


Article

# D<sup>3</sup>-YOLOv10: Improved YOLOv10-Based Lightweight Tomato Detection Algorithm Under Facility Scenario

Ao Li <sup>1,\*</sup> , Chunrui Wang <sup>1</sup>, Tongtong Ji <sup>1</sup>, Qiyang Wang <sup>2</sup> and Tianxue Zhang <sup>3,4</sup>

<sup>1</sup> School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China; 2320410131@stu.hrbust.edu.cn (C.W.); 2320400008@stu.hrbust.edu.cn (T.J.)

<sup>2</sup> School of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China; qywang@ujs.edu.cn

<sup>3</sup> School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China; tianxuezhang@buaa.edu.cn

<sup>4</sup> Institute of Medical Robotics, Shanghai Jiaotong University, Shanghai 200240, China

\* Correspondence: ao.li@hrbust.edu.cn

**Abstract:** Accurate and efficient tomato detection is one of the key techniques for intelligent automatic picking in the area of precision agriculture. However, under the facility scenario, existing detection algorithms still have challenging problems such as weak feature extraction ability for occlusion conditions and different fruit sizes, low accuracy on edge location, and heavy model parameters. To address these problems, this paper proposed D<sup>3</sup>-YOLOv10, a lightweight YOLOv10-based detection framework. Initially, a compact dynamic faster network (DyFasterNet) was developed, where multiple adaptive convolution kernels are aggregated to extract local effective features for fruit size adaptation. Additionally, the deformable large kernel attention mechanism (D-LKA) was designed for the terminal phase of the neck network by adaptively adjusting the receptive field to focus on irregular tomato deformations and occlusions. Then, to further improve detection boundary accuracy and convergence, a dynamic FM-WIoU regression loss with a scaling factor was proposed. Finally, a knowledge distillation scheme using semantic frequency prompts was developed to optimize the model for lightweight deployment in practical applications. We evaluated the proposed framework using a self-made tomato dataset and designed a two-stage category balancing method based on diffusion models to address the sample class-imbalanced issue. The experimental results demonstrated that the D<sup>3</sup>-YOLOv10 model achieved an  $mAP_{0.5}$  of 91.8%, with a substantial reduction of 54.0% in parameters and 64.9% in FLOPs, compared to the benchmark model. Meanwhile, the detection speed of 80.1 FPS more effectively meets the demand for real-time tomato detection. This study can effectively contribute to the advancement of smart agriculture research on the detection of fruit targets.

**Keywords:** tomato detection; YOLOv10; occlusion recognition; attention mechanism; knowledge distillation



**Citation:** Li, A.; Wang, C.; Ji, T.; Wang, Q.; Zhang, T. D<sup>3</sup>-YOLOv10: Improved YOLOv10-Based Lightweight Tomato Detection Algorithm Under Facility Scenario. *Agriculture* **2024**, *14*, 2268.

<https://doi.org/10.3390/agriculture14122268>

academic editor: Maciej Zaborowicz

Received: 13 November 2024

Revised: 25 November 2024

Accepted: 26 November 2024

Published: 11 December 2024



**Copyright:** © 2024 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

Due to the presence of various essential nutrients in tomatoes, which are beneficial for human health, tomatoes hold significant nutritional value [1]. Globally, the mass cultivation and consumption of tomatoes as a vegetable crop play a significant role in agricultural economies and diets, highlighting their importance in both food security and agricultural productivity [2]. However, the tomato harvesting process still largely depends on manual labor, resulting in high labor costs and inefficiencies. With the rapid advancement of artificial intelligence, it is expected that automated robots will increasingly replace manual labor for tasks such as harvesting, identification, and yield estimation. The accuracy of computer vision detection is crucial for optimizing the efficiency of these automated systems [3]. In facility agriculture, the complex growing environment of tomatoes is characterized by occlusion of leaves and branches, varying ripeness stages, and overlapping

fruits, which pose significant challenges to detection accuracy [4]. Thus, improving the robustness and precision of tomato detection in such challenging conditions is essential.

Currently, with the development of extensive theoretical advancements in target detection and image segmentation technologies [5], deep learning offers superior speed and accuracy over traditional machine vision, thanks to its advanced attribute extraction and self-learning abilities [6,7]. For example, Rong et al. [8] proposed the detection and segmentation of the connections between tomato fruit, stem, and calyx using an improved Swin Transformer V2, which achieves a mean pixel accuracy (MPA) of 89.79%. The authors in Sun et al. [9] used multi-scale feature integration to merge intricate low-level features with high-level semantic information for tomato recognition in occluded conditions, resulting in an 8.8% increase in mean average precision. Within the domain of modern object detection, YOLO [10,11], as a single-stage target detection algorithm, is highly regarded by scholars due to its superior execution speed and precision; it is extensively employed in agricultural object detection. A model named DSW-YOLO was proposed by Du et al. [12], which enhances YOLOv7 by incorporating DCNv3 to identify occlusions of strawberry fruits in agricultural settings. Furthermore, enhancing the bounding box loss function to Wise-IoU v3 (WIoU v3) speeds up the network's convergence rate [13,14], achieving a mean Average Precision (mAP) of 0.86. The authors in Zheng et al. [2] developed a novel RC-YOLOv4 network for the identification of tomatoes under complex occlusions in natural environments, achieving an overall accuracy of 94.44% with a detection rate of 10.71 frames per second. For the identification of cherry tomatoes in states of occlusion, an improved DSP-YOLOv7-CA network was proposed by Hou et al. [15], which can achieve 98.86% accuracy with a model size of 33.71 MB and 104.61 GFLOPs.

The aforementioned algorithms have demonstrated promising detection results through the design of performance-rich feature extraction modules and attention mechanisms [16–18]; however, they frequently encounter challenges related to large model scale, numerous parameters, and high computational demands. Consequently, a substantial amount of research has shifted focus toward the development of lightweight target detection algorithms. The authors in Cheng et al. [19] proposed YOLOLite-CSG, a lightweight approach for detecting crop pests derived from YOLOv3Lite, which enhances the residual block architecture by incorporating sandglass blocks and coordinate attention mechanisms. This technique attained a detection accuracy of 82.9% when evaluated on the CP15 dataset for crop pests, with a computational complexity of 9.8 GFLOPs, which represented an 8.1% reduction compared to YOLOv3. The authors in Gao et al. [20] proposed a light-scale LACTA architecture for cherry tomato target detection, which incorporates an adaptive feature extraction network (AFEN) along with a cross-layer feature fusion network (CFFN), and the approach achieved a 97.3% detection accuracy with 11.4 GFLOPs of computation while significantly reducing parameters by 72%. The authors in Zeng et al. [21] introduced an enhanced YOLOv5 network for detecting tomatoes, using MobileNetV3 as the backbone and pruning the neck layer to minimize model parameters while employing a genetic algorithm for hyperparameter optimization to boost detection accuracy; the model's mAP was just 0.5% lower than YOLOv5, while its size was reduced to one-fifth. A lightweight apple detection method based on an improved YOLOv8 was proposed by Liu et al. [22], named Faster-YOLO-AP, which integrates partial depthwise convolution (PDWConv) and depthwise separable convolution (DWSCConv) and reduces network parameters by 2.35 M, compared to YOLOv8n, thus achieving an average accuracy of 84.12%.

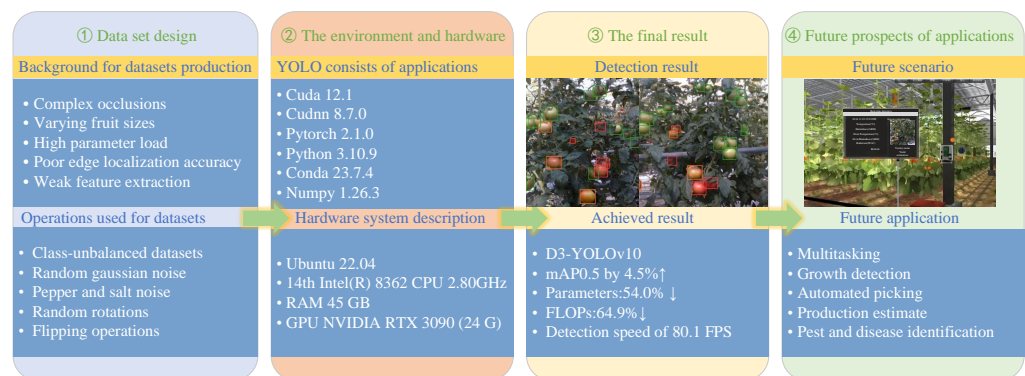
However, the above research has successfully reduced model parameters and floating-point computations; this simplification of the architecture or reduction in parameters inevitably leads to a loss of crucial features. As a result, there is a consequential decline in detection accuracy, accompanied by diminished feature representation capability and reduced computational precision, which together impact the overall effectiveness of the detection system.

Despite numerous researchers having made significant explorations in the tomato detection field, several challenges remain, particularly in achieving real-time detection

under occlusion. These issues can be summarized as follows: (1) The issue of reduced accuracy caused by occlusion has yet to be fully addressed due to dense tomato plantings and complex environments in facility agriculture. (2) The designed detection algorithms ought to be easily deployable in real-world applications, thus guaranteeing effective performance in resource-limited settings. (3) Lightweight models often demonstrate diminished detection accuracy, which constrains their effectiveness in high-demand applications that require both superior performance and efficiency.

In order to solve these problems, this paper proposes a lightweight and efficient framework called D<sup>3</sup>-YOLOv10 (Dynamic + Deformable + Distillation). The workflow diagram for this paper is shown in Figure 1. The primary contributions of this paper can be summarized as follows:

1. We constructed a self-made tomato image dataset under a facility environment that encompassed varying maturity stages of tomato with different illumination conditions. To alleviate the natural maturity class-imbalanced issue that existed in this dataset, a two-stage category balancing method based on diffusion models was proposed for balancing the samples with different maturity stages.
2. We proposed a novel tomato detection model for the facility environment. Specifically, dynamic faster network and deformable attention were separately designed to improve the feature extraction capabilities on different tomato sizes and occlusion conditions. Moreover, a scalable, dynamic non-monotonic focusing mechanism, WIoU, was also developed to further facilitate edge detection accuracy and convergence.
3. We developed a knowledge distillation scheme, utilizing a semantic frequency prompt for optimizing the detection model. It enabled our model to be more suitable for lightweight deployment requirements in practical applications.
4. We conducted detection experiments on our self-made datasets. Compared to existing methods, D<sup>3</sup>-YOLOv10 achieves superior performance in detecting heavily occluded tomatoes, exhibiting enhanced target localization and refined bounding box accuracy, thereby significantly improving overall detection efficacy.



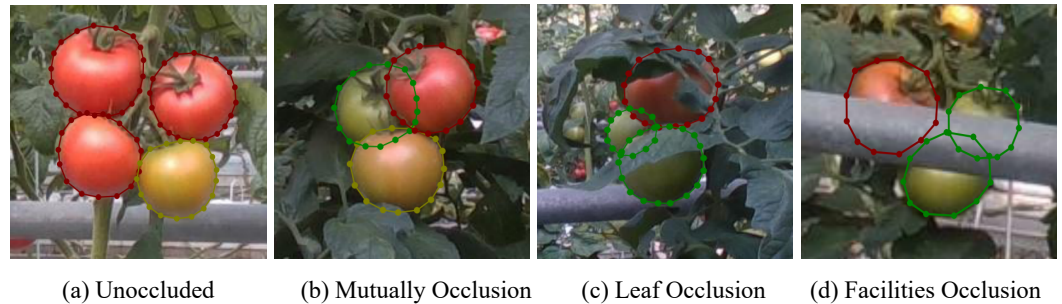
**Figure 1.** The workflow diagram for this paper.

## 2. Materials and Methods

### 2.1. Image Acquisition and Datasets Establishment Strategy

Tomato images were collected at the Modern Agricultural Industrial Park New District in Zhenjiang, Jiangsu Province, China, from May to June 2023. Using an Intel Realsense D435 depth industrial camera, this study captured 878 original images of tomatoes under natural light at a distance of 1.2 m, which were taken in the morning, noon, and afternoon, encompassing diverse growth stages, levels of occlusion, as shown in Figure 2a–d. The images of tomatoes were collected and saved as ‘.jpg’ files, with their resolution resized to 640 × 640 pixels. The acquired images were manually labeled using Labelme software (version 4.5.6), with experts classifying maturity stages into three categories: immature, semi-mature, and mature, represented by green, yellow, and red markings, respectively, as shown in Figure 2. To enhance dataset diversity and prevent model overfitting, data

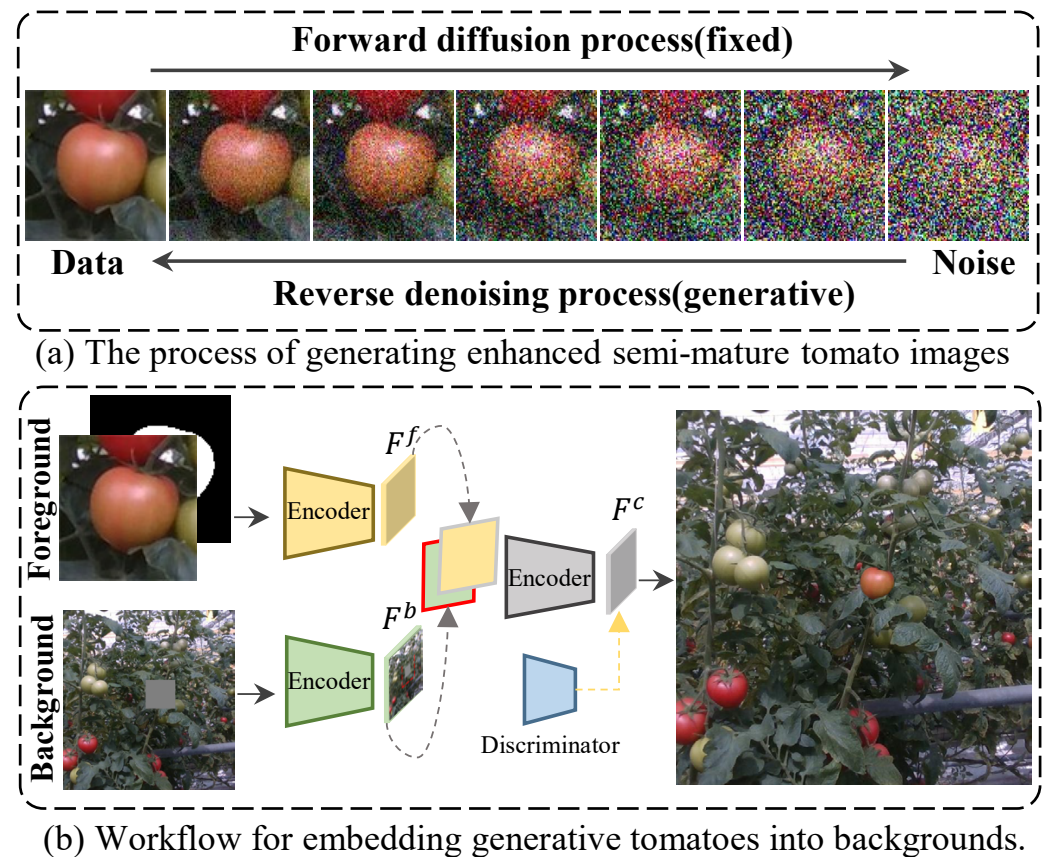
augmentation techniques, including random Gaussian noise, pepper-and-salt noise, random rotations, and flipping operations, were applied, resulting in a dataset size increase of approximately threefold. Subsequently, the dataset was divided into training, validation, and test subsets according to a 6:2:2 ratio.



**Figure 2.** Tomato occlusion in the facility scenario, including the unoccluded, mutual occlusion, leaf occlusion, and facility occlusion.

### 2.2. Dataset Category Balancing Methods

To address the class-imbalanced issue of semi-mature tomatoes in the original datasets, we proposed a two-stage class-balancing method based on diffusion models. The first stage of the method developed a generative model that was trained to learn the patterns of standard normal distribution noise, sampled from this distribution, and progressively denoised to generate high-quality target data samples, ultimately producing enhanced semi-mature tomato images, as shown in Figure 3a.



**Figure 3.** Structure diagram of the two-stage class balancing method based on the diffusion model.

In the second stage, we designed a semi-mature tomato object synthesis method using foreground object search. Two encoders were trained to extract background and foreground



features, and their compatibility was assessed by calculating feature similarity, which was then employed to synthesize the compatible images, depicted in Figure 3b. The generated images from the first stage, along with their corresponding mask images, were used as foreground inputs to the network, while background inputs were selected from the original tomato datasets images for further processing and analysis. In order to improve the realism of the generated images, we propose a method that randomly selects one to five foreground images and carefully selects synthesis points with dim lighting and dense branches and leaves. By implementing the two-stage class balancing method based on diffusion models in the original training dataset, the proportions of three types for mature, semi-mature, and immature tomatoes were adjusted to approximately 16:15:18, significantly addressing the issue of class imbalance.

### 2.3. Improved YOLOv10s Network Structure

The YOLOv10 architecture comprises three core components: the backbone serves for feature extraction, the neck facilitates feature refinement and aggregation, while the head is responsible for producing the final predictions. YOLOv10 utilizes a dual assignment strategy that obviates the necessity for non-maximal suppression (NMS) in post-processing during inference, resulting in faster inference times. In addition, YOLOv10 optimizes the network architecture by reducing redundant computations and parameters, rendering it more appropriate for implementation in real-world applications. Notwithstanding this, the detection accuracy of YOLOv10 for specific scenarios has failed to reach the desired level. Consequently, this paper presents an advanced variant of YOLOv10, designated as D<sup>3</sup>-YOLOv10.

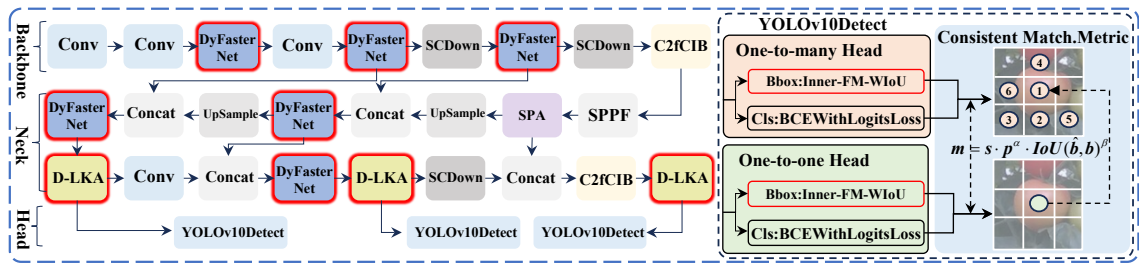
Figure 4 illustrates the complete architecture of the proposed study, comprising two components: D<sup>3</sup>-YOLOv10 structure and knowledge distillation workflow. D<sup>3</sup>-YOLOv10 structure primarily comprises DyFasterNet (Dynamic Faster Network), D-LKA (Deformable Large Kernel Attention), C2fCIB (Faster CSP Bottleneck and Compact Inverted Blocks), SPPF (Spatial Pyramid Pooling Fast). The knowledge distillation based on the semantic frequency prompt involves using D<sup>3</sup>-YOLOv10 as both the teacher and student networks. Specifically, YOLOv10 comprises six scale versions to optimize the balance; we chose the pre-trained S scale (depth: 0.33, width: 0.50, max-channels: 1024) as the teacher model and the N scale (depth: 0.33, width: 0.25, max-channels: 1024) as the student model.

### 2.4. Dynamic Faster Network

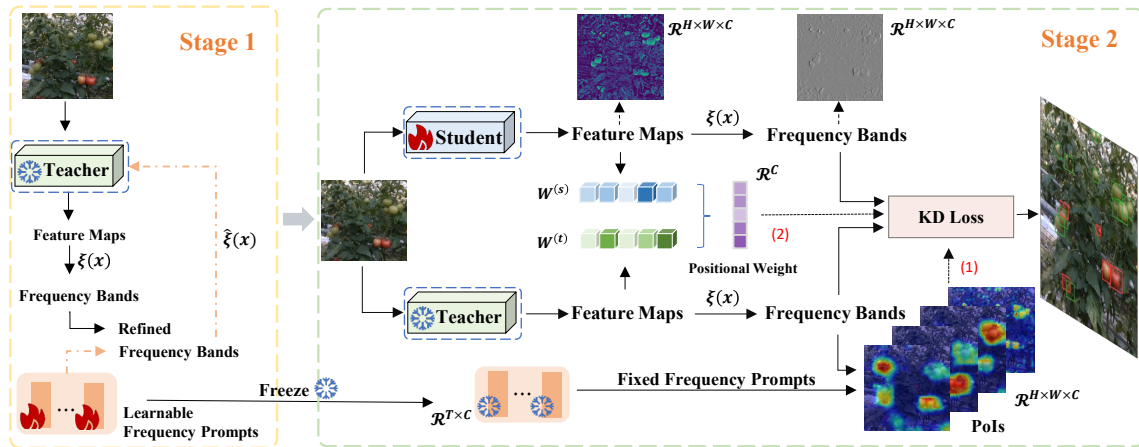
The traditional convolutional method relies on a single kernel for processing diverse features, which limits its effectiveness in addressing variations in feature size and characteristics. To address these limitations and effectively capture the diverse size and shape characteristics of tomato growth at different stages, this study proposes the Dynamic Faster Network (DyFasterNet), which splits and concatenates multiple dynamic convolution modules that are adaptively aggregated using a range of convolution kernels. The workflow of DyFasterNet is illustrated in Figure 5. Equation (1) illustrates the output feature map  $Y \in \mathbb{R}^{C_{out} \times H' \times W'}$  obtained from a dynamically weighted fusion method based on multi-kernel information.

$$Y = X * \sum_{i=1}^M \alpha_i W_i \quad (1)$$

where,  $X$  is Input features,  $M$  is the number of convolution kernel sets,  $*$  represents the convolution operation, a predefined set of kernels  $W_i \in \mathbb{R}^{C_{out} \times C_{in} \times K \times K}$ ,  $\alpha_i$  denotes the  $i$ -th convolutional kernels weight coefficient. The dynamic coefficient  $\alpha_i$  is obtained the input  $X$ . Specifically, The dynamic coefficient  $\alpha_i$  is derived by employing global average pooling on the input  $X$ , processing the resulting vector through a two-layer MLP, and then applying a softmax activation function. Comparatively to the original convolutional layer, the process of coefficient generation brings negligible FLOPs and The process of calculation is illustrated in Figure 5b.



(a) Schematic of the proposed D<sup>3</sup>-YOLOv10 network structure.

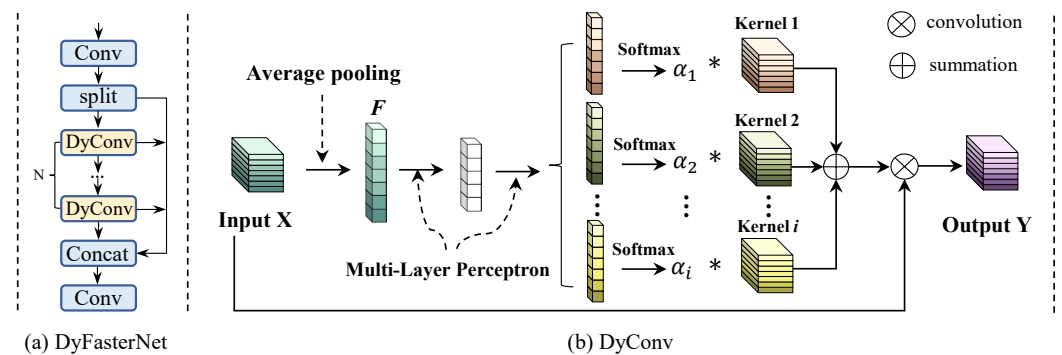


(b) Frequency prompts interact.

(c) Workflow of the teacher(D<sup>3</sup>-YOLOv10-s) and student(D<sup>3</sup>-YOLOv10-n) network distillation process.

**Figure 4. Overview of the experimental framework.** Stage 1: With the teacher model’s guidance, the learnable frequency prompts interact with the frequency bands. Stage 2: The feature maps distilled from both the student and teacher are initially transformed into the frequency domain. The frequency prompts from Stage 1 are then applied, with the frozen prompts multiplied by the teacher’s frequency bands to generate points of interest (PoIs). Finally, the spatial weights for each channel are determined by the teacher and student spatial gates. Process (1) in the figure identifies the distillation locations, while Process (2) measures the distillation extent.

The workflow of DyFasterNet is illustrated in Figure 5. DyFasterNet employs a split-and-concatenated network structure with partial connections to reduce redundant gradient computations and integrates DyConv modules to further lower computational complexity and enhance feature extraction. To optimize the network efficiency of YOLOv10, we replaced the original C2F module (shown in Figure 4a) with the DyFasterNet proposed in this section, enhancing the network’s ability to extract features from diverse features of tomatoes and reducing floating-point operations.

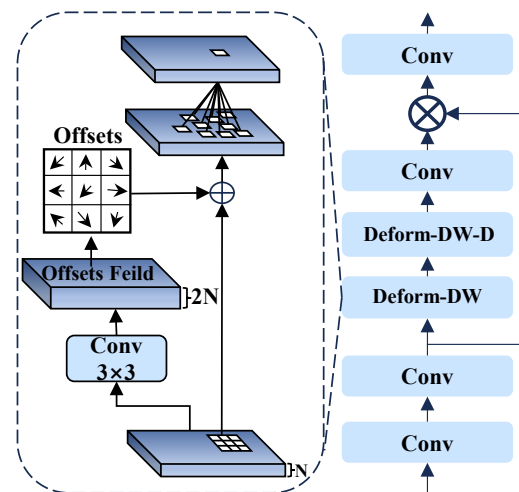


**Figure 5.** Architecture of the DyFasterNet module: (a) DyFasterNet; (b) Dynamic convolution.

### 2.5. Deformable Large Kernel Attention Mechanism

In facility agriculture, tomato detection faces several challenges, including occlusion caused by overlapping fruits and leaves, variability in fruit shape and size at different growth stages, and interference from complex background environments. This study developed the deformable large-kernel attention (D-LKA) mechanism in the final phase of the neck network as depicted in Figure 4a. This attention mechanism combines the receptive field of large convolution kernels with the flexibility of deformable convolutions, enhancing the ability to capture objects of irregular shapes and sizes. It improves the extraction of tomato features with irregular shapes caused by occlusions from leaves, the environment, and inter-fruit occlusion.

It is depicted in Figure 6 that the mechanism described above consists of several components. Specifically, the D-LKA model integrates Deformable Convolutions through large-kernel convolution, where the large kernel is composed of depth-wise convolution (DW conv) and depth-wise dilated convolution (DW-D conv). Deformable Convolutions (Deform-DW) modify the sampling grid by applying integer offsets to facilitate flexible deformation. This process involves an additional convolutional layer that learns the deformation from feature maps and generates an offset field, ultimately leading to an adaptive convolution kernel. An activation function, GELU, is applied to enhance nonlinearity.



**Figure 6.** Deformable large kernel attention mechanism structure.

DyFasterNet and D-LKA complement each other in the model, effectively improving detection performance and robustness through the division of labor and cooperation between low-level features and high-level features. DyFasterNet focuses on efficiently extracting basic features such as edges and textures, optimizes computing resources, and accelerates inference speed. Meanwhile, D-LKA enhances the model's ability to identify complex deformations and occluded targets by combining the receptive field of large convolutional kernels and the flexibility of deformable convolutions. DyFasterNet alleviates the computational overhead of D-LKA, allowing the model to maintain high precision while optimizing overall efficiency, thus achieving efficiency and robustness in complex scenarios such as tomato detection.

### 2.6. Knowledge Distillation

Knowledge distillation assists in tackling the issue of implementing large, intricate models on edge devices with limited resources, where such models can be cumbersome and inefficient [23–25]. However, traditional knowledge distillation often relies on downsampling within the spatial domain of the teacher model, which can corrupt the features and limit effectiveness. Frequency distillation addresses the challenge of identifying valuable

pixels across different frequency bands, improving generalization and robustness while overcoming limitations of spatial-based methods in dense prediction tasks [26–28].

This study developed an offline knowledge distillation scheme, utilizing a semantic frequency prompt for optimizing the detection model D<sup>3</sup>-YOLOv10. It enabled our model to be more suitable for lightweight deployment requirements than the dense prediction requirements for tomatoes in practical applications. By assimilating rich semantic frequency features information from the teacher model, the student demonstrates improved adaptability to complex environments, thereby enhancing detection reliability and stability. Furthermore, through the distillation of fine-grained feature information, the student model achieves more precise identification and classification of tomatoes across varying stages of maturity.

### Semantic Frequency Prompt

The knowledge distillation method is based on a semantic frequency prompt, which is divided into two main stages. Firstly, raw frequency prompts are inserted into the teacher model, which absorbs semantic frequency context during fine-tuning. This process generates pixel-by-pixel frequency masks to identify pixels of interest (PoIs) across various frequency bands. In the second stage, during the distillation process, these frequency prompts are used to generate pixel-by-pixel masks that pinpoint PoIs within specific frequency bands. Furthermore, a location-aware relational frequency loss is employed to enhance spatial resolution for the student model, improving its performance on dense prediction tasks. An overview of FreeKD's two-stage process is shown in Figure 4b,c.

### 2.7. Loss Function

The YOLOv10 model uses CIoU loss for bounding box regression, where geometric factors like overlap, centroid distance, and aspect ratio can penalize low-quality samples, reducing generalization. To tackle this, WIoU optimizes distance attention by concentrating on the center between the anchor and target boxes, as demonstrated in Equation (3).

$$\mathcal{L}_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{wh + w^{st} h^{st} - W_i H_i} \quad (2)$$

where  $w$  and  $h$  are the width and height of the anchor box, respectively, and  $w^{st}$ ,  $h^{st}$  are the width and height of the target box, respectively. Where  $W_i$  and  $H_i$  denote the width and height of the intersecting rectangles formed by the anchor box and the target box, respectively.

$$\mathcal{R}_{WIoU} = \exp\left(\frac{(x_c - x_c^{st})^2 + (y_c - y_c^{st})^2}{(W_g^2 + H_g^2)^*}\right) \quad (3)$$

where,  $W_g$  and  $H_g$  represent the dimensions of the smallest enclosing box,  $x_c$  and  $y_c$  denote the coordinates of the center for the anchor box,  $x_c^{st}$  and  $y_c^{st}$  indicate the coordinates of the center for the target box, as demonstrated in Figure 7.  $\mathcal{R}_{WIoU} \in [1, e)$ , which will significantly increase  $\mathcal{L}_{IoU} \in [0, 1]$  of the ordinary quality anchor box.

To more effectively minimize the impact of poor-quality samples affecting the loss value, dynamic non-linear frequency modulation is incorporated into  $\mathcal{R}_{WIoU}$ , resulting in the construction of  $\mathcal{R}_{FM-WIoU}$ , which effectively prevents large harmful gradients from low-quality examples, as demonstrated in Equations (4) and (5)

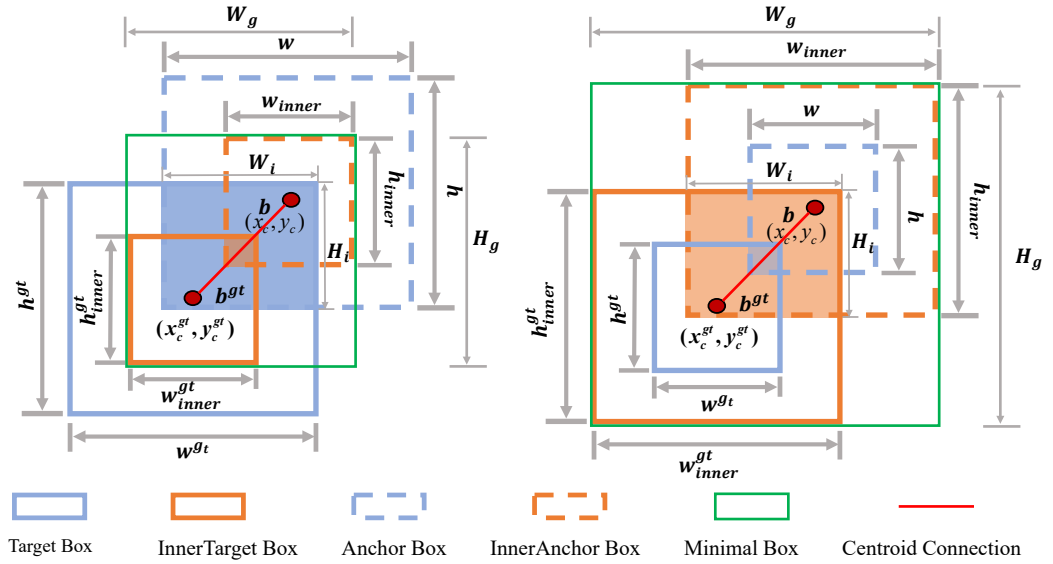
$$\mathcal{L}_{FM-WIoU} = \mathcal{L}_{IoU} \mathcal{R}_{WIoU} \gamma \quad (4)$$

$$\gamma = \frac{\beta}{\delta \alpha^{\beta-\delta}}, \beta = \frac{\mathcal{L}_{IoU}^*}{\mathcal{L}_{IoU}} \in [0, +\infty) \quad (5)$$

where,  $\gamma$  is the non-linear focus factor,  $\beta$  symbolizes the outlier within the anchor box, and an increased outlier value indicates a reduced quality of the anchor box.  $\mathcal{L}_{IoU}^*$  denotes the monotonic focusing coefficient.



To further expedite the network's convergence and improve detection precision, this paper proposes a dynamic non-monotonic FM-WIoU boundary box regression (BBR) loss function based on a scaling factor (Inner-FM-WIoU). Specifically, convergence is accelerated for high-IoU samples through the use of smaller auxiliary bounding boxes, whereas larger auxiliary bounding boxes are better suited for low-IoU samples.



**Figure 7.** The proposed Inner-FM-WIoU.

As shown in Figure 7, The target box and the anchor box are denoted as  $b^{gt}$  and  $b$ , respectively. The center coordinates of both the target box and the inner target box are represented by  $(x_c^{gt}, y_c^{gt})$ , while  $(x_c, y_c)$  denotes the center of the anchor point and the inner anchor point. The width and height of the target box are represented as  $w^{gt}$  and  $h^{gt}$ , respectively, whereas the width and height of the anchor are represented by  $w$  and  $h$ . The variable ratio relates to the scaling factor, typically ranging from 1 to 1.5. The Inner-FM-WIoU BBR loss function ( $\mathcal{L}_{Inner-FM-WIoU}$ ) calculation process is shown in Equations (6)–(13).

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} * ratio}{2}, b_r^{gt} = x_c^{gt} + \frac{w^{gt} * ratio}{2} \quad (6)$$

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} * ratio}{2}, b_b^{gt} = y_c^{gt} + \frac{h^{gt} * ratio}{2} \quad (7)$$

$$b_l = x_c - \frac{w * ratio}{2}, b_r = x_c + \frac{w * ratio}{2} \quad (8)$$

$$b_t = y_c - \frac{h * ratio}{2}, b_b = y_c + \frac{h * ratio}{2} \quad (9)$$

$$inter = (\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l)) * (\min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t)) \quad (10)$$

$$union = (w^{gt} * h^{gt}) * (ratio)^2 + (w * h) * (ratio)^2 - inter \quad (11)$$

$$IoU^{inner} = \frac{inter}{union} \quad (12)$$

$$\mathcal{L}_{Inner-FM-WIoU} = \mathcal{L}_{FM-WIoU} + IoU - IoU^{inner} \quad (13)$$

**Distillation with Frequency.** A classic approach of frequency-based distillation is to imitate the tensors at the pixel level. Typically, the feature maps from the teacher and student

networks, respectively, are denoted by  $\mathbf{F}^{(t)} \in \mathbb{R}^{C \times H \times W}$  and  $\mathbf{F}^{(s)} \in \mathbb{R}^{C_s \times H \times W}$ , and the frequency band can be emulated by the following:

$$\mathcal{L}_{FKD} = \sum_{k=1}^L \|a_k - b_k\|_1, a_k \in \xi(\mathbf{F}^{(t)}), b_k \in \xi(\phi(\mathbf{F}^{(s)})) \quad (14)$$

where  $L$  is number of frequency bands and  $F(s)$  is adapted to the same resolution as  $F(t)$ , using  $\phi$ , a linear projection layer.  $\xi$  is the Discrete Wavelet Transformation (DWT) Semantic Frequency Prompt distillation. The distillation loss based on semantic frequency prompt can be represented as Equation (15).

$$\mathcal{L}_{FreeKD} = \sum_{k=1}^L \omega^{(r)} \|M \otimes a_k - M \otimes b_k\|_1. \quad (15)$$

Among them,  $\omega^{(r)} = \omega^{(t)} \otimes \omega^{(s)}$  is generated by the position-aware relationship weights between teachers and students, ensuring that the channels in the distillation should include channels that are meaningful to both teachers and students. The mutual information exists within the set of real numbers  $\mathbf{M} \in \mathbb{R}^{C \times H_{HH} \times W_{HH}}$ , which is found between the prompt pixel  $P$  and the frequency pixel  $R^{(t)}$  in the teacher band. This is represented by the gating weight, which is produced by the multi-layer perceptron (MLP).

The overall loss is represented by Equation (16), which encompasses the regression loss, classification loss, and distillation loss.

$$\mathcal{L} = \mathcal{L}_{Inner-FM-WIoU} + \mathcal{L}_{BCEWithLogitsLoss} + \mu \mathcal{L}_{FreeKD}, \quad (16)$$

where  $\mu$  is the factor that equalizes the loss.

## 2.8. Evaluation Metrics

Based on the labeled and model-predicted samples, a confusion matrix was created that included four indicators:  $TP$ ,  $TN$ ,  $FP$ , and  $FN$ . This study used the mean Average Precision ( $mAP$ ) was employed to assess the performance. The relevant formulas are provided below:

$$P = \frac{TP}{TP + FP} \times 100\% \quad (17)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (18)$$

The  $AP$  indicates the area beneath the precision-recall curve. The equations utilized to calculate each evaluation metric are as follows:

$$AP = \int_0^1 P(R) dR \quad (19)$$

The  $mAP$  (mean average precision) represents the average of  $AP$  across all categories and directly indicates the model's classification capability. The formula for computing  $mAP$  is as follows:

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \times 100\% \quad (20)$$

FPS evaluates the model's real-time inference performance, while GFLOPs (Giga Floating Point Operations) quantify the computational complexity [29].

$$FPS = \frac{1}{T_{100}} \quad (21)$$

$$GFlops = \frac{Nflops/10^9}{T} \quad (22)$$

where  $N_{fpos}$  refers to the total number of floating-point operations performed by the model, including addition, multiplication, and division.  $T$  refers to the time taken for a single model execution, while  $T_{100}$  indicates the total time required to process 100 images, both expressed in seconds.

### 2.9. Performance of the Two-Stage Category Balancing Method

To assess the efficacy of the suggested two-stage category balancing approach utilizing diffusion models, it was applied to 614 sample images from the initial training dataset (which had class imbalance). This process resulted in a balanced dataset comprising 614 images, each containing an increased number of samples. The 614 initial images (representing the class-imbalanced dataset) and the 614 modified images (forming the class-balanced dataset) were then enhanced using common data augmentation methods. These included adding random Gaussian noise, applying pepper-and-salt noise, adjusting brightness, enhancing contrast, and performing flipping operations. This augmentation resulted in both an expanded original dataset and an augmented class-balanced dataset, which were utilized for training the YOLOv10s network. The performance assessment of the two trained YOLOv10s models was carried out using 264 original images from the testing dataset, as illustrated in Figure 8. The experiments show that the YOLOv10s model trained on the augmented class-balanced dataset outperformed the one trained on the augmented original dataset, especially in detecting semi-mature tomatoes. Specifically, the Average Precision (AP) for semi-ripe tomatoes increased by 19.3%, while the AP for ripe and unripe tomatoes improved by 6.0% and 1.9%, respectively. Overall, the mean Average Precision (mAP) for all tomato fruits increased by 7.4%. These results confirm that the proposed two-stage class balancing method based on the diffusion model, which enhances the precision in detecting semi-mature tomatoes, minimizes inter-class recognition disparities, and strengthens the model’s overall robustness.

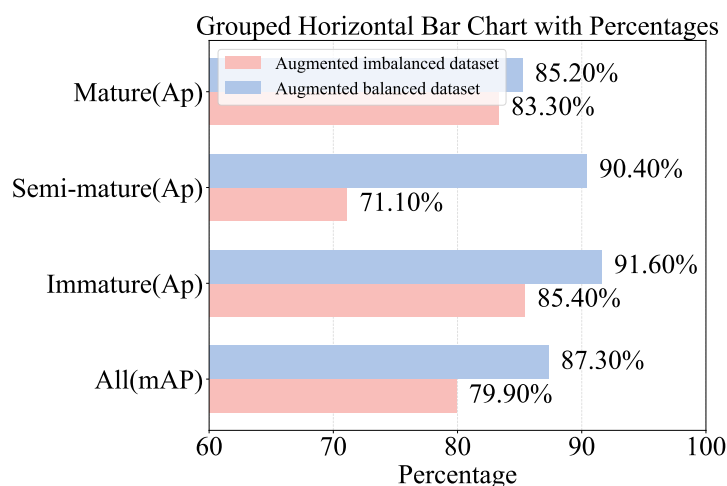


Figure 8. YOLOv10s model performance for class-balanced datasets vs. class-imbalanced datasets.

### 2.10. Ablation Experiments on the Model’s Performance

This research aims to create a lightweight and efficient algorithm for detecting and classifying tomatoes by enhancing the YOLOv10s network (D<sup>3</sup>-YOLOv10). The improvements involve proposing the DyFasterNet for feature extraction, the D-LKA attention mechanism, optimizing bounding box loss function FM-WIoU based on a scaling factor, and knowledge distillation based on semantic frequency prompts (FreeKD) for lightweighting. To assess the impact of each enhancement on the network, we performed ablation experiments while maintaining consistent training conditions and hyperparameters. And then, the aforementioned improvements were sequentially implemented, with YOLOv10s serving as the baseline.

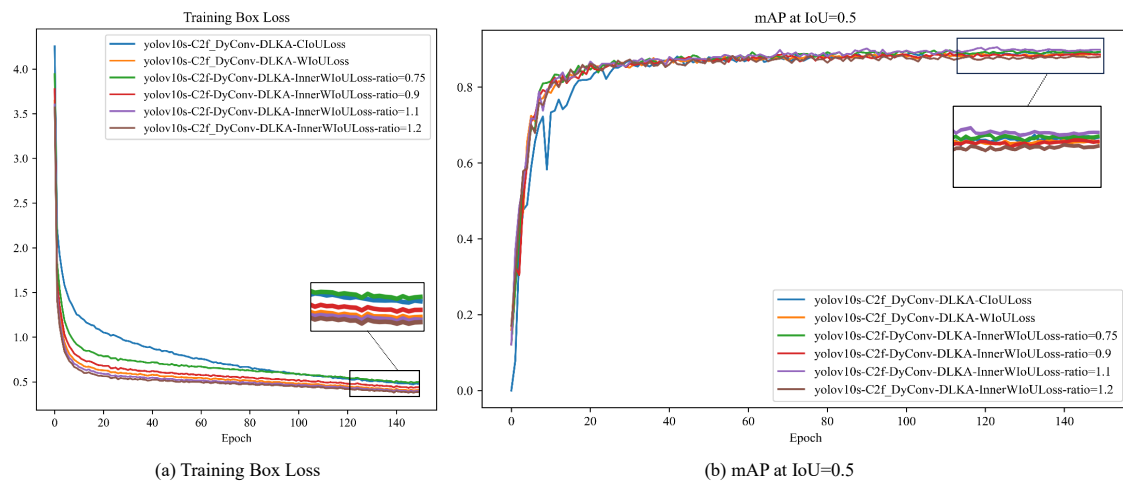
As shown in Table 1, the integration of DyFasterNet as the revised feature extraction component improved  $mAP_{0.5}$  by 1.9% and concurrently diminished FLOPs by 15.1% in comparison to the baseline. The design of an attention mechanism (D-LKA) based on the above improvement escalates the  $mAP_{0.5}$  of the network by an additional 1.4%. The proposed bounding box loss function (Inner-FM-WIoU), incorporating the FM-WIoU and a scaling factor, improves the  $mAP_{0.5}$  by 0.8% and 0.2%, respectively. However, the implementation of these efficient attention mechanisms and feature extraction methods has led to a slight rise in the count of parameters compared to YOLOv10s. Notably, the introduction of semantic frequency prompt knowledge distillation method (FreeKD), effectively mitigated this increase in parameters during the training of our modified YOLOv10s network. This approach not only achieved a 0.2% improvement in  $mAP_{0.5}$  but also significantly reduced the parameters and FLOPs to 46.3% and 35.1% of the Benchmark model, respectively.

To validate the efficacy of the proposed bounding box loss function (Inner-FM-WIoU), a comparative analysis was performed on different regression loss functions, including CIoU, FM-WIoU, and FM-WIoU, with different scaling factors. Meanwhile, Extensive experiments were conducted to identify the optimality ratio for the scaling factor, which compares the performance of ratios 0.75, 0.9, 1.1, and 1.2 on the D<sup>3</sup>-YOLOv10 model as illustrated in Figure 9. Figure 9a displays the BoxLoss curve during training, showing that when the ratio exceeds 1.0, the scalable FM-WIoU converges faster than the original CIoU and WIoU, achieving a lower loss value. Figure 9b presents the corresponding  $mAP_{0.5}$  curve, demonstrating that FM-WIoU with a scaling factor of 1.1 not only reaches convergence more quickly but also attains the highest  $mAP_{0.5}$  when compared to alternative methods. Thus, the FM-WIoU loss for bounding boxes utilizing a scaling factor of 1.1 optimizes edge detection accuracy and convergence, enhancing boundary precision and accelerating training for improved detection performance. As illustrated in Figure 11b–f, our algorithm outperforms others by producing more precise bounding boxes that are more accurately aligned with the true positions of the tomatoes, thereby showcasing the efficacy of the suggested bounding box loss function.

**Table 1.** Ablation study with D<sup>3</sup>-YOLOv10 when DyFasterNet, D-LKA, FM-WIoU, Inner-FM-WIoU, and FreeKD are applied to baseline.

	DyFasterNet	D-LKA	FM-WIoU	Inner-FM-WIoU	FreeKD	$mAP_{0.5}$ (%) ↑	Parameters (M) ↓	FLOPs(G) ↓
Baseline						87.3	8.04	24.5
DyFasterNet	✓					89.2	10.17	20.8
D-LKA		✓				88.3	11.84	31.5
DyFasterNet+ D-LKA	✓	✓				90.6	12.78	25.9
DyFasterNet+ D-LKA+FM-WIoU	✓	✓	✓			91.4	12.78	25.9
DyFasterNet+ D-LKA+FM-WIoU+Inner	✓	✓	✓	✓		91.6	12.78	25.9
DyFasterNet+ D-LKA+FM-WIoU+Inner+ FreeKD	✓	✓	✓	✓	✓	91.8	3.72	8.6





**Figure 9.** The loss curve and  $mAP_{0.5}$  under different scaling ratios.

The results suggest that, with the effective implementation of the feature extraction component, the attention mechanism module, and the Inner-FM-WIoU bounding box loss function, the model experienced a 4.5% improvement in its  $mAP_{0.5}$ . The development of the distillation method, FreeKD, to train our improved network effectively counterbalanced the increase in parameters and computation. Furthermore, consistent improvement in the  $mAP_{0.5}$  of tomato fruits was observed by modifying the YOLOv10s model with various combinations of components.

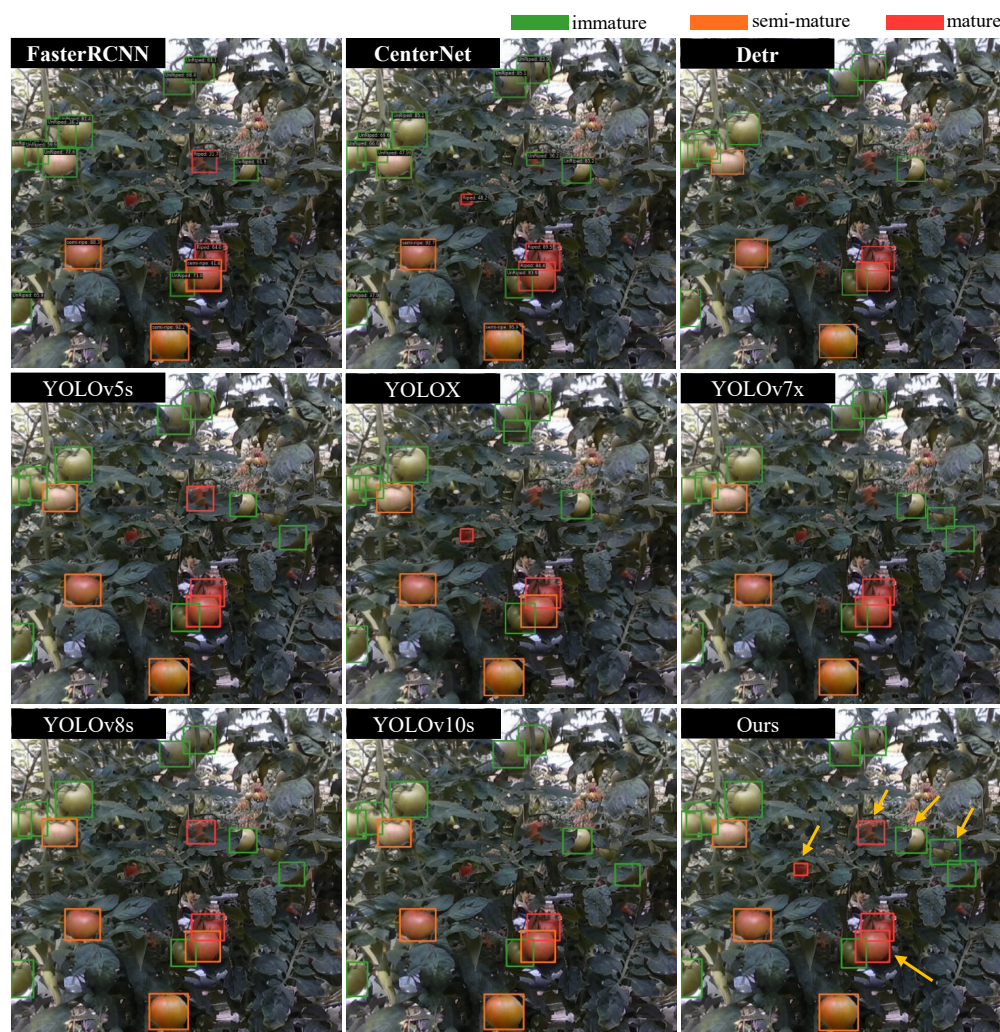
### 2.11. Performance Analysis of Different Models

Further substantiating our proposed model's superiority, we conduct comparison experiments with several prominent object detection models, including FasterRCNN [30], CenterNet [31], DETR [32], YOLOv5 [33], YOLOX [34], YOLOv7 [35], YOLOv8s [36], and YOLOv10s [11]. FasterRCNN and CenterNet use ResNet50 as their backbone networks, while the YOLO series maintains scale consistency across versions like YOLOv5s, YOLOv7x, YOLOv8s, and YOLOv10s. This approach aligns the quantity of participants with the comparative models and our baseline network, thereby enhancing the validity and reliability of the experiments while showcasing the superiority of our model (D<sup>3</sup>-YOLOv10). As shown in Table 2, the results are summarized.

**Table 2.** Comparative results of various lightweight detection models on our self-made tomato dataset.

Model	$mAP_{0.5}$	$mAP_{0.5-0.95}$	FPS	Parameters (M)	FLOPs (G)	Model Size (MB)
FasterRCNN [30]	83.5	56.5	12.74	28.31	948.16	320
CenterNet [31]	84.5	57.2	24.57	32.67	70.22	130
DETR [32]	86.4	53.1	26.17	36.74	101.4	474
YOLOv5 [33]	88.3	63.0	99.03	7.03	16	13.7
YOLOX [34]	86.0	60.3	163.93	8.94	26.76	68.5
YOLOv7 [35]	87.1	58.8	62.1	36.49	103.2	71.3
YOLOv8s [36]	88.3	62.0	169.00	11.13	28.4	21.4
YOLOv10s [11]	87.3	63.1	159.00	8.04	24.5	15.7
D <sup>3</sup> -YOLOv10	91.8	63.8	80.1	3.72	8.6	7.54

With respect to precision, our model demonstrates a significant improvement over the comparative models. As shown in Figure 10, we chose an image featuring various types of occlusions for the case study to evaluate detection performance in challenging situations. In this scenario, our model uniquely attained the highest detection rate, whereas other models experienced significant issues with false positives or missed detections, thus compromising their overall performance.



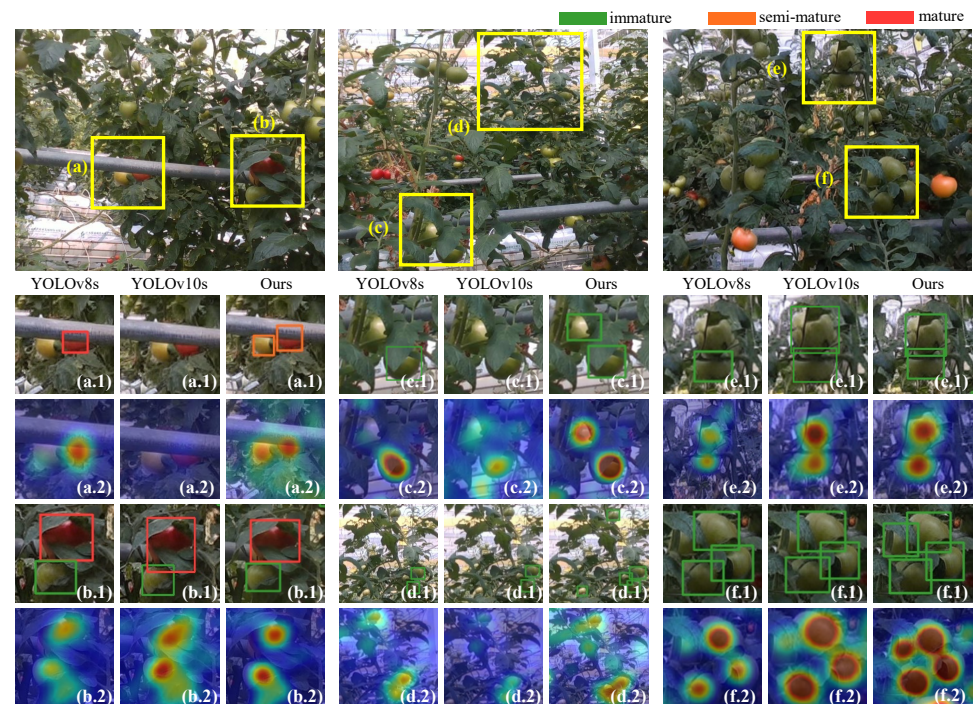
**Figure 10.** Examples of detection capabilities of comparative models in facility environment.

With reference to the parameters delineated in Table 2, our model necessitates the minimum computational effort and minimum parameter amount and exhibits the most superior mAP performance. While the FPS of our model was marginally less than that of YOLOv10s, it exhibited considerable computational efficiency, with our model's computation at just 64.9% of YOLOv10s. It is noteworthy that our model surpasses other lightweight models in performance. Even though our model's FPS is slightly lower than YOLOv10s, it exhibits high computational efficiency, with the computational load being only 46.1% and the parameter count merely 35.1% of the latter. Despite the increased FPS due to improved accuracy, far exceeding real-time requirements, meeting the high-precision needs for fruit counting, sorting, and quality control in complex scenarios such as high density of growth or obstruction in facility agriculture. In summary, our model achieves notable accuracy improvements while substantially reducing both parameter count and computational complexity, all while preserving a reasonable inference speed.

Figure 11 shows three samples under different occlusion types, comparing YOLOv8s, YOLOv10s, and the D<sup>3</sup>-YOLOv10 model proposed in this study, along with the tomato detection results and heat map visualizations generated across six different regions of interest. Due to the detection outcomes of these models, unripe tomatoes that resembled the color of leaves posed a considerable challenge. The D<sup>3</sup>-YOLOv10 exhibited enhanced abilities in identifying and detecting these unripe tomatoes. In group (a), where tomatoes were sheltered by the facility environment, only our network successfully detected and correctly classified them. In group (b), the heat map produced by D<sup>3</sup>-YOLOv10 demonstrates a



higher degree of concentration. In group (d), we can see that D<sup>3</sup>-YOLOv10 exhibited superior performance in detecting a greater number of small target tomatoes. The D<sup>3</sup>-YOLOv10 model excelled in identifying and detecting tomatoes with background color similarity, as demonstrated in groups (c), (e), and (f). Specifically, in group (c), where two tomatoes were obscured by multiple leaves, only D<sup>3</sup>-YOLOv10 successfully detected them. In group (f), where four densely packed tomatoes were subject to mutual occlusion, this scenario posed a significant detection challenge faced by all models. Only D<sup>3</sup>-YOLOv10 succeeded in detecting all of the tomatoes. These results highlighted D<sup>3</sup>-YOLOv10's superior capability in extracting feature information such as edges and colors.



**Figure 11.** Robustness experiment in various scenarios: Area (a) demonstrates the performance of the model in a facility agriculture environment with occlusion. Areas (b,c,e) show tomatoes partially occluded by branches or leaves. While area (d) highlights small target recognition. Finally, area (f) presents cases of mutual occlusion between fruits.

### 3. Conclusions

To address the challenge of tomato classification and detection in facility scenarios with complex occlusions, which includes weak feature extraction for varying fruit sizes and occlusion conditions, poor edge localization accuracy, and a high parameter load, this study proposed a novel YOLOv10-based detection framework called D<sup>3</sup>-YOLOv10 for realizing high-precision tomato detection. By implementing the DyFasterNet module and the D-LKA attention mechanism, the framework's detection accuracy was substantially increased. An innovative bounding box regression loss function incorporating a scaling factor for FM-WIoU significantly improved the model's convergence speed and increased the accuracy of boundary box regression. The aforementioned improvements increased the model's  $mAP_{0.5}$  by 4.5%. Furthermore, by developing a knowledge distillation scheme, which utilizes semantic frequency prompts for lightweighting the models, the parameters and FLOPs were concurrently compressed by 54.0% and 64.9%, respectively. Meanwhile, a detection speed of 80.1 FPS was achieved by the model, satisfying the criteria for tomato detection in real time. This study underscored the model's exceptional performance in densely occluded environments, which provides novel technical insights for fruit target detection in smart agriculture. Additionally, it facilitated the deployment and application of target detection algorithms on devices with limited computational power.

Notwithstanding the promising outcomes, the model has certain limitations. Its generalization capability is constricted by a homogeneous dataset, which demands future training on various crops and environments. The dependence on labeled data complicates the preparation process, which can be alleviated through weakly supervised learning. Future endeavors will broaden the model's applications to multi-crop detection, facilitating adaptability across diverse crops and multi-task learning for tasks such as pest detection and yield prediction. Dynamic monitoring of growth stages will further refine detection strategies, thus enhancing precision agriculture and promoting sustainability.

**Author Contributions:** A.L.: conceptualization, methodology, funding acquisition, writing—original draft, writing—review and editing. C.W.: methodology, software, investigation, validation, writing—original—final draft, writing—review and editing. T.J.: visualization, data curation. Q.W.: resources, investigation, writing—review and editing. T.Z.: investigation, validation. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Key Research and Development Program of China (grant number 2022YFD2000500), the National Natural Science Foundation of China (grant number 62071157).

**Institutional Review Board Statement:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Data Availability Statement:** Data will be made available on request.

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

## References

1. Jun, S.; Yating, L.; Xiaohong, W.; Chunxia, D.; Yong, C. SSC prediction of cherry tomatoes based on IRIV-CS-SVR model and near infrared reflectance spectroscopy. *J. Food Process. Eng.* **2018**, *41*, e12884. [[CrossRef](#)]
2. Zheng, T.; Jiang, M.; Li, Y.; Feng, M. Research on tomato detection in natural environment based on RC-YOLOv4. *Comput. Electron. Agric.* **2022**, *198*, 107029. [[CrossRef](#)]
3. Zhang, F.; Chen, Z.; Ali, S.; Yang, N.; Fu, S.; Zhang, Y. Multi-class detection of cherry tomatoes using improved Yolov4-tiny model. *Int. J. Agric. Biol. Eng.* **2023**, *16*, 225–231.
4. Sun, J.; He, X.; Ge, X.; Wu, X.; Shen, J.; Song, Y. Detection of key organs in tomato based on deep migration learning in a complex background. *Agriculture* **2018**, *8*, 196. [[CrossRef](#)]
5. Li, Y.; Feng, Q.; Liu, C.; Xiong, Z.; Sun, Y.; Xie, F.; Li, T.; Zhao, C. MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting. *Eur. J. Agron.* **2023**, *146*, 126812. [[CrossRef](#)]
6. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [[CrossRef](#)]
7. Tao, K.; Wang, A.; Shen, Y.; Lu, Z.; Peng, F.; Wei, X. Peach flower density detection based on an improved cnn incorporating attention mechanism and multi-scale feature fusion. *Horticulturae* **2022**, *8*, 904. [[CrossRef](#)]
8. Rong, Q.; Hu, C.; Hu, X.; Xu, M. Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing. *Comput. Electron. Agric.* **2023**, *210*, 107923. [[CrossRef](#)]
9. Sun, J.; He, X.; Wu, M.; Wu, X.; Shen, J.; Lu, B. Detection of tomato organs based on convolutional neural network under the overlap and occlusion backgrounds. *Mach. Vis. Appl.* **2020**, *31*, 31. [[CrossRef](#)]
10. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. [[CrossRef](#)]
11. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458. Available online: <https://github.com/THU-MIG/yolov10> (accessed on 12 November 2024).
12. Du, X.; Cheng, H.; Ma, Z.; Lu, W.; Wang, M.; Meng, Z.; Jiang, C.; Hong, F. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels. *Comput. Electron. Agric.* **2023**, *214*, 108304. [[CrossRef](#)]
13. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IOU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051. Available online: <https://arxiv.org/abs/2301.10051v3> (accessed on 12 November 2024).
14. Zhang, H.; Xu, C.; Zhang, S. Inner-IOU: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877. Available online: <https://arxiv.org/abs/2311.02877v4> (accessed on 12 November 2024).
15. Hou, G.; Chen, H.; Ma, Y.; Jiang, M.; Hua, C.; Jiang, C.; Niu, R. An occluded cherry tomato recognition model based on improved YOLOv7. *Front. Plant Sci.* **2023**, *14*, 1260808. [[CrossRef](#)] [[PubMed](#)]
16. Han, K.; Wang, Y.; Guo, J.; Wu, E. ParameterNet: Parameters Are All You Need for Large-scale Visual Pretraining of Mobile Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 15751–15761.



17. Peng, Y.; Wang, A.; Liu, J.; Faheem, M. A comparative study of semantic segmentation models for identification of grape with different varieties. *Agriculture* **2021**, *11*, 997. [CrossRef]
18. Azad, R.; Niggemeier, L.; Hüttemann, M.; Kazerouni, A.; Aghdam, E.K.; Velichko, Y.; Bagci, U.; Merhof, D. Beyond self-attention: Deformable large kernel attention for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 1287–1297. [CrossRef]
19. Cheng, Z.; Huang, R.; Qian, R.; Dong, W.; Zhu, J.; Liu, M. A lightweight crop pest detection method based on convolutional neural networks. *Appl. Sci.* **2022**, *12*, 7378. [CrossRef]
20. Gao, J.; Zhang, J.; Zhang, F.; Gao, J. LACTA: A lightweight and accurate algorithm for cherry tomato detection in unstructured environments. *Expert Syst. Appl.* **2024**, *238*, 122073. [CrossRef]
21. Zeng, T.; Li, S.; Song, Q.; Zhong, F.; Wei, X. Lightweight tomato real-time detection method based on improved YOLO and mobile deployment. *Comput. Electron. Agric.* **2023**, *205*, 107625. [CrossRef]
22. Liu, Z.; Abeyrathna, R.R.D.; Sampurno, R.M.; Nakaguchi, V.M.; Ahamed, T. Faster-YOLO-AP: A lightweight apple detection algorithm based on improved YOLOv8 with a new efficient PDWConv in orchard. *Comput. Electron. Agric.* **2024**, *223*, 109118. [CrossRef]
23. Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; Zhou, E. General instance distillation for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7842–7851. [CrossRef]
24. Zhang, B.; Sui, J.; Niu, L. Foreground Object Search by Distilling Composite Image Feature. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 22986–22995. [CrossRef]
25. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4643–4652. [CrossRef]
26. Zhu, S.; Yang, J.; Cai, C.; Pan, Z.; Zhai, W. Application of dynamic vibration absorbers in designing a vibration isolation track at low-frequency domain. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2017**, *231*, 546–557. [CrossRef]
27. Zhang, L.; Chen, X.; Tu, X.; Wan, P.; Xu, N.; Ma, K. Wavelet knowledge distillation: Towards efficient image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12464–12474. [CrossRef]
28. Zhang, Y.; Huang, T.; Liu, J.; Jiang, T.; Cheng, K.; Zhang, S. FreeKD: Knowledge Distillation via Semantic Frequency Prompt. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–22 June 2024; pp. 15931–15940.
29. Li, Y.; Wu, S.; He, L.; Tong, J.; Zhao, R.; Jia, J.; Chen, J.; Wu, C. Development and field evaluation of a robotic harvesting system for plucking high-quality tea. *Comput. Electron. Agric.* **2023**, *206*, 107659. [CrossRef]
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
31. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578. [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*; Springer: Cham, Switzerland, 2020; pp. 213–229.
33. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv5. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 12 November 2024).
34. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
35. 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. [CrossRef]
36. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 12 November 2024).

**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.