



OPEN ACCESS

EDITED BY

Lei Shu,
Nanjing Agricultural University, China

REVIEWED BY

Wen Zhang,
Southwest University of Science and
Technology, China
Jieli Duan,
South China Agricultural University, China
Teera Phatrapornnant,
National Electronics and Computer
Technology Center, Thailand

*CORRESPONDENCE

Jinzhu Lu
✉ lujinzhu1103@163.com

RECEIVED 29 August 2024

ACCEPTED 05 November 2024

PUBLISHED 28 November 2024

CITATION

Wang Y, Lu J, Wang Q and Gao Z (2024)

A method of identification and
localization of tea buds based on
lightweight improved YOLOV5.

Front. Plant Sci. 15:1488185.

doi: 10.3389/fpls.2024.1488185

COPYRIGHT

© 2024 Wang, Lu, Wang and Gao. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A method of identification and localization of tea buds based on lightweight improved YOLOV5

Yuanhong Wang^{1,2}, Jinzhu Lu^{1,2*}, Qi Wang^{1,2} and Zongmei Gao³

¹Modern Agricultural Equipment Research Institute, Xihua University, Chengdu, China, ²School of Mechanical Engineering, Xihua University, Chengdu, China, ³Department of Biological Systems Engineering, Washington State University, Prosser, WA, United States

The low degree of intelligence and standardization of tea bud picking, as well as laborious and time-consuming manual harvesting, bring significant challenges to the sustainable development of the high-quality tea industry. There is an urgent need to investigate the critical technologies of intelligent picking robots for tea. The complexity of the model requires high hardware computing resources, which limits the deployment of the tea bud detection model in tea-picking robots. Therefore, in this study, we propose the YOLOV5M-SBSD tea bud lightweight detection model to address the above issues. The Fuding white tea bud image dataset was established by collecting Fuding white tea images; then the lightweight network ShuffleNetV2 was used to replace the YOLOV5 backbone network; the up-sampling algorithm of YOLOV5 was optimized by using CARAFE modular structure, which increases the sensory field of the network while maintaining the lightweight; then BiFPN was used to achieve more efficient multi-scale feature fusion; and the introduction of the parameter-free attention SimAm to enhance the feature extraction ability of the model while not adding extra computation. The improved model was denoted as YOLOV5M-SBSD and compared and analyzed with other mainstream target detection models. Then, the YOLOV5M-SBSD recognition model is experimented on with the tea bud dataset, and the tea buds are recognized using YOLOV5M-SBSD. The experimental results show that the recognition accuracy of tea buds is 88.7%, the recall rate is 86.9%, and the average accuracy is 93.1%, which is 0.5% higher than the original YOLOV5M algorithm's accuracy, the average accuracy is 0.2% higher, the Size is reduced by 82.89%, and the Params, and GFlops are reduced by 83.7% and 85.6%, respectively. The improved algorithm has higher detection accuracy while reducing the amount of computation and parameters. Also, it reduces the dependence on hardware, provides a reference for deploying the tea bud target detection model in the natural environment of the tea garden, and has specific theoretical and practical significance for the identification and localization of the intelligent picking robot of tea buds.

KEYWORDS

tea buds, target detection, YOLOV5M-SBSD, lightweight modeling, deep learning

1 Introduction

Tea is the second most consumed beverage globally; its unique aroma and characteristic flavor make it famous worldwide. With the booming economy of the tea market, the economic benefits of tea are also increasing. According to the statistical data report of the China Tea Circulation Association, in 2021, the total area of tea gardens in the country was 32640.6 km², with an increase of 5.51% compared with that of 2020, of which the harvestable area was 29163.8667 km², with a weighting of about 89.35%; the output was about 3,029,400 tons, with an increase of 2% compared with that of 2020; the total amount of domestic sales was 2,319,900 tons, with an increase of about 4.56%; total domestic sales amounted to 43,798,694,461.99 dollars, an increase of about 8.0% year-on-year; the average price of domestic sales was 19.02 dollars/g, an increase of 3.3% year-on-year. Tea buds with high nutritional value can be made into high-quality tea with high economic value. The plucking of tea buds must be graded and plucked on the tea buds, which are generally classified into three main categories: single bud, one bud and one leaf, and one bud and two leaves. The image of the tea bud classification is shown in [Supplementary Figure 1](#).

With the globalization of the tea buds industry, countries worldwide are gradually researching the intelligent plucking of tea buds, especially for the target detection of tea buds, which has become a research hotspot. The tea buds' target recognition methods can be roughly divided into three categories: traditional image processing algorithms based on color space, recognition methods based on traditional machine learning, and recognition methods based on deep learning ([Bojie et al., 2019](#)) achieved the segmentation of tea bud targets in tea bud images by extracting the RGB channels of the tea bud images and then performing HIS and HSV spatial conversion of the RGB color space, respectively, and calculating the channel component thresholds of the converted spatial model ([Zhang et al., 2019](#)) utilized the improved B-G algorithm for tea tree canopy processing to segment the tea bud image in the canopy image and then combined it with Bayesian discrimination to realize the recognition of tea buds and harvesting status. Under natural conditions, traditional image processing methods based on color or shape are difficult to perform well in natural and complex infield environments due to problems such as lighting and background complexity.

With the rapid development of machine vision technology, it has received more and more attention as it has demonstrated excellent capabilities in processing image features ([Liu et al., 2019](#)). ([Karunasena and Priyankara, 2020](#)) proposed a stacked class classifier based on the histogram of gradient features (HOG) combined with a support vector machine (SVM) for tea bud detection with an average detection rate of 55% ([Wang et al., 2018](#)) used the HIS model to identify and separate tea buds using the improved K-means algorithm after gray scaling the tea images with the S-factor. A comprehensive analysis of the traditional machine learning-based recognition method found that it relies on image pre-processing and data conversion; pre-processing is crucial, and if the processing is not reasonable, it will seriously impact the model's accuracy. Secondly, the external environment

dramatically affects the method and performs poorly in the natural complex tea garden environment.

Regarding the recognition of tea bud targets, most scholars use deep learning-based methods to realize the recognition of tea bud targets ([Kamilaris and Prenafeta-Boldú, 2018](#); [Chen and Chen, 2020](#)) used Faster RCNN to identify the one tip with two leaves regions in tea bud images and then used the fully convolutional model FCN to identify the tea bud picking points in the one tip with two leaves regions ([Yang et al., 2019](#)) used an improved Yolo-V3 deep convolutional neural network to recognize tea bud picking points, combined with the K-means method to cluster the target box sizes and trained the model to recognize correctly up to more than 90% ([Qian et al., 2020](#)) proposed a tea bud segmentation method based on an improved deep convolutional coding network (TS-Segnet), and the segmentation results were approximately the same as the actual results ([Yan et al., 2022](#)) realized the recognition and localization of tea bud targets by building an improved Mask RCNN model (MR3P-TS model), and their experimental results showed that the picking point localization precision was 0.949 and the recall rate was 0.910. ([Li et al., 2023](#)) proposed a deep learning-based method for tea bud yield estimation by augmenting the YOLOV5 model with a squeeze and excitation network (SENet) and then combining the Hungarian matching algorithm and Kalman filtering algorithm to achieve reliable tea bud counts. The final results found that the model has an average accuracy of 91.88% ([Chen et al., 2021](#)) proposed a fresh tea bud detection method based on image enhancement fusion SSD (FTSD-IEFSSD). The authors used both the enhanced image and the original image for the detection sub-network through the image enhancement algorithm of RGB channel transformation, combined with multi-layer semantic fusion and adaptive score fusion, to nearly improve the target recognition accuracy ([Gong and Wang, 2021](#)) proposed an improved YOLOV4 tea bud target recognition method based on improved YOLOV4, and the final experimental results showed that the average accuracy of the model was 93.08% and the recall rate was 86.94%.

Currently, the main representative models for real-time target detection algorithms include RNN series [Faster RCNN ([Ren et al., 2017](#)), Mask RCNN ([He et al., 2017](#))], YOLO series (V3 ([Redmon and Farhadi, 2018](#)), V4 ([Bochkovskiy et al., 2020](#)), V5), and DETR series [RTMDet ([Lyu et al., 2022](#)), DETRV2 ([Chen et al., 2022](#))]. Considering the special growth environment and physiological characteristics of tea buds, we have chosen the single-stage object detection model YOLOV5 with faster detection speed as the benchmark model to meet the needs of real-time detection of tea buds. Meanwhile, most of the tea bud detection targets belong to regular growth. More research is needed on the multi-target detection of tea buds in complex environments. In contrast, most research has focused on improving the accuracy of tea bud detection for the tea buds detection model without considering the difficulty and cost of model deployment. Detecting tea buds in complex environments using lightweight models is a great challenge; this paper proposes a lightweight tea bud detection model (YOLOV5M-SBSD) for identifying tea buds in complex tea plantation environments, and the method proposed in this paper achieves fast real-time detection of tea buds.

2 Materials and methods

2.1 Data acquisition

In this study, the images of Fuding No. 4 white tea were collected on May 08, 2022, in the open tea garden of Chengdu Liangfeng Tea Co. in Pujiang County, Chengdu City, Sichuan Province, China (N: 30°09'56.45" E: 103°23'49.90")—Pujiang Liangfeng Tea Plantation as shown in [Supplementary Figure 2](#). After a series of processing of the collected raw images, the final white tea dataset contained 5,287 images of tea buds, with a size of 960*1080 pixels, and saved in JPEG format. The iPhone 12 rear camera took the dataset's original images, and the camera's specific parameters are shown in [Table 1](#).

TABLE 1 Camera Setup Parameters for collecting tea buds images.

Variable value	Status
Image size	1920*1080 Pixels
Flash	Off
Zoom	Off
Aperture	f/1.6
Exposure time	1/180 s
Focal distance	26mm
Operation	Manual
Macro	Off
Type	JPG

In this paper, when acquiring images, an acquisition device was used to capture images of tea buds at a distance of 300 mm-800 mm from the tea garden, and the shooting conditions included the background complexity [Figure 1A](#), the shelter [Figure 1B](#), and the camera angle [Figure 1C](#), and an example of a sample image is shown in [Figure 1](#).

In this paper, manually labeling is used to annotate the tea bud images to ensure the effectiveness of the annotation. Considering the quality of tea bud picking and the speed of tea bud positioning, we mainly focus on picking single buds and one bud with one leaf, and label their identification tags as burgeen uniformly. The tea buds with more than 2/3 occlusion are not annotated, and the XML file containing the coordinate information of the tea buds is generated after the annotation. The labeled dataset is divided into training set, validation set, and test set in the ratio of 8:1:1, and there is no repetition between each group.

2.2 Algorithm description of YOLOV5

The YOLOV5 model with relatively balanced accuracy and speed is selected in the first stage of the target detection algorithm model. However, the network width and depth will affect the training and detection time of the model, and there are four versions of the YOLOV5 model with differences in the network width and depth, namely YOLOV5S, YOLOV5M, YOLOV5L, and YOLOV5X. To meet the model's lightweight deployment and real-time requirements and to consider image inputs of arbitrary size, YOLOV5M version 6.0 was finally selected as the benchmark model for Fuding white tea, and its network structure is shown in [Figure 2](#).

The network structure of YOLOV5 is divided into the input side, as shown in the Backbone network [Figure 2A](#), Neck network

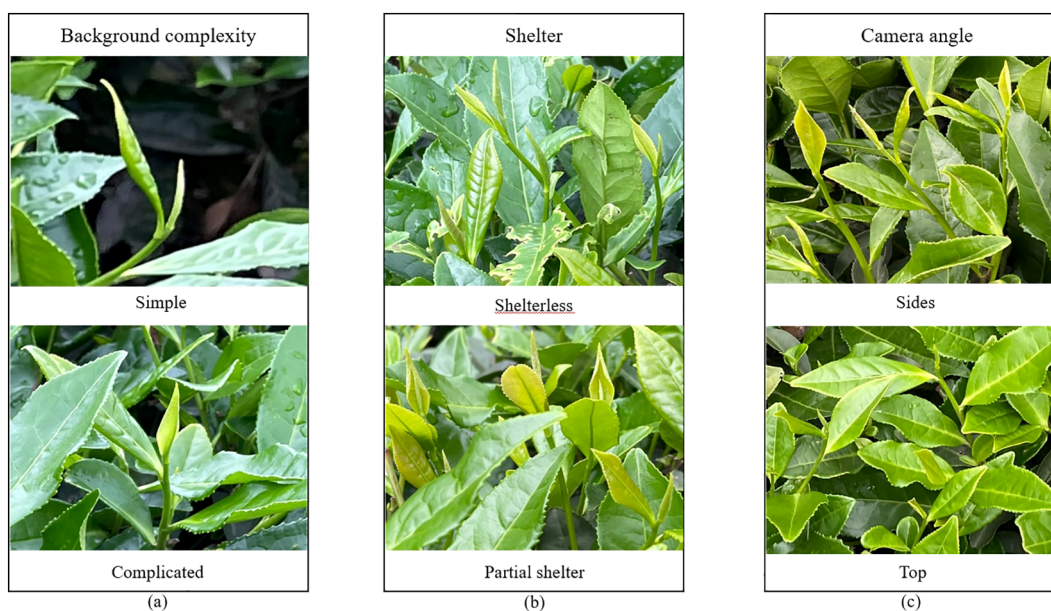


FIGURE 1 Acquisition of tea bud images under different conditions. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

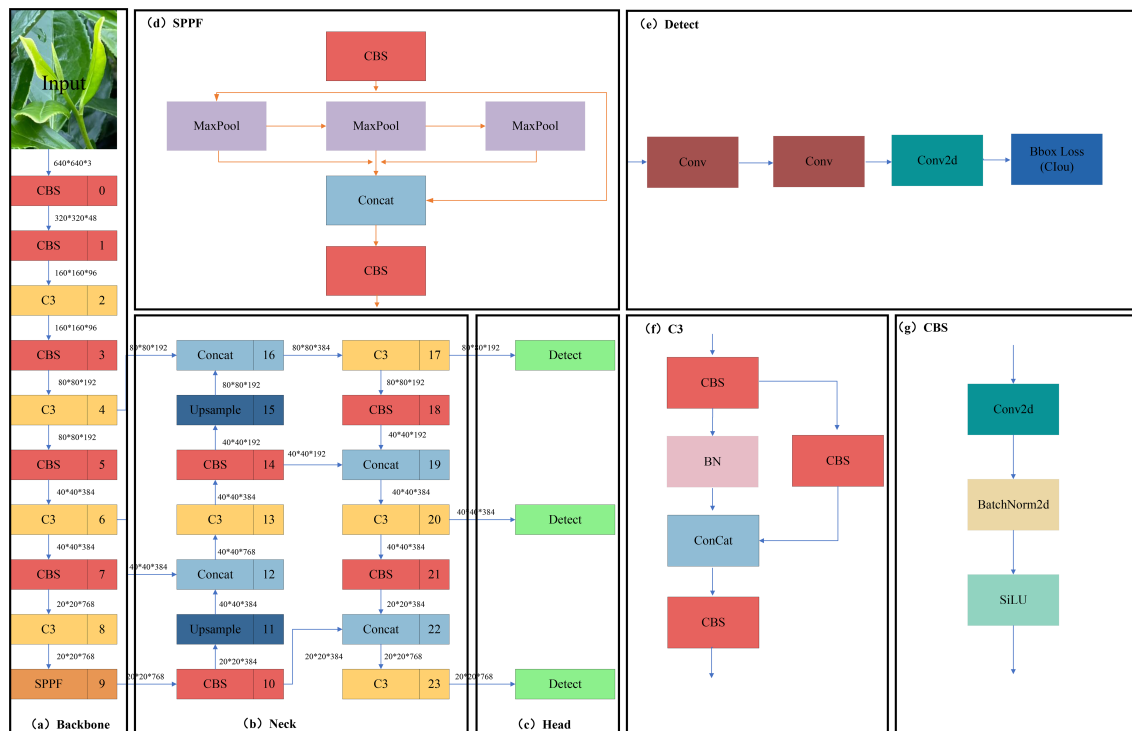


FIGURE 2 YOLOV5 network structure. (A) Backbone module; (B) Neck module; (C) Head module; (D) SPPF module; (E) Detect module; (F) C3 module; (G) CBS module.

Figure 2B, and Head network Figure 2C. It mainly includes Mosaic data enhancement, adaptive anchor frame computation, and adaptive picture scaling at the input side; the Backbone mainly consists of the CBS module Figure 2G, C3 module Figure 2F, and SPPF module Figure 2D; Neck mainly consists of the CBS, up-sampling module (Upsample), Concat module, and C3; the Head mainly comprises three detection head Detect module Figure 2E.

The CBS mainly consists of a convolutional layer, batch-normalized (BN) layer, and sigmoid weighted linear unit (SiLU) activation function, in which the BN layer solves the problems of gradient vanishing and gradient explosion through data normalization. The SiLU activation function is a smooth and non-monotonic function that prevents the gradient from diminishing to 0 during the slow training process. C3 (Park et al., 2018) is a convolution module in YOLOV5, which serves to increase the receptive field of the network and improve the feature extraction capability of the network. The SPPF refers to a feature extraction module for target detection. The SPPF structure improves the model’s detection ability for targets of different sizes by pooling and fusing the feature maps of different sizes of receptive fields to obtain feature information of different scales. At the same time, the SPPF structure also has a specific downsampling effect, which can reduce the resolution of the feature map and improve the computational speed.

The Neck network is an intermediate feature extraction network added based on the Backbone, which is mainly used to enhance the feature expression ability and sensory field of the model to improve the detection performance of the model further. It mainly fuses the

image features through the structure of the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) and transmits them to the detection end. The top-level semantic features are passed down through the top-down FPN (Lin et al., 2017), concatenating the lower and higher-level features using the bottom-up PAN (Liu et al., 2018). Finally, the feature information of different scales is fused, and the CIoU loss function is used at the output to measure the degree of gap between the predicted and natural frames. At the same time, the weighted non-maximum suppression (NMS) method is used for post-processing to remove the redundant candidate frames. The CIoU loss function (Zheng et al., 2022) increases the loss of the detection frame scale based on DIOU and increases the loss of the length and width, and the predicted frames are more in line with the actual frames, which improves the regression accuracy. The formula is shown in (Equation 1):

$$CIOU = IOU - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \tag{1}$$

where $\rho^2(b, b^{gt})$ represents the Euclidean distance between the centroids of the prediction and real frames, respectively, and c represents the diagonal distance of the smallest closure region that can contain both the prediction and real frames. The equations for α and v are shown in (Equations 2, 3):

$$\alpha = \frac{v}{1-IOU+v} \tag{2}$$

$$v = \frac{4}{\pi^2} \left(\tan^{-1} \frac{w^{gt}}{h^{gt}} - \tan^{-1} \frac{w}{h} \right)^2 \tag{3}$$

At the prediction layer, CIOU_Loss is utilized to transfer the loss and weighted NMS is used to obtain the optimal target frame.

The model computation and complexity of the YOLOV5 algorithm in this study are too high, and the arithmetic power of agricultural embedded devices needs to be higher. Therefore, it is imperative to reduce the amount of computation, improve the detection speed, and ensure the detection accuracy.

2.3 Improving the YOLOV5 network design

The YOLOV5M-SBSD network proposed in this paper consists of a Backbone network, Neck network, and Head network. The Backbone network replaces the original backbone network by utilizing ShuffleNetV2 (Ma et al., 2018), which reduces the computational effort and the number of parameters through channel rearrangement and group convolution. Channel rearrangement introduces cross-group connections in the network, reducing information transmission paths and improving feature interaction. Grouped convolution divides the input channels into multiple groups for convolutional operations, reducing the amount of computation in a single convolutional layer. The CARAFE (Wang et al., 2019) module structure was used in the Neck network to optimize the up-sampling algorithm of YOLOV5, which increases the sensory wildness of the network while maintaining lightweight. Due to the complex background environment of the tea plantation, significant differences in light intensity at different times and weather, inconsistent angles of the dataset, and other disturbing factors, the SimAM (Yang et al., 2021) attention module was embedded in the Neck network to improve the focusing on the tea bud target. To improve the accuracy of tea bud detection. In this study, SimAM is embedded into the tandem layer of PANet after giving higher weights to the semantic information of tea buds. Then, the C2F module replaces the C3 module in the Neck network. The C2F module can better adapt to targets of different sizes and shapes by using a variety of convolutional kernel sizes and step sizes, as well as a feature pyramid structure to capture feature information at different scales, improving the model's detection capability and accuracy. The weighted bidirectional feature pyramid network (BiFPN) (Tan et al., 2020) replaces the PANet in the original model to achieve more efficient multi-scale feature fusion. In the Head network, we incorporate a dynamic target detection head (Dyhead) (Dai et al., 2021) to significantly improve the performance of the model target detection head without increasing the computational effort. Dyhead significantly improves the representation of the target detection head without any computational overhead by coherently combining the multi-head self-attention mechanism within the scale-aware feature layer, the spatial location for spatial awareness, and the output channel for task awareness. Finally, we replace the CIOU loss function of this model with MPDIU (Ma and Xu, 2023) and incorporate the NWD (Wang et al., 2021) module, which can obtain faster convergence and more accurate regression results, effectively improving the detection accuracy and localization speed of the tea

bud target. The YOLOV5M-SBSD network structure is shown in Figure 3.

2.3.1 Backbone network improvements

We replace the original backbone of YOLOV5M with the ShuffleNetV2 lightweight Backbone network, and the V2 version introduces a new operation, Channel Split. First, at the beginning, the input feature map is divided into two branches in the channel dimension, with the channel numbers C' and $C-C'$, and the actual implementation is $C' = C/2$. The left branch is mapped equally; the right branch contains three consecutive convolutions and has the same input and output channels, while the two 1×1 convolutions are not group convolutions; the two branches are equivalent to two groups. The output of these two branches will not be an Add element but a Concat operation for both branches. Then, Channel shuffle for the result of the Concat operation to ensure the exchange of information between the two branches. Moreover, the Concat and Channel shuffle can be combined with the Channel Split of the next module to form an element-level operation. Instead of having Channel Split, the downsampling module has one copy of the input for each branch, and each branch has a stride=2 downsampling. Finally, after Concat together, the feature map space size is halved, but the number of channels doubles. Meanwhile, V2 adds a Conv5 convolution before global Pooling, which differs from the V1 version. Under the same condition, ShuffleNetV2 is slightly faster and more accurate than other lightweight models. The network structure of ShuffleNetV2 is shown in Figure 4.

2.3.2 Lightweight upsampling CARAFE module

CARAFE is divided into two main modules, which are the up-sampling prediction module and feature reorganization module; assuming that the multiplicity of up-sampling is σ , given an input feature map with shape $H \times W \times C$, CARAFE firstly measures the up-sampling kernel by using the up-sampling prediction module and then completes the up-sampling by using the feature reorganization module, to get the output feature map with the shape of $\sigma H \times \sigma W \times C$, and the network structure diagram of CARAFE is shown in Figure 5.

In the up-sampling feature prediction module the feature map channels are first compressed, and for input shape feature maps with shapes such as $H \times W \times C$, the number of channels is compressed to C_m using 1×1 convolution to reduce the amount of computation in the subsequent steps. Then it is subjected to content encoding and up-sampling kernel prediction, assuming that the size of the up-sampling kernel is $k_{up} \times k_{up}$, and the amount of computation increases with the increase of the up-sampling kernel, and if we use a different up-sampling kernel for each position of the output feature map, then we need to predict the shape of the up-sampling kernel as $\sigma H \times \sigma W \times k_{up} \times k_{up}$, and the shape of the up-sampling kernel is $\sigma H \times \sigma W \times k_{up} \times k_{up}$, for the compressed input feature maps in the first step, utilizing a $k_{encoder} \times k_{encoder}$ convolutional layer to predict the upsampling kernel, the number of input channels is C_m , the number of output channels is $\sigma^2 K_{up}^2$, and then the channel dimensions are expanded in the spatial dimension, and finally we get an upsampling kernel with the

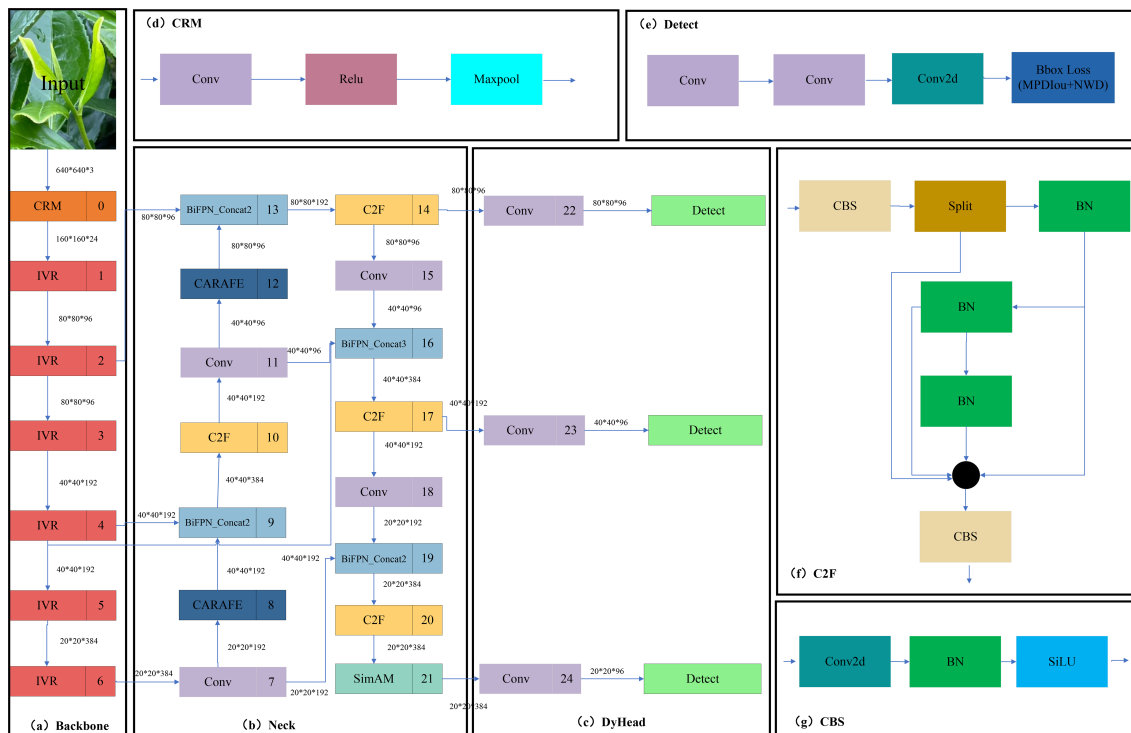


FIGURE 3 YOLOV5M-SBSD network structure diagram. (A) Backbone module; (B) Neck module; (C) DyHead module; (D) CBM module; (E) Detect module; (F) C2F module; (G) CBS module.

shape of $\sigma H * \sigma W * k_{up}^2$, and then finally the upsampling kernel in the second step is subjected to a normalization operation using softmax to make the convolution kernel’s weights sum to 1. In the feature reorganization module, each position of the output feature map is mapped back to the input feature map, the $k_{up} * k_{up}$ region centered on it is taken out, and at the same time the up-sampling kernel of the point is predicted as a dot product, to get its output value, and different channels at the same position share the same up-sampling kernel.

2.3.3 C2F module

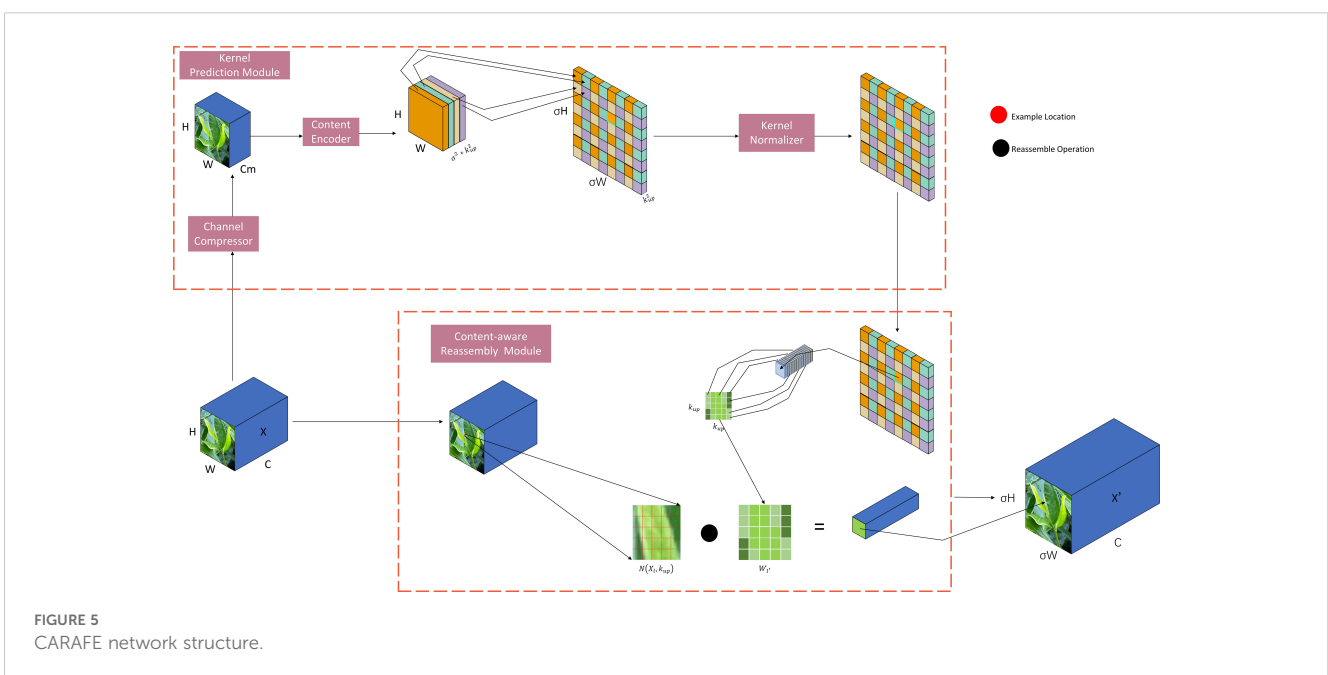
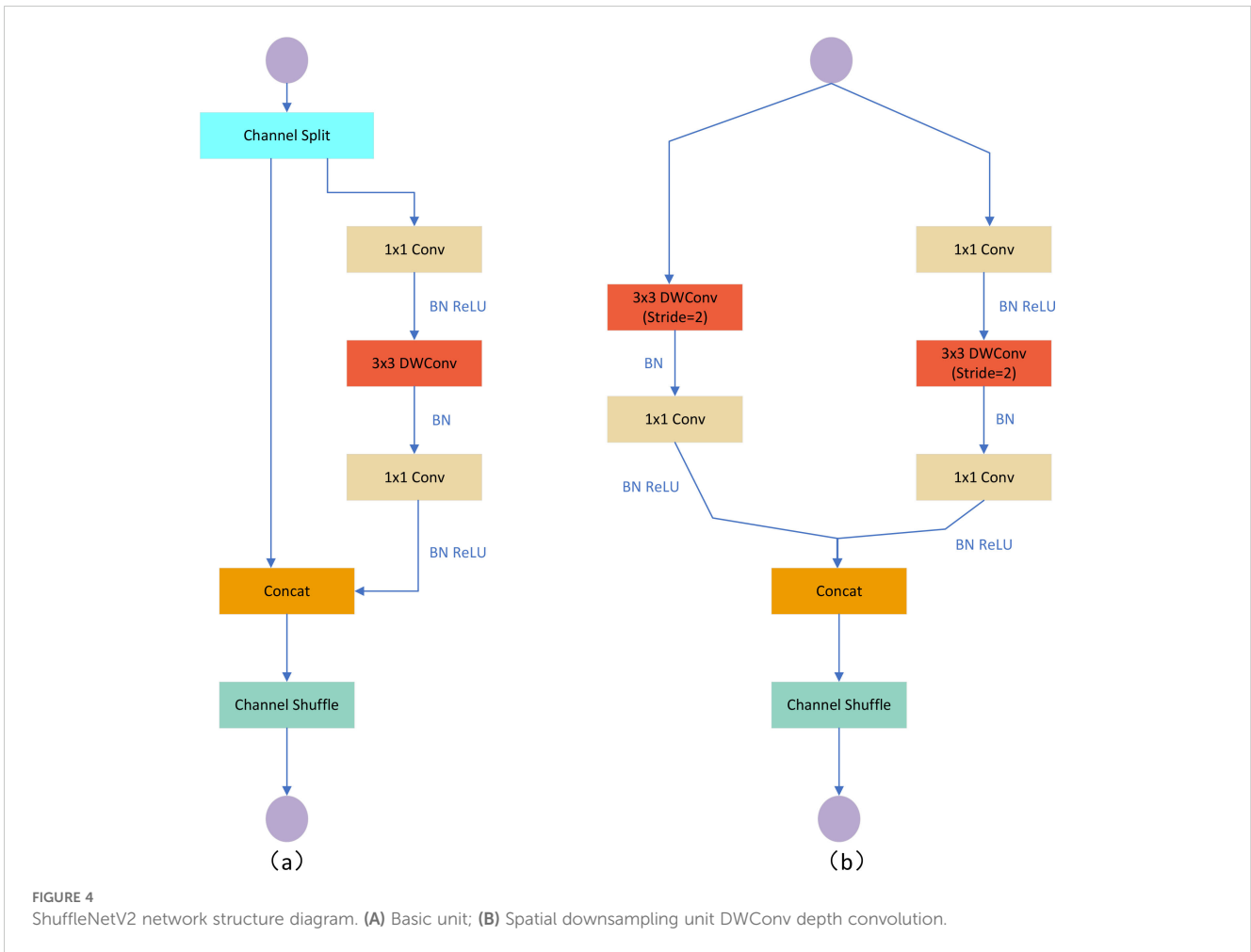
We use the idea of a C2F module in YOLOV8 to replace the original C3 module; the C2F module is a stage Partial Network, which is used for feature fusion, and its central role is to fuse different levels of features to improve the performance of target detection. C2F module mainly consists of two parts, the upsampling and feature fusion two. The up-sampling part matches the size of the high-level feature map by scaling the low-resolution feature map to high resolution through interpolation operation. The feature fusion part adds the up-sampled feature maps with the corresponding low-level feature maps element by element to fuse the semantic information of different levels. Through the operation of C2F, YOLOV5M-SBSD can fuse multi-level feature information while maintaining high resolution, thus improving the accuracy and robustness of the tea bud target detection, better capturing the detailed information of the target, and reducing the leakage and misdetection. The network structure of C2F is shown in Supplementary Figure 3.

2.3.4 BiFPN module

We introduce a BiFPN in the Neck network to replace the original Concat layer. The BiFPN network structure is weighted and bidirectionally connected, i.e., top-down and bottom-up structures, and cross-scale connectivity is achieved by constructing bidirectional channels, which directly fuse the features in the feature extraction network with the relative-size features in the bottom-up paths, retaining shallower semantic information and less loss of deep semantic information. At the same time, BiFPN sets different weights according to the importance of different input features while repeatedly adopting this structure to enhance the feature fusion. The weighted fusion in the BiFPN structure adopts the fast normalized fusion, which is proposed for the slow training speed, and scales down the weights to the range of 0~1, and the training is fast because it does not use the Softmax method. The cross-scale connection is realized by adding a jump connection and a bi-directional path; the weighted fusion and bi-directional cross-scale connection have been realized. The structure of FPN is shown in Figure 6A, the structure of PANet is shown in Figure 6B, and the structure of BiFPN is shown in Figure 6C.

2.3.5 SimAm attention mechanism

We introduce a parameterless attention mechanism, SimAm, in the new model, which is simple and efficient compared to other attention modules. Unlike existing channel or spatial attention modules, this module does not require additional parameters to derive 3D attention weights for the feature map. Currently, existing attention modules are usually inherited into each block to improve



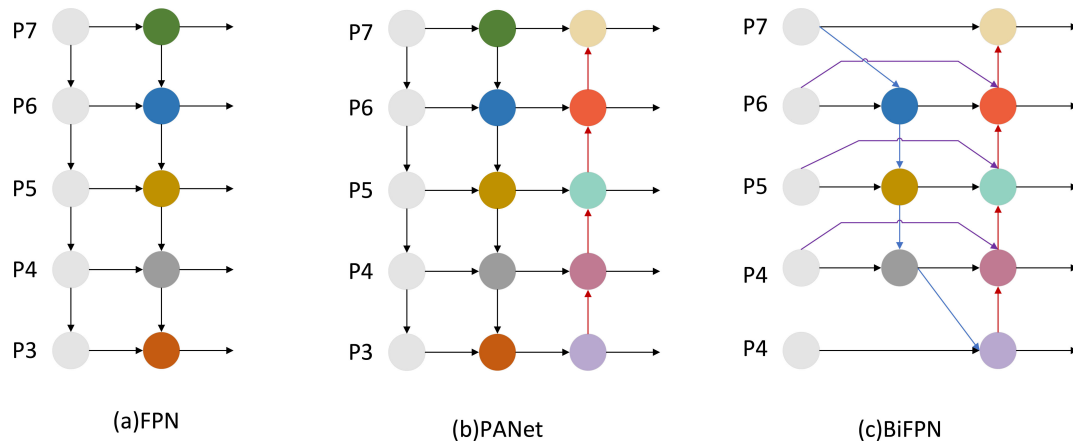


FIGURE 6

BiFPN network design. (A) FPN introduces a top-down path to fuse multi-scale features from P4 to P8; (B) PANet adds a bottom-up path on top of FPN; (C) BiFPN removes redundant nodes from PANet and adds additional connections.

the output from previous layers. This refinement step usually operates along the channel dimension (Figure 7A) or the spatial dimension (Figure 7B), and these methods generate one- or two-dimensional weights and treat neurons at each channel or spatial location equally. Among them, channel attention belongs to 1D attention, which treats different channels differently and treats all locations equally. Spatial attention belongs to 2D attention, which treats different locations differently and treats all channels differently. SimAm belongs to 3D weighted attention (Figure 7C), which can learn more discriminative cues and is significantly better than traditional 1D and 2D weighted attention. Compared to other mainstream attention mechanisms, SimAm performs best and does not introduce additional parameters. The attention to different dimensional weights is shown in Figure 7.

2.3.6 MPDIOU loss function

We introduced the MPDIOU loss function in YOLOV5M-SBSD to replace the original CIOU loss function. The MPDIOU loss function is a kind of bounding box regression loss function, which is used to measure the difference between the predicted box and the real box, and the use of MPDIOU can effectively solve the optimization problem of the bounding box regression loss function in the case that the predicted box and the actual labeled box have the same aspect ratio, but the width value and the height value are completely different. and height values are completely different, and can obtain faster convergence and more accurate regression results. MPDIOU is computed as follows, two arbitrary convex shapes: $A, B \subseteq S \in \mathbb{R}^n$, the width and height of the input images are w, h . For A and B , $(x_1^A, y_1^A), (x_2^A, y_2^A)$ denote the coordinates of the upper-left and lower-right points of A , and $(x_1^B, y_1^B), (x_2^B, y_2^B)$ denote the coordinates of the upper-left and lower-right points of B . The MPDIOU is computed as follows.

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \quad (4)$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \quad (5)$$

$$MPDIOU = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \quad (6)$$

This MPDIOU loss function is defined as follows:

$$L_{MPDIOU} = 1 - MPDIOU \quad (7)$$

All factors of the existing bounding box regression loss function can be determined from the coordinates of the 4 points, and the conversion formula is shown below.

$$|C| = (\max(x_2^{gt}, x_2^{prd}) - \min(x_1^{gt}, x_1^{prd})) * (\max(y_2^{gt}, y_2^{prd}) - \min(y_1^{gt}, y_1^{prd})) \quad (8)$$

$$x_c^{gt} = \frac{x_1^{gt} + x_2^{gt}}{2}, y_c^{gt} = \frac{y_1^{gt} + y_2^{gt}}{2}, y_c^{prd} = \frac{y_1^{prd} + y_2^{prd}}{2}, x_c^{prd} = \frac{x_1^{prd} + x_2^{prd}}{2} \quad (9)$$

$$w_{gt} = x_2^{gt} - x_1^{gt}, h_{gt} = y_2^{gt} - y_1^{gt}, h_{prd} = y_2^{prd} - y_1^{prd}, w_{prd} = x_2^{prd} - x_1^{prd} \quad (10)$$

In the above equation:

$|C|$ denotes the minimum outer rectangle area covering Bgt and Bprd; (x_c^{gt}, y_c^{gt}) and (x_c^{prd}, y_c^{prd}) denote the coordinates of the centers of the real labeled bounding box and the predicted bounding box, respectively; w_{gt} and h_{gt} denote the width and height of the real labeled bounding box; w_{prd} and h_{prd} denote the width and height of the predicted bounding box.

The correlation diagram of the MPDIOU loss function is shown in Figure 8.

2.3.7 Dyhead dynamic inspection head

We introduce the Dyhead dynamic detection head in our new algorithmic detection model, which significantly improves the representation of the target detection head without additional computational overhead by coherently combining the multi-head Self-attention mechanism within the scale-aware feature layer, spatial-aware spatial location, and task-aware output channel. By embedding Dyhead into the YOLOV5M-SBSD one-stage detection algorithm model, tea bud detection performance is significantly

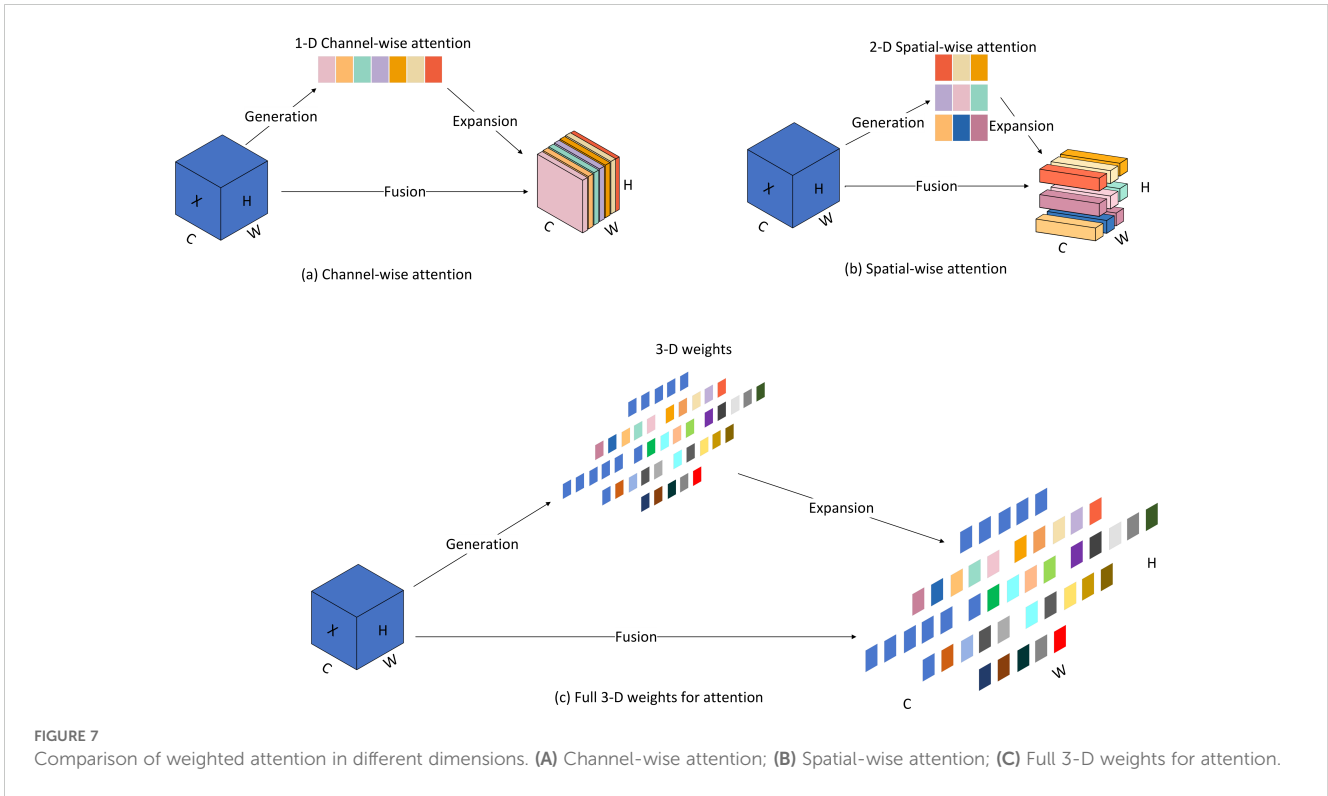


FIGURE 7 Comparison of weighted attention in different dimensions. (A) Channel-wise attention; (B) Spatial-wise attention; (C) Full 3-D weights for attention.

improved. The above three attention modules are defined as follows: for the scale-aware attention module, Scale-aware Attention (π_L), which fuses features of different scales through semantic importance to enhance the scale-awareness of target detection.

$$\pi_L(F) * F = \sigma\left(f\left(\frac{1}{SC} \sum_{s,c} F\right)\right) * F \quad (11)$$

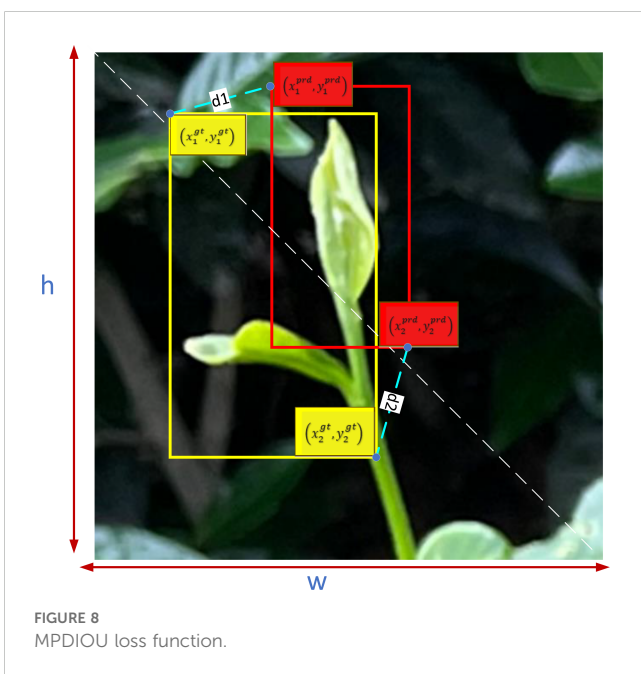


FIGURE 8 MPDIU loss function.

where $f(\cdot)$ is a linear function, approximated with the use of $1*1$ convolution; $\sigma(x) = \max(0, \min(1, \frac{x+1}{2}))$ is a hard-sigmoid function.

For the spatial-aware attention module Spatial-aware Attention (π_S), focusing on the discriminative ability of different spatial locations, deformable convolutional sparsification is first used. Then, the aggregated features of the feature layer are acquired at the exact location to enhance the spatial location awareness of target detection.

$$\pi_S(F) * F = \frac{1}{L} \sum_{l=1}^L \sum_{k=1}^K w_{l,k} * F(l; p_k + \Delta p_k; c) * \Delta m_k \quad (12)$$

where K is the number of sparsely sampled locations; $p_k + \Delta p_k$ does a location shift to focus on discriminative regions; and Δm_k is a self-learnable importance metric factor with respect to location p_k . All of the above features are obtainable by learning the input features from the F intermediate layer.

For the task-aware attention module Task-aware Attention (π_C), different tasks are selected by dynamically turning on or off the feature channel to enhance the target detection's ability to perceive different tasks.

$$\pi_C(F) * F = \max(\alpha^1(F) * F_C + \beta^1(F), \alpha^2(F) * F_C + \beta^2(F)) \quad (13)$$

where $[\alpha^1, \alpha^2, \beta^1, \beta^2]^T = \theta(\cdot)$ is the hyperfunction to control the activation thresholds; $\theta(\cdot)$ first performs global pooling over the dimensions of $L*S$, then uses two fully-connected layers, a normalization layer, and finally normalizes the outputs using the Shifted sigmoid function.

The network structure of Dyhead is shown in Figure 9, and the structure of the embedded one-stage target detection model is shown in Figure 10.

2.4 Experimental environment and parameter configuration

This paper performs tests and runs on the same device, i.e., a desktop mainframe, with specific accessory configurations and experimental environments, as shown in Table 2.

Based on the hardware conditions, we set the hyperparameters with a learning rate of 0.01, momentum of 0.937, weight_decay of 0.0005, batch size of 16, workers of 2, and the optimizer uses stochastic gradient descent (SGD) and a single graphics processing unit (GPU) to accelerate training.

2.5 Evaluation indicators

The evaluation metrics used in this paper employ both performance and complexity; for the performance of the model, there are four metrics: Precision (P), Recall (R), Mean Accuracy (mAP), and F1-Score. P has been used to measure the performance of the model detection, and R has been used to assess the comprehensiveness of the detection (Hsu and Lin, 2021). The combination of P and R is the average precision (AP), and the average accuracy (map) is the mean value of AP, which is used to measure the performance of the whole model (Guo et al., 2022).

mAP(0.5,0.95) denotes the mAP for different thresholds ranging from 0.5 to 0.95, in steps of 0.05. Since P and R are conflicting performance metrics, the F1-Score is the P and R, the reconciled mean of P and R with a range value of (0,1), which uses both P and R to assess the quality of the model. The specific formula is as follows: the quality of the model. The specific formula is as follows:

$$P = \frac{TP}{TP+FP} * 100\% \tag{14}$$

$$R = \frac{TP}{TP+FN} * 100\% \tag{15}$$

where TP is a positive sample predicted to be a positive class, FN is a positive sample predicted to be a negative class, and FP is a negative sample predicted to be a positive class:

$$AP = \int_0^1 P(R)dR \tag{16}$$

$$mAP(0.5) = \frac{\sum_{i=1}^n AP_i}{n} \tag{17}$$

where n is the number of categories

$$F1 = \frac{2 * P * R}{P + R} \tag{18}$$

For the model complexity aspect, there are three metrics: Params, GFlops and Size (Rampriya et al., 2022), which are

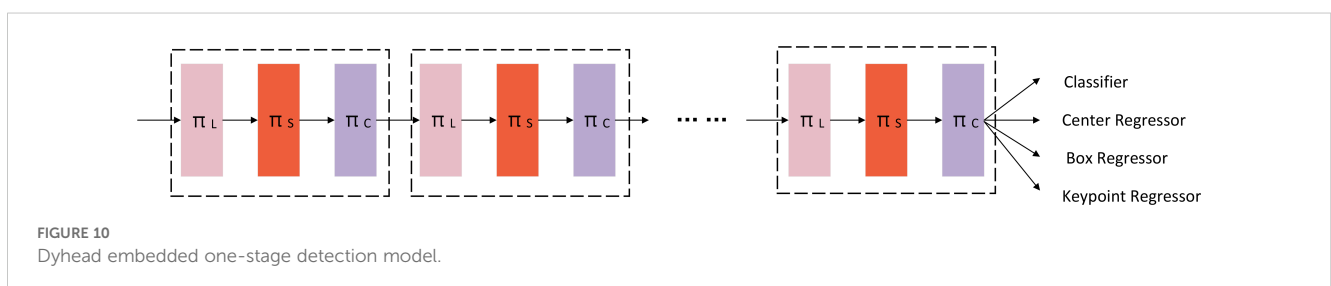
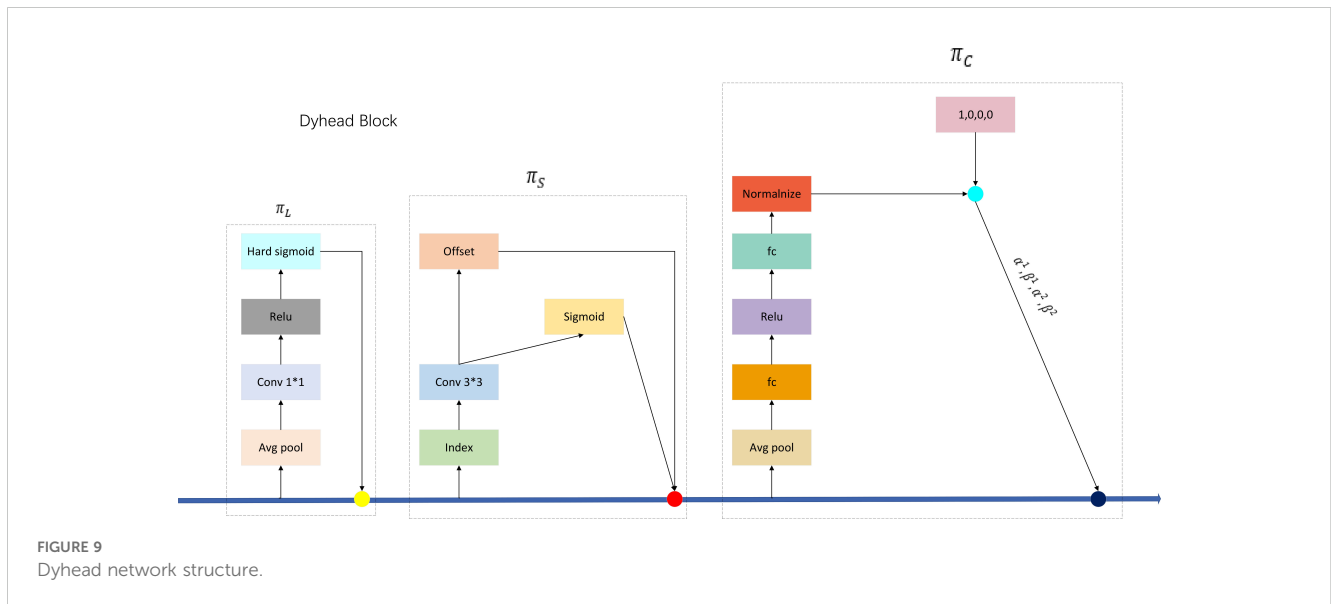


TABLE 2 Hardware configuration and operating environment.

Hardware	Configure	Environments	Version
System	Windows10	python	3.10.11
CPU	Intel(R) i5-13400F	pytorch	1.13.1
GPU	NVIDIA RTX 3060 Ti	pycharm	2023.1
RAM	32.0 GB	CUDA	11.7
Hard-disk	1TB	CUDNN	8.4.1.5

formulated as follows:

$$Params = [r * (f * f) * o] + o \quad (19)$$

$$GFlops = o \left(\sum_{i=1}^n k_i^2 * C_{i-1}^2 * C_i + \sum_{i=1}^n M^2 * C_i \right) \quad (20)$$

Where r is the input size, f is the convolution kernel size, o is the constant order, K is the convolution kernel size, C is the number of channels, M is the input image size and i is the number of iterations.

3 Results and discussion

3.1 YOLOV5M modeling improvement experiments

3.1.1 Lightweighting module experiments

In this paper, YOLOV5M is chosen as the Backbone network, and its Backbone network is replaced by MobileNetV3, ShuffleNetV2, GhostNetV1, and GhostNetV2 for a comparison test. The careful consideration of the lightweight effect and the recognition accuracy finally determines the Backbone network of YOLOV5M-SBSD. It can be seen from Table 3 that the comprehensive rate of the lightweight effect and the recognition accuracy of YOLOV5M-ShuffleNetV2 are significantly better than those of other types. Although the P and mAP of the benchmark model Backbone network are improved by replacing the Ghost series, the Params and GFlops are higher than those of MobileNetV3 and ShuffleNetV2 as the Backbone networks. After replacing the Backbone network of the original benchmark model with a lightweight network, the Params and GFlops of the tea bud detection model are reduced to a certain extent, while the P and

mAP of the Ghost series are improved, the P and mAP of MobileNetV3 and ShuffleNetV2 are reduced, and the R of the tea bud detection model is reduced to a certain extent. For the accuracy P of YOLOV5M-MobileNetV3, the mAP is 1.4% and 0.8% lower than those of the original benchmark model, and the number of Params, GFlops and Size are reduced by 89.99%, 91.65% and 89.16%, respectively; For the accuracy P of YOLOV5M-ShuffleNetV2, the mAP is 1.0% and 0.6% lower than those of the original benchmark model, and the number of Params, GFlops and Size were reduced by 90.31%, 91.03% and 89.51%, respectively; For the P and mAP of YOLOV5M-GhostNetV1 is 0.7% and 0.1% higher than those of the original benchmark model, and the number of Params, GFlops and Size were reduced by 59.19%, 62.11% and 54.73%, respectively; For the P and mAP of YOLOV5M-GhostNetV2 is 0.3% and 0.8% higher than those of the original benchmark model, and the number of Params, GFlops and Size were reduced by 4.05%, 10.44% and 3.49%, respectively. Considering the detection performance of the model, the detection accuracy and average accuracy of YOLOV5M-ShuffleNetV2 are better than those of YOLOV5M-MobileNetV3, and finally the Backbone network of YOLOV5M-SBSD is ShuffleNetV2.

3.1.2 Comparative experiments on attention mechanisms

After replacing the Backbone of YOLOV5M with ShuffleNetV2, the P and mAP are subsequently lost. Therefore, we consider introducing BiFPN to replace the original Concat layer of the model first and then introducing an attention mechanism to improve the recognition effect of YOLOV5M-ShuffleNetV2-BiFPN. This experiment adds CBAM, CA, ShuffleAttention, NAM, and SimAM attention mechanisms at the exact position of the Neck network of the YOLOV5M-ShuffleNetV2 base network to conduct comparative experiments. As seen from Table 4, the model's detection is improved by embedding the attention mechanism, and the addition of all types of attention mechanisms except the SimAM attention mechanism leads to a slight increase in Params and GFlops. Although the improvement of CA for mAP is 0.1% higher than that of SimAM, its P is 1% lower than that of SimAM, and its Params, GFlops, and Size are 0.026616M, 0.1G, and 0.05M higher than that of SimAM, respectively. Other types of attentional mechanisms have lower P and mAP than that of SimAM, and their Params, GFlops, and Size are higher than that of SimAM. They are all higher than SimAM. The introduction of

TABLE 3 Comparison of YOLOV5M results under different backbone networks.

Model	P	R	mAP	Params/M	GFlops/G	Size/M
YOLOV5M	88.2%	89.3%	92.9%	20.852934	47.9	40.2
+MobileNetV3	86.8%	87.5%	92.1%	2.296612	4.0	4.76
+ShuffleNetV2	87.2%	87.9%	92.3%	2.020998	4.3	4.22
+GhostNetV1	88.9%	88.7%	93.0%	8.509854	18.2	16.8
+GhostNetV2	88.5%	88.3%	93.7%	20.009622	42.6	38.8

TABLE 4 Comparison of YOLOV5M-ShuffleNetV2-BiFPN results with the addition of different attention mechanisms.

Model	P	R	mAP	Params/M	GFlops/G	Size/M
YOLOV5M+Sh+BiFPN	87.4%	87.1%	92.4%	2.057871	4.4	4.29
+CBAM	87.4%	87.4%	92.1%	2.076810	4.5	4.33
+CA	86.8%	88.3%	92.5%	2.084487	4.5	4.34
+ShuffleAttention	87.2%	87.7%	92.3%	2.061183	4.5	4.29
+NAM	87.6%	86.6%	92.1%	2.061807	4.5	4.29
+SimAM	87.8%	87.1%	92.4%	2.057871	4.4	4.29

the SimAM attention mechanism improves the P of YOLOV5M-ShuffleNetV2 by 0.4%. The experimental results show that using SimAM, a participantless attention mechanism, improves the feature extraction ability of the tea bud target, suppresses the interference of the complex background, and effectively improves the detection effect in detecting tea buds without increasing the complexity of the model.

3.1.3 Comparative experiments on loss functions mechanisms

The introduction of parameter-free attention SimAM improves the detection effect of the target detection model on tea buds, and the loss function of the target detection model is replaced to improve the robustness of the training model. In this experiment, the EIOU, SIOU, WIOU, α _IOU, F_CIOU, F_EIOU, and MPDIU loss functions were replaced for comparison experiments based on YOLOV5M-ShuffleNetV2-BiFPN-SimAM. From Table 5, it can be seen that different loss functions have no effect on the model complexity and have some effect on the P and mAP. The experimental results show that when the loss function is replaced with MPDIU, the base P and mAP are increased by 0.1% and 0.2%, respectively, and nothing changes. The effect of other types of loss function relative to the base model loss function of the P have some reduction for the mAP, except for the loss function EIOU and WIOU remain unchanged, the rest of them are reduced

than the base model. The MPDIU loss function can obtain faster convergence speed and more accurate regression results.

3.1.4 Ablation experiments

We conducted ablation experiments on the lightweight module ShuffleNetV2, BiFPN, SimAM, loss function MPDIU, C2F, NWD, lightweight Upsample CARAFE, and Dyhead, to evaluate the effectiveness of the YOLOV5M-SBSD detection algorithm in the detection of tea buds. It is worth noting that in order to make Table 6 more aesthetically pleasing and concise, we have abbreviated the relevant modules in the ablation experiment results of Table 6, where Tag represents the serial number, Basic represents the baseline model, SNetV2 represents ShuffleNetV2, CAFE represents CARAFE, and Dhead represents Dyhead. From Table 6, it can be seen that a variety of model structures and algorithmic strategy-based improvement methods are effective, and compared with the original YOLOV5M, the P is improved by 0.5%, the mAP is improved by 0.2%, the Size is reduced by 82.89%, and the Params and GFlops are reduced by 85.6% and 83.7%, respectively.

Replacing the Backbone of YOLOV5M using ShuffleNetV2 significantly reduced the performance and complexity of the detection model, with the Size reduced by 82.89%, the Params and GFlops, reduced by 85.6% and 83.7%, respectively. However, the cost of changing the model to reduce the model complexity and computational volume was to reduce the effectiveness of tea bud

TABLE 5 Comparison of YOLOV5M-ShuffleNetV2-BiFPN-SimAM results under different loss functions.

Model	P	R	mAP	Params/M	GFlops/G	Size/M
YOLOV5M+Sh+Bi+Si(CIOU)	87.8%	87.1%	92.4%	2.057871	4.4	4.29
+EIOU	87.6%	87.2%	92.4%	2.057871	4.4	4.29
+SIOU	86.8%	87.9%	92.2%	2.057871	4.4	4.29
+WIOU	87.4%	87.3%	92.4%	2.057871	4.4	4.29
+ α _IOU	86.0%	86.8%	90.6%	2.057871	4.4	4.29
+F_CIOU	86.9%	88.2%	92.2%	2.057871	4.4	4.29
+F_EIOU	87.0%	86.9%	92.0%	2.057871	4.4	4.29
+MPDIU	87.9%	87.1%	92.4%	2.057871	4.4	4.29

TABLE 6 Ablation experiment results.

Tag	Basic	SNetV2	BiFPN	SimAM	MPIOU	C2F	NWD	CAFE	Dhead	P	mAP	Params/M	GFlops/G	Size/M
0	√									88.2%	92.9%	20.85293	47.9	40.2
1	√	√								87.2%	92.3%	2.020998	4.3	4.22
2	√	√	√							87.4%	92.4%	2.057871	4.4	4.29
3	√	√	√	√						87.8%	92.4%	2.057871	4.4	4.29
4	√	√	√	√	√					87.9%	92.4%	2.057871	4.4	4.29
5	√	√	√	√	√	√				87.9%	92.5%	2.057871	4.4	4.29
6	√	√	√	√	√	√	√			88.4%	92.9%	2.659599	5.7	5.43
7	√	√	√	√	√	√	√	√		88.4%	93.0%	2.793559	5.9	5.70
8	√	√	√	√	√	√	√	√	√	88.7%	93.1%	3.400287	6.9	6.88

identification, resulting in a 0.7% reduction in mAP. By replacing the original Concat layer with the introduction of BiFPN in the Neck network, the P was improved by 0.2%, and the mAP was improved by 0.1%, which improved the ability of mAP in the model to recognize tea bud targets. In addition, introducing the parameter-free attention mechanism SimAM without increasing the model complexity improves the P by 0.4%, and the mAP is unchanged. By replacing the loss function MPDIU combined with NWD to achieve the optimization of the loss function, the P is improved by 0.6%, the mAP is improved by 0.5%, and finally, the Dyhead is replaced with the original detection head. Finally, the P is improved by 0.3%, and the mAP is improved by 0.2%. As shown in Figure 11,

during the improvement process of the above method, the mAP is gradually improved, and the model complexity changes are minor, finally forming the YOLOV5M-SBSD detection model.

3.2 Experiments comparing the performance of different models

To verify the comprehensive performance of the YOLOV5M-SBSD model proposed in this paper, we used a total of Seven detection models: Faster RCNN, YOLOV3, YOLOV4, YOLOV4-tiny, YOLOV5S, YOLOV5M, and YOLOV5M-SBSD, for

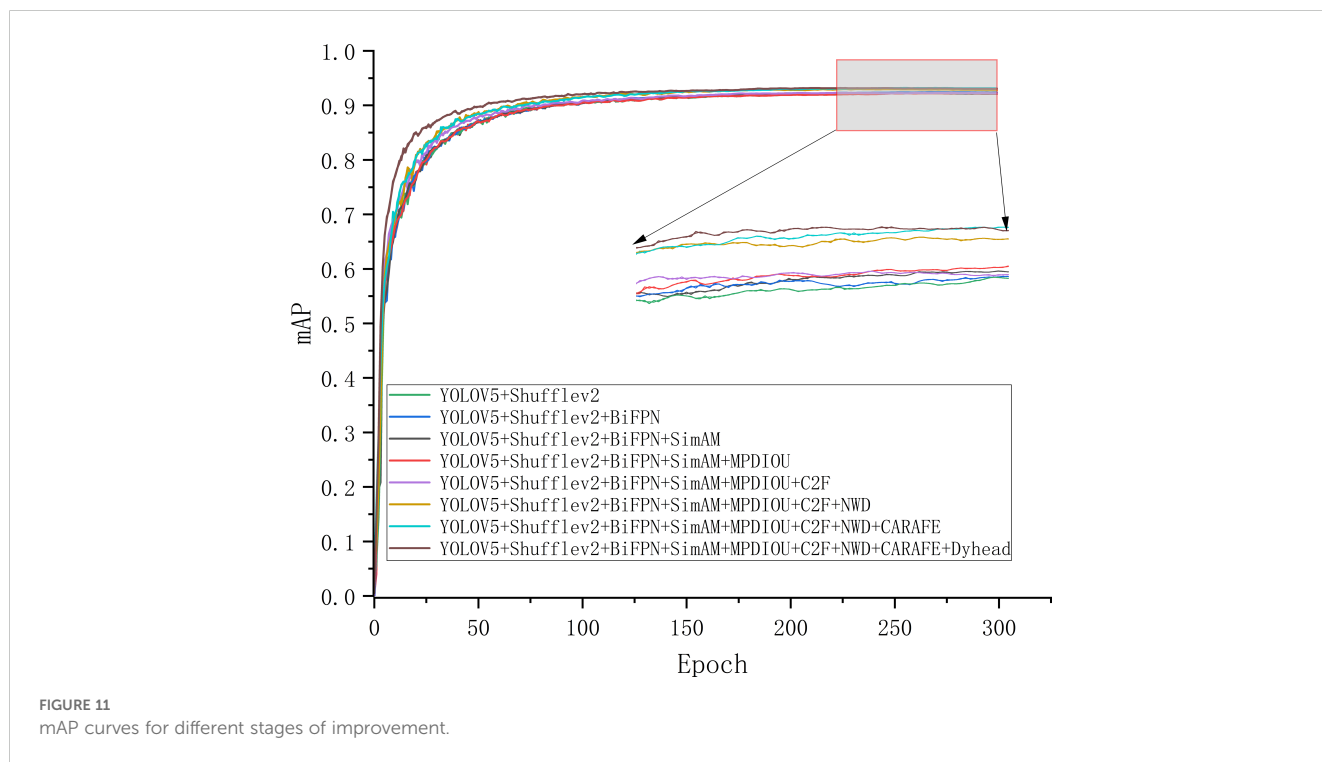


FIGURE 11 mAP curves for different stages of improvement.

comparison, and the results of the performance tests are shown in Table 7.

Table 7 shows that the YOLOV5M-SBSD target detection model has the best performance in terms of precision P, mAP, Params, GFlops, and Size. The YOLOV5M-SBSD target detection model has the best performance in terms of P, mAP, Params, GFlops, and Size than Faster RCNN, YOLOV3, YOLOV4, YOLOV4-tiny, YOLOV5S and YOLOV5M. The P is 39.22%, 0.5%, 15%, 21.1%, 17.1% and 0.5% higher respectively, the mAP is 9.08%, 1.7%, 5.2%, 12%, 22.2% and 0.2% higher respectively, and the Params is 98.53%, 94.48%, 93.53%, 42.12%, 51.52% and 83.7% lower respectively, and the GFlops is 98.14%, 95.54%, 92.33%, 42.98%, 56.33% and 85.6% lower respectively, and the Size is 93.63%, 94.43%, 96.56%, 52.23%, 84.68%, and 82.89% lower respectively. The YOLOV5 version of the follow-up is widely used due to its flexibility, and the model complexity of YOLOV5M is high, which is not conducive to model deployment on low-computing-power devices, but after lightweight its detection performance will be reduced to a certain extent, so by making a series of improvements to the lightweight YOLOV5M to achieve its detection performance, which shows that YOLOV5M-SBSD has the best performance among the one-stage target detection algorithms.

3.3 Analysis of model detection effect

The improved model YOLOV5M-SBSD has a better overall performance for tea bud detection, and it is the best for different background complexity, different tea bud complexity, and different shooting angles; it has the lowest leakage and misrecognition rate, and the average detection rate of the model is above 80%, as shown in Figures 12–18. Among them, Faster RCNN, YOLOV3, YOLOV4, YOLOV4-tiny, YOLOV5S, and YOLOV5M have different degrees of misrecognition and even some leakage detection, and YOLOV4-tiny has the most severe leakage detection. In this research, black dashed boxes were utilized to indicate false detections, purple dashed boxes were used to indicate missed detections, and blue dashed boxes were used to indicate duplicate detections. For the Faster RCNN target detection algorithm, certain misdetections and omissions exist, such as Simple in Figure 12A and Top in Figure 12C. For the YOLOV3 target detection algorithm, certain duplicate

identifications and omissions exist, such as Shelterless in Figure 13B and Top in Figure 13C. There are certain omissions for the YOLOV4 target detection algorithm, such as Shelterless and Partial shelter in Figure 14B and Top in Figure 14C. There are serious omissions for the YOLOV4-tiny target detection algorithm, such as Simple in Figure 15A, Shelterless and Partial Shelter in Figure 15B and Top in Figure 15C. The YOLOV5S target detection algorithm has certain misdetections and omissions, such as Top in Figure 16C and Shelterless in Figure 16B. The YOLOV5M target detection algorithm has certain duplicate identifications and misdetections, such as Simple in Figure 17A, Shelterless in Figure 17B and Top in Figure 17C. For our proposed YOLOV5M-SBSD target detection model, there is no such existing situation as mentioned above, which effectively illustrates the excellent performance of our proposed model.

3.4 Discussion

Visual recognition is a prerequisite for the intelligent picking of tea buds, which is of great significance for developing intelligent picking equipment for tea buds. With the advancement of computer technology and the development of agricultural robots, the application scope of deep learning has become more and more extensive. The single-stage target detection algorithm YOLO series has received extensive attention from scholars to meet the need for real-time detection and adapt to the real-time detection of intelligent picking equipment (Li et al., 2022) proposed a method to achieve real-time detection of tea buds using the YOLOV3SPP deep learning algorithm combined with channel pruning; they achieved this by adding a pyramid pooling module to the YOLOV3 model while combining the channel pruning algorithm and then fine-tuning the model, and through experiments, it was ultimately found that the size of the model and the detection time was reduced relative to the previous model by 96.81% and 59.62%. The detection speed of the compressed model is 15.9 fps, which is 3.18 times that of the original model (Zhang et al., 2023) proposed a ShuffleNetV2-YOLOV5Lite-E-based edge device detection method for one-bud and two-leaf tea. The final experimental results show that the file size of the improved model is reduced by 27% relative to the previous model, and the detection speed of the improved model

TABLE 7 Comparison of experimental results for different models.

Model	P	mAP	Params/M	GFlops/G	Size/M
Faster RCNN	49.48%	84.02%	137.099	370.210	108
YOLOV3	88.2%	91.4%	61.497430	154.5	123.5
YOLOV4	73.7%	87.9%	52.496000	89.8	200
YOLOV4-tiny	67.6%	81.1%	5.874210	12.1	44.9
YOLOV5S	71.6%	70.9%	7.012822	15.8	14.4
YOLOV5M	88.2%	92.9%	20.852934	47.9	40.2
YOLOV5M-SBSD	88.7%	93.1%	3.400287	6.9	6.88

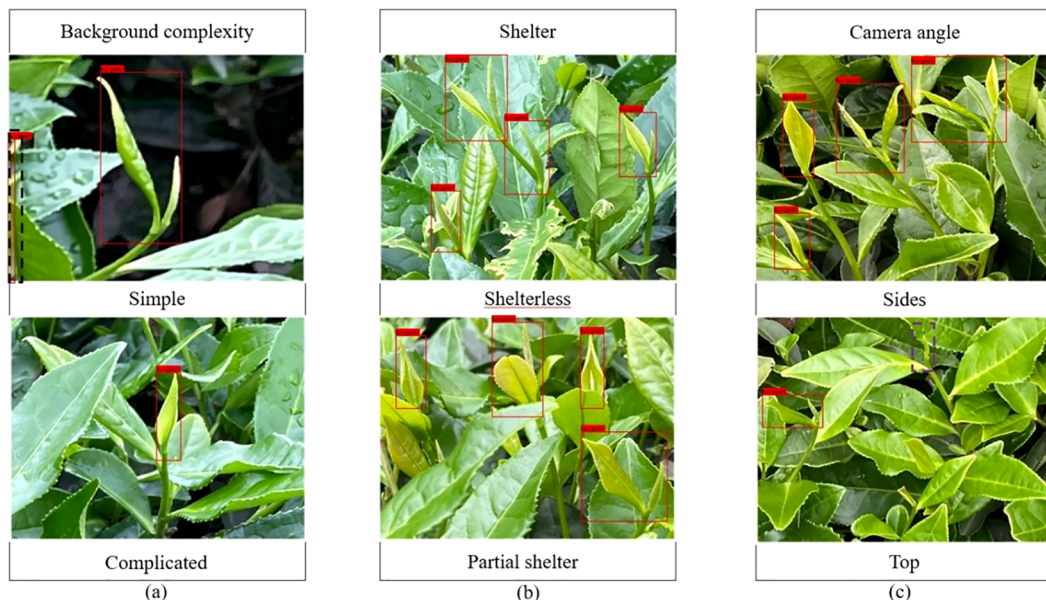


FIGURE 12 Faster RCNN detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

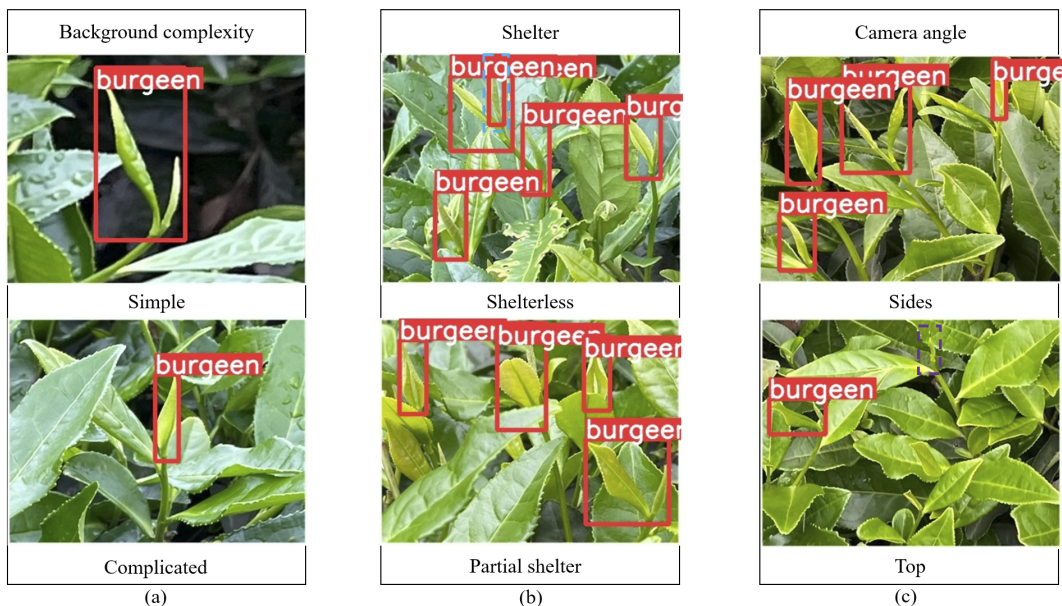


FIGURE 13 YOLOV3 detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

is 3.2 times faster than the original YOLOV5 model (Cao et al., 2022) proposed a tea bud detection algorithm combining GhostNet and YOLOV5 by comparing the newly improved model with Faster RCNN, YOLOV5, and YOLOV5- lite correlation models, and the final experimental results showed that the target recognition accuracy of the newly improved model was improved by 1.31%, 4.83%, and 3.59%, respectively, concerning the compared models.

The mAP of the YOLOV5M-SBSD target detection model proposed in this paper is 93.1%, and the average detection speed of a single image is 15.41ms, which meets the requirement of real-time detection. In addition, from Table 7, it can be seen that compared with Faster RCNN, YOLOV3, YOLOV4, YOLOV4 tiny, YOLOV5S, YOLOV5M and YOLOV5M-SBSD the model proposed in this paper has higher detection accuracy and average detection rate, as

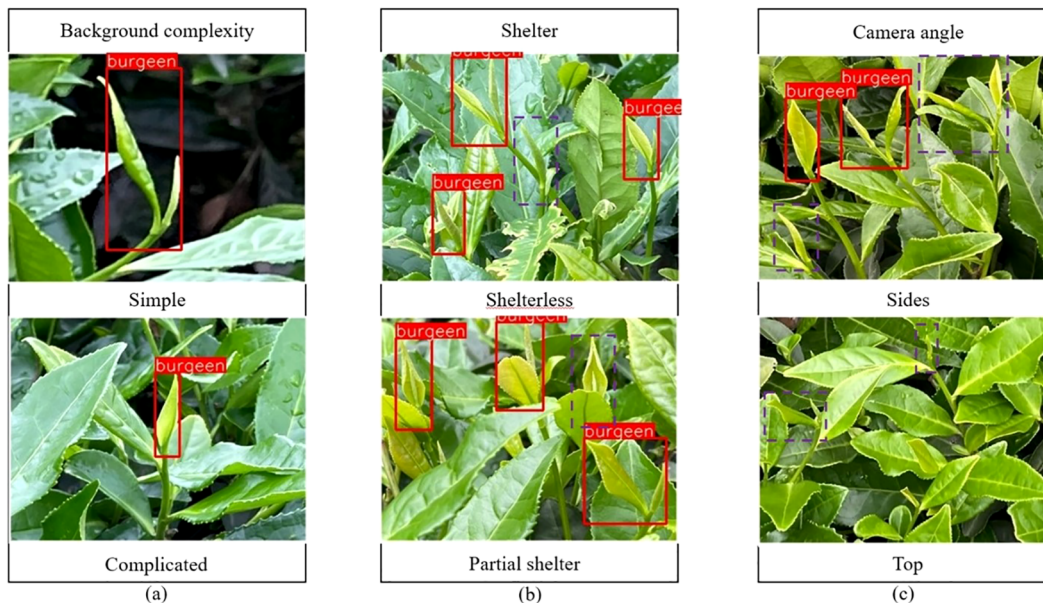


FIGURE 14 YOLOV4 detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

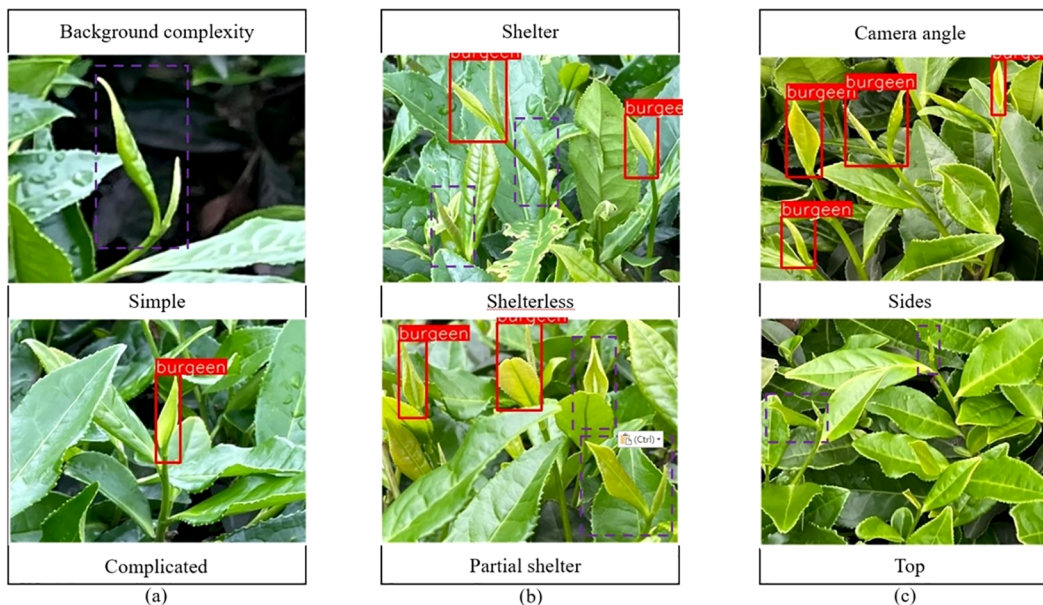


FIGURE 15 YOLOV4-tiny detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

well as lower model volume, Params and GFlops. In order to distinguish the difference between the newly proposed lightweight detection model and other detection models, we took the initials of the main modules added by the improvement as the suffix, and renamed the newly proposed lightweight detection model. Therefore, we named the newly proposed lightweight tea bud

detection model YOLOV5M-SBSD. Our newly proposed lightweight tea bud detection model, YOLOV5M-SBSD, has a wide range of application prospects, such as crop yield estimation and intelligent picking robot equipment development. The YOLOV5M-SBSD tea bud detection model can effectively adapt to the equipment with low computing power and reduce the impact

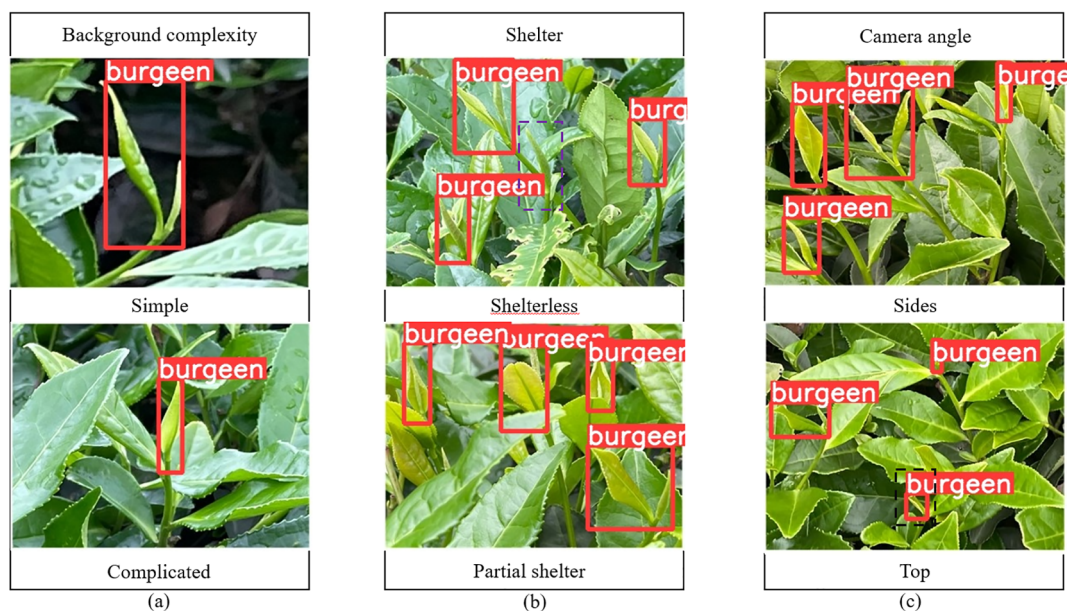


FIGURE 16
YOLOV5S detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

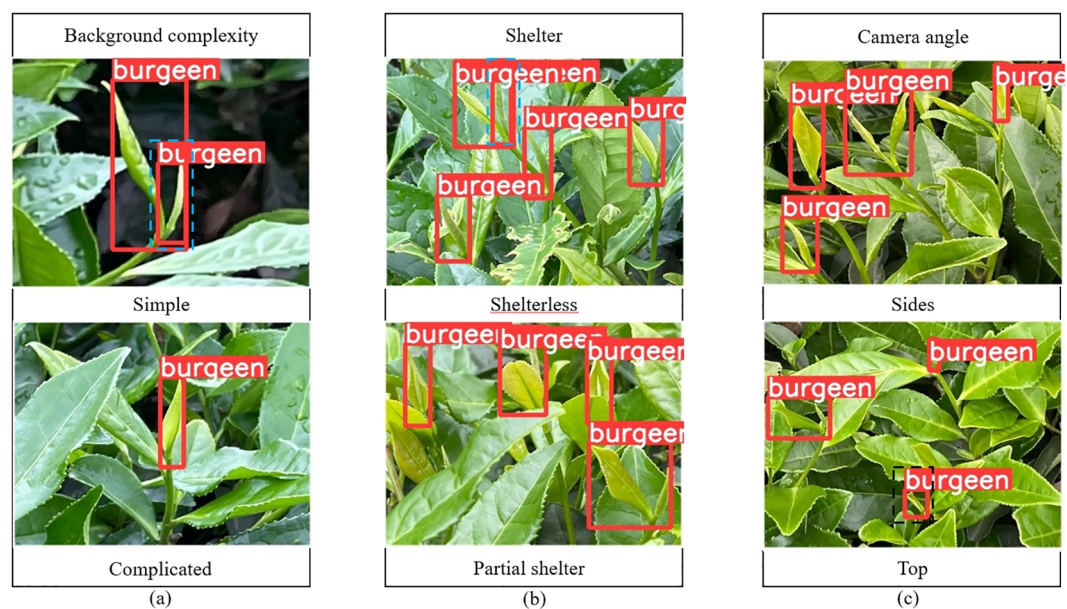


FIGURE 17
YOLOV5M detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

of insufficient computing power on the detection effect of tea buds. In addition, the model can provide a new idea for the detection of other target crops, and provide technical support for target detection under low computing power equipment.

In future research, the tea bud detection model will be optimized through transfer learning by combining the characteristics of other

representative tea varieties. In addition, multi-source information fusion methods are used to reduce the influence of factors such as solid light on tea buds and to improve the ability to extract features from tea buds. We try to collect images of tea buds of the same variety in different periods and make corresponding data sets to reduce the influence of the growth characteristics of tea buds on the

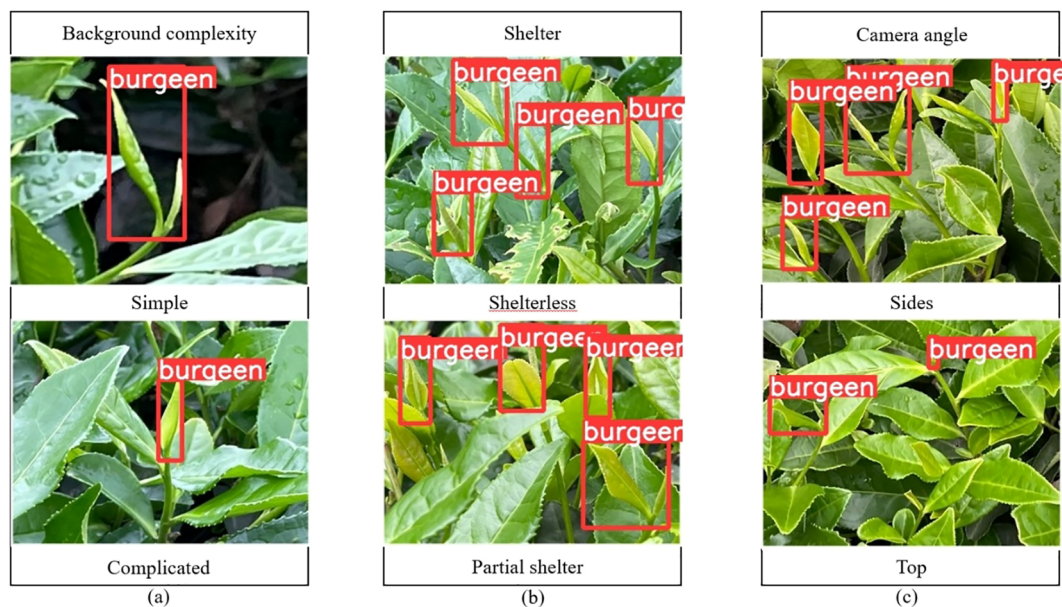


FIGURE 18
YOLOV5M-SBSD detection accuracy. (A) Tea bud background complexity; (B) Tea bud shelter; (C) Tea bud camera angle.

identification of tea buds. Finally, the improved lightweight model is deployed to a low-computing-power device to carry out picking experiments in the complex environment of tea gardens to verify the excellent performance of the improved algorithm.

4 Conclusions

In order to achieve accurate detection of tea buds in the complex environment of limited computing power equipment and tea gardens, this paper proposes an improved target detection model YOLOV5M-SBSD. The experimental results show that YOLOV5M-SBSD outperforms the YOLOV5M target detection algorithm model, with Params, GFlops, and Size decreasing from 20.852934M to 3.400287M, from 47.9G to 6.9G, and from 40.2M to 6.88M, respectively. Params, GFlops, and Size are reduced by 83.7%, 85.6%, and 82.89%, respectively. Meanwhile, the P of the target detection model improves by 0.5%, and the mAP improves by 0.2%. Compared with other mainstream target detection models YOLOV3, YOLOV4, YOLOV4-tiny, YOLOV5S, YOLOV5M, and Faster RCNN, YOLOV5M-SBSD has the highest detection accuracy of 88.7%, the highest average detection rate of 93.1%, and the lowest Params, GFlops, and Size, respectively, of 3.400287M, 6.9G, and 6.88 M. This effectively demonstrates that YOLOV5M-SBSD can effectively and accurately detect tea buds in complex environments and on computationally underpowered devices, provide technical support for the development of intelligent picking equipment for high-quality tea, and promote the intelligent development of the high-quality tea industry.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YW: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. JL: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – review & editing. QW: Investigation, Software, Writing – review & editing. ZG: Writing – review & editing, Methodology.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Key projects of Chengdu Science and Technology Bureau (2022YF0501127SN).

Acknowledgments

We are very grateful for the help of other students in the laboratory.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv*. doi: 10.48550/arXiv.2004.10934
- Bojie, Z., Dong, W., Weizhong, S., Yu, L., and Ke, W. (2019). "Research on tea bud identification technology based on HSI/HSV color transformation." *2019 6th International Conference on Information Science and Control Engineering (ICISCE)*, Shanghai, China, pp. 511–515. doi: 10.1109/ICISCE48695.2019.00108
- Cao, M., Fu, H., Zhu, J., and Cai, C. (2022). Lightweight tea bud recognition network integrating GhostNet and YOLOv5. *Math. Biosci. Eng.* 19, 12897–12914. doi: 10.3934/mbe.2022602
- Chen, B., Yan, J., Wang, K., and Matusů, R. (2021). Fresh tea sprouts detection via image enhancement and fusion SSD. *J. Control Sci. Eng.* 2021, 1–11. doi: 10.1155/2021/6614672
- Chen, X., Wei, F., Zeng, G., and Wang, J. (2022). Conditional DETR V2: efficient detection transformer with box queries. *arXiv preprint arXiv:2207.08914*.
- Chen, Y.-T., and Chen, S.-F. (2020). Localizing plucking points of tea leaves using deep convolutional neural networks. *Comput. Electron. Agric.* 171. doi: 10.1016/j.compag.2020.105298
- Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., et al. (2021). Dynamic Head: Unifying Object Detection Heads with Attention[J]. doi: 10.48550/arXiv.2106.08322
- Gong, T., and Wang, Z. (2021). "A tea tip detection method suitable for tea pickers based on YOLOv4 network," in *2021 3rd International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT)*. doi: 10.1109/ISRIMT53730.2021.9596977
- Guo, S., Li, L., Guo, T., Cao, Y., and Li, Y. (2022). Research on mask-wearing detection algorithm based on improved YOLOv5. *Sensors* 22 (13), 4933. doi: 10.3390/s22134933
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2980–2988. doi: 10.1109/ICCV.2017.322
- Hsu, W.-Y., and Lin, W.-Y. (2021). Adaptive fusion of multi-scale YOLO for pedestrian detection. *IEEE Access* 9, 110063–110073. doi: 10.1109/ACCESS.2021.3102600
- Kamilaris, A., and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 147, 70–90. doi: 10.1016/j.compag.2018.02.016
- Karunasena, G. M. K. B., and Priyankara, H. D. N. S. (2020). Tea bud leaf identification by using machine learning and image processing techniques. *Int. J. Sci. Eng. Res.* 11, 624–628. doi: 10.14299/ijser.2020.08.02
- Li, Y. T., He, L. Y., Jia, J. M., Chen, J. N., Lyu, J., and Wu, C. A. Y. (2022). High-efficiency tea shoot detection method via a compressed deep learning model. *Int. J. Agric. Biol. Eng.* 15, 159–166. doi: 10.25165/ijabe.20221503.6896
- Li, Y., Ma, R., Zhang, R., Cheng, Y., and Dong, C. (2023). A tea buds counting method based on YOLOV5 and Kalman filter tracking algorithm. *Plant Phenomics* 5, 0030. doi: 10.34133/plantphenomics.0030
- Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. *IEEE Computer Society*. doi: 10.1109/CVPR.2017.106
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., et al. (2019). Deep learning for generic object detection: A survey. *Int. J. Comput. Vision* 128, 261–318. doi: 10.1007/s11263-019-01247-4
- Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE*, 8759–8768. doi: 10.1109/CVPR.2018.00913
- Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., et al. (2022). RTMDet: an empirical study of designing real-time object detectors. doi: 10.48550/arXiv.2212.07784.
- Ma, N., Zhang, X., Zheng, H. T., and Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[J]. Springer, Cham. doi: 10.1007/978-3-030-01264-9_8
- Ma, S., and Xu, Y. (2023). MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression[J].
- Park, H., Yoo, Y. J., Seo, G., Han, D., Yun, S., and Kwak, N. (2018). C3: Concentrated-Comprehensive Convolution and its application to semantic segmentation. *arXiv: Comput. Vision Pattern Recognit.* doi: 10.48550/arXiv.1812.04920
- Qian, C., Li, M., and Ren, Y. (2020). Tea sprouts segmentation via improved deep convolutional encoder-decoder network. *IEICE Trans. Inf. Syst.* E103.D, 476–479. doi: 10.1587/transinf.2019EDL8147
- Rampriya, R. S., Suganya, R., Nathan, S., and Perumal, P. S. (2022). A comparative assessment of deep neural network models for detecting obstacles in the real time aerial railway track images. *Appl. Artif. Intell.* 36, 1–33. doi: 10.1080/08839514.2021.2018184
- Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv e-prints preprint arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767
- Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031
- Tan, M., Pang, R., and Le, Q. V. (2020). EfficientDet: Scalable and Efficient Object Detection[C]/2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). *IEEE*. doi: 10.1109/CVPR42600.2020.01079
- Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C., Lin, D., et al. (2019). CARAFE: Content-Aware ReAssembly of FEatures," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 3007–3016. doi: 10.1109/ICCV.2019.00310
- Wang, Y., Shao, P., Wu, M., Wang, X., Zhou, J., and Liu, S. (2018). Research on the tea bud recognition based on improved k-means algorithm. *MATEC Web Conferences* 232. . doi: 10.1051/mateconf/201823203050
- Wang, J., Xu, C., Yang, W., and Yu, L. (2021). A Normalized Gaussian Wasserstein Distance for Tiny Object Detection[J]. doi: 10.48550/arXiv.2110.13389
- Yan, L., Wu, K., Lin, J., Xu, X., Zhang, J., Zhao, X., et al. (2022). Identification and picking point positioning of tender tea shoots based on MR3P-TS model. *Front. Plant Sci.* 13, 962391. doi: 10.3389/fpls.2022.962391
- Yang, L., Zhang, R. Y., Li, L., and Xie, X. (2021). SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks[J].PMLR.
- Yang, H., Chen, L., Chen, M., Ma, Z., Deng, F., Li, M., et al. (2019). Tender tea shoots recognition and positioning by picking robot using improved YOLO-V3 model. *IEEE Access* 7, 180998–181011. doi: 10.1109/Access.6287639
- Zhang, S., Yang, H., Yang, C., Yuan, W., Li, X., Wang, X., et al. (2023). Edge device detection of tea leaves with one bud and two leaves based on shuffleNetv2-YOLOv5-lite-E. *Agronomy* 13, 577. doi: 10.3390/agronomy13020577
- Zhang, L., Zhang, H., Chen, Y., Dai, S., Li, X., Imou, K., et al. (2019). Real-time monitoring of optimum timing for harvesting fresh tea leaves based on machine vision. *Int. J. Agric. Biol. Eng.* 12, 6–9. doi: 10.25165/ijabe.20191201.3418
- Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., et al. (2022). Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybernetics* 52, 8574–8586. doi: 10.1109/TCYB.2021.3095305

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2024.1488185/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Classification image of tea buds. The blue box is single tea; the purple box is one leaf; the red box is two leaves.

SUPPLEMENTARY FIGURE 2

Pujiang Liangfeng Tea Plantation.

SUPPLEMENTARY FIGURE 3

C2F network structure. (A) C2F module; (B) CBS module.