Improved YOLOv8 network using multi-scale feature fusion for detecting small tea shoots in complex environments

Yatao Li[1,2,3], Liuhuan Tan[1], Zhenghao Zhong[1], Leiying He[1,4], Jianneng Chen[1,4*], Chuanyu Wu[1,4], Zhengmin Wu[2*]

(1. *School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China*;
2. *State Key Laboratory of Tea Plant Germplasm Innovation and Resource Utilization, Anhui Agricultural University, Hefei 230036, China*;
3. *Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry (Wuyi University), Wuyishan 354300, Fujian, China*;
4. *Key Laboratory of Transplanting Equipment and Technology of Zhejiang Province, Hangzhou 310018, China*)

Abstract: Tea shoot segmentation is crucial for the automation of high-quality tea plucking. However, accurate segmentation of tea shoots in unstructured and complex environments presents significant challenges due to the small size of the targets and the similarity in color between the shoots and their background. To address these challenges and achieve accurate recognition of tea shoots in complex settings, an advanced tea shoot segmentation network model is proposed based on You Only Look Once version 8 segmentation (YOLOv8-seg) network model. Firstly, to enhance the model's segmentation capability for small targets, this study designed a feature fusion network that incorporates shallow, large-scale features extracted by the backbone network. Subsequently, the features extracted at different scales by the backbone network are fused to obtain both global and local features, thereby enhancing the overall information representation capability of the features. Furthermore, the Efficient Channel Attention mechanism was integrated into the feature fusion process and combined with a reparameterization technique to refine and improve the efficiency of the fusion process. Finally, Wise-IoU with a dynamic non-monotonic aggregation mechanism was employed to assign varying gradient gains to anchor boxes of differing qualities. Experimental results demonstrate that the improved network model increases the AP50 of box and mask by 4.33% and 4.55%, respectively, while maintaining a smaller parameter count and reduced computational demand. Compared to other classical segmentation algorithms models, the proposed model excels in tea shoot segmentation. Overall, the advancements proposed in this study effectively segment tea shoots in complex environments, offering significant theoretical and practical contributions to the automated plucking of high-quality tea.

Keywords: tea shoot segmentation, multi-scale fusion, attention mechanism, reparameterization technique, YOLOv8-seg
DOI: 10.25165/j.ijabe.20251805.9475

Citation: Li Y T, Tan L H, Zhong Z H, He L Y, Chen J N, Wu C Y, et al. Improved YOLOv8 network using multi-scale feature fusion for detecting small tea shoots in complex environments. Int J Agric & Biol Eng, 2025; 18(5): 223–233.

1 Introduction

Tea holds substantial economic and cultural value worldwide[1], and tea harvesting is a crucial aspect of the tea industry. However, labor shortages and short harvesting cycles pose significant challenges for tea harvesting, necessitating the demand for automated tea harvesting solutions[2]. Due to its high economic value, high-quality tea requires stringent standards for plucking tea shoots, which are generally classified into three categories: single bud, one bud with one leaf, and one bud with two leaves[3]. Therefore, the harvesting of high-quality tea necessitates selective plucking methods, requiring the harvesting machines to possess a certain level of intelligence. The detection and segmentation of tea shoots are crucial for realizing intelligent picking. Consequently, research on tea shoot detection and segmentation is of great value and significance to the development of the tea industry.

Current crop segmentation primarily relies on traditional image processing and deep learning. Traditional methods segment targets based on local features, geometry, and pixel-level processing, but their reliance on manual feature design limits accuracy in complex tea shoot segmentation[4-6]. Deep learning has achieved remarkable success in computer vision, driving advancements in crop segmentation[7]. Kang et al.[8] combined instance and semantic segmentation in a first-order detection network, achieving high efficiency in apple detection. Liao et al.[9] employed background transfer learning and a color attention module to improve dandelion segmentation. For tea shoot detection, Xu et al.[10] proposed a two-level fusion network integrating YOLOv3[11] and DenseNet201, balancing speed and accuracy. Gui et al.[12] enhanced detection by introducing Ghost convolution[13] and a bottleneck attention module, reducing computational costs while improving precision. Li et al.[14] applied a pruned YOLOv3 model for real-time tea shoot picking. In

Received date: 2024-10-29 **Accepted date:** 2025-07-15
Biographies: Yatao Li, PhD, research interests: agricultural robot vision, Email: ytli@zstu.edu.cn; **Liuhuan Tan**, Postgraduate, research interests: image identification, Email: 2024210501044@mails.zstu.edu.cn; **Zhenghao Zhong**, Postgraduate, research interests: deep learning, Email: 202230503320@mails.zstu.edu.cn; **Leiying He**, Associate Professor, research interests: agricultural robot vision, Email: hlying@zstu.edu.cn; **Chuanyu Wu**, Professor, research interests: intelligent agricultural equipment, Email: cywu@zstu.edu.cn.
***Corresponding author: Jianneng Chen**, Professor, research interest: agricultural machinery equipment and technology. School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China. Email: jiannengchen@zstu.edu.cn; **Zhengmin Wu**, PhD, research interests: tea science. State Key Laboratory of Tea Plant Germplasm Innovation and Resource Utilization, Anhui Agricultural University, Hefei 230036, China. Email: wzmin@ahau.edu.cn.

the field of small target crop detection, Wu et al.[15] expanded the receptive field and fused multi-scale features via a multi-branch structure, improving weed detection. Xu et al.[16] designed a feature extraction module with grid resampling to enhance detection of inconspicuous small targets. Liu et al.[17] integrated a multi-scale extraction module and a dedicated small target detection layer, boosting accuracy for unopened cotton bolls. Research demonstrates that multi-scale feature extraction and fusion enhance small object segmentation in agriculture. However, tea bud images captured by picking robots present challenges: their small pixel proportion and color similarity to old leaves increase missed and false detections. Limited small-target features are further weakened or lost in deep network layers. While multi-scale fusion improves detection, cross-scale integration often fails to preserve fine-grained details, exacerbating spatial information loss. Additionally, the high color similarity between tea buds and background demands robust shape and texture discrimination, yet traditional feature weighting struggles to differentiate noise from true signals, increasing false positives. Thus, optimizing shallow feature utilization and minimizing information loss during multi-scale fusion is critical to improving small tea bud detection in complex environments.

Early target segmentation algorithms relied only on final feature maps for predictions, neglecting feature fusion in neck networks. Shallow features provide positional accuracy but lack semantics, while deep features offer rich semantics but poor localization—creating challenges for small target segmentation. To address this, Lin et al.[18] proposed Feature Pyramid Network (FPN), enabling unidirectional fusion by upsampling deep features and merging them with shallow ones. This balances high-level semantics and low-level details, improving small-target detection efficiently. Liu et al.[19] enhanced FPN with path aggregation network, adding bottom-up fusion to strengthen multi-scale representation. Tan et al.[20] advanced bidirectional fusion further via a weighted feature pyramid network, pruning low-contribution nodes and iterating layers for adaptive feature weighting. Zhang et al.[21] introduced TopFormer, using multi-scale tokens as inputs to generate perceptive features, which are injected back into original tokens. This token-based design enhances cross-scale perception. Wang et al.[22] refined TopFormer with Gather-and-Distribute (GD) modules for granular feature fusion across scales. Qian et al.[23] expanded feature pyramids by adding five layers and introducing max pooling and up-sampling pooling modules. These enable flexible multi-scale fusion while preserving critical spatial information. The development of feature fusion techniques has progressed from top-down fusion to simple bidirectional fusion, then to complex bidirectional fusion, and finally to the use of specialized modules to aid in fusion. Top-down fusion often results in the continuous dilution of information from the topmost layers as it progresses downward. In contrast, bidirectional fusion, while more effective, does not fully utilize secondary information for cross-layer information fusion, especially across multiple layers. Moreover, the introduction of specially designed complex modules, despite improving fusion capabilities, can lead to a reduction in efficiency.

To address these issues, this study proposes an enhanced tea shoot segmentation model based on the YOLOv8 target segmentation algorithm. The model is targeted to a series of improvements on YOLOv8 according to the complex environment of the actual growth of tea. The modified model enhances the segmentation accuracy of tea shoots in the field, thereby facilitating precise plucking by tea plucking robots. The key contributions of this study are as follows:

1) Designed a neck fusion network for small targets. The use of large feature maps of the backbone network was added to the neck fusion network. The features are divided into global and local features for multi-scale fusion.

2) Adding ECA attention mechanism and combining with reparameterization technique in the feature fusion process to capture key features and improve the computational efficiency. Realize the efficient and fine fusion of features.

3) Use Wise-IoU instead of CIoU. Reduce the competitiveness of high-quality anchor boxes while reducing the harmful gradient generated by low-quality examples. The new loss function focuses on average-quality anchor boxes, thus improving the overall performance of segmentation.

2 Materials and methods

2.1 Data preparation

2.1.1 Data acquisition

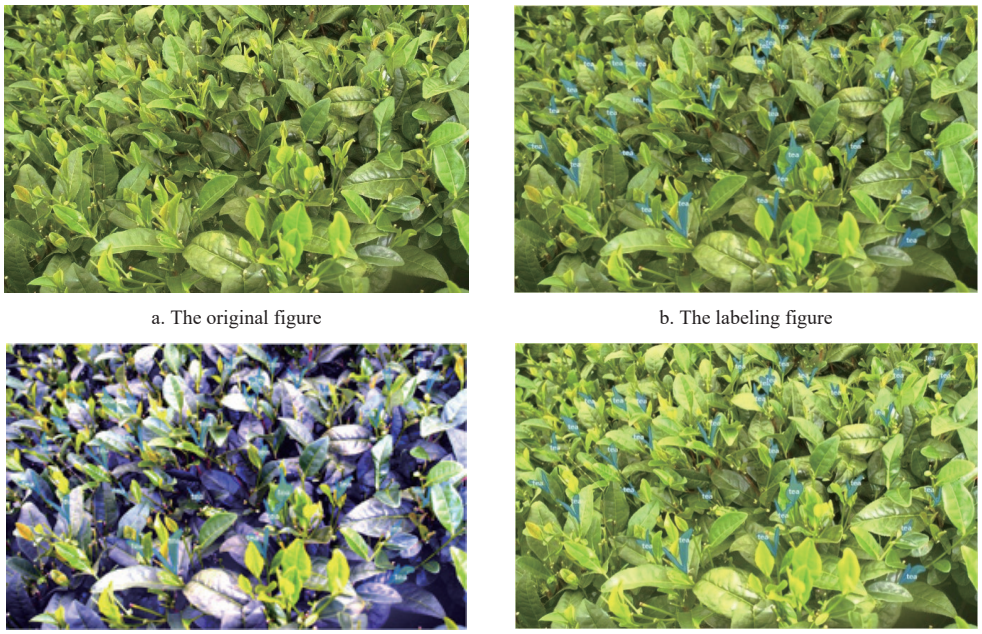
The tea data utilized in this study were collected in August 2023 from Songyang County, Lishui City, Zhejiang Province, China. The tea variety is Longjing. During the collection process, the center of the camera lens is positioned at a vertical distance of 0.25-0.30 m from the surface of the tea plant. This distance ensures that a single frame image fully covers the target tea buds within the standard tea row width range and maintains the tea buds within an ideal size range in the image. The camera's optical axis was oriented at an angle of 55°±2° relative to the ground plane. This angle maximizes the visible surface area of the tea buds to minimize leaf obstruction while effectively suppressing distant, unclear interference targets. Data collection was conducted in two sessions, morning and afternoon, systematically covering typical tea garden lighting scenarios, including direct sunlight, diffuse reflection, and leaf transmission. This ensured diversity in the lighting conditions of the dataset. A total of 850 target images were collected.

2.1.2 Data generation

Using the data clarity feature of the Baidu Machine Learning (BML) platform, 850 images from the initial dataset were screened. The screening criteria were to remove images with low clarity due to environmental disturbances and those with excessive similarity. The annotation criteria followed the "one bud, one leaf" standard for premium tea picking, annotating all areas above the node and 5 cm below the node. All images in the dataset were annotated using the instance segmentation annotation tool on the BML platform, and the annotation results were manually reviewed to verify the accuracy of the annotation of bud and leaf morphological features. Following the annotation process, offline data augmentation was performed on these images. The augmentation techniques employed were autocontrast and brightness adjustments, each of which effectively doubled the original dataset with labels. This augmentation process culminated in a total of 2340 labeled tea shoots images. AutoContrast enhances the contrast of images, thereby increasing the distinction between tea shoots and the background, which is particularly beneficial for differentiating tea shoots from older leaves. Brightness adjustment modifies the image brightness, simulating the varying lighting conditions encountered in real outdoor environments from morning to night. This adjustment is essential for increasing the dataset's diversity with respect to light intensity. The visual comparison of images before and after augmentation is depicted in Figure 1.

After completing the augmentation, the dataset was divided, resulting in a tea segmentation dataset comprising 2100 training

images and 240 validation images. Although the dataset size is relatively modest, the tea targets are notably small and densely packed, with each image containing between 20 to 40 targets. The statistics show that there are 62 985 labeled targets in the dataset. These tea buds exhibit significant diversity in terms of growth posture, spatial arrangement, and lighting conditions, demonstrating a certain degree of broad applicability. Most of these targets are between 5%-10% of the height of the picture, and the width is between 2%-5% of the width of the picture, which is typical of small targets.



a. The original figure b. The labeling figure

c. The labeling figure after autocontrast enlargement d. The labeling figure after brightness enlargement

Figure 1 Figure enlargement and labeling

2.2 YOLOv8-seg model introduction

YOLOv8, developed by Ultralytics in January 2023, represents a significant enhancement and optimization over previous YOLO versions, delivering notable improvements in image classification, object detection, and instance segmentation[24]. The backbone network employs the CSPDarknet architecture[25], designed to extract feature information at various scales. The neck network incorporates the Path Aggregation Network (PAFPN) structure, an extension of the FPN. PAFPN adds a bottom-up pathway to the traditional FPN, addressing the issue of insufficient detail in deeper features extracted from shallow features. Consequently, PAFPN can capture richer feature information. Within the neck network, the C2f module facilitates feature fusion and enhancement, improving feature expression capability and network efficiency through cross-stage local connections. The head of YOLOv8 utilizes a decoupled head structure, which separates target location and category information into distinct output layers. For its loss function, YOLOv8 employs the Complete Intersection over Union (CIoU)[26] as the regression loss. CIoU improves regression accuracy by considering the relative proportions of detection boxes and incorporating aspect ratios. Overall, the YOLOv8 algorithm demonstrates exceptional performance across various tasks, achieving state-of-the-art accuracy on multiple datasets with rapid detection speeds. Consequently, this model was selected as a benchmark for tea shoot detection.

2.3 Tea segmentation model

2.3.1 Model improvement and optimization methods

Based on the superior performance of the YOLOv8 algorithm, this study chose YOLOv8 as the baseline model for the tea shoot segmentation algorithm, as shown in Figure 2. The backbone of this algorithm follows the structure of the baseline model, utilizing the CBS module and the C2f module to extract features. The backbone network extracts four types of feature maps: 160×160, 80×80, 40×40, and 20×20. The 160×160 feature map has a small receptive field and high resolution, excelling at capturing low-level information such as the fine textures and blurred edges of tea buds, but it has weak semantic discrimination capabilities and may confuse tea buds with similar-shaped old leaves. The 80×80 feature combines both detail and structural information, effectively distinguishing tea buds from interfering objects such as leaves and branches. The 40×40 feature possesses strong semantic representation capabilities, enhancing target discrimination in complex backgrounds, but at the cost of some detail loss. The 20×20 feature, though weakened in small target detection due to its large receptive field, provides critical supplementary contextual information for higher-level semantic understanding through multi-scale fusion. Since this segmentation target, tea buds, are mostly small objects in the actual environment, the shallow large features extracted by the main network are very helpful for the segmentation of small objects. Therefore, compared with the baseline model, this study increased the use of feature maps with a size of 160×160. This study input all four feature maps into the neck multi-scale feature fusion network for feature fusion. A novel multi-scale feature fusion structure is introduced to facilitate the effective fusion of the four sizes of feature maps. This structure incorporates the fusion of global features, local features, and their combination. Initially, the feature maps extracted from the backbone network are fused to generate global features, which encapsulate both shallow and deep information. Subsequently, feature maps of the same scale as the output header are fused with their neighboring feature maps to obtain local features. To optimize the fusion of global and local features, a parameter-free attention mechanism and reparameterization technique was employed[27]. Finally, the fused features are passed to the detection head using a bottom-up

approach, maintaining the same structure as the segmentation head of YOLOv8 without any modifications. This comprehensive approach ensures the effective segmentation of tea shoots while leveraging the strengths of the YOLOv8 framework.

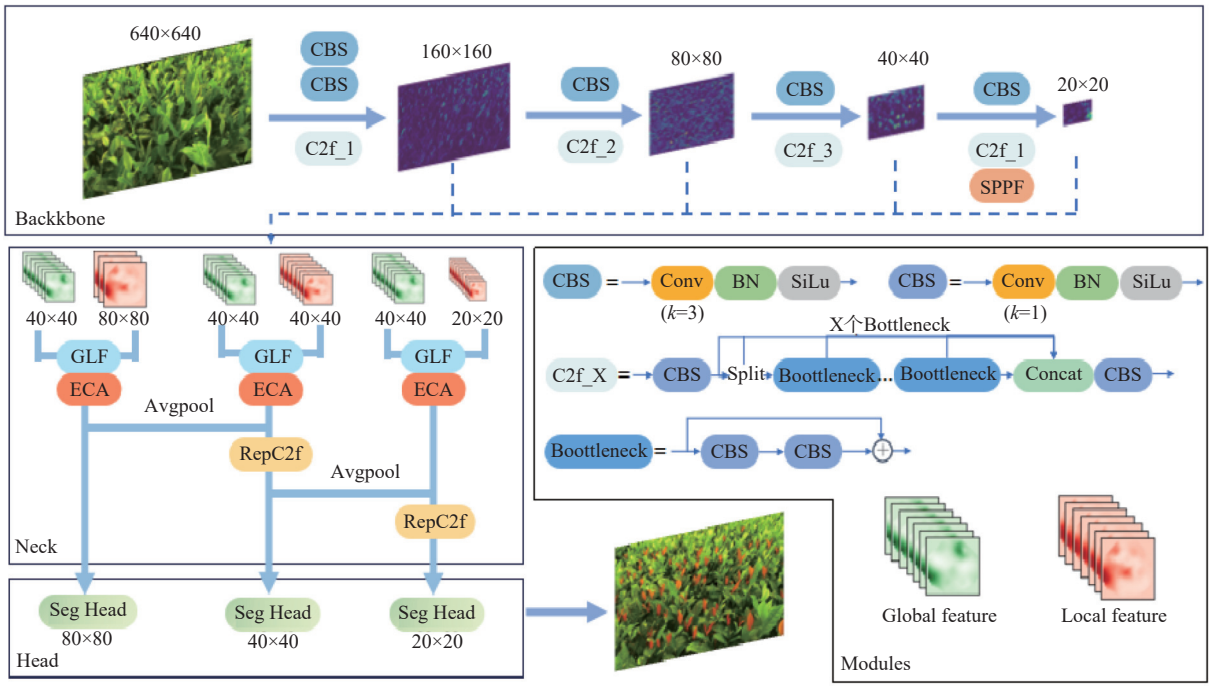


Figure 2 Improvement of model network structure

2.3.2 Multi-scale feature fusion

The specific acquisition of global features is illustrated in Figure 3. Prior to the acquisition of global features, it is necessary to select the scale of the global features. In this case, 40×40 was chosen as the scale for feature fusion. However, if 20×20 had been selected, a significant amount of underlying information would have been lost, which would have been disadvantageous for the detection of small targets. The use of scales such as 80×80 or even 160×160 would necessitate the allocation of greater computational resources when processing in subsequent modules. In light of the aforementioned considerations, it can be posited that a fusion scale of 40×40 is optimal in terms of both accuracy and speed. Once the fusion scale has been established, the feature maps of the remaining scales are aligned. The 160×160 and 40×40 feature maps are then downsampled using average pooling, while the 20×20 feature maps are upsampled using bilinear interpolation. Finally, the four aligned features are fused by the CBS module. The formula for obtaining global features is as follows:

$$\hat{X}_l = \text{AvgPool}(X_l, \text{output\_size} = (H, W)) \quad (1)$$

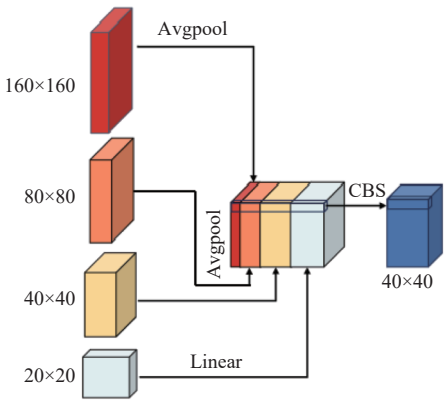


Figure 3 Structure of global feature fusion module

$$\hat{X}_m = \text{AvgPool}(X_m, \text{output\_size} = (H, W)) \quad (2)$$

$$\hat{X}_n = \text{Linear}(X_n, \text{output\_size} = (H, W)) \quad (3)$$

$$\hat{X}_{\text{global}} = \text{CBS}\left(\text{Concat}\left(\hat{X}_l, \hat{X}_m, \hat{X}_s, \hat{X}_n\right) \in R^{B\times 4C\times H\times W}\right) \quad (4)$$

where, X_l, X_m, X_s, X_m are four different feature scales from largest to smallest, H and W are the height and width of the features, AvgPool is average pooling, Linear is linear interpolation, and X_{global} is the global feature.

With regard to the acquisition of local features, the fusion method employed is analogous to that used for global features. The scale of the local features determines the final output to the detection head of the feature map scale. The processing of 160×160 size features requires a significant amount of computational resources, which is why the choice of 80×80, 40×40, 20×20 as the scale of the three local features was made. The selected scales are employed as a benchmark for aligning features of varying sizes. Following alignment, the fusion process is conducted. It should be noted that this study has attempted to determine the optimal combination of features extracted from the backbone network for use in generating local features. The selected combinations and the generated local features are presented in Section 3.3. Thus far, three types of local features have been obtained, with sizes of 80×80, 40×40, and 20×20, respectively.

Once the requisite global and local features have been obtained, they need to be fused using the global-local feature fusion (GLF) module. Figure 4 demonstrates the specific fusion method using an 80×80 local feature as an example. This process involves the fusion of three distinct types of local features, each with a different size, with the global features. Firstly, the local features are subjected to a convolutional block with a 1×1 convolution. This operation is intended to combine and transform the features in each channel, thereby increasing the nonlinearity of the network. At the same time, the channels of the local features are adjusted for subsequent

weighted fusion. Global features are divided into two parallel branches. The first branch integrates local features, performs scale-based feature alignment, and then processes them through a 1×1 convolution within the CBS block, activated by the Sigmoid function to produce a set of weights. The second branch consists only of local features that likewise undergo scale-based feature alignment.Once the aforementioned steps have been completed, the local features will be processed and the first branch will obtain the weight, which will then be multiplied with the second branch to obtain the results of the addition. Subsequently, the processed local features will be multiplied with the weights obtained from the first branch and then added with the results obtained from the second branch. Finally, the features will be further extracted and fused by a RepC2f module. At this juncture, the fusion of global and local features is complete. The formula for combining global and local features is as follows:

$$F_l = \mathrm{CBS}_{1\times1}(X_{\mathrm{local}}) \tag{5}$$

$$A_g = \mathrm{sigmoid}(\mathrm{CBS}_{1\times1}(\mathrm{Linear}(X_{\mathrm{global}}, (H, W)))) \tag{6}$$

$$F_g = \mathrm{CBS}_{1\times1}(\mathrm{Linear}(X_{\mathrm{global}}, (H, W))) \tag{7}$$

$$Y = F_l \otimes A_g \oplus F_g \tag{8}$$

where, $\mathrm{CBS}_{1\times1}$ is the CBS module using 1×1 convolution, F_l is the local feature part in the fusion process, F_g is the entire feature part in the fusion process, A_g is the global feature weight in the fusion process, $\otimes$ is element-wise multiplication, and $\oplus$ is element-wise addition.

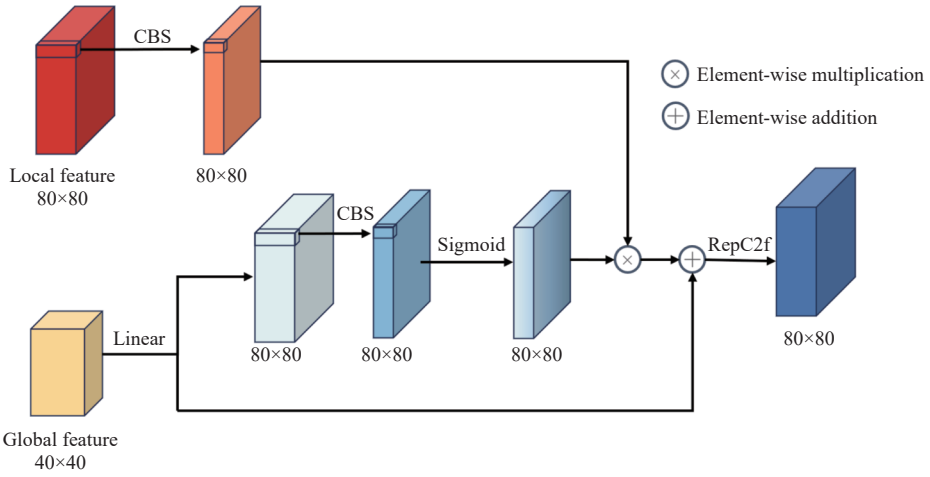


Figure 4 Structure of global-local feature fusion module

2.3.3 Lightweight fusion

At the end of the local feature fusion, global feature fusion, and local feature and global feature fusion modules, this study has built-in Efficient Channel Attention[28]. ECA is a lightweight channel attention mechanism, which, through a local cross-channel interaction strategy without dimensionality reduction, can help it to focus on the really important parts of the channel during the feature fusion process and thus improve the performance of the model. The structure of ECA is shown in Figure 5. ECA first performs Global Average Pooling on the input feature maps to obtain a global feature description for each channel. Then one-dimensional convolution is used to capture the inter-channel dependencies. Finally, the results obtained by 1D convolution are used as channel weights to weight the original feature maps. ECA introduces almost no parameters, and in this model, a single use of ECA introduces only three parameters, and ultra-lightweight is the main reason why it is chosen.

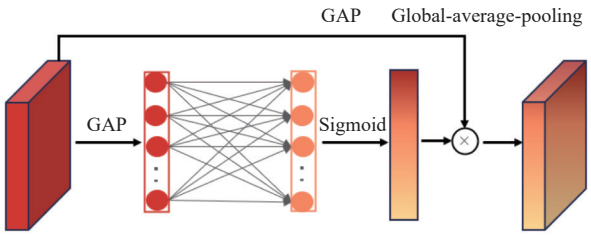


Figure 5 Structure of efficient channel attention module

In the context of local and global feature fusion, as well as top-down fusion, the reparameterization technique is employed to adjust the convolutional layers within the C2f residual block. The fundamental principle of the reparameterization technique is to decompose a complex convolutional operation into a series of elementary convolutional operations for training purposes, and subsequently to merge these elementary convolutional operations into an equivalent complex convolutional operation in the inference stage. The conversion diagram is shown in Figure 6. During training, a multi-branch structure is adopted to enhance training effectiveness. The main branch consists of a 3×3 convolution followed by a BN activation function; the auxiliary branch consists

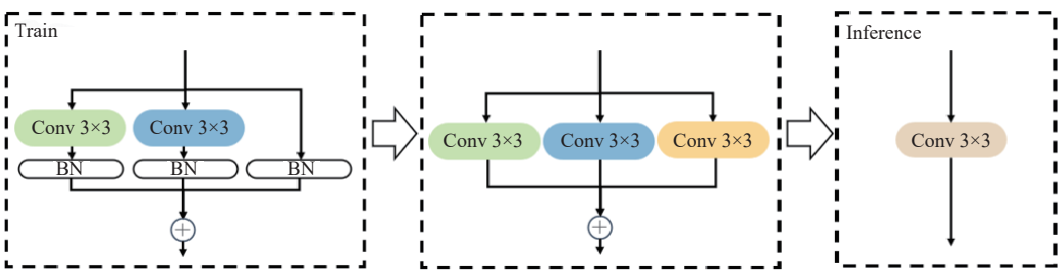


Figure 6 Schematic diagram of reparameterization training inference conversion

of a 1×1 convolution followed by a BN activation function; the third branch is an identity mapping branch, used when the input and output channel counts are the same. During inference, the three trained branches are converted into a single 3×3 convolution. The identity branch can be viewed as a 1×1 convolution. First, the conv and BN operations are converted into a convolution with a bias term. Then, the edges of the two 1×1 convolution kernels are padded with zeros to form a 3×3 convolution kernel. Finally, the three convolution kernels are summed together. This method enhances the expressive capacity and training efficiency of the model while maintaining the inference efficiency.

2.3.4 WIoU loss algorithm

Accurate target localization is a pivotal step in target detection algorithms, achieved through the regression of the bounding box. Wise-IoU (WIoU)[29] introduces a dynamic non-monotonic aggregation mechanism and proposes evaluating the quality of the anchor box based on its degree of outlier. This method employs a gradient gain assignment strategy to diminish the competitiveness of high-quality anchor boxes while mitigating the detrimental gradient effects of low-quality samples.

WIoUv3 is improved by WIoUv1; WIoUv1 can be obtained from Equation (9)-(11). IoU is utilized to measure the degree of overlap between the prediction box and the real box. P_{WIoU} reflects the degree of attention to the center distance.

$$\mathcal{L}_{\mathrm{IoU}} = 1 - \mathrm{IoU} = 1 - \frac{W_i H_i}{S_u} \in [0,1] \tag{9}$$

$$\mathcal{R}_{\mathrm{WIoU}} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{\left(W_g^2 + H_g^2\right)}\right) \in [1,e) \tag{10}$$

$$\mathcal{L}_{\mathrm{WIoUv1}} = \mathcal{R}_{\mathrm{WIoU}}\mathcal{L}_{\mathrm{IoU}} \tag{11}$$

where W_i and H_i represent the width and height of the overlapping part, and S_u denotes the area of the prediction box and the real box minus the overlapping part. The coordinates (x, y) and (x_{gt}, y_{gt}) correspond to the centers of the prediction and real boxes. W_g and H_g are the width and height of the smallest rectangular box that encircles the prediction box and the real box.

WIoUv3 introduces the concept of outlier β to assess the quality of the anchor box. Utilizing β, along with the predefined fixed values α and δ, the non-monotonic aggregation coefficient r is constructed and subsequently applied to WIoUv1. Λ^*_{IoU} denotes the monotonic focus factor. This dynamic gradient assignment strategy optimizes gradient allocation according to real-time conditions, thereby mitigating the influence of low-quality samples that could produce detrimental gradients. The formula for WIoUv3 is as follows:

$$\beta = \frac{\mathcal{L}^*_{\mathrm{IoU}}}{\mathcal{L}_{\mathrm{IoU}}} \in [0,+\infty) \tag{12}$$

$$\mathcal{L}_{\mathrm{WIoUv3}} = r\mathcal{L}_{\mathrm{WIoUv1}}, \quad r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{13}$$

The study focuses on tea shoots, which present a challenging environment characterized by dense growth, small target size, variable angles, and frequent occlusions, resulting in the presence of some low-quality samples. In such scenarios, applying a static approach to all samples, including low-quality ones, can inadvertently enhance the fitting loss and compromise the model's generalization capability. To address this issue, this study employed WIoUv3 in place of CIoU within the YOLOv8 framework, leveraging the advanced dynamic mechanisms of WIoU to improve detection performance under these complex conditions.

3 Experimental results and analysis

3.1 Evaluation indicators

The complexity and performance evaluation of deep learning models typically employs metrics such as precision, recall, average precision (AP), the number of parameters, and floating-point operations per second (FLOPs).

Precision quantifies the probability of true positive samples among all samples predicted to be positive, while recall measures the probability of true positive samples among all actual positive samples. Average precision (AP) is calculated as the area under the precision-recall curve, representing the mean precision across varying recall levels. These metrics collectively provide a comprehensive evaluation of model performance and are computed using the following equations:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{AP}\alpha = \int_0^1 p(r)\,\mathrm{d}r \tag{16}$$

where, TP, FP, and FN are the number of true positive cases, false positive cases, and false negative cases, respectively. $P(r)$ denotes the precision when the recall is r, and α denotes the IoU threshold. A larger value of α represents a more stringent prediction requirement; in this paper, α is set to 0.5, and AP50 is used to denote the average accuracy of the model in subsequent experiments in this paper.

Parameters directly influence the storage demands of a model, while FLOPs quantify the computational workload during model inference. These metrics serve as pivotal benchmarks for assessing model complexity and computational resource utilization. Computation of these metrics is facilitated by the following equations:

$$\mathrm{Parameter} = C_{\mathrm{in}} \times C_{\mathrm{out}} \times K \times K \tag{17}$$

$$\mathrm{FLOPs} = \left(2C_{\mathrm{in}} \times K^2 - 1\right) \times H_{\mathrm{out}} \times W_{\mathrm{out}} \times C_{\mathrm{out}} \tag{18}$$

where, C_{in} and C_{out} are the number of channels of the input and output convolutional layers, K denotes the size of the convolutional kernel, and H_{out} and W_{out} are the height and width of the output feature map of the convolutional layer.

3.2 Experimental platform and training settings

This experiment was performed on a computer with 64 GB of RAM on Ubuntu 16.04 LTS system, an Intel i7-9800 8-core CPU and four NVIDIA RTX2080Ti 11GB GPUs. The learning task was performed on python 3.8.17 using the computing platform CUDA10.2, cuDNN 7.6.5 with pytorch 1.8.0.

The experimentation leverages the YOLOv8 codebase, which offers a spectrum of model scales designated as *n*, *s*, *m*, *l*, and *x*, where *n* denotes the smallest model. To ensure methodological consistency, all training parameters remain uniform, with pre-training weights being omitted. The training was performed for a total of 200 epochs, the batch size was set to 32, and the image input size was 640×640. Training is executed employing the SGD optimizer, initialized with a learning rate of 0.01. To speed up the training process and to avoid overfitting, the impulse parameter was set to 0.937 and the weight decay coefficient to 0.0005.

3.3 Local feature fusion selection experiments

An ablation study is performed in this section to investigate

which block or blocks of features extracted using the backbone network can lead to the best performance of the model when fused to obtain local features.

The backbone network extracts feature maps of four different sizes, labeled sequentially from shallow to deep as 1 through 4. Specifically, the values 1, 2, 3, and 4 correspond to the dimensions of 160×160, 80×80, 40×40, and 20×20 feature maps, respectively. The neck network requires three distinct sizes of local features. The notation (2,3,4) indicates that the feature maps labeled 2, 3, and 4 are used independently as local features. The notation (12,23,34) signifies the fusion of feature maps 1 and 2, 2 and 3, and 3 and 4, serving as the first, second, and third local features, respectively. Similarly, (123,234,134) represents the fusion of three different sizes of feature maps to form each local feature. Figure 7 and Figure 8 display the AP50 and loss metrics of the model during training with various local feature fusion strategies. These figures demonstrate that the new neck network facilitates faster convergence and improved AP50.

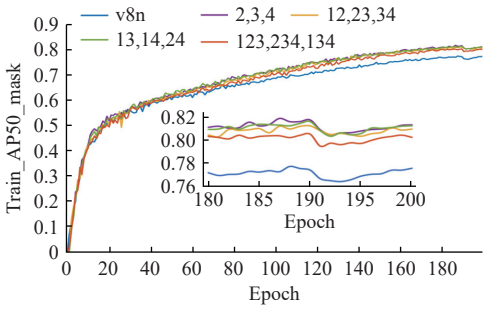


Figure 7 Train_AP50_mask with different local feature fusion strategies

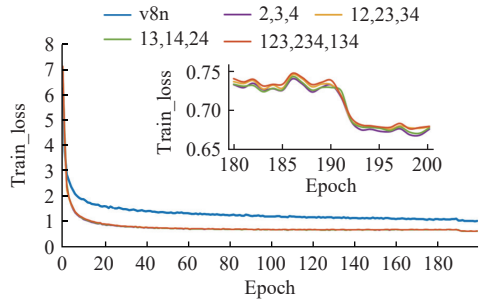


Figure 8 Train_loss with different local feature fusion strategies

Table 1 presents the performance outcomes of models employing various local feature fusion strategies on the validation set. The results indicate a consistent improvement in performance over YOLOv8n-seg's PAFPN, irrespective of the local feature fusion strategy used. Notably, the model utilizing a single feature map as the local feature achieves the highest performance. This may be due to the global feature already encapsulating information from the four different-sized feature maps, and excessive fusion leading to information redundancy. Compared to YOLOv8n-seg, the model with the optimal fusion strategy shows improvements of 4.33% in AP50 and 4.08% in Recall for box, and 4.55% in AP50 and 3.98% in Recall for mask. The neck structure introduced in this study increases the fusion operations of feature maps of varying sizes. However, due to the reparameterization of C2f, the overall number of parameters for the model under the optimal fusion strategy is less than that of YOLOv8n-seg. Models using the fusion of two or three feature maps as local features exhibit a gradual increase in parameter count compared to the original model. Regarding FLOPs performance, YOLOv8n-seg achieves the best results, primarily because the fusion of different-sized feature maps necessitates operations such as feature alignment, thereby increasing computational effort. Additionally, YOLOv8n-seg, as the smallest model in the YOLOv8 series, achieves an inference speed of 1.21 millisecond per image on A100 TensorRT. Therefore, given that 12.61G FLOPs require only 1.21 millisecond, the additional 0.78G FLOPs for a nearly 4% improvement in AP50 and Recall is a worthwhile trade-off.

Table 1 Results of different local feature fusion strategies

Local feature	Parameters/ M	FLOPs/ G	AP50_ box/%	Recall_ box/%	AP50_ mask/%	Recall_ mask/%
(2,3,4)	**3.22**	13.39	**86.02**	**79.21**	**81.60**	**75.99**
(12,23,34)	3.33	13.82	85.38	78.54	80.28	74.46
(13,14,24)	3.35	13.99	85.58	78.62	80.92	75.93
(123,234,134)	3.35	14.10	84.87	78.46	79.82	74.74
YOLOv8n-seg	3.26	**12.61**	81.69	75.13	77.05	72.01

3.4 Ablation study of module

To better understand the contributions of each improvement to the model's performance enhancement, this study conducted a related ablation study. The results of this study are presented in Table 2. The data clearly demonstrate that the model incorporating the new neck network outperforms the original YOLOv8n-seg model. This finding underscores the pivotal role of the multi-scale feature fusion method in boosting model performance. Furthermore, the inclusion of the ECA attention mechanism allows the model to better focus on crucial information during the feature fusion process, with minimal impact on parameter count. Additionally, the dynamic non-monotonic aggregation mechanism of WIoU contributes to further performance gains. The hexagram of the ablation study results, shown in Figure 9, clearly illustrates that each module enhances the model's performance. The optimal performance is achieved when all modules are integrated, highlighting the synergistic effect of the proposed improvements.

Table 2 Ablation study of module

Model	YOLOv8n-seg	YOLOtea	YOLOtea-ECA	YOLOtea-ECA-WIoU
New neck	-	√	√	√
ECA	-	-	√	√
WIoU	-	-	-	√
Parameters/M	3.26	3.22	3.22	**3.22**
FLOPs/G	**12.61**	13.39	13.39	13.39
AP50_ box/%	81.69	84.01	84.41	**86.02**
Recall_ box/%	75.13	77.46	78.21	**79.21**
AP50_ mask/%	77.05	79.42	80.91	**81.60**
Recall_ mask/%	72.01	74.24	75.05	**75.99**

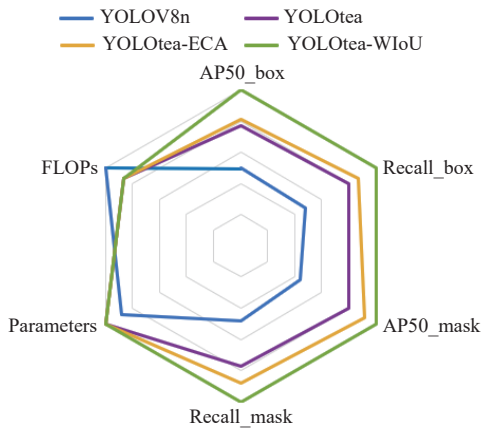


Figure 9 Ablation research hexagram

3.5 Experimental comparison of different segmentation algorithms

This section presents a comparison between this improved model and several other models, including Mask R-CNN, SOLOv2, Point rend, YOLOv5-seg, and YOLOv8-seg. This improved model selected the fusion strategy of (2,3,4) as the optimal local feature fusion strategy. The experimental results presented in Table 3 demonstrate that the improved model proposed in this paper exhibits clear advantages in terms of the number of parameters, AP50_box, AP50_mask, and inference speed.

Table 3 Experimental comparison of different segmentation algorithm models

Model	Parameters/M	FLOPs/G	Backbone scale	Box_AP50/%	Mask_AP50/%	FPS/img·s^{-1}
Mask R-CNN	43.97	115.25	Resnet50	88.64	84.78	15.38
SOLOv2	46.23	248.20	Resnet50	-	85.60	14.49
Point rend	55.76	64.76	Resnet50	91.89	89.12	13.70
YOLOv5-seg	2.76	11.09	*n*	78.98	74.65	129.87
	9.78	38.09	*s*	90.62	86.87	57.14
YOLOv8-seg	3.26	12.61	*n*	81.69	77.05	121.95
	11.79	42.69	*s*	92.52	89.03	53.76
TeaYolo	3.22	13.39	*n*	86.02	81.60	98.04
	9.80	41.58	*s*	93.83	89.78	53.48

In this study, the final improved model YOLOtea-ECA-WIoU is defined as TeaYolo, the *n* and *s* in the table indicate the scale of the model, and the larger the scale the deeper the network is. The combination of the YOLOv8n-seg scale backbone network and the proposed improvements yields the most pronounced enhancements, as evidenced by the 4.33% and 4.55% increases in AP50_box and AP50_mask, respectively, compared to the original YOLOv8n-seg model. However, the inference speed is slightly reduced, yet the 98.04 FPS remains sufficient for real-time detection. The enhancement of the YOLOv8s backbone network has diminished, which may be attributed to the fact that the AP50 _box has reached 93.83% and the AP50 _mask is approaching 90%, which represents a high level of performance. In comparison to the non-YOLO series of Mask R-CNN, SOLOv2, and Point rend, the TeaYolo-*n* model does not demonstrate an advantage in terms of accuracy, but it is significantly faster. The TeaYolo-*s* model has surpassed the aforementioned non-YOLO series models in terms of accuracy while maintaining a substantial lead in speed. The parameters T and FPS indicate that the YOLO series model has a significant advantage in real-time performance while maintaining high accuracy. This is due to the advantages of its algorithmic structure, which is also the reason why it is widely acknowledged. This proves that it is correct in choosing the YOLO model as the baseline model. In comparison with the analogous series of YOLO models, the YOLOv8-seg model exhibits greater strength than the YOLOv5-seg model of the same scale in both *n* and *s*. Compared with the YOLOv8-seg model of the same scale, the model proposed in this study shows the intensity of the enhancement, thus confirming the efficacy of the augmentation of the model of this study.

In comparison with other models, the enhanced model in this study achieves the best balance between performance and inference speed. In particular, the TeaYolo-*n* model achieves AP50_box and AP50_mask of 86.02% and 81.60%, respectively, at a small scale of 3.22 M parameters and a high inference speed of 98.04 FPS. In comparison with the non-YOLO series and the same scale YOLO series, the TeaYolo-*s* model achieves the optimal performance in all metrics.

3.6 Visualization results on the tea dataset

To intuitively compare the model performance before and after improvements, the tea shoot samples were selected for visual analysis of the segmentation results. To simulate the real-world tea plucking environment and evaluate the generalization performance of the model, in addition to the dataset images, this study also uses images from the perspective of the tea plucking robot developed by our team. The location of the picking robot and camera is shown in Figure 10. This study represents dataset images as class I and robot viewpoint images as class II. In this study, the models developed using various local feature fusion strategies were compared with the original YOLOv8n-seg model, and the results are shown in Figure 11. The targets segmented by these five models were collated and evaluated. Correct targets were manually marked on the original images with red detection boxes as references. The segmentation results from each model were then compared against these references. Orange masks represent the segmentation results of the respective models, while white and black boxes indicate missed and incorrectly detected targets, respectively, compared to the reference examples.

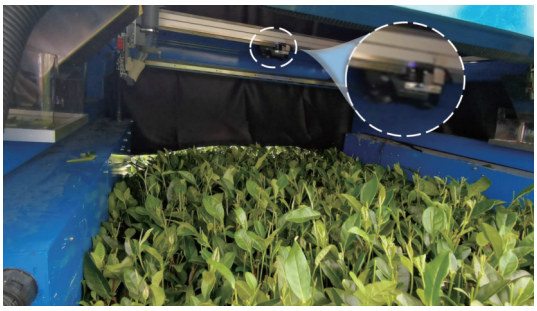
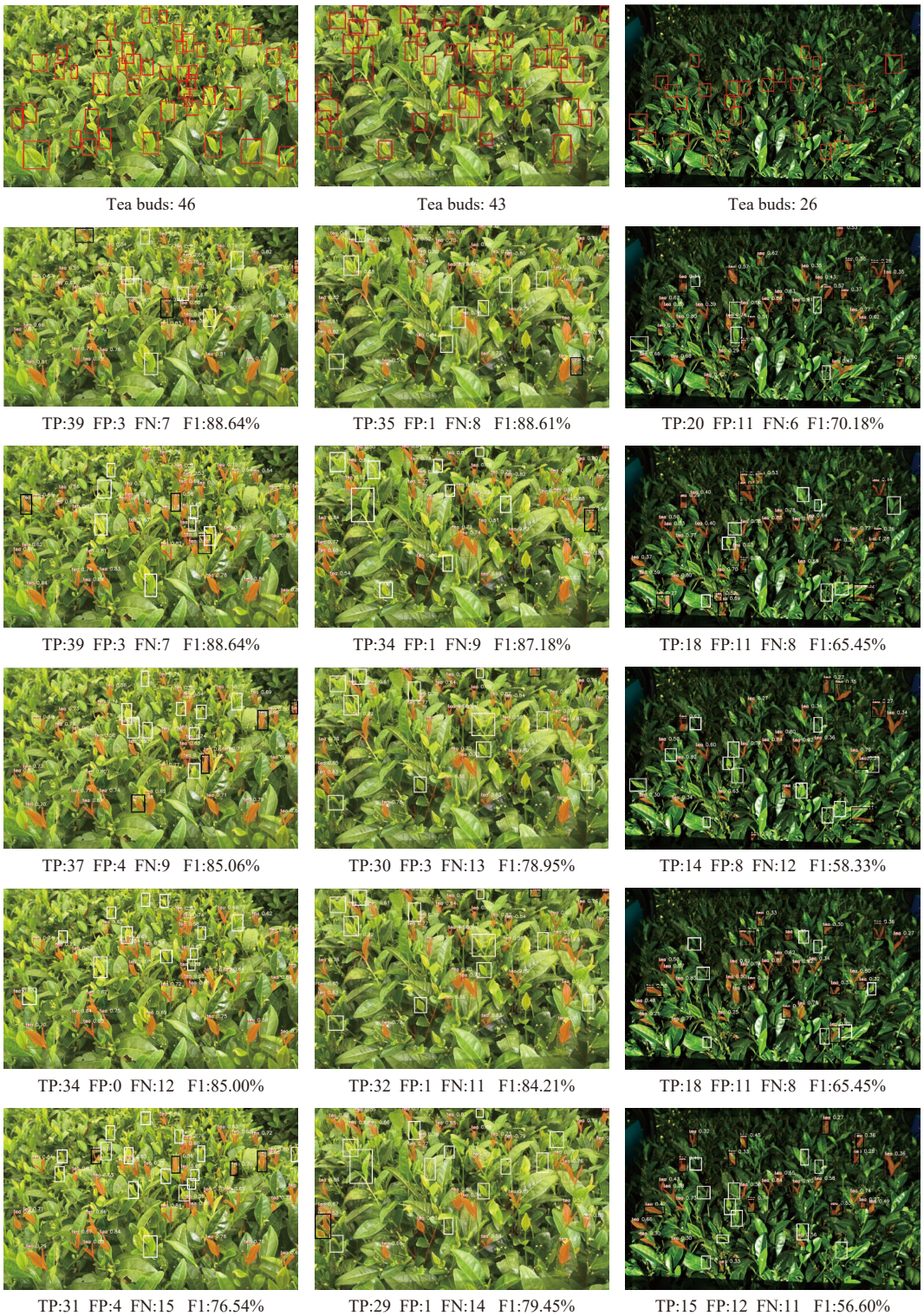


Figure 10 Spatial arrangement of camera and tea plucking robot

The results reveal that for Class I images, the model employing the optimal local feature fusion strategy maximizes the segmentation of target tea shoots in the images. Despite a decrease in performance, models utilizing non-optimal local feature fusion strategies still outperform the YOLOv8n-seg model. Moreover, the improved model exhibits fewer instances of mask breakage, indicating higher mask quality. Even when confronted with more small targets in distant views, the improved model demonstrates the capability to segment a considerable number of tea shoots, with significantly better results than the YOLOv8n-seg model. For Class II images, the improved model exhibits improved performance compared to the original model. However, due to the limitation of the number of datasets and the large gap in style between the image types and the training images, there are more missed and wrongly detected targets than Class I images.

4 Discussion

From the above results, it can be seen that since the model used 160×160 large feature maps, the improvements proposed by this model have produced good results for tea buds, especially for smaller tea buds. Figure 12 presents a visualization of the four distinct sizes of feature maps extracted from the backbone network, ranging from shallow to deep. Notably, the shallow network features exhibit higher resolution, encapsulating richer location and detailed information, but lack semantic depth. Conversely, deeper network features possess stronger semantic information yet exhibit lower resolution and limited perception of details. Furthermore, variations in the receptive fields of different-sized feature maps are discernible. Figure 13 illustrates the results of Grad-CAM[30] heat



Note: a. Manual box labeling b. Improved model under (2,3,4) local feature strategy c. Improved model under (12,23,34) local feature strategy d. Improved model under (24,13,14) local feature strategy e. Improved model under (123,243,134) local feature strategy f. YOLOv8n-seg model

Figure 11 Visualization results of the original YOLOv8n-seg model and the improved model using different local features on tea samples

map visualization for various layers of the backbone network, with the second layer housing the 160×160 feature maps and the ninth layer housing the 20×20 feature maps. As depicted in the figure, as the neural network delves deeper, the sensory fields of the extracted feature maps progressively expand. For small targets occupying only a fraction of the image's local area, a larger receptive field may overlook these targets, leading to inaccurate localization. Thus, the utilization of large feature maps extracted from the shallow network layers proves beneficial in assisting the detection of such small targets.

Recognizing that large feature maps are favorably consistent with the goals of this study, this study initially tried to directly utilize large feature maps within the framework of the original model. This entailed extending the bottom-up branch of the neck network upwards, facilitating the fusion of the 80×80 feature maps with the 160×160 shallow large feature maps before top-down fusion. Subsequently, the resulting four different-sized feature maps were output to the four detection heads. However, this method

posed challenges. Not only did it necessitate an additional detection head, but the alignment and fusion of feature maps based on the 160×160 scale significantly escalated model inference computation and imposed higher demands on the memory of the training device. Consequently, training costs surged under equivalent equipment conditions. Furthermore, this study explored augmenting the utilization of 160×160 large feature maps while discarding the use of 20×20 small feature maps. Regrettably, this approach yielded subpar results, with a notable decline in model performance. This outcome underscores the indispensability of information contained within the small feature maps extracted from deeper network layers for effective inference tasks. It aligns with the prevailing understanding that deeper networks generally yield superior results due to the crucial information embedded within deeper layers. Consequently, this study maintained the number of detection heads and the size of feature maps output to the detection heads while augmenting the utilization of shallow, large 160×160 feature maps. To circumvent 160×160 large-scale alignment fusion, these local features were stratified into dimensions of 80×80, 40×40, and 20×20, with global features standardized at 40×40. This strategic configuration ensured the efficiency in multi-scale feature fusion. The model's performance is optimized when these diverse features are efficiently combined, enabling it to effectively address various challenging detection tasks.

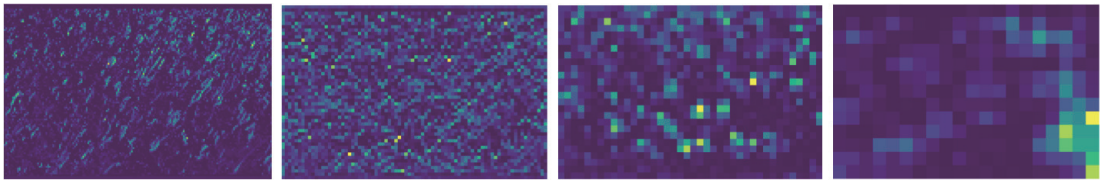
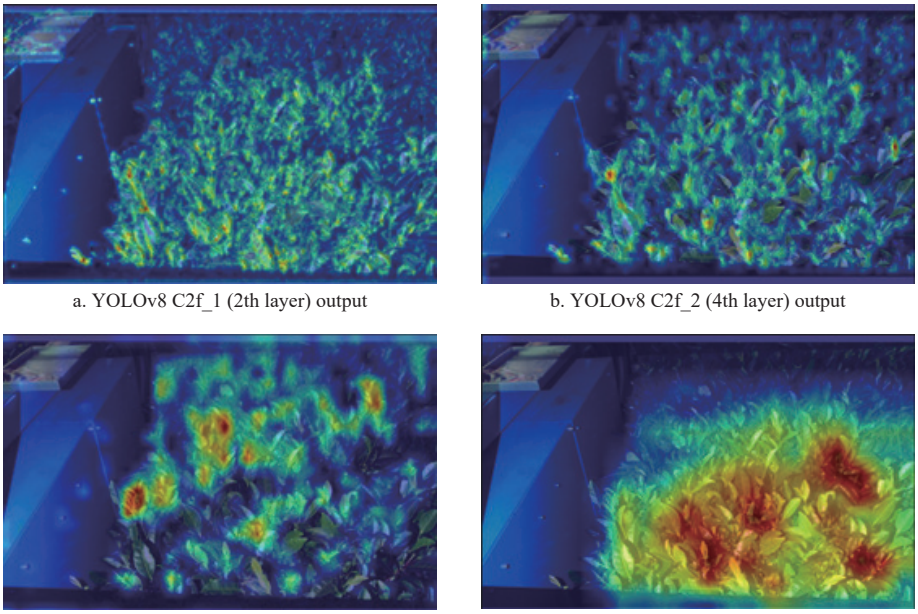


Figure 12 Visualization results of feature maps of different sizes; from left to right the sizes of feature maps are 160×160, 80×80, 40×40, and 20×20



a. YOLOv8 C2f_1 (2th layer) output b. YOLOv8 C2f_2 (4th layer) output

c. YOLOv8 C2f_3 (6th layer) output d. YOLOv8 SPPF (9th layer) output

Figure 13 Visualization of Grad-CAM heat maps for different backbone network layers

Although the proposed improvements show some superiority, the complexity and variability of the environment impose greater demands on the model. As can be seen from the images of tea samples, tea buds and tea leaves often shade each other, while the target and the background share similar colors. In addition, variations in light intensity throughout the day can significantly impact the appearance of tea buds, even within the same variety. These variations are particularly pronounced when the target is overexposed, causing the buds to appear whitened and resulting in a substantial loss of feature texture. To address this issue, future research will explore multimodal data fusion methods. By integrating depth information to enhance shape feature representation, the sensitivity of traditional RGB features to lighting conditions can be reduced, while fully leveraging the stable perception advantages of near-infrared spectroscopy in complex lighting environments such as low light and high reflectivity. Additionally, to address the issue of tea bud occlusion in complex picking scenarios, this study proposes introducing a Next-Best-View active perception strategy. By real-time assessment of the occlusion status of a tea bud from the current viewpoint, the camera's viewpoint is dynamically adjusted through lateral and pitch micro-adjustments to obtain more complete tea bud observation data.

5 Conclusions

This study performed a series of optimizations on the YOLOv8-seg model to enhance its segmentation capability for small targets in complex environments. The performance of the improved model was validated with the following experimental results. Firstly, the new neck network is more efficient in fusing multi-scale information and achieves optimal performance when using single feature maps as local features, resulting in improvements of 2.32% and 2.37% in AP50 for box and mask, respectively. Furthermore, when ECA and reparameterization techniques are applied to feature

fusion, the number of model parameters is reduced from 3.26 million to 3.22 million, and performance continues to improve. Finally, with WIoU used for gradient assignment in loss computation, the model's performance reaches its peak in this experiment. The final improved model demonstrates enhancements of 4.33% and 4.55% in AP50 for box and mask, respectively, compared to the pre-improvement model. In comparisons of performance and efficiency with other classical segmentation algorithms models, this model consistently retains its advantage. This demonstrates the superiority of the improved model in tea shoot segmentation. This study has significant theoretical and practical implications for the intelligent and precise plucking of high-quality tea shoots.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (Grant No. 52305289, Grant No. U23A20175), the Open Fund of State Key Laboratory of Tea Plant Biology and Utilization (Grant No. NKLTOF20240103), the earmarked fund for CARS-19 and the Open Project Program of Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University.

[References]

[1] 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.

[2] Yang J W, Li X, Wang X, Fu L Y, Li S W. Vision-Based Localization Method for Picking Points in Tea-Harvesting Robots. Sensors, 2024; 24(21): 6777.

[3] Zheng H, Fu T, Xue X L, Ye Y X, Yu G H. Research status and prospect of tea mechanized picking technology. Journal of Chinese Agricultural Mechanization, 2023; 44(9): 28–35. (in Chinese)

[4] Zhao L L, Deng H B, Zhou Y C, Miao T, Zhao K, Yang J, et al. Instance segmentation model of maize seedling images based on automatic generated labels. Transactions of the Chinese Society of Agricultural Engineering, 2023; 39(11): 201–211. (in Chinese)

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

[6] Fan P, Lang G D, Yan B, Lei X Y, Guo P J, Liu Z J, et al. A method of segmenting apples based on gray-centered rgb color space. Remote Sensing, 2021; 13(6): 1211.

[7] Akbar J U M, Kamarulzaman S F, Muzahid A J F, Rahman M A, Uddin M. A comprehensive review on deep learning assisted computer vision techniques for smart greenhouse agriculture. IEEE ACCESS, 2024; 12: 4485–4522.

[8] Kang H W, Chen C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. Computers and Electronics in Agriculture, 2020; 171: 105302.

[9] Liao J C, Babiker I, Xie W F, Li W, Cao L B. Dandelion segmentation with background transfer learning and RGB-attention module. Computers and Electronics in Agriculture, 2022; 202: 107355.

[10] 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.

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

[12] Gui Z Y, Chen J N, Li Y, Chen Z W, Wu C Y, Dong C W. A lightweight tea bud detection model based on Yolov5. Computers and Electronics in Agriculture, 2023; 205: 107636.

[13] Han K, Wang Y H, Tian Q, Guo J Y, Xu C J, Xu C. GhostNet: More features from cheap operations. arXiv: 1911.11907, 2019; In press. doi: 10.48550/arXiv.1911.11907.

[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] Wu H Y, Wang Y S, Zhao P F, Qian M B. Small-target weed-detection model based on YOLO-V4 with improved backbone and neck structures. Precision Agriculture, 2023; 24(6): 2149–2170.

[16] Xu H, Zhong S, Zhang T X, Zou X. Multiscale multilevel residual feature fusion for real-time infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing, 2023; 61: 1–16.

[17] Liu Q H, Zhang Y, Yang G P. Small unopened cotton boll counting by detection with MRF-YOLO in the wild. Computers and Electronics in Agriculture, 2023; 204: 107576.

[18] Lin T Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. arXiv: 1612.03144, 2016; In press. doi: 10.48550/arXiv.1612.03144.

[19] Liu S, Qi L, Qin H F, Shi J P, Jia J Y. Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, 2018; pp.8759–8768. doi: 10.48550/arXiv.1803.01534.

[20] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020; pp.10778–10787. doi: 10.1109/CVPR42600.2020.01079.

[21] Zhang WQ, Huang Z L, Luo G Z, Chen T, Wang X G, Liu W Y, et al. TopFormer: Token pyramid transformer for mobile semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, 2022; 12083–12093. doi: 10.1109/CVPR52688.2022.01177.

[22] Wang Y W, Ren Y Q, Kang S Y, Yin C B, Shi Y, Men H. Identification of tea quality at different picking periods: A hyperspectral system coupled with a multibranch kernel attention network. Food Chemistry, 2024; 433: 137307.

[23] Qian H M, Wang H L, Feng S, Yan S Y. FESSD: SSD target detection based on feature fusion and feature enhancement. J Real-Time Image Process, 2023; 20: 2.

[24] Song M X, Liu C, Chen L Q, Liu L C, Ning J M, Yu C Y. Recognition of tea buds based on an improved YOLOv7 model. Int J Agric & Biol Eng, 2024; 17(6): 238–244.

[25] Wang C-Y, Liao H-Y M, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H. CSPNet: A new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA: IEEE, 2020; pp.1571–1580. doi: 10.48550/arXiv.1911.11929.

[26] Zheng Z H, Wang P, Ren D W, Liu W, Ye R G, Hu Q H, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics, 2022; 52(8): 8574–8586.

[27] Din X H, Zhang X Y, Ma N N, Han J G, Ding G G, Sun J. Repvgg: Making vgg-style convnets great again. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, 2021; pp.13728–13737. doi: 10.48550/arXiv.2101.03697.

[28] Wang Q L, Wu B G, Zhu F P, 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, WA, USA: IEEE, 2020; pp.11531–11539. doi: 10.1109/CVPR42600.2020.01155.

[29] Tong Z J, Chen Y H, Xu Z W, Yu R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv: 2301.10051, 2023; In press. doi: 10.48550/arXiv.2301.10051.

[30] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. arXiv: 1610.02391, 2016; In press. doi: 10.48550/arXiv.1610.02391.