IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2023, accepted 3 August 2023, date of publication 15 August 2023, date of current version 23 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305405

PERSPECTIVE

Tea Buds Detection in Complex Background Based on Improved YOLOv7

JUNQUAN MENG, FENG KANG, YAXIONG WANG, SIYUAN TONG, CHENXI ZHANG, AND CHONGCHONG CHEN

School of Technology, Beijing Forestry University, Beijing 100083, China
Key Laboratory of State Forestry Administration for Forestry Equipment and Automation, Beijing 100083, China

Corresponding author: Feng Kang (kangfeng98@bjfu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant BLX201820, and in part by the Ningxia Hui Autonomous Region Key Research and Development Plan Project 2019BBF02009.

ABSTRACT Aiming at the problem that the color of tea buds is highly similar to the background in complex scenes and it is difficult to identify the buds, this study proposed an improved YOLOv7 algorithm by replacing the original convolution blocks with Depth Separable Convolution (DS Conv) blocks, and adding Convolutional Block Attention Modules (CBAM) and Coordinate Attention (CA) modules. The method improved mean Average Precision (mAP) by 1.28% and mean Recall (mR) rate by 2.92%, the final mAP and mR reached 96.70% and 93.88%, respectively, and 30.62 Frame Per Second (FPS) of the improved model meets the requirements of real-time detection. The results show that the detection accuracy of the improved YOLOv7 algorithm for tea buds was higher than that of other target detection algorithms, and the detecting performance is not significantly affected by the light conditions, and the recognition accuracy of tea buds at each growing period was excellent and balanced. This study provides experience for the realization of intelligent tea picking.

INDEX TERMS Attention module, deep learning, tea buds, target detection, YOLOv7.

I. INTRODUCTION

Tea is the second most consumed beverage in the world, of which the market has a large potential [1], [2]. As one of the world's most important beverages, the way in which tea is harvested needs further refinement. At present, the tea harvesting methods in China are mechanical and manual picking, and mechanical picking is suitable for bulk tea with low economic value, which is operated by reciprocating cutting, and the harvested tea leaves are incomplete and mixed with a large number of old leaves and broken branches [3]. Harvesting method of famous tea with high economic value is still mainly manual picking, which is time-consuming, intensive and inefficient with high labor cost [4], so the demand for automated and intelligent picking is increasingly urgent, in which the identification of tea shoots is the basis of intelligent picking [5]. Compared with other target detection tasks, tea shoots have characteristics such as close color to the background, small sizes, high density, and large differences between individuals at different periods, which lead to more difficulties to identify.

The standard of tea picking is divided into one bud with one leaf, one bud with two leaves, and one bud with three leaves, one bud with one leaf means the tender bud and the first leaf, one bud with two leaves means the tender bud and the first two leaves, and one bud with three leaves means the tender bud and the first three leaves. All three are of good quality and have relatively high economic value, and are picked according to different market demands. It is considered that one bud with one leaf has the highest content of tea polyphenols and nitrogen compounds, so the economic value of one bud with one leaf is higher than that of one bud with two leaves and three leaves, and one bud with one leaf is generally picked as the highest grade [6].

The mainstream target detection algorithms include Faster RCNN [7], SSD [8], YOLO (You Only Look Once) [9]

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

algorithms, Faster RCNN algorithm is a two-stage detector with relatively high accuracy, SSD and YOLO algorithms are both single-stage detectors, of which SSD has the fastest detection speed, but lower accuracy, the detection speed of YOLO series is much faster than the two-stage detector, and YOLO's accuracy is gradually approaching the two-stage detector through continuous development [10].

YOLO was first introduced in 2016, and in contrast to the two-stage detector, YOLO does not need to find the possible target regions in advance [9]. YOLOv7 [11] was released in July 2022. Similar to its predecessors, YOLOV7 is still structured in three parts: backbone, neck, and head. The RGB images are compressed in size and expanded in number of channels in the backbone section for feature extraction, and then three feature layers are output to the neck section for further feature extraction and feature fusion through a series of up-sampling and down-sampling operations to aggregate feature information at different scales. The COCO dataset is clustered to obtain anchor boxes of different sizes and aspect ratios, through K-means clustering method, the anchor boxes are continuously adapted to the labels during the training process, and finally the most suitable boxes are retained through the non-maximal suppression method. The prediction result is obtained by the head part, which determines whether the box contains an object and the category of target. The whole YoloV7 workflow consists of feature extraction, enhanced feature extraction and feature fusion, and prediction of the targets.

The YOLOv7 algorithm had excellent performance on public datasets, its accuracy was higher than other target detection algorithms [11]. The model had achieved good results in pedestrian detection [12], kidney stone detection [13], and rack detection [14], whereas it had not been applied to tea bud detection. This study proposes an improved YOLOv7 model to detect tea buds in complex background where the color of tea buds is similar to the background and their growth stages are different, making it difficult to identify.

The remaining parts of this paper are organized as follows. Section II discusses the relevant research on tea bud detection. Section III explains the method of constructing the dataset and the details of the improvement method of YOLOv7. Section IV discusses and analyzes the experimental results. Section V summarizes the methods of this study and makes prospects.

II. RELATED WORKS

The current methods for identifying tea shoots based on image processing are classified into traditional image segmentation and deep learning. Related studies using traditional image segmentation: Zhang et al. [15] combined improved B-G algorithm with Bayes to identify and classify the shoots. Lei et al. [16] extracted the G-B features of the images, Otsu method was used to extract the skeleton of tea shoots after the second time segmentation, and Shi-Tomasi algorithm was used to detect the skeleton corners and mark the picking points, the recognition success rate was 85.12%. Xuemei Wu et al. [17] used the K-means clustering method based on a and b components in the Lab color model to segment tea shoots, the results showed that for tea images collected at different distances, the method performed better than the Otsu method, the average recognition rate of the method was 94% and the segmented shoots were more complete. The traditional image segmentation method extracts the shoot parts based on the difference of color between the shoots and the background, which has poor robustness, slow computing speed, and is difficult to detect in real time.

With the rapid development of computer vision, deep learning methods have been gradually applied to the field of tea bud recognition [18], [19], [20], [21], [22]. The main research is about picking point detection, some of them are based on deep learning, combining image segmentation methods to identify picking points: Yang et al. [23] used the improved PSO-SVM algorithm to extract tea bud foreground and the improved YOLOv3 model was used to recognize the intersection points of detected boxes and tea stems as picking points, the accuracy rate was over 90%. Chen et al. [24] developed an intelligent vision system applied to tea buds picking and used the YOLOv3 algorithm to identify tea buds, and extracted the picking points by combining skeleton extraction with minimum bounding rectangle algorithms, the average accuracy of tea bud recognition was 71.96% and the picking point extraction accuracy was 83%. Yan et al. [25] proposed an MR3P-TS model based on improved Mask RCNN algorithm, the picking points were obtained by processing the minimum bounding rectangle of the segmented tea bud region. The mAP and f2 values of segmentation were 44.9% and 31.3% respectively, and the accuracy and recall rate of picking point localization were 94.9% and 91.0% respectively. The studies above used convolutional neural networks to extract the tea bud images and explore the picking point based on the shape characteristics of the tea bud. However, the tea buds in these studies were relatively less affected by background and therefore easier to extract the foreground. Fewer studies have been conducted for tea bud target detection. Xu et al. [26] used the YOLOv3 algorithm for tea bud detection, and the Densenet201 algorithm was used for further classification. The detection accuracy of this method for tea buds was 95.71%. Jun Lv et al. [27] proposed a tea shoot detection model based on region luminance adaptive correction with chunking of different sizes and local region gamma luminance adaptive correction for high luminance images. YOLOv5 was used for recognition, the accuracy and recall were 92.4% and 90.4% respectively and the method had strong robustness to different light intensities. Zhu et al. [28] used the Faster RCNN algorithm to detect tea shoots in complex backgrounds, and the average accuracy was 54% when no distinction was made between shoot types, the average accuracy was 22% and 75% when single shoots and one/two leaves were distinguished, it was illustrated by comparison that the deep learning method outperformed

TABLE 1. Hyperparameter settings.

Freezing epoch	Freezing batch size	Unfreezing epoch	Unfreezing batch size	Optimizer	Initial learning rate	momentum	Weight decay	Learning rate decay type
50	4	300	2	SGD	1e-2	0.937	5e-4	Cos

the traditional segmentation algorithm in terms of detection accuracy and speed. The tea buds of above studies were clearly differentiated from the background in terms of color, and the interference caused by the background is relatively minor. Moreover, the tea buds in these studies are all in mature stage, making them relatively easy to identify.

The tea buds of above studies were clearly differentiated from the background in terms of color, and they were all in the mature growth stage. The research object of this paper is Fuyun 6 tea species in complex background, the buds are close to the old leaves in color, which makes it more difficult to pre-process by traditional segmentation algorithm, and the buds of Fuyun 6 in natural state contain multiple growth stages from initial spreading to totally mature, and the postures of tea buds are inconsistent. The dataset of this study is closer to the situation of real picking. This study used the deep learning method to identify buds of Fuyun 6 species in complex background.

III. MATERIALS AND METHODS

A. EXPERIMENTAL EQUIPMENT AND DATA PRE-PROCESSING

The experimental site of this study was the Wangu tea plantation (23°19′38″N, 108°39′24″E) in Mingliang Town, Shanglin County, Nanning City, Guangxi Zhuang Autonomous Region, China. The images were taken by the rear camera of Huawei mate30 cell phone (super-sensitive main camera with 40 megapixels, telephoto camera with 8 megapixels). The experimental object was Fuyun 6 tea species. The best time for tea bud picking was early April (around the Qingming Festival) [29]. The image collecting period was from April 4th to 7th, 2022. The images were taken at horizontal angle, 20cm-50cm away from the buds. The number of original images was 1063, of which 533 images were taken under strong light (12.am-14.pm) and 530 images were taken under weak light (17.pm-19.pm), all images were saved in JPG format.

The experimental hardware was AMD Ryzen5 3600X CPU and NVIDIA GeForce RTX 3060 Ti, running under Windows 10 operating system, based on Anaconda environment, the deep learning framework was Pytorch1.10.0, the input size of images was 640 pixel × 640 pixel, and the hyperparameters of training settings for detection algorithms are shown in Table 1. The SGD optimizer was chosen because it achieved higher accuracy than the ADAM optimizer in previous experiments. The advantage of the ADAM optimizer is that for large datasets it can more effectively make use of its adaptive learning rate and momentum adjustment to speed up convergence. However, for small datasets, this adaptive mechanism may lead to performance degradation of the Adam optimizer, the SGD has greater generalization and resistance to over-fitting [30]. Our dataset of 2126 images is a relatively small one, on which the SGD optimizer performed better. The dataset was produced in PascalVOC format and labeled using the Labelimg tool [31].

It is considered that one bud with one leaf has high economic value [6], [22]. In this study, one bud with one leaf was labeled. There are two forms of one bud with one leaf: spreading and developed [32], they are not distinguished during actual harvesting and are targets to harvest, so both spreading and developed forms were considered as objects to identify in this study. As shown in Fig. 1, the tea buds were divided into four classes for labeling.

Tea buds are mostly grown vertically upward, so the horizontal mirroring method was used to augment the dataset. The original dataset consisted of 1063 images, after horizontal mirroring there were 2126 images. The dataset was randomly divided into training set and validation set in the ratio of 4:1, 1700 images were assigned to the training set and 426 images to the validation set.

B. THE IMPROVED YOLOV7 MODEL

The network structure of the improved YOLOv7 model is shown in Fig. 2, which contains backbone network, neck network, and detection head [11]. RGB images of 640 pixels × 640 pixels are input to the backbone part, then the feature maps of size 20 pixels × 20 pixels, 40 pixels × 40 pixels, and 80 pixels × 80 pixels are output after operated by BConv blocks, ELAN layers, and MPConv layers respectively.

Change of the backbone structure causes the pre-training weights to be inapplicable, so only the neck part was improved. The YOLOv7 network structure was improved through adding the Depth Separable Convolution (DS Conv), CBAM attention modules, and Coordinate Attention (CA) modules. The conventional convolution blocks in the black dashed boxes in Fig. 2 were replaced with the DS Conv blocks, the CBAM attention modules were embedded before the DS Conv blocks and MPConv layers, and the CA modules were added after the small and medium-sized feature layers which were output from the backbone.

1) DEPTH SEPARABLE CONVOLUTION

DS Conv consists of depth convolution and point convolution [33]. DS Conv has lower number of parameters and lower operation cost compared with conventional convolution. RGB images are first convolved by depth convolution with kernel size of 3 × 3, the number of convolution kernels

FIGURE 1. Classes to be labeled. Class 1: developed one bud with one leaf, the bud and leaf are clearly separated and the leaf is completely stretched. Class 2: spreading one bud with one leaf, the bud and the leaf are not completely separated and the leaf is curled up. Class 3: primordial spreading one bud with one leaf, the junior form of Class 2, with bud contained within the leaf. Class c: buds in lateral posture.

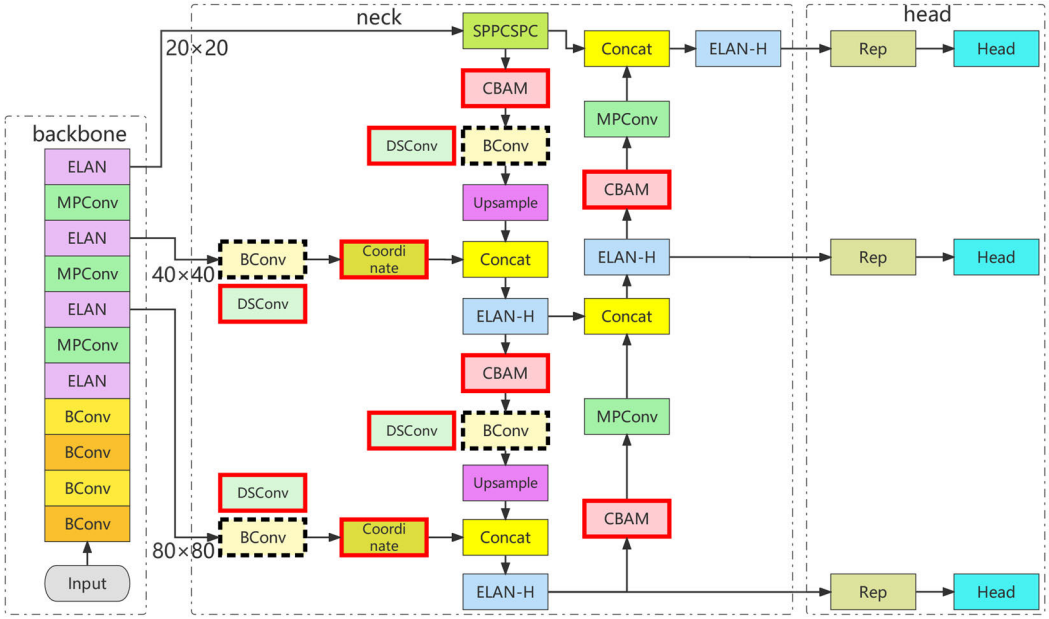

FIGURE 2. Structure of the improved YOLOv7 model. Black boxes are modules of the original YOLOv7, Red boxes are the added modules, Black dotted boxes are the replaced modules, BConv blocks with different colors differ from the convolution kernels and step sizes.

is the same as the number of input channels, so after the operation a 3-channel image generates 3 feature maps, which cannot expand the number of feature maps, and it convolves each channel independently, which cannot effectively take advantage of the feature information of different channels in the same spatial location. Therefore, point convolution is used to combine the feature maps in a weighted manner. The size of the convolution kernel of point convolution is 1×1, and the number of output channels is determined by the number of convolution kernels, avoiding the influence of the number of conventional convolution kernels and step size, thus improving the detection efficiency.

2) CBAM ATTENTION MODULE

CBAM [34] consists of channel attention and spatial attention. Channel attention focuses on the noteworthy part of the image and spatial attention focuses on its location, operating on the channel and spatial dimensions, respectively, to give greater weight to the feature information. The obstructions and background of tea buds in the natural state tend to cause interference, which can easily lead to false detection such as target omission or inaccurate recognition boxes. The CBAM attention module was introduced, with the channel attention module focusing on feature information and the spatial attention module focusing on location information of tea buds. Through the synergistic effect of the two modules, the weight of the background part was reduced, to enhance the feature extraction ability towards tea buds and to frame the target areas more accurately. The CBAM module was added to the feature-fusing part of the network, and in order to explore the best position for embedding the CBAM module, it was added to the two positions as shown in Fig. 3, and their detection effect were compared.

3) COORDINATE ATTENTION MODULE

CA [35] fuses location and channel information by aggregating features input from horizontal and vertical directions, a 1D global pooling operation was used to obtain 2 direction-aware and location-sensitive feature maps, improving the

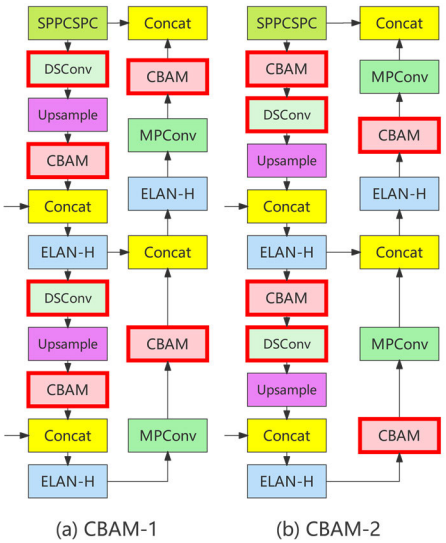

FIGURE 3. Different positions of CBAM Attention. (a) CBAM-1, adding CBAM modules after Upsampling and MPConv modules, (b) CBAM-2, adding CBAM modules before DS Convs and MPConv modules.

loss of location information caused by 2D global pooling, promoting the target feature extraction capability, and expanding the perceptual field. CA can precisely identify and locate targets, which is more friendly to small targets [36]. Tea shoots are of different sizes, the color of small-sized buds is not well distinguished from background and they contain less area of pixels. The YOLOv7 model is prone to lose feature information of small-sized shoots in complex backgrounds. This study introduced the CA module that can efficiently recognize small-sized and dense targets. Through comparison of experiments, the best position of Coordinate module was explored. To improve the detection ability of the model for small and medium-sized tea buds, the Coordinate modules were added to the neck and head part, of which the specific position is shown in Fig. 4.

The structures with CBAM modules and Coordinate modules added at different positions are detailed in Table 2.

C. K-MEANS++ CLUSTERING BOUNDING BOXES SIZES

The sizes of the bounding boxes pre-defined for YOLOv7 are obtained using K-means method to cluster the coco dataset, whereas the coco dataset differs significantly from the tea buds dataset used in this study, so the modified K-means clustering method (K-means++) was used to cluster our tea buds dataset to obtain sizes of anchor boxes that suitable for our study. The results are shown in Table 3, and were used as the pre-defined sizes of anchor boxes for all algorithms in this study.

The K-means algorithm randomly selects k clustering centers, and the clustering effect is affected by the center selection, the result cannot converge to the global minimum when several centers are initialized to the same cluster [37]. The K-means++ algorithm improves the initialization of cluster centers by selecting k centers one by one. It randomly

TABLE 2. Details of structures.

Module	Structure	Detail
CBAM Module	CBAM-1	adding CBAM modules after Upsampling and MPConv modules, before Concat modules
	CBAM-2	adding CBAM modules before DS Convs and MPConv modules
Coordinate Module	COR-1	Coordinate modules are added into the neck part, after DS Convs
	COR-2	Coordinate modules are added into the neck part, before DS Convs
	COR-3	Coordinate modules are added into the head part, before Rep modules

TABLE 3. Defaulted sizes of anchor boxes.

Feature Map	20×20	40×40	80×80
Sizes clustered by K-means++	[125,142]	[75,83]	[33,50]
	[151,188]	[88,153]	[51.69]
	[214,245]	[93,111]	[59,110]

selects the first center point, and then selects the next data point with a higher probability if it has a larger Euclidean distance from the previous center. The distance between each of its k centers is far enough to avoid the shortcomings of K-means clustering.

D. MODEL EVALUATION INDICATORS

The detection accuracy is evaluated using *mAP* (mean Average Precision) and *mR* (mean Recall), the FPS (Frame Per Second) evaluates the detection speed. Precision (*P*) is calculated by the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

The Recall rate (*R*) is calculated by the following equation:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where *TP* (True Positive) indicates that the target is identified as the target, *FP* (False Positive) indicates that the background is identified as the target, *TN* (True Negative) indicates that the background is identified as the background, and *FN* (False Negative) indicates that the target is identified as the background [38].

There is an interaction between Precision and Recall, and different Precision and Recall are calculated from different confidence threshold values. The area enclosed by the *P-R* curve is denoted by *AP* (Average Precision):

$$AP = \int_0^1 P(R)dR \tag{3}$$

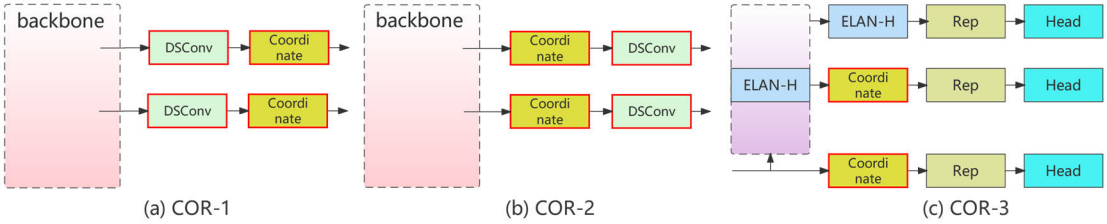

FIGURE 4. Different positions of Coordinate Attention. Coordinate modules are all added into the medium and small feature layers, (a) COR-1: Coordinate modules are added into the neck part, after DS Convs, (b) COR-2: Coordinate modules are added into the neck part, before DS Convs, (c) COR-3: Coordinate modules are added into the head part, before Rep modules.

TABLE 4. Performance comparison between different target detection models.

	mAP0.5(%)	mR(%)	FPS
Retinadet -resnet50	94.28	89.21	32.82
Efficientdet -d1	95.08	90.09	18.33
Faster RCNN -vgg	92.75	95.10	26.87
SSD-vgg	86.42	77.53	92.52
YOLOv5s	89.54	80.41	66.76
YOLOv7	95.42	90.96	34.59

mAP denotes the average value of *AP* for all classes (assuming there are k classes):

$$mAP = \frac{\sum_{1}^{k} AP(i)}{k} \tag{4}$$

mR denotes the average of Recall for all classes (assuming there are k classes):

$$mR = \frac{\sum_{1}^{k} R(i)}{k} \tag{5}$$

FPS denotes the number of images processed per second.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. COMPARISON OF PERFORMANCE OF DIFFERENT DETECTION ALGORITHMS

The YOLOv7 model was compared with other target detection algorithms such as Retinadet, Efficientdet, SSD, Faster RCNN, and YOLOv5s. The results are shown in Table 4.

As shown in the table, the mAP and mR of Efficientdet and Faster RCNN are 95.08% and 90.09%, 92.75% and 95.10%, respectively, where Faster RCNN has the highest mR value, but the detection speed of the above 2 networks is too slow (18.33 FPS and 26.87 FPS), which cannot meet the demand for real-time detection (30 FPS) [39]. SSD has the fastest detection speed, but its mAP and mR values are the smallest and its detection accuracy is low. YOLOv5s and Retinadet have balanced performance in each performance indicator. Although the detection speed of YOLOv5s reaches 66.76 FPS, its mAP and mR values have a large gap with YOLOv7. Retinadet is lower in both accuracy and speed than YOLOv7. YOLOv7 has the highest recognition accuracy, with at least 0.34% higher in terms of mAP than other algorithms.

TABLE 5. Results of ablation experiment.

Structure	Details
YOLOv7	The original YOLOv7 model
YOLOv7-I	YOLOv7 with conventional convolutions replaced by DS Convs
YOLOv7-II	YOLOv7 with CBAM modules added at the position shown in the structure CBAM-2
YOLOv7-III	YOLOv7 with Coordinate modules added at the position shown in the structure COR-1
YOLOv7-IV	YOLOv7 with conventional convolutions replaced by DS Convs, and CBAM modules added at the position shown in the structure CBAM-2
YOLOv7-V	YOLOv7 with conventional convolutions replaced by DS Convs, CBAM modules added at the position shown in the structure CBAM-2, and Coordinate modules added at the position shown in the structure COR-1

B. ABLATION EXPERIMENT OF IMPROVED YOLOV7

Improved YOLOv7 promoted the detection ability of YOLOv7 by adding DS Conv, CBAM attention modules, and CA modules. The details of different structures are shown in Table 5, and the results of the ablation experiment are shown in Table 6.

The mAP, mR, and FPS of the YOLOv7 model are 95.42%, 90.96%, and 34.59, respectively, which are respectively improved by 0.97%, 2.08%, and 1.81 after replacing conventional convolutions with DS Convs, optimizing the detection performance in both speed and accuracy compared with the original model. The result reflects the efficiency of DS Conv. Compared to the result of YOLOv7, the mAP and mR are improved by 1.03% and 1.45% respectively after adding CBAM modules only. The mAP and mR are improved by 0.85% and 1.32% respectively after adding

TABLE 6. Results of ablation experiment.

	mAP0.5(%)	mR(%)	FPS
YOLOv7	95.42	90.96	34.59
YOLOv7-I	96.39	93.04	36.40
YOLOv7-II	96.45	92.41	33.87
YOLOv7-III	96.27	92.28	34.02
YOLOv7-IV	96.62	93.11	35.22
YOLOv7-V	96.70	93.88	30.62

Coordinate modules only. The addition of CBAM and Coordinate modules respectively leads to a certain amount of detection time cost, but both of them effectively enhance the feature extraction ability and detection accuracy of YOLOv7 model. According to the results of structure YOLOv7-I, YOLOv7-IV, and YOLOv7-V, based on the addition of DS Convs, the addition of CBAM attention modules in YOLOv7-IV improves the mAP by 0.23% and mR by 0.07%. CBAM modules reduce the background weight, optimize the efficiency of the network for tea bud feature information, improve the model's focus on the tea-bud regions and achieve more accurate localization. Finally, compared with YOLOv7-IV, the addition of the CA modules in YOLOv7-V improves the mAP and mR by 0.08% and 0.77%, respectively. Coordinate modules enhance the detection ability for tea buds of the model and improve the robustness for targets of different scales. Adding CBAM and Coordinate modules simultaneously results in better detection performance than adding them separately, the result indicates that the two modules working together can produce better performance, fully utilizing their advantages. The final improved YOLOv7 model has a 1.28% and 2.92% improvement in mAP and mR values over the YOLOv7 model, the final mAP and mR of the improved YOLOv7 reaches 96.70% and 93.88%, respectively. The detection speed slightly decrease with the FPS value of 30.62, which still meets the demand of real-time detection.

The AP and R values for each class are shown in Table 7 and Table 8, respectively. The YOLOv7 model has AP values higher than 93% and R values higher than 88% for all classes, the improved model has AP values higher than 95% and R values higher than 92% for all classes, showing that the model has high and balanced detection accuracy for all classes of targets. The AP and R values of Class 1 are the highest because Class 1 has the largest number (2717) and high similarity between individuals, and they are highly differentiated from other classes and less disturbed by the background.

1) COMPARISON EXPERIMENT BETWEEN DIFFERENT POSITIONS OF CBAM ATTENTION MODULES

The comparison results of adding CBAM modules at different positions based on replacing conventional convolutions with DS Convs are shown in Table 9. The effect is different when CBAM embedded at different positions.

TABLE 7. Comparison of ap values of all classes.

	1	2	3	c	mAP0.5(%)
YOLOv7	96.93	94.72	93.70	96.33	95.42
Improved YOLOv7	97.84	96.28	96.77	95.92	96.70

TABLE 8. Comparison of recall values of all classes.

	1	2	3	c	mR(%)
YOLOv7	94.53	88.73	90.00	90.61	90.96
Improved YOLOv7	96.45	94.91	92.11	92.06	93.88

Compared with YOLOv7-I, the addition of CBAM modules at CBAM-1 position respectively reduces mAP and mR by 0.01% and 0.35%, CBAM-2 position improves 0.23% and 0.07% respectively, CBAM-2 position has better performance. CBAM modules added between convolutional blocks extract features from both the preceding and following convolutional blocks, can more effectively improve the model's focus on feature information. The CBAM module is added to the feature fusion part of each scale to deeply utilize the feature information and further improve the detection performance. The channel attention module weights different channels of the feature map, giving more weights to the channels containing tea-shoot features, the spatial attention module weights different positions of the feature map giving more weights to the position features of tea shoots, which reduces the influence caused by complex background. The combination of the two modules improves the detection accuracy and Recall rate of the model, and the boundaries of its recognition boxes are more accurate.

2) COMPARISON EXPERIMENT BETWEEN DIFFERENT POSITIONS OF COORDINATE ATTENTION MODULES

On the basis of adding DS Conv and CBAM modules, the comparison results of adding Coordinate modules at different positions are shown in Table 10.

Compared with the mAP and mR of YOLOv7-IV, the COR-1 position improves mAP and mR respectively by 0.08% and 0.77%, the COR-2 position decreases mAP by 0.01% and improves mR by 0.77%, and the COR-3 position decreases mAP by 0.31% and improves mR by 0.78%. Although the COR-2 and COR-3 structure improve the Recall rate, the detection precision of the model decreases. COR-1 structure improve both mAP and mR values, COR-1 structure has the best performance. As shown in Fig. 4(a), Coordinate modules are embedded after the backbone output the small and medium-sized feature layers, which can improve the detection capability of the model for small and medium-sized

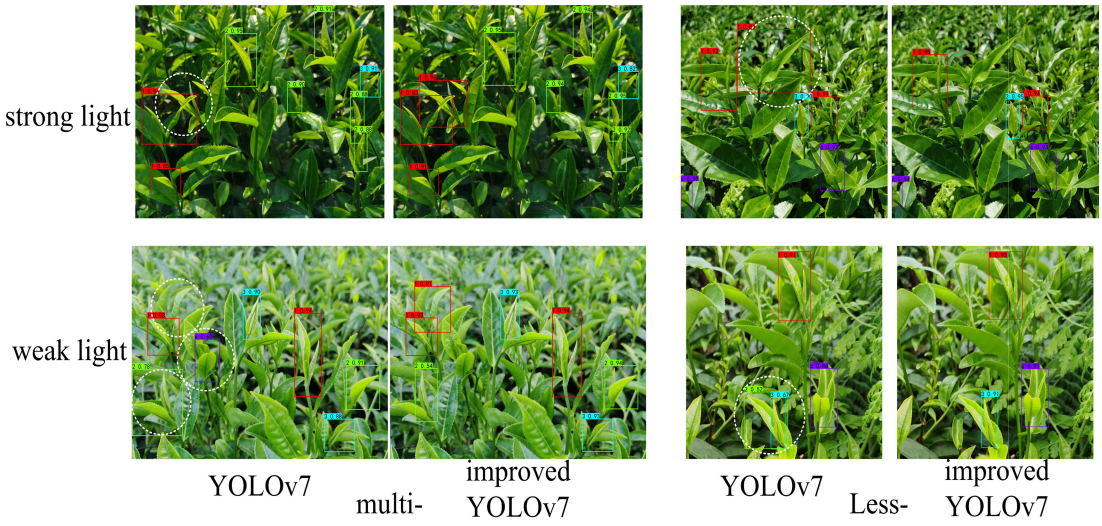

FIGURE 5. Detection results. The white circles are incorrectly detected targets.

TABLE 9. Comparison of CBAM attention at different positions.

	mAP0.5(%)	mR(%)
CBAM-1	96.38	92.69
CBAM-2	96.62	93.11

TABLE 10. Comparison of coordinate attention at different positions.

	mAP0.5(%)	mR(%)
COR-1	96.70	93.88
COR-2	96.61	93.88
COR-3	96.31	93.89

TABLE 11. Results under different lighting conditions.

	mAP0.5(%)	mR(%)
Strong light	97.20	94.74
Weak light	97.56	95.33

targets. The feature maps first undergo DS Conv, and then undergo pooling operation of Coordinate module to aggregate the features from horizontal and vertical directions with global perception and precise location. As a more complete information, the feature maps are input for concating in the neck part to further strengthen the network's ability to fuse feature information, the expression of the tea-bud regions is enhanced, and the detection accuracy of the model was more effectively improved.

C. COMPARISON OF DIFFERENT LIGHTING ENVIRONMENTS

To verify that the improved YOLOv7 algorithm still has good detection performance under different lighting conditions, a comparison experiment was conducted for 2 lighting conditions. The results are shown in Table 11.

The mAP values are both higher than 97% and the mR values are both higher than 94% under two different lighting conditions. The difference in mAP is 0.36%, and the difference in mR is 0.59%. These results indicate that the detection performance of the improved model is not significantly affected by the lighting conditions, and it maintains high detection accuracy under different lighting conditions, without a significant difference in detection ability.

D. DETECTION RESULTS OF TEA BUDS

The tea bud samples used in most of the current studies [21], [22], [23], [24], [25], [26], [27], [28] had obvious color differences from the background, which made it easier to identify and segment, and their growth stages were all in a fully matured and developed state. The background of the Fuyun 6 dataset in this study was complex, and the tea buds were close to the background in color, which was difficult to distinguish. Our tea buds dataset included all growth stages from tender to mature, and the tea buds were in different postures with cases of lateral postures, the dataset used in this study is closer to the real working conditions of Fuyun 6 tea picking.

Some of the detection results are shown in Fig. 5. The background in the images are complex, the color of tea buds is not significantly distinguished from the background, and the tea buds are of different sizes, unevenly distributed, and obscured from each other. For the multi-target image under strong light, YOLOv7 misses an obscured Class 1 target, which is accurately identified by the improved model. For the less-target image under strong light, YOLOv7 mistakenly detects a weed as a Class 1 target, because the weed is V-shaped bifurcation, shaped like developed one bud with one leaf, the improved model correctly rejects the weed. For the multi-target image under weak light, YOLOv7 mistakenly detects an old leaf as a Class c target, the box of a Class 2 target is not accurate in which there is an old leaf, and misses an obscured Class 1 target, the improved model identifies and

correctly frames these three targets. For the less-target image of under weak light, YOLOv7 misidentifies a background as a Class 2 target because an old leaf is located right behind a Class 3 target, resembling a spreading one bud with one leaf, the improved model accurately identifies the background. The results show that the improved model has improved the detection performance compared with YOLOv7 for images with different number of targets under different lighting conditions. The improved YOLOv7 is able to identify all targets and classify them correctly with greater detection ability.

V. CONCLUSION

This article focuses on one bud with one leaf targets of the Fuyun 6 tea species. Due to the high similarity in color between the tea bud with the background, the inconsistency of tea bud growth stages, and the difference in target appearance between the initial and mature stages, the difficulty of recognition is increased.

To solve this problem, we proposed an improved YOLOv7 model for detecting one bud with one leaf. Compared with other detection algorithms, including Retinadet, Efficientdet, SSD, Faster RCNN, and YOLOv5s algorithms, YOLOv7 performs the best on our dataset. In addition, this article enhances the YOLOv7 network structure by adding DS Convs, CBAM attention modules, and CA modules. The CBAM and Coordinate modules were added into different positions in order to explore the best detection performance. Through comparative experiments on different embedding positions of attention mechanisms, the results indicate that adding the CBAM attention modules before the DS Convs and MPConv layers, and adding the CA modules after the DS Convs in the neck part, can effectively improve the detection performance. The improved YOLOv7 model meets the requirements of real-time detection, and the mAP and mR reach 96.70% and 93.88%, respectively. The improved YOLOv7 model achieves a balanced detection effect under two lighting conditions with high accuracy. By comparing the detection results, compared with YOLOv7, the improved YOLOv7 model has better detection ability for images with multiple or few targets under both strong and weak light.

This article has a limited dataset, the research object is only Fuyun 6 tea species. Subsequent research can incorporate datasets of more tea species to enrich the robustness of this method for other teas. This article only studies the one bud with one leaf targets, because this type of targets has higher economic value. In future research, single bud and one bud with two leaves can be included, so that the model can recognize and classify tea buds at different levels.

REFERENCES

[1] A. Shevchuk, L. Jayasinghe, and N. Kuhnert, "Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data," *Food Res. Int.*, vol. 109, pp. 387–402, Jul. 2018, doi: 10.1016/j.foodres.2018.03.059.

[2] J. Wang, X. Li, G. Yang, F. Wang, S. Men, B. Xu, Z. Xu, H. Yang, and L. Yan, "Research on tea trees germination density detection based on improved YOLOv5," *Forests*, vol. 13, no. 12, p. 2091, Dec. 2022, doi: 10.3390/f13122091.

[3] R. Z. Zhu, "Structural design and optimization of famous tea picking robot," M.S. thesis, Dept. S.E, Jiangxi Agric. Univ., Jiangxi, China, 2022.

[4] M. T. Chen, "Recognition and location of high-quality tea buds based on computer vision," M.S. thesis, Dept. Tech., Qingdao Univ. Sci. Tech., Qingdao, China, 2019.

[5] H. K. Xia, B. G. Shi, and H. X. Huang, "Application progress of image processing in intelligent picking of tea sprouts," *Anhui Agric. Sci. Bull.*, vol. 25, no. 9, pp. 133–134, May 2019, doi: 10.16377/j.cnki.issn1007-7731.2019.09.054.

[6] M. Ren, B. Zhao, G. Zhao, and X. Chen, "Quality of green tea from different leaf position on new shoots and its influence factors," *Guizhou Agric. Sci.*, vol. 38, no. 12, pp. 77–79, Dec. 2010, doi: 10.1016/S1002-0160(10)60014-8.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[10] S. F. Zhang, "Recent advances of object detection using CNNs: An overview," *J. Nanjing Univ. Posts Telecommun.*, vol. 39, no. 5, Jul. 2019, doi: 10.14132/j.cnki.1673-5439.2019.05.010.

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[12] D. Patel, S. Patel, and M. Patel, "Application of image-to-image translation in improving pedestrian detection," 2022, *arXiv:2209.03625*.

[13] "Böbrek hastalýklarýiçin acýklanabilir yapay zeka destekli derin ören-meye dayalýbir tespit ve tahmin modeli," *Eur. J. Sci. Tech.*, vol. 40, pp. 67–74, Sep. 2022, doi: 10.31590/ejosat.1171777.

[14] M. Hussain, H. Al-Aqrabi, M. Munawar, R. Hill, and T. Alsboui, "Domain feature mapping with YOLOv7 for automated edge-based pallet racking inspections," *Sensors*, vol. 22, no. 18, p. 6927, Sep. 2022, doi: 10.3390/s22186927.

[15] L. Zhang, H. Zhang, Y. Chen, S. Dai, X. Li, I. Kenji, Z. Liu, and M. Li, "Real-time monitoring of optimum timing for harvesting fresh tea leaves based on machine vision," *Int. J. Agric. Biol. Eng.*, vol. 12, no. 1, pp. 6–9, Jan. 2019, doi: 10.25165/j.ijabe.20191201.3418.

[16] L. Zhang, L. Zou, C. Wu, J. Chen, and H. Chen, "Locating famous tea's picking point based on Shi-Tomasi algorithm," *Comput., Mater. Continua*, vol. 69, no. 1, pp. 1109–1122, 2021, doi: 10.32604/cmc.2021.016495.

[17] X. Wu, X. Tang, F. Zhang, and J. Gu, "Tea buds image identification based on lab color model and K-means clustering," *J. Chin. Agric. Mech.*, vol. 36, no. 5, pp. 161–164&179, Sep. 2015, doi: 10.13733/j.jcam.issn.2095-5553.2015.05.040.

[18] X. X. Sun, "The research of tea buds detection and leaf diseases recognition based on deep learning," M.S. thesis, Shandong Agric. Univ., Shandong, China, 2019.

[19] Y. T. Chen and S. F. Chen, "Localizing plucking points of tea leaves using deep convolutional neural networks," *Comput. Electron. Agric.*, vol. 171, Apr. 2020, Art. no. 105298, doi: 10.1016/j.compag.2020.105298.

[20] J. Tian, H. Zhu, W. Liang, J. Chen, F. Wen, and Z. Long, "Research on the application of machine vision in tea autonomous picking," *J. Phys., Conf.*, vol. 1952, Jun. 2021, Art. no. 022063, doi: 10.1088/1742-6596/1952/2/022063.

[21] Z. Qingqing, L. Lianzhong, N. Jingming, W. Guodong, J. Zhaohui, L. Mengjie, and L. Dongliang, "Tea buds recognition under complex scenes based on optimized YOLOv3 model," *Acta Agric. Zhejiangensis*, vol. 33, no. 9, pp. 1740–1747, Sep. 2021, doi: 10.3969/j.issn.1004-1524.2021.09.18.

[22] X. Wang, C. Han, W. Wu, J. Xu, Q. Zhang, M. Chen, Z. Hu, and Z. Zheng, "Fundamental understanding of tea growth and modeling of precise tea shoot picking based on 3-D coordinate instrument," *Processes*, vol. 9, no. 6, p. 1059, Jun. 2021, doi: 10.3390/pr9061059.

[23] H. Yang, L. Chen, M. Chen, Z. Ma, F. Deng, M. Li, and X. Li, "Tender tea shoots recognition and positioning for picking robot using improved YOLO-v3 model," *IEEE Access*, vol. 7, pp. 180998–181011, 2019, doi: 10.1109/ACCESS.2019.2958614.

[24] C. Chen, J. Lu, M. Zhou, J. Yi, M. Liao, and Z. Gao, "A YOLOv3-based computer vision system for identification of tea buds and the picking point," *Comput. Electron. Agric.*, vol. 198, Jul. 2022, Art. no. 107116, doi: 10.1016/j.compag.2022.107116.
[25] L. Yan, K. Wu, J. Lin, X. Xu, J. Zhang, X. Zhao, J. Tayor, and D. Chen, "Identification and picking point positioning of tender tea shoots based on MR3P-TS model," *Frontiers Plant Sci.*, vol. 13, Aug. 2022, Art. no. 962391, doi: 10.3389/fpls.2022.962391.
[26] W. Xu, L. Zhao, J. Li, S. Shang, X. Ding, and T. Wang, "Detection and classification of tea buds based on deep learning," *Comput. Electron. Agric.*, vol. 192, Jan. 2022, Art. no. 106547, doi: 10.1016/j.compag.2021.106547.
[27] L. Jun, F. Mengrui, and Y. Qing, "Detection model for tea buds based on region brightness adaptive correction," *Trans. Chin. Soc. Agric. Eng.*, vol. 37, no. 22, pp. 278–285, Dec. 2021, doi: 10.11975/j.issn.1002-6819.2021.22.032.
[28] H. C. Zhu, "Tea bud detection based on faster RCNN network," *Trans. Chin. Soc. Agric. Mach.*, vol. 53, no. 5, pp. 217–224, May 2022, doi: 10.6041/j.issn.1000.-1298.2022.05.022.
[29] Z. W. Wang, "Design of intelligent tea picking robot for complex working conditions," M.S. thesis, Dept. M.E., Shandong Univ., Shandong, China, 2021.
[30] N. S. Keskar and R. Socher, "Improving generalization performance by switching from Adam to SGD," 2017, *arXiv:1712.07628*.
[31] Z. Tian, G. Zhang, Y. Liao, R. Li, and F. Huang, "Corrosion identification of fittings based on computer vision," in *Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Dublin, Ireland, Oct. 2019, pp. 592–597, doi: 10.1109/AIAM48774.2019.00123.
[32] W. Q. Zheng, Q. Shen, and D. F. Zheng, "Preliminary investigation in the picking rate of top quality fresh tea leaves by machine," *J. Mount. Agric. Biol.*, vol. 29, no. 4, pp. 304–307, Apr. 2010, doi: 10.15958/j.cnki.sdnyswxb.2010.04.012.
[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[34] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.
[35] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.
[36] C. Zhang, F. Kang, and Y. Wang, "An improved apple object detection method based on lightweight YOLOv4 in complex backgrounds," *Remote Sens.*, vol. 14, no. 17, p. 4150, Aug. 2022, doi: 10.3390/rs14174150.
[37] S. Hong, Z. Jiang, J. Zhu, Y. Rao, W. Zhang, and J. Gao, "A deep learning-based system for monitoring the number and height growth rates of Moso bamboo shoots," *Appl. Sci.*, vol. 12, no. 15, p. 7389, Jul. 2022, doi: 10.3390/app12157389.
[38] J. Yang, Z. Qian, Y. J. Zhang, Y. Qin, and H. Miao, "Real-time recognition of tomatoes in complex environments based on improved YOLOv4-tiny," *Trans. Chin. Soc. Agric. Eng.*, vol. 38, no. 9, pp. 215–221, May 2022, doi: 10.11975/j.issn.1002-6819.2022.09.023.
[39] M. Sun, "Moving object detection based on FPGA," Ph.D. dissertation, Dept. Tech., Beijing Jiaotong Univ, Beijing, China, 2012.

JUNQUAN MENG was born in Guangxi Zhuang Autonomous Region, China, in 1996. He received the B.S. degree from Beijing Forestry University, Beijing, China, in 2019, where he is currently pursuing the master's degree in forestry equipment and informatization. His research interests include computer vision and intelligent picking robot.

FENG KANG was born in Hebei, China, in 1981. He received the B.S., M.E., and Ph.D. degrees from China Agricultural University, Beijing, China, in 2003, 2006, and 2011, respectively. He has been a Professor with Beijing Forestry University, Beijing, since 2018. He is currently the Director of Forestry Machinery Branch and Forest Engineering Branch of Chinese Forestry Society. He has published more than 20 academic articles. His research interests include forestry intelligent equipment and artificial intelligence.

YAXIONG WANG received the B.S. degree from the Taiyuan University of Technology, Shanxi, China, in 2010, and the Ph.D. degree from Beijing Forestry University, Beijing, China, in 2018. He is currently a Lecturer with Beijing Forestry University. He has published eight SCI/EI academic articles. His research interests include forestry intelligent equipment and forest engineering.

SIYUAN TONG received the B.S. degree from Beijing Information & Technology University, Beijing. He is currently pursuing the Ph.D. degree with Beijing Forestry University. His research interests include intelligent pruning and target detection.

CHENXI ZHANG was born in Gansu, China, in 1998. He received the B.S. degree from Beijing Forestry University, Beijing, China, in 2020, where he is currently pursuing the master's degree in forestry equipment and informatization. His research interests include computer vision and target detection.

CHONGCHONG CHEN received the B.S. degree from Shijiazhuang Tiedao University, Hebei. He is currently pursuing the Ph.D. degree with Beijing Forestry University. His research interests include forestry intelligent equipment and artificial intelligence.