Phyton-
International Journal of
Experimental Botany

Tech Science Press

DOI: 10.32604/phyton.2024.046331

ARTICLE

Multi-Scale Mixed Attention Tea Shoot Instance Segmentation Model

Dongmei Chen[1], Peipei Cao[1], Lijie Yan[1], Huidong Chen[1], Jia Lin[1], Xin Li[2], Lin Yuan[3] and Kaihua Wu[1,*]

[1]School of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China

[2]Tea Research Institute, Chinese Academy of Agriculture Sciences, Hangzhou, 310018, China

[3]School of Information Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou, 310018, China

[*]Corresponding Author: Kaihua Wu. Email: wukaihua@hdu.edu.cn

Received: 27 September 2023 Accepted: 25 December 2023 Published: 27 February 2024

ABSTRACT

Tea leaf picking is a crucial stage in tea production that directly influences the quality and value of the tea. Traditional tea-picking machines may compromise the quality of the tea leaves. High-quality teas are often handpicked and need more delicate operations in intelligent picking machines. Compared with traditional image processing techniques, deep learning models have stronger feature extraction capabilities, and better generalization and are more suitable for practical tea shoot harvesting. However, current research mostly focuses on shoot detection and cannot directly accomplish end-to-end shoot segmentation tasks. We propose a tea shoot instance segmentation model based on multi-scale mixed attention (Mask2FusionNet) using a dataset from the tea garden in Hangzhou. We further analyzed the characteristics of the tea shoot dataset, where the proportion of small to medium-sized targets is 89.9%. Our algorithm is compared with several mainstream object segmentation algorithms, and the results demonstrate that our model achieves an accuracy of 82% in recognizing the tea shoots, showing a better performance compared to other models. Through ablation experiments, we found that ResNet50, PointRend strategy, and the Feature Pyramid Network (FPN) architecture can improve performance by 1.6%, 1.4%, and 2.4%, respectively. These experiments demonstrated that our proposed multi-scale and point selection strategy optimizes the feature extraction capability for overlapping small targets. The results indicate that the proposed Mask2FusionNet model can perform the shoot segmentation in unstructured environments, realizing the individual distinction of tea shoots, and complete extraction of the shoot edge contours with a segmentation accuracy of 82.0%. The research results can provide algorithmic support for the segmentation and intelligent harvesting of premium tea shoots at different scales.

KEYWORDS

Tea shoots; attention mechanism; multi-scale feature extraction; instance segmentation; deep learning

1 Introduction

Tea leaf harvesting is a critical phase in tea production. The quality of harvested tea directly affects its overall quality and value. Currently, tea leaf harvesting can be broadly categorized into bulk tea harvesting and premium tea harvesting, whose quality and value are much higher than bulk tea. According to statistics, in 2020, the production and value of premium tea in Zhejiang Province were 102,000 tons and 21.34 billion yuan, accounting for 53.18% and 89.43% of the total in the province, respectively [1]. Traditional tea harvesting machines often use mechanical power to cut tea leaves rapidly, lacking selectivity and prone to damage the woody parts of the tea bushes [2]. Such machines are typically suitable for bulk tea harvesting. However, premium tea requires a selective plucking approach, with only the tender shoots [3], one shoot-one leaf, or one shoot-two leaves being harvested, and manual harvesting has been the primary method. With the increasing demand for tea and rising labor costs, it is necessary to replace manual labor with automated harvesting technologies, leading to the emergence of intelligent tea harvesting machines. In these devices, the accurate recognition and detection of tea shoots through computer vision is essential for precise harvesting [4–6]. Thus, research of algorithms for recognition of the premium tea shoots holds both theoretical value and practical significance.

In the agricultural application scenarios, the background types of the detected objects are divided into two categories: clean background and complex background. A clean background refers to a solid color plane with significant color deviation from the object to be detected, and a complex background refers to indistinguishable colors similar to the detected object [7]. Early studies focused on the identification and detection of tea shoots primarily using traditional machine learning and image processing techniques to analyze RGB images [8], but they have certain drawbacks. The double threshold method is sensitive to image noise, and can only be applied to single-object segmentation [9]. Regarding HIS color space and region growing, limitations arise from the involvement of only three spectral bands, resulting in fewer hierarchical levels in the fused color scheme and impacting the interpretation of land cover types [10]. Otsu threshold method fails to accurately separate targets and backgrounds when there is substantial grayscale overlap and k-means mean clustering, with its completely random selection, may lead to slow algorithm convergence [11]. Furthermore, these traditional methods often required manual parameter selection and were mostly applicable to specific scene conditions, such as clean backgrounds, limiting their applicability in complex agricultural scenarios [12]. In recent years, advancements in parallel computing hardware have made deep learning algorithms, especially convolutional neural networks (CNN), increasingly popular in the field of harvesting robotics [13,14]. Deep learning methods have superior feature extraction capabilities and strong generalization abilities, more suitable for practical harvesting tasks, including target recognition and detection in unstructured environments [15,16]. To seek more effective and accurate solutions in the field of tea shoot image processing and intelligent tea leaf harvesting, numerous scholars utilized deep learning and computer vision technologies. Yang et al. [17] used an improved YOLOv3 model to obtain feature maps of tea leaves at different scales, enabling the model to better identify tea leaf types and positions. Luo et al. [18] established an automatic detection model for tea shoots based on the Faster R-CNN [19] network framework with a VGG16 network [20], improving detection accuracy and algorithm robustness. Chen et al. [3] combined Faster R-CNN with the fully convolutional network (FCN) to identify harvesting points in tea shoot regions using a complex two-stage model. Qian et al. [21] introduced an improved deep convolutional encoder-decoder network with skip connections and a contrastive center loss function for semantic segmentation of tea shoots. Lv et al. [22] addressed the sensitivity of existing tea shoot detection algorithms to changes in lighting conditions by proposing a tea shoot detection model based on region brightness adaptive correction, enhancing shoot feature saliency and detection accuracy.

However, most of the aforementioned deep learning-based methods primarily focus on crop detection. Further steps are required for crop segmentation and precise harvesting point localization before providing information to the harvesting equipment. In recent years, to achieve precise segmentation of tea shoots, several algorithm modules have emerged in the field of instance segmentation, such as the Mask R-CNN [23] network, which builds upon Faster R-CNN by adding an FCN (Fully Convolutional Network) branch. The method enables simultaneous object detection and instance segmentation, allowing for pixel-level recognition of multiple object outlines. Researchers have already applied instance segmentation methods to agricultural scenarios, including strawberries [16], cucumbers [24], and apples [25]. Wu et al.

proposed an approach based on the CNN model Deeplab V3+ and classical image processing algorithms to segment banana bunches [26]. Additionally, there are also algorithms like YOLACT: Real-time Instance Segmentation [27], which divides instance segmentation into two parallel tasks: generating a dictionary of non-local prototype masks across the entire image and predicting a set of linear combination coefficients for each instance. Currently, Liu et al. [28] have achieved real-time instance segmentation of tomato plants based on the YOLACT. Wang et al. introduced SOLOv2: Dynamic and Fast Instance Segmentation [29], which transforms the segmentation problem into a position classification problem. This eliminates the need for anchors, normalization, and bounding box detections in instance segmentation. To address the target similarity problem in complex backgrounds, Chen et al. proposed a rapid visual gender detection method for pigeon features based on the YOLOv5 model [30]. The above-mentioned research relies primarily on CNN models. However, the convolutional operations in CNN can only capture local information and struggle to establish long-distance connections for a global image. In contrast, the Transformer models allow for direct global relationship modeling, thereby expanding the image's receptive field, obtaining more contextual information, and supporting parallel computation [31]. Recently Cheng et al. [32] proposed the Masked-attention Mask Transformer for Universal Image Segmentation (Mask2Former) model for image segmentation. This is a new architecture capable of addressing any image segmentation task. When compared with state-of-the-art models on the COCO [33] dataset, Mask2Former produces final results with better boundary quality and further achieves end-to-end object detection.

However, the tea shoots are relatively small in the regular real-field pictures and are very similar to the complex background. Thus, it is very challenging to obtain an accurate edge image of the tea shoots as well as the plucking information, which is very important in intelligent picking machines. So based on the self-made small tea shoots dataset. We construct a multi-scale instance segmentation network Mask2FusionNet with attention mechanisms [34] at its core to comprehensively extract features from tea shoot images based on a self-constructed tea shoot dataset. Firstly, we analyzed the characteristics of the tea shoot harvesting dataset. Confronting the challenge of distinguishing small targets in the complex backgrounds within this dataset, we proposed a detection method employing a multi-scale edge point selection strategy. This method effectively utilizes contextual information at different scales to differentiate the relationship between captured targets and the green leaf background. By separating overlapping targets and the edge contours of tea trees, we enhanced the clarity of contour information extraction and the accuracy of small target identification. Subsequently, we conducted comparative experiments, pitting our proposed algorithm against mainstream algorithms, and revealing improvements in both accuracy and convergence speed. The ablation experiments were then performed to validate the effectiveness of ResNet50, PointRend, and FPN, further affirming the superiority of our algorithm in detecting small-scale targets in complex backgrounds. Finally, we conducted a visual analysis of different modules, and experimental results indicated that our model exhibits superior detection capabilities in tea shoot recognition.

2 Materials and Methods

2.1 Image Acquisition and Processing

All the tea leaf images were collected from multiple tea gardens in the West Lake Scenic Area of Hangzhou, Zhejiang Province, China, in March 2021. The tea variety was "Longjing 43," developed by the Tea Research Institute of the Chinese Academy of Agricultural Sciences. The images were captured under natural light conditions with the rear dual camera of mobile phones. These images were captured with the rear dual cameras of a smartphone under natural lighting conditions on both cloudy and sunny days. The shooting distance ranged from 60 to 80 centimeters, employing both oblique and vertical shooting angles in a 1:1 ratio. The RGB images of tea shoots had a size of 3024 * 4032 pixels and were stored in jpg format.

The collected image data had a relatively large field of view, containing a high number of dispersed tender tea shoots. To annotate the images, images with distinct features of tea shoots and stems and visible harvesting points were selected and divided into 2 * 4 segments. Thus, a total of 464 images were chosen, with the size of 1512 * 1008 pixels. These images were then divided into a training set, a validation set, and a test set in a 7:2:1 ratio. The training set was used for learning the weight parameters during the model training process, the validation set was used to optimize the network model's structure, and the test images were used to verify the accuracy of the proposed method. To improve the accuracy of tea shoot labeling, we consulted with the tea experts and local farmers, ultimately establishing the tea shoot image annotation scheme. We employed Labelme software for instance segmentation labeling of tea shoots, creating a tea shoot dataset in COCO format with a total of 1703 samples. Fig. 1a depicts the original harvested tea shoot image, while Fig. 1b illustrates the annotated tea shoot image. To prevent overfitting during the model training process, an online data augmentation method was applied to expand the dataset. Image augmentation helped reduce the imbalance in sample distribution and improved the model's generalization ability. The image augmentation methods used in the study mainly included scale transformation, flip transformation, and pixel value normalization. The computer configuration for this experiment includes an Intel(R) Xeon(R) Platinum 8124 M CPU @ 3 GHz, 64 GB of RAM, an NVIDIA GeForce RTX 2080Ti GPU with 11 GB of VRAM, and the Linux Ubuntu 18.04 operating system.

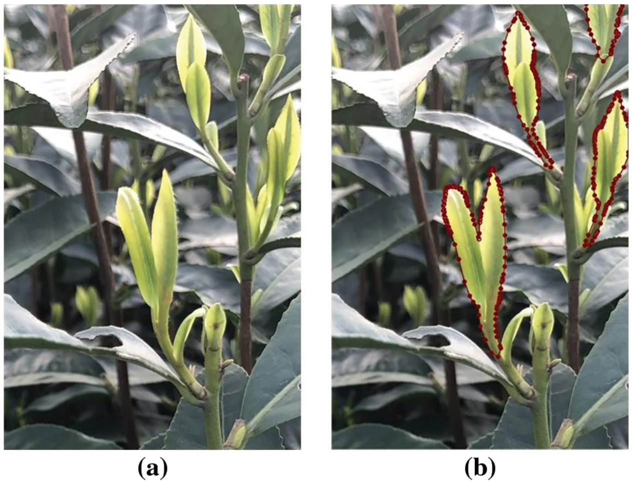


Figure 1: Segmentation sample of tea shoot dataset (a) original image (b) annotation image

2.2 Mask2FusionNet Tea Shoot Instance Segmentation Model

The tea shoot segmentation model based on Mask2FusionNet is illustrated in Fig. 2. Its structure primarily consists of several modules, including the feature extraction network, pixel decoder, self-attention [31], mask attention, and point selection strategy [35]. The tea shoots images are input in the ResNet50 structure and the features specific to tea shoots are extracted with the feature extraction network. Subsequently, the feature maps are passed through the pixel decoder module, which gradually restores the low-resolution feature maps to the original image's resolution using operations such as deconvolution or upsampling. In each upsampling layer, convolutional layers and normalization layers are added to generate high-resolution pixel embeddings. Among the four feature maps, the first three are input into the transformer decoder. This decoder processes object queries to generate region predictions and mask predictions. Within the predicted regions, a point selection strategy is applied to extract feature points. These feature points are then mapped to the final mask predictions using a point-head network.

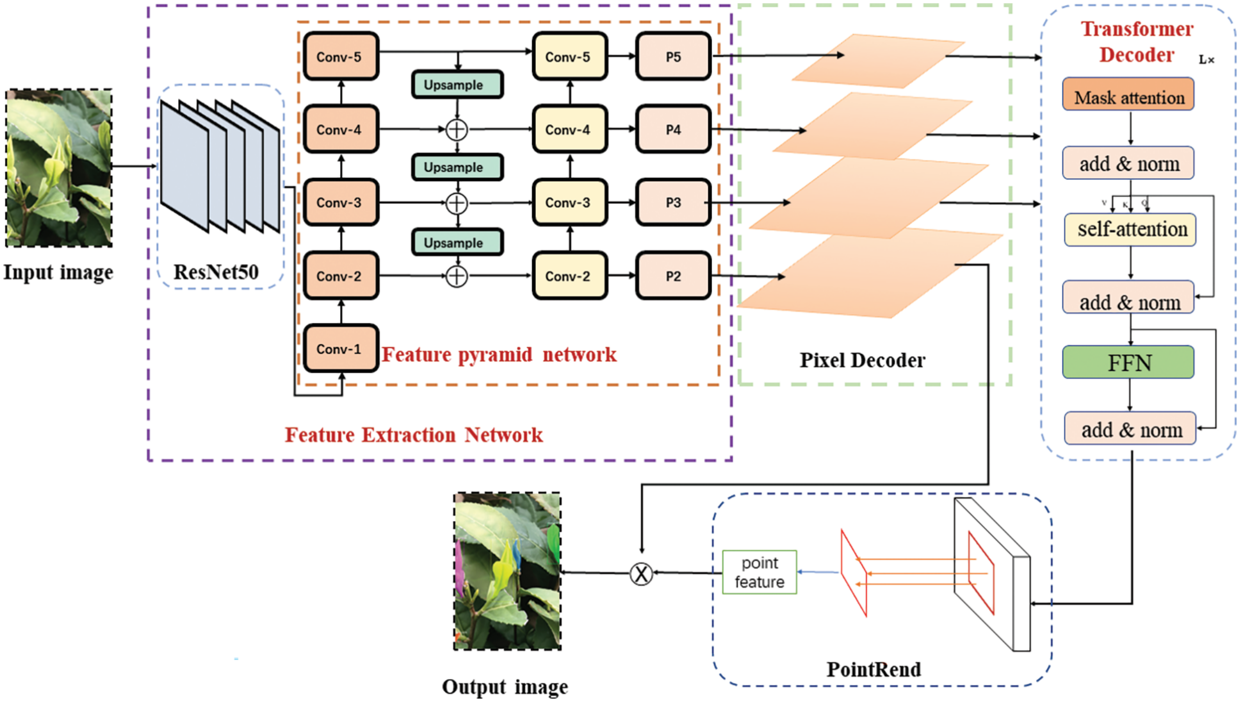


Figure 2: Framework of tea shoot segmentation model based on Mask2FusionNet

Tea shoots often vary in shape and size during the detection process due to different shooting angles and growth stages. Therefore, a network architecture that combines input tea shoot images with the Residual Networks (ResNet) [36] and the Feature Pyramid Network (FPN) [37] is employed to capture more details from feature maps at different levels. The main backbone network in this setup is ResNet50, which performs convolutional operations on the input images, reducing the image size by half and doubling the number of channels with each convolutional operation. Feature maps are extracted at different stages, denoted as C1 to C5, from shallow to deep. The middle part of the network is the feature pyramid structure. Initially, a 1 × 1 convolution is applied to the deep-level feature map C5, followed by downsampling. In the feature fusion phase, a 3 × 3 convolution is used to eliminate aliasing effects introduced during upsampling, resulting in feature map P4. Similarly, a 1 × 1 convolution is applied to C3, and after a 1 × 1 convolution with C4, the corresponding feature maps are element-wise summed to produce P3. This process continues, yielding P2 and P1. The improved feature extraction network enhances its focus capabilities and can be directly integrated into the model, allowing it to participate in the iterative update of the entire model's network parameters. In the feature extraction network, FPN combines the detailed positional information of lower-level features with the rich semantic information of higher-level features, enhancing the representation capacity of features. This provides richer and more useful information for detecting small targets and segmentation. Consequently, in the context of tea shoot recognition and segmentation, the FPN network can more accurately and efficiently detect tea shoots of different scales, further improving the accuracy of instance segmentation [38,39]. The formulas for the improved P2, P3, P4, and P5 feature maps are as follows:

$$P_5 = Conv_{1\times1}(C_5) \otimes M_5(Conv_{1\times1}(C_5)) \tag{1}$$

$$P_i = Conv_{1\times1}(C_i) \otimes M_i(Conv_{1\times1}(C_i)) \oplus f^{2\times2}_{upsampling}(P_{i+1}) \tag{2}$$

In the equation, C_i means the feature map output from the i-th stage of feature extraction, M_i means the weight matrix of the attention module corresponding to the feature map of the i-th stage, P_i means the feature map after feature fusion, ⊕ is Element-wise addition, ⊗ is the element-wise multiplication and means performing 2 × 2 upsampling.

The Pixel Decoder transforms the feature maps extracted by the model into pixel-level segmentation masks. This is achieved through a series of convolutional and upsampling operations to generate segmentation masks with the same resolution as the input image. The convolutional layers are used to process the feature maps, extracting rich semantic information. Upsampling operations, either through deconvolution or interpolation, restore the spatial dimensions of the feature maps to match those of the input image. Finally, a classifier is applied using a 1 × 1 convolutional layer to convert the feature maps into pixel-level segmentation masks. The objective of the Pixel Decoder is to produce high-quality segmentation results, accurately delineating regions of different classes. It effectively translates abstract feature representations into semantically meaningful segmentation masks, enabling precise image instance segmentation.

By taking the feature maps generated by the Pixel Decoder as input, we employ masked attention to process the foreground regions of the predicted masks. Following a series of convolutional operations, we apply self-attention to the input feature maps through the following steps: Firstly, we calculate queries (Q), keys (K), and values (V) through linear transformations. This yields corresponding Q, K, and V vectors. Subsequently, by computing the similarity between queries and keys, attention weights are obtained. These weights determine the importance of each position relative to others. Next, based on the attention weights and the representation of values, we calculate weighted sums to obtain context feature representations for each position. Finally, we utilize a Feed Forward Network (FFN) structure, comprising a combination of nonlinear activation functions and multiple linear transformations, to produce the corresponding image features. The Transformer Decoder module directs the model's attention more toward image regions relevant to the target, filtering out irrelevant features for the task. This helps in selecting effective features more precisely, thus enhancing the model's performance.

Compared to predicting all pixels in the high-resolution output grid, selecting only a few points can significantly reduce computational and memory consumption. For the recognition of the tea shoots, which often resemble the background, point selection strategies can sample uncertain regions, focusing on areas with uncertain predictions for further analysis. This approach can enhance the model's performance in predicting blurry or boundary regions, thereby improving the overall segmentation results. After extracting fine-grained features from the fused image and conducting rough predictions, point features are generated, resulting in the final point prediction image. The point selection strategy can also maintain a certain degree of uniform coverage, aiming to cover the entire prediction area as much as possible. This ensures that the model learns sufficient features from the entire region and avoids excessive bias towards specific areas.

2.3 Loss Function Calculation and Model Evaluation Methods

The model's loss function consists of both a classification loss and a mask loss. For the classification loss, binary cross-entropy loss is utilized, while for the mask loss, binary cross-entropy loss for a single class is employed. The formulas for each loss function are as follows:

$$L_{cls}(p_i^*, p_i) = -p^* \log p \tag{3}$$

$$L_{mask}(s_i^*, s_i) = -(s^* \log s + (1 - s^*) \log(1 - s)) \tag{4}$$

In the equation, p means the probability of anchor box prediction being the target, p^* is the actual labels corresponding to the anchor boxes, s is binary predicted masks, s^* is the corresponding ground truth masks.

The text uses mean Average Precision (mAP) as the performance metric for evaluating shoot recognition and segmentation. This metric is calculated based on the values of True Positives (TP) and False Positives (FP) and is computed using the following formula:

$$mAP = mean\left(\frac{TP}{TP + FP}\right) \tag{5}$$

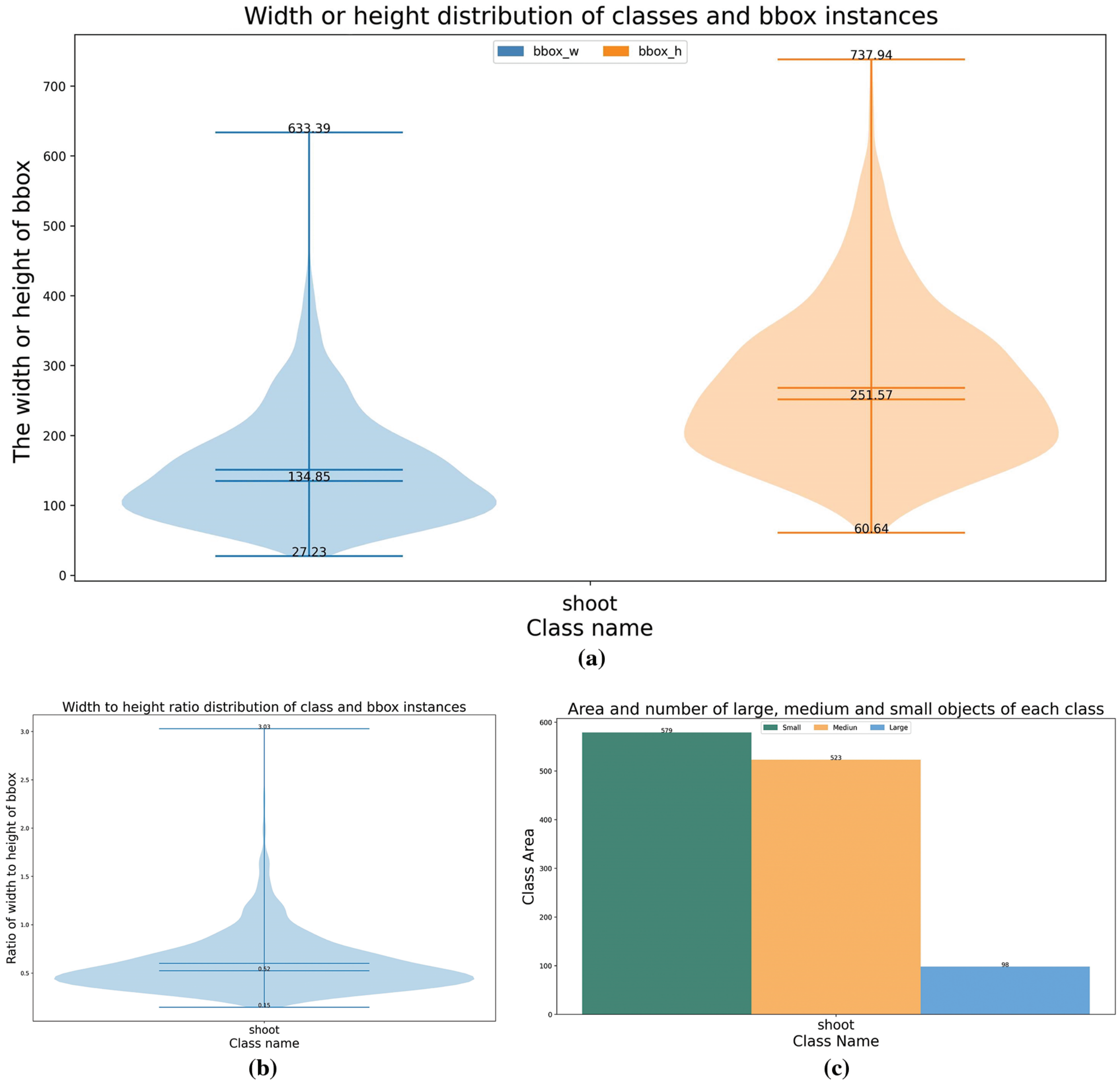
In the evaluation of shoot recognition results using bounding boxes, TP represents the number of target boxes that were accurately detected and recognized as target boxes, while FP represents the number of non-target boxes that were incorrectly detected and recognized as target boxes. In the evaluation of shoot segmentation results using masks, TP represents the number of target pixels that were accurately detected and segmented as target pixels, while FP represents the number of non-target pixels that were incorrectly detected and segmented as target pixels.

3 Experimental Process and Results Analysis

3.1 Tea Shoot Image Dataset

To gain a more intuitive understanding of the characteristics of the tea shoot dataset, we performed data analysis by calculating and plotting the width and height distribution of all object-bounding instances within the dataset, as well as the area distribution of object-bounding instances based on area rules. Fig. 3a displays the width and height distribution of target box instances, showing that most of the target boxes in the dataset have widths ranging from around 60 to 120 and heights ranging from around 150 to 220. This indicates that the target sizes in the dataset are generally small. Fig. 3b presents the distribution of aspect ratios of target box instances, indicating that most of the aspect ratios in the dataset fall between 0.2 and 0.8. The smallest aspect ratio observed was 0.15, while the largest was 3.03. This suggests that the dataset has a relatively uniform distribution of target box shapes. Fig. 3c shows the area distribution of target box instances based on area rules. In this experiment, the area rules classified targets with resolutions smaller than 32 * 32 pixels as small targets, resolutions between 32 * 32 pixels and 96 * 96 pixels as medium targets, and resolutions greater than 96 * 96 pixels as large targets. The results indicate that small and medium targets constitute 89.9% of the dataset.

The results indicate that the tea shoot dataset primarily consists of small-scale objects. In the COCO dataset, the proportion of small and medium-sized targets is approximately 75% [33]. In contrast, our dataset has a higher proportion of small targets. Additionally, our dataset presents some challenging scenarios, such as high similarity between small targets and the background, as well as target overlap. These factors significantly increase the difficulty of detection and segmentation. As shown in Fig. 4, the first row consists of original images, and the second row contains annotated images. Figs. 4a–4c show examples of the small tea targets, target overlap, and high background similarity in the images. To further study whether the proposed Mask2FusionNet model is suitable for this dataset, we conducted comparative experiments in the following subsection.



(a)

(b)

(c)

Figure 3: Data set analysis results. (a) Instance width and height distribution chart, (b) instance width/height distribution chart, (c) instance area distribution chart

3.2 Comparison of Results with Mainstream Segmentation Algorithms

In this study, the proposed tea shoot instance segmentation model was compared with other mainstream object detection algorithms to evaluate its performance in recognizing tea shoots, particularly in terms of false negatives and overlapping instances. The comparative models selected for this evaluation were Mask R-CNN [23], YOLACT [27], SOLOv2 [29], and Mask2Former [34]. The accuracy curve graph of the final models is presented in Fig. 5. Fig. 4 displays the results of Average Precision (AP) on the validation dataset for the four object detection algorithms. From the graph, it can be observed that our model achieves the highest accuracy in recognizing tea shoots. Table 1 provides a summary of the results for the four object detection algorithms in terms of Average Precision (AP) on the validation dataset, along with the evaluation metrics for our algorithm's tea shoot segmentation results. The results show that our model outperforms the Mask R-CNN model with a 6.7% improvement in AP_{50}, compared to a 4.1% improvement over the YOLACT model, and a remarkable 23.1% improvement over the SOLOv2 model.

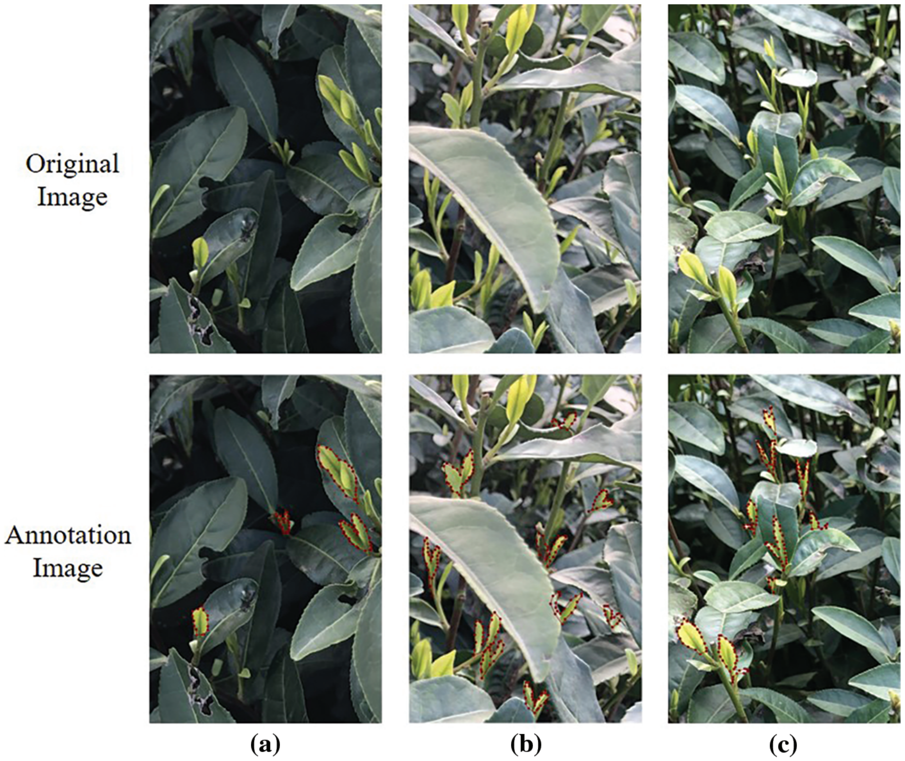


Figure 4: Segmentation sample of tea shoot dataset (a) small targets (b) overlap targets (c) background similarity targets

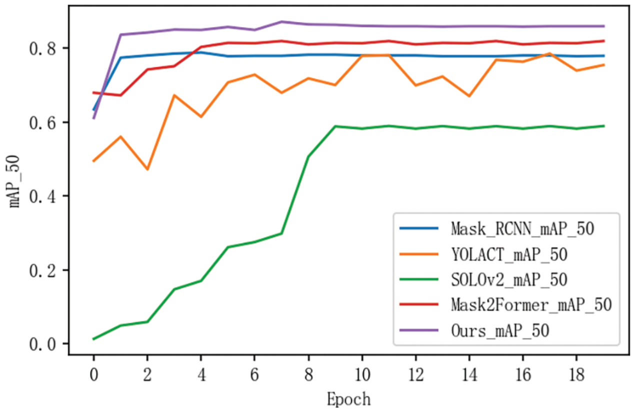


Figure 5: The accuracy of mainstream instance segmentation algorithms on the validation set

Table 1: Tea shoot segmentation evaluation indicators of different modules

Models	Segmentation			FPS
	AP	AP_{50}	AP_{75}	
Mask R-CNN	47.9	75.3	54.1	4.15
YOLACT	51.7	77.9	56.3	6.75
SOLOV2	26.2	58.9	22.5	7.36
Mask2Former	56.8	78.8	62.0	4.23
Ours	**63.6**	**82.0**	**69.1**	**4.26**

The above experiments indicate that the proposed Mask2FusionNet model's small-object detection scheme is suitable for this dataset and demonstrates a superior recognition accuracy compared to other models. To further validate which modules and parameters are crucial for the accurate detection of overlapping tea shoots, we will conduct a series of ablation experiments in the following steps.

3.3 Ablation Experiments

We introduced a multi-scale hybrid attention model based on ResNet50, achieving precise segmentation of tea shoots by incorporating FPN and PointRend networks. To validate the effectiveness of different modules in small-object detection, we conducted a series of ablation experiments.

As shown in Table 2, experimental results demonstrate that our method significantly outperforms the original Mask2Former network in terms of precision, achieving an AP_{25} of 63.6%, AP_{50} of 82.0%, and AP_{75} of 69.1%. In Table 2, EXP1 to EXP4 used ResNeXt as the backbone, and the average precision at AP_{50} for the four combinations is 79.2. EXP5 through EXP8 used ResNet50 as the backbone, and the average AP_{50} for the four combinations is 80.9, which had a better performance than ResNeXt. In EXP2, the PointRend was introduced based on EXP1, resulting in a 1.8% improvement in AP_{50} accuracy. Similarly, EXP6 with PointRend, compared with EXP5, led to a 2.3% increase in AP_{50} accuracy. In EXP3, the FPN was introduced based on EXP1, resulting in a 3.2% improvement in AP_{50} accuracy. Similarly, EXP7, built upon experiment five, introduced the FPN, resulting in a 2.7% increase in AP_{50} accuracy. EXP4, which simultaneously introduced the PointRend and FPN architectures based on EXP1, led to a 4.4% increase in AP_{50} accuracy. Similarly, EXP8, based on EXP5, simultaneously introduced the PointRend and FPN, resulting in a 3.2% increase in AP_{50} accuracy. These results indicate that whether using ResNeXt or ResNet50 as the backbone, the PointRend and FPN are effective for tea leaf detection.

Table 2: Tea shoot segmentation evaluation metrics based on the Mask2Former model

Experiment	Backbone ResNeXt	Backbone ResNet50	PointRend	FPN	Segmentation		
					AP_{25}	AP_{50}	AP_{75}
EXP1	+				56.6	76.8	60.8
EXP2	+		+		58.3	78.6	61.7
EXP3	+			+	59.9	80.3	63.5
EXP4	+		+	+	60.5	81.2	65.8

(Continued)

Table 2 (continued)

Experiment	Backbone ResNeXt	Backbone ResNet50	PointRend	FPN	Segmentation		
					AP_{25}	AP_{50}	AP_{75}
EXP5		+			56.8	78.8	62.0
EXP6		+	+		60.9	81.1	65.7
EXP7		+		+	61.9	81.5	68.1
EXP8		+	+	+	**63.6**	**82.0**	**69.1**

Compared with Backbone ResNeXt (EXP1-EXP4), the average AP_{50} of ResNet50 (EXP5-EXP8) improved by 1.6%. Compared with the original model, the inclusion of the PointRend structure (EXP2, 4, 6, 8) resulted in an average AP_{50} increase of 1.4%. The inclusion of the FPN structure (EXP3, 4, 7, 8) resulted in an average AP_{50} increase of 2.4%. In conclusion, compared to the Mask2Former network, our network architecture approach improves precision by 3.2%, and the experimental results demonstrate the superiority of our model in the instance segmentation task of tea shoots.

3.4 Visualization of Mask2FusionNet Results

To further demonstrate the advantages of our model in tea shoot recognition, we conducted model visualization to analyze the tea shoot recognition problem. As shown in Fig. 6, experimental results indicate that our model has better detection capabilities in image recognition compared to other models.

Fig. 6 displays tea shoot instance segmentation results based on Mask2Former with the introduction of different modules. The first column represents the original images, the second column shows the results of the Mask2Former original network without adding any modules, the third column presents the results of combining Mask2Former with the PointRend network, the fourth column shows the results of combining Mask2Former with the FPN network structure, and the last column displays the results of our network. In the first row of images, the baseline network made erroneous candidate box selections during tea shoot recognition, while the other networks had more accurate selections. In the second row, the baseline network made errors in recognizing pickable tea shoots, mistakenly identifying unpickable shoots. Other networks avoided this issue, and both the baseline network and the PointRend structure network had cases of missed detection while introducing the FPN structure achieved more precise recognition. In the third row of images, the baseline network's recognition of tea shoot contours was incomplete, and it incorrectly grouped two shoots together as one. The PointRend network and FPN network had false positives and missed detections in recognizing shoots, with lower accuracy compared to the fused PointRend and FPN networks. In the fourth row of images, the baseline network failed to completely recognize the tea shoots, while other networks could recognize them, with more accurate contour recognition. These experimental results demonstrate that our network structure, compared to the Mask2Former network structure, extracted more information from the target regions, achieved more accurate and complete tea shoot edge contours, and improved the model's accuracy by excluding older leaves that are easily confused with the shoots.

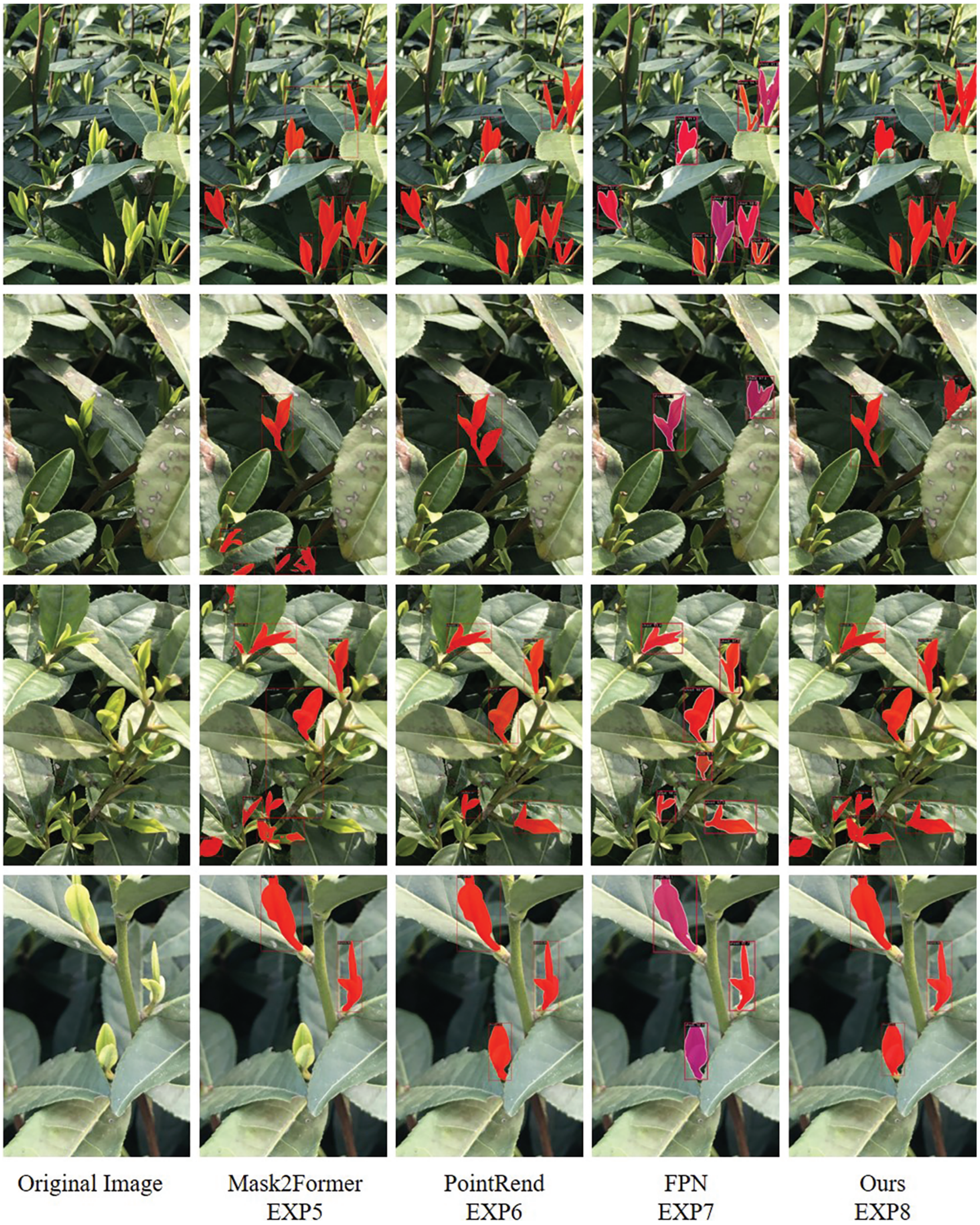


Figure 6: Example segmentation results of tea shoots with different modules

4 Conclusion

In addressing the intelligent harvesting of tea shoots, there is a high demand for accuracy in identification and precise edge segmentation. Consequently, we constructed a tea shoot dataset, analyzed its characteristics, and proposed a multi-scale feature edge detection algorithm. Based on the

Mask2Former architecture, we employed a multi-scale feature fusion method to enhance the detection capability of small targets. A point selection strategy was employed to refine the clarity of edge contours, ultimately achieving a higher accuracy and finer edge output in an end-to-end network. Through comparative experiments, our model achieved a 3.2% improvement in accuracy compared to the original Mask2Former. Specifically, the ResNet50, PointRend strategy and FPN architecture showed performance improvements of 1.6%, 1.4%, and 2.4%, respectively. The Mask2FusionNet model obtains more accurate and complete tea shoot edge contours, enhancing the model's accuracy. Visual interpretation of the model's recognition results ultimately demonstrates that our proposed Mask2FusionNet model effectively enhances the model's feature extraction capabilities for tea shoot images, improves detection performance, and reduces false positive rates. For tea shoots, segmenting multi-scale and overlapping samples is crucial.

This module could have further application in conjunction with a depth camera for tea bud plucking. Currently, Glenn Jocher et al. have proposed YOLOv5: v7.0, applying the Intel RealSense D455 depth camera for depth image collection and achieving SOTA results [40], which will be a focal point in our subsequent research. Currently, some studies have been applied in the actual picking of tea shoots. Li et al. achieved a reliable algorithm based on red, green, and blue-depth (RGB-D) camera images developed to detect and locate tea shoots in fields for tea harvesting robots [41]. In future work, we can deploy the Mask2FusionNet algorithm further on the depth camera, exploring the potential application of depth camera systems in tea shoot recognition. This will provide essential foundational information for the automatic identification and precise harvesting of tea shoots.

Acknowledgement: The authors sincerely thank the anonymous reviewers who made valuable comments on this paper.

Funding Statement: This research was supported by the National Natural Science Foundation of China No. 62276086, the National Key R&D Program of China No. 2022YFD2000100, Zhejiang Provincial Natural Science Foundation of China under Grant No. LTGN23D010002.

Author Contributions: D. Chen: Conceptualization, Methodology, Investigation, Data curation, Supervision, Writing–original draft, Writing–review & editing. P. Cao: Software, Methodology, Data curation, Formal analysis, Investigation, Writing–original draft. L. Yan: Methodology, Validation, Writing–review & editing. H. Chen: Methodology, Validation. J. Lin: Validation, Writing–review & editing. X. Li: Methodology, Validation. L. Yuan: Validation. K. Wu: Conceptualization, Methodology, Supervision, Investigation, Funding acquisition, Writing–review & editing.

Availability of Data and Materials: The datasets used and/or analyzed during the current study are available from the author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declared that they have no conflicts of interest to report regarding the present study.

References

1. Luo LW, Feng HQ, Hu S. "Thirteenth Five-Year Plan" review and "Fourteenth Five-Year Plan" prospect of Zhejiang Tea Industry. Chin Tea Process. 2021;1:6–12. doi:10.15905/j.cnki.33-1157/ts.2021.01.002.
2. Muruganand S, Sureshkumar AA. Design and development of selective tea leaf plucking robot. ACIS. 2014;2(2):45–8.
3. Chen Y, Chen S. Localizing plucking points of tea leaves using deep convolutional neural networks. Comput Electron Agric. 2020;171:105298.

4. Ya G. Research on the application of automation software control system in tea garden mechanical picking. In: International Conference on Applications and Techniques in Cyber Security and Intelligence, 2019; Berlin, Springer; p. 1830–6.
5. Zhu YP, Wu CY, Tong JH, Chen JN, He LY, Wang RY, et al. Deviation tolerance performance evaluation and experiment of picking end effector for famous tea. Agric. 2021;11(2):128
6. Han Y, Xiao HR, Song ZY, Ding WQ, Mei S. Design and experiments of 4CJ-1200 self-propelled tea plucking machine. Int J Agr Biol Eng. 2021;14(6):75–84.
7. Tang YC, Qiu JJ, Zhang YQ, Wu DX, Cao YH, Zhao KX, et al. Optimization strategies of fruit detection to overcome the challenge of unstructured background in field orchard environment: a review. Precis Agric. 2023;24:1183–219.
8. Tang YC, Zhou H, Wang HJ, Zhang YQ. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. Expert Syst Appl. 2022;2011:118573.
9. Yang FZ, Yang LL, Tian YN, Yang Q. Recognition of the tea sprout based on color and shape features. Trans Chin Soc Agric Mach. 2009;40(S1):119–23 (In Chinese).
10. Fan JP, Yau DK, Ekmagarmid AK, Aref WG. Automatic image segmentation by integrating color-edge extraction and seeded region growing. IEEE Trans Image Process. 2001;10(10):1454–66.
11. Shao PD, Wu MH, Wang XW, Zhou J, Liu S. Research on the tea bud recognition based on improved k-means algorithm. In: 2018 2nd International Conference on Electronic Information Technology and Computer Engineering (EITCE 2018), 2018.
12. Li CG, Tang YC, Zou XJ, Zhang P, Lin JQ, Lian G, et al. A novel agricultural machinery intelligent design system based on integrating image processing and knowledge reasoning. Appl Sci. 2022;12(15):7900.
13. Zhou HX, Wang X, Au W, Kang H, Chen C. Intelligent robots for fruit harvesting: recent developments and future challenges. Precis Agric. 2021;23:1856–907.
14. Zhou YZ, Tang YC, Zou XG, Wu ML, Tang W, Meng F, et al. Adaptive active positioning of camellia oleifera fruit picking points: classical image processing and YOLOv7 fusion algorithm. Appl Sci. 2022;12:12595.
15. Karunasena G, Priyankara H. Tea bud leaf identification by using machine learning and image processing techniques. Int J Sci Eng Res. 2020;11(8):624–8.
16. Yu Y, Zhang KL, Yang L, Zhang DX. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput Electron Agr. 2019;163:104846.
17. Yang HL, Chen L, Chen MT, Ma Z, Deng F, Li MZ, et al. Tender tea shoots recognition and positioning for picking robot using improved YOLO-V3 model. IEEE Access. 2019;7:180998–1011.
18. Luo HL, Feng ZL, Ran ZN, Ma J, Lv J. Research on automatic detection of tender tea buds based on VGG16 network. Agric Technol. 2020;40(1):15–7. doi:10.19754/j.nyyjs.20200115005 (In Chinese).
19. Ren SQ, He KM, Girshick RB, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal. 2015;39:1137–49.
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Comput Vis Pattern Recognit. 2014. doi:10.48550/arXiv.1409.1556.
21. Qian C, Li MY, Ren Y. Tea sprouts segmentation via improved deep convolutional encoder-decoder network. IEICE Trans Inf Syst. 2020;103(2):476–9.
22. Lv J, Fang MR, Yao Q, Wu CY, He YL, Bian L, et al. Detection model for tea buds based on region brightness adaptive correction. Trans Chin Soc Agric Eng. 2021;37(22):278–85 (In Chinese).
23. He KM, Gkioxari G, Dollár P, Girshick RB. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2017; Venice, Italy. p. 2961–9.
24. Liu XY, Zhao DA, Jia WK, Ji W, Ruan CZ, Sun YP. Cucumber fruits detection in greenhouses based on instance segmentation. IEEE Access. 2019;7:139635–42.
25. Jia WK, Tian YY, Luo R, Zhang ZH, Lian J, Zheng YJ. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. Comput Electron Agric. 2020;172:105380.

26. Wu FY, Yang Z, Mo XK, Wu ZH, Tang W, Duan JL, et al. Detection and counting of banana bunches by integrating deep learning and classic image-processing algorithms. Comput Electron Agric. 2023;209:107827.
27. Bolya D, Zhou C, Xiao FY, Lee YJ. Yolact: real-time instance segmentation. In: IEEE/CVF International Conference on Computer Vision, 2019; Seoul, Korea (South). p. 9156–65.
28. Liu C, Feng QC, Sun YH, Li YJ, Ru MF, Xu LJ. YOLACTFusion: an instance segmentation method for RGB-NIR multimodal image fusion based on an attention mechanism. Comput Electron Agric. 2023;213:108186.
29. Wang X, Zhang RF, Kong T, Li L, Shen CH. SOLOv2: dynamic and fast instance segmentation. Adv Neural Inf Process Syst. 2020;33:17721–32.
30. Chen SY, Tang YC, Zou XJ, Huo HL, Hu KW, Hu B, et al. Identification and detection of biological information on tiny biological targets based on subtle differences. Mach. 2022;10(11):996.
31. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Comput Lang. 2017; p. 6000–10.
32. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; Red Hook, NY, USA. p. 1280–89.
33. Lin TY, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. Comput Vis Pattern Recognit. 2014;8693:740–55.
34. Mnih V, Heess NM, Graves A, Kavukcuoglu K. Recurrent models of visual attention. Mach Learn. 2014;2:2204–12. ArXiv:1406.6247.
35. Kirillov A, Wu YX, He KM, Girshick RB. Pointrend: image segmentation as rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; Seattle, WA, USA. p. 9799–808.
36. He KM, Zhang X, Ren SQ, Sun J. Deep residual learning for image recognition. Comput Vis Pattern Recognit. 2016; p. 770–8.
37. Lin TY, Dollár P, Girshick RB, He KM, Hariharan B, Belongie SJ. Feature pyramid networks for object detection. Comput Vis Pattern Recognit. 2017; p. 936–44.
38. Wang XL, Girshick R, Gupta A, He KM. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; Salt Lake City, UT, USA. p. 7794–803.
39. Chen YP, Kalantidis Y, Li JS, Yan SC, Feng JS. A^2-Nets: double attention networks. Neural Inf Process Syst. 2018;31:350–59.
40. Glenn J, Ayush C, Alex S, Jirka B, NanoCode012, Yonghye K. ultralytics/yolov5: v7.0-YOLOv5 SOTA realtime instance segmentation. Zenodo. 2022. doi:10.5281/zenodo.7347926.
41. Li YT, He LY, Jia JM, Lv J, Chen JM, Qiao X, et al. In-field tea shoot detection and 3D localization using an RGB-D camera. Comput Electron Agric. 2021;185:106149.