Journal of Agricultural Engineering

Journal of Agricultural Engineering 2026; LVII:1637
doi:10.4081/jae.2025.1637

Automated tea shoot picking using the YOLO network and Mamba images segmentation for top-view detection with a monocular camera

Zhiqiang Wang,[1] Hui Niu,[1] Jing Zhang,[1] Wu Zhang,[1,2,3] Jian Mao,[1] Shengqi Zhang,[1] Jun Liu,[1] Guanpeng Zuo,[1] Zhe Zheng,[1] Zhenxiang Chi[1]

[1]School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei; [2]Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs, Hefei; [3]Anhui Province Key Laboratory of Smart Agricultural Technology and Equipment, Hefei, China

Abstract

Detection and localization of tea shoots (one bud with two leaves) are critical steps in the automation of tea harvesting. Using red, green, blue-depth (RGB-D) camera to detect and locate tea shoots from side angles results in significant occlusion of tea shoots, as well as loss of depth information. To achieve automated, intelligent, and precise tea harvesting, this paper proposes a method for detecting and locating tea shoots from the top using a monocular camera. Firstly, the "You Only Look Once" (YOLO) network is employed to detect tea shoots regions in images collected by the monocular camera and to crop individual tea shoot top images. For these cropped images, a U-shaped images segmentation model based on Mamba is proposed. This model achieves a mean intersection over union (MIoU) of 87.80% and an accuracy (ACC) of 95.63%, precisely locating the specific tea shoots top regions. The center of the circumscribed circle of this region is used as the position for the next step in the picking process, accurately guiding the picking effector to the top of the tea shoot. Finally, the picking effector, controlled by feedback signals from infrared sensors, performs up-and-down reciprocation and cutting actions to complete the picking process. This method effectively avoids the problem of depth information loss during localization with RGB-D camera. To verify the effectiveness of the proposed approach, picking experiments were conducted on HouKui tea within a simulated tea garden environment, achieving a tea shoot picking success rate of 75.54%. The results indicate that this method offers significant application value and provides a new perspective for the development of automated tea shoots picking.

Key words: monocular camera; Mamba; picking effector; tea shoots picking.

Correspondence: Wu Zhang, PhD, School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China. E-mail: zhangwu@ahau.edu.cn

Introduction

In traditional HouKui tea picking methods, the short picking cycle results in low production efficiency and high costs (Han *et al.*, 2014). Automated tea picking has been proposed to effectively alleviate this problem (Tang *et al.*, 2016).

Automated tea picking necessitates obtaining accurate picking locations, and the first step is to detect tea shoots. Currently, most scholars identify tea shoots based on their color and shape characteristics. Wu *et al.* (2015) used the K-means clustering algorithm for image segmentation of tea shoots, effectively distinguishing tea shoots in images to achieve tea shoots identification. Zhang *et al.* (2019) used Bayesian discrimination to construct a model for identifying the state of tea shoots, and its high recognition rate provides basic conditions for automated tea garden construction. Karunasena and Priyankara (2020) proposed a new method for detecting tea shoots using machine vision and machine learning. They combined histogram of oriented gradients (HOG) features based on cascade classifiers with support vector machine (SVM) classification for tea shoots detection. However, most tea images in these studies were obtained under simple background conditions or indoor conditions. In actual tea garden environments, there are complex backgrounds and variable lighting conditions. Additionally, the characteristics of tea shoots vary significantly across different picking periods (Wu *et al.*, 2018; Zhou and Ca, 2022). Hence, methods that identify tea shoots based solely on color are not suitable for real tea garden environments. With the advancement of deep learning technology, the benefits of deep learning algorithms in target detection are becoming more evident (Liu *et al.*, 2020). Many researchers have started applying deep learning algorithms to agricultural robots for task target recognition and detection (Kamilaris and Prenafeta-Boldú, 2018). Zhu *et al.* (2022) used fast region-based convolutional network (Faster R-CNN) to address the simultaneous problem of target localization and classification, achieving better detection speed and accuracy. Zou *et al.* (2022) proposed an improved YOLO network-based method for detecting tea shoots to enhance detection accuracy, with improvements in both precision and recall rates.

After detecting tea shoots, accurately obtaining the picking position is crucial for realizing automated tea picking. Long *et al.* (2022) segmented tea images based on tea shoots characteristics, then used a combination of edge detection and skeletonization algorithms to locate tea shoots picking points. Han *et al.* (2019) used an improved YOLO model to detect tea shoots, extracted the skeleton from the detection box, and then determined the lowest point of the skeleton as the picking point. This method accurately locates the two-dimensional coordinates of the picking point. However, achieving automated picking requires obtaining complete three-dimensional coordinates, so acquiring 3D data is crucial for automated picking. Wang *et al.* (2022) used a combination of monocular vision ranging technology and the OpenCV library to help a fruit picking robot achieve 3D localization of target fruits. With the advent of RGB-D cameras that can acquire R, G, B, and depth information of target regions, the obtained depth information can be used for 3D localization of target points. Consequently, RGB-D cameras have been integrated into agricultural robots in recent years to complete the localization of picking points in automated harvesting (Fu *et al.*, 2020). Mai *et al.* (2015) fused the color images and depth images obtained by RGB-D cameras to generate 3D point clouds of apple trees, then used point cloud segmentation algorithms to extract the 3D point clouds of the fruits and obtain their corresponding spatial location information and radii. After completing 2D localization of oranges, Yang *et al.* (2019) used the Kinect V2 depth camera to acquire depth images of the target region, mapping the depth information to the picking point to obtain the 3D information of the picking point. Li *et al.* (2021) proposed a 3D localization method for tea picking points by obtaining 3D point clouds of tea shoots and based on the growth morphology of tea shoots, using RGB-D cameras for on-site tea shoots detection and 3D localization in tea gardens. However, in the actual working environment of automated picking equipment, RGB-D cameras may experience depth information loss due to unstable lighting conditions (Li *et al.*, 2022), affecting picking efficiency. Thus, this paper proposes a method for detecting and locating tea shoots from the top using a monocular camera. The main research content includes: i) using HouKui tea as an example, collect a certain amount of top-view images of HouKui tea shoots, process and annotate the image data, build a tea dataset, and utilize the YOLOV7 network to detect the top-view images of HouKui tea captured by the monocular camera; ii) develop a U-shaped images segmentation model based on Mamba to accurately segment the top regions of tea shoots in the images; iii) based on the segmentation results, locate the picking position and propose a method of picking by controlling the picking effector with infrared sensors, thus avoiding the issue of depth information loss that occurs when using RGB-D cameras for localization; iv) construct a specific experimental platform for picking experiments to validate the method's effectiveness.

Materials and Methods

Method overview

In real tea garden environments, unstructured and unstable conditions can lead to depth information loss during data collection with RGB-D cameras. Moreover, upward growth is a biological trait we can observe. Therefore, in addressing the issue of depth information loss and based on the characteristics of tea shoots, this paper independently develops and designs a picking actuator equipped with an infrared sensor. A method using a monocular camera to detect and locate tea shoots at the top of the tea plants is employed for the picking task. The specific structural design and working principle are illustrated in Figure 1. First, the top-view images of tea shoots captured by the monocular camera are detected, and individual tea shoot top images are cropped based on the detection results. Then, the individual tea shoot top images are fed into the image segmentation model to obtain precise tea shoot top mask images, which are used to locate the central position of the tea shoot top. Subsequently, the picking effector is guided by a sliding rail to position itself directly above the specified tea shoot. Finally, the infrared sensor controls the picking effector to complete the picking task. The detailed discussion of the work presented in this paper follows.

Image dataset construction

The subject of this study is HouKui tea, which is typically har-

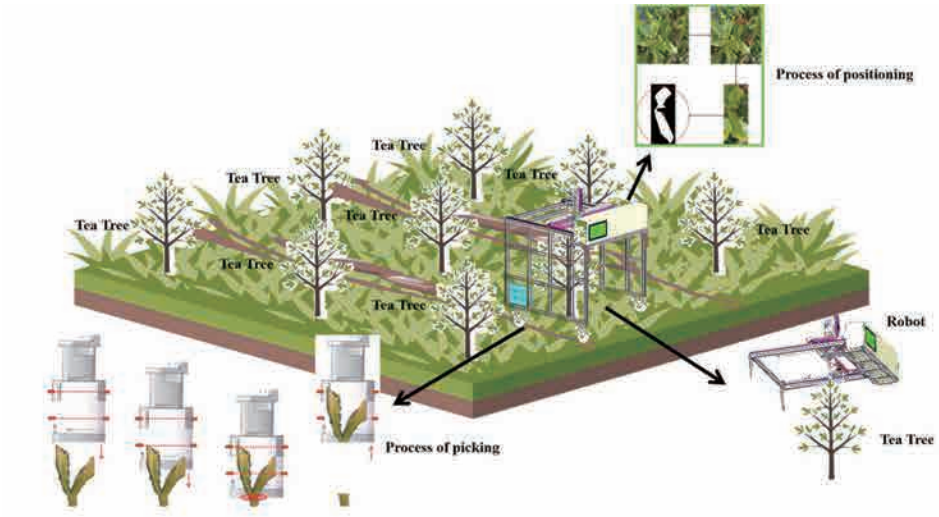

Figure 1. System overview diagram.

vested around April each year. During this period, HouKui tea exhibits distinct color and morphological characteristics. Therefore, we collected a certain number of top-view images of HouKui tea shoots at the Anhui Agricultural University Central Anhui Experimental Station, National Key Laboratory of Tea Tree Biology and Resource Utilization Tea Tree Germplasm Resource Nursery, as shown in Figure 2.

After obtaining the image data, this study targets the color and top morphological characteristics of HouKui tea shoots for detection. The original images are annotated based on these characteristics, by enclosing the top region of the tea shoots in the images and labeling it as «tea». For the image segmentation model's data annotation, the top region of a single tea shoot in the original image is cropped to create a new image, with the top region outlined and annotated in the new image. To achieve better detection results, we performed data augmentation on the original image data to increase the diversity of the sample images. 70% of the initial images were randomly selected and processed by methods such as flipping, brightness enhancement, brightness reduction, and adding noise to obtain new images. This method allows for data augmentation of the dataset, thereby reducing the risk of overfitting during training and improving the detection accuracy of the model. After data augmentation, the dataset is divided into training, validation, and test sets in an 8:1:1 ratio for training and testing the object detection and image segmentation models, as shown in Table 1.

Tea shoots detection and image segmentation

Tea shoots detection

YOLO, as a popular real-time object detection model, can quickly and accurately detect the location of tea shoots in the process of automated tea picking. The YOLOv7 (Wang *et al.*, 2023) network is an upgraded version of the YOLO series, employing a more complex network structure and introducing skip connections, which helps capture semantic information in images more effectively. Consequently, it enhances the accuracy and performance of object detection, enabling real-time, rapid, and precise detection of tea shoot locations during automated tea picking process.

Tea shoots image segmentation

After detecting the tea shoot, precise localization of its apex region requires segmentation of the single tea shoot top image obtained during detection. In the domain of image segmentation, models based on CNN (Santos *et al.*, 2020) and Transformer (Wang *et al.*, 2023) have been extensively employed. However, during application, both have limitations in constructing long-term dependencies in high-resolution tea shoots images, affecting the accuracy of image segmentation. Recently, Mamba (Gu and Dao, 2024) has been proposed as a new selective structural state-space model, excelling in long-sequence modeling tasks. Therefore, this paper introduces a U-shaped images segmentation model based on Mamba to accurately segment the top region of tea shoots in images.

Specifically, the model includes Patch Embedding, Encode, Decode, Linear Projection, and Skip Connection, using a symmetric structure, as shown in Figure 3A. Firstly, Patch Embedding is used to divide the input image into small patches and map the image dimensions to C, resulting in the embedded image. Then, in the four stages of the Encode, two VSS blocks are used for feature extraction and fusion, and patch merging is performed at the end of the first three stages, reducing the height and width of the feature map by half while doubling the number of channels. Similarly, Decode is a structure symmetric to Encode. Before the last three stages, Patch Expanding is used for up sampling, gradually restor-

Figure 2. Tea garden and tea shoots top image.

Table 1. Tea shoots images.

Model	Initial images	Extended images	Training set	Validation set	Test set
Object detection	3216	6753	7975	997	997
Image segmentation	2991	6281	7417	927	927

ing the original image resolution, and finally, a projection layer restores the number of feature map channels. Additionally, simple addition operations are used for skip connections at each layer, effectively retaining the feature information extracted at each stage of the encoder, which helps enhance the decoder's feature recovery capability. The VSS blocks in this model are mainly derived from Visio Mamba (Zhu *et al.*, 2024), as shown in Figure 3B. The input feature map first undergoes normalization through Layer Normalization, after which the features are processed separately. For the first processing method, features are initially processed through a linear layer, depth wise separable convolution and an activation function, then input into the SS2D module for further feature extraction. After normalization with Layer Normalization, the features are merged with another set of features processed through a linear layer and activation function, and finally, a linear layer is used for feature fusion. Unlike the traditional VIT, this VVS block avoids position embedding, allowing more feature information to be processed with the same computational resources, making the model run faster. Combined with its excellent performance in long-sequence modeling tasks, this method can segment the top region of tea shoots in images more accurately and quickly, providing strong theoretical support for its configuration in automated tea picking equipment.

Localization and picking method

After detecting tea shoots and segmenting the tea shoot top image, we finally obtained a more accurate tea shoots position. To ensure the picking effector reaches the optimal picking position, we use the circumcenter of the segmented region as the final picking position, as shown in Figure 4A. After locating this position, it is necessary to calculate the coordinate position suitable for the picking effector to complete the picking task through coordinate transformation. First, convert the location point from the pixel coordinate system to the camera coordinate system, and then from the camera coordinate system to the picking effector coordinate system, as shown in Figure 5.

The specific calculation for converting from the pixel coordinate system to the camera coordinate system is as follows:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \quad \text{(Eq. 1)}$$

Simplifying, we obtain:

$$\begin{cases} X_c = (u - u_0) * Z_c / f_x \\ Y_c = (v - v_0) * Z_c / f_y \\ Z_c = Z_c \end{cases} \quad \text{(Eq. 2)}$$

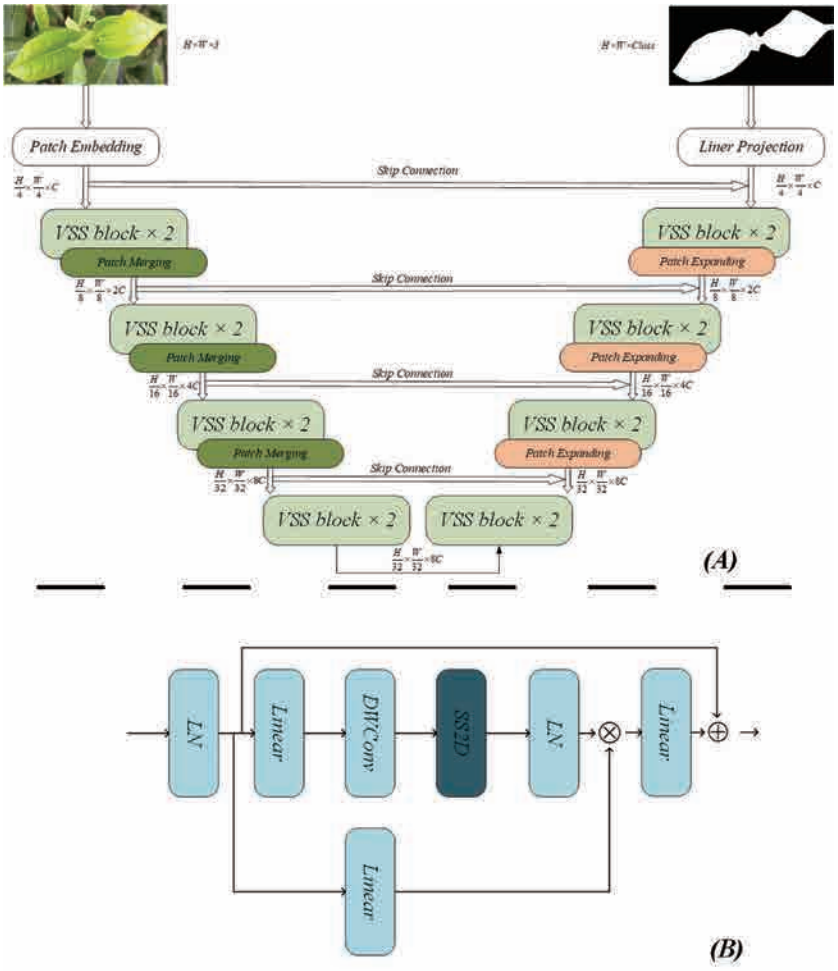

Figure 3. U-shaped images segmentation model (**A**). The VSS blocks (**B**).

Among them, u_0, v_0, f_x and f_y are derived from the camera's intrinsic parameters; u and v are the pixel coordinates of the previously determined picking position; Z_C is set as the height from the camera to the top of the tea shoots.

Through the above calculations, the positioning point is successfully converted to the 2D coordinates (X_c, Y_c) in the camera coordinate system, then the coordinates are translated according to the relative position of the camera and the picking effector, finally obtaining the 2D coordinates (X,Y) in the picking effector coordinate system. Based on this coordinate information, the picking effector can reach the position directly above the target tea shoot. starts moving downwards quickly without needing sensor feedback control. This stage can use a simple open-loop control by setting the maximum control signal value as the initial control signal.

At this point, we use a sensor feedback-based control method, utilizing the infrared sensor feedback signals to control the picking effector, aiming to achieve efficient and precise picking of the target tea shoot. This control process can be divided into three main stages, each with specific control strategies and response mechanisms. The details are as follows: i) as shown in Figure 4B-1, in the initial stage, when neither end infrared sensor can detect the tea shoot, the picking effector starts moving downwards quickly without needing sensor feedback control. This stage can use a simple open-loop control by setting the maximum control signal value as the initial control signal. ii) When the bottom infrared sensor detects the tea shoot, it indicates that the tea shoot is beginning to enter the picking effector, as shown in Figure 4B-2. The picking effector slows down but continues to move downwards, with the control system entering a closed-loop control state to slow down the picking effector. At this point, proportional control is used to adjust the effector speed. The control formula is as follows:

$$u(t) = K_p \cdot \left(1 - y_{bottom}(t) \cdot y_{top}(t)\right) \quad \text{(Eq. 3)}$$

where K_p is the proportional gain, $y_{bottom}^{(t)}$ is the feedback signal from the bottom infrared sensor, and $y_{top}^{(t)}$ is the feedback signal from the top infrared sensor. When the shoot is detected, $y_{bottom}^{(t)} = 1$ and $y_{top}^{(t)} = 0$. iii) When both sensors detect the tea shoot, it indicates that the shoot has entered the picking effector to a length that meets the picking requirements. At this time, the infrared sensor feedback signals $y_{bottom}^{(t)} = 1$, $y_{top}^{(t)} = 1$, so the calculated $u(t) = 0$ based on the control formula, the picking effector stops moving and triggers the cutting action, as shown in Figure 4B-3. The picking task is successfully completed and the picking effector returns to its original position, as shown in Figure 5B-4.

Combining the picking position obtained by the positioning method with the above picking process, it is possible to accurately complete the automated tea picking task in an unstructured and unstable agricultural environment without using depth information. This provides a novel approach and application method for automated tea picking. Algorithm 1 shows the pseudocode for the picking effector control method.

Among them, operating state of the picking effector is shown in Figure 6.

Algorithm 1
Input: t,y_bottom( ),y_top( )
Params:u_max,u_return,K_p
Operator:u( ),Cut( ),ReturnHome( )
Output: Picking completion status

```
u(t)=u_max
while (True) do
if y_bottom(t)==1 then
u(t) = K_p * (1 - y_bottom(t))
if y_bottom(t)==1 and y_top(t) then
u(t)=0
Cut( )
end if
end if
t=t+1
end while
u(t) = u_return
ReturnHome( )
```

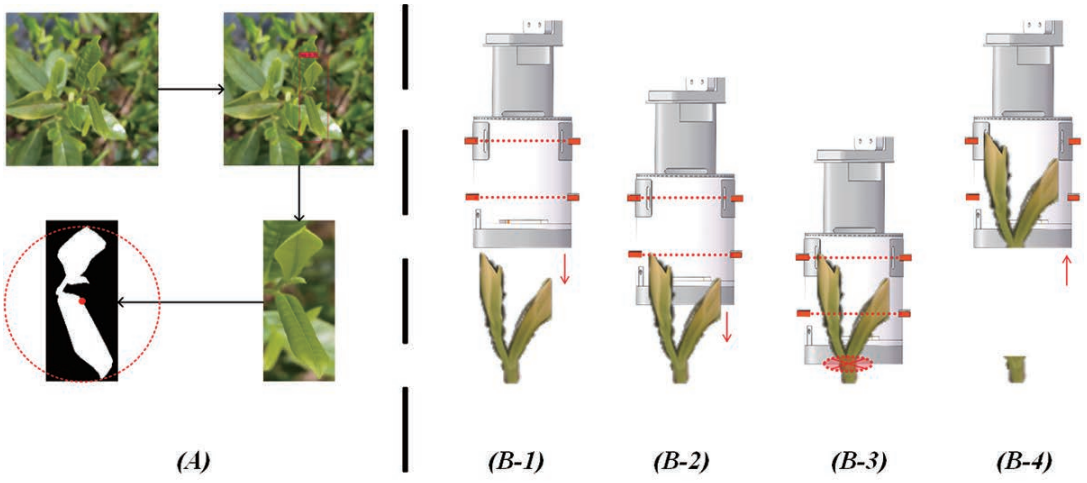

Figure 4. Localization method (**A**). Picking method (**B**): the picking effector starts moving downwards quickly (**B-1**), the bottom infrared sensor detects the tea shoot and the picking effector slows down but continues to move downwards (**B-2**), both sensors detect the tea shoot and the picking effector triggers the cutting action (**B-3**); the picking effector returns to its original position (**B**).

Experiments and analysis

Evaluation of the image segmentation model

The training platform used in this study is a computer running the Linux operating system, equipped with an Intel i5-11400F (2.6GHz) CPU, an NVIDIA RTX 3080TI GPU, and 32GB of memory running. The image input size is 256×256, the training cycle is 300 epochs, the batch size is 32, the initial learning rate is 0.001, and the weight decay coefficient is 0.0001. To test the performance of the segmentation model, this study uses a wide range of evaluation metrics, including MIoU, DSC, ACC, Specificity, and Sensitivity. Among them, MIoU is used to measure the degree of overlap between the predicted results and the ground truth labels, and its specific calculation formula is as follows:

$$MIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad \text{(Eq. 4)}$$

In the formula, and represent the predicted region and the ground truth region of the i-th class, respectively, and N denotes the number of classes. DSC is used to evaluate the degree of overlap between the predicted segmentation region and the ground truth segmentation region, and its specific calculation formula is as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \quad \text{(Eq. 5)}$$

In the formula, A denotes the predicted region, and B denotes the ground truth region.

ACC is a commonly used metric for evaluating the overall performance of a model, representing the proportion of correctly classified pixels. Spe is used to evaluate the model's ability to correctly identify negative classes, while Sen measures the model's ability to identify positive classes. The specific calculation formulas are as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Eq. 6)}$$

$$Spe = \frac{TN}{TN + FP} \quad \text{(Eq. 7)}$$

$$Sen = \frac{TP}{TP + FN} \quad \text{(Eq. 8)}$$

In the formula, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. The specific performance of the model is shown in Table 2. This table illustrates a direct performance comparison between this model and other segmentation models. For metrics with an upward arrow (↑), a higher value indicates better performance. The best-performing metrics and models are shown in bold. At the same time, the model's advantages in terms of both accuracy and efficiency are compared, with the results shown in Figure 7. The results indicate that the segmentation model proposed in this study can more accurately predict the top region of tea shoots. To further verify the effectiveness of the model, we

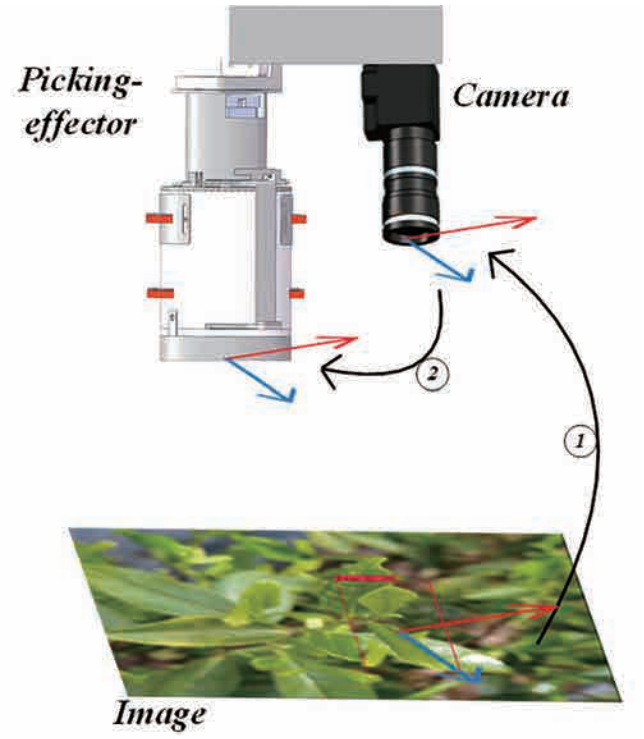

Figure 5. Coordinate system transformation. 1, the pixel coordinate system to the camera coordinate system; 2, the camera coordinate system to the picking effector coordinate system

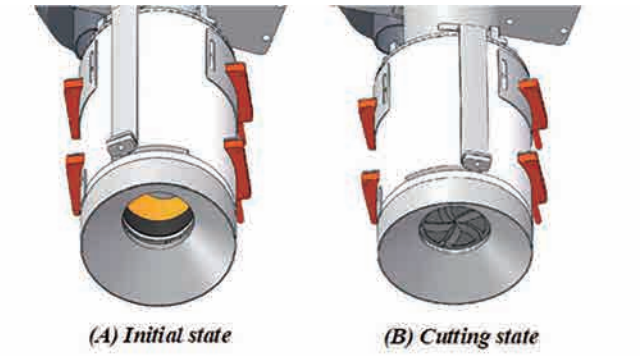

Figure 6. The operating state of the picking effector.

Table 2. Comparison of different images segmentation models.

Model	MIoU(%)↑	DSC(%)↑	ACC(%)↑	Spe(%)↑	Sen(%)↑	Parameters (M)	GFLOPs	FPS↑
Unet	84.02%	90.92%	93.99%	93.98%	94.84%	2.011	3.231	36.5
Swin-Unet	83.11%	90.51%	93.93%	95.42%	91.03%	46.912	14.178	8.3
Ours	87.80%	93.50%	95.63%	95.74%	95.42%	27.427	4.118	28.6

used different segmentation models to predict the top region of the tea shoots, as shown in Figure 8. Clearly, in the visualization results, the segmentation model proposed in this study has a significant advantage in segmenting the top region of tea shoots, fully meeting the requirements for precise positioning in the process of automated tea picking.

Picking experiment

Overview of the experimental system

The experimental equipment mainly consists of a controller, a Jetson Nano development board, a monocular camera, and a picking effector, as shown in Figure 9. The Siemens PLC controller is used to control the picking effector's movement and picking actions on the slide rail device. Additionally, the Jetson Nano development board is used for image acquisition, image processing, coordinate calculation, and communication with the PLC controller via TCP. The monocular camera uses an AH7500MG010 industrial region scan camera with a resolution of 2448×2448 and a frame rate of 24FPS. It communicates with the development board via GigE. The picking effector is independently designed and developed, equipped with two sets of infrared array sensors that communicate with the PLC controller via RS234 to control the up-and-down reciprocating motion and picking actions of the effector.

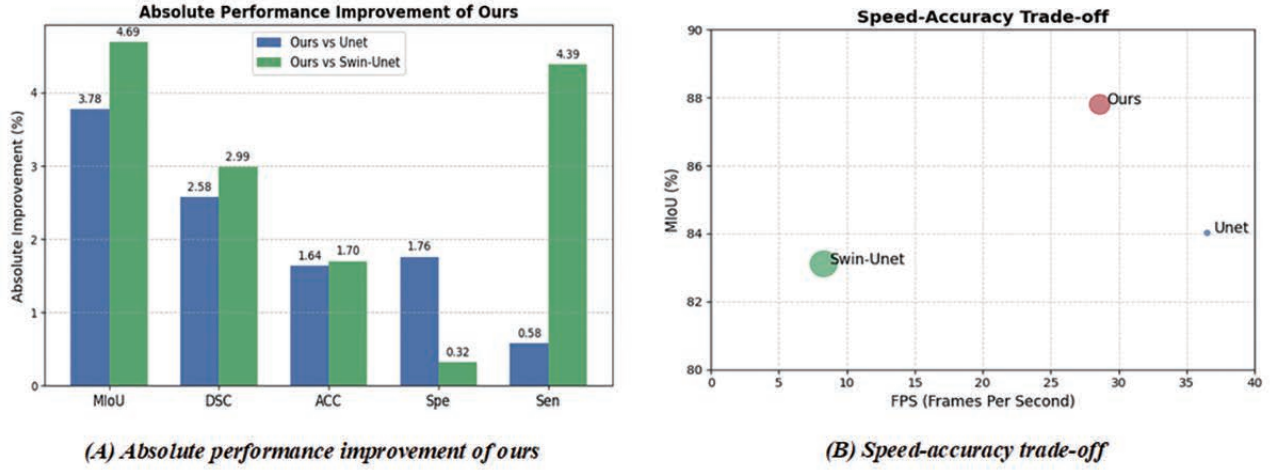

Figure 7. Comparison of the model's accuracy and efficiency.

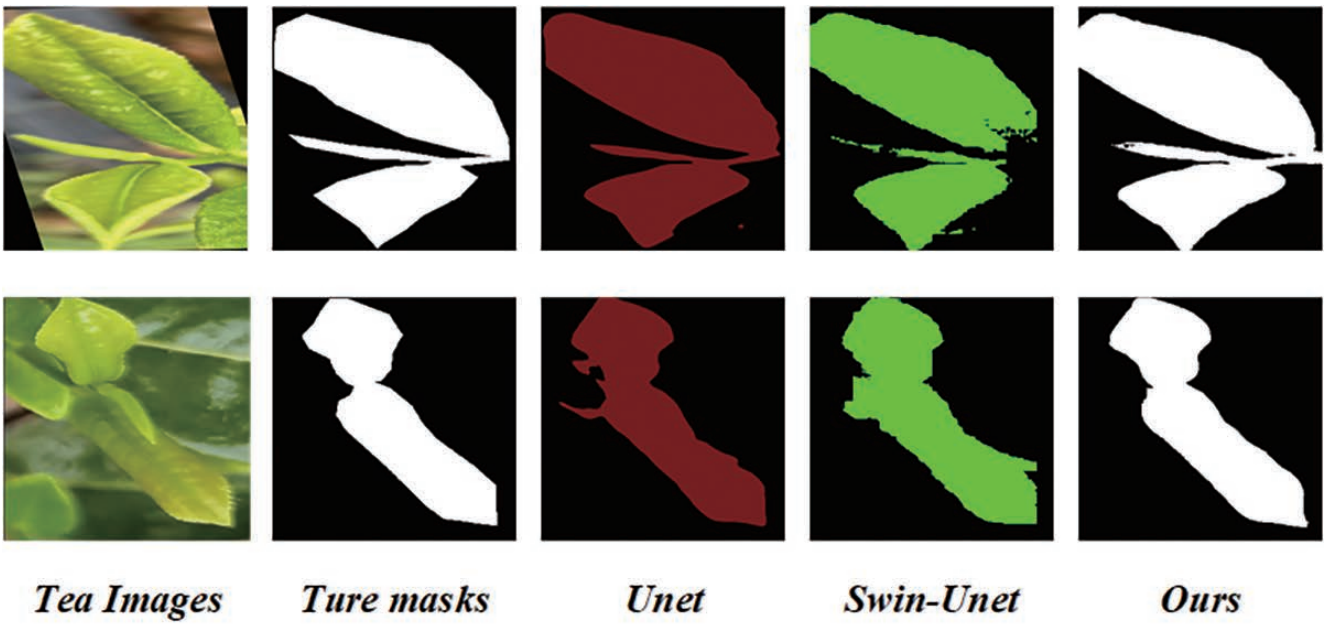

Figure 8. Segmentation predictions of tea shoots by different segmentation models.

Picking experiment results

To preliminarily validate the effectiveness of the method proposed in this paper, an initial simulation experiment platform was set up for picking experiments, as shown in Figure 10.

The specific experimental procedure is as follows:

Start the experimental equipment, turn on the auxiliary light source, and use the monocular camera to capture images from the top of the tea shoots, as shown in Figure 11A.

Detect the images, crop the detection results to generate new images of single tea shoot, segment the top region of the tea shoot using the segmentation model to locate the picking point, and convert the picking point coordinates into coordinates usable by the picking effector.

Package and send the coordinate data to the PLC controller, which controls the motor to drive the picking effector to the specified coordinate position, as shown in Figure 11-B.

Upon reaching the specified position, the picking effector begins to move downward and cuts the tea shoot after moving a certain distance. After the cutting is completed, the picking effector resets, thus completing a picking operation, as shown in Figure 11 C,D.

A total of 10 experiments were conducted, and the experimental results are divided into three parts: tea shoot detection, tea shoot segmentation, and picking performance on the simulation experiment platform.

The first part of the experimental results presents the statistical analysis of tea shoot detection. First, the number of tea shoots on the experimental tea plants was manually counted and recorded. Subsequently, a monocular camera was used to capture multiple images of the tea plants to ensure that the camera's field of view covered the entire tea plant. The acquired images were then input into a tea shoot detection model for processing. The inference results and inference time of the model were recorded. A detection was considered successful if the entire top region of a tea shoot was completely enclosed within the predicted bounding box, and the number of successfully detected tea shoots was then counted. The detection accuracy Pc was calculated as the ratio of the number of successfully detected tea shoots to the previously recorded actual number of tea shoots, while the detection speed per tea shoot Tc was obtained by dividing the inference time by the number of successfully detected tea shoots. The statistical results of tea shoot detection are presented in Figure 12A.

Secondly, the tea shoot segmentation results were statistically evaluated. The top regions of tea shoots successfully identified by the detection model were cropped using the corresponding prediction bounding boxes, producing images containing only a single tea shoot. These cropped images were then fed into the tea shoot segmentation model, with both the inference results and inference speed recorded. Segmentation success was judged based on the criteria of image completeness and contour clarity. The number of successfully segmented tea shoots was thus determined. The statistical results of tea shoot segmentation are presented in Figure 12B.

Thirdly, the picking performance on the simulation experiment platform was analyzed. Picking points were localized based on the combined results of the detection and semantic segmentation models, and the number of successfully located picking points was recorded. These points were used as targets to guide the picking actuator to perform the picking operation. A picking attempt was considered successful if the tea shoot was completely cut by the actuator and delivered into the collection box. For each experimental group, the total number of successfully picked tea shoots and the total time consumed were recorded. Based on these records, the picking success rate Pp and the picking speed per tea shoot Tp were calculated. The statistical results of the simulation experiment platform's picking performance are presented in Figure 12C.

Based on the recorded experimental results, the tea shoot

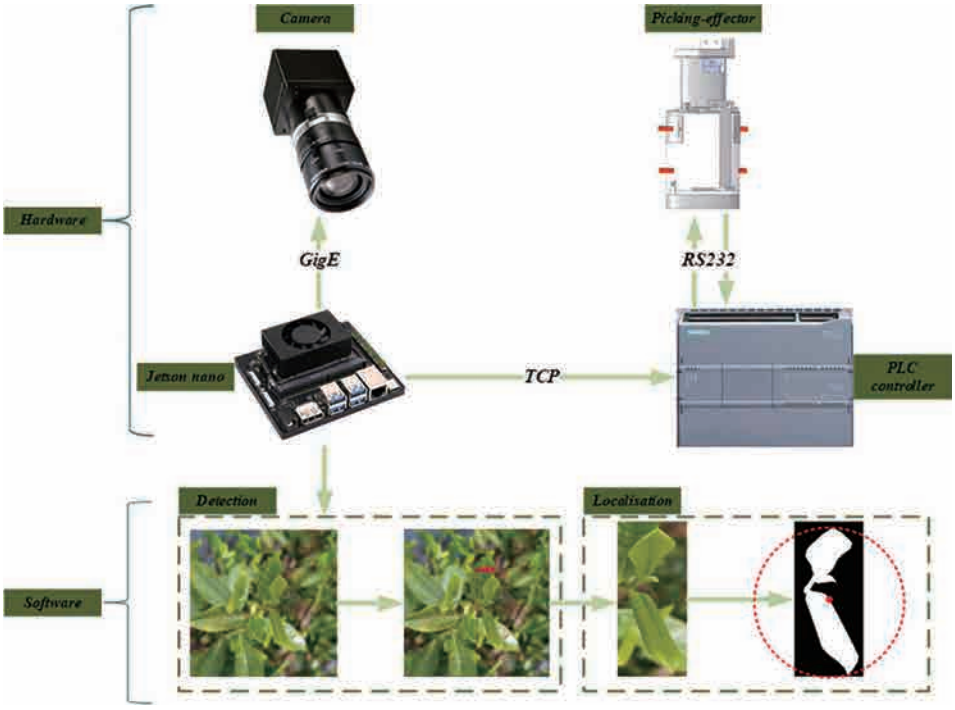

Figure 9. Picking equipment system.

detection task achieved an accuracy of 83.49% with a detection speed of 0.039 seconds per tea shoot. This outcome demonstrates that the proposed tea shoot detection model effectively meets the requirements of rapid and accurate detection necessary for intelligent tea shoot harvesting. Meanwhile, the tea shoot segmentation model exhibited a notable advantage with an accuracy of 78.87% and a relatively fast segmentation speed, indicating its strong capability for precise segmentation of the tea shoot top regions. The failures in detection and segmentation were primarily attributed to incomplete tea shoot images captured by the camera, which resulted from the limited field of view of the imaging device. Furthermore, the statistical analysis of the picking results on the simulation experiment platform showed that the proposed intelligent tea shoot harvesting method achieved a success rate of 75.54%, with an average picking speed of approximately 13 seconds per tea shoot.

Among the tea shoots that were successfully localized but failed to be picked, one main cause was that some tea shoots developed thicker or tougher stems, which increased resistance during the cutting process by the picking actuator, resulting in incomplete cuts and failure to separate the tea shoot. Another reason for picking failure was that tea shoots, although successfully cut, failed to enter the collection box. This phenomenon was caused by moisture adhering to the tea shoots, which led to adhesion between the leaves and the inner wall of the actuator's end sleeve. Such adhesion prevented the negative pressure pump from effectively suctioning the tea shoots through the pipeline into the collection box. Consequently, when the picking actuator reset and the cutting blade retracted, the cut tea shoots dropped, resulting in picking failure. Additionally, before the picking operation, the length of the tea shoots to be picked can be controlled by adjusting the installation distance of the infrared sensors, as shown in Figure 13, with lengths of 10, 7, and 5 cm being picked ,respectively. This method can meet the processing requirements of different tea varieties in practical applications.

Analysis

The method proposed in this paper, which utilizes a monocular camera to detect and locate tea shoots from the top, provides a novel application approach for the automated picking of tea shoots. It eliminates the need for RGB-D cameras to obtain depth information and locate picking points, making it more broadly applicable and stable in practical applications. However, in the experimental process, some tea shoots may not be successfully detected, primarily because their growth shape is inclined, which is different from most tea shoots. Additionally, the recognition model dataset constructed in this paper includes only one type of tea. In subsequent research, to enhance the model's robustness, datasets of various tea types should be created for model training.

In the study, we proposed a U-shaped images segmentation model based on Mamba, which can accurately segment individual tea shoot top regions. However, in the actual working process, a slow picking speed issue arises due to the poor computational power of edge devices and the lack of model acceleration processing. In the future, to enhance the efficiency of the picking task, lightweight processing of the model will be an excellent optimiza-

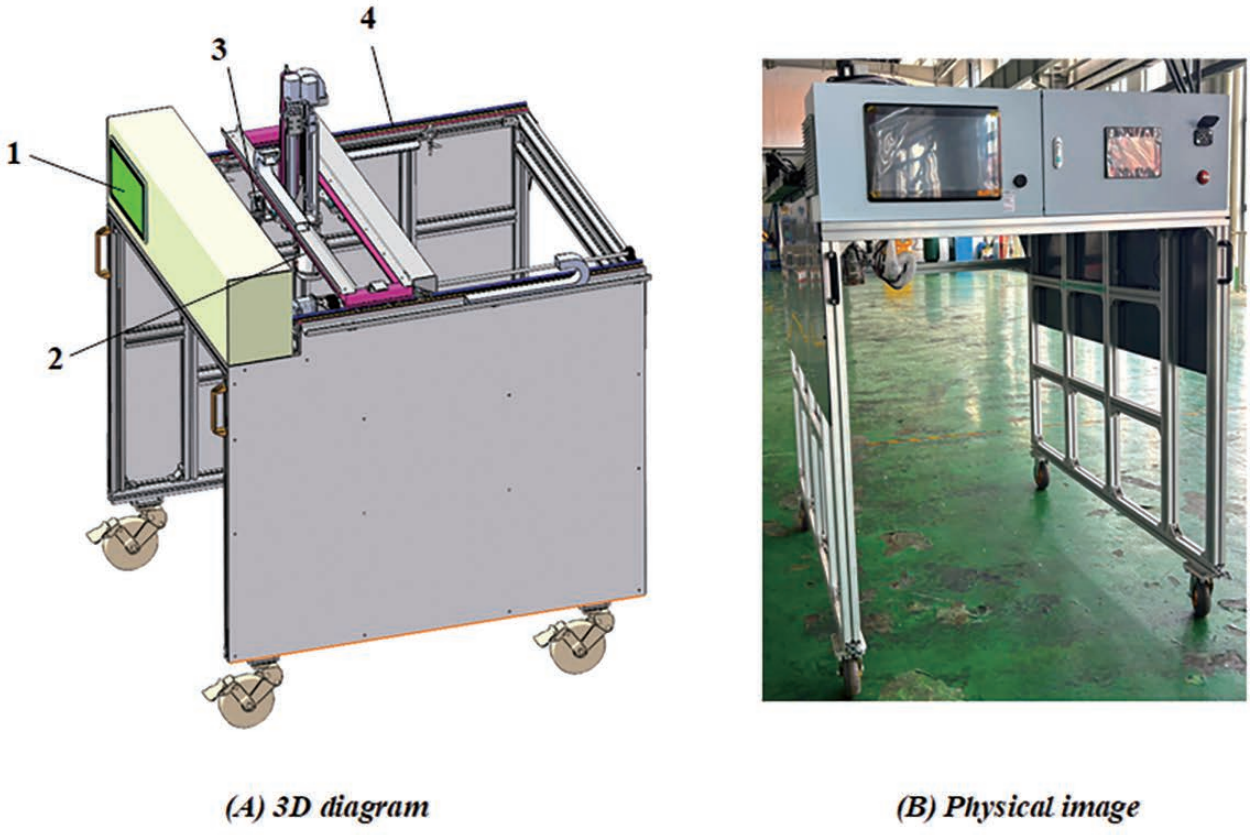

Figure 10. Experiment platform. 1, Controller; 2, picking effector; 3, monocular camera; 4, slide rail device.

tion approach. Additionally, to verify the effectiveness of the method proposed in this paper, we independently designed and developed a picking framework and picking effector. For the picking effector, the presence of tea shoots is detected by the config infrared sensors, which then control the up-and-down movement, cutting, and speed of the picking effector. This structure and picking method are clearly suitable for automated tea picking, but in future optimization schemes, it can be installed and config on a framework capable of autonomous movement in a normal tea garden environment, and multiple picking effectors working together will undoubtedly improve picking efficiency. Therefore, designing a more reasonable automated picking framework based on this picking effector is crucial.

Conclusions

The method proposed in this paper, which uses a monocular camera to detect and locate at the top of tea shoots, aims to solve the issues of extensive occlusion of shoots and loss of depth information encountered when using an RGB-D camera for side-angle detection and localization. First, the top-view images of tea shoots captured by the monocular camera are detected, and individual tea shoot images are cut based on the detection results. Then, the individual tea shoot top images are input into the image segmentation model to obtain accurate mask images, which are used to locate the center position of the tea shoot top. Finally, to verify the effectiveness of the method, we independently designed and developed a picking framework and effector. Then, we conducted experimental verification in a simulated tea garden environment. The following conclusions were drawn from the experimental results: in the study, the image segmentation model proposed in this paper achieved MIoU and ACC of 87.80% and 95.63%, respectively. Compared to other segmentation models, this model demonstrates significant advantages in both performance and visualization results, fully meeting the requirements for precise localization in the automated picking of tea shoots. Additionally, based on the method of using a monocular camera for top-view recognition and localization of tea shoots, a picking effector was independently designed and developed, along with a suitable picking plan. In the simulated tea garden environment, achieving a picking success rate of 75.54%. Furthermore, before the picking operation, the length of the tea shoots to be picked can be controlled by adjusting the installation distance of the infrared sensors to meet the processing requirements of different tea varieties.

In summary, this method avoids the issue of depth information loss and inability to locate picking points encountered with RGB-D cameras, while also providing a new application solution for the automated picking of tea shoots. However, in real tea garden environments, enabling the picking equipment to autonomously move within the garden will be the next challenge. Additionally, we need to improve the picking efficiency of this method and using multiple effectors for picking work will be a promising development direction. Therefore, in future work, designing a reasonable autonomous mobility framework and multi-effector picking strategy for the proposed method will be our focus.

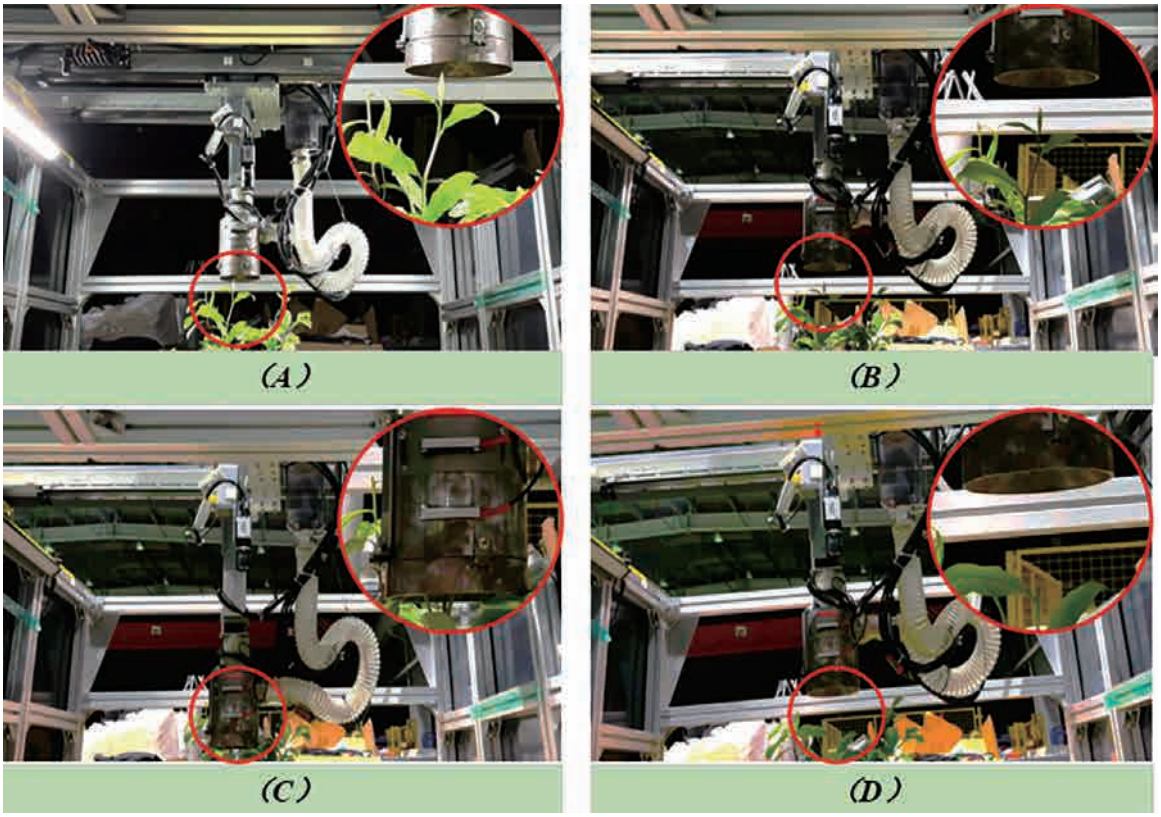

Figure 11. Picking experiment steps: capturing images from the top of the tea shoots (**A**), the picking effector move to the position directly above the tea shoot (**B**), the picking effector cuts the tea shoot (**C**), and the picking effector resets (**D**).

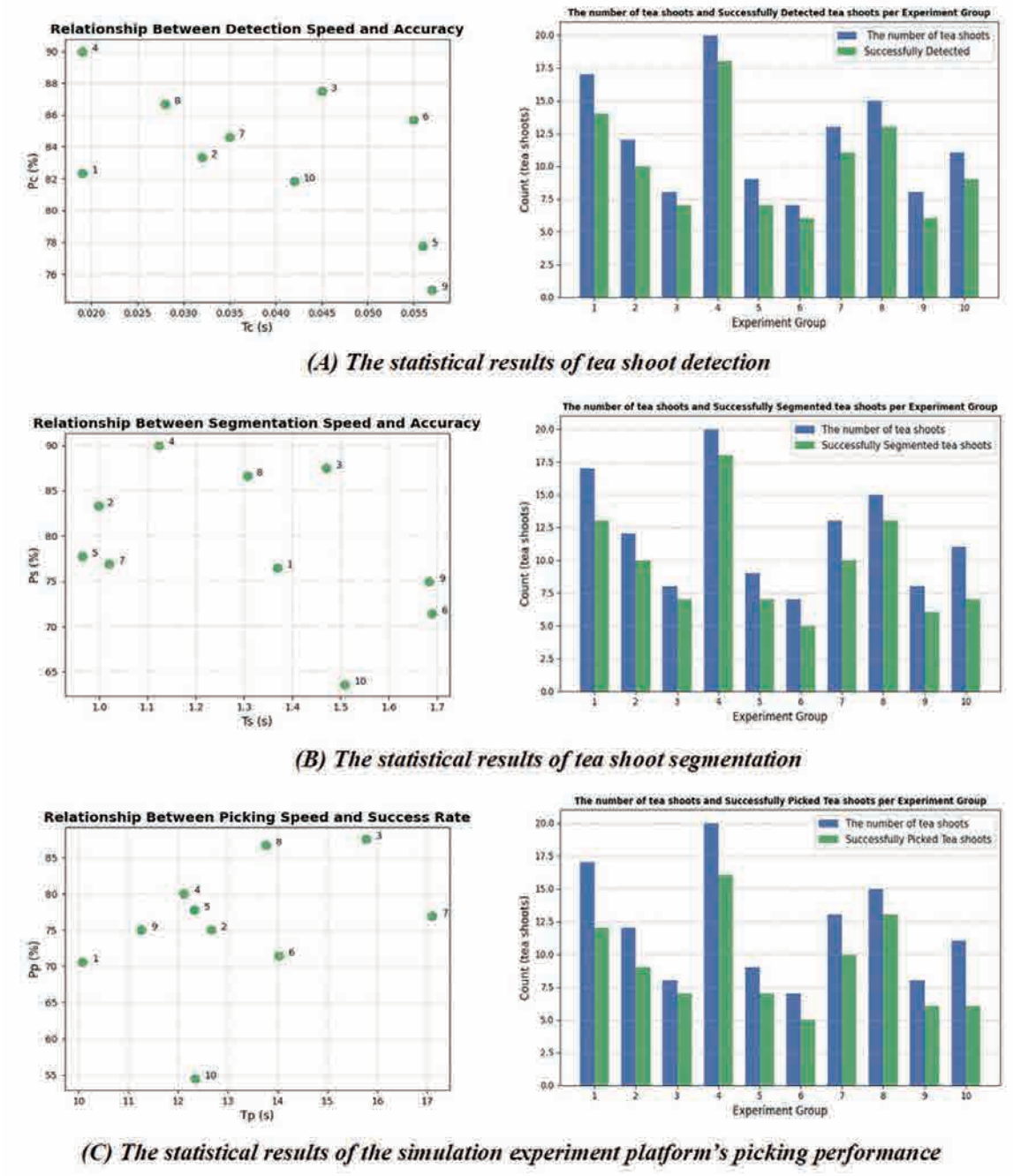

Figure 12. Statistical results of the picking experiment.

Figure 13. Picking results of tea shoots of different lengths.

References

Fu L, Gao F, Wu J, Li R, Karkee M, Zhang Q, 2020. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. Comput Electron Ag. 177:105687.

Gu A, Dao T, 2024. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752.

Han Y, Xiao H, Qin G, Song Z, Ding W, Mei S, 2014. Developing situations of tea plucking machine. Engineering 6:268-273.

Kamilaris A, Prenafeta-Boldú FX, 2018. Deep learning in agriculture: A survey. Comput. Electron Agr 147:70-90.

Karunasena GMKB, Priyankara HDNS, 2020. Tea bud leaf identification by using machine learning and image processing techniques. Int J Sci Eng Res 11:624-628.

Li J, Gao W, Wu Y, Liu Y, Shen Y, 2022. High-quality indoor scene 3D reconstruction with RGB-D cameras: A brief review. Comput Vis Media 8:369-393.

Li Y, He L, Jia J, Lv J, Chen J, Qiao X, Wu C, 2021. In-field tea shoot detection and 3D localization using an RGB-D camera. Comput Electron Agr 185:106149.

Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M, 2020. Deep learning for generic object detection: a survey. Int J Comput Vis 128:261-318.

Long Z, Jiang Q, Wang J, Zhu HL, Li B, Wen F, 2022. Research on method of tea flushes vision recognition and picking point localization. Transducers Microsyst Technol 41:39-41.

Mai CY, Zheng LH, Sun H, Yang W, 2015. Research on 3D reconstruction of fruit tree and fruit recognition and location method based on RGB-D camera. J Agric Mech Eng 46:35-40.

Santos TT, de Souza LL, dos Santos AA, Avila S, 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput Electron Agr 170:105247.

Tang YP, Han WM, Hu AG, Wang WY, 2016. Design and experiment of intelligentized tea-plucking machine for human riding based on machine vision. Trans Chin Soc Agric Mach 47:15-20.

Wang CY, Bochkovskiy A, Liao HYM, 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Vancouver. pp. 7464-7475.

Wang J, Zhang Z, Luo L, Wei H, Wang W, Chen M, Luo S, 2023. DualSeg: Fusing transformer and CNN structure for image segmentation in complex vineyard environment. Comput Electron Agr 206:107682.

Wang ZF, Lu HL, Geng WT, Sun ZQ, 2022. Research on object detection and positioning system of fruit picking robot based on OpenCV. Electr Technol Softw Eng 137-140.

Wu SP, Ca YM, Wang S, Fu QY, Feng Y, Lv LZ, 2018. Effects of different picking periods on biochemical components and quality of Xinyang Maojian tea. Shandong Agric Sci 50:57-60.

Wu XM, Tang X, Zhang FG, Gu JM, 2015. Tea buds image identification based on lab color model and K-means clustering. Trans Chin Soc Agric Machin 36:161-164.

Yang CH, Liu YP, Wang Y, Xiong LY, Xu HB, Zhao WH, 2019. Research and experiment on recognition and location system for citrus picking robot in natural environment. Trans Chin Soc Agric Machin 50:14-22.

Yang H, Chen L, Chen M, Ma Z, Deng F, Li M, Li X, 2019. Tender tea shoots recognition and positioning for picking robot using improved YOLO-V3 model. IEEE Access 7:180998-181011.

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

Zhou SY, Ca L, 2022. Effects of different picking periods and tenderness of tea leaves on the quality of black tea. Chin Food Saf Mag 149-152.

Zhu HC, Li X, Meng Y, Yang HB, Xu Z, Li ZH, 2022. Tea bud detection based on Faster R CNN network. Trans Chin Soc Agric Mach 53:217–224.

Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X, 2024. Vision Mamba: Efficient visual representation learning with bidirectional state space model. arXiv:2401.09417.

Zou Q, Lu AJ, Zhou H, Zhao Q, 2022. An improved YOLOV3 algorithm model for tea bud detection. Laser J 43:70-75.

Received: 21 November 2024; Accepted: 4 November 2025.

Contributions: Zhiqiang Wang, conceptualization, methodology, software, validation, formal analysis, data curation, writing - original draft. Hui Niu, validation, formal analysis, data curation writing - review and editing. Jing Zhang, conceptualization, investigation, methodology, supervision, writing - review & editing. Wu Zhang, investigation, supervision, project administration, funding acquisition. Jian Mao, conceptualization, data curation, investigation. Shengqi Zhang, conceptualization, methodology, software. Jun Liu, data curation, resources, supervision. Guanpeng Zuo, writing - review and editing. Zhe Zheng, writing - review and editing. Zhenxiang Chi, writing - review and editing.

Conflict of interest: the authors declare no competing interests, and all authors confirm accuracy.

Funding: the financial support provided by Key Research and Development Project of Anhui Province (202204c06020022, 202104a06020012), Independent Project of Anhui Key Laboratory of Smart Agricultural Technology and Equipment (APKLSATE2019X001), and The National Natural Science Foundation of China (32371993).

Acknowledgements: this work was supported by Key Research and Development Project of Anhui Province (202204c06020022, 202104a06020012); Independent Project of Anhui Key Laboratory of Smart Agricultural Technology and Equipment (APKLSATE2019X001); and The National Natural Science Foundation of China (32371993). The authors also acknowledge the Anhui Agricultural University Central Anhui Experimental Station, National Key Laboratory of Tea Tree Biology and Resource Utilization Tea Tree Germplasm Resource Nursery.