DMT: A model detecting multispecies of tea buds in multi-seasons

Taojie Yu[1], Jianneng Chen[1,2*], Zhiwei Chen[1], Yatao Li[1,2], Junhua Tong[1,2], Xiaoqiang Du[1,2]

(1. *School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China*;
2. *Key Laboratory of Transplanting Equipment and Technology of Zhejiang Province, Hangzhou 310018, China*)

Abstract: In China, tea products made from fresh leaves characterized by one leaf with one bud (1L1B) are classified as "Famous Tea", which has better taste and higher economic value, but suffers from a labor shortage. Aiming at picking automation, existing studies focus on visual detection of 1L1B, but algorithm validation is limited to a specific variety of tea sprouting in a certain harvest season at a certain location, which limits the engineering application of developed tea picking robots working in various natural tea fields. To address this gap, a deep learning model DMT (detecting multispecies of tea) based on YOLOX-S was proposed in this paper. The DMT network takes YOLOX-S as a baseline and adds ECA-Net to the CSP Darknet and FPN of YOLOX-S. The average precision (AP), precision, and recall of DMT are 94.23%, 93.39%, and 88.02%, respectively, for detecting 1L1B sprouting in spring; 93.92%, 93.56%, and 87.88%, respectively, for detecting 1L1Bsprouting in autumn. These experimental results are better than those of the five current object detection models. After fine-tuning the DMT network with another dataset composed of multiple tea varieties, the DMT network can detect 1L1B for different varieties of tea in multiple picking seasons. The results can promote the engineering application of picking automation of fresh tea leaves.

Keywords: tea buds, detection model, multispecies of tea, multi-season
DOI: 10.25165/j.ijabe.20241701.8021

Citation: Yu T J, Chen J N, Chen Z W, Li Y T, Tong J H, Du X Q. DMT: A model detecting multispecies of tea buds in multi-seasons. Int J Agric & Biol Eng, 2024; 17(1): 199–208.

1 Introduction

Tea is a traditional drink that originated in China, and it possesses excellent health and economic values[1-4]. According to the various modes of harvesting fresh tea leaves and postharvest processing, tea is usually classified as bulk tea or famous tea. Harvesting equipment for bulk tea adopts non-selective mechanized harvesting methods[5-7]. Famous tea is famous for its higher economic value and better taste, becoming a representative of high-quality tea across China[8]. The famous tea product is made of fresh tea leaves featuring one leaf with one bud (1L1B). Famous tea has a better taste but is difficult to pick. The picking process completely relies on hands. The aging of the population and the continuous transfer of rural labor to the cities has resulted in labor shortages for picking fresh leaves, hence, in turn, has restricted the development of the famous tea industry.

Recently, facing these challenges, researchers have been focusing on applying machine vision technology to identify the tender buds of the famous tea and subsequently accelerate the mechanization and automation of famous tea harvesting. Zhang et al.[9] used the Bayesian posterior probability criterion to monitor the buds of purple rose tea trees in real-time in April, and they estimated the best picking time for the buds accordingly. Chen et al.[10] used a binary search tree to describe the shape of 1L1B, and a support vector machine (SVM) classifier graded the quality of fresh tea leaves after harvesting, which minimized the mixing of old leaves and broken leaves with the fresh leaves of raw materials that are harvested using mechanical tea-picking machines. Lu et al.[11] proposed an improved Artificial Color Contrast/Principal Component Analysis (ACC/PCA) method to solve the impact of the changes in illumination on tea bud detection. Similarly, Zhang et al.[12] proposed an improved watershed function to suppress the adverse impact of different illumination on the segmentation of famous tea buds, and to obtain a better segmentation effect. Chen et al.[13] used Region-based Convolutional Neural Network (R-CNN) and Fully Convolutional Network (FCN) models to identify autumn tea with one bud and two leaves in November 2017 and July 2018 along with its picking points, and they determined the three-dimensional coordinates of the picking points. Li et al.[14] used YOLOV3 to identify the buds of autumn tea in September that were captured by an RGB-Depth (RGB-D) camera, and they estimated the 3D coordinates of the bud-picking points in the depth map corresponding to the bounding box. Yang et al.[15] used the SVM algorithm and the YOLOV3 target detection network to identify the tender buds of famous tea, and they completed the picking experiment on the Delta parallel manipulator. Xu et al.[16] proposed a deep learning model for tea bud detection, which maximizes the rapid detection capability of YOLOV3 and the high-precision classification capability of DenseNet201 to detect tea buds.

The above studies adopt a variety of classification methods in classic image processing such as SVM, watershed function, binary search, and quadtree[17,18]. They also use different deep learning models, such as R-CNN, YOLOV3, and FCN to improve the

Received date: 2022-11-06 **Accepted date:** 2023-07-15
Biographies: Taojie Yu, MS, research interest: deep learning and design of agricultural machinery, Email: 202130605347@mails.zstu.edu.cn; **Zhiwei Chen**, MS, research interest: design of agricultural machinery, Email: 934138476@qq.com; **Yatao Li**, PhD, research interest: robot 3D vision and agricultural robot intelligent equipment technology, Email: ytli@zstu.edu.cn; **Junhua Tong**, PhD, Associate professor, research interest: design and optimization of agricultural machinery, Email: jhtong@zstu.edu.cn; **Xiaoqiang Du**, PhD, Professor, research interest: design and optimization of agricultural machinery, Email: xqiangdu@zstu.edu.cn

***Corresponding author: Jianneng Chen**, PhD, Professor, research interest: agricultural machinery equipment and technology. Faculty of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China. Email: jiannengchen@zstu.edu.cn.

accuracy and recognition of tea bud detection. In China, the picking season of green tea starts in late March and ends in late August, including spring tea (from March to April), summer tea (from May to June), and autumn tea (from July to September). The main tea varieties include No. 43 Longjing, No. 108 Zhongcha, and Cuifeng and there are differences in the physical characteristics of fresh leaves between different seasons and tea varieties. Existing studies only involve a single tea variety in a specific picking season (or unspecified). Therefore, the detection effect of the models is unknown and the accuracy may be reduced for different picking seasons or different picking varieties. In addition, the size of the tea bud object is small, and the detection object is relatively close to the surrounding environment, so the deep learning model may not be applicable. Previous research used popular learning models in combination with classical digital image processing algorithms without directly improving the model structure, limiting the generalization ability of the detection model and hindering the integration of tea bud target detection to picking automation.

This study proposed an object detection model, the DMT network (detection of multispecies of tea), which can identify 1L1B of multispecies of famous tea during the three picking seasons of spring, summer, and autumn in the natural environment. The manuscript is organized as follows. First, the addition of the ECA-Net[19] lightweight attention module to the original YOLOX-S model is presented. Second, the use of the No. 43 Longjing tea dataset to train the DMT network to detect No. 43 Longjing tea for two picking seasons is evaluated. Finally, the MVT dataset is used to generalize the detection ability of the DMT network.

2 Materials and methods

2.1 Dataset generation

Common deep learning models are often adapted for public datasets, such as PASCAL VOC datasets (Figure 1a)[20]. The number of label boxes for this type of dataset image is generally only 1 to 3, and there is a significant difference between the target and the background features. The color, texture, and shape of tea buds in pictures taken in the natural environment are similar to those of old leaves, and there are usually 10-40 targets in a single picture, as shown in Figure 1b). Therefore, conventional deep-learning models that have achieved high-performance evaluations using various public datasets may not be suitable for identifying fresh tea leaves and buds. Thus, to detect the 1L1B of famous tea, it is necessary to create a training dataset and make targeted improvements to the existing deep-learning model.

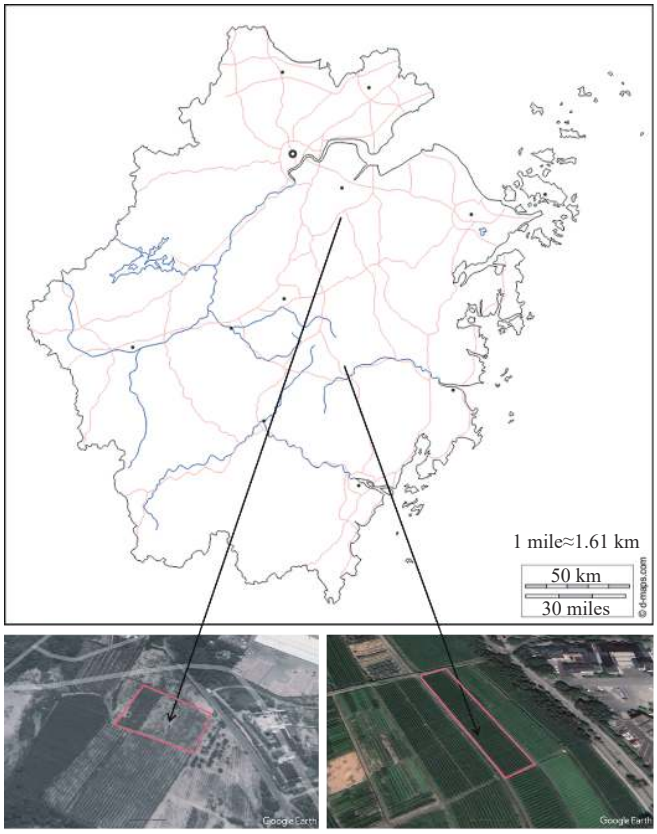
a. Several types of pictures are arbitrarily selected from the PASCAL VOC dataset and their annotation boxes

b. A picture of No. 43 Longjing tea with annotation boxes from the tea dataset

Figure 1 Comparison between common public datasets and homemade datasets

2.1.1 Image collection of tea buds

A total of 10 000 tea images were collected in April 2020, April 2021 (Spring), August 2020, and August 2021 (Autumn) at the Tea Research Institute of the Chinese Academy of Agricultural Sciences and the Shengzhou Tea Comprehensive Experimental Base of the Tea Research Institute of the Chinese Academy of Agricultural Sciences (Figure 2), to create an RGB image dataset for training and validation of the network model. The sampling devices include mobile phones (HUAWEI Mate30, iPhone12) and industrial cameras (ZIVID two).



a. Tea Research Institute of China Agricultural Sciences, located in Xihu District, Hangzhou City, Zhejiang Province, China (coordinates 30.185883°N, 120.099789°E). The red box shows the sampled tea fields

b. Shengzhou Tea Comprehensive Experimental Base, Tea Research Institute, Chinese Academy of Agricultural Sciences, located in Shengzhou City, Zhejiang Province, China (coordinates 29.748373°N, 120.825306°E). The red box shows the sampled tea fields.

Note: 1 mile≈1.61 km.

Figure 2 Tea bud image collection locations

Figure 3a is an image of the fresh leaves of No. 43 Longjing tea taken in spring. There are a few obvious differences in size and color between the tea buds and the surrounding old leaves. Figure 3b is an image of the fresh leaves of No. 43 Longjing taken in autumn. Compared to the tea buds of No. 43 Longjing in spring, the color of the buds in autumn is darker, and the texture is easily distinguishable. This dataset is used as the main training dataset and is referred to as the LJ43 dataset (No. 43 Longjing dataset).

A total of 1000 pictures of Cuifeng tea, 1000 pictures of No. 43 Longjing, and 1000 pictures of No. 108 Zhongcha were collected as a multispecies small dataset named MVT dataset (Multi-Variety Tea dataset), as shown in Figure 4. The MVT dataset is used to examine the generalization ability of the model trained on the LJ43 dataset. The variety, quantity, sampling time, and sampling location of the original tea images in the LJ43 dataset and the MVT dataset are presented in Table 1 and Figure 2.

a. Compared with the surrounding old leaves, the young buds of spring tea have a lighter color and smoother texture

b. Tender buds of autumn tea are darker in color and rougher in texture than those of spring tea

Figure 3 Representative images of tea buds of No. 43 Longjing tea in spring and autumn

a. No. 43 Longjing tea photo taken in May 2020

b. Cuifeng green tea photo taken in August 2020

c. No. 108 Zhongcha tea photo taken in March 2022

Figure 4 Three different tea buds in the MVT dataset

Table 1 Properties of generated datasets

Dataset	Tea varieties	Sampling time	Number
LJ43 Dataset	No. 43 Longjing	April 2020 April 2021	7000
	No. 43 Longjing	August 2020 August 2021	3000
MVT dataset	No. 43 Longjing	May 2020	1000
	No. 108 Zhongcha	March 2022	1000
	Cuifeng	August 2020	1000

No. 43 Longjing tea has less fuzz and the root of the bud shows light red. Cuifeng green tea has a lot of fuzz and slender dark green buds. No. 108 Zhongcha tea has less fuzz and is green with a hint of yellow buds. On one hand, these different phenotypic characteristics will hinder the detection of different tea varieties by deep learning models, on the other hand, deep learning models trained with multi-tea varieties data sets tend to have better generalization.

2.1.2 Dataset annotation

Both the LJ43 dataset and MVT dataset use the PASCAL VOC dataset format. Tea classified as famous tea requires 1L1B as raw materials. A labeling software was used to label all the areas of 1L1B in the original image. The label boxes are classified into two categories: spring tea and autumn tea. The tea buds are significantly different in terms of posture (Figure 5), degree of occlusion (Figure 6), and clarity (Figure 7).

2.1.3 Data augmentation

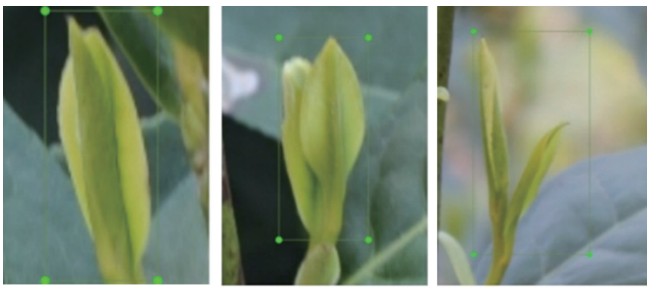
The recognition accuracy can be improved by increasing the number of pictures and data diversity in the dataset. Various image processing methods were randomly combined to increase or decrease the brightness, contrast, and color saturation, flip the image horizontally, and adjust the aspect ratio of the original tea pictures. Finally, the original 10 000 images of No. 43 Longjing tea were expanded to a sample size of 60 000 images in the LJ43 dataset. A random combination application process of the data augmentation method on the LJ43 tea dataset is shown in Figure 8.

a. Bud axis is inclined to the right (more common in the right half of the tea tree)

b. Bud axis vertical (more common on the treetop)

c. Bud axis is inclined to the left(more common in the right half of the tea tree)

Figure 5 Detected 1L1B with different postures in the LJ43 dataset

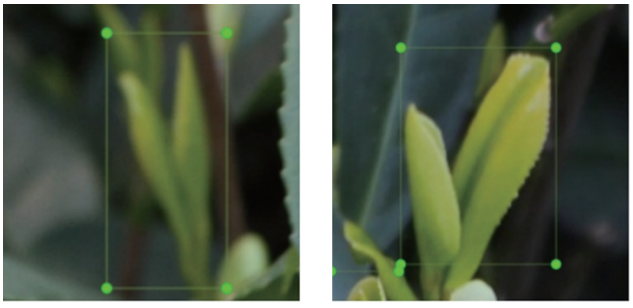


a. Complete occlusion between tea bud and tea leaf

b. Partial occlusion between tea bud and tea leaf

c. No occlusion between the tea bud and tea leaf

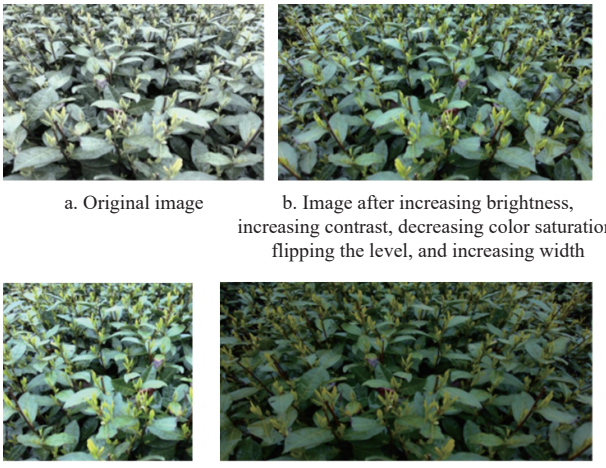
Figure 6 Detected 1L1B with different degrees of occlusion in the LJ43 dataset



a. Fuzzy tea bud

b. Clear tea bud

Figure 7 Detected 1L1B with different degrees of ambiguity



a. Original image

b. Image after increasing brightness, increasing contrast, decreasing color saturation, flipping the level, and increasing width

c. Image after increasing brightness, decreasing contrast, decreasing color saturation, and decreasing width

d. Image after decreasing brightness, lowering contrast, increasing saturation, flipping level, and increasing width

Figure 8 Random data augmentation

2.2 Construction of DMT network

2.2.1 Basic structure of Network

YOLOX-S is a YOLO (you only look once) series network model developed by MEGVII Technology (Beijing, China) consisting of CSPDarknet, FPN, and YOLO HEAD[21].

In the YOLOX-S model, the RGB image of size 640×640×3 is first input to the Focus module (Figure 9). The Focus module selects the spaced pixels in an image to form four independent feature layers with the same number of channels and an area size that is reduced by a factor of 1/4 from the original image. These four independent feature layers are then superimposed to obtain a feature layer with a size of 320×320×12.

CSPDarknet adopts the structure of SPPBottleneck and CSPNet (Figure 10), which amplifies the difference in gradient joint, eliminates the repeated information that is learned by different network layers, and effectively extracts the depth of tea buds in the changeable open-air tea garden environment information while increasing only a small amount of model volume. Once the CSPDarknet extracts the depth image information of the tea bud, the model will output the depth feature information of the tea bud with sizes of 80×80×256, 40×40×512 and 20×20×1024 for two CSP2 layers, and one CSP3 layer to the FPN layer, respectively, to achieve feature enhancement extraction.

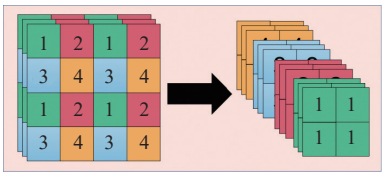
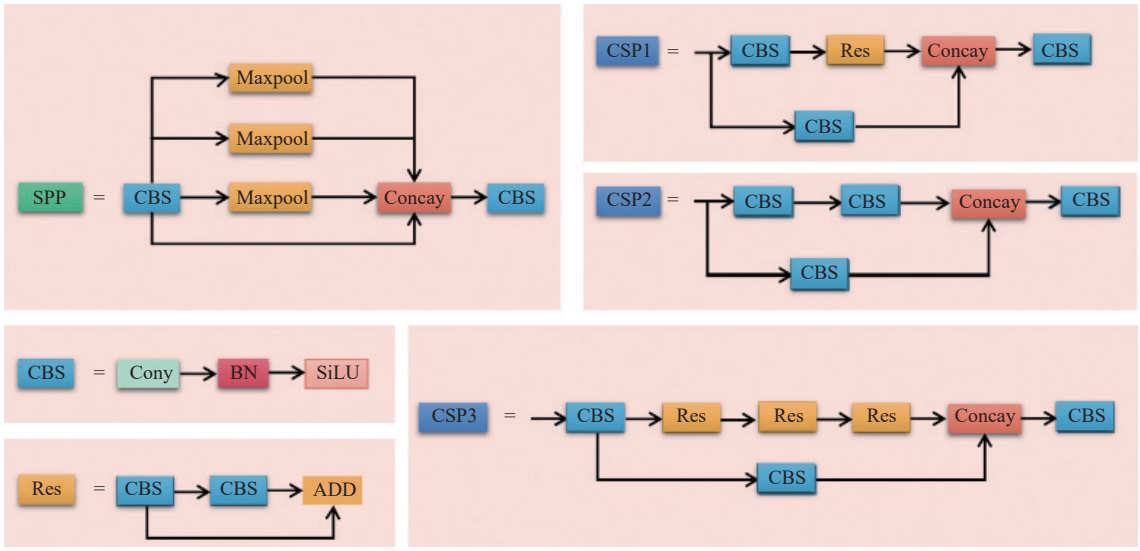


Figure 9 Schematic of Focus module



Note: CBS: Conv+BN+SiLU; Res: Residual Block; CSP: CSPDarknet; BN: Batch Normalization; SiLU: Sigmoid Weighted Liner Unit; ADD: Element-wise add. Same below.

Figure 10 Schematic of the specific structure of SPPBottleneck, CBS, Res, and three types of CSP

In the YOLOX-S, the CSPNet structure is used in the FPN layer. Further, the multi-feature extraction layer is designed as a pyramid, and this can help fuse the feature layers of different shapes and numbers of channels and subsequently result in enhanced feature extraction. The FPN layer of YOLOX-S can not only perform efficient feature extraction for the relatively complex and changeable background as well as the target in the LJ43 dataset but also take into account the characteristics of the top and bottom layer feature information, which helps the YOLOX-S predict and detect the position of small targets. The obtained accuracy is excellent making the model suitable for use in the study of the LJ43 dataset and the MVT dataset.

As shown in Figure 11, in the YOLO-HEAD section, a decoupled head, which has not been used in the YOLO series before, is used. The decoupled head does not use the same depth feature information for detection classification and target box regression for prediction, but it outputs three prediction results for each feature layer. It further predicts the detection classification and target box regression results separately and stacks them accordingly. The deep feature information of the classification branch is more concise, and thus, it is suitable for classification tasks in cases that involve only two types of labels. There are more contour boundary features in the depth feature information of the regression branch, and this factor can help distinguish the difference in shape between old leaves and tea buds, and in doing so, reduce the possibility of false detection.

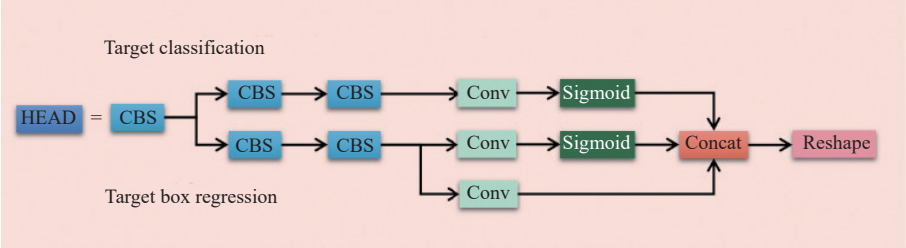


Figure 11 Schematic of YOLO HEAD module

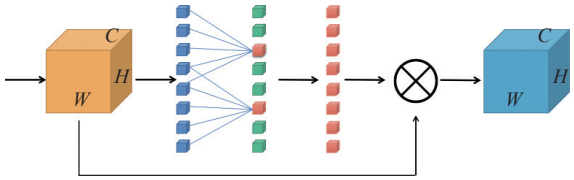
In terms of anchor box selection, YOLOX-S applies the anchor-free method. In the anchor-based algorithm, to obtain a more suitable anchor frame value, it is often necessary to perform clustering algorithm processing on the annotation frame of the dataset before training. However, this may result in an increase in the complexity of the detection head and the number of generated results. The algorithm of the anchor-free method is simple, and it is suitable for datasets that require intensive prediction, such as the LJ43 dataset and the MVT dataset.

Based on the comprehensive analysis of the network structure and characteristics, this study uses YOLOX-S as the basic model and makes targeted improvements to YOLOX-S to ensure that it caters to small-sized targets.

2.2.2 Addition of attention mechanism

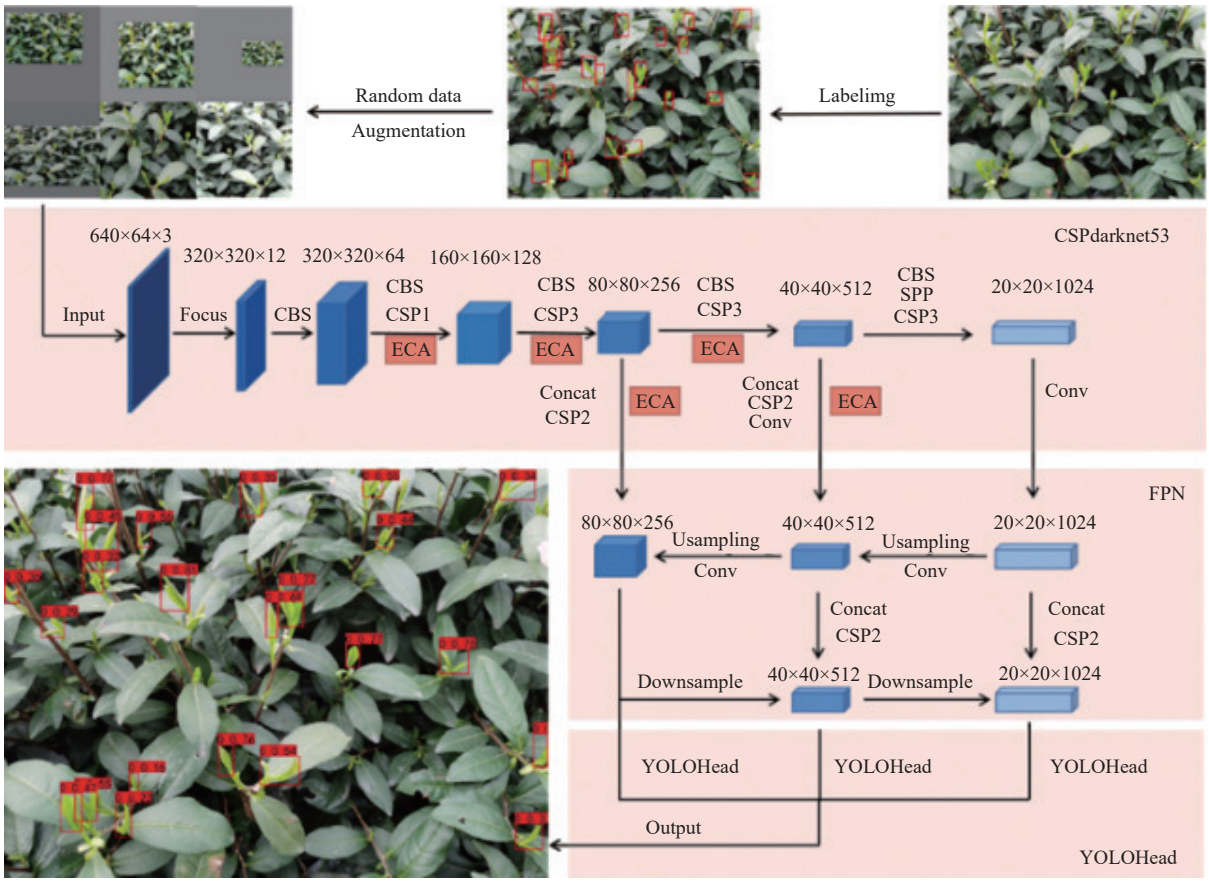
An attention mechanism is a good approach for improving the detection ability of the target detection network.

Based on a comprehensive analysis of the literature, the ECA-Net attention mechanism was added to YOLOX-S for model optimization[20]. The ECA-Net channel attention mechanism, shown in Figure 12, can yield excellent detection results based on the addition of only a few parameters. ECA-Net uses a 1×1 convolutional layer directly after the global average pooling layer and thus removes the fully connected layer. The ECA-Net module does not reduce the information dimensions, and therefore, the DMT network can learn the top-level information more effectively and improve the tea bud position prediction ability. At the same time, the DMT network can effectively capture cross-channel interaction information, and it can also extract the middle and bottom information, thus improving the detection ability for small targets such as tea buds. Therefore, the ECA-Net attention mechanism is introduced in the Backbone and RPN stages to enhance the feature extraction effect; the overall structure of the DMT network is shown in Figure 13.



Note: W is the width of the image; H is short for the height of the image; C is short for the number of channels of the image. The blue square is the original channel, the green square is the channel without change, and the red square is the new channel formed by multiple blue channels after local cross-channel information exchange.

Figure 12 ECA-Net structure



Note: Compared to the original CSP Darknet, the DMT network adds one ECA-Net each after the first and second CSP layers in CSP Darknet. Compared to the original FPN layer, the DMT network adds ECA Net between the second and third CSP layers in CSP Darknet and the Concat module of FPN, the fourth in CSP Darknet and the first in FPN layer. ECA-Net module is added between the CBS modules.

Figure 13 Overall detection flow chart of DMT network

2.3 Experiment environment and evaluation index

2.3.1 Hardware and software environment

All experiments use a computer with an Intel i7 12700 CPU, NVIDIA 3080ti GPU, and 32 GB of memory. The computer runs Ubuntu 18.04, Python 3.8, and PyTorch 1.11.

2.3.2 Evaluation metrics of deep learning models

This study used three commonly used indicators in the field of machine learning to evaluate the target detection effect, such as

Equations (1)-(3).

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

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

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

where, P is the precision; R is the recall; TP denotes True Positive, the true category is positive, the predicted category is positive; FP denotes False Positive, the true category is negative, the predicted category is positive; FN denotes False Negative, true class is positive, predicted class is negative.

Due to the high quality of famous tea raw materials, it is more unacceptable to detect unqualified leaves as qualified tea buds than to omit qualified tea buds. Therefore, precision and AP are more important than recall.

2.4 Model training

2.4.1 Model training parameter settings

The LJ43 dataset which consists of 60 000 images is classified as a training set and a validation set according to a ratio of 9:1. The training parameters are set as follows: the batch size of 161 200 epochs, the initial learning rate of 0.02, and minimum learning rate of 0.0002. Stochastic gradient descent (SGD) with a momentum of 0.937 and weight decay of 0.000 05 as well as cosine annealing learning is adopted to optimize the training effect.

2.4.2 Adding pre-trained model

The convergence speed during model training can be improved by avoiding local optimum points or saddle points and by increasing the generalization ability of the model. The PASCAL VOC dataset was used for pre-training on the DMT network before training with the LJ43 dataset. A comparison of the YOLOX-S model and the pre-trained model performance is shown in Figure 14, which indicates the favorable impact of the addition of a pre-training model on the DMT network. At the same time, adding a pre-training model under the same hyperparameters ensures that the DMT network converges more quickly during training.

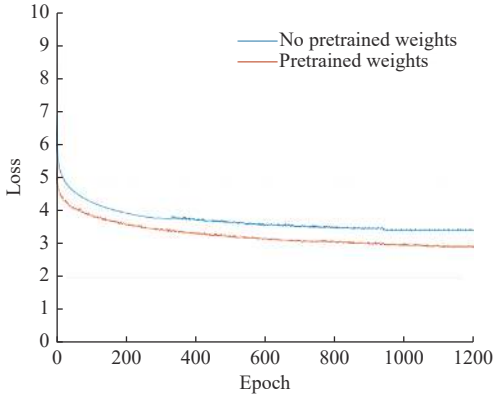


Figure 14 Loss curves of the validation set to check whether pre-training is required

2.5 Experimental methods

2.5.1 Comparison of attention mechanisms

The impact of different attention mechanisms on the model is studied to verify the superiority of the ECA-Net attention mechanism over the YOLOX-S model. Without adding gain methods such as pre-training models, three different attention mechanisms are added to the YOLOX-S model to evaluate the overall improvement. The three different attention mechanisms are convolutional block attention module (CBAM)[22], squeeze and excitation networks (SE-Net)[23], and ECA-Net.

2.5.2 Mosaic data augmentation method

The mosaic data augmentation method is enabled by default in the YOLOX model. The mosaic method is implemented as follows: first, four images are randomly selected from the training set, and image flipping, image scaling, color gamut adjustment, and other image adjustment operations are performed on these images. The four images are then spliced with target frames to generate a new image with a complex background and a higher number of target frames. On most public datasets, the mosaic data enhancement method enriches the background of the detected objects, which can be equivalent to detecting four pictures at the same time during the match normalization calculation. This further enriches the dataset and saves computing resources, thus resulting in a significant performance improvement. The results of the mosaic data augmentation method on the LJ43 dataset are shown in Figure 15. It is observed that some target boxes in the image become very small, and the background of the image becomes more complex. The impact of these factors on the detection performance of the DMT network is not known. Therefore, on the LJ43 dataset, other parameters were fixed, and experiments were performed with mosaic data augmentation enabled and mosaic data augmentation disabled.

Figure 15 Mosaic data enhanced image effect

2.5.3 Comparison of popular target detection models

To verify the applicability of the DMT network to the target detection of tea bud in spring and autumn in the natural environment, the proposed DMT network was compared with the five popular target detection models (YOLOX-S, YOLOV5-S, YOLOV4, Faster-RCNN, and SSD)[24-26].

2.5.4 Target detection generalization ability

As listed in Table 2, the study created an MVT dataset of 3000 images to examine the generalization ability of YOLOX-S after the addition of the attention mechanism. A dataset consisting of 1000 images could be considered “few-shot learning”. Three methods are commonly used for few-shot learning: model fine-tuning, data augmentation, and transfer learning. Among these methods, transfer learning requires high computing power, and it takes a long time to train. For the data enhancement method, if the number of enhanced pictures is significantly small, the accuracy improvement will be small; if the number of enhanced pictures is too large, it will also require substantial training time. Therefore, this study adopts the DMT network to train the LJ43 dataset as a pre-training model and uses the MVT dataset for model fine-tuning. The LJ43 dataset generated by a large number of Longjing 43 tea images can greatly improve the detection of most green tea varieties in different picking seasons. In detecting other varieties of green tea and famous

tea varieties, because there are fewer samples, the data fine-tuning method can achieve better detection results in a short time.

Table 2 Comparison of quantity and proportion of spring tea and autumn tea between original images and LJ43 dataset

Item	Original image	LJ43 dataset
Spring tea	7000	30 000
Autumn tea	3000	30 000
Proportion	7:3	1:1

3 Results and discussion

3.1 Attention mechanism comparison results

Compared to the original model, the addition of the three attention mechanisms provided a significant and positive impact on the performance of the YOLOX-S model. As listed in Table 3, among these models, DMT shows the best AP and precision in detecting spring tea and autumn tea. In the detection of spring tea and autumn tea, SE-YOLOX-S shows the best performance on recall. At the same time, the size of the DMT network only increased by 440 Params, while that of the SE-YOLOX-S and CABM-YOLOX-S increased by 166 200 and n Params, respectively. Therefore, ECA-Net is more suitable for the task of identifying famous tea buds compared to the CBAM and SE-Net.

Table 3 Recognition performance of DMT, YOLOX-S, and YOLOX-S with three attention mechanisms added

Model	Spring tea bud			Autumn tea bud			Total Params
	AP /%	Precision /%	Recall /%	AP /%	Precision /%	Recall /%	
DMT	91.58	92.74	81.06	90.92	89.81	85.81	8 968 695
SE-YOLOX-S	90.45	92.23	82.42	89.35	89.64	86.62	9 134 455
CBAM-YOLOX-S	90.06	91.96	82.63	90.92	89.23	84.91	9 019 565
YOLOX-S	89.63	90.88	82.46	89.27	87.96	83.83	8 968 255

3.2 Mosaic comparison experiment results

As shown in Table 4, after enabling the mosaic data augmentation method in the LJ43 dataset, the AP, precision, and recall of spring tea decreased by 1.72%, 0.53%, and 3.98%, respectively. The AP, precision, and recall of autumn tea decreased by 2.58%, 0.33%, and 3.77%, respectively. Contrary to the random combination data enhancement method adopted in Section 2.1.3, the image background that is generated using the Mosaic data enhancement method is extremely complex. In some public datasets, because of the single background of the detected object, the Mosaic data enhancement method is very effective. However, when the LJ43 dataset under study already has a fairly complex environmental background, the use of the Mosaic data enhancement method will result in an increase in the complexity of the background, subsequently affecting the detection effect. At the same time, because of Mosaic data enhancement, four pictures are spliced into one picture, which further increases the difficulty of small target detection, especially for tea bud detection.

Table 4 Tea bud detection performance of DMT with and without mosaic data augmentation method

Model	Spring tea bud			Autumn tea bud		
	AP/%	Precision/%	Recall/%	AP/%	Precision/%	Recall/%
No Mosaic DMT	94.23	93.39	88.02	93.92	93.56	87.88
Mosaic DMT	92.51	92.86	84.04	91.34	93.23	84.11

3.3 Comparison of popular target detection models

The YOLO series models have been widely used in various target detection tasks. This study compares YOLOX-S, YOLOV5-S, and YOLOV4 models with DMT; the results are listed in Table 5. Compared to the YOLOX-S model, the DMT network results in a 4.60% increase in AP value, a 5.56% increase in recall, and a 2.51% increase in precision value in the case of spring tea. For the autumn tea, the AP value increased by 4.65%, recall increased by 4.05%, and precision increased by 5.60%. An important factor behind the increase in these values is the addition of appropriate cross-channel interaction to the deep information learning of the target because of the use of the ECA-Net module, which improves the model's ability to acquire deep information. At the same time, the application of the anchor-free method to calculate the anchor box based on the prediction of the network also makes the YOLOX model and the improved YOLOX model more suitable for target detection tasks with complex backgrounds and small detection targets such as wild tea recognition. The DMT network is superior to existing YOLO series models in terms of most indicators.

Table 5 Detection performance results of DMT and commonly used target detection models

Model	Spring tea bud			Autumn tea bud			Total Params
	AP/%	Precision /%	Recall /%	AP/%	Precision /%	Recall /%	
No Mosaic Pre-training DMT	94.23	93.39	88.02	93.92	93.56	87.88	8 968 695
No Mosaic No pre-training DMT	91.58	92.74	81.06	90.92	89.81	85.81	8 968 695
Mosaic Pre-training DMT	92.51	92.86	84.04	91.34	93.23	84.11	8 968 695
SE-YOLOX-S	90.45	92.23	82.42	89.35	89.64	86.62	9 134 455
CBAM-YOLOX-S	90.06	91.96	82.63	90.92	89.23	84.91	9 019 565
YOLOX-S	89.63	90.88	82.46	89.27	87.96	83.83	8 968 255
YOLOV5-S	87.29	85.47	78.33	85.15	84.98	73.64	7 276 605
YOLOV4	84.84	80.89	75.61	79.89	80.15	74.42	64 363 101
Resnet50-Faster-RCNN	80.05	79.22	78.19	76.95	77.68	74.43	137 078 239
Vgg16-SSD	67.18	65.61	71.37	63.21	61.59	60.48	26 151 824

The performance of the DMT network is also significantly better than that of the Faster-RCNN target detection model. The main reason is that in Faster-RCNN, Resnet50 is used as the backbone, and its feature maps are only sourced from top-level features, which enables information that is not conducive to the positioning of the target frame. Because of the small target of the LJ43 dataset, and some targets are occluded and blurred, the prediction method that only uses the features of the top layer of the network to predict the target is not ideal for predicting the position of tea buds. As shown in Figure 16, when the tea bud is very small, the position prediction error causes inaccurate spatial positioning for the picking machinery, resulting in picking failure.

Note: A large number of errors indicate that the approximate position has been predicted, but the deviation between the specific position or the size of the prediction frame and the standard frame is too large to be used as a qualified prediction frame. Here, the blue and red boxes are the labeling and error prediction boxes, respectively.

Figure 16 Faster-RCNN and VGG16-SSD 300 error examples of tea bud prediction

The DMT network also outperforms VGG16-SSD 300 in terms of precision, AP, and recall. Unlike Faster-RCNN, VGG16-SSD 300 adopts a deep learning network with a multi-scale feature fusion method but does not fully utilize the features of lower layers, resulting in its inability to identify and detect small targets. This model is thus not applicable to the LJ43 dataset. At the same time, the resolutions of Faster-RCNN (512×512×3) and VGG16-SSD 300 (512×512×3) are also relatively small, which is not conducive to small targets such as tea buds.

3.4 DMT generalization ability on MVT dataset

The data indicators are shown in Table 6. For detecting No. 108 Zhongcha tea in March 2021, the AP, precision, and recall are 81.49%, 77.79%, and 71.41%, respectively, which are the lowest among the three MVT datasets. This is because the image collection method involved fixing a tarpaulin behind the tea buds for flash photography, which is different from the open-air shooting method in the LJ43 dataset, resulting in a large difference in the images. In future research, a variety of different image-shooting methods can be added to expand the diversity of the dataset.

Table 6 Tea bud detection performance of DMT using the MVT dataset

MVT dataset	AP/%	Precision/%	Recall/%
May 2020 No. 43 Longjing	83.43	80.37	76.12
August 2020 Cuifeng	84.24	82.32	75.33
March 2022 No 108. Zhongcha	81.49	77.79	71.41

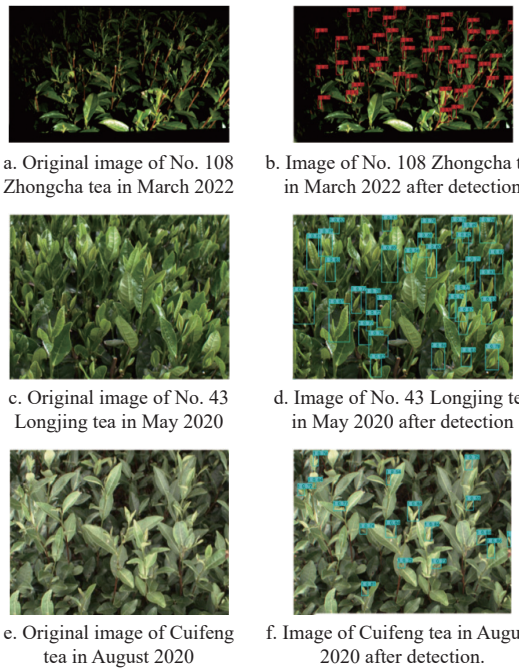
Compared to the recognition performance of DMT on No. 43 Longjing tea in April and No.43 Longjing in August, the AP, precision, and recall of No. 43 Longjing in May 2020 decreased by 10.8%, 13.02%, 11.9%, and 10.49%, 13.19%, 11.76%, respectively. Although there are two types of May data for No. 43 Longjing and the data originate from the same variety of tea trees, the characteristics of summer tea in May are slightly different from those of spring tea and autumn tea. In May, the tender bud leaves of No. 43 Longjing tea are larger and more relaxed compared to those of spring tea, and their color is similar to the surrounding old leaves, but different from autumn tea. Therefore, summer tea images can be added to the LJ43 dataset in subsequent studies to cover all harvest seasons of local tea. Here, only the method of training spring tea and autumn tea cannot accurately detect summer tea.

The detection performance for Cuifeng Green Tea in August 2020 was relatively good, and AP, precision, and recall were 84.24%, 82.32%, and 75.33%, respectively. This is because the characteristics of autumn Cuifeng Green Tea are similar to those of No. 43 Longjing, and the photographic sampling methods are also similar.

Figure 17 shows the original pictures of three different tea varieties in different picking seasons and at different times along with their detection results. A majority of the tea buds can be accurately identified, and this proves that the DMT network can be used on the MVT dataset. After fine-tuning, the DMT network has the ability to detect spring tea, summer tea, and autumn tea of different varieties of green tea in three picking seasons.

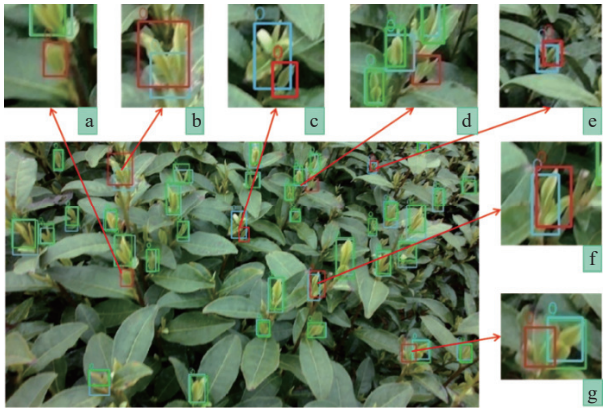
3.5 Analysis of error detection results

From the analysis of a detected image that is arbitrarily selected (Figure 18), some detection errors (red boxes) are observed in the image. The error detection type shown in Figure 18b accounts for the largest proportion of errors. During the production of the tea dataset, 20 or more tea buds often appear in a single image, and some tea buds are blurred. Moreover, some tea buds are located in corners, and some overlap with other tea buds. Such unfavorable situations make it impossible to completely select all tea buds that meet the standard during manual labeling. The detection of under-labeled shoots will result in an increase in the FP value, and this further affects precision and AP.



a. Original image of No. 108 Zhongcha tea in March 2022
b. Image of No. 108 Zhongcha tea in March 2022 after detection
c. Original image of No. 43 Longjing tea in May 2020
d. Image of No. 43 Longjing tea in May 2020 after detection
e. Original image of Cuifeng tea in August 2020
f. Image of Cuifeng tea in August 2020 after detection.

Figure 17 Original image of three types of tea with different varieties, different plantation locations, and different times, along with their detection results



Note: The blue box is the label box when making dataset, the green box is the prediction box that is judged to be correct after prediction, and the red box is the prediction box that is judged to be wrong after prediction. Boxes (a)-(d), and (g) are ignored in the labeling process but are actually the correct tea buds. Box (b) is labeled incorrectly but is detected correctly in the detecting process. Box (f) is judged as false detection because the detection box is out of range of label box.

Figure 18 Detection map of randomly selected single tea buds randomly selected various error boxes

3.6 Performance comparison between YOLOX-S and DMT network on the TinyPerson dataset

In order to further verify that the DMT network has a better detection effect than YOLOX-S on a small target dataset. In this study, the detection effect of YOLOX-S and DMT network was compared on a small target public dataset TinyPerson[27] as shown in Figure 19 . At the same time, in the research that proposed this dataset, it was tested using several common deep learning models, and the results are listed in Table 7.

Figure 19 Example of TinyPerson dataset

Table 7 Comparisons of AP_{50} of various models on TinyPerson from "Scale Match for TinyPerson Detection"

Detector	AP_{50}
FCOS	16.90
RetinaNet	30..82
DSFD	31.15
Adaptive RetinaNet	41.25
Adaptive FreeAnchor	41.36
Faster RCNN-FPN	43.55

The training parameters are set as follows: the batch size of 16, 500 epochs, the initial learning rate of 0.02, and the minimum learning rate of 0.0 002. Stochastic gradient descent (SGD) with a momentum of 0.937 and weight decay of 0.00 005 as well as cosine annealing learning is adopted to optimize the training effect. No additional data augmentation is used in training.

It can be seen from Table 8 and Figure 20, DMT has a higher AP than YOLOX-S. DMT network not only adds ECA-Net each after the first and second CSP layers in CSP Darknet but also adds ECA Net between the second and third CSP layers in FPN. Therefore, the DMT network has certain advantages in dealing with the problem that features of small target data sets are difficult to extract.

Table 8 Comparisons of AP_{50} on TinyPerson from this study

detector	Label varieties	AP_{50}
YOLOX-S	Sea_Person	23.30
	Earth_Person	28.67
DMT network	Sea_Person	27.46
	Earth_Person	34.55

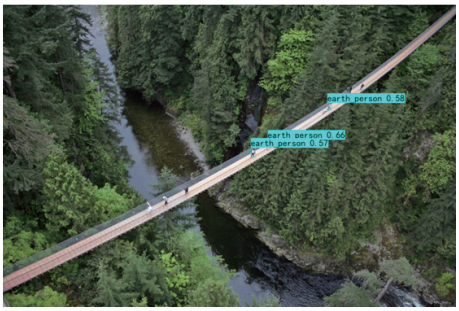
The detection effect of DMT and YOLOX-S in TinyPerson is not good. The reasons may be as follows:

1) When the original image size is (1280, 720) and the detection target is only dozens of pixels in size, the size of DMT and YOLOX-S is still (600, 600) when input. It would make detection targets more difficult to identify;

2) The number of images in TinyPerson is 793. After the 9:1 division of the training part and testing part, only 713 pictures participated in the training. It would lead to insufficient learning of data features by deep learning algorithms;

3) The training parameters used in training are not good enough parameters of DMT and YOLOX-S for the TinyPerson data set, and the data expansion method adopted in "Scale Match for TinyPerson Detection" is not adopted in this study.

a. Original image in the TinyPerson dataset, which shows 9 targets



b. Detection effect of the DMT network, which correctly detected 3 targets and missed 6 targets without error detection

c. Detection effect of YOLOX-S, which correctly detected 2 targets and missed 7 targets without error detection.

Figure 20 Comparisons of detection effect on TinyPerson

4 Conclusions

This study proposed an object detection network (DMT network) based on YOLOX-S to improve the generalization ability of the deep learning model on detecting multispecies of tea buds in multi-season. The DMT network integrates attention mechanism to realize cross-channel information interaction. Experiment results show that in the LJ43 dataset (single variety), the AP, precision, and recall of the DMT network are 94.23%, 93.39%, and 88.02%, respectively, for spring tea; and 93.92%, 93.56%, and 87.88%, respectively, for autumn tea. The results are better than other comparison models (YOLOX, YOLOV5-S, YOLOV4, Faster RCNN, and SSD). After fine-tuning the DMT network with the MVT dataset, the DMT network has the ability to detect buds of No. 43 Longjing tea in May, Cuifeng tea in August and No. 108 Zhongcha tea in March. The experimental results show that the model has good generalization ability and can be used for detecting multiple varieties of tea bud sprouting in multi-season. In future studies, ways to improve the network structure and add summer tea data to the dataset will be explored to further enhance generalization detection ability. This study could promote the engineering application of picking automation of fresh tea leaves.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (Grants No. U23A20175; No. 52305289); "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2022C02052); China Agriculture Research System of MOF and MARA and Basic. The authors also acknowledge the anonymous reviewers for their critical comments and suggestions for improving the manuscript.

[References]

[1] Chacko S M, Thambi P T, Kuttan R, Nishigaki I. Beneficial effects of green tea: A literature review. Chinese Medicine, 2010; 5(1): 13.

[2] Chan S P, Yong P Z, Sun Y, Mahendran R, Wong J C M, Qiu C, et al. Associations of long-term tea consumption with depressive and anxiety symptoms in community-living elderly: Findings from the diet and healthy aging study. The Journal of Prevention of Alzheimer's Disease, 2018; 5(1): 21–25.

[3] Zhu H K, Liu F, Ye Y, Chen L, Liu J Y, Gui A H, et al. Application of machine learning algorithms in quality assurance of fermentation process of black tea-based on electrical properties. Journal of Food Engineering, 2019; 263: 165–172.

[4] Wang Y J, Li L Q, Liu Y, Cui Q Q, Ning J M, Zhang Z Z. Enhanced quality monitoring during black tea processing by the fusion of NIRS and computer vision. Journal of Food Engineering, 2021; 304: 110599.

[5] Han Y, Xiao H R, Qin G M, Song Z Y, Ding W, Mei S, et al. Developing situations of tea plucking machine. Engineering, 2014; 6(6): 268–273.

[6] Tang Y P, Han W M, Hu A G, Wang W Y. Design and experiment of intelligentized tea-plucking machine for human riding based on machine vision. Transactions of the CSAM, 2016; 47(7): 15–20. (in Chinese)

[7] Zhao R M, Bian X B, Chen J N, Dong C W, Wu C Y, Jia J M, et al. Development and test for distributed control prototype of the riding profiling tea harvester. Journal of Tea Science, 2022; 42(2): 263–276. (in Chinese)

[8] Lee J E, Lee B J, Hwang J A, Ko K S, Chung J O, Kim E H, et al. Metabolic dependence of green tea on plucking positions revisited: A metabolomic study. Journal of Agricultural and Food Chemistry, 2011; 59(19): 10579–10585.

[9] Zhang L, Zhang H D, Chen Y D, Dai S H, Li X M, Kenji I, et al. Real-time monitoring of optimum timing for harvesting fresh tea leaves based on machine vision. Int J Agric & Biol Eng, 2019; 12(1): 6–9.

[10] Chen Z W, He L Y, Ye Y, Chen J N, Sun L, Wu C Y, et al. Automatic sorting of fresh tea leaves using vision-based recognition method. Journal of Food Process Engineering, 2020; 43(9): e13474.

[11] Lu J, Huang Y, Lee K M. Feature-set characterization for target detection based on artificial color contrast and principal component analysis with robotic tealeaf harvesting applications. International Journal of Intelligent Robotics and Applications, 2021; 5: 494–509.

[12] Zheng L, Zou L, Wu C Y, Jia J M, Chen J N. Method of famous tea bud identification and segmentation based on improved watershed algorithm. Computers and Electronics in Agriculture, 2021; 184: 106108.

[13] Chen Y-T, Chen S-F. Localizing plucking points of tea leaves using deep convolutional neural networks. Computers and Electronics in Agriculture, 2020; 171: 105298.

[14] Li Y T, He L Y, Jia J M, Lv J, Chen J N, Qiao X, et al. In-field tea shoot detection and 3d localization using an RGB-D camera. Computers and Electronics in Agriculture, 2021; 185: 106149.

[15] Stein E, Shakarchi R. Fourier analysis an introduction. Princeton University Press. 2002; 309p.

[16] Xu W K, Zhao L G, Li J, Shang S Q, Ding X P, Wang T W. Detection and classification of tea buds based on deep learning. Computers and Electronics in Agriculture, 2022; 192: 106547. doi:c10.1016/j.compag 2021.106547.

[17] Gill G S, Kumar A, Agarwal R. Monitoring and grading of tea by computer vision - A review. Journal of Food Engineering, 2011; 106(1): 13–19.

[18] Laddi A, Sharma S, Kumar A, Kapur P. Classification of tea grains based upon image texture feature analysis under different illumination conditions. Journal of Food Engineering, 2013; 115(2): 226–231.

[19] Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M, Hu Q H. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle: IEEE, 2020; pp.11531–11539. doi: 10.1109/CVPR 42600.2020.01155.

[20] Everingham M, Eslami S, Gool L V, Williams C, Winn J, Zisserman A. Assessing the significance of performance differences on the PASCAL VOC challenges via bootstrapping. Technical Note, 2013; pp.1–4.

[21] Ge Z, Liu S T, Wang F, Li Z M, Sun J. YOLOX: Exceeding YOLO Series in 2021. arXiv e-Print archive, 2021. arXiv: 2107.08430.

[22] Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional Block Attention Module. arXiv e-Print archive, 2018. arXiv: 1807.06521.

[23] Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020; 42(8): 2011–2023.

[24] Liu W, Anguelov D< Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: Single Shot MultiBox Detector. In: Computer Vision - ECCV 2016, Springer, 2016; 9905: 21–37. doi: 10.1007/978-3-319-46448-0_2.

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

[26] Bochkovskiy A, Wang C Y, Liao H. YOLOv4: Optimal speed and accuracy of object detection. arXiv e-print archive, 2020; arXiv: 2004.10934.

[27] Yu X H, Gong Y Q, Jiang N, Ye Q X, Han Z J. Scale match for tiny person detection. In: 2020 IEEE Winter Conference on Application of Computer Vision (WACV), Snowmass: IEEE, 2020; pp.1246–1254. doi: 10. 1109/WACV45572.2020.9093394.