agriculture

Article

Pear Fruit Detection Model in Natural Environment Based on Lightweight Transformer Architecture

Zheng Huang [1,2], Xiuhua Zhang [1,2,*], Hongsen Wang [1,2], Huajie Wei [1,2], Yi Zhang [1,2] and Guihong Zhou [2,3]

[1] College of Mechanical and Electrical Engineering, Hebei Agricultural University, Baoding 071001, China; 20237091054@pgs.hebau.edu.cn (Z.H.); 20237091051@pgs.hebau.edu.cn (H.W.); 20227090993@pgs.hebau.edu.cn (H.W.); zhangyi@hebau.edu.cn (Y.Z.)

[2] Hebei Intelligent Agriculture Technology Innovation Center, Baoding 071001, China; zhouguihong@hebau.edu.cn

[3] College of Information Science and Technology, Hebei Agricultural University, Baoding 071001, China

[*] Correspondence: jdzhxh@hebau.edu.cn or zhang72xh@163.com; Tel.: +86-139-3089-3593

Abstract: Aiming at the problems of low precision, slow speed and difficult detection of small target pear fruit in a real environment, this paper designs a pear fruit detection model in a natural environment based on a lightweight Transformer architecture based on the RT-DETR model. Meanwhile, Xinli No. 7 fruit data set with different environmental conditions is established. First, based on the original model, the backbone was replaced with a lightweight FasterNet network. Secondly, HiLo, an improved and efficient attention mechanism with high and low-frequency information extraction, was used to make the model lightweight and improve the feature extraction ability of Xinli No. 7 in complex environments. The CCFM module is reconstructed based on the Slim-Neck method, and the loss function of the original model is replaced with the Shape-NWD small target detection mechanism loss function to enhance the feature extraction capability of the network. The comparison test between RT-DETR and YOLOv5m, YOLOv7, YOLOv8m and YOLOv10m, Deformable-DETR models shows that RT-DETR can achieve a good balance in terms of model lightweight and recognition accuracy compared with other models, and comprehensively exceed the detection accuracy of the current advanced YOLOv10 algorithm, which can realize the rapid detection of Xinli No. 7 fruit. In this paper, the accuracy rate, recall rate and average accuracy of the improved model reached 93.7%, 91.9% and 98%, respectively, and compared with the original model, the number of params, calculation amount and weight memory was reduced by 48.47%, 56.2% and 48.31%, respectively. This model provides technical support for Xinli No. 7 fruit detection and model deployment in complex environments.

Keywords: fruit detection; lightweight Transformer; RT-DETR; Xinli No. 7; FasterNet; Slim-Neck

check for updates

Academic Editors: Xiuguo Zou, Xiaochen Zhu, Wentian Zhang, Yan Qian and Yuhua Li

Received: 3 December 2024
Revised: 23 December 2024
Accepted: 24 December 2024
Published: 25 December 2024

Citation: Huang, Z.; Zhang, X.; Wang, H.; Wei, H.; Zhang, Y.; Zhou, G. Pear Fruit Detection Model in Natural Environment Based on Lightweight Transformer Architecture. *Agriculture* **2025**, *15*, 24. https://doi.org/10.3390/agriculture15010024

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

Pear is one of the main fruits in China. It has a cultivation history of more than 3000 years in China. Its cultivation area and output rank first in the world, and it is known as the king of fruits in China. China's pear production has a positive growth trend, and the export volume has always been above 500,000 tons, which plays an important role in China's agricultural economy. The mechanization level of the pear industry is low, and the picking process requires a large amount of labor, accounting for 35–45% of the total labor input [1], and the picking cost is 50–70% of all the links [2]. The overall process of intelligent and mechanized picking can be generally divided into two parts: one is to realize rapid identification of pear fruits, and the other is to accurately and losslessly grasp the identified fruits [3]. However, due to the influence of light changes, branches and leaves occlusion, fruit overlap and distance environment changes, the model in the actual picking environment is prone to problems such as identification difficulties and slow detection speed. Therefore, how to quickly and accurately identify pear fruits has become the primary problem of automatic picking.

Real-time target detection is an important technology field that is widely used in all walks of life. Integrating visual recognition technology into global agriculture can carry out real-time monitoring and accurate analysis of various factors in the agricultural environment, thus significantly improving agricultural production efficiency, quality and sustainability. In the agricultural field, it is embodied in forest fruit quality detection [4], animal breeding detection [5], plant pest detection [6], fruit target detection [7], fruit density classification [8] and orchard road example segmentation [9]. Existing real-time target detection is generally based on CNN architecture, the most famous of which is the YOLO model [10] because it reasonably balances the trade-off between speed and accuracy. Tan et al. [11] proposed a fragrant pear object detection method based on improved YOLOv8n. Taking YOLOv8n as the base model, he used residual convolution module to optimize C2 f for feature fusion, optimized Spatial Pyramid Pooling Fast (SPPF) to Simplified Spatial Pyramid Pooling Fast (simSPPF), introduced PConv convolution and used Inner-CIoU loss function. Weight parameter sharing is proposed to achieve a lightweight detection head. The average accuracy of the self-built Sweet pear dataset is 94.7%, and the reasoning time of the original model on the self-built dataset is 62.9 ms. Zheng et al. [12] designed a lightweight pear target detection M-YOLOv7-SCSN+F model. The data enhancement method based on the Fourier transform generated new image data by analyzing image frequency domain information and reconstructing image amplitude components, thereby improving the model generalization ability. Liu et al. [13] proposed a detection method based on MAE-YOLOv8 for small objects in a real complex orchard environment. By replacing the feature pyramid network, the authors improved the detection accuracy of small target objects. In order to alleviate the problem of missing detection and inaccurate positioning caused by overlapping occlusion, the minimum point distance intersection is introduced as a regression loss function. Chen et al. [14] proposed an improved YOLOv8-based multi-objective segmentation method for the apple tree at the emerging stage and combined advanced convolutional network modules (ConvNeXt V2, Multi-Scale Dilated Attention (MSDA) and Distribution Shifting Convolution (DSConv)) to enhance YOLOv8 and improve the accuracy of organ segmentation in complex natural environments. The above models show certain advantages in accuracy and speed, but YOLO models usually require Non-Maximum Suppression (NMS) for post-processing, and the introduction of hyperparameters makes the accuracy and speed of the model unstable, slowing down the reasoning speed of the model [15]. In addition, the need to select reasonable NMS thresholds has hindered the development of real-time object detection.

DETR (Detection Transformer) [16] was first proposed by Facebook, and Transformer architecture [17] was introduced into the target detection network, treating detection as a collection prediction problem without the need to generate candidate regions and post-processing steps. In recent years, the application of Transformers in real-time target detection has become an important research direction in the field of computer vision. The introduction of the Transformer provides a new way of thinking for traditional Convolutional Neural Networks (CNN). It is particularly good when it comes to handling long-term dependencies, global context information, and enhancing the expressiveness of the model. Although DETR is very successful, its training convergence speed is slow, and its detection performance for small targets is poor. Therefore, RT-DETR [18], an end-to-end real-time target detector based on Transformer architecture, provides a method to solve such problems. The hybrid encoder of DETR introduces multi-scale features to accelerate training convergence and improve performance [19], but the sharply increased sequence length still causes the encoder to become a computational bottleneck. An efficient hybrid encoder is designed in RT-DETR to replace the original Transformer encoder. Multi-scale features can be rapidly processed by cross-scale fusion and decoupling intra-scale interaction. The encoder can effectively process features of different scales, greatly reduce the computational load of the encoder and significantly improve the reasoning speed. In order to reduce the difficulty of the object query, DETR uses the confidence score to select the best feature in the encoder to initialize the object query [20,21]. However, the current query selection leads to uncertainty in the selection feature, which affects the model performance. Therefore, RT-DETR (The full name of RT-DETR is in the Appendix A Table A1) proposes the IoU-aware query selection in the decoder. By providing IoU constraints in the training process, higher quality initial object queries are provided for the decoder and the detection accuracy is improved. Zhao et al. [22] proposed a lightweight cherry tomato ear state recognition model based on an improved Transformer. By replacing the trunk structure and adding an adaptive detail fusion module, the calculation and model parameters are significantly reduced, and the average accuracy of 90% is guaranteed, while low calculation and fast detection are realized. Hu et al. [23] proposed an improved RT-DETR detection model, RIC-DETR, in which ResNet18 was selected as the backbone feature network, and the reverse residual mobile module was introduced while the second innovation and improvement were carried out. Under the condition that the average accuracy of 97.2% was maintained, the computation, parameter count and memory footprint are greatly reduced. Li et al. [24] proposed a chicken target detection model with high precision and strong generalization based on improved Real-Time Detection Transformer (RT-DETR) Efficient Multi-Scale-Conv Detection Transformer (EMSC-DETR). In order to solve the problem of small target features being easily lost, a Study Data Tabulation Model (SDTM) was introduced. The module significantly improves the computational efficiency of the converter, with $\mathrm{mAP}_{0.5}$ being 98.6%. Li et al. [25] proposed an end-to-end semi-supervised object detection method based on DEtection TRansformer (DETR), which simplified the post-processing process without the need for Non-Maximum Suppression (NMS) and adopted a more advanced binary matching allocation strategy. The proposed method only used 5% of the total data to achieve 74.1% of the mAP. Although the performance of RT-DETR is slightly inferior to that of YOLO in small target detection, it is superior in speed and accuracy to current real-time detectors of similar scale.

Aiming at the problems of low precision and slow detection speed of pear fruit in a real environment and difficult detection of small target fruit in a long-term environment, this study improved and designed a pear fruit detection model in a natural environment based on RT-DETR. Firstly, ResNet-r18 was replaced with a lightweight FasterNet network. Secondly, HiLo, an efficient attention mechanism for extracting high and low-frequency information, is used to improve the Attention-based Intrascale Feature Interaction module (AIFI), which achieves higher performance and faster speed. A new convolutional GSConv is introduced into the Compact Convolutional Feature Fusion Module (CCFM) to reduce the complexity of the model and maintain high recognition accuracy. Finally, the loss function in RT-DETR is replaced with Shape-IoU and used in combination with the Normalized Wasserstein Distance (NWD) small target detection mechanism to further improve the detection performance of small target pears.

2. Materials and Methods

2.1. Construction of Data Sets

2.1.1. Image Acquisition

The image data was collected from the modern pear garden demonstration base of Hebei Wokang Agricultural Science and Technology Co., Ltd., Xingtai City, Hebei Province, China. The collected pear variety was Xinli No. 7, and the collection time was from July to August 2024, during the fruit ripening and picking period. Fruit characteristic parameters of Xinli No. 7; The shape is oval, the fruit is medium to large, the average single fruit weight is 185.85 g, the longitudinal diameter is 72.61 mm, and the transverse diameter is 71.73 mm. This variety has the characteristics of precocious maturity and a long natural harvesting period, as well as resistance to disease and insect pests and storage resistance [26]. It is rare for Chinese pear crossbreeding to combine these four fine traits in a single fruiting individual at the same time, which is a germplasm innovation in Chinese pear breeding.

Through field investigation, the plant spacing of a modern pear garden is 1~2 m, row spacing is 4 m, tree height is 3~4 m, and the main cultivation mode is wide row close planting. Taking the JAKA C12 robotic arm as the picking robotic arm, the picking working radius of the robotic arm is 1327 mm, so two shooting distances are designed for close-range shooting (100~500 mm) and long-range shooting (700~1350 mm). In order to make the time distribution of the collected data set closer to the actual picking time of fruit farmers, the shooting time of this data set was selected in two periods: 7:00–12:00 and 13:00–19:00. The image shooting equipment was iPhone 13, Honor Magic 3 pro and ZED 2i binocular cameras, which were saved as jpg format. A total of 7468 images of Xinli No. 7 fruit in different environments and scenes were collected in the experiment. The data set was classified according to near-vision conditions, including 4001 close-range images and 3467 long-range images. According to lighting conditions, 3357 low-light images, 1890 high-light images and 2221 backlight images were classified. In the data set, the factors that influenced the collection, such as branch and leaf occlusion, fruit overlap, different distances, and lighting conditions, were comprehensively considered, and part of the collected images are shown in Figure 1.

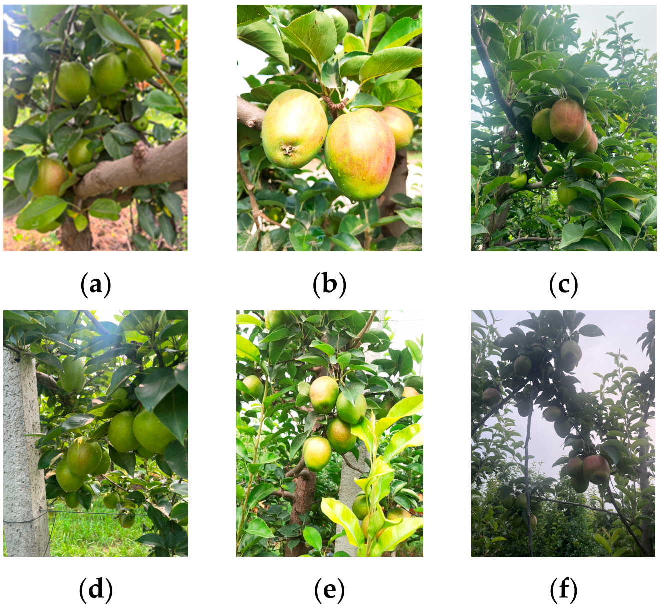


Figure 1. Fruit image of Xinli No. 7 under different environment. (**a**) Branches and leaves cover. (**b**) Frontlight close view. (**c**) Backlight close view. (**d**) Fruit overlap. (**e**) Frontlight distant view. (**f**) Backlight distant view.

2.1.2. Data Set Creation

In order to improve the annotation efficiency, YOLO semi-automatic annotation was adopted [27]. First, labelImg was used to annotate 2000 fruit images of Xinli No. 7 in different scenes, and the YOLOv7 model was applied to train the annotated images to obtain the optimal weight file, and then semi-automatic annotation was performed on the remaining fruit images of Xinli No. 7. Finally, labelImg was used to manually adjust the wrong label and missing label of the annotation result.

After manual and semi-automatic annotation, a total of 7468 XML files of Xinli No. 7 were obtained and randomly divided according to the ratio of 8:1:1 to form a training set, verification set and test set, in which the training set contained 5974 images, the verification set and the test set each contained 747 images.

2.2. Experimental Method

2.2.1. Target Detection Selection Index

In this paper, Precision (P, %), Recall (R, %), mean Average Precision (mAP, %), Frames Per Second (FPS, frames/s), Floating-point Operations Per second (FLOPs, G), parameter number (Params, M) and Model size (MB) are used as model evaluation indicators, and P, R and mAP are used to measure the detection accuracy of the model [28], FPS is used to measure the detection speed of the model, model lightweight is measured by the number of parameters and the amount of computation, and deployment cost is assessed using the space occupied by the model.

2.2.2. Training of Network Models

In this study, the Intel core i7-14700 KF Win11 operating system is used, the main frequency is 3.4 GHZ, the running memory is 32 G, and the Nvidia GeForce RTX 4070 Ti SUPER graphics card is installed. Experiments were carried out on the Pycharm platform, the PyTorch deep learning framework was configured for environment construction, and Python language was used for algorithm writing. Model hyperparameters were set as follows: the default image is 640 × 640 pixels, the number of model training iteration cycles was set to 200, the number of samples processed in each batch was set to 8, the number of worker threads during data loading was set to 4, and the initial learning rate was set to 0.0001. All other training hyperparameters were used as default values, and all tests of the model were carried out under the same environment.

3. Network Model and Improvement

This paper improves a pear detection model under a natural environment based on the end-to-end real-time target detector RT-DETR based on Transformer architecture. First, the backbone of the original model is replaced with a lightweight FasterNet network. Secondly, HiLo, an efficient attention mechanism that can extract high and low-frequency information, is used to improve the Attention-based Intrascale Feature Interaction (AIFI), and a GSConv convolution is introduced into the Cross-Scale Feature Fusion Module (CCFM). Finally, the loss function in the original model is replaced with the loss function of the Shape-NWD small target detection mechanism. The model performance is improved. The network structure diagram of the improved model is shown in Figure 2.

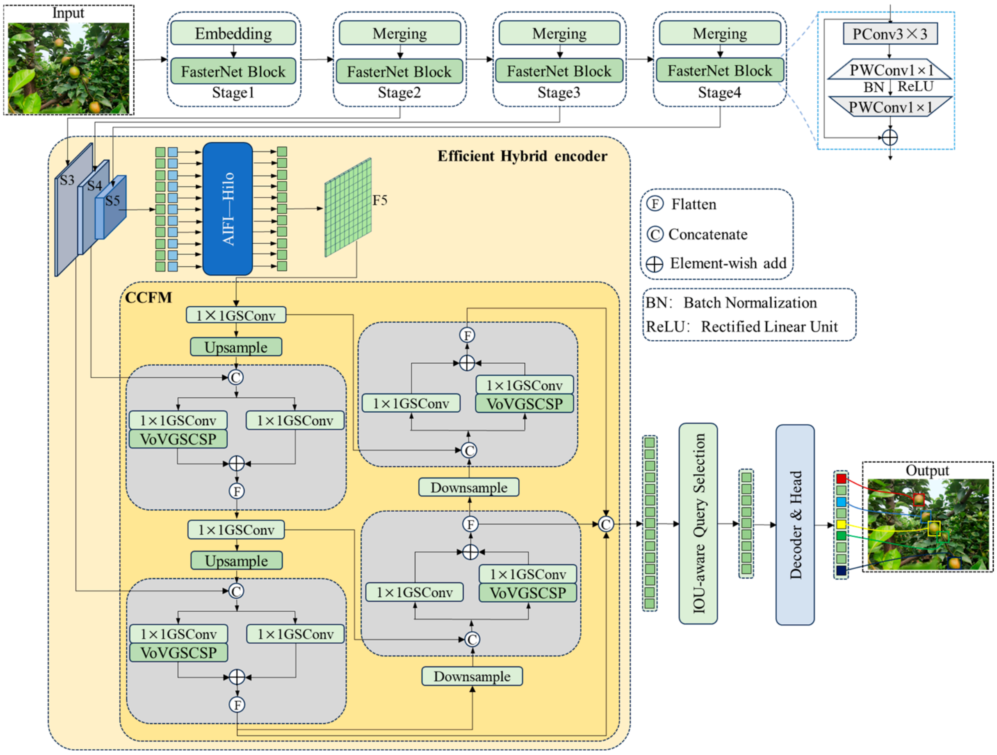


Figure 2. Improved network model structure diagram.

3.1. Detect Model Backbone Network Replacement

In order to make the improved RT-DETR model closer to practical applications, this paper re-studies the commonly used backbone network model and finds that most models focus on reducing Floating point operations (FLOPs), but the reduction of FLOPs does not necessarily mean the reduction of the same horizontal delay, as shown in Formula (1). The main reason is that Floating point operations per second (FLOPS) need to be optimized at the same time to achieve truly low latency. Therefore, this paper introduces a lightweight FasterNet network as the backbone network for feature extraction and introduces a simple but fast and effective convolutional PConv, which can extract spatial features more efficiently while reducing redundant computation and memory access. The FasterNet Block is shown in the blue dashed box in Figure 2.

$$\text{Latency} = \frac{\text{FLOPs}}{\text{FLOPS}} \tag{1}$$

Compared with the traditional ResNet network, the FasterNet network can greatly reduce the parameter number and calculation amount of the model when the precision rate, recall rate, and mean average precision are slightly reduced [29], realizing the light weight of the model, which is more conducive to deployment on the pear fruit picking robot. There are four hierarchical stages in the FasterNet backbone network, each of which is preceded by an embedding layer (step size 4, Conv4×4) or a merging layer (step size 2, Conv2×2) for spatial downsampling and channel number expansion, respectively [30]. In order to fully and efficiently utilize the information from all channels, each directional residual block consists of a PConv layer at the front end and two PWConv layers at the back end, in which a Batch Normalization (BN) is placed in the middle of the two PWConv layers. The Rectified Linear Unit (ReLU) activation function is used to enhance the input features of the Xinli No. 7 fruit image, thereby improving model performance and training speed.

3.2. Introduction of Improved Efficient Attention Mechanism HiLo

In order to solve the problem of the huge computational cost of throughput in high-resolution images in the original RT-DETR model, especially in this task, a new efficient attention mechanism, HiLo, was introduced, and the scaled dot product attention of the attention branch was re-improved to adapt the two high and low-frequency attention branches. One path encodes high-frequency interactions by scaling dot-product attention and relatively high-resolution feature maps, while the other encodes low-frequency interactions by dot-product attention and downsampled features and finally incorporates improved attention mechanisms into the in-scale feature interaction module. The HiLo attention mechanism uses two kinds of effective attention to decoupling the high and low frequencies in the feature graph [31], eliminating the huge computational burden of different lower frequencies in the standard Multi-head Self-Attention layer (MSA) features.

Compared to existing standard attention mechanisms, such as self-attention mechanisms, CBAM or Transformer models, HiLo attention mechanisms improve the representation and computational efficiency of the model through more sophisticated processing and fusion of high-level and low-level information. As shown in Figure 3, the HiLo attention mechanism divides the MSA layer into two components: One is used to encode high-frequency attention branch Hi-Fi with local self-attention mechanism and high-resolution feature map, and the other is used to globally encode low-frequency attention branch Lo-Fi by subsampling features. The output of each HiLo attention mechanism is a series of high-frequency attention branches and low-frequency attention branches, as shown in formula (2). Thus, the information extraction efficiency [32] is effectively improved, which is more effective than standard MSA.

$$\mathrm{HiLo}(\mathrm{X}) = [\mathrm{Hi\text{-}Fi}(\mathrm{X});\ \mathrm{Lo\text{-}Fi}(\mathrm{X})] \tag{2}$$

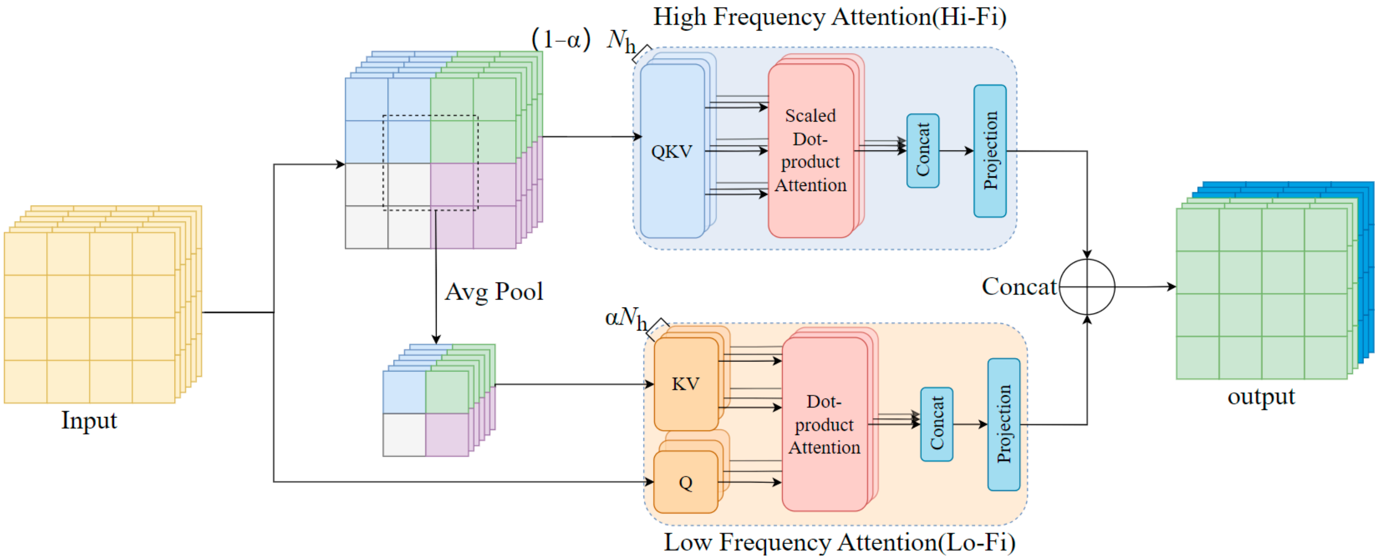


Figure 3. High-efficiency attention mechanism HiLo framework diagram.

HiLo allocates Hi-Fi and Lo-Fi with the same structure as the standard Multi-head Self-Attention layer (MSA). In order to make the allocation scheme more favorable, the same number of magnetic heads in the MSA are divided into two groups, as shown in Formula (3), and the division ratio is: where $(1 - \alpha)\mathrm{N_h}$ heads are used for Hi-Fi, Other $\alpha\mathrm{N_h}$ heads are used for Lo-Fi, and each attentional architecture is less complex than a standard MSA, so the overall framework of the HiLo attentional mechanism guarantees a low complexity model with high throughput for high-resolution images.

$$\text{HiLo} = \alpha\text{Hi-Fi} + (1 - \alpha)\text{Lo-Fi} \tag{3}$$

3.3. Reconstruction of Cross-Scale Feature Fusion Module Based on Slim-Neck Method

In order to further meet the real-time detection requirements of the pear fruit-picking robot, this study introduced lightweight convolutional GSConv into the Cross-Scale Feature Fusion Module (CCFM), which can utilize local information and global information at the same time and realize the fusion of these different scale information. It can further reduce the complexity of the model while maintaining accuracy and solve the problem of the speed of prediction calculation in convolutional neural networks. As shown in Figure 4 below, GSConv first inputs an ordinary convolution undersampling, then uses DWConv deep convolution to concatenate the output results of the two CONVs, and finally performs data distribution shuffle operation to concatenate the corresponding channel numbers of the previous two convolution. Therefore, when the spatial information of the input image is gradually transferred to the channel, GSConv convolution avoids the phenomenon of partial loss of semantic information caused by spatial compression and channel expansion of each feature image [33].

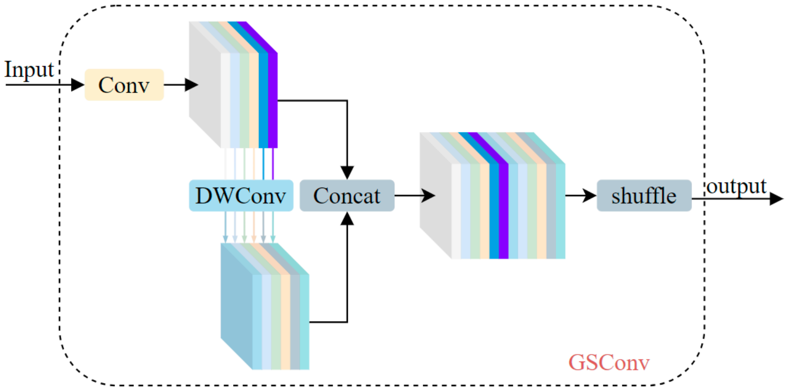


Figure 4. GSConv module framework diagram.

After introducing lightweight GSConv convolution, we continue to introduce GS bottleneck and interstage partial network modules VoV-GSCSP, which are designed to further improve feature utilization efficiency and network performance. An improved transformation structure of the VoV-GSCSP module is proposed, as shown in Figure 5. Due to its simple structure, this module only consists of lightweight GSConv convolution and GS bottleneck, which requires less training hardware while ensuring performance.

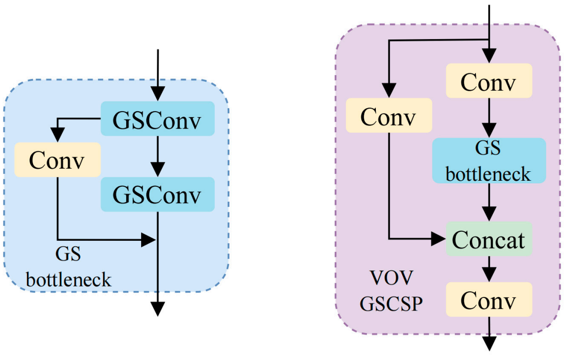


Figure 5. GS bottleneck and VoV-GSCSP cross-stage network modules.

Therefore, this cross-scale feature Fusion Module based on Slim-Neck architecture is more in line with the requirements of this paper for a lightweight model and low computing cost and can achieve significant accuracy improvement while meeting the deployment conditions of mobile terminals.

3.4. Shape-NWD Small Target Detection Mechanism Loss Function

In the RT-DETR algorithm, the prediction frame regression loss function uses GIoU. When the two prediction frames have the same height and width and are in the same horizontal plane, GIoU degenerates into IoU loss function, which leads to the problem of slow convergence and inaccurate regression. To solve the above problems, this study adopts the shape-IoU loss function to replace the GIoU loss function used by RT-DETR [34]. This method can calculate the loss by focusing on the shape and scale of the bounding box itself so as to make bounding box regression more accurate. Figure 6 shows the schematic diagram of Shape-IoU parameters.

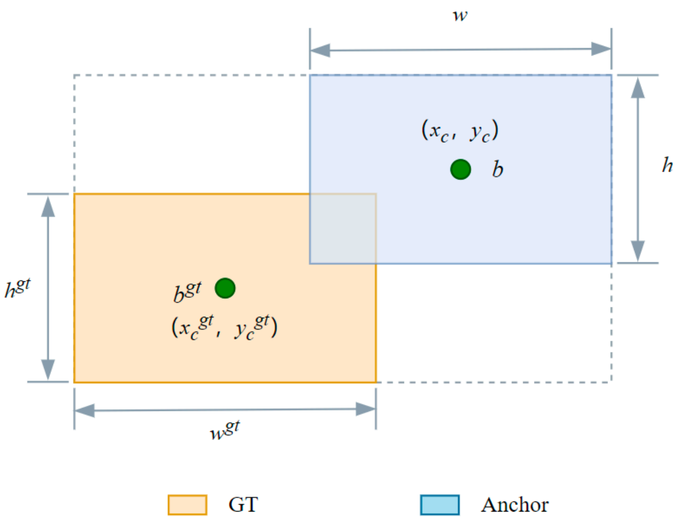


Figure 6. Schematic diagram of Shape-IoU parameters.

The formula of Shape-IoU can be derived from Figure 6 and Formulas (4)–(6).

$$\mathrm{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (4)$$

$$distance^{shape} = hh \times \left(x_c - x_c^{gt}\right)^2 / c^2 + ww \times \left(y_c - y_c^{gt}\right)^2 / c^2 \quad (5)$$

$$\Omega^{shape} = \sum_{t=w,h} \left(1 - e^{-wt}\right)^{\theta}, \theta = 4 \quad (6)$$

where IoU is the actual crossover ratio, ww and hh represent the weight coefficients in the horizontal direction and vertical direction respectively, and their values are related to the shape of the GT frame. Shape-IoU loss is defined by Formula (7).

$$L_{\text{Shape-IoU}} = 1 - \mathrm{IoU} + distance^{shape} + 0.5 \times \Omega^{shape} \quad (7)$$

The NWD [35] small target detection method was introduced into the target detection of Xinli No. 7 fruit in the test, and the NWD small target detection mechanism and Shape-IoU loss function were combined. In this experiment, the strategy with 50% of each is selected. As shown in Formula (8), the loss function of the Shape-NWD small target detection mechanism can not only improve the target detection accuracy but also ensure the model detection speed.

$$\text{Shape-NWD} = (1 - \text{IoU})(1 - 0.5\text{NWD}) + \text{IoU}(1 - 0.5\text{Shape-IoU}) \tag{8}$$

4. Test Results and Analysis

4.1. Performance Comparison Test of RT-DETR Model

In order to select the optimal model file, the performance of ResNet-r18, ResNet-r34 and ResNet-r50 backbone networks commonly used in the ResNet series were compared under the same conditions, taking Precision, Recall, $\text{mAP}_{0.5}$, Params, FLOPs, Model size and FPS as the basic evaluation indexes. The test results are shown in Table 1 below. Under the condition that the accuracy rate, recall rate and average accuracy mean are not different, the RT-DETR-r18 model file meets the requirements of rapid detection of pear fruits while meeting the lightweight requirements, and the generated weight model is 38.5 MB, which is more conducive to deployment on the mobile end of picking equipment.

Table 1. Performance comparison of RT-DETR models.

Model	Params/M	FLOPs/G	P/%	R/%	mAP/%	Weights Size/MB	FPS·(f s^{-1})
RT-DETR-r18	19.87	57.3	89.7	90.1	96.6	38.5	75.7
RT-DETR-r34	31.1	88.8	89.5	90.6	96.7	60	61.6
RT-DETR-r50	41.96	129.5	90.4	89.8	96.6	164	49.8

4.2. Comparative Analysis of Different Algorithms

YOLOv5m, YOLOv7, YOLOv8m, YOLOv10m, Deformable-DETR and RT-DETR-r18 models were used to perform performance comparison on Xinli No. 7 data set, and the test results were shown in Table 2. According to the results in Table 2 below, it can be seen that the number of parameters and calculation amount displayed by YOLOv5m are small, but they do not meet the requirements for rapid detection of pear fruit. Although YOLOv7 maintains high recognition accuracy, the model volume is large, which does not meet the requirements of a lightweight model. Although YOLOv8m and YOLOv10m have high detection accuracy, they do not meet the requirements of lightweight models due to the large amount of calculation. Compared with the YOLOv8m model, RT-DETR-r18 has a 5.97 M and 21.4 G reduction in parameter number and computation amount, a 0.8% increase in average accuracy, and an 11.1 MB reduction in weight memory. Compared with the YOLOv10m model, RT-DETR-r18 is superior to YOLOv10m in precision, recall and average accuracy, although the params and weight memory are increased by 3.42 M and 6.6 MB, the computation amount is decreased by 6.1 G and the FPS is increased by 5.28 f/s.

Table 2. Analysis of comparison results of different algorithms.

Model	Params/M	FLOPs/G	P/%	R/%	mAP/%	Weights Size/MB	FPS·(f·s^{-1})
YOLOv5m	21.04	50.2	87.7	91	96.4	40.4	15.9
YOLOv7	36.48	103.2	91.8	89.5	97.1.	71.3	24.04
YOLOv8m	25.84	78.7	89.1	89.7	95.8	49.6	75.19
YOLOv10m	16.45	63.4	89.3	90	96.3	31.9	70.42
Deformable-DETR	40	196	88.6	88.3	95.7	86	29.5
RT-DETR-r18	19.87	57.3	89.7	90.1	96.6	38.5	75.7

Deformable-DETR has a significant gap with RT-DETR in all evaluation indexes in this paper, especially in terms of parameter number, computation amount and weight memory, 20.13M, 138.7 G and 47.5 MB higher than the RT-DETR model. However, RT-DETR can achieve a good balance in terms of model lightweight and recognition accuracy, and the transmission frame per second can reach 75.7 f/s, meeting the needs of practical applications and realizing the real-time accurate detection of Xinli No.7 fruit.

4.3. Ablation Study

The ablation test is an important means to evaluate the effectiveness of model improvement. In order to verify the validity of the modules and loss functions proposed in this paper, the ablation test was designed and carried out under the same experimental environment following the principle of the control variable method. As shown in Table 3, test 1 represents the performance indicators of RT-DETR before improvement. Through the following test improvements, the comparison of key evaluation indicators of the improvement points can be visually observed. The results show that the identification performance of the model is improved compared with that of the original model.

Table 3. RT-DETR ablation results based on Xinli No. 7 data set.

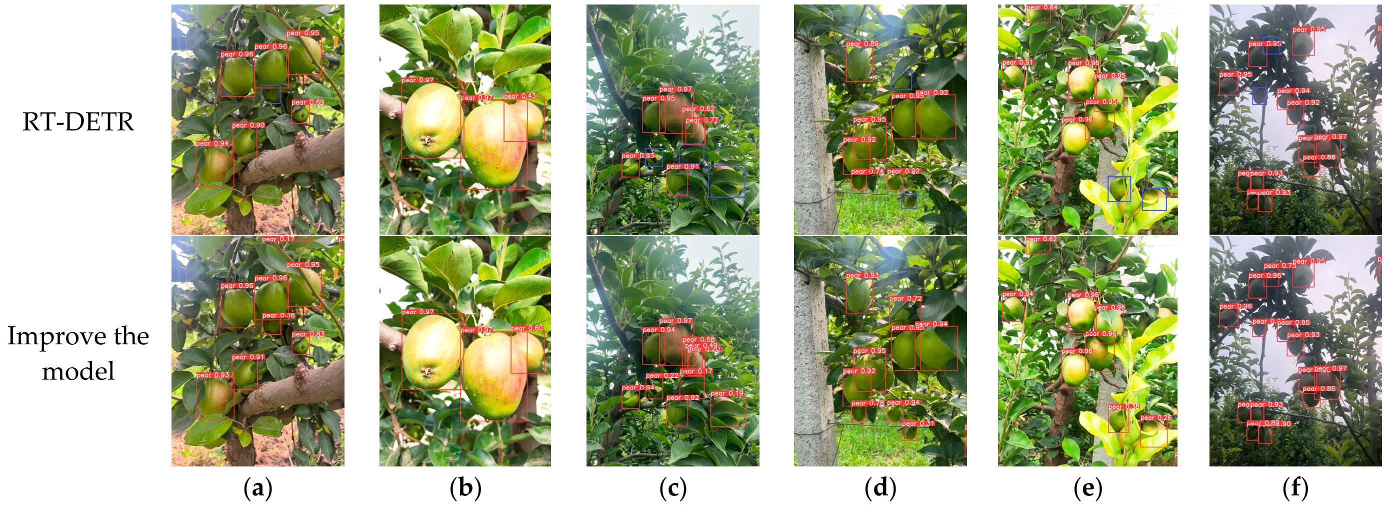
Model	FN	HiLo	SN	Shape-NWD	Params/M	FLOPs/G	P/%	R/%	mAP/%	Weights Size/MB
Experiment1	×	×	×	×	19.87	57.3	89.7	90.1	96.6	38.5
Experiment2	√	×	×	×	10.81	28.5	89.2	89.1	96.1	21.2
Experiment3	√	√	×	×	10.78	28.6	92	90.1	97.1	21.2
Experiment4	√	√	√	×	10.24	25.1	91.3	91.5	97.3	19.9
Experiment5	√	√	√	√	10.24	25.1	93.7	91.9	98	19.9

It can be seen from experiment 2 in Table 3 that after introducing the lightweight FasterNet (FN) network as the backbone network for feature extraction, the number of params and calculation amount of the model is reduced by 45.6% and 50.26%, respectively, and the weight memory is also significantly reduced, which realizes the lightweight of the model. This improvement is attributed to the simple and efficient extraction of spatial features of the Xinli No. 7 fruit image after the introduction of PConv convolution, the reduction of computational redundancy and memory access, and the introduction of new convolution can maintain high FLOPs while reducing FLOPS. Secondly, after introducing HiLo, an improved and efficient attention mechanism for extracting high and low-frequency information, in experiment 3, the recognition accuracy of the algorithm was greatly improved, with the recall rate, accuracy rate and average accuracy increased by 1%, 2.8% and 1%, respectively, while the number of params, calculation amount and weight memory remained basically unchanged. This improvement is due to the attention mechanism. HiLo uses two efficient attention types to decouple the high/low frequencies in the feature map, refining the high/low-frequency features and connecting them by capturing the fine grain features of the high/low-frequency Xinli No.7 fruit image through high-frequency attention (Hi-Fi) and low-frequency attention (Lo-Fi). Moreover, in experiment 4, the Slim-Neck (SN) method was used to reconstruct the Cross-Scale Feature Fusion Module (CCFM). Although the accuracy was slightly reduced, the weight memory and computing cost were reduced to 19.9 MB and 25.1 G, respectively, because GSConv was used to replace Conv convolution. Because the structure of the VoV-GSCSP module is improved and transformed, the model structure is simple and efficient, and the simultaneous recall rate and average accuracy value are improved slightly.

Finally, in experiment 5, the loss function was replaced by the Shape-NWD small target detection mechanism loss function. Through the improvement of the above module, the model recognition accuracy is continuously improved. Finally, the accuracy rate, recall rate and average accuracy of the improved model are increased by 4%, 1.8% and 1.4%, respectively, compared with the original model in experiment 1.

After the ablation test, all four modules in this paper are effective, and the average accuracy of the improved model reaches 98%, which is 9.7% and 0.77% higher than the 88.3% and 97.23% in previous pear fruit detection model literature [11,12], respectively, showing the advanced and deployable of the improved model in this paper.

Figure 7 shows the effect of the original model and the improved model on the target detection of Xinli No. 7 fruit under different scenario conditions. Both the RT-DETR model and the improved model can accurately identify the fruit target of Xinli No. 7 in the image under the frontlight close view, but the improved model could detect the missed fruit to various degrees under branches and leaves cover, backlight close view, fruit overlap, frontlight distant view and backlight distant view, blue boxes missed due to fruit occlusion in Figure 7a,f and blue baskets missed due to branches and leaves occlusion in Figure 7c–e is shown. Moreover, the improved model has higher confidence in different environments. As shown in Figure 7b, the fruit confidence detected by the improved model is 17% higher than that of RT-DETR for Xinli No. 7 fruit, that is, fruit overlap. The test results show that the improved model has a better detection effect and higher confidence in the complex environment of branches and leaves cover, fruit overlap and reverse light near and far, and meets the recognition accuracy requirements of Xinli No. 7 fruit picking robot.



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

Figure 7. Fruit target detection results of Xinli No.7 under different environments. (**a**) Branches and leaves cover. (**b**) Frontlight close view. (**c**) Backlight close view. (**d**) Fruit overlap. (**e**) Frontlight distant view. (**f**) Backlight distant view.

4.4. Construction of Grading Detection Model

The improved model in this paper tested Xinli No. 7 fruit in 6 typical environments, and the test results are shown in Table 4 below. As can be seen from the results in the table, there is a large difference in recognition accuracy between recall rate and accuracy rate in frontlight close view. The reason for this result is that the data collected in frontlight close view is small, and the model cannot be fully learned and verified, and it is difficult to ensure generalization. Another reason is that the color and shape characteristics of the Xinli No. 7 fruit are not prominent enough in the close view of the frontlight, which makes it easy to confuse the appearance of branches and leaves, resulting in the missing some fruits in the environment. The model showed good detection ability in the environment of backlight near prospect, branches and leaves occlusion and fruit overlap. The detection effect of pear fruit in the frontlight distant view is especially the best, and the accuracy rate, recall rate, and average accuracy average have reached a good balance. This result can be said that the improved model is suitable for the recognition distance requirement of the robot arm picking pear fruit set in this paper. No matter what kind of typical pear fruit environment, the average accuracy of the improved model is the lowest, 98.4%, which realizes the accurate identification of Xinli No. 7 fruit in the natural environment.

Table 4. An improved model for pear fruit detection results in different environments.

Typical Pear Fruit Environment	P/%	R/%	mAP/%
Frontlight close view	98.8	91.6	98.4
Frontlight distant view	96	94.3	98.8
Backlight close view	94.3	94.8	98.4
Backlight distant view	95.7	93.1	98.7
Branches and leaves cover	94.7	93.9	98.5
Fruit overlap	95	93	98.6

5. Discussion

In terms of data set construction, although this paper comprehensively considered the influence factors such as branch and leaf occlusion, fruit overlap, different distances and light conditions, it did not build the diverse Xinli No. 7 fruit dataset under complex environmental conditions in different pear orchards in different regions, and it also lacked the construction of other pear fruit variety data sets. Therefore, the subsequent data set construction can comprehensively consider the above requirements. Continue to increase the stability and generalization of the model to detect the pear fruit. In terms of limitations, the model in this paper was only trained and applied in the experimental equipment and environment, and it can be adapted to other hardware devices such as laptops, smartphones or embedded devices in the later stage. It is also necessary to run stably under extreme environmental conditions with limited computing resources and declining data quality to verify the adaptive ability and delay of the improved model. Overall, the pear detection model, through the application of deep learning and computer vision technology, can greatly improve the efficiency of agriculture, food safety and supply chain management, promote the development of precision agriculture, and play an important role in the sustainability of the agricultural industry.

In the future practical application, we will continue to study how to deploy the model algorithm in the mobile end of the pear fruit picking equipment and integrate debugging with the robot arm, the chassis of the picking device and the end effector so as to achieve the long-term goal of automatic and non-destructive pear fruit picking [3].

6. Conclusions

(1) Based on the end-to-end real-time target detector RT-DETR model of Transformer architecture, this paper designs a pear fruit detection model in a natural environment based on lightweight Transformer architecture, aiming at the problems of low detection accuracy, slow speed and difficult detection of small target pears in a real environment. The accuracy rate, recall rate and average accuracy of the model reach 93.7%, 91.9% and 98%, respectively, and the number of parameters, calculation amount and weight memory reach 10.24 M, 25.1 G and 19.9 MB, respectively. Therefore, this model not only achieves high recognition accuracy but also has the requirements for deployment in automated pear-picking robots, fruit measurement devices and mobile terminals of automatic sorting systems.

(2) The performance comparison tests of YOLOv5m, YOLOv7, YOLOv8m and YOLOv10m, Deformable-DETR and RT-DETR-r18 models on Xinli No. 7 dataset were designed, and three comprehensive evaluation indexes of model lightweight, recognition accuracy and detection speed were used to evaluate the self-built Xinli No. 7 fruit dataset. The results show that the RT-DETR-r18 model can achieve a good balance in

terms of model lightweight and recognition accuracy compared with other models. The transmission frame per second is 75.7 f/s, which can realize rapid and accurate detection of Xinli No.7 fruit.

(3) The ablation experiment was divided into five groups. Based on the original model of the first group, ResNet-r18 was replaced with a lightweight FasterNet backbone network respectively. Secondly, the AIFI module was improved by using HiLo, an improved and efficient attention mechanism with high and low-frequency information extraction. A simple and efficient GSConv convolution is introduced into the CCFM module, and the loss function GIoU in RT-DETR is replaced by the Shape-NWD small target detection mechanism loss function. The results show that compared with the original model, the accuracy, recall and average accuracy of the improved model are increased by 4%, 1.8% and 1.4%, respectively, and the number of params, calculation and weight memory is reduced by 48.47%, 56.2% and 48.31%, respectively, so as to meet the requirements of model lightweight and accurate identification of pear fruits.

Author Contributions: Conceptualization, X.Z. and Z.H.; Methodology, Z.H. and Y.Z.; Software, Z.H., H.W. (Hongsen Wang) and H.W. (Huajie Wei); Validation, X.Z., Z.H. and H.W. (Hongsen Wang); Formal analysis, Z.H. and G.Z.; Investigation, H.W. (Hongsen Wang) and (Huajie Wei); Resources, X.Z. and G.Z.; Data curation, X.Z., Y.Z. and H.W. (Hongsen Wang); Writing—original draft preparation, X.Z., Z.H. and G.Z.; Writing—review and editing, Z.H., Y.Z. and H.W. (Hongsen Wang) and H.W. (Huajie Wei); Visualization, Z.H., Y.Z., G.Z. and H.W. (Huajie Wei); Supervision, X.Z.; Project administration, X.Z.; Funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Department of Agriculture and Rural Affairs of Hebei Province (Hebei Agriculture Research System HBCT2024170207).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors gratefully acknowledge the editors and anonymous reviewers for their constructive comments on our manuscript.

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

Appendix A

Table A1. Abbreviation list.

Abbreviation	Full Title
RT-DETR	Real-Time Detection Transformer
SPPF	Spatial Pyramid Pooling Fast
simSPPF	Simplified Spatial Pyramid Pooling Fast
MSDA	Multi-Scale Dilated Attention
NMS	Non Maximum Suppression
CNN	Convolutional Neural Networks
EMSC-DETR	Efficient Multi-Scale-Conv Detection Transformer
SDTM	Study Data Tabulation Model
FPS	Frames Per Second
AIFI	Attention-based Intrascale Feature Interaction
DSConv	Distribution Shifting Convolution
Shape-NWD	Shape-Normalized Wasserstein Distance
CCFM	Cross-Scale Feature Fusion Module
GSConv	Generalized-Sparse Convolution
FLOPs	Floating point operations
FLOPS	Floating point operations per second
BN	Batch Normalization
ReLU	Rectified Linear Unit
MSA	Multi-head Self-Attention

References

1. Yuan, Y.; Bai, S.; Niu, K.; Zhou, L.; Zhao, B.; Wei, L.; Xiong, S.; Liu, L. Research progress on mechanized harvesting technology and equipment for forest fruit. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 53–63.
2. Wang, B. Research on Key Technologies of Pear Fruit Picking Robot Based on ROS and YOLOv5. Master's Thesis, Hebei University, Hebei, China, 2024.
3. Li, M.; Liu, P. A bionic adaptive end-effector with rope-driven fingers for pear fruit harvesting. *Comput. Electron. Agric.* **2023**, *211*, 107952. [CrossRef]
4. Hai, T.; Zhang, N.; Lu, X.; Xu, J.; Wang, X.; Hu, J.; Ji, M.; Zhao, Z.; Wang, J.; Dong, M. Implementation and Evaluation of Attention Aggregation Technique for Pear Disease Detection. *Agriculture* **2024**, *14*, 1146. [CrossRef]
5. Liu, L.; Xu, S.; Chen, J.; Wang, H.; Zheng, X.; Shen, M.; Liu, L. Detection of Feeding Behavior in Lactating Sows Based on Improved You Only Look Once v5s and Image Segmentation. *Agriculture* **2024**, *14*, 1402. [CrossRef]
6. Dai, G.; Tian, Z.; Fan, J.; Sunil, C.; Dewi, C. DFN-PSAN: Multi-level deep information feature fusion extraction network for interpretable plant disease classification. *Comput. Electron. Agric.* **2024**, *216*, 108481. [CrossRef]
7. Zhao, Y.; Li, Y.; Xu, X. Object Detection in High-Resolution UAV Aerial Remote Sensing Images of Blueberry Canopy Fruits. *Agriculture* **2024**, *14*, 1842. [CrossRef]
8. Jiang, L.; Wang, Y.; Wu, C.; Wu, H. Fruit Distribution Density Estimation in YOLO-Detected Strawberry Images: A Kernel Density and Nearest Neighbor Analysis Approach. *Agriculture* **2024**, *14*, 1848. [CrossRef]
9. Wu, W.; He, Z.; Li, J.; Chen, T.; Luo, Q.; Luo, Y.; Wu, W.; Zhang, Z. Instance Segmentation of Tea Garden Roads Based on an Improved YOLOv8n-seg Model. *Agriculture* **2024**, *14*, 1163. [CrossRef]
10. 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. [CrossRef]
11. Tan, H.; Ma, W.; Tian, Y.; Zhang, Q.; Li, M.; Li, M.; Yang, X. Improved YOLOv8n object detection of fragrant pears. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 178–185.
12. Zheng, W.; Yang, Y. Mature Pear Target Detection Method Based on Frequency Domain Data Enhancement and Lightweight YOLO v7 Model. *J. Agric. Mach.* **2024**, *55*, 244–253.
13. Liu, Q.; Lv, J.; Zhang, C. MAE-YOLOv8-based small object detection of green crisp plum in real complex orchard environments. *Comput. Electron. Agric.* **2024**, *226*, 109458. [CrossRef]
14. Chen, J.; Ji, C.; Zhang, J.; Feng, Q.; Li, Y.; Ma, B. A method for multi-target segmentation of bud-stage apple trees based on improved YOLOv8. *Comput. Electron. Agric.* **2024**, *220*, 108876. [CrossRef]
15. Chen, F.; Chen, C.; Zhu, X.; Shen, D.; Zhang, X. Detection of Camellia oleifera fruit maturity based on improved YOLOv7. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 177–186.
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872. [CrossRef]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
18. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
19. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159. [CrossRef]
20. Yao, Z.; Ai, J.; Li, B.; Zhang, C. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *arXiv* **2021**, arXiv:2104.01318. [CrossRef]
21. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]
22. Zhao, B.; Liu, S.; Zhang, W.; Zhu, L.; Han, Z.; Feng, X.; Wang, R. Research on Performance Optimization of Lightweight Transformer Architecture for Cherry Tomato Picking. *J. Agric. Mach.* **2024**, *55*, 1–13.
23. Hu, J.; Zhang, G.; Shen, M.; Li, W. Detecting surface defects of pine wood using an improved RT-DETR model. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 210–218. [CrossRef]
24. Li, X.; Cai, M.; Tan, X.; Yin, C.; Chen, W.; Liu, Z.; Wen, J.; Han, Y. An efficient transformer network for detecting multi-scale chicken in complex free-range farming environments via improved RT-DETR. *Comput. Electron. Agric.* **2024**, *224*, 109160. [CrossRef]
25. Li, H.; Shi, F. A DETR-like detector-based semi-supervised object detection method for Brassica Chinensis growth monitoring. *Comput. Electron. Agric.* **2024**, *219*, 108788. [CrossRef]
26. Wang, R.; Zhang, B.; Guo, T.; He, T.; Cui, H.; Wang, Z.; Man, S. Physiological response and cold resistance evaluation of 5 pear varieties under low temperature stress. *Shandong Agric. Sci.* **2023**, *55*, 57–63.

27. Qiao, C.; Han, M.; Gao, W.; Gao, W.; Li, K.; Zhu, X.; Zhang, L. Quantitative Detection of Cucumber Downy Mildew Spores at Multi-scale Based on Faster-NAM-YOLO. *J. Agric. Mach.* **2023**, *54*, 288–299.
28. Fu, C.; Ren, L.; Wang, F. Recognizing beef cattle behavior under automatic scene distinction using lightweight FABF-YOLOv8s. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 152–163.
29. Yang, F.; Li, X.; Cheng, H.; Guo, Y.; Chen, L.; Li, J. MSB-FCN: Multi-Scale Bidirectional FCN for Object Skeleton Extraction. *IEEE Trans. Image Process.* **2020**, *30*, 2301–2312. [CrossRef]
30. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.; Chan, S.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
31. Pan, Z.; Cai, J.; Zhuang, B. Fast Vision Transformers with HiLo Attention. *arXiv* **2020**, arXiv:2205.13213. [CrossRef]
32. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography. *Drones* **2024**, *8*, 240. [CrossRef]
33. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A lightweight-design for real-time detector architectures. *J. Real-Time Image Process.* **2022**, *21*, 62. [CrossRef]
34. Zhang, H.; Zhang, S. Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale. *arXiv* **2023**, arXiv:2312.17663. [CrossRef]
35. Wang, J.; Xu, C.; Yang, W.; Yu, L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. *arXiv* **2021**, arXiv:2110.13389. [CrossRef]

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.