v3 achieves dynamic adjustment of the quality of anchor boxes through a non-monotonic focusing factor, which enables the model to allocate gradients more reasonably during the training process, thereby improving the stability and accuracy of the detection box positioning.

Overall, LDW-DETR achieves a good balance between detection accuracy and model complexity through backbone network optimization, feature fusion module design, and loss function improvement. The model structure diagram of LDW-DETR is shown in Figure 3.

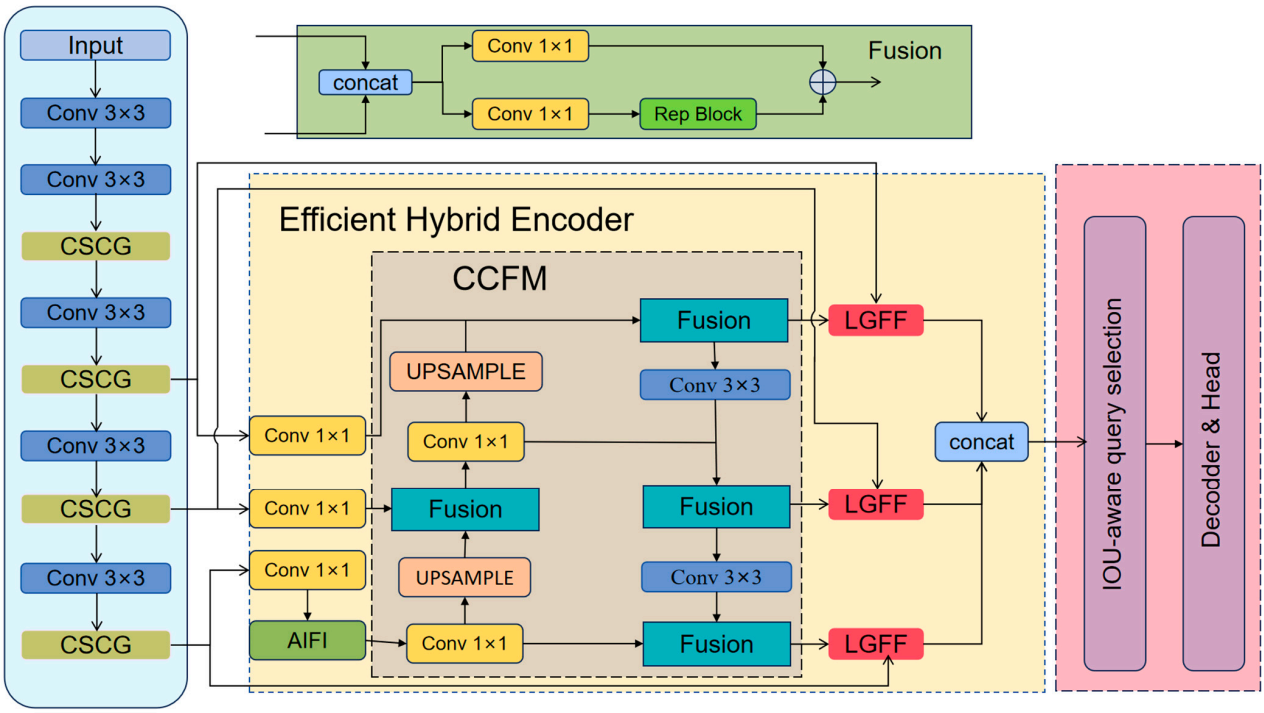


Figure 3. LDW-DETR network architecture diagram.

2.4. LGFF Module

To address the limitations of the baseline model in detecting small-scale lesions and distinguishing fine-grained disease categories in tomato leaves, we introduce a novel LGFF module, inspired by the patch-aware mechanism in the PPA module [38]. The LGFF module utilizes a multi-branch design to extract features across various scales, greatly enhancing the detection performance for minor lesions in tomato leaves. The structure of this module is illustrated in Figure 4.

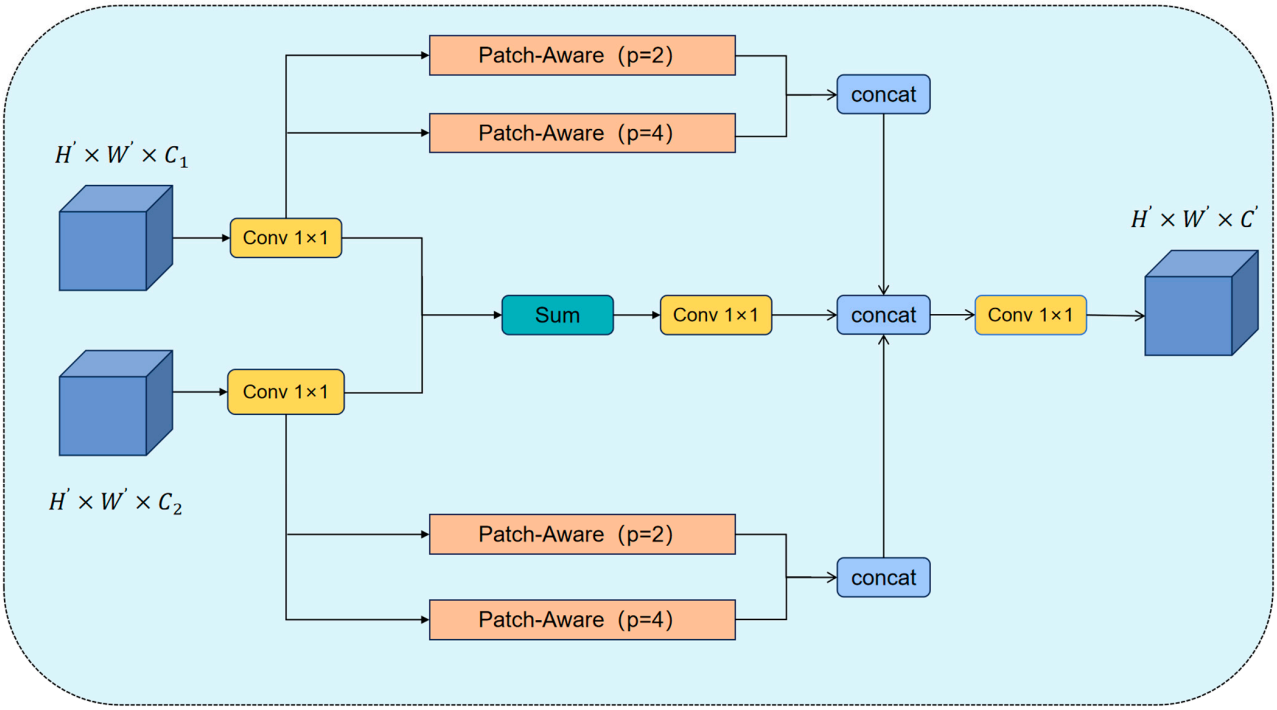


Figure 4. LGFF network architecture diagram.

The LGFF module takes two input tensors, $F_1 = H' \times W' \times C_1$ and $F_1 = H' \times W' \times C_2$. Each input is first processed by a 1×1 convolution to reduce the channel dimensionality, thereby decreasing computational cost and facilitating subsequent feature fusion. Subsequently, the features will branch into three parallel paths. One branch is dedicated to fusing the two input features by performing element-wise addition and dimension reduction operations to construct the baseline feature $F_{conv} = H' \times W' \times K$, where K denotes the number of channels after dimension reduction. The other two branches independently process the two input tensors to extract diverse spatial information. Each of these branches is further divided into a local sub-branch and a global sub-branch, yielding outputs $F_{local} \in H' \times W' \times K_i'$ and $F_{global} \in H' \times W' \times K_i'$, where i denotes the branch for the ith input tensor. This design enables the extraction of multi-scale information under different receptive fields. Specifically, the local sub-branch focuses on capturing fine-grained features within smaller regions, such as leaf textures and edges, thereby enhancing the recognition of small lesions or diseases with highly similar lesion patterns. In contrast, the global sub-branch emphasizes contextual information over larger regions, which facilitates the localization of lesion areas and improves robustness against challenging factors such as illumination variations and leaf occlusions. Finally, the outputs $F_{local} \in H' \times W' \times K_i'$ and $F_{global} \in H' \times W' \times K_i'$ from the two branches are concatenated with the baseline feature $F_{conv} = H' \times W' \times K$ along the channel dimension to fuse multi-scale information. The concatenated result is then passed through a 1×1 convolution for channel reduction, yielding the final output $F' = H' \times W' \times C'$.

In the LGFF module, the local and global sub-branches are primarily distinguished by controlling the parameter p of the patch-aware module, whose structure is illustrated in Figure 5. Specifically, the input tensor is initially processed using reshape and unfold operations to divide it into a sequence of consecutive spatial patches, followed by mean aggregation along the channel dimension. Subsequently, a linear transformation is applied using a feedforward network (FFN), combined with a softmax activation function to obtain a probability distribution in the spatial dimension, thereby completing the feature weighting.

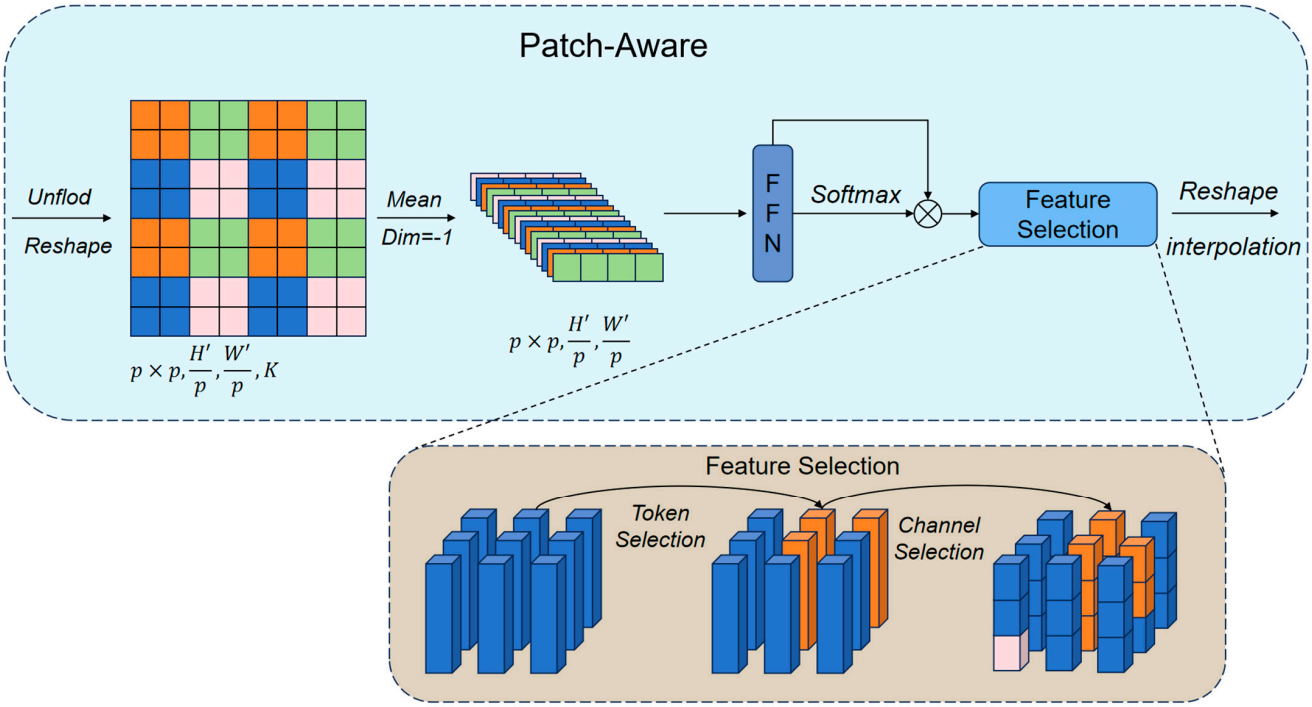


Figure 5. Patch-Aware architecture diagram.

Within the weighted results, feature selection is performed to extract representations that are relevant to the task from both the tokens and the channels. Specifically, let $d = \frac{H' \times W'}{p \times p}$ and let the weighted result be represented as $(t_j)_{j=1}^{c'}$, where $t_j \in R^d$ represents the tokens representing the j-th output. Feature selection is applied to each token, yielding an output denoted as $t'_j = P \times sim(t_j, \xi)$, where $\xi \in R^K$ and $P \in K \times K$ are task-specific parameters. $sim(\cdot, \cdot)$ is a cosine similarity function within the range [0, 1]. Here, ξ as the task embedding, specifying which tokens are relevant to the task. Each token t_j is reweighted according to its relevance to the task embedding, measured by cosine similarity, thereby effectively simulating token selection. Finally, a linear transformation is applied to P to perform channel selection for each token, followed by reshape and interpolation operations to generate the local and global features.

2.5. Improved Backbone

The ResNet-18 architecture, employed as the backbone network in our baseline model, is a classic and widely adopted feature extraction framework for various object detection tasks. However, its structure still has limitations in fine-grained tomato leaf disease detection tasks with small lesion areas. Second, the BasicBlock design of ResNet-18 primarily relies on local convolutions with a limited receptive field, making it difficult to capture long-range dependencies between lesions and surrounding tissues, which in turn constrains the representation of global contextual information.

To address the aforementioned issues, we proposes an improved backbone architecture. First, a lightweight CSP-Darknet is adopted as the base framework. By employing the Cross Stage Partial (CSP) strategy, the network maintains strong feature extraction capability while significantly reducing model computational and parameters complexity, thereby making it suitable for edge deployment and real-time applications. Second, the Strip Block [39] is innovatively combined with the CGLU [40] mechanism to construct the SCG (Strip-CGLU) module, which simultaneously captures both horizontal and vertical strip features as well as fine-grained details, enhancing ability of the model to perceive lesion edges and texture patterns. Finally, the SCG module is deeply integrated with the C2f

module in CSPDarknet to form the CSCG module, enabling effective integration of local detail features with global contextual information. Its architecture diagram is shown in Figure 6.

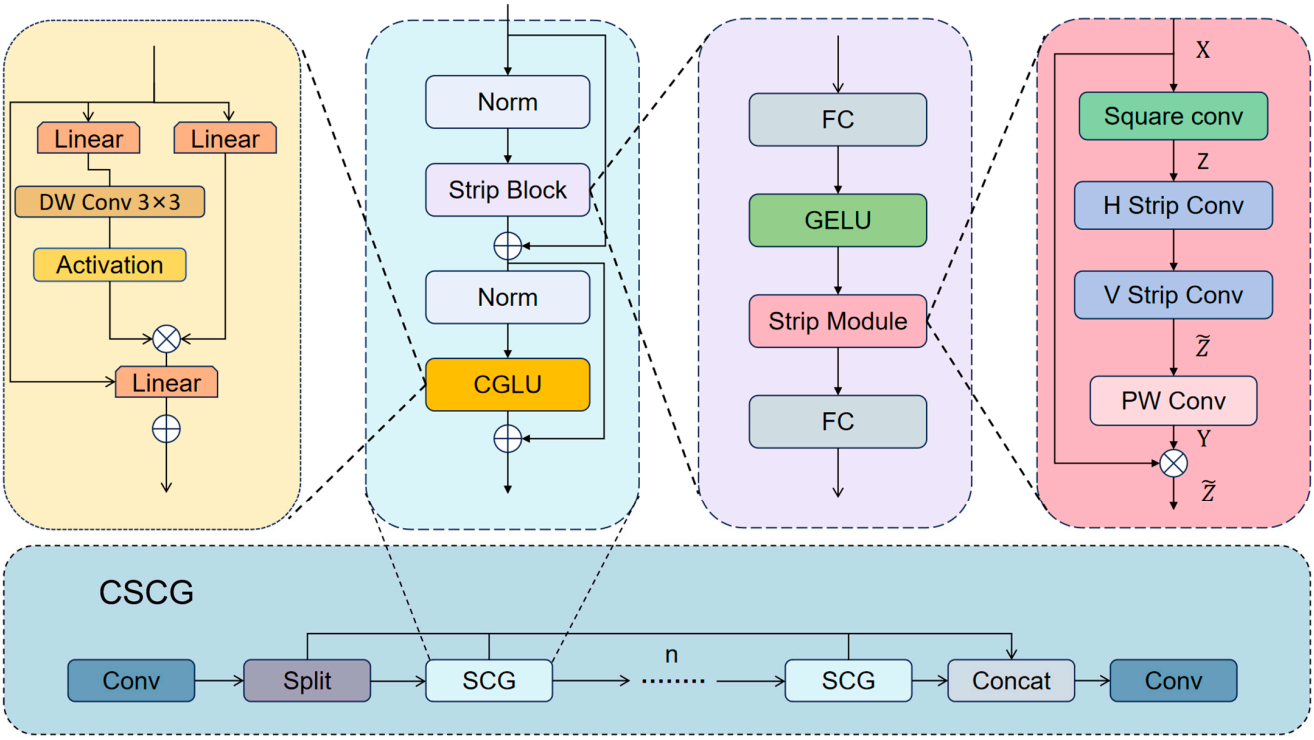


Figure 6. CSCG architecture diagram. ⊗: Element-wise multiplication of matrices; ⊕: Element-wise addition of matrices.

The CSCG module inherits the CSP architecture design of C2f but replaces its core feature extraction unit, the Bottleneck. The original Bottleneck primarily relies on standard convolution operations, which are insufficient for effectively capturing small lesions on tomato leaves as well as elongated lesion patterns along leaf edges. To address this limitation, we completely substitute the Bottleneck with the SCG module. By integrating the Strip Block with the CGLU, the SCG module is capable of extracting strip-like features in both horizontal and vertical directions while adaptively regulating feature responses through a gating mechanism. In this way, the CSCG module not only preserves the efficiency of the CSP framework but also significantly enhances the ability to recognize lesions of varying shapes on tomato leaves. Specifically, the processing flow of the CSCG module is as follows: first, the input feature map undergoes a 1×1 convolution for channel adjustment. The transformed features are then split into two parallel branches: one branch passes through n sequential SCG modules for deep feature extraction, while the other branch is directly forwarded to the fusion stage through cross-layer connection. Subsequently, the outputs from both branches are merged along the channel dimension, and a 1×1 convolution is applied to fuse the combined features. Finally, the refined features are produced as the module output.

The SCG mainly consists of two residual sub-blocks: a Strip Block and a Convolutional CGLU. The Strip Block incorporates a standard small-kernel convolution alongside two large strip-kernel convolutions. This design enables it to effectively extract robust features from objects of diverse aspect ratios. At the core of the Strip Block lies the strip module, which effectively merges the strengths of both standard and strip convolutions, enabling efficient acquisition of core features from objects of varying aspect ratios. Specifically, given an input tensor $X \in R^{C \times H \times W}$ with C channels, a depthwise convolution with

a square kernel of size $K \in R^{c \times k_H \times k_W}$ is first applied to extract local contextual features, resulting in Z. Subsequently, two consecutive depth convolutions with large striped kernels are employed to better capture objects with high aspect ratios. The output is denoted as $\widetilde{Z}$. By incorporating strip convolutions, the network gains the ability to capture directional features along two spatial axes. Compared to traditional convolutions with uniform spatial perception, this approach focuses more intently on an object's shape orientation and edge structure, making it particularly suitable for detecting elongated targets with strong local continuity. Subsequently, a pointwise convolution is used to map $\widetilde{Z}$ into Y. Consequently, every point in the output feature map Y contains information about both horizontal and vertical features gathered from a broad area. Finally, the feature map Y is employed as an attention weight to reweight the input X and the output $\widetilde{Y}$ is obtained through element-wise multiplication.

CGLU is another key component of the SCG module. It extends the original GLU by introducing a lightweight 3×3 depthwise convolution before the activation function in the gating branch. To align with the principle of channel-wise gated attention, this structural adjustment evolves the mechanism into one that utilizes nearest-neighbor features for gated channel attention. In CGLU, a distinct gating signal is assigned to each token, derived from its nearest fine-grained features, thereby improving the model's ability to capture local context. This enables more effective extraction of detailed characteristics such as lesion edges and textures, while maintaining low computational overhead. Consequently, CGLU provides dynamic regulation of feature responses, thereby improving the model's adaptability to complex lesion patterns.

2.6. Wise-IOU

In object detection, the performance of the bounding box regression task is highly dependent on the design of the loss function. Although the RT-DETR model utilizes the GIoU loss, its metric mechanism has inherent drawbacks. A key limitation arises when the predicted box is completely inside the ground truth box; the GIoU loss degenerates to the IoU loss and loses its ability to guide the regression based on positional information. Similarly, the GIoU loss fails to provide discriminative guidance for predicted boxes that are equidistant from the ground truth but in different locations.

To address the aforementioned issues, we introduce the Wise IoU (WIoU) [41] loss function as a replacement for the GIoU loss function. The parameter scheme of WIoU is shown in Figure 7. The green area indicates the predicted bounding box, while the blue area indicates the ground truth bounding box.

WIoU comprises three versions: WIoU v1, WIoU v2, and WIoU v3. Specifically, WIoU v1 loss function builds upon the traditional IoU loss function by incorporating a penalty term for the distance between the center points of the predicted and ground truth bounding boxes. This approach reduces the influence of geometric factors, thereby enhancing the model's generalization capability.

The loss function of WIoU v1 is calculated as follows:

$$L_{WIoU\ v1} = R_{WIoU} L_{IoU} \tag{1}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2}\right) \tag{2}$$

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{S_u} \tag{3}$$

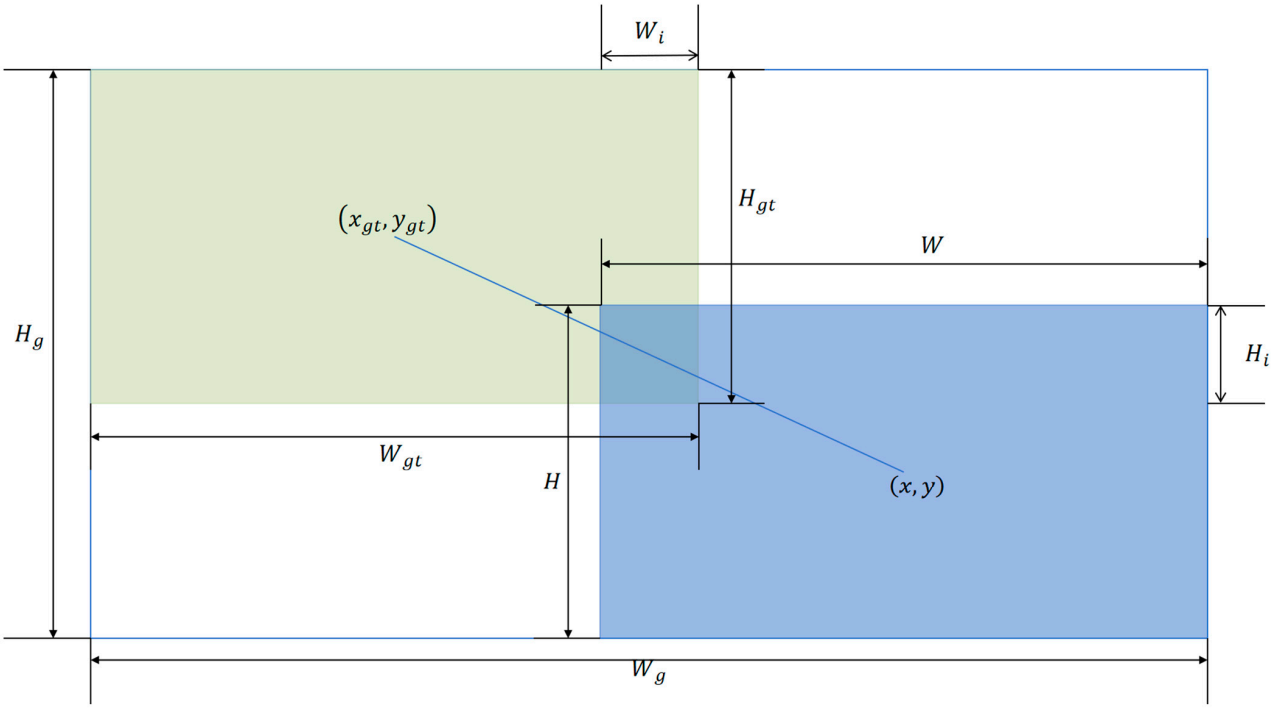


Figure 7. Schematic diagram of WIoU parameters.

Here, W_i and H_i denote the width and height of the overlapping region between the predicted bounding box and the ground truth bounding box, respectively, while S_u represents the area of the overlapping region. x and y are used to represent the center coordinates of the ground truth box, with x_{gt} and y_{gt} representing those of the predicted box. W_g and H_g correspond to the width and height of the smallest enclosing rectangle for both the ground truth and predicted boxes.

On the basis of WIoU v1, WIoU v3 is constructed by introducing a non-monotonic focusing factor. The loss function of WIoU v3 is calculated as follows:

$$L_{WIoU\ v3} = rL_{WIoU\ v1} \tag{4}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{5}$$

$$\beta = \frac{L_{IoU}^{*}}{L_{IoU}^{-}} \in (0, +\infty) \tag{6}$$

Here, r denotes the non-monotonic focusing factor, while α and β represent hyper-parameters. β represents the outlier degree. A smaller β indicates higher anchor quality, in which case a smaller gradient gain is assigned to focus the regression on anchors of regular quality. Conversely, a larger β corresponds to lower-quality anchors, for which a smaller gradient gain is allocated to reduce their adverse impact on model accuracy. L_{IoU}^{*} is the monotonic focusing factor, and L_{IoU}^{-} denotes the exponential moving average with a momentum value of m. Since L_{IoU}^{-} is dynamically updated, the criterion for anchor quality division also changes adaptively. This enables WIoU v3 to continuously adjust its gradient gain allocation strategy, ensuring optimal adaptation to the current training context.

WIoU v3 introduces a non-monotonic focusing factor, which allows the model to pay more attention to medium-quality anchors and dynamically adjust anchor quality partitioning for optimal gradient gain allocation. This design mitigates the excessive penalties caused by distance and aspect ratio factors, thereby improving detection accuracy.

3. Results

3.1. Experimental Setup

The training and testing of the model were conducted on the Ubuntu 22.04.4 operating system with an NVIDIA GeForce RTX 3080 GPU. The software environment comprised Python 3.8.10, CUDA 12.1, and PyTorch 2.0.1. No pre-trained weight files were utilized during model training. Hyperparameter tuning was performed prior to model training, with specific hyperparameter values detailed in Table 2.

Table 2. Model training parameters.

Type	Value
Epoch	100
Batch Size	16
Image Size	640 × 640
Lr0	0.0001
Lrf	1
Momentum	0.9
Weight Decay	0.0001
Optimizer	AdamW

3.2. Evaluation Metrics

To comprehensively and thoroughly evaluate the model's detection effectiveness and performance, we selected a scientifically sound and representative set of evaluation metrics. Specifically, these include Precision (P), Recall (R), mAP@0.5, mAP@0.5–0.95, FPS, the number of model parameters and GFLOPs.

Precision measures how many of the predicted positive samples are actually true positives, and recall shows the percentage of actual positive samples that the model correctly identifies. The number of parameters serves as a key indicator of model complexity, while GFLOPs evaluate the computational cost during model inference. mAP@0.5 and mAP@0.5–0.95 denote the mean average precision at different IoU thresholds. The mathematical formulas for these metrics are as follows:

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

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \frac{1}{c}\sum_{t=1}^{c} AP_i \tag{10}$$

In these equations, TP (True Positive) refers to the count of positive samples correctly identified by the model; FP (False Positive) indicates the count of negative samples incorrectly classified as positive; and FN (False Negative) denotes the count of positive samples incorrectly classified as negative. AP denotes Average Precision, the average precision of the i-th object category, where C represents the total number of categories.

3.3. Per-Class Detection Performance

To provide a detailed evaluation of the proposed LDW-DETR model, we analyzed its detection performance for each disease category in the main test set. Table 3 presents the detection results of LDW-DETR for various diseases in the dataset.

Table 3. Per-class detection performance evaluation.

Class	P/%	R/%	mAP@0.5 (%)	mAP@0.5–0.95 (%)
brown spot	0.825	0.818	0.889	0.743
bacterial	0.912	0.904	0.948	0.805
healthy	0.846	0.862	0.908	0.758
late blight	0.869	0.817	0.893	0.747
yellow curl virus	0.784	0.743	0.869	0.672
mean	0.847	0.829	0.901	0.745

As shown in Table 3, the model demonstrates relatively suboptimal performance in detecting yellow leaf curl virus, with all its evaluation metrics significantly lower than other categories. This is attributed to the high visual similarity between the characteristics of this disease and other categories. For example, the overall yellowing and leaf curling symptoms induced by yellow leaf curl virus can be easily confused with natural yellowing patterns occurring during the growth of healthy leaves, which directly leads to a higher false detection rate. In contrast, the model achieves the most outstanding detection performance for bacterial diseases, with its mAP@0.5 reaching 94.8%, indicating its capability to accurately learn the unique and highly discriminative feature patterns of this category. The prediction results for other categories remain stable, with final precision, recall, mAP@0.5, and mAP@0.5–0.95 reaching 84.7%, 82.9%, 90.1%, and 74.5%, respectively, demonstrating excellent overall detection performance and a balanced capability in identifying different categories of tomato leaf diseases.

3.4. Ablation Study

To assess the effect of the three proposed modules, an ablation study was performed using RT-DETR-R18 as the baseline model. The experimental results are presented in Table 4.

Table 4. Ablation protocol.

No.	Improved CSPDarknet	LGFF	WIoU v3	P/%	R/%	mAP@0.5 (%)	mAP@0.5–0.95 (%)	Params (M)	FPS
1				82.2	79.3	87.5	70.8	19.8	68.10
2	√			81.7	81.2	88.2	71.2	13.31	73.74
3		√		83.2	81.7	89.1	73.3	22.79	66.23
4			√	83.9	82.6	89.4	71.6	19.8	71.25
5	√	√		82.3	83.1	88.8	71.8	16.25	72.59
6	√	√	√	84.7	82.9	90.1	74.5	16.25	76.01

As shown in Table 4, replacing the backbone network of the baseline model with Improved CSPDarknet resulted in a slight decrease of 0.4% in precision. However, recall, mAP@0.5, and mAP@0.5–0.95 all exhibited slight improvements, while the model's parameter count decreased by 32.8%. This demonstrates that the Improved CSPDarknet achieves model lightweighting and detection efficiency enhancement without compromising detection performance. When the LGFF module was integrated into the baseline model, the number of parameters and FPS increased by 15.1% and 2.7%; however, precision, recall, mAP@0.5 and mAP@0.5:0.95 improved by 1.0%, 3.9%, 1.6% and 2.5%, respectively. This demonstrates the effectiveness of the global–local feature fusion strategy specifically designed for small-scale lesions and fine-grained features, which significantly enhances detection accuracy. WIoU v3 loss function, by leveraging a non-monotonic focusing factor to differentiate anchor box quality, further improved precision, recall, mAP@0.5, mAP@0.5:0.95 and FPS by 1.7%, 4.8%, 1.9%, 0.8% and 4.6%, respectively.

When Improved CSPDarknet and LGFF were used jointly, precision, recall, mAP@0.5, mAP@0.5:0.95 and FPS increased by 0.1%, 5.3%, 1.3%, 1.0%, and 6.6%, respectively, while the parameter count decreased by 17.9%, highlighting their complementary advantages. Finally, when Improved CSPDarknet, LGFF, and WIoU v3 were integrated together, the model achieved the best performance, with precision, recall, mAP@0.5, mAP@0.5:0.95 and FPS improved by 2.5%, 5.1%, 2.6%, 3.7% and 11.6%, respectively, while the parameter count decreased by 17.9%. These results demonstrate that, compared with the original RT-DETR model, our improved model effectively combines the advantages of Improved CSPDarknet, LGFF, and WIoU v3, achieving higher detection accuracy while significantly reducing model parameters and computational complexity, thereby striking an optimal balance between detection performance and computational efficiency.

In the improved CSPDarknet, we proposed the CSCG module, composed of Strip Block and CGLU, to replace the bottleneck in C2f. To further validate the effectiveness of this module, we conducted an ablation study, and the results are shown in Table 5.

Table 5. Ablation protocol on the improved CSPDarknet.

No.	Strip Block	CGLU	P/%	R/%	mAP@0.5 (%)	mAP@0.5–0.95 (%)	Params (M)	FPS
1			82.2	79.3	87.5	70.8	19.8	68.10
2	√		82.3	80.1	87.6	70.6	14.3	73.41
3		√	81.9	80.8	87.9	71.1	18.2	70.33
4	√	√	81.7	81.2	88.2	71.2	13.31	73.74

As shown in Table 5, after incorporating the Strip Block module, the model's precision, recall, and mAP@0.5 increased by 0.1%, 0.8%, and 0.1%, respectively, while the number of model parameters decreased by 25.2% and the FPS improved by 7.8%. Although mAP@0.5–0.95 slightly decreased by 0.2%, all other metrics showed improvement. After using CGLU, the model's precision slightly decreased, but the other metrics experienced modest gains. When Strip Block and CGLU were used together to form the CSCG module, the model's precision decreased by 0.5%, but the key mAP@0.5 increased by 0.7%, and both recall and mAP@0.5–0.95 showed slight improvements. Additionally, the model parameters decreased by 32.8%, and FPS increased by 7.6%. These results indicate that after replacing the backbone with the improved CSPDarknet, the model achieves higher accuracy while also benefiting from faster inference speed and reduced parameter count.

3.5. Component Evaluation

3.5.1. Comparative Experiments on Different Loss Functions

To investigate the impact of different loss functions on the RT-DETR model and identify the most suitable one, we replaced the GIoU function in RT-DETR with SIoU, PIoU, DIoU, Focaler IoU, and WIoU v3, respectively, and compared their performance in tomato leaf disease detection. The results are shown in Table 6.

Table 6. Comparison with other Loss Functions.

Model	P/%	R/%	mAP@0.5 (%)
GIoU	82.2	79.3	87.5
SIoU	81.7	82.4	88.0
PIoU	83.3	80.3	86.3
DIoU	81.2	82.7	88.3
Focaler IoU	82.3	83.8	88.8
WIoU v3	83.9	82.6	89.4

SIoU takes geometric factors into account. When the GIoU loss was replaced with SIoU, the model's precision slightly decreased to 81.7%, while other evaluation metrics showed modest improvements. PIoU incorporates uncertainty modeling to enhance the performance of the traditional IoU. After using PIoU, the model's precision and recall increased to 83.3% and 80.3%, respectively, although mAP@0.5 decreased to 86.3%. DIoU takes into account not only the overlap between the predicted and ground-truth boxes but also the distance between their centers, which enhances both the convergence speed and localization accuracy of bounding box regression. With DIoU, precision dropped to 81.2%, but recall and mAP@0.5 improved to 82.7% and 88.3%, respectively. Focaler IoU modifies the standard IoU loss by mapping it onto a linear region, enabling the model to focus on hard-to-detect samples and improving both generalization and detection accuracy. Using Focaler IoU, precision, recall, and mAP@0.5 increased to 82.3%, 83.8%, and 88.8%, respectively. WIoU v3 achieved the best overall balance, with a precision of 83.9%, recall of 82.6%, and mAP@0.5 of 89.4%, all higher than the original GIoU. Although its recall is slightly lower than that of DIoU and Focaler IoU, WIoU v3 outperforms all other loss functions in terms of precision and mAP@0.5, demonstrating its effectiveness for tomato leaf disease detection.

3.5.2. Comparative Experiments on Different Backbone

To further investigate the effectiveness of our proposed improved CSPDarknet architecture, we conducted comparative experiments with several other mainstream lightweight backbones. The results are shown in Table 7.

Table 7. Comparison with other Backbone.

Model	P/%	R/%	mAP@0.5 (%)	mAP@0.5–0.95 (%)	Params (M)	FPS
MobileNetv4	74.3	75.0	84.6	68.1	10.8	90.1
ShuffleNetV2	79.2	73.4	83.2	70.7	12.1	80.5
Fasternet	77.9	76.3	85.3	70.3	11.3	86.3
Ours	81.7	81.2	88.2	71.2	13.31	73.74

As shown in Table 7, after replacing the backbone with other mainstream lightweight networks, the model achieves a certain improvement in FPS and a reduction in parameter count compared with our proposed backbone. However, this comes at the cost of a significant drop in key metrics such as precision, recall, and mAP@0.5. This indicates that such lightweight backbones improve computational efficiency at the expense of accuracy, failing to achieve an optimal balance between the two. In contrast, our improved CSPDarknet backbone not only enhances computational efficiency based on the baseline model but also improves detection accuracy, achieving a better balance between efficiency and precision.

3.6. Comparative Experiments on Different Detection Models

To thoroughly assess the performance of the proposed LDW-DETR model for tomato leaf disease detection, we compared it with several leading object detection algorithms, such as YOLOv5m, YOLOv6m, YOLOv8m, YOLOv10m, YOLOv11m, DINO and Deformable-DETR. The results are shown in Table 8.

From the data presented in Table 8, the proposed LDW-DETR model achieved a precision of 84.7%, a recall of 82.9%, a mAP@0.5 of 90.1%, a mAP@0.5–0.95 of 74.5% and a FPS of 76.1, with a parameter count of 16.25M and a computational cost of 54.7 GFLOPs. Compared with YOLOv5m, YOLOv6m, YOLOv8m, YOLOv10m, YOLOv11m, DINO and Deformable-DETR, the LDW-DETR model improved precision by 1.1%, 3.0%, 3.7%, 1.7%, and 0.9%, respectively; recall by 2.9%, 1.7%, 0.7%, 3.8%, and 0.8%; mAP@0.5 by 3.7%, 1.9%,

2.9%, 3.0%, and 0.7%; and mAP@0.5–0.95 by 4.0%, 3.3%, 2.0%, 2.0%, and 1.4%. Regarding computational cost, reductions of 14.5%, 66.1%, 30.5%, 7.1%, and 19.2% were observed, respectively. In terms of model size and FPS, although LDW-DETR has a larger model size and lower FPS compared to YOLOv10m, it achieves higher precision and mAP@0.5 in key metrics. Moreover, when compared to other algorithms, it still maintains a more compact model size and higher FPS. Overall, the LDW-DETR model demonstrates improvements in core metrics such as mAP@0.5, mAP@0.5–0.95, precision, and recall, while simultaneously reducing computational cost and maintaining a compact model size. These results indicate that LDW-DETR maintains stable recognition performance in complex field scenarios, validating its applicability and effectiveness for tomato leaf disease detection.

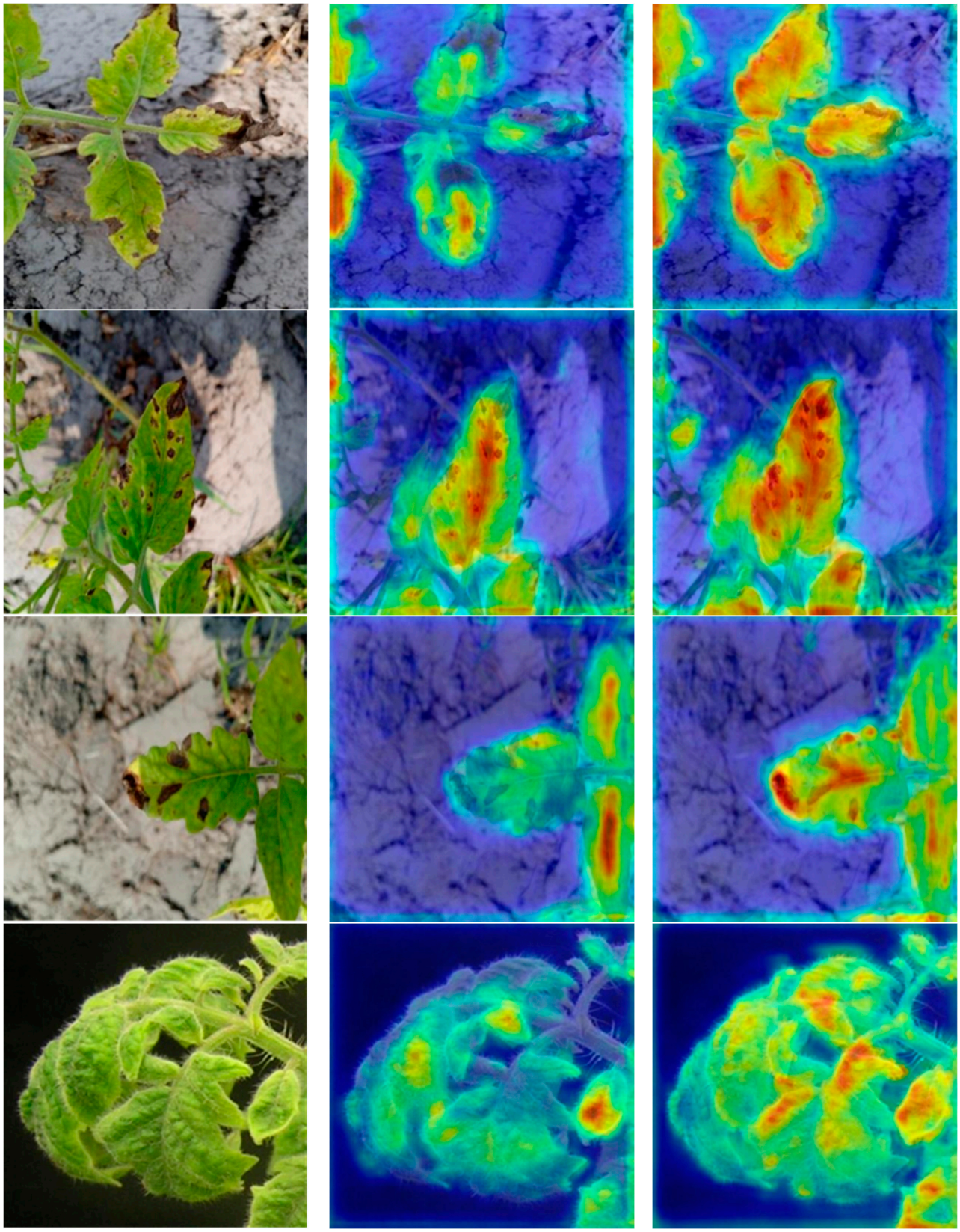
Table 8. Comparison with other models.

Model	P/%	R/%	mAP@0.5 (%)	mAP@0.5–0.95 (%)	Params (M)	GFLOPs (G)	FPS
Yolov5m	83.6	80.0	86.4	70.5	25.0	64.0	62.3
Yolov6m	81.7	81.2	88.2	71.2	51.9	161.2	50.1
Yolov8m	81.0	82.2	87.2	72.5	25.8	78.7	70.3
Yolov10m	83.0	79.1	87.1	72.5	15.3	58.9	84.1
Yolov11m	83.8	82.1	89.4	73.1	20.0	67.7	72.3
DINO	82.3	79.5	87.5	72.1	47.5	-	9.1
Deformable-DETR	82.8	81.7	88.6	73.4	40.1	-	11.2
Ours	84.7	82.9	90.1	74.5	16.2	54.7	76.1

3.7. Visual Analysis

To further demonstrate the effectiveness of LDW-DETR in tomato leaf disease detection, Grad-CAM [42] was employed for visualization analysis to compare the heatmaps generated by RT-DETR and LDW-DETR, as shown in Figure 8. In the heatmaps, warmer colors indicate regions that the model focuses on more and that have a greater influence on its predictions, whereas cooler colors correspond to regions with lower attention, which typically represent background or irrelevant areas.

As shown in the heatmaps presented in Figure 8, both the RT-DETR and LDW-DETR models exhibit higher attention to the target regions compared with background areas when detecting tomato leaf diseases. This indicates that both models have sufficiently learned feature representations during training and are capable of accurately identifying the diseased regions. However, analysis of Figure 8b reveals certain limitations of the RT-DETR model. Although it roughly focuses on the target areas, the attention is relatively dispersed, with lower focus on some disease spots. Furthermore, the model fails to capture key details such as lesion edges and small disease spots accurately and comprehensively, leading to omissions in feature extraction. For instance, in the first and third rows of Figure 8b, cooler tones cover some obvious diseased regions of the tomato leaves, indicating low attention to these areas. In the second and fourth rows, the warm-tone coverage is limited, suggesting insufficient focus on fine lesions. In contrast, Figure 8c demonstrates that the LDW-DETR model significantly enhances attention to the core regions of leaf diseases. The warm-tone regions align more closely with the actual disease distribution and can more sensitively capture fine lesions, lesion boundaries, and the edges between diseased and healthy tissue. The higher correspondence between warm-tone regions and actual disease features indicates that the improved model extracts critical features of tomato leaf diseases more effectively, resulting in more accurate identification.



(**a**) (**b**) (**c**)

Figure 8. Heatmaps: (**a**) Original images; (**b**) Baseline model heatmap; (**c**) Our model heatmap.

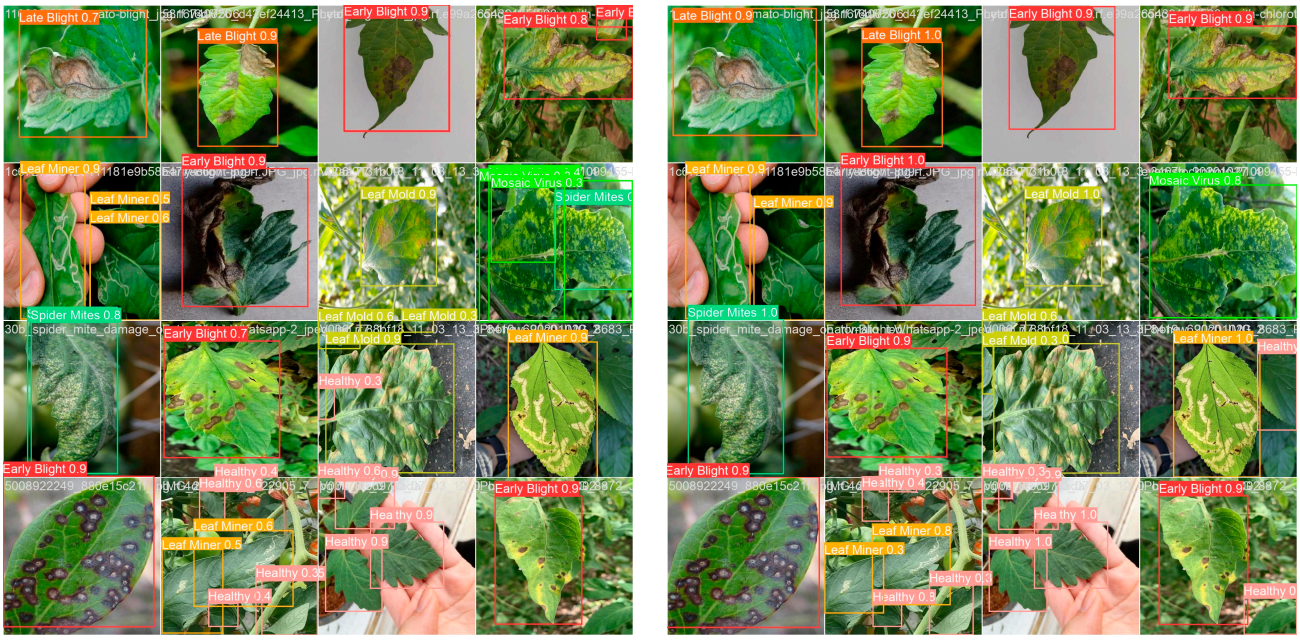
3.8. Generalization Experiment

Tomato leaf diseases exhibit diverse and complex patterns, and their appearance can vary under different cultivation environments, lighting conditions, and imaging devices. To further evaluate the robustness and generalization capability of the proposed model, a generalization experiment was conducted using a publicly available dataset from Roboflow. This dataset contains nine different types of tomato leaf diseases, with images collected under various natural conditions and using different imaging devices, providing a comprehensive test for the model's generalization ability. The experimental environment and parameter settings were kept consistent with those described in Section 3.1, and the results are presented in Table 9.

Table 9. Generalization Experiment Results.

Model	P (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5–0.95 (%)	GFLOPs (G)	Paras (M)
Baseline	95.7	95.0	97.2	91.0	57.0	19.9
Ours	96.1	95.1	98.1	91.5	54.7	16.2

From Table 9, it can be observed that the proposed LDW-DETR model remains effective on the public Roboflow dataset. Specifically, the precision and recall increased by 0.4% and 0.1%, respectively, while mAP@0.5 and mAP@0.5–0.95 improved by 0.9% and 0.5%. In addition, both GFLOPs and the number of parameters are lower than those of the baseline model, which fully demonstrates the generalization capability and effectiveness of the proposed model in tomato leaf disease detection. To provide a more intuitive comparison of detection performance, the results of both models are visualized in Figure 9.



(**a**) (**b**)

Figure 9. Visualization of detection results: (**a**) Baseline model; (**b**) Our model.

As shown in Figure 9a, the baseline model exhibits several issues, including misaligned bounding boxes with lesion areas, false positives, and missed detections. In some cases, it also generates multiple category labels accompanied by redundant bounding boxes, which leads to reduced stability and lower accuracy. In contrast, Figure 9b demonstrates that our proposed model produces bounding boxes that more precisely align with the actual lesion

regions. Even under challenging conditions such as complex backgrounds and overlapping leaves, the model can accurately identify tomato leaf disease categories while significantly reducing false and missed detections. Moreover, the predicted confidence scores of our model are generally higher than those of the baseline model, enhancing the reliability and stability of the results. These findings clearly demonstrate that the proposed model possesses strong effectiveness and robustness for tomato leaf disease detection, making it well-suited for applications under varying environmental conditions.

4. Conclusions

In this paper, an improved tomato leaf disease detection model LDW-DETR based on RT-DETR is proposed, which is mainly used for the detection of various diseases of tomato leaves. In order to cope with the challenges of low recognition ability of small-sized lesions and insufficient ability to distinguish diseases with similar characteristics, we have made the following innovations: First, the LGFF module is designed to realize the adaptive fusion of local fine-grained features and global context features thru a multi-branch structure, which effectively enhances the model's ability to identify small lesions and similar diseases in complex backgrounds. Secondly, CSPDarknet was introduced as the backbone network, and the bottleneck layer of C2f was improved by combining Strip Block and CGLU, which significantly improved the perception ability of slender lesions and edge texture features while reducing the complexity of the model. Finally, the WIoU v3 loss function is introduced, and the dynamic adjustment mechanism of the non-monotonic focusing factor is used to optimize the bounding box regression process, so that the model can allocate gradients more reasonably during training, thereby improving the positioning accuracy and stability.

The experimental results show that LDW-DETR improves mAP@0.5 by 2.6% and mAP@0.5–0.95 by 3.7% compared with the baseline model RT-DETR-R18, while the model parameter amount is reduced by 17.9%, realizing an effective balance between detection accuracy and model lightweight. Ablation experiments further confirmed the independent contribution and synergistic effect of each core module, and the comparison results with the mainstream detection algorithms YOLOv5m, YOLOv6m, YOLOv8m, YOLOv10m and YOLOv11m further verified the superiority of the proposed model. LDW-DETR shows the best comprehensive performance in terms of accuracy, recall rate and computational cost, and still maintains high robustness and generalization ability in the generalization experiment. The visualization results of the heat map show that the improved model can pay more attention to the lesion area, and the response to small-scale lesions and edge areas is stronger, showing higher superiority.

Although this study demonstrated strong performance in laboratory settings, bridging the gap to real-world applications remains a key challenge. Leveraging the lightweight and efficient design of LDW-DETR, our future work will focus on its deployment in practical agricultural environments. A major direction is to explore the model's real-time use on UAV-based remote sensing platforms. To achieve this, we plan to apply model pruning and quantization techniques to optimize LDW-DETR for embedded edge computing devices, enabling on-board analysis of UAV video streams for large-scale, dynamic monitoring and localization of crop diseases.

At the same time, we aim to bring this technology to mobile platforms by developing a lightweight diagnostic tool for farmers and agricultural technicians. Using a smartphone camera, users would be able to quickly and accurately identify plant diseases directly in the field, greatly enhancing the accessibility of intelligent agricultural services.

To further improve robustness under complex field conditions, we also plan to explore multimodal data fusion approaches, such as combining RGB images with infrared or

hyperspectral data to overcome the limitations of single-modality inputs. In addition, we intend to build a larger and more diverse field dataset to strengthen the model's generalization and real-world adaptability.

We believe that these application-driven efforts will help transform precision agriculture from algorithmic research into practical, field-ready solutions.

Author Contributions: Conceptualization, H.Y. and H.X.; methodology, H.X.; software, Y.L.; validation, M.M.; formal analysis, T.T. and Z.H.; investigation, H.X. and T.T.; resources, H.Y.; data curation, Z.H.; writing—original draft preparation, H.X.; writing—review and editing, H.Y.; visualization, H.X. and Y.L.; supervision, M.M. and T.T.; project administration, H.Y.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hubei Provincial Natural Science Foundation (2025AFC122); the Ministry of Education Industry-University Cooperation Education Project (231106627155856); 2025 ESI (Engineering) (01003009); the School Enterprise Cooperation Project (No. whpu-2024-kj-4582, No. whpu-2024-kj-4639, No. whpu-2025-kj-5347); University Graduate Teaching and Research Project, Research on Interdisciplinary Collaborative Innovation and Joint Training Mechanism in Artificial Intelligence (2025JY05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: These datasets can be found at https://www.kaggle.com/datasets/nirmalsankalana/plantdoc-dataset (accessed on 28 September 2025), https://universe.roboflow.com/plant-disease-detection/fieldplant/dataset/11 (accessed on 28 September 2025) and https://www.kaggle.com/datasets/emmarex/plantdisease (28 September 2025).

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

References

1. Al-gaashani, M.S.; Shang, F.; Muthanna, M.S.; Khayyat, M.; Abd El-Latif, A.A. Tomato leaf disease classification by exploiting transfer learning and feature concatenation. *IET Image Process.* **2022**, *16*, 913–925. [CrossRef]
2. Li, N.; Wu, X.; Zhuang, W.; Xia, L.; Chen, Y.; Wu, C.; Rao, Z.; Du, L.; Zhao, R.; Yi, M. Tomato and lycopene and multiple health outcomes: Umbrella review. *Food Chem.* **2021**, *343*, 128396. [CrossRef]
3. Panno, S.; Davino, S.; Caruso, A.G.; Bertacca, S.; Crnogorac, A.; Mandić, A.; Noris, E.; Matić, S. A review of the most common and economically important diseases that undermine the cultivation of tomato crop in the mediterranean basin. *Agronomy* **2021**, *11*, 2188. [CrossRef]
4. Batool, A.; Hyder, S.B.; Rahim, A.; Waheed, N.; Asghar, M.A. Classification and identification of tomato leaf disease using deep neural network. In Proceedings of the 2020 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 22–23 February 2020; pp. 1–6.
5. Wang, W.; Liu, M.; Zhao, Z.; Wang, J. Improved convolutional neural network method for classificationof apple leaves. *J. South-Cent. Minzu Univ.* **2022**, *41*, 71–78.
6. Wang, Y.; Zhang, P.; Tian, S. Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion. *Front. Plant Sci.* **2024**, *15*, 1382802. [CrossRef] [PubMed]
7. Dutot, M.; Nelson, L.; Tyson, R. Predicting the spread of postharvest disease in stored fruit, with application to apples. *Postharvest Biol. Technol.* **2013**, *85*, 45–56. [CrossRef]
8. Trippa, D.; Scalenghe, R.; Basso, M.F.; Panno, S.; Davino, S.; Morone, C.; Giovino, A.; Oufensou, S.; Luchi, N.; Yousefi, S. Next-generation methods for early disease detection in crops. *Pest Manag. Sci.* **2024**, *80*, 245–261. [CrossRef]
9. Yang, H.; Shu, Z.; Li, Z. Enhanced Wind Power Forecasting Using a Hybrid Multi-Strategy Coati Optimization Algorithm and Backpropagation Neural Network. *Sensors* **2025**, *25*, 2438. [CrossRef]
10. Jiang, F.; Han, X.; Wen, S.; Tian, T. Spatiotemporal interactive learning dynamic adaptive graph convolutional network for traffic forecasting. *Knowl. Based Syst.* **2025**, *311*, 113115. [CrossRef]
11. Jiang, F.; Wu, B.; Wen, S.; Tian, T. MSDIPN: Multi-scale Deep Interval Prediction Network for Multivariate Time Series. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 5452–5466. [CrossRef]
12. Yang, H.; Zhang, S.; Shen, H.; Zhang, G.; Deng, X.; Xiong, J.; Feng, L.; Wang, J.; Zhang, H.; Sheng, S. A multi-layer feature fusion model based on convolution and attention mechanisms for text classification. *Appl. Sci.* **2023**, *13*, 8550. [CrossRef]

13. Yang, H.; Lin, D.; Zhang, G.; Zhang, H.; Wang, J.; Zhang, S. Research on detection of rice pests and diseases based on improved YOLOV5 algorithm. *Appl. Sci.* **2023**, *13*, 10188. [CrossRef]
14. Saleem, M.H.; Potgieter, J.; Arif, K.M. Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precis. Agric.* **2021**, *22*, 2053–2091. [CrossRef]
15. Zhang, M.; Liu, C.; Li, Z.; Yin, B. From Convolutional Networks to Vision Transformers: Evolution of Deep Learning in Agricultural Pest and Disease Identification. *Agronomy* **2025**, *15*, 1079. [CrossRef]
16. Mao, T.; Wu, J. Classification and adulteration degree of Maojian tea based on multi-scale and attention mechanism. *J. South-Cent. Minzu Univ.* **2024**, *43*, 790–796. [CrossRef]
17. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1440–1448. [CrossRef]
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
20. Gong, X.; Zhang, S. A high-precision detection method of apple leaf diseases using improved faster R-CNN. *Agriculture* **2023**, *13*, 240. [CrossRef]
21. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
22. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [CrossRef]
23. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
24. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
25. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–21.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
27. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
28. Wang, Y.; Wang, Y.; Zhao, J. MGA-YOLO: A lightweight one-stage network for apple leaf disease detection. *Front. Plant Sci.* **2022**, *13*, 927424. [CrossRef] [PubMed]
29. Abulizi, A.; Ye, J.; Abudukelimu, H.; Guo, W. DM-YOLO: Improved YOLOv9 model for tomato leaf disease detection. *Front. Plant Sci.* **2025**, *15*, 1473928. [CrossRef]
30. Yang, S.X.; Yao, J.F.; Teng, G.F. Corn Leaf Spot Disease Recognition Based on Improved YOLOv8. *Agriculture* **2024**, *14*, 666. [CrossRef]
31. Xiong, C.; Ma, S.; Gao, Z.; Li, F.; Chen, W. Deep reconstruction for compressed sensing based on multi-resolution feature fusion with Transformer. *J. South Cent. Minzu Univ.* **2025**, *44*, 400–406. [CrossRef]
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
33. Yang, H.; Deng, X.; Shen, H.; Lei, Q.; Zhang, S.; Liu, N. Disease detection and identification of rice leaf based on improved detection transformer. *Agriculture* **2023**, *13*, 1361. [CrossRef]
34. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
35. Hughes, D.; Salathé, M. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv* **2015**, arXiv:1511.08060.
36. Moupojou, E.; Tagne, A.; Retraint, F.; Tadonkemwa, A.; Wilfried, D.; Tapamo, H.; Nkenlifack, M. FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning. *IEEE Access* **2023**, *11*, 35398–35410. [CrossRef]
37. Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad, India, 5–7 January 2020; pp. 249–253.
38. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. Hcf-net: Hierarchical context fusion network for infrared small object detection. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 15–19 July 2024; pp. 1–6.
39. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. *arXiv* **2025**, arXiv:2501.03775. [CrossRef]

40. Shi, D. Transnext: Robust foveal visual perception for vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17773–17783.
41. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

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