Journal of Real-Time Image Processing (2024) 21:88
https://doi.org/10.1007/s11554-024-01473-1

RESEARCH

YOLOv5s-BC: an improved YOLOv5s-based method for real-time apple detection

Jingfan Liu[1] · **Zhaobing Liu**[1]

Received: 10 November 2023 / Accepted: 30 April 2024 / Published online: 10 May 2024

Abstract
The current apple detection algorithms fail to accurately differentiate obscured apples from pickable ones, thus leading to low accuracy in apple harvesting and a high rate of instances where apples are either mispicked or missed altogether. To address the issues associated with the existing algorithms, this study proposes an improved YOLOv5s-based method, named YOLOv5s-BC, for real-time apple detection, in which a series of modifications have been introduced. First, a coordinate attention block has been incorporated into the backbone module to construct a new backbone network. Second, the original concatenation operation has been replaced with a bi-directional feature pyramid network in the neck network. Finally, a new detection head has been added to the head module, enabling the detection of smaller and more distant targets within the field of view of the robot. The proposed YOLOv5s-BC model was compared to several target detection algorithms, including YOLOv5s, YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet50), and Faster R-CNN (VGG), with significant improvements of 4.6%, 3.6%, 20.48%, 23.22%, 15.27%, and 15.59% in mAP, respectively. The detection accuracy of the proposed model is also greatly enhanced over the original YOLOv5s model. The model boasts an average detection speed of 0.018 s per image, and the weight size is only 16.7 Mb with 4.7 Mb smaller than that of YOLOv8s, meeting the real-time requirements for the picking robot. Furthermore, according to the heat map, our proposed model can focus more on and learn the high-level features of the target apples, and recognize the smaller target apples better than the original YOLOv5s model. Then, in other apple orchard tests, the model can detect the pickable apples in real time and correctly, illustrating a decent generalization ability. It is noted that our model can provide technical support for the apple harvesting robot in terms of real-time target detection and harvesting sequence planning.

Keywords Apple detection · YOLOv5s · Deep learning · Robot · Real-time detection

1 Introduction

Apple, as one of the top four fruits in the world, is rich in many vitamins and minerals and has been popular with consumers around the world. According to the statistics of the Food and Agriculture Organization of the United Nations, apples rank second after grapes in global fruit production [7]. However, most apple fruits are hand-picked, and such production methods are very inefficient. In addition, with an aging population and a large influx of rural labor into the cities, labor costs in the fruit cultivation industry have risen accordingly. All these factors significantly impact the market competitiveness of fruit products. Therefore, it is imperative to harvest apple and other fruits efficiently in real time and reduce harvesting costs. Fruit harvesting robot based on machine vision can use its information perception to identify and pick fruits. Thus, it can improve efficiency and increase economic benefits, which has become a research hotspot for intelligent agricultural equipment [12]. However, there are still few products of fruit-harvesting robots applied in agriculture, and most of them are relatively low in intelligence and even less in large-scale applications [3, 30]). In view of the above situation, it is of great practical significance to study the technology related to fruit-harvesting robots.

Within the laboratories, different fruit-harvesting robots are studied. Although these fruit-harvesting robots have unique features suitable for specific application scenarios, they all rely on the same core technologies, such as stable

✉ Zhaobing Liu
zhaobingliu@whut.edu.cn

[1] Hubei Digital Manufacturing Key Laboratory, School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China

mobile platforms, multi-sensor collaboration, advanced machine vision technology, and flexible motion control. Among them, machine vision has been drawing great attention due to the rapid development of artificial intelligence. Over the years, researchers have combined vision technology to recognize and locate fruits to provide technical support for fruit-harvesting robots. A comprehensive survey revealed that in the field of machine vision, target detection algorithms have tremendous potential for growth by virtue of their high detection accuracy and easy deployment [28]. Note that the mainstream two-stage target detection algorithms include Faster R-CNN [17] and Mask R-CNN [5], while one-stage target detection algorithms are SSD [10] and YOLO (You Only Look Once) series including YOLOv3 [15], YOLOv4 [2], YOLOv5 [20], etc.. Noticeably, the two-stage target detection algorithms generally have higher detection accuracy, but the trained model is large, leading to slow detection speed during practical detection. In contrast, the one-stage target detection algorithm is increasingly used as the preferred solution due to the advantages of the small number of model parameters and rapid detection speed.

The following section will focus on discussion of YOLO, applied to agriculture in the last three years. The YOLO series was pioneered by Redmon and his colleagues and developed on the darknet in versions YOLOv1, YOLOv2, and YOLOv3 [14–16]. Numerous iterations have emerged since then, and Bochkovskiy et al. [2] have continued to build on the darknet and come up with YOLOv4. Unlike previous versions, Ultralytics developed YOLOv5 with the Pytorch framework. YOLOv5 is favored by researchers for its ease of deployment and well detection performance. YOLOv5 has four basic network models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Their feature map depths are progressively deeper, and the model parameters are increased sequentially. Table 1 summarizes the performance of the improved YOLOv5 model in the agricultural domain. In terms of apple detection, Yan et al. [25] proposed a light target detection method for apple-picking robots based on an improved YOLOv5s algorithm. The bottleneck Cross Stage Partial (CSP) module is redesigned as a bottleneck CSP-2 module. In addition, the squeeze and excitation module in the visual attention mechanism network is inserted into the improved backbone network. The average detection accuracy is 86.75%. Lv et al. [12] proposed a visual recognition method for detecting apple growth patterns in orchards using the YOLOv5s algorithm. The authors replaced the SiLU activation function in the network with the ACON-C activation function, which improved the accuracy of the algorithm without sacrificing real-time performance. Sun et al. [18] proposed an improved lightweight apple detection method YOLOv5-PRE for fast apple yield detection in an orchard environment, and introduced ShuffleNet and GhostNet lightweight structures in the YOLOv5-PRE model to reduce the size of the model. Xu et al. [23] proposed an improved YOLOv5 apple grading method. The Mish activation function replaced the original YOLOv5 activation function, and the squeeze excitation module was added to the YOLOv5 backbone. The average accuracy of the improved YOLOv5 algorithm for grading apples under the test set is 90.6%. For the detection of other fruits and vegetables, Yao et al. [26] developed a YOLOv5-based Kiwifruit defect detection model, called YOLOv5-Ours. The proposed model added a small target detection layer by embedding SELayer attention to different channels. The average detection accuracy of YOLOv5-Ours reached 94.7%. Wu et al. [22] constructed a new YOLOv5-B model by enhancing the loss function. Then, the optimal truncation point is obtained by segmenting the contours

Table 1 Performance of the improved YOLOv5 models

Detection object	Networks model	F1 (%)	mAP (%)	Detection speed (FPS)	GPU	References
Apple	Improved YOLOv5s	87.49	86.75	66.7	Nvidia Geforce RTX 2060	Yan et al. [25]
Apple	YOLOv5-B	92.8	98.4	71	Nvidia Geforce GTX 1080	Lv et al. [12]
Apple	YOLOv5-PRE	88.88	94.03	37.04	Nvidia Quadro P620	Sun et al. [18]
Apple	Im-YOLOv5	90.74	90.6	59.63	Nvidia Geforce GTX 1660Ti	Xu et al. [23]
Kiwifruit	YOLOv5-Ours	–	94.7	10	Nvidia GeForce GTX 1050Ti	Yao et al. [26]
Banana	YOLOv5-B	94.44	93.2	111.1	Nvidia Tesla V100 SXM2	Wu et al. [22]
Zanthoxylum	Improved YOLOv5s	–	94.5	88.33	Nvidia GeForce RTX 3060 Laptop	Xu et al. [24]
Shoots of litchi	YOLOv5-SBiC	–	79.56	55.6	Nvidia GeForce RTX 3090	Liang et al. [9]
Fusarium head blight in wheat	YOLOv5-DIOU	87.95	91.18	–	Nvidia GeForce RTX 3060	Zhang et al. [27]
Tomato virus disease	SE-YOLOv5	89.39	94.1	50.63	Nvidia GeForce RTX 2060 Super	Qi et al. [13]
Tea leaf blight	DDMA-YOLO	71.6	76.8	–	Nvidia GeForce RTX 2060	Bao et al. [1]
Passion fruit pests	Improved YOLOv5	95.54	96.51	129.87	–	Li et al. [8]

of the axes utilizing an edge detection algorithm. Experiments show that the average detection accuracy of the proposed model for banana multi-target recognition is 93.2%. Xu et al. [24] proposed an improved YOLOv5s-based target detection method for Zanthoxylum-picking robots. An improved CBF module is proposed based on the backbone CBH module, and a Specter module is proposed to replace the bottleneck CSP module. Test experiments conducted on NVIDIA Jetson TX2 show that the average inference time is 0.072 s. Liang et al. [9] developed a YOLOv5-SBiC algorithm for late-autumn bud recognition. In the algorithm, a transformer module was introduced to speed up the network convergence. Besides, an attention mechanism module was used to help the model extract more useful information. Test results show that the proposed algorithm improves the recognition accuracy by 4.0% over the original YOLOv5 algorithm, reaching 79.6%. In the field of pests and diseases detection of fruits and vegetables, Zhang et al. [27] proposed a new method based on a target detection network, feature extraction, and classifier to detect adjacent wheat ears. The proposed algorithm combines distance-interlinked non-maximum suppression based on the original YOLOv5 to form an improved YOLOv5 target detection network with an average detection accuracy of 90.67% and a detection time of 0.73 ms. Qi et al. [13] implemented the extraction of key features by inserting a squeeze stimulus module into the original YOLOv5 network framework, drawing on the human visual attention mechanism. The model was evaluated on the tomato virus disease test set, and the average detection accuracy was 94.10%. Bao et al. [1] proposed a DDMA-YOLO-based UAV remote sensing method to detect and monitor tea leaf blight. The algorithm added a multi-scale RFB module based on the original YOLOv5 network with dual-dimensional mixed attention (DDMA) in the neck, and the average detection accuracy was 76.8%. Li et al. [8] proposed a fast and lightweight improved YOLOv5 detection algorithm. Based on the original YOLOv5 model, a new point-line distance loss function was presented. In addition, an attention module was added to the network for adaptive attention, which can attend to the target object passion fruit pests in both channel and spatial dimensions. The average detection accuracy was 96.51%.

Considering the above-mentioned discussions, real-time apple detection methods with lightweight models need to be further developed. Through a comprehensive survey of the improved YOLO target detection methods in the agricultural field, although most of the existing detection models have relatively high recognition accuracy, their increased complexity, parameters, and hardware requirements usually lead to low real-time performance. Therefore, designing a lightweight real-time apple detection algorithm is necessary to meet the requirements of real-time recognition of picking robots while ensuring recognition accuracy. In this paper, we propose an improved YOLOv5s-based real-time apple detection method to overcome the limitations of current apple recognition techniques. The major contributions of this paper are as follows:

(i) The CA attention mechanism module has been incorporated into the backbone network and the neck network. In the backbone network, the CA attention mechanism can help the model automatically screen and focus on key feature channels, reduce unnecessary information redundancy, optimize model parameters, and reduce computational costs, thereby improving the efficiency and speed of the model. In the neck network, the CA attention mechanism can weight different feature channels during the feature fusion process, allowing the model to better integrate multi-scale and multi-level information, and enhance the diversity of features and the robustness of the model.

(ii) The BiFPN block has been designed in the neck network, which first receives feature maps of different scales from across the region in the backbone, and then performs concat operation on these feature maps, which is named Bi-concat. BiFPN combines the mechanism of bi-directional feature propagation, can effectively fuse the features of different scales, and thus improves the ability of the model to characterize the object at different scales and levels. Besides, the BiFPN block makes the information transfer of the feature pyramid more balanced and effective through multiple iterations of feature fusion and updating, contributing to improved accuracy and stability of the object detection model. At the same time, it assists the model to better understand the location and size of the object in the image, thus improving the accuracy of object localization.

(iii) The new detection head has been added in the head network. As the resolution of the feature maps used for small object detection increases, the local receptive field of the feature maps shrinks accordingly, which allows the network to detect more small objects with lower resolution. The addition of this detection head enables the use of high-resolution feature maps to detect smaller objects that are farther away, thereby improving the accuracy of object detection and localization.

2 Data acquisition and preprocessing

2.1 Apple images acquisition

In this research, the dataset was from the Agricultural Automation and Robotics Laboratory at Washington State

University that was originally utilized to estimate yields in robotic harvesting [11]. To obtain the dataset, the lab researchers installed the image sensor behind the prismatic gantry of the robot. The distance between the sensor and the tree was nearly 1.5 m. Figure 1 displays the apple images in the dataset from early morning to dusk. In this work, we took 1750 apple RGB images from the original dataset as the new dataset and initially divided this dataset according to the ratio of 0.85:0.15, where 1487 images were in the training set and 263 images were in the test set. There was no overlapping between the two sets.

2.2 Image labelling

The labelling software (Labelimg) is utilized to classify and label apple images that are visible to the human eyes after acquisition, as shown in Fig. 2. Due to the intricate environment in apple orchards, the apple images are separated into two categories: graspable and ungraspable, with the corresponding labels 'apple' and 'block'. In detail, apples are categorized based on the following criteria:

(i) Classification 1: Apples that are obstructed by leaves and branches are categorized as ungraspable apples.
(ii) Classification 2: Small target apples that can be observed despite being far away are categorized as graspable apples, which provides valuable data for training models for small targets.
(iii) Classification 3: Large target apples that are close enough and can be observed directly are categorized as graspable apples.
(iv) Classification 4: Apples that are recognized in both bright daylight and insufficient light are categorized as graspable. This adds complexity to the data and enhances the robustness of the model.

2.3 Image augmentation

The quality of the training set plays a pivotal role in determining the capability of the convolutional neural network (CNN) model to identify apples accurately. If the training set is too small, it can lead to overfitting of the model, which may impede its performance in new or unknown environments. Image augmentation involves enhancing the visual quality of an image and augmenting its specific features by applying a series of processes. This methodology can effectively enlarge the size and diversity of the training set and improve the generalization ability of the CNN model. Specific image enhancement methods were selected based on application scenarios and data characteristics. We have selected eight data augmentation methods based on our own scenario requirements. These methods included random contrast, edge enhancement, contrast-limited adaptive histogram equalization (Clahe), motion

Fig. 1 Apple images from various points in time [11]

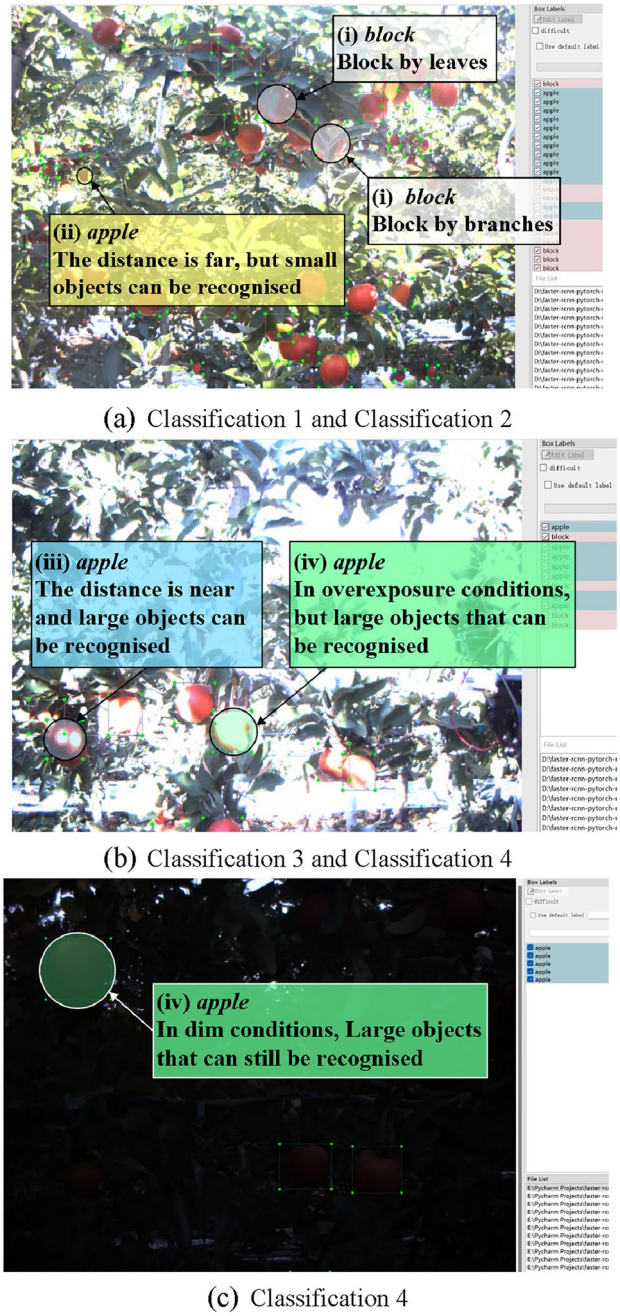

Fig. 2 Apple image labelling in the software

blur, perspective transformation, adding salt and pepper noise, max pool, and changing color temperature [25]. A total of 11,896 enhanced images were generated by these methods from the initial training set of 1487 images, so the new training set consists of 13,383 images. Figure 3 illustrates the eight different image augmentation methods that have been used on each image.

3 Methods

3.1 YOLOv5s

The YOLOv5 algorithm is a one-stage target detection algorithm that generates class probabilities and position coordinate values for objects without requiring region proposals. It is one of the most popular target detection algorithms among agricultural researchers and its network structure can be divided into four modules: input, backbone, neck, and head. The input module uses mosaic data augmentation, adaptive anchor frame calculation, and adaptive image scaling operations. The backbone network consists of focus and Cross Stage Partial (CSP) structures. The neck network utilizes a Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structure. The CIoU loss function is used as the loss function of the bounding box in the head module. Among them, CSP draws on CSPNet to address the problem of excessive computation in inference from the perspective of network structure design. Note that the problem of excessive inference computation is mainly due to the duplication of gradient information in network optimization [21]. Taking the YOLOv5s network as an example, the first CSP structure, namely C3-1, is applied in the backbone, and the other CSP structure, namely C3-2, is applied in the neck, which enhances the ability of network feature fusion. Their structural compositions are shown in Fig. 4. The step size of the convolution kernel in front of each CSP structure is two, so that it can play the role of undersampling. In addition, YOLOv5 uses the Spatial Pyramid Pooling Fast (SPPF) module instead of the SPP module, which uses a cascade of multiple small-size pooling kernels instead of a single large-size pooling kernel in the SPP module. Thus, it further improves the running speed while fusing feature maps of different sensory fields to enrich the expression capability of feature maps.

Readers can refer to the official code (https://github.com/ultralytics/YOLOv5) for more details. Because the YOLOv5s model has the fewest parameters among the four models officially provided by YOLOv5, it is in line with the trend of lightweight and easier to deploy on fruit-harvesting robots, thereby satisfying the effect of real-time grasping. Therefore, in this study, we choose it as the research candidate. Figure 5 depicts the network architecture of the enhanced algorithm in detail.

3.2 CA block

The attention mechanism is essential in identifying targets as it enables the model to concentrate on crucial parts

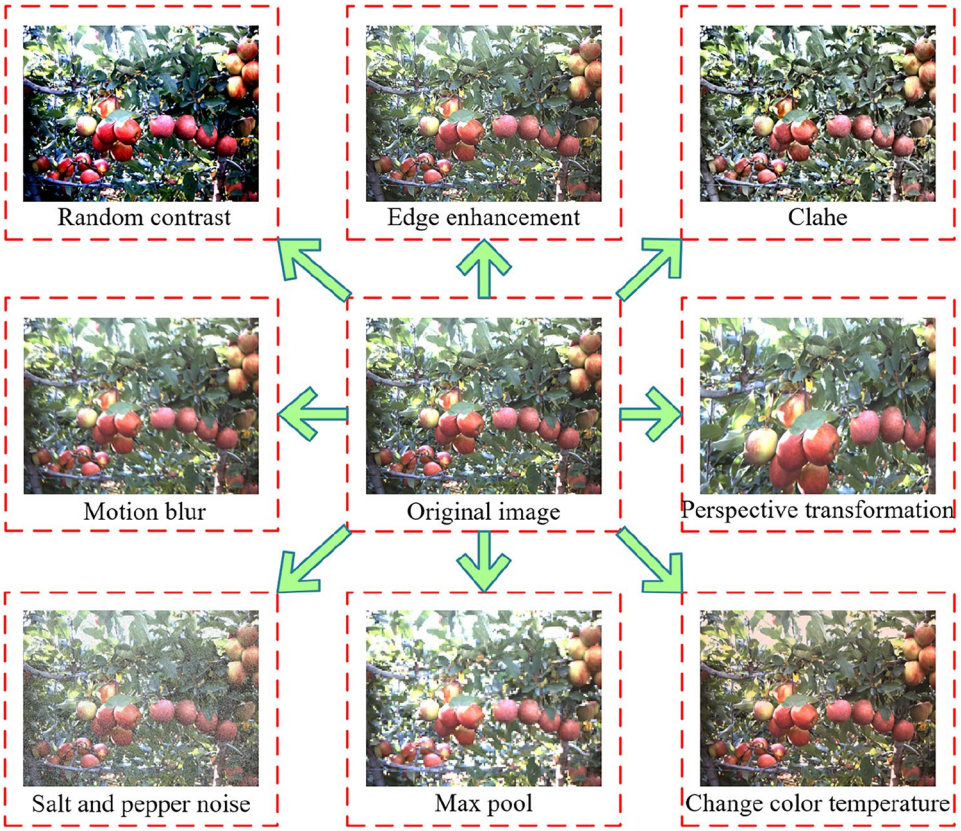

Fig. 3 Enhanced images using different image enhancement methods

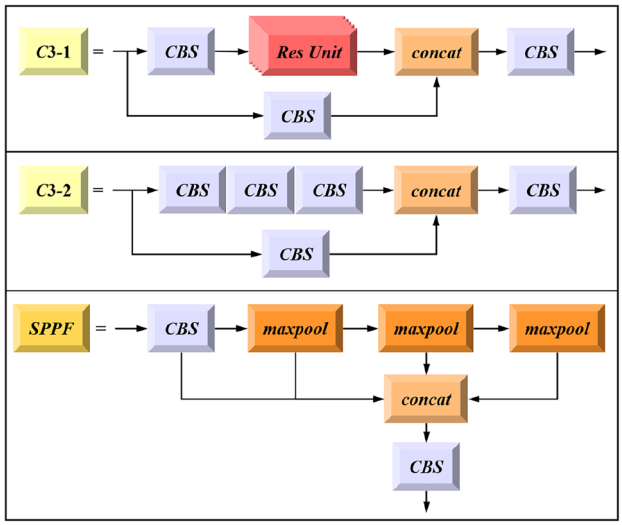

Fig. 4 YOLOv5s partial component structure

of images, thereby improving accuracy and efficiency in detection. Hou et al. [6] proposed a CA mechanism that integrates position information into channel attention. In more detail, CA decomposes channel attention into two one-dimensional feature encoding processes that aggregate features along two spatial directions, respectively. This allows for capturing remote dependencies in one spatial direction while maintaining accurate location information in the other spatial direction. The resulting feature maps are then encoded as a pair of direction-aware and position-sensitive attention maps, respectively, which can be applied complementarily to the input feature maps to enhance the representation of the object of interest. In addition, CA has the property of portability and can be flexibly embedded into CNN. Considering collectively, we choose it as our attention mechanism component to be introduced into the YOLOv5 network in this experiment. The specific operation of CA is divided into two steps: coordinate information embedding and CA generation. Figure 6 shows the structure of the CA block.

3.2.1 Coordinate information embedding

The CA block is designed to obtain attention to the width and height of the image and encode the exact position information. First, the input feature map X is divided into two directions, height h, and width w, and is pooled globally to obtain the feature maps in both directions. Thus, the output of the c th channel with height h can be expressed as

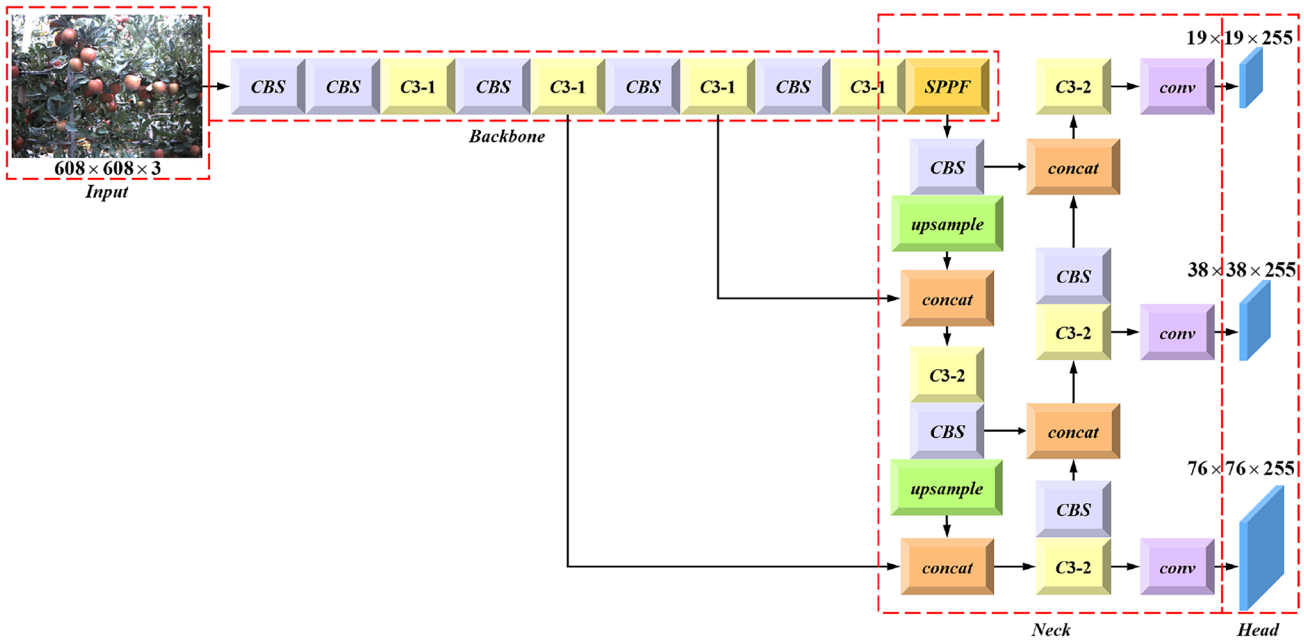

Fig. 5 The network architecture of YOLOv5s

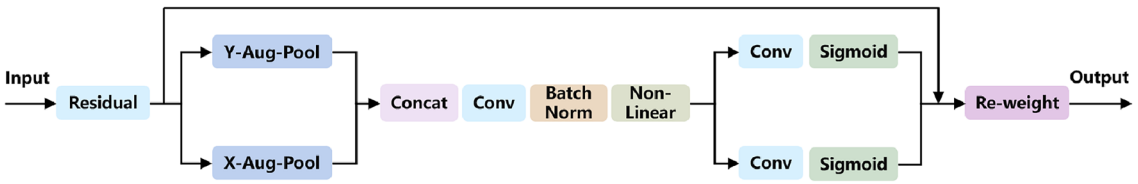

Fig. 6 Structure of the CA block. The terms 'X-Aug-Pool' and 'Y-Aug-Pool' refer to the one-dimensional horizontal and vertical global pools, respectively

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

Similarly, the output of the c-th channel with width w can be written as

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{2}$$

3.2.2 CA generation

After the transformation in information embedding, the obtained feature maps in two directions are concatenated together. Then, a 1×1 convolution kernel is used to convolve the concatenated feature maps. Further, the batch normalization of the convolved feature maps is performed to obtain the feature map F_1, and the non-linear activation function is used to activate the feature map F_1 to obtain the feature map f. The above process can be expressed as follows:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right). \tag{3}$$

Then, f is sliced into two separate tensors f^h and f^w along the spatial dimension, and next, the feature maps f^h and f^w are transformed to the same number of channels as the input feature map X using two 1×1 convolutions F_h and F_w. The sigmoid activation function is used to activate it. The equation is expressed as follows:

$$g^h = \sigma\left(F_h\left(f^h\right)\right), \tag{4}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right). \tag{5}$$

Finally, the output of CA block Y can be written as

$$Y = X \times g^h \times g^w. \tag{6}$$

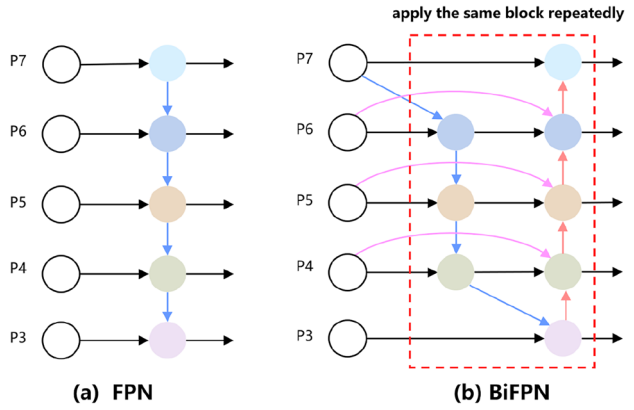

Fig. 7 Comparison of FPN and BiFPN. For P3–P7, it is multi-scale features from level 3 to level 7

3.3 BiFPN

To enhance the efficiency of the model, Tan et al. [19] developed a weighted bi-directional feature pyramid network. The main objective of this structure is to create a bi-directional connectivity mechanism based on the FPN, which allows information to flow in both directions and gradients to propagate throughout the network. This network is capable of multi-scale feature fusion of feature maps from different resolutions, thus improving the overall performance of the network. Since it is a versatile network, BiFPN can be seamlessly integrated with different neural network architectures, thus enhancing the generalization capability and stability of the network, for a wide range of image segmentation tasks. Figure 7 meticulously illustrates the comparison of FPN and BiFPN structures, where P_i represents a feature level with resolution of $1/2^i$ of the input images.

3.4 YOLOv5s-BC

In actual detection, the YOLOv5s algorithm can detect apples with high recognition. However, due to the interference and influence of the complex environment in the orchard, small target apples that are far away are usually ignored by the algorithm. Considering that the obscured apples are mistaken as targets, this leads to the robot that is not able to grab the apples by estimating their position and posture correctly. Consequently, we propose an improved YOLOv5s algorithm, named YOLOv5s-BC, by making several modifications as follows:

(i) The CA attention mechanism block is introduced in the backbone network and the neck network. In the backbone network, the CA attention mechanism can help the model to automatically filter and focus on key feature channels, reduce unnecessary information redundancy, optimize model parameters, and reduce computational costs, thus improving the efficiency and speed of the model. In the neck network, the CA attention mechanism can weight different feature channels in the feature fusion process, which enables the model to better integrate multi-scale and multi-level information, and enhances the diversity of features and the robustness of the model. The C3-CA block is shown in Fig. 8.

(ii) The BiFPN block is designed in the neck network, which first receives feature maps of different scales from across the region in the backbone, and then performs concat operation on these feature maps, which is named Bi-concat. BiFPN combines the mechanism of bi-directional feature propagation, can effectively fuse the features of different scales, and thus improves the ability of the model to characterize the object at different scales and levels. In addition, the BiFPN block makes the information transfer of the feature pyramid more balanced and effective through multiple iterations of feature fusion and updating, contributing to improved accuracy and stability of the object detection model. At the same time, it assists the model to better understand the location and size of the object in the image, thus improving the accuracy of object localization.

(iii) A new detection head is added to the head network. As the resolution of the feature maps used for small object detection increases, the local receptive field of the feature maps shrinks accordingly, which allows the network to detect more small objects with lower resolution [31]. The addition of this detection head enables the use of high-resolution feature maps to detect smaller objects that are farther away, thereby improving the accuracy of object detection and localization.

Figure 9 depicts the network architecture of the enhanced algorithm in detail.

3.5 Model evaluation index

To evaluate the performance of the established model, several indexes are discussed in this section. True positives (TP) refer

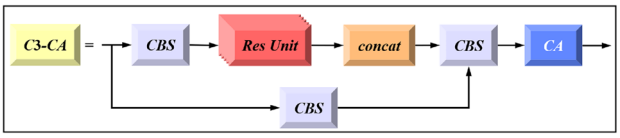

Fig. 8 Structure of the C3-CA block

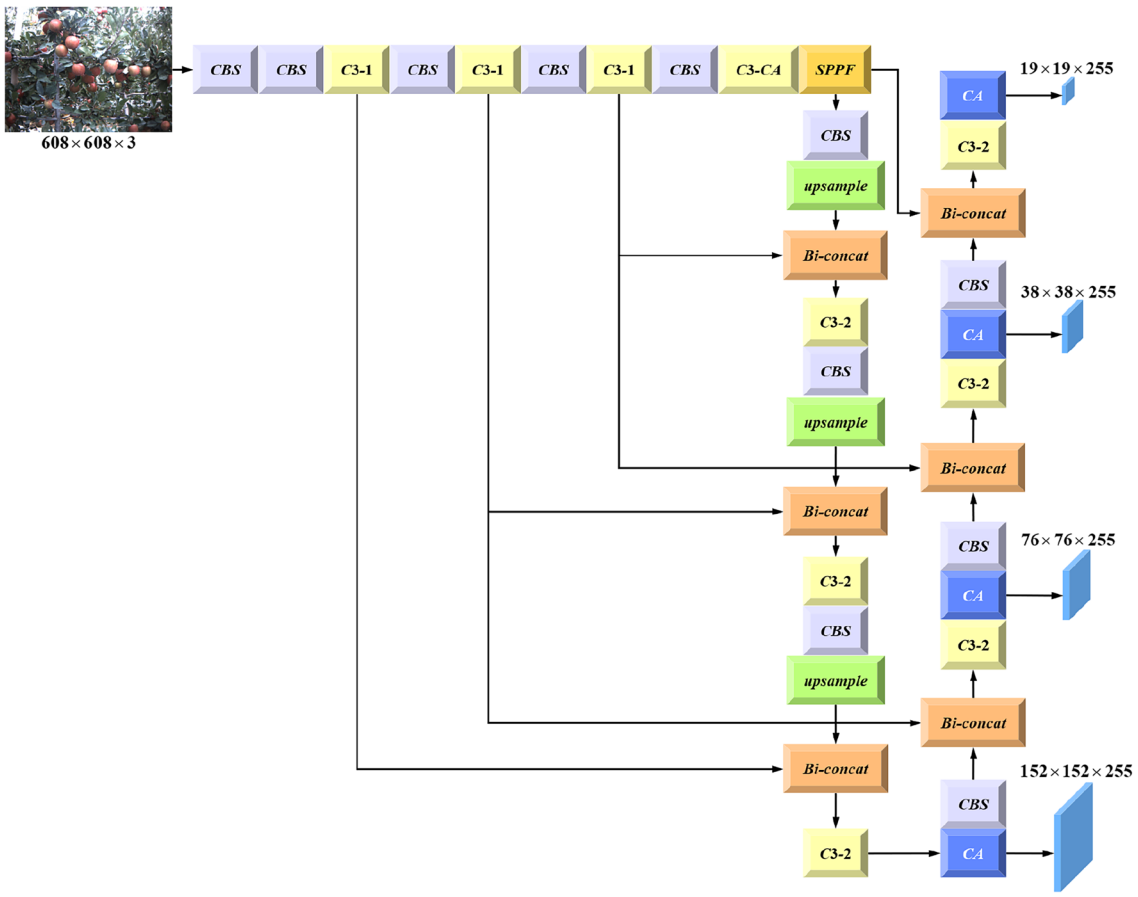

Fig. 9 The network architecture of YOLOv5s-BC

to positive samples that are categorized correctly. True negatives (TN) are negative samples that are identified accurately. False positives (FP) are negative samples that are mislabeled as positives. False negatives (FN) occur when positive samples are wrongly labelled as negative. Precision and Recall are defined in Eqs. (7) and (8), respectively. The calculation formula for Accuracy is shown in Eq. (9). $F1$ Score has become a metric often used in statistics to measure the accuracy of classification models due to the fact that it combines both the precision and recall of a classification model. It can be calculated by Eq. (10)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{9}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{10}$$

The precision–recall (P–R) curve is plotted with the horizontal coordinate as the recall rate R and the vertical coordinate as the precision rate P. The area enclosed by this curve is the average precision (AP). The calculation of AP is based on Eq. (11)

$$AP = \int_0^1 P_{(R)} \mathrm{d}R. \tag{11}$$

Mean average precision (mAP) is another commonly used evaluation metric in target detection models, which is the average of AP rate of each category.

The Intersection over Union (IoU) ratio is a crucial concept in target detection. It is the ratio of the intersection areas and union areas between the predicted box and the ground truth box, with non-deformation and non-negativity on the scale, as shown in Eq. (12). B^{gt} is the area of ground truth box, and B is the area of predicted box. w^{gt}

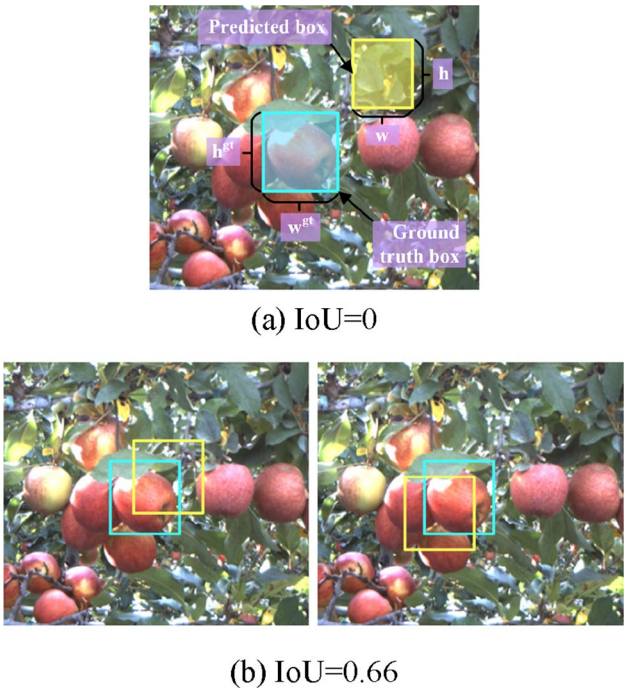

Fig. 10 IoU values for different situations

and h^{gt} are the width and height of the ground truth box respectively. Similarly, w and h are the width and height of the predicted box, respectively (see Fig. 10a)

$$\text{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}. \tag{12}$$

In Fig. 10a, it can be observed that two boxes did not intersect and as a result, IoU value is equal to zero which is insufficient to indicate their mutual distance. Additionally, when the loss is equivalent to zero, there is no backward transfer of gradient, and therefore, no learning progress can occur. Consequently, IoU falls short in providing a robust representation of their intersection. Figure 10b demonstrates that in both cases the IoU remains equal, however, their degree of overlap is different. To address this, numerous solutions have been proposed recently to enhance the IoU calculation. In this research, we adopt the original YOLOv5s selection, namely Complete IoU (CIoU), to compute the box loss (Box Loss).

The CIoU loss is proposed considering that the consistency of the bounding box aspect ratio is an important geometric factor [29]. α is a positive trade-off parameter (See Eq. (13)), υ is the similarity of the metric aspect ratio (see Eq. (14)), where b and b^{gt} denote the central points of ground truth box and predicted box, distance(·) is the Euclidean distance, and length(·) is the diagonal length of the smallest closed box that covers both boxes. CIoU loss is calculated by Eq. (15)

Table 2 Server configuration parameters

Parameters	On the server
Operating system	Ubuntu 18.04
GPU	RTX A4000 (16GB)
CPU	Intel Xeon Gold 5320
Deep learning framework	Pytorch 1.8.1
Programming language	Python 3.8

$$\alpha = \frac{\upsilon}{(1 - \text{IoU}) + \upsilon}, \tag{13}$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2, \tag{14}$$

$$\text{Loss}_{\text{CIoU}} = 1 - \text{CIoU} = 1 - \text{IoU} + \left(\frac{\text{distance}(b, b^{gt})}{\text{length}(B, B^{gt})} + \alpha\upsilon\right). \tag{15}$$

The Binary Cross Entropy (BCE) loss function is utilized to compute both classification loss (Cls Loss) and object loss (Obj Loss), as demonstrated in Eq. (16). n is the total number of samples, y_i is the category which the i sample belongs to, and x_i is the predicted probability of the i sample

$$\text{Loss}_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln\left(x_i\right) + \left(1 - y_i\right)\ln\left(1 - x_i\right)\right]. \tag{16}$$

The loss function measures the difference between predicted and ground truth information. A lower loss function value indicates greater similarity between the predicted and ground truth information. Therefore, the loss function is also an essential index to evaluate the target detection model. The loss function for our model is divided into three major parts: Box Loss, Cls Loss, and Obj Loss. It is the weighted sum of these losses (See Eq. (17))

$$\text{Loss} = \text{coef}_{\text{box}} \times \text{Loss}_{\text{box}} + \text{coef}_{\text{cls}} \times \text{Loss}_{\text{cls}} + \text{coef}_{\text{obj}} \times \text{Loss}_{\text{obj}}. \tag{17}$$

4 Experiments and discussion

4.1 Experimental setup

In this experiment, the training and testing of the model were done on the server. The server configuration parameters are shown in Table 2. In addition, the YOLOv5s-BC network

Table 3 Hyper-parameters

Hyper-parameters	Value
Initial value of learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Box loss coefficient	0.05
Cls loss coefficient	0.5
Obj loss coefficient	1.0

employs stochastic gradient descent (SGD) as the optimizer, with specific hyper-parameters shown in Table 3. Additionally, the model was trained for 200 epochs, and the batch size for model training was set to 16. The input image had a size of 640 pixels by default. After the above training parameters are determined, the model can be trained accordingly.

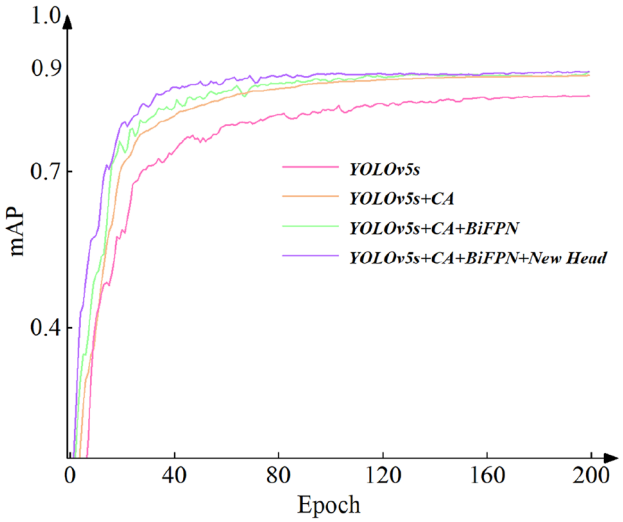

Fig. 11 The mAP values of different strategies in ablation experiments

4.2 Experimental results

4.2.1 Comparison of different target detection algorithms

In this work, a series of ablation experiments were performed on the apple dataset generated in Sect. 2 to assess the efficacy of the improved YOLOv5s-BC model. The results of the experiments are presented in Table 4.

After the inclusion of three additional modules, an enhanced detection model superior to YOLOv5s was achieved after training, albeit with a slight decrease in inference speed. Concurrently, the augmented complexity of the model leads to a minor increase in the generated weight files. To compensate for information loss during the transmission from the backbone network to the neck network and enhance the representation ability of the feature map. The network incorporates the CA attention mechanism module in series, leading to a significant increase in the average detection accuracy of the model. By further integrating shallower features and deep features, and incorporating the BiFPN module into the network, the convergence speed of model is notably enhanced, with the curve beginning to converge by the 120th epoch. The mAP has seen a 0.2% increase post-integration, and the AP for the block category has also improved by 0.4%. Finally, to solve the problem that small targets in the image are easily overlooked, a new detection head is added to the network. The convergence speed is doubled compared to the case without the detection head. The curve begins to converge at the 60th epoch. Experimental findings demonstrate that the proposed improved strategy effectively boosts the detection accuracy of apple targets in complex environments.

Figure 11 shows the change in average detection accuracy mAP during the training process of 200 epochs when adding different improvement strategies to the model. Figure 12 shows the changes in the loss function loss during the training process of 200 epochs when adding different improvement strategies to the model.

The efficacy of the YOLOv5s-BC was further tested by evaluating its performance against several other prominent target detection models, namely, YOLOv8, YOLOv4, YOLOv3, SSD, Faster R-CNN (VGG), and Faster R-CNN (ResNet50). Specifically, Faster R-CNN (VGG) and Faster R-CNN (ResNet50) employ VGG and ResNet50 as their respective backbone networks. All eight models were trained using the same training dataset and parameters determined previously. The training results of different target detection algorithms are presented in Table 5. It shows that the improved YOLOv5s-BC model achieves the 88.7% mAP on the test sets, outperforming the original YOLOv5s, YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet50),

Table 4 Ablation experiments on apple dataset

YOLOv5s	CA	BiFPN	New head	mAP	AP_{block}	FPS	Mb
✓				84.8	74.2	88.5	13.7
✓	✓			88.3	79.9	68.04	14.0
✓	✓	✓		88.5	80.3	62.12	14.9
✓	✓	✓	✓	88.7	80.5	55.25	16.7

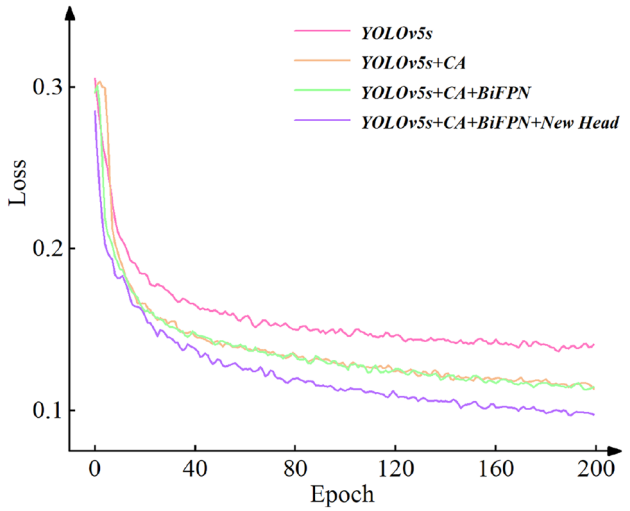

Fig. 12 The loss values of different strategies in ablation experiments

and Faster R-CNN (VGG) models by 4.6%, 3.6%, 20.48%, 23.22%, 15.27%, and 15.59%, respectively.

On the other hand, the YOLOv5s-BC model demonstrates 37.57% decrease in detection speed in comparison to the original YOLOv5s. Nevertheless, it provides a significant improvement over YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet50), and Faster R-CNN (VGG), by 434%, 331%, 240%, 201%, and 72%, respectively. This observation highlights the superiority of detection speed of the one-stage target detection algorithm over the two-stage target detection algorithm. It is essential to note that models designed for mobile devices with limited resources require lightweight. Therefore, the number of model parameters is an important index that assesses the model performance. The number of network layers and model parameters are increased in our proposed model due to the CA blocks embedded and a new detection head added. It is noted that the weight file of YOLOv5s-BC is 21.9% bigger than the original YOLOv5s. However, it is smaller than YOLOv8, YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet), and Faster R-CNN (VGG) by 4.7, 239.3, 229.6, 74.4, 96.8, and 530.2 Mb, respectively. In conclusion, although our proposed method is slightly inferior to the original YOLOv5s model in terms of detection speed and the number of model parameters, it is higher than the original YOLOv5s model in terms of detection accuracy. The overall performance of our proposed model is also the highest when compared with other target detection algorithms.

The *P–R* curves in Fig. 13 depict the performance of the proposed model by comparing its prediction results with the true labels at different thresholds. A model is considered to perform better when its *P–R* curves for different categories of targets are closer to the upper right corner. Specifically, Table 6 displays the *P–R* curve and F1 values of the proposed model, while Table 7 illustrates the accuracy of the proposed model when performed on the test set. For the pickable apples category, *F*1 is 91.6%. For the non-pickable apples category, *F*1 is 77.0%. This is due to the presence of leaves or overlapping apples obscuring them, making it difficult for the model to learn the complex high-level features. The overall *F*1 reaches 84.32%, which is the highest score among these algorithms. The detection accuracy reaches 99.8% for the class of graspable apples and 98.55% for the

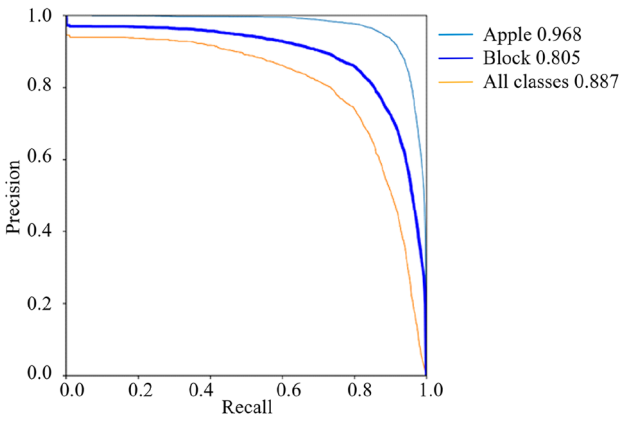

Fig. 13 The *P–R* curves of the proposed YOLOv5s-BC model

Table 5 Comparison of detection results of different models

Models	mAP (%)	AP_{block} (%)	F1 (%)	Detection speed (FPS)	Weight size (Mb)
Faster R-CNN (VGG)	76.74	63.18	70.50	32.09	546.9
Faster R-CNN (ResNet50)	76.95	62.44	70.50	18.34	113.5
SSD	68.1	50.86	68.12	16.27	91.1
YOLOv3	73.62	58.72	68.88	12.81	246.3
YOLOv4	85.62	76.11	81.31	10.35	256.0
YOLOv5s	84.8	74.2	79.83	88.5	13.7
YOLOv8s	88.6	80.4	84.0	56.49	21.4
YOLOv5s-BC	88.7	80.5	84.32	55.25	16.7

Table 6 Results of the proposed model

Category	Apple (%)	Block (%)	Mean (%)
Precision	91.3	74.7	83.0
Recall	91.9	79.5	85.7
F1	91.6	77.0	84.3

Table 7 Accuracy of the proposed model

Category	Apple	Block
Ground truth	6259	5101
Detection results	6249	5027
Accuracy	99.8%	98.55%

class of ungraspable apples. It is indicated that our model does not overfit on the test set and can detect new apple images well. Additionally, the model achieves a detection speed of over 55 FPS during video detection, demonstrating excellent recognition accuracy and efficiency in the real-time detection. Therefore, our model meets the standard requirements for mobile deployment.

4.2.2 Further test of the apple detection model

After conducting the aforementioned experiments, we have determined that the YOLOv5s-BC model offers the most optimal comprehensive performance, which satisfies the prerequisites for detecting apples in real time.

To further assess its accuracy in identifying the morphological attributes of apples, the feature maps of the detection layers were exhibited as heat maps. We chose picture number 1328 from the test set as the display image for conducting comparison experiments of YOLOv5s-BC and YOLOv5s models. Figure 14 illustrates the heat maps of both YOLOv5s and YOLOv5s-BC at the minimum detection layer. The YOLOv5s-BC model includes a new prediction head to enhance recognition of smaller objects that may be concealed by leaves or located far away. Detection results at small and medium scales reveal that the YOLOv5s model only provides a rough indication of the target location, which includes unnecessary information like leaves and branches. In comparison, our proposed model can identify the target more accurately while avoiding incorporating irrelevant details like leaves and branches, especially on small and medium scales. This is due to the CA mechanism, which enables the model to better focus on the most relevant parts of the image and thus enhances the overall detection accuracy.

To demonstrate the effectiveness of the improved model more visually, testing has been performed based on the apple images randomly selected from the produced apple image test set. In addition, to further illustrate that the improved model is also applicable to other apple orchards, apple images were downloaded from the publicly available apple

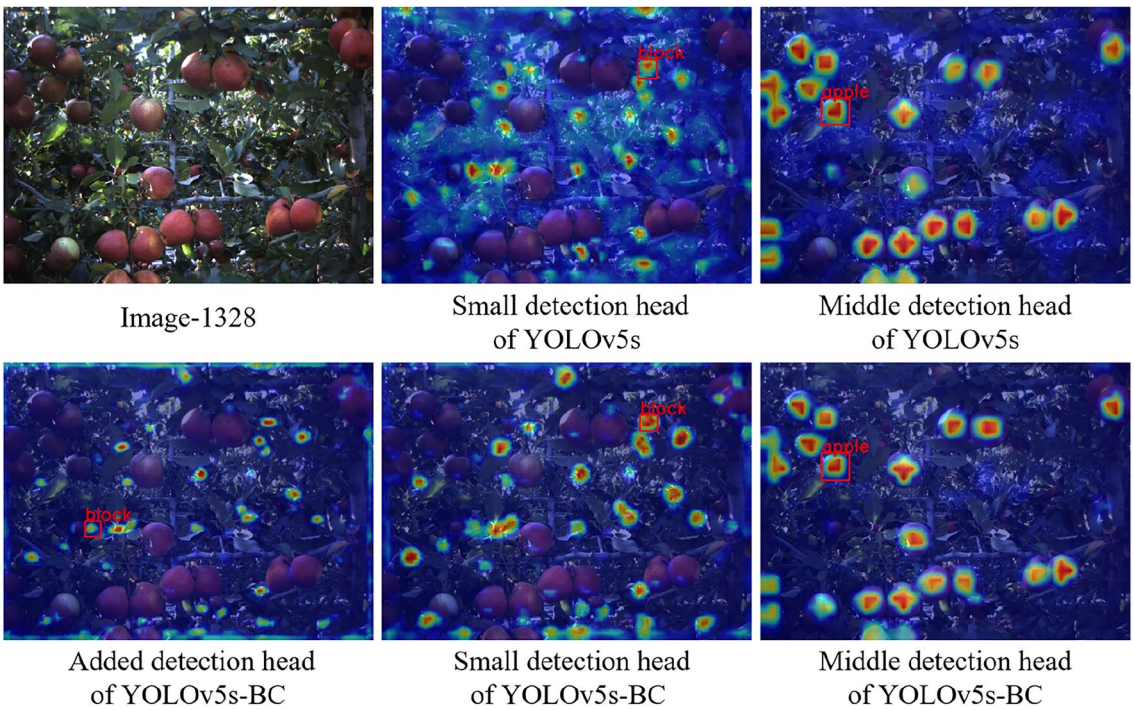

Fig. 14 Heat maps of YOLOv5s-BC and YOLOv5s

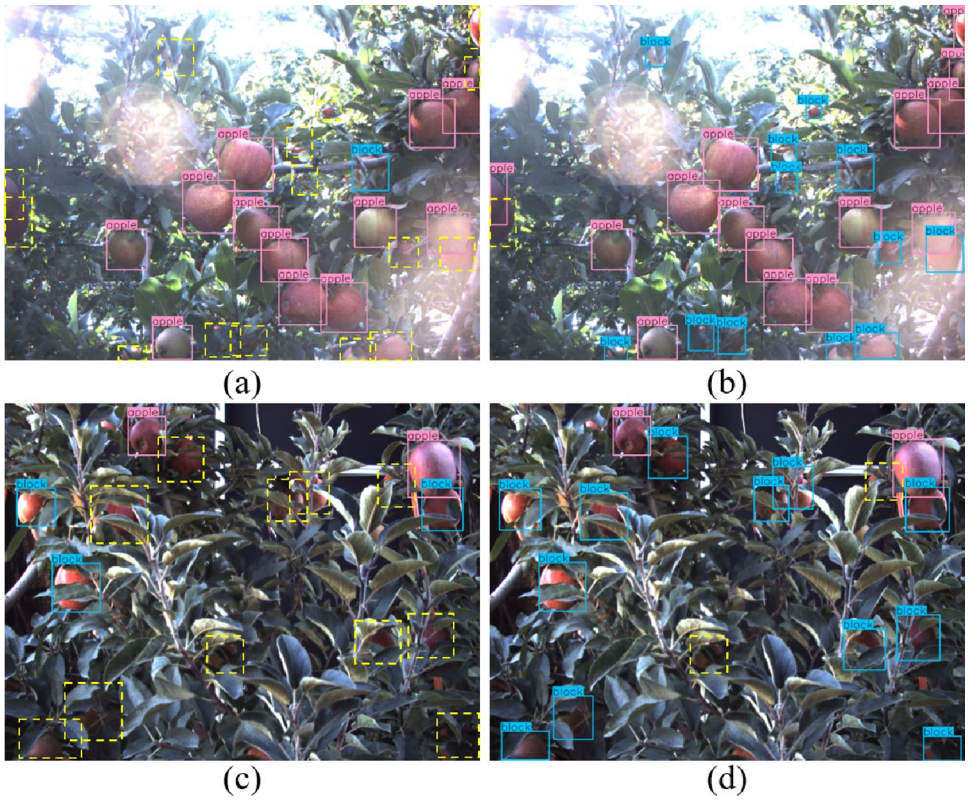

Fig. 15 Comparison of detection results before and after model improvement in near field of view scenes (short distance: 55–65 cm)

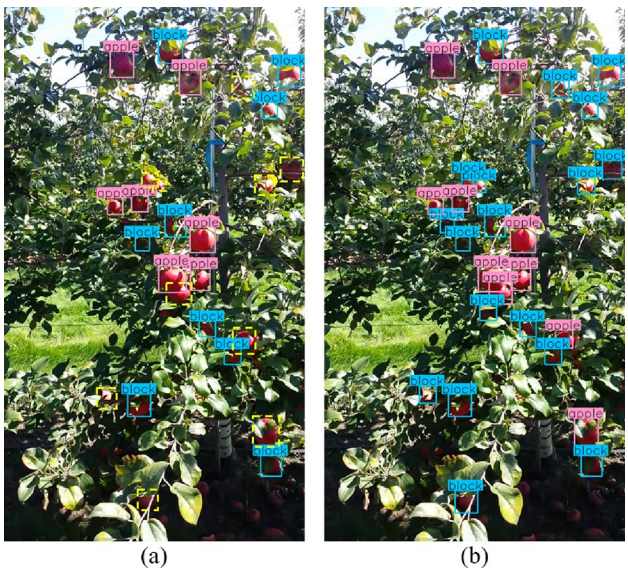

Fig. 16 Comparison of detection results before and after model improvement in far field of view scenarios (long distance: 110–130 cm)

dataset published by the Robotic Sensor Networks Laboratory at the University of Minnesota to supplement the test set [4]. The images tested in Fig. 15 are from the produced apple image test set, and the images tested in Fig. 16 are from the MinneApple dataset. The red box in the figure indicates that the test object is of the category "apple". The blue box indicates that the test object is of the category "block", and the yellow dashed box indicates a missed detection. In the figure, the long distance represents the distance from the depth camera to the apple tree between 110 and 130 cm, and the short distance represents the distance from the depth camera to the apple tree between 55 and 65 cm.

Specifically, Fig. 15 presents the detection results before and after the improvement of the YOLOv5s model in the near field of view scene. Figure 15a, c depicts the detection outcomes of the YOLOv5s model, while Fig. 15b, d illustrates the detection outcomes of the YOLOv5s-BC model. It is evident from the figures that the original model had numerous instances of missing apples during detection, with a total of 15 and 11 apples being missed, respectively. The improved model YOLOv5s-BC has significantly improved this situation. In particular, it can accurately identify and classify small target apples that are far away and apples obscured by leaves. In addition, the confidence level of the improved model for the apple target is above 75%, which fully demonstrates that the improved model has better detection results for apple targets in close-range scenes. Figure 16 illustrates the detection results before and after the model YOLOv5s improvement in the far-field scene. Where Fig. 16a shows the detection results of model YOLOv5s and Fig. 16b shows the detection results of model YOLOv5s-BC. It can be observed from the figure that the original model

misclassifies the distracting objects obscured by leaves as "apple" due to the distance of the depth camera from the apple target. In addition, the original model misses ten targets in the detection when the targets are similar in color to the background. In contrast, the improved model has any high detection rate for apples in the new orchard environment. Especially, for the small target apples in the image, the improved model can basically recognize and classify all of them correctly. The improved model YOLOv5s-BC constructs a more efficient feature fusion network, which enables full feature fusion of high-level and lower-level information, providing more detection information for small targets. The experimental results show that the improved model performs better than the original model in detecting small targets in long-range scenes.

5 Conclusions

In this paper, a real-time detection method based on YOLOv5s-BC is presented for apple detection. By adding a new detection head and combining the CA and BiFPN modules to optimize the YOLOv5s network model, the image features of target apples can be effectively extracted and the detection capability of smaller target apples can be enhanced. The detailed conclusions are summarized as follows.

The mAP performance of the YOLOv5-BC model on the test set reaches 88.7%, improving over the YOLOv5s, YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet50), and Faster R-CNN (VGG) models by 4.6%, 3.6%, 20.48%, 23.22%, 15.27%, and 15.59%. The weight size of the model is only 16.7 Mb, larger than the original YOLOv5s by 3 Mb, but smaller than YOLOv8, YOLOv4, YOLOv3, SSD, Faster R-CNN (ResNet), and Faster R-CNN (VGG) by 4.7, 239.3, 229.6, 74.4, 96.8, and 530.2 Mb. The detection of an image takes only 0.018 s, which guarantees the real-time requirements for apple detection. In the heat map, adding a new detection head to the model can detect apples from smaller targets. In addition, adding the CA mechanism makes the model pay more attention to and learn the high-level information of the detected targets and abandon other irrelevant information. In the test experiments at short and long distances, the proposed model can detect all the targets more perfectly, displaying the well robust performance of the model.

However, there are still certain limitations of the YOLOv5-BC model, such as the existence of a small number of missed or false detections. Therefore, the attention mechanism of the model needs to be further optimized, and the backbone network of the model needs to be modified as well, to further improve the detection accuracy of the proposed model.

Acknowledgements The present work is supported by the Start-up Funds from Wuhan University of Technology and the National Innovation and Entrepreneurship Training Program for College Students (S202310497143). Research participants (Jie Lin, Yu Pei, and Rongzhen Yang) are also appreciated.

Author contributions Jingfan Liu: Methodology, Software, Investigation, Visualization, Writing – original draft, Writing – review & editing.Zhaobing Liu: Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bao, W.X., Zhu, Z.Q., Hu, G.S., et al.: UAV remote sensing detection of tea leaf blight based on DDMA-YOLO. Comput. Electron. Agric. [J] **205**, 17 (2023). https://doi.org/10.1016/j.compag.2023.107637
2. Bochkovskiy, A., Wang, C.-Y., Mark Liao, H.-Y.: YOLOv4: Optimal Speed and Accuracy of Object Detection (2020). Arxiv [J]. arXiv:2004.10934
3. Fountas, S., Mylonas, N., Malounas, I., et al.: Agricultural robotics for field operations. Sensors [J] **20**(9), 27 (2020). https://doi.org/10.3390/s20092672
4. Häni N., Roy P., Isler, V.: MinneApple Data [M] (2019)
5. He, K.M., Gkioxari G., Dollar P., et al.: Mask R-CNN[C]. In: 16th IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, Italy, pp. 2980–2988 (2017). https://doi.org/10.1109/iccv.2017.322
6. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021). https://doi.org/10.48550/arXiv.2103.02907
7. Jia, W.K., Zhang, Y., Lian, J., et al.: Apple harvesting robot under information technology: a review. Int. J. Adv. Robot. Syst. [J] **17**(3), 16 (2020). https://doi.org/10.1177/1729881420925310
8. Li, K.S., Wang, J.C., Jalil, H., et al.: A fast and lightweight detection algorithm for passion fruit pests based on improved YOLOv5. Comput. Electron. Agric. [J] **204**, 11 (2023). https://doi.org/10.1016/j.compag.2022.107534
9. Liang, J.T., Chen, X., Liang, C.J., et al.: A detection approach for late-autumn shoots of litchi based on unmanned aerial vehicle (UAV) remote sensing. Comput. Electron. Agric. [J] **204**, 10 (2023). https://doi.org/10.1016/j.compag.2022.107535
10. Liu, W., Anguelov, D., Erhan, D., et al.: Ssd: single shot multibox detector[C]. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2

11. Lu, Y.Z., Young, S.: A survey of public datasets for computer vision tasks in precision agriculture. Comput. Electron. Agric. [J] **178**, 13 (2020). https://doi.org/10.1016/j.compag.2020.105760
12. Lv, J.D., Xu, H., Han, Y., et al.: A visual identification method for the apple growth forms in the orchard. Comput. Electron. Agric. [J] **197**, 9 (2022). https://doi.org/10.1016/j.compag.2022.106954
13. Qi, J.T., Liu, X.N., Liu, K., et al.: An improved YOLOv5 model based on visual attention mechanism: application to recognition of tomato virus disease. Comput. Electron. Agric. [J] **194**, 12 (2022). https://doi.org/10.1016/j.compag.2022.106780
14. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection[C]. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, pp. 779–788 (2016). https://doi.org/10.1109/cvpr.2016.91
15. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision & Pattern Recognition [J], pp. 6517–6525 (2017). https://doi.org/10.1109/CVPR.2017.690
16. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. Arxiv [J] (2018). arXiv:1804.02767
17. Ren, S.Q., He, K.M., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. [J] **39**(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031
18. Sun, L.J., Hu, G.R., Chen, C., et al.: Lightweight apple detection in complex orchards using YOLOV5-PRE. Horticulturae [J] **8**(12), 15 (2022). https://doi.org/10.3390/horticulturae8121169
19. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. Arxiv [J], pp. 10778–10787 (2020). arXiv:1911.09070
20. Ultralytics yolov5 [M]
21. Wang, C.-Y., Liao, H.-Y. M., Wu ,Y.-H., et al.: CSPNet: A new backbone that can enhance learning capability of CNN[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
22. Wu, F.Y., Duan, J.L., Ai, P.Y., et al.: Rachis detection and three-dimensional localization of cut off point for vision-based banana robot. Comput. Electron. Agric. [J] **198**, 12 (2022). https://doi.org/10.1016/j.compag.2022.107079
23. Xu, B., Cui, X., Ji, W., et al.: Apple grading method design and implementation for automatic grader based on improved YOLOv5. Agric. Basel [J] **13**(1), 18 (2023). https://doi.org/10.3390/agriculture13010124
24. Xu, Z.B., Huang, X.P., Huang, Y., et al.: A real-time zanthoxylum target detection method for an intelligent picking robot under a complex background, based on an improved YOLOv5s architecture. Sensors [J] **22**(2), 15 (2022). https://doi.org/10.3390/s22020682
25. Yan, B., Fan, P., Lei, X.Y., et al.: A real-time apple targets detection method for picking robot based on improved YOLOv5. Remote Sens. [J] **13**(9), 23 (2021). https://doi.org/10.3390/rs13091619
26. Yao, J., Qi, J.M., Zhang, J., et al.: A real-time detection algorithm for kiwifruit defects based on YOLOv5. Electronics [J] **10**(14), 13 (2021). https://doi.org/10.3390/electronics10141711
27. Zhang, D.Y., Luo, H.S., Wang, D.Y., et al.: Assessment of the levels of damage caused by Fusarium head blight in wheat using an improved YoloV5 method. Comput. Electron. Agric. [J] **198**, 16 (2022). https://doi.org/10.1016/j.compag.2022.107086
28. Zhao, Y.S., Gong, L., Huang, Y.X., et al.: A review of key techniques of vision-based control for harvesting robot. Comput. Electron. Agric. [J] **127**, 311–323 (2016). https://doi.org/10.1016/j.compag.2016.06.022
29. Zheng, Z.H., Wang, P., Liu, W., et al.: Distance-IoU loss: faster and better learning for bounding box regression[C]. In: 34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence. Assoc Advancement Artificial Intelligence, New York, pp. 12993–13000 (2020). https://doi.org/10.48550/arXiv.1911.08287
30. Zhou, H.Y., Wang, X., Au, W., et al.: Intelligent robots for fruit harvesting: recent developments and future challenges. Precis. Agric. [J] **23**(5), 1856–1907 (2022). https://doi.org/10.1007/s11119-022-09913-3
31. Zhu, X.K., Lyu, S.C., Wang, X., et al.: TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]. In: 18th IEEE/CVF International Conference on Computer Vision (ICCV). Electr Network: Ieee Computer Soc, pp. 2778–2788 (2021). https://doi.org/10.1109/iccvw54120.2021.00312