Citrus fruit detection based on an improved YOLOv5 under natural orchard conditions

Yu Tang[1†], Wenxuan Huang[1†], Zhiping Tan[1*], Weilin Chen[2], Sheng Wei[3], Jiajun Zhuang[4], Chaojun Hou[4], Jinchang Ren[1,5]

(1. *Academy of Interdisciplinary Studies, Guangdong Polytechnic Normal University, Guangzhou 510665, China*;
2. *School of Mechatronics Engineering and Automation, Foshan University, Foshan 528000, Guangdong, China*;
3. *Engineering Research Center for Intelligent Robotics, Jihua Laboratory, Foshan 528200, Guangdong, China*;
4. *Academy of Contemporary Agriculture Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China*;
5. *National Subsea Center, Robert Gordon University, Aberdeen, AB21 0BH, UK*)

Abstract: Accurate detection of citrus can be easily affected by adjacent branches and overlapped fruits in natural orchard conditions, where some specific information of citrus might be lost due to the resultant complex occlusion. Traditional deep learning models might result in lower detection accuracy and detection speed when facing occluded targets. To solve this problem, an improved deep learning algorithm based on YOLOv5, named IYOLOv5, was proposed for accurate detection of citrus fruits. An innovative Res-CSPDarknet network was firstly employed to both enhance feature extraction performance and minimize feature loss within the backbone network, which aims to reduce the miss detection rate. Subsequently, the BiFPN module was adopted as the new neck net to enhance the function for extracting deep semantic features. A coordinate attention mechanism module was then introduced into the network's detection layer. The performance of the proposed model was evaluated on a home-made citrus dataset containing 2000 optical images. The results show that the proposed IYOLOv5 achieved the highest mean average precision (93.5%) and F1-score (95.6%), compared to the traditional deep learning models including Faster R-CNN, CenterNet, YOLOv3, YOLOv5, and YOLOv7. In particular, the proposed IYOLOv5 obtained a decrease of missed detection rate (at least 13.1%) on the specific task of detecting heavily occluded citrus, compared to other models. Therefore, the proposed method could be potentially used as part of the vision system of a picking robot to identify the citrus fruits accurately.

Keywords: occluded citrus fruits detection, improved YOLOv5, coordinate attention mechanism, object detection
DOI: 10.25165/j.ijabe.20251803.8935

Citation: Tang Y, Huang W X, Tan Z P, Chen W L, Wei S, Zhuang J J, et al. Citrus fruit detection based on an improved YOLOv5 under natural orchard conditions. Int J Agric & Biol Eng, 2025; 18(3): 176–185.

1 Introduction

Citrus reigns as one of the most prolifically cultivated fruits across the globe, with an annual production of 140 million tons[1]. However, the harvesting for citrus fruits is still a labor-intensive procedure in China. Recently, with the advancement of computer vision and deep neural networks, automated intelligent harvesting of citrus fruits has become feasible[2]. However, in natural orchard environments, the accurate detection of citrus fruits is influenced by the factors such as branch and mutual occlusion between fruits. These complex environmental factors could lead to the loss of valuable information for robust target detection using traditional deep learning algorithms, which typically exhibit lower detection accuracy and speed when addressing occluded target detection problems[3].

In recent years, scholars have proposed many target detection models for different types of fruits. Most object detection algorithms comprise both one-stage and two-stage algorithms. For two-stage object detection algorithms, Wan et al.[4] introduced a Faster R-CNN for multi-class fruit variety classification. They fine-tuned the architecture of the model's convolutional and pooling layers to effectively identify apples, mangoes, and oranges within orchard environments. The proposed model achieved 90.72% mAP. Li et al.[5] introduced focal loss in the Region Proposal Network (RPN) of the Faster R-CNN. This adaptation empowers the network to tackle the issue of skewed sample distribution in seedling classification, particularly in cases of complex and easy samples. The average accuracy of automatic detection of hydroponic lettuce seedlings reached 86.2%. He et al.[6] introduced a deep boundary box regression forest detection method for immature fruit, which assisted in the detection of immature fruits by three different features: shape, texture, and color. The average accuracy of the method was 87.6%. Nevertheless, the detection method mentioned above exhibits a slow processing speed, rendering it unsuitable for real-time monitoring endeavors. In scenarios requiring real-time

Received date: 2024-03-19 **Accepted date:** 2025-05-26
Biographies: Yu Tang, Professor, research interest: Precision agriculture, Email: yutang@gpnu.edu.cn; **Wenxuan Huang**, MS, research interest: Object detection, Email: 54261803@qq.com; **Weilin Chen**, Associate professor, research interest:Flexibility mechanism, Email: weilin.chen@fosu.edu.cn; **Sheng Wei**, PhD, research interest: Object detection, Email: weisheng@jihualab.com; **Jiajun Zhuang**, Associate professor, research interest: Deep learning, Email: jiajunzhuang@126.com; **Chaojun Hou**, Associate professor, research interest: Machine vision, Email: houchaojun@zhku.edu.cn; **Jinchang Ren**, Professor, research interest: Machine vision, Email: jinchang.ren@ieee.org.
†The first two authors contributed equally to this manuscript and should be considered co-first authors.
***Corresponding author: Zhiping Tan**, Associate professor, research interest: Object detection, Email: tanzp@gpnu.edu.cn.

fruit picking surveillance, the detection speed must be maintained at a minimum of 10-15 frames per second[7]. Despite the commendable detection accuracy exhibited by the two-stage algorithm, its response speed falls short, thereby posing challenges in meeting the demands of real-time detection scenarios.

In contrast, single-stage object detection algorithms, notably the YOLO series, have gained significant traction in object detection applications owing to their rapid detection speed. Tian et al. introduced an enhanced version of the YOLOv3 model designed for identifying the growth stages of apples. This improved model incorporated a denser network structure, leading to better propagation and reuse of features. The improved YOLOv3 achieved an F1-score of 0.817[8]. For rapid recognition and precise localization of fig fruits within intricate surroundings, Wu et al.[9] introduced a fig detection algorithm that leverages YOLOv4 deep learning technology. Acknowledging the challenges posed by identifying numerous clusters of kiwifruit within intricate field conditions, Rui et al.[10] presented an enhanced YOLOv4 convolutional neural network. This improved model addresses the identification of individual fruits hindered by factors such as leaves, branches, overlapping instances, and obscured views. Their recognition rates for separated fruits were 90.5%, 85.9%, 93.5%, and 96.1%, respectively. To detect litchi and its stalks at night, Liang et al.[11] improved YOLOv3 with U-Net. Used in different lighting conditions, the algorithm achieved 96.1% accuracy and 89.0% recall. However, the method was not tested in daylight. Xia et al.[12] used an attention mechanism to improve YOLOv7 by increasing visual features' weight while suppressing invalid features' weight. The proposed method had 93.1% accuracy on their kiwi dataset. Wu et al.[13] introduced a YOLOv7 network along with several data augmentation techniques to detect camellia fruits within intricate field settings. The fusion of the YOLOv7 network and diverse data augmentation approaches led to a heightened mean average precision (mAP) of 96.0%. Chen et al.[14] introduced a technique based on YOLOv5 to identify varying levels of ripeness in citrus fruits. Wang et al.[15] selected the YOLOv5 algorithm to identify the stem and calyx of an apple in real time. Through the application of exploration of detection heads, layer trimming, and channel reduction techniques, accurate detection of apple varieties' stems and calyxes achieved an impressive 93.89% accuracy rate. Qian et al.[16] pre-trained the YOLOv5 model by converting the RGB image dataset to a pseudo-color dataset of the selected channel (Cr channel), resulting in an 8% improvement in the final experimental results over the mean of the original YOLOv5 mAP. Zhang et al.[17] introduced the grape cluster detection algorithm named YOLOv5-GAP, which integrated a transformer module to enhance the backbone network. Experimental findings demonstrated that the average accuracy of YOLOv5-GAP is increased by 4.34% compared with YOLOv5. Gao et al.[18] introduced a multi-class detection approach utilizing fast regional convolutional neural networks. This method can effectively detect apples in various occlusion scenarios, encompassing scenarios without occlusion, apples obscured by leaves, and those partially hidden by branches and other fruits. The achieved mapping accuracy for these four classes was 0.879, and the average image processing time stood at 0.241 s. Li et al.[19] devised a hierarchical positive sample selection mechanism to enhance the YOLOv5 model's fitting capacity. This innovation resulted in an impressive 12.47% increase in the F1-score for the enhanced YOLOv5 variant.

Chen et al.[20] proposed an improved multi-task deep convolutional neural network detection model, MD-YOLOv7, for the detection of cherry and tomato fruit ripeness. The total score in multi-task learning was 86.6%, and the average reasoning time was 4.9 ms. Nan et al.[21] developed a new WGB-YOLO network and used it for dragon fruit multi-category detection. WGB-YOLO showed a good detection effect in the orchards of dragon fruit intensive planting, and the mAP value was 86.0%.

In summary, the improvements made to YOLO-based models in recent years primarily fall into several core directions: 1) enhancing multi-scale feature fusion to improve detection in complex environments (e.g., BiFPN, PANet); 2) reducing model complexity to enable real-time deployment in resource-limited scenarios (e.g., lightweight backbones and pruning strategies); 3) improving detection accuracy through the integration of attention mechanisms and transformer modules; and 4) incorporating domain-specific enhancements for agricultural tasks such as occlusion handling and color-based segmentation. Despite these advances, challenges such as performance trade-offs, increased model size, or poor generalizability under varying environmental conditions remain. These issues underscore the necessity of further research into models tailored for occlusion-heavy, natural orchard environments.

Although many improved YOLO-based object detection models have been developed for various fruit detection tasks, they often focus on general detection performance under specific conditions. As summarized above, these improvements target aspects such as feature fusion, model efficiency, and attention mechanisms. However, little emphasis has been placed on the problem of object detection where significant information loss is caused by fruit and branch occlusions. Therefore, it is essential to investigate appropriate deep learning-based architecture to accurately and rapidly detect occluded citrus fruits in natural orchard environments. In this study, an improved deep learning algorithm based on YOLOv5, named IYOLOv5, is proposed for the detection of citrus fruits. The main contributions are as follows: 1) an innovative Res-CSPDarknet network is employed to enhance feature extraction and minimize feature loss within the backbone network; 2) the BiFPN module is adopted as the new neck net to extract deep semantic features; and 3) a coordinate attention mechanism module is introduced into the network's detection layer to improve the detection accuracy.

2 Data collection and processing

A variety of citrus fruits named "Tangerine" was investigated in collaboration with Guangzhou Conghua Hualong Fruit & Vegetable Freshness Co. Ltd. (113°39′2.38″E, 23°33′12.48″N). To verify the validity of the newly proposed citrus detection method, this study used digital cameras to take 4239 images in natural citrus orchards over a period of about two weeks in late December 2021, covering different conditions on sunny and cloudy days. All of these images have a resolution of 3000×4000 pixels, and the camera position is about 30-100 cm away from the citrus fruit. The RGB images of citrus fruits are shown in Figure 1.

To ensure data quality and enhance the robustness of the detection model, this study excluded images that were out of focus or contained partially damaged citrus fruits. In the selection process, this study prioritized images with occlusions caused by branches or overlapping fruits, as well as samples captured from diverse angles and distances, to improve the model's generalization ability.

1600 images were randomly selected for the training set, another 200 for validation, and another 200 for the test set. Labelimg software was used to manually draw boundary boxes for

Figure 1 RGB images of citrus fruits

citrus fruits in the dataset. For specific operations, please refer to Figure 2. The bounding boxes delineating the labeled citrus fruits are annotated with red rectangles, serving as the foundation for generating ground truth. Meanwhile, the bounding boxes encompassing labeled occluding objects are annotated with blue rectangles. The area ratio of the blue rectangle to the red rectangle is used to define the degree of occlusion of citrus fruits.

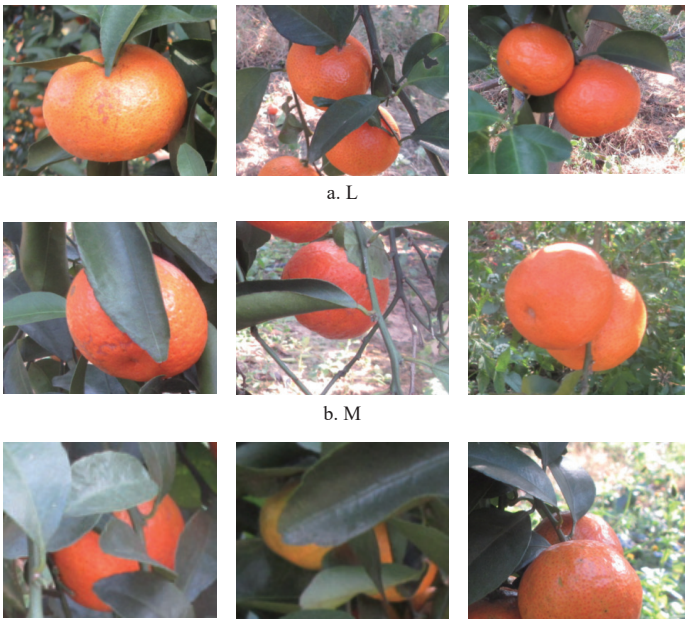
The lightly occluded samples (L) are those with an average occlusion of less than 30%. The moderately occluded samples (M) are those with an average occlusion of greater than 30% and less than 60%. The heavily occluded samples (H) have an average occlusion of greater than 60% and less than 90%. The samples with occlusion degrees more than 90% were not considered in the experiment due to their feature information being almost completely lost. The number of citrus fruits with different occlusion degrees is listed in Table 1. The citrus fruit samples with different degrees of occlusion are shown in Figure 3.

a. b. c.

Figure 2 Citrus fruits data annotation diagram

Table 1 Number of citrus fruits with different occlusion degrees in the dataset

Dataset	Number of images	Lightly occluded (L)	Moderately occluded (M)	Heavily occluded (H)	Total
Train	1600	7308	4002	3789	15 099
Val	200	1089	613	450	2152
Test	200	1112	599	479	2190



a. L

b. M

c. H

Figure 3 Citrus fruits samples with different degrees of occlusion

3 The proposed method

An improved deep learning algorithm based on YOLOv5, named IYOLOv5, is proposed for the detection of citrus fruits in natural orchard environments. The overall structure of the proposed model is illustrated in Figure 4. The improvements introduced in IYOLOv5 consist of three major components: 1) Residual connections are integrated into the backbone network to reduce feature information loss during training. This enhancement facilitates the construction of deeper network structures and helps mitigate overfitting. 2) A BiFPN structure is adopted as the neck to fuse multi-scale features and assign adaptive weights, enabling better feature representation across scales. 3) A coordinate attention mechanism is embedded before the detection layer, which significantly enhances the model's ability to localize occluded citrus fruits more precisely. These design choices are made considering the challenges posed by complex orchard conditions, including fruit occlusion, background clutter, and variable lighting. The details of each component - namely the backbone network, BiFPN module, and detection layer - are presented in Sections 3.1, 3.2, and 3.3, respectively.

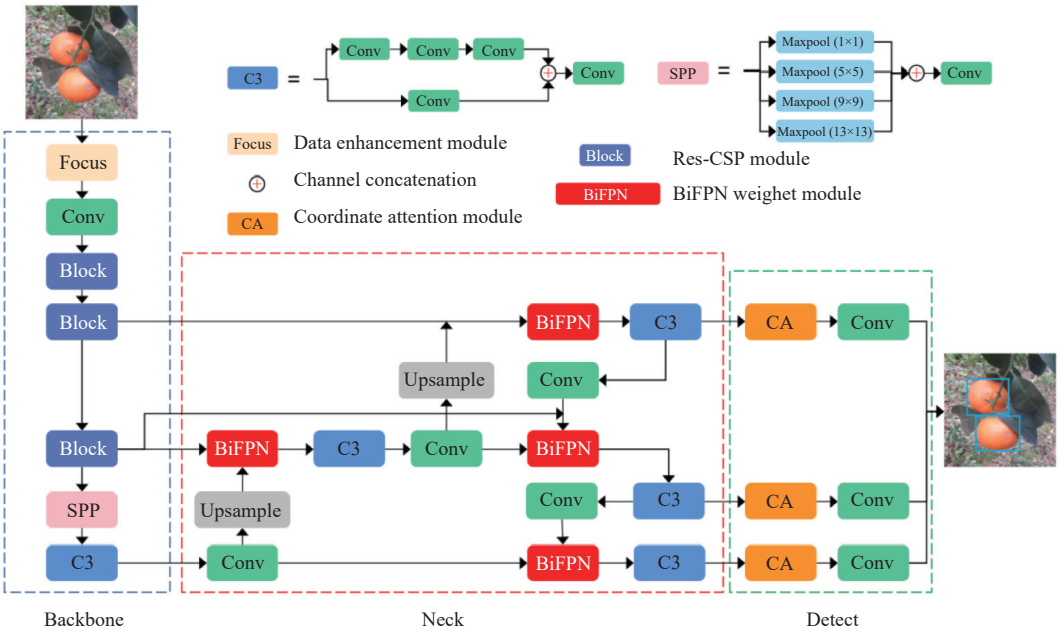


Figure 4 Architecture of improved YOLOv5 network

3.1 Res-CSPDarknet53

A stronger capability for image feature extraction is achieved in the architecture of YOLOv5 compared to previous models in the YOLO series by employing a deep backbone network. However, in the process of information transmission, deep backbone networks are vulnerable to information loss. This situation is caused by the lack of image information, which may lead to the decrease of detection efficiency, which leads to the limitations of deep backbone networks in practical applications. In this study, a backbone network named Res-CSPDarknet53 is proposed, and its network structure is illustrated in Figure 5. Unlike the CSPDarknet53 backbone network used in the original YOLOv5, improvements are made to the existing C3 layer by establishing a connection between the adjacent two layers, and utilizing a 1×1 convolutional layer for the purpose of modifying the channel count. During the training phase, the matrix information post-convolution is input into the Block module. It not only undergoes feature extraction through the C3 module but also undergoes channel adjustment through the convolutional layer. Here, unlike the 3×3 convolutional layer used for extracting image features, the 1×1 convolutional layer maximally retains the matrix's feature information. The information from both paths is then fused into a new matrix and input into the next convolutional layer for integration and works as the input for the subsequent network modules. Compared to CSPDarknet53, Res-CSPDarknet53 utilizes a 1×1 convolutional layer to preserve image feature information and contribute to the subsequent network training, thereby reducing feature loss during the transmission of image features in the backbone network.

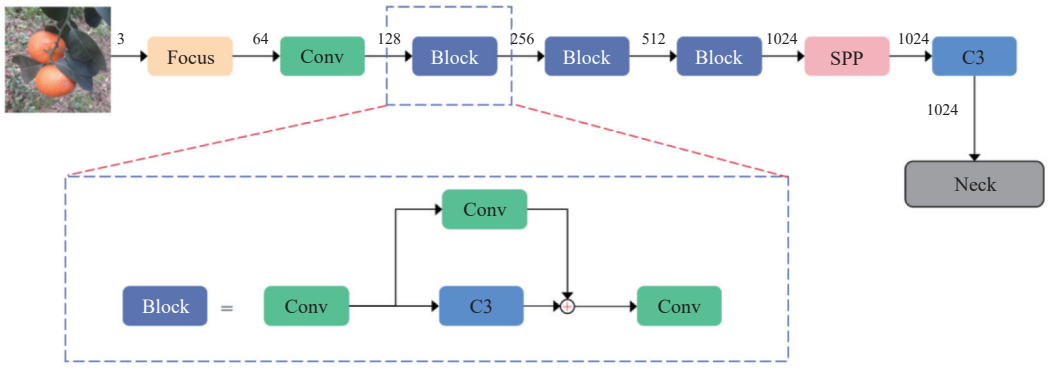


Figure 5 Network structure of Res-CSPDarknet53

This design helps Res-CSPDarknet53 retain fine-grained feature information and facilitates smoother gradient propagation during training. A balance between model complexity and inference speed is maintained, as supported by the ablation experiments presented in Section 4.

3.2 BiFPN

Although precise target localization has been provided by the backbone-transmitted feature map, the low-level feature map contains less semantic information due to the use of fewer convolutional layers. In contrast, the high-level feature map encompasses abundant semantic information following a series of successive extraction convolutions. Semantic information is used to determine the detection target category. The feature fusion methodology in YOLOv5 employs the frameworks of FPN and PANet, where FPN is assigned with conveying profound semantic information from the low layers to the high layers. Based on the FPN, PANet operates in a reverse manner, transmitting surface-level positional details from the upper layers to deeper layers. The combination of the two kinds of information realizes bidirectional feature fusion so that the prediction results have both semantic and location information. However, only simple addition was conducted during the above procedures in YOLOv5, resulting in lower computational efficiency.

To address the above problem, this study introduces the BiFPN structure for effective feature fusion, which combines bidirectional cross-connections with weighted feature fusion. An additional path is introduced in the feature extraction network, that is, an edge is added to the bottom-up path, and the extracted feature is fused with the corresponding node. The portion in PANet that does not receive backbone network feature information has been improved. Meanwhile, BiFPN deletes the nodes with only one input direction in the FPN structure because such nodes contribute less to feature fusion, which is also conducive to simplifying the network. BiFPN provides a weight factor for each feature branch and obtains the optimal weight through autonomous network learning. Finally, the BiFPN is further optimized by reducing the input nodes to fit the output of the effective feature layers of the backbone, as shown in Figure 6.

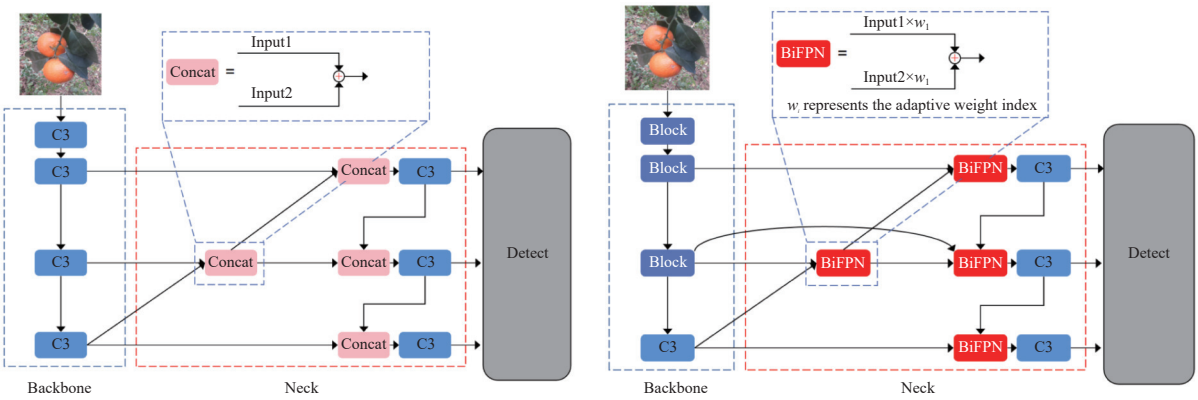


Figure 6 Network structure of PANet and BiFPN in the Neck

This structural enhancement is particularly beneficial for citrus fruit detection under natural orchard conditions, where fruit size varies significantly, and occlusions caused by branches and overlapping fruits are frequent. By adaptively weighting multi-scale features, BiFPN helps the model focus on the most informative scales, improving detection accuracy for both small and partially obscured fruits, which are common in real-world orchard environments.

3.3 Coordinated attention mechanism

In order to elevate the object detection network's detection precision, this study introduces a novel and lightweight attention mechanism termed as coordinate attention[22]. In comparison to previous attention mechanisms such as squeeze-and-excitation (SE)[23] and convolutional block attention module (CBAM)[24], coordinate attention demonstrates superior efficiency and reduced computational burden.

The coordinate attention mechanism is particularly well-suited for fruit detection tasks under natural orchard conditions. Due to frequent occlusion by branches and overlapping fruits, as well as background clutter from foliage and varying lighting, conventional attention mechanisms may struggle to capture precise spatial dependencies. By encoding positional information along both spatial directions while preserving channel dependencies, coordinate attention helps the model concentrate on relevant fruit regions, thereby improving localization accuracy in cluttered and occluded scenarios.

In particular, coordinate attention combines spatial position information with channel weights, enabling the network to obtain both channel weights and spatial position information simultaneously, which helps the target detection network return more accurate location results. In the computation of coordinate attention, the traditional attention mechanism usually adopts the method of global pooling, which compresses the information of the whole space into a single scalar value. For a given input X, the compression step of channel c^{th} can be expressed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{i=1}^{W} x_c(i, j) \tag{1}$$

where, z_c is the output associated with the c^{th} channel and $X=[x_1, x_2, ..., x_C]$ is intermediate feature tensor.

In contrast, the coordinate attention method transforms the global pooling step into a coding operation involving two one-dimensional vectors. In this approach, for the given input X, horizontal features are encoded by pooling kernels (H, 1), and vertical features are encoded by pooling kernels (1, W), resulting in a c-dimensional feature output as described by Equations (2) and (3):

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < w} x_c(h, i) \tag{2}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < h} x_c(j, w) \tag{3}$$

where, z_c^h represents the output of channel c^{th} in the feature diagram

when the vertical height is h; while z_c^w represents the output of channel c^{th} in the feature diagram when the horizontal width is w. The integration of feature information from distinct directions generates a pair of directional feature maps. This alternative method helps improve the accuracy of detection by acquiring remote correlations along one direction while retaining spatial information in the other direction.

The generation of coordinate attention involves concatenating the outputs from Equations (2) and (3), followed by a sequence of transformations, as outlined in Equation (4):

$$f = \delta(F_1([z^h, z^w])), \quad f \in R^{C/r\times(H+W)} \tag{4}$$

where, δ is a nonlinear activation function and F_1 is a transform function. The intermediate feature f, encompassing both horizontal and vertical spatial details, is divided into two separate features: $f^h \in R^{C/r\times H}$ and $f^w \in R^{C/r\times W}$. Additional 1×1 convolutions and sigmoid functions align the dimensions of f^h and f^w with the input X, detailed in Equations (5) and (6):

$$g^h = \delta(F_h(f^h)) \tag{5}$$

$$g^w = \delta(F_w(f^w)) \tag{6}$$

The fusion of g^h and g^w results in the formation of a weighting matrix. F_h and F_w are transform functions employed to calculate the output $y_c(i, j)$, as exemplified in Equation (7):

$$y_c(i, j) = x_{c(i,j)} \times g_c^h(i) \times g_c^w(j) \tag{7}$$

In this investigation, this study integrates the coordinate attention mechanism ahead of the convolutional layer within the detect layer to enhance the object detection network's focus on the regions depicting fruit images. Figure 7 shows the structure in the detect layer.

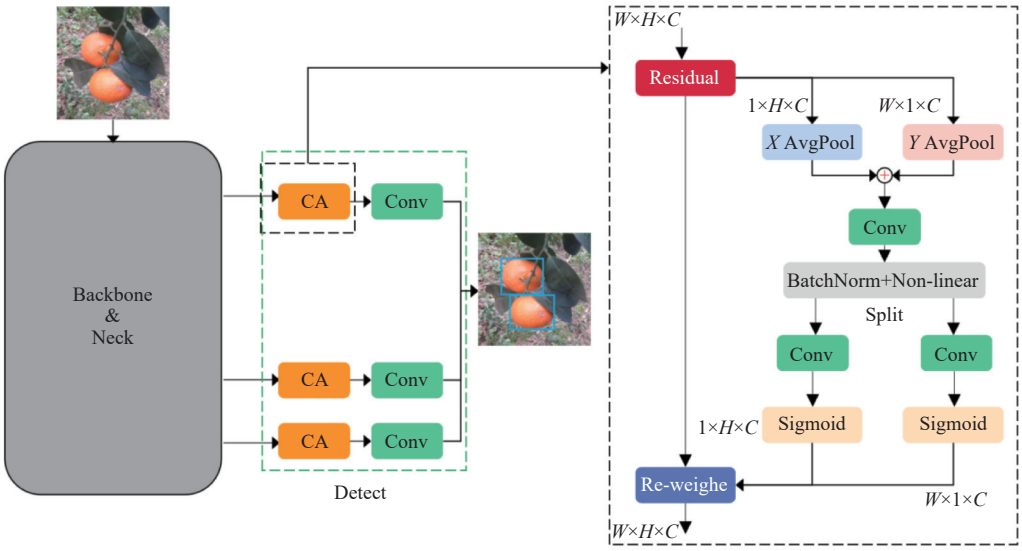


Figure 7 Coordinated attention mechanism in the Head

4 Experimental results and analysis

4.1 Implementation setup

All the models were trained and tested on a workstation featuring dual Intel Xeon Silver 4210R processors, and a total of 256 GB memory. The framework for the experiments was Pytorch1.9.0, and the CUDA11.5 parallel computing framework was used in conjunction with the CUDNN8.3.0 deep neural network acceleration library. The proposed IYOLOv5 network conducted end-to-end federated training using stochastic gradient descent (SGD), where the training process employed a batch size of 32, incorporating batch normalization for regularization during weight updates. The initial attenuation parameter was 0.01, the attenuation rate was 0.9, and the training epochs were 600. Meanwhile, the default values were assigned to other parameters during the training process. Frames Per Second (FPS), depicted in Equation (8), was adopted to assess the pace of model inference:

$$\text{FPS} = \frac{N}{t_N} \tag{8}$$

where, t_N signifies the cumulative time spent by the model for detection across all images, and N denotes the total image count.

The model detection efficacy was evaluated using average precision (AP) and F1-score. AP and F1-score were calculated using precision (P) and recall (R). Equations (9)-(12) define the precision (P), recall (R), F1-score, and AP, respectively:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2\times P\times R}{P+R} \tag{11}$$

$$\text{AP} = \int_0^1 P(R)\text{d}R \tag{12}$$

where, TP signifies the instances correctly identified as true, FP denotes the instances incorrectly identified as true, and FN represents the instances mistakenly identified as false.

4.2 Ablation experiment

Three improved structures including the Res-CSPDarknet53 backbone, BiFPN structure, and coordinate attention mechanism are involved in the proposed IYOLOv5, as illustrated in Section 3. Therefore, ablation experiments are used to determine the performance gain obtained by the structures. Table 2 shows that the embedding of all three improved structures brought obvious improvement on YOLOv5, in which the values of P, R, AP, and F1-score increased by 2.7%, 4.6%, 3.8%, and 3.8%, respectively. Specifically, the introduction of BiFPN obtained the best performance gain compared to YOLOv5, where the values of R,

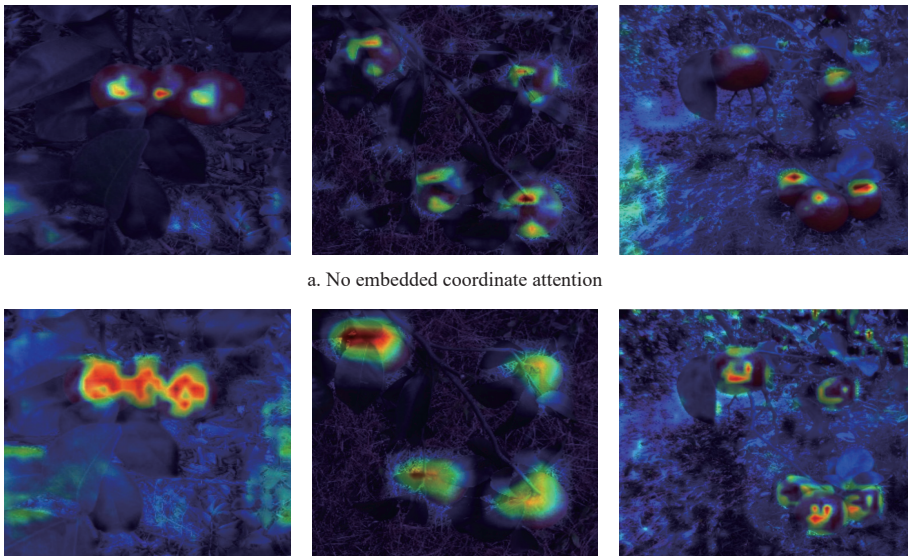
AP, and F1-score increased by 3%, 2%, and 1.5%, respectively. However, the value of *P* decreased by 0.3%, one possible reason for which is that the network incorporated with the BiFPN structure is prone to recall more potential bounding boxes containing citrus fruits, including the image regions containing only the background objects, which thus affects the detection accuracy. On the other hand, the embedding of the coordinate attention mechanism obtained an increase of *P* value by 0.4%. The findings suggest that the incorporation of the attention mechanism notably intensified YOLOv5's focus on the fruit region within the entire image, which was verified by the heat map shown in Figure 8.

Table 2 Results of the ablation experiment

Coordination attention	BiFPN	Res-CSP Darknet53	*P*	*R*	AP	F1-score
-	-	-	0.960	0.881	0.897	0.918
√	-	-	0.964	0.896	0.907	0.928
-	√	-	0.957	0.911	0.917	0.933
-	-	√	0.954	0.903	0.905	0.927
√	√	-	0.961	0.912	0.919	0.935
-	√	√	0.970	0.914	0.922	0.941
√	-	√	0.969	0.915	0.924	0.946
√	√	√	0.987	0.927	0.935	0.956

According to the results of ablation experiments, the performance of YOLOv5 is not significantly improved by Res-CSPDarknet53 alone. However, when Res-CSPDarknet53 is used in combination with other structures, the performance of YOLOv5 is significantly improved. When combined with BiFPN, *P* value, *R* value, AP value, and F1-score increased by 1%, 3.3%, 2.5%, and 2.3%, respectively. After embedding the coordinated attention structure, these indicators increased by 0.9%, 3.4%, 2.7%, and 2.8%, respectively. This indicates that the new Res-CSPDarknet53 backbone network is helpful in reducing the loss of image feature information, which makes the Neck and Head parts of YOLOv5 need to process more information from the backbone network. However, the original structure is not sufficient to efficiently process this additional information, as demonstrated by the joint embedding experiment of the BiFPN structure and coordinated attention structure.

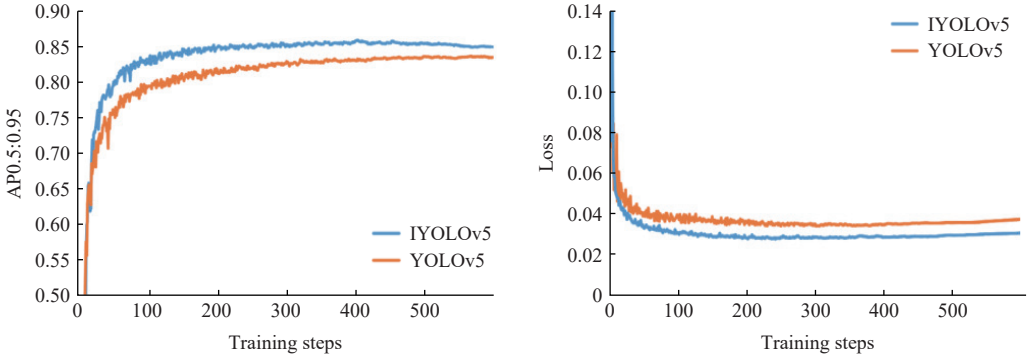
Figure 9 shows the comparison between the training loss curve of IYOLOv5 and YOLOv5 in the training process and the progress curve of mAP0.5:0.95. In the first 200 training steps, YOLOv5's AP0.5:0.95 curve shows the AP it reached compared to IYOLOv5. In addition, the loss function of YOLOv5 showed more fluctuations, indicating that the convergence of IYOLOv5 was improved. At about the 400th step, the two models tend to be stable, the loss curve is consistent, and the model gradually converges. After the model converges, the indices of IYOLOv5 are all higher than those of YOLOv5. These results indicate that IYOLOv5 has faster convergence in the training process.



a. No embedded coordinate attention

b. Embedded coordinate attention

Figure 8 Heat map in the head



a. mAP 0.5:0.95 curves

b. Loss curves

Figure 9 Comparison of training curve between YOLOv5 and IYOLOv5

4.3 Comparison with other algorithms

Experiments using five other target detection models including YOLOv3[25], YOLOv5, YOLOv7[26], CenterNet[27], and Faster R-CNN[28] were conducted to further evaluate the performance of the proposed IYOLOv5. The experimental results are listed in Table 3.

The results show that IYOLOv5 has the best performance in *R* value, AP score, and F1-score, which are 92.7%, 93.5%, and 95.6%, respectively. It can be concluded that the network structure methodology proposed in this paper demonstrates effectiveness in accurately detecting citrus fruits within intricate natural environments. On the other hand, Faster R-CNN had the highest *P*-value (99.6%), followed by improved YOLOv5 (98.7%). One possible reason is that the Faster R-CNN is a two-stage target detection network with a built-in RPN structure that guarantees accuracy. However, the RPN structure is to propose the recommended region in the picture, and then judge the category of the target in the region through the subsequent structure of the network, which is easy to miss the overlapping citrus fruit. The Faster R-CNN's lower *R*-value (62.1%) could validate this claim. In terms of FPS, IYOLOv5 (35.9) is lower than YOLOv5 (46.9), which is better than the other four target detection models. This is due to the fact that IYOLOv5 has embedded the Res-CSPDarknet53 and BiFPN structures to increase the number of parameters in the network, thus increasing the time consumed by the network calculation.

To ascertain the detection capabilities of IYOLOv5 concerning citrus fruits with varying degrees of occlusion, this study collected the above five object detection models and the detection results of IYOLOv5 on citrus fruits with different occlusion degrees. The statistical results are shown in Table 4. The partial detection results are shown in Figure 10, with each detected object presented in the form of a bounding box determined by the smallest closed rectangle containing the visible portion of the citrus fruit. For citrus fruits with lightly occluded and clear features, all six networks can accurately detect citrus fruits. According to the statistical results, IYOLOv5 detected 32 more targets than YOLOv7, which had the best performance in the comparison algorithm, and the missed detection rate decreased by 2%. IYOLOv5 has the best performance in boundary box positioning accuracy, with a confidence of more than 90%. This enhancement may be due to the spatial coordinate weighted information brought by the coordinate attention mechanism, which makes the bounding box more accurate.

Table 3 Results of comparison with other algorithms

Method	*P*	*R*	AP	F1-score	FPS
Faster R-CNN	**0.996**	0.621	0.842	0.765	3.0
CenterNet	0.978	0.701	0.814	0.816	32.5
YOLOv3	0.968	0.809	0.905	0.881	6.9
YOLOv5	0.960	0.881	0.897	0.918	**46.9**
YOLOv7	0.958	0.882	0.918	0.914	11.2
IYOLOv5	0.987	**0.927**	**0.935**	**0.956**	35.9

Table 4 Detection results of different methods for citrus fruits with different levels of occlusion

Number of citrus fruits	Test set	Faster R-CNN	CenterNet	YOLOv3	YOLOv5	YOLOv7	IYOLOv5
L	1112	976	955	1001	1008	1050	**1082**
M	599	398	375	431	482	499	**550**
H	479	221	192	243	230	312	**375**
Total	2190	1595	1522	1675	1720	1819	**2007**

YOLOv3 YOLOv5 YOLOv7

Faster R-CNN CenterNet IYOLOv5

Figure 10 Detection of citrus fruits with occluded branches and leaves

For citrus fruits with occluded branches and leaves, the detection results are shown by the green arrows in Figure 10. Citrus fruits show a small number of visible features, which puts forward higher requirements for feature extraction capability of target detection networks. Experimental results show that IYOLOv5 can successfully detect more citrus fruits obscured by branches and leaves than other target detection models.

For multiple citrus fruits with similar backgrounds that are

occluded from each other, some targets are heavily occluded, the detection results of which are as shown by the blue arrows in Figure 11. The effective features of citrus fruits are few and overlap each other, which makes it difficult for the object detection network to distinguish the target and increases the missed detection rate. Other target detection models are not effective in detecting citrus fruits that are occluded from each other. In contrast, the IYOLOv5 successfully detected heavily occluded citrus fruits, even when they were mutually occluded and surrounded by complex background interference.

Figure 11 Detection of mutually occluded citrus fruits

According to the statistical results in Table 4, IYOLOv5 detected the most heavily occluded citrus fruits, detecting 63 more objects than YOLOv7, and the missed detection rate decreased by 13.1%. Therefore, the detection effect of IYOLOv5 on heavily occluded citrus in natural environment is superior to other target detection models. This may be due to the fact that Res-CSPDarknet53 backbone can effectively reduce the loss of image feature information, so that IYOLOv5 has enough feature information to distinguish citrus fruits in complex environments.

Generally, the IYOLOv5 model proposed in this study demonstrates strong generalization and robustness, enabling accurate citrus fruit detection across various levels of occlusion. Especially in the background of similar color texture, the detection of heavily occluded and overlapping target objects is good.

5 Conclusions

In this study, enhancements were made to the architecture of YOLOv5, leading to the proposal of an IYOLOv5 model tailored for precise detection of citrus fruits within orchard environments. Based on the findings from this study, the subsequent specific conclusions can be delineated:

1) The IYOLOv5 proposed by this study includes three major improvements: (1) a new backbone network Res-CSPDarknet network is used; (2) BiFPN is adopted as a new neck network; and (3) the coordinate attention mechanism is embedded. The conducted ablation experiment corroborated a notable enhancement in the performance metrics of IYOLOv5. Specifically, there was an increase of 2.7% in accuracy, 4.6% in recall rate, 3.8% in AP, and a commensurate rise of 3.8% in the F1-score when juxtaposed with the original YOLOv5 model.

2) IYOLOv5 outperformed five other commonly used networks in comparative experiments, including YOLOv7, YOLOv5, YOLOv3, Faster R-CNN, and CenterNet. The experimental findings underscore the notable advantages of the IYOLOv5 algorithm in target detection accuracy, showcasing the attainment of the highest average detection accuracy (93.5%). In particular, for heavily obscured citrus fruits, IYOLOv5 showed at least a 13.1% reduction in missed detection compared to other models.

Therefore, the proposed IYOLOv5 model is highly suitable for detecting citrus targets in natural orchard environments. In future investigations, the team aims to explore the detection of citrus targets in natural orchard settings across different scales, lighting conditions, and levels of occlusion.

Nevertheless, the proposed model still has some limitations. Its detection performance may degrade under extreme lighting conditions such as backlight or nighttime environments. In addition, while IYOLOv5 performs well on citrus fruit detection, its generalizability to other fruit types has not yet been validated. Future work will focus on improving the model's adaptability across a wider range of orchard scenarios, incorporating more advanced attention mechanisms, and optimizing the model architecture for lightweight deployment on edge or mobile devices.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Guangdong Province, China (Grant No. 2020B1515120070, Grant No. 2022A1515010885), the Innovation Team Project of Universities in Guangdong Province, China (Grant No. 2021KCXTD010), the Key Construction Discipline Research Capacity Enhancement Project of Guangdong Province, China

(Grant No. 2022ZDJS014), the Key Construction Discipline Research Capacity Enhancement Project of GPNU, China (Grant No. 22GPNUZDJS11), and the Characteristic Innovation Project of Universities in Guangdong Province, China (Grant No. 2023KTSCX066).

[References]

[1] Liu Y, Ma X Y, Shu L, Hancke G P, Abu-Mahfouz A M. From industry 4.0 to agriculture 4. 0: Current status, enabling technologies, and research challenges. IEEE Transactions on Industrial Informatics, 2020; 17(6): 4322–4334.

[2] Onishi Y, Yoshida T, Kurita H, Fukao T, Arihara H, Iwai A. An automated fruit harvesting robot by using deep learning. Robomech Journal, 2019; 6: 13.

[3] Tang Y, Dananjayan S, Hou C J, Guo Q W, Luo S M, He Y. A survey on the 5G network and its impact on agriculture: Challenges and opportunities. Computers and Electronics in Agriculture, 2021; 180: 105895.

[4] Wan S H, Goudos S. Faster R-CNN for multi-class fruit detection using a robotic vision system. Computer Networks, 2020; 168: 107036.

[5] Li Z B, Li Y, Yang Y B, Guo R H, Yang J Q, Yue J, et al. A high-precision detection method of hydroponic lettuce seedlings status based on improved Faster RCNN. Computers and Electronics in Agriculture, 2021; 182: 106054.

[6] He Z L, Xiong J T, Chen S M, Li Z X, Chen S F, Zhong Z, et al. A method of green citrus detection based on a deep bounding box regression forest. Biosystems Engineering, 2020; 193: 206–215.

[7] Tu S Q, Pang J, Liu H F, Zhuang N, Chen Y, Zheng C, et al. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images. Precision Agriculture, 2020; 21: 1072–1091.

[8] Tian Y N, Yang G D, Wang Z, Wang H, Li E, Liang Z Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computers and Electronics in Agriculture, 2019; 157: 417–426.

[9] Wu Y J, Yang Y, Wnag X F, Cui J, Li X Y. Fig fruit recognition method based on YOLO v4 deep learning. In: 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand: IEEE, 2021; pp.303–306. doi: 10.1109/ECTI-CON51831.2021.9454904.

[10] Suo R, Gao F F, Zhou Z X, Fu L S, Song Z Z, Dhupia J, et al. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Computers and Electronics in Agriculture, 2021; 182: 106052.

[11] Liang C X, Xiong J T, Zheng Z H, Zhong Z, Li Z H, Chen S M, et al. A visual detection method for nighttime litchi fruits and fruiting stems. Computers and Electronics in Agriculture, 2020; 169: 105192.

[12] Xia Y, Nguyen M, Yan W Q. A real-time kiwifruit detection based on improved YOLOv7. In: Yan W Q, Nguyen M, Stommel M, editors. Image and Vision Computing. Cham: Springer. 2022; pp.48–61.

[13] Wu D L, Jiang S, Zhao E L, Liu Y L, Zhu H C, Wang W W, et al. Detection of Camellia oleifera fruit in complex scenes by using YOLOv7 and data augmentation. Applied Sciences, 2022; 12(22): 11318.

[14] Chen S M, Xiong J T, Jiao J M, Xie Z M, Huo Z W, Hu W X. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. Precision Agriculture, 2022; 23: 1515–1531.

[15] Wang Z P, Jin L Y, Wang S, Xu H R. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. Postharvest Biology and Technology, 2022; 185: 111808.

[16] Wang N, Qian T T, Yang J, Li L Y, Zhang Y Y, Zheng X G, et al. An enhanced YOLOv5 model for greenhouse cucumber fruit recognition based on color space features. Agriculture, 2022; 12(10): 1556.

[17] Zhang T, Wu F Y, Wang M, Chen Z Y, Li L Y, Zou X J. Grape-bunch identification and location of picking points on occluded fruit axis based on YOLOv5-GAP. Horticulturae, 2023; 9(4): 498.

[18] Gao F F, Fu L S, Zhang X, Majeed Y, Li R, Karkee M, et al. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. Computers and Electronics in Agriculture, 2020; 176: 105634.

[19] Li Z S, Xie W Q, Zhang L Z, Lu S, Xie L, Su H Y, et al. Toward efficient safety helmet detection based on YoloV5 with hierarchical positive sample selection and box density filtering. IEEE transactions on instrumentation and measurement, 2022; 71: 1–14.

[20] Chen W B, Liu M C, Zhao C J, Li X X, Wang Y Q. MTD-YOLO: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. Computers and Electronics in Agriculture, 2024; 216: 108533.

[21] Nan Y L, Zhang H C, Zeng Y, Zheng J Q, Ge Y F. Intelligent detection of Multi-Class pitaya fruits in target picking row based on WGB-YOLO network. Computers and Electronics in Agriculture, 2023; 208: 107780.

[22] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE. 2021; pp. 13713–13722.

[23] Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019; 42(8): 2011–2023.

[24] Woo S, Park J, Lee J-Y, Kweon I S. Cbam: Convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. editors. Computer Vision - ECCV 2018. Springer. 2018; doi: 10.1007/978-3-030-01234-2_1.

[25] Redmon J, Farhadi A. Yolov3: An incremental improvement. arxiv preprint arxiv: 1804.02767, 2018; In press. doi: 10.48550/arXiv.1804.02767

[26] Wang C Y, Bochkovskiy A, Liao H Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, 2023; pp.7464–7475.

[27] Zhou X, Wang D, Krähenbühl P. Objects as points. arxiv preprint arxiv: 1904.07850. 2019; In press.

[28] Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 39(6): 1137–1149.