agriculture

MDPI

Article

Instance Segmentation and 3D Pose Estimation of Tea Bud Leaves for Autonomous Harvesting Robots

Haoxin Li [1], Tianci Chen [1], Yingmei Chen [1], Chongyang Han [1], Jinhong Lv [1], Zhiheng Zhou [2,*] and Weibin Wu [1,3,*]

[1] National Key Laboratory of Agricultural Equipment Technology, College of Engineering, South China Agricultural University, Guangzhou 510642, China
[2] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China
[3] Guangdong Engineering Technology Research Center for Mountainous Orchard Machinery, Guangzhou 510642, China
* Correspondence: zhouzh@scut.edu.cn (Z.Z.); wuweibin@scau.edu.cn (W.W.)

Abstract: In unstructured tea garden environments, accurate recognition and pose estimation of tea bud leaves are critical for autonomous harvesting robots. Due to variations in imaging distance, tea bud leaves exhibit diverse scale and pose characteristics in camera views, which significantly complicates the recognition and pose estimation process. This study proposes a method using an RGB-D camera for precise recognition and pose estimation of tea bud leaves. The approach first constructs an for tea bud leaves, followed by a dynamic weight estimation strategy to achieve adaptive pose estimation. Quantitative experiments demonstrate that the instance segmentation model achieves an mAP@50 of 92.0% for box detection and 91.9% for mask detection, improving by 3.2% and 3.4%, respectively, compared to the YOLOv8s-seg instance segmentation model. The pose estimation results indicate a maximum angular error of 7.76°, a mean angular error of 3.41°, a median angular error of 3.69°, and a median absolute deviation of 1.42°. The corresponding distance errors are 8.60 mm, 2.83 mm, 2.57 mm, and 0.81 mm, further confirming the accuracy and robustness of the proposed method. These results indicate that the proposed method can be applied in unstructured tea garden environments for non-destructive and precise harvesting with autonomous tea bud-leave harvesting robots.

Keywords: YOLOv8s-seg model; adaptive pose estimation; RGB-D camera; precise harvesting

Academic Editor: Vincenzo Alfano

Received: 7 December 2024
Revised: 8 January 2025
Accepted: 16 January 2025
Published: 17 January 2025

Citation: Li, H.; Chen, T.; Chen, Y.; Han, C.; Lv, J.; Zhou, Z.; Wu, W. Instance Segmentation and 3D Pose Estimation of Tea Bud Leaves for Autonomous Harvesting Robots. *Agriculture* **2025**, *15*, 198. https://doi.org/10.3390/agriculture15020198

1. Introduction

Tea is a globally important economic crop, highly regarded for its pharmacological properties and rich nutritional content, making it popular among consumers [1]. Among these, high-quality tea, known for its high nutritional and economic value, is particularly sought after. However, the harvesting of high-quality tea relies on manual labor, with a short harvesting period and strict standards, which makes labor shortages a critical bottleneck that restricts production capacity and hampers the improvement of production efficiency. Especially in China, with the acceleration of population aging and urbanization, the reduction in labor force has further driven up production costs, leading to a decrease in the yield of high-quality tea [2,3]. Therefore, developing a tea-bud-leaves autonomous-harvesting robot is not only a necessary means to address current challenges, but also an inevitable trend to drive the intelligent transformation of the high-quality tea industry and improve production efficiency. In the unstructured tea garden environment, the shape,

scale, and density of tea bud leaves vary significantly, presenting great challenges for the harvesting task [4,5]. While positional information of tea bud leaves provides some spatial data, it is insufficient for achieving precise and non-destructive harvesting. Autonomous harvesting robots may experience failures in harvesting and damage to the target when relying solely on positional data. Hence, three-dimensional (3D) pose estimation of the harvesting target is crucial. The autonomous harvesting robot can adjust its harvesting posture based on the 3D pose information of tea bud leaves, enabling precise and non-destructive harvesting.

The target recognition technology of tea-bud-leaves harvesting robots is crucial for autonomous harvesting. It primarily includes two major approaches: traditional digital image processing techniques and deep learning methods [6]. Traditional digital image processing techniques require manual feature extraction [7], which can provide tea-bud-leaves segmentation results but often suffers from lower accuracy and higher demands for image quality [8,9]. Karunasena et al. [10] combined digital image processing techniques with machine learning to identify tea bud leaves, but their method achieved only 55% accuracy. This approach has limited effectiveness, and fails to adequately address the tea-bud-leaves recognition problem in complex environments. Zhang et al. [11] applied an improved watershed algorithm for tea-bud-leaves segmentation under varying lighting conditions. However, their study tested the method only on tea bud leaves that were not occluded, and the performance of this approach in unstructured tea garden environments has not been fully verified, particularly regarding its ability to handle irregular growth and overlapping leaves. Therefore, traditional digital image processing methods still exhibit significant shortcomings in complex environments. In contrast, deep learning methods extract multi-level features through training, offering significant advantages when facing challenges such as varying lighting conditions and complex backgrounds in unstructured tea gardens [12,13]. Li et al. [14] optimized the Backbone, Neck, and loss function of the YOLOv4 model, achieving tea-bud-leaves target detection in unstructured tea garden environments. Chen et al. [15] used an improved YOLOv7 model for multi-scale and multi-target tea-bud-leaves detection in such environments, attaining an excellent average precision of 94.43%. Other researchers have also reported related work on target detection for tea bud leaves [16–19]. However, target detection of tea bud leaves only provides the bounding box of the target, and the background information within the bounding box can interfere with the harvesting robot, impacting its performance.

Semantic segmentation and instance segmentation are effective methods for accurately determining the locations of tea bud leaves. Lu et al. [20] utilized four methods—Deeplabv3+, U-Net, HRNet_W18, and Fast-SCNN—to segment tea bud leaves, achieving mean intersection-over-union (mIoU) scores of 78.59%, 79.64%, 81.00%, and 74.80%, respectively. These results provide a solid foundation for pinpointing harvesting points for tea bud leaves. Zhang et al. [21] applied an improved version of DeepLabv3+ to generate masks for tea bud leaves, and used YOLOv7 to extract harvesting point locations from these masks. However, the error rate increased significantly when the system faced diverse leaf shapes and occlusion situations. In unstructured tea garden environments, the camera often captures targets that vary in scale and shape, which complicates detection. To address this challenge, Chen et al. [22] incorporated an attention mechanism and a multi-path aggregation model into a semantic segmentation framework for tea bud leaves. This approach substantially enhanced segmentation performance, especially in cases involving leaves with different shapes. While these methods successfully identify the fine-grained positions of tea bud leaves from RGB images, they do not consider the 3D information of the targets, which limits the robot's ability to perform precise operations in complex environments.

The 3D pose of the harvesting target provides the robotic harvester with more accurate spatial information, enabling it to navigate to the specified location and adjust its harvesting posture accordingly [23]. This is critical for achieving non-destructive and precise harvesting [24–26], as illustrated in Figure 1. In recent years, numerous researchers have focused on pose estimation for harvesting targets. Lin et al. [27] proposed a guava pose estimation method using an RGB-D camera, which involved recognizing both the fruit and branches. By leveraging the spatial relationship between the two, they successfully estimated the fruit's pose, achieving an angular error of 23.43° ± 14.18°. Luo et al. [28] utilized Mask R-CNN to detect the grape and flower stem, and then applied the locally weighted scatterplot smoothing (LOWESS) algorithm and geometric analysis for stem pose estimation. The average angular error in stem pose estimation was 22.2°. Zhu et al. [29] introduced a novel method for pitaya pose estimation, using the positional relationship between the fruit and branches to estimate the fruit's pose, achieving an average angular error of 8.8°. While these studies primarily target fruits with regular shapes, applying similar methods to tea bud leaves, which have highly variable morphology, presents significant challenges.

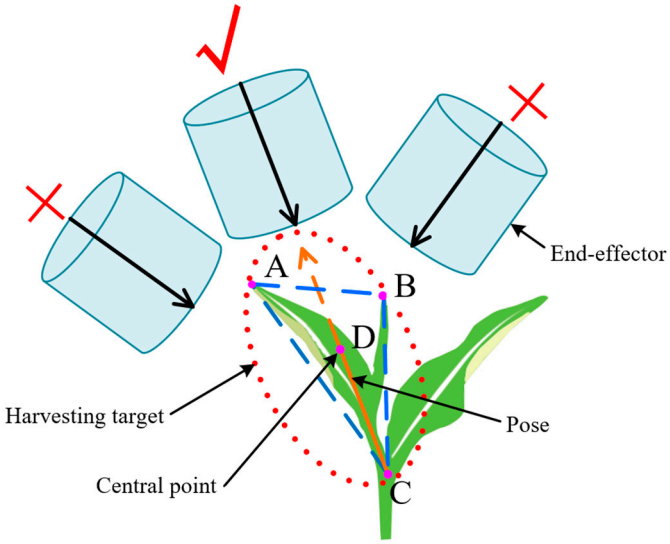

Figure 1. Robot adjusting picking pose based on estimated pose of tea bud leaves. In the figure, A is the apex of the leaf, B is the apex of the tea bud, and C is the lowest point of the stem. D is the centroid of the growth plane formed by A, B, and C. The line connecting D and C defines the pose of the tea bud leaves.

This study addresses the challenging task of pose estimation for tea bud leaves in unstructured tea garden environments by proposing a pose estimation method based on an RGB-D camera. First, an improved YOLOv8s-seg instance segmentation model is employed to obtain precise localization of the tea bud leaves. Then, a dynamic weight-based adaptive estimation method is introduced for the pose estimation of tea bud leaves, enabling accurate pose estimation in unstructured tea garden environments and providing a decision-making basis for the autonomous tea-bud-leaves picking robot.

2. Materials and Methods

2.1. Data Acquisition and Processing

This study focuses on the Yinghong No. 9 cultivar, with image data collected in a natural environment at South China Agricultural University in Guangzhou, Guangdong Province, China. Due to the complex background in tea gardens, the appearance of tea bud images varies with different viewing angles. To enhance the robustness and generalization of the visual model proposed in this study, a multi-angle image data collection experiment

was conducted using the Intel RealSense D405 depth camera. Multiple RGB and depth images were captured from various angles. The RGB images were used for visual model recognition, while the depth images provided 3D information to estimate the pose of the tea bud leaves. The dataset consists of 1434 raw RGB images, which were labeled using the LabelMe software (https://github.com/labelmeai/labelme, accessed on 17 August 2023). The label information is stored in "json" format. During the labeling process, the one-bud-one-leaf approach was applied, with the labels "tea_Y", "tea_I", and "tea_V" assigned to represent three distinct forms of tea bud leaves in the camera's field of view. Specifically, "tea_Y" indicates that the tea bud, tea leaf, and stem are clearly visible; "tea_I" indicates that the tea bud and tea leaf overlap at a given angle, with the stem clearly visible; and "tea_V" indicates that the tea bud and tea leaf are clearly visible, while the stem is occluded. The tea garden environment for data collection and the labeled RGB images are shown in Figure 2.

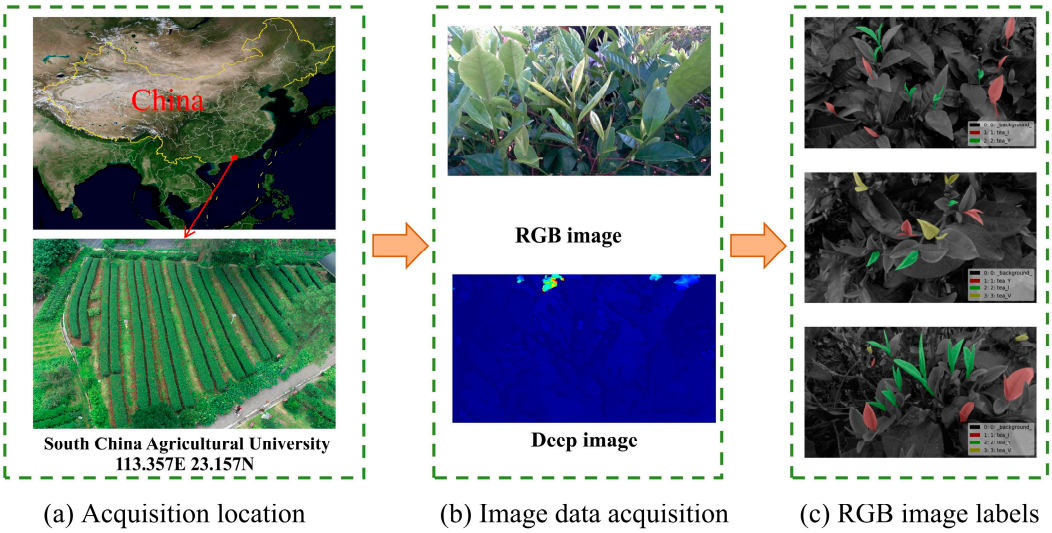

Figure 2. Data collection schematic.

In unstructured tea garden environments, variations in lighting conditions and the diverse morphology of tea bud leaves pose significant challenges to the accurate recognition of tea bud leaves by the visual system. To address these challenges and improve the robustness of the visual recognition model, data augmentation techniques such as mirroring, contrast adjustment, and saturation modification were applied to the original dataset. This resulted in a total of 2868 RGB images. The dataset statistics are shown in Table 1. The dataset includes 6912 samples labeled as "tea_Y", 3762 samples labeled as "tea_I", and 3716 samples labeled as "tea_I". The dataset was split into training, validation, and test sets in an 8:1:1 ratio.

Table 1. Dataset statistics.

Datasets	Image Samples	Label Category		
		tea_Y	tea_I	tea_V
train	2294	5519	3037	2965
val	286	719	353	355
test	288	674	372	396
Total	2868	6912	3762	3716

2.2. Instance Segmentation Model for Tea Bud Leaves

2.2.1. YOLOv8 Segmentation Model

The YOLOv8 model is a SOTA model that has demonstrated significant advantages in tasks such as image classification, object detection, pose estimation and instance segmentation. In this study, the YOLOv8s-seg model is used as the baseline for obtaining multi-morphological masks of tea bud leaves from RGB images. YOLOv8s-seg consists of three main components: the Backbone, the Neck, and the Head. The Backbone is responsible for feature extraction from the input image, producing feature maps at different scales. The Neck integrates these multi-scale feature maps from the Backbone to enhance the model's feature representation capability. The Head then makes accurate predictions based on these features. To address the challenges posed by the diverse morphological characteristics of tea bud leaves and the varying target scales in the complex, unstructured tea garden environment, this study proposes a novel multi-morphological segmentation model for tea bud leaves.

(i) The E-GELAN module is constructed and integrated into the Backbone for feature extraction, excelling at capturing the detailed morphological features and contextual information of tea bud leaves.
(ii) DCNv2 and the Dynamic Head are employed to enhance the Neck and YOLO Head, improving the differential representation of global and local features.
(iii) The Wise-IoUv3 loss function is used to train the model, dynamically adjusting the weights based on the varying shapes and scales of the targets, thereby enhancing the model's adaptability to the unstructured tea garden environment.

The instance segmentation model architecture is shown in Figure 3.

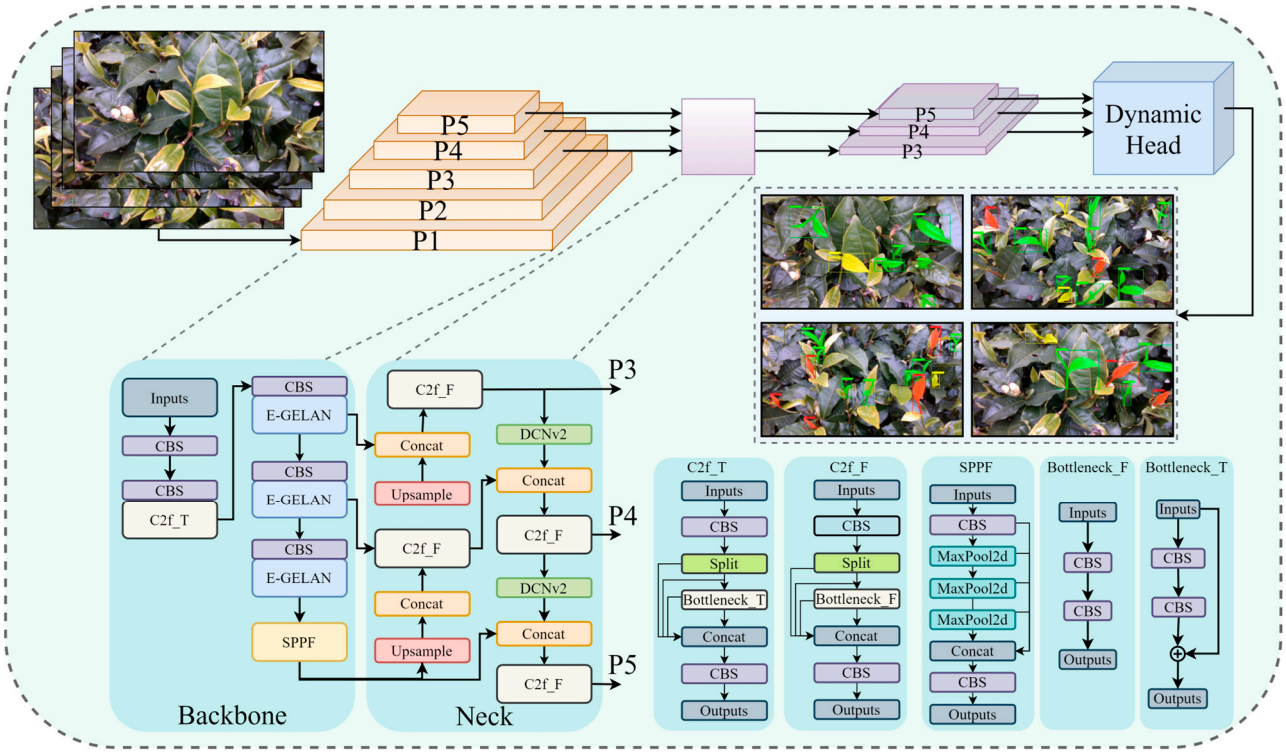

Figure 3. Tea-bud-leaves instance segmentation model.

2.2.2. E-GELAN Module

In this section, we introduce a novel E-GELAN module. The ELAN and Extended-ELAN (E-ELAN) modules proposed in YOLOv7 [30] extract more diverse features along

different gradient paths, thereby enhancing the model's feature extraction capability. YOLOv9 [31] balances inference speed and accuracy by combining the CSPNet module with the ELAN module to design a generalized ELAN (GELAN). Inspired by this approach, we propose a new Extended-GELAN (E-GELAN) module tailored for the multi-morphological characteristics of tea bud leaves in the field of view. The E-GELAN module adopts a multi-path aggregation strategy to enhance the ability to extract diverse features, enabling the comprehensive capture of the morphological details and contextual information of tea bud leaves. The structure of the E-GELAN module is shown in Figure 4.

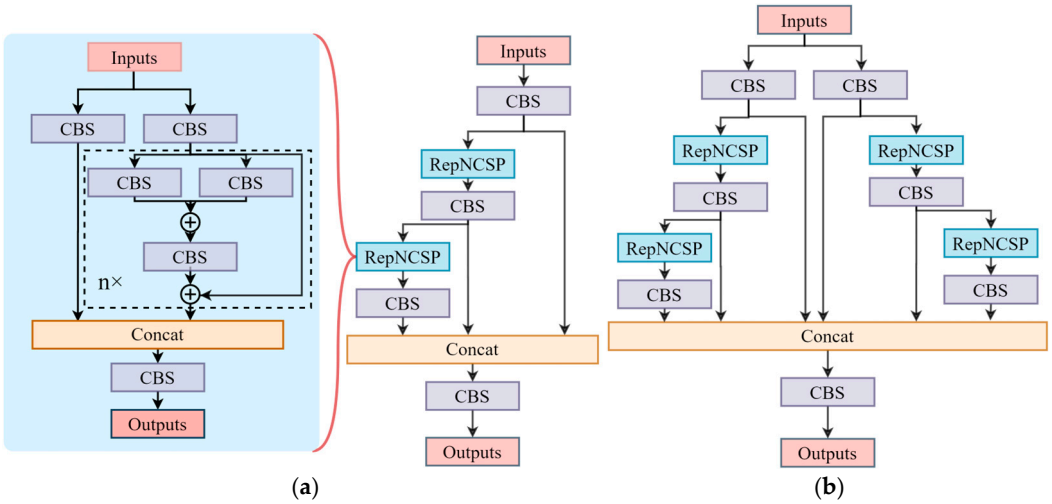

Figure 4. GELAN and E-GELAN modules. (**a**) GELAN module; (**b**) E-GELAN module.

2.2.3. DCNv2 and Dynamic Head

In unstructured tea garden environments, tea bud leaves exhibit diverse geometric characteristics, making it difficult for traditional convolution operations to capture the detailed morphological features. To improve the model's ability to express both global and local features distinctly, the convolutional module in the Neck of the YOLOv8s-seg model is replaced with Deformable ConvNets v2 (DCNv2) [32]. This modification aims to guide the model in adapting to variations in receptive field size across different morphological shapes of tea bud leaves. The structure of the modified model is shown in Figure 5.

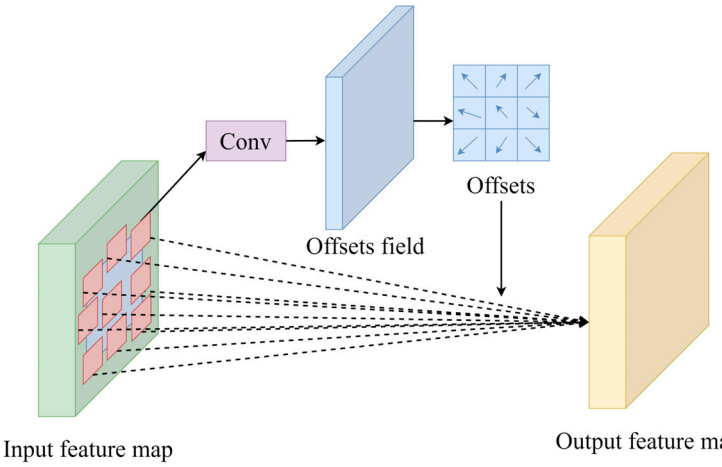

Figure 5. DCNv2.

DCNv2 employs learnable offsets and modulation scalars to design a deformable convolution kernel that adaptively adjusts its shape, enabling it to capture different morphological features of the targets in the input feature map. First, the size of the deformable

convolution kernel and the predefined offsets are specified. For instance, with a 3×3 deformable convolution kernel, the number of sampling positions is $K = 9$, and the predefined offset is $p_k \in \{(-1,-1),(-1,0),\cdots(1,1)\}$. The input feature map x is then processed through DCNv2, and the output feature map y is computed as shown in Equation (1).

$$y(p) = \sum_{k=1}^{K} \mu_k \bullet x(p + p_k + \Delta p_k) \bullet \Delta m_k \tag{1}$$

where p represents the position in the feature map, μ_k represents the weight at the sampling position, and Δp_k and Δm_k represent the learnable offset and modulation scalar at the sampling position, respectively. Δp_k is an unconstrained real number, while $\Delta m_k \in [0,1]$.

For the tea-bud-leaves instance segmentation model, an efficient and accurate Head is particularly crucial. Although the YOLOv8s-seg Head has achieved notable success in object detection, it still faces limitations in predicting tea bud leaves in unstructured tea garden environments, primarily due to its reliance on a single feature for prediction. To address this issue, this study adopts the Dynamic Head [33] for prediction, which incorporates scale-awareness, spatial-awareness, and task-awareness capabilities. This approach effectively enhances the model's recognition accuracy for multi-morphological tea bud leaves in unstructured tea garden environments. The structure of the Dynamic Head is shown in Figure 6.

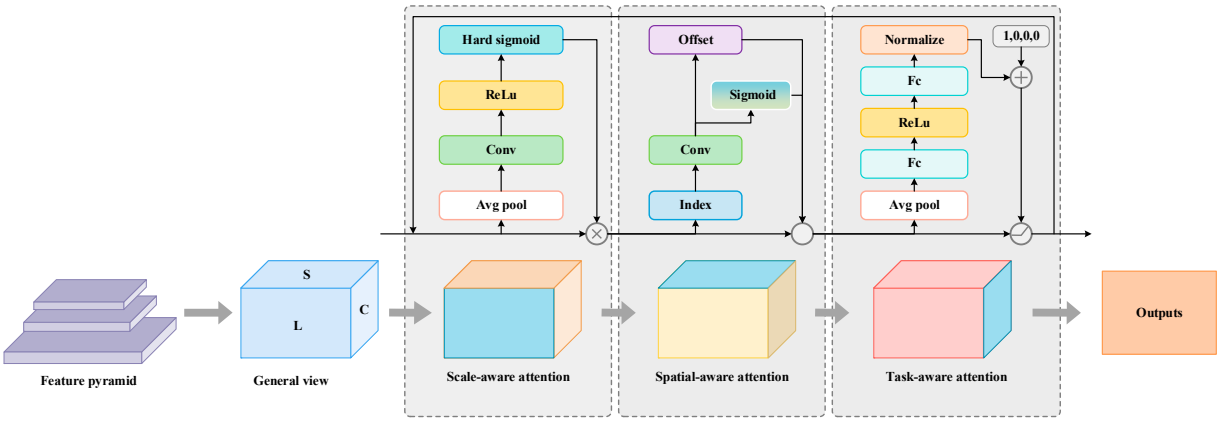

Figure 6. Dynamic Head.

The Dynamic Head receives outputs from different levels of the Neck and performs upsampling or downsampling to unify the scale, constructing a 4-dimensional tensor $F \in R^{L \times H \times W \times C}$. If expressed as $S = H \times W$, the output is a 3-dimensional tensor $F \in R^{L \times S \times C}$. The Dynamic Head applies attention mechanisms to different dimensions, with the specific calculations shown in Equation (2).

$$W(F) = \pi_C(\pi_S(\pi_L(F) \bullet F) \bullet F) \bullet F \tag{2}$$

where π_L, π_S and π_C represent attention modules for the three different dimensions: L, S and C, respectively. The scale-aware attention module π_L is used to fuse features across different scales in the feature map, the spatial-aware attention module π_S enhances the model's spatial position discrimination capability, and the task-aware attention module π_C dynamically activates, according to the specific task. Their expressions are given by the following equation:

$$\pi_L(F) \bullet F = \sigma(f(\frac{1}{SC}\sum_{S,C} F)) \bullet F \tag{3}$$

$$\pi_S(F)\bullet F = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}\mu_{l,k}\bullet F(l; p_k + \Delta p_k; c)\bullet \Delta m_k \tag{4}$$

$$\pi_C(F)\bullet F = \max(\alpha^1(F)\bullet F_c + \beta^1(F)\bullet F_c, \alpha^2(F)\bullet F_c + \beta^2(F)\bullet F_c) \tag{5}$$

where f represents the linear transformation function of the 1×1 convolution, with $\sigma(x) = \max(0, \min(1, \frac{x+1}{2}))$. During the calculation of π_S, the parameter variables are consistent with those in DCNv2. In the calculation of π_C, $[\alpha^1, \beta^1, \alpha^2, \beta^2]^T$ is a hyper function that controls the activation threshold, and F_c refers to the feature slices of the channels.

2.2.4. Wise-IoUv3 Loss Function

In image segmentation tasks, the loss function is used to evaluate the discrepancy between the predicted values and the ground truth. For tea-bud-leaves instance segmentation, the targets are relatively small, and often become obscured by the complex background, which makes traditional IoU loss functions inadequate for meeting the recognition requirements. To address this challenge, this study employs Wise-IoUv3 [34] as the loss function. Wise-IoUv3 adaptively adjusts the weights based on the size and category of the targets, thereby improving the detection performance for difficult-to-detect targets. The calculation of Wise-IoUv3 is shown in Equation (6).

$$\begin{cases} L_{WIoU} = L_{IoU}\bullet R_{WIoU}\bullet r \\ L_{IoU} = 1 - IoU \\ R_{WIoU} = \exp(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)^*}) \\ r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \end{cases} \tag{6}$$

where L_{IoU} represents the traditional IoU loss function, and a smaller value indicates better model prediction performance. R_{WIoU} denotes the distance metric function between the predicted and ground-truth bounding boxes. x, y, x_{gt}, and y_{gt} represent the coordinates of the predicted and ground-truth boxes, while W_g and H_g denote the width and height of the minimum enclosing rectangle for the predicted and ground-truth boxes, respectively. $(W_g^2 + H_g^2)^*$ represents the normalization factor, ensuring that the results are not influenced by the size of the enclosing rectangle. r refers to the gradient gain, and β denotes the outlier degree. α and δ are hyperparameters.

2.3. Dynamic Weight-Based Adaptive-Pose Estimation Method for Tea Bud Leaves

2.3.1. Tea-Bud-Leaves Local Point-Cloud Acquisition Based on ORBSLAM3

In the process of constructing the complete shape of tea bud leaves using an RGB-D camera, significant challenges arise due to the discrepancies between the RGB images and depth images obtained from a single-position sampling. As a result, it is essential to combine point clouds from multiple-position sampling to create a comprehensive 3D representation of the target [35]. Vision-based Simultaneous Localization and Mapping (VSLAM) is a technology that enables real-time estimation of both the 3D structure of the environment and the camera's position. ORBSLAM3 [36], an advanced VSLAM algorithm, performs exceptionally well in complex environments by detecting and matching ORB feature points to extract key points from the field of view, and by employing graph optimization techniques to construct dense point clouds. In this study, ORBSLAM3 is used to estimate the camera position in real time, obtaining the rotation matrix R and translation matrix T, as shown in Figure 7. The depth image is then converted into a 3D point cloud using the camera's intrinsic parameters. By utilizing the rotation matrix R and translation matrix T from multiple-position sampling, the 3D point clouds from these different posi-

tions are accurately aligned to generate a global, dense point cloud of the target. Finally, the instance segmentation model's output mask is used to delineate the target area, extracting the local point cloud representing the complete shape of the tea bud leaves.

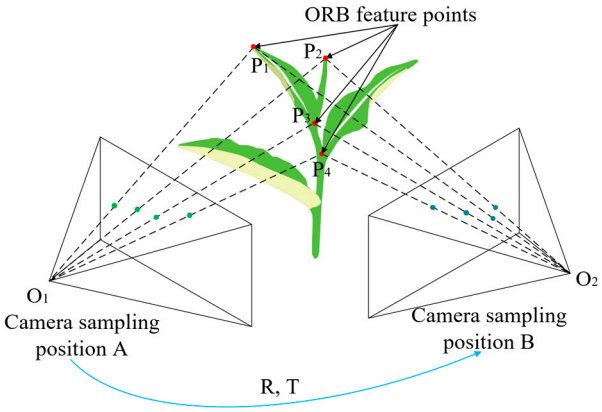

Figure 7. ORBSLAM3 algorithm overview.

A comparison between the point clouds obtained from a single sampling position and those fused from multiple sampling positions, as shown in Figure 8, clearly demonstrates the advantages of using ORBSLAM3 to estimate the camera's pose from multiple sampling positions when constructing the complete shape of the tea-bud-leaves point cloud.

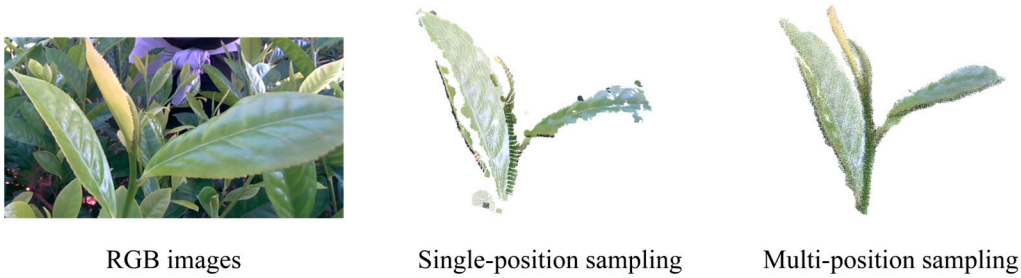

Figure 8. Comparison of local point cloud obtained from single-position sampling and multiple-position sampling for tea bud leaves.

2.3.2. Point Cloud Pre-Processing

Dense point clouds provide a more accurate representation of the morphological features of tea bud leaves. However, due to environmental factors such as camera limitations, local point clouds of tea bud leaves may contain noise and outliers, which can affect the accuracy and speed of pose estimation. Therefore, filtering and down-sampling the acquired tea-bud-leaves point clouds are crucial steps in the pose estimation process. Statistical Outlier Removal (SOR) filtering is an effective method for outlier removal. This approach calculates the average distance between each point and its neighboring points, as well as the global standard deviation of distances, to determine whether a point is an outlier. Specifically, for the tea-bud-leaves point cloud set $\{T_1, T_2, \ldots, T_n\}$, the average distance of a target point cloud to its neighboring points and the global standard deviation of distances are computed as follows:

$$\begin{cases} d_i = \frac{1}{k}\sum_{j=1}^{k}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \\ \varepsilon = \frac{1}{n}\sum_{i=1}^{n} d_i \\ \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \varepsilon)^2} \end{cases} \tag{7}$$

where d_i represents the average distance between point cloud $T_i(x_i, y_i, z_i)$ and the k-nearest points within the neighborhood, ε denotes the global average distance of the tea-bud-leaves point-cloud set consisting of n points, and σ is the standard deviation of the global average distance. For a spatial point cloud $T_i(x_i, y_i, z_i)$ with distance d_i falling within the standard range $[\varepsilon - \lambda\sigma, \varepsilon + \lambda\sigma]$, the point is retained; otherwise, it is considered an outlier, and removed. In the process of SOR filtering for 3D point clouds, a larger k value helps improve the smoothness of the filtering and effectively reduces the influence of noise on the results. However, it also significantly increases the computational cost, resulting in longer processing times. On the other hand, a smaller λ value leads to more points being classified as outliers and subsequently filtered out, which aids in removing anomalies and noise. However, if λ is too small, it may erroneously discard normal points, thus affecting the quality of the point cloud. Based on multiple tests, in this study, $k = 30$ and $\lambda = 2$ were found to effectively balance noise suppression and computational efficiency. This combination not only removes a large amount of noise, but also preserves the essential morphological features of the tea bud leaves in the point cloud.

Due to the significantly larger surface area of the tea leaves compared to the tea bud and stem, the point-cloud density on the leaves is considerably higher. This imbalance may lead to the tea-bud-leaves point cloud becoming trapped in local optima during pose estimation. To mitigate this issue, it is necessary to down-sample the tea-bud-leaves point cloud. Voxel Grid Down-sampling (https://github.com/PointCloudLibrary/pcl, accessed on 20 August 2023) is a widely used method for this purpose, which divides the 3D space into voxels and replaces all points within each non-empty voxel with the voxel's centroid.

2.3.3. Dynamic Weight-Based Adaptive Pose Estimation for Tea Bud Leaves

The use of the symmetry axis of fruits as an indirect representation of their pose has been reported in numerous studies. However, high-quality tea with one-bud-one-leaf does not exhibit symmetry. As shown in Figure 1, since the growth of tea bud leaves follows the direction of the stem, this study adopts three vertices, A, B, and C, of the tea bud leaves to define the primary growth plane. The center point of this growth plane is denoted as D. The pose of the tea bud leaves is determined by the line connecting D and C.

To represent the fundamental morphological features of tea bud leaves, the point cloud density on the leaves is typically higher than that of the bud and stem. In order to accurately estimate the pose of the tea bud leaves, the point cloud can be segmented into longitudinal layers, and the centroid of each layer's point cloud is used to represent the entire point-cloud set of the tea bud leaves. The weighted least squares method is advantageous for linear fitting of low-noise data, as it assigns different weights to various point-cloud data, fully considering the differences between data points and their influence on the fitting results. Let P represent the point-cloud set of centroids for each layer, where $(x_i, y_i, z_i) \in P$, $i = 1, 2, \cdots, n$. The objective function for the weighted least squares method is expressed as follows:

$$\begin{cases} D_x = (X - AM_x)^T W (X - AM_x) \\ \frac{\partial D_x}{\partial M_x} = -2AW(X - AM_x) \end{cases} \tag{8}$$

$$\begin{cases} D_y = (X - AM_y)^T W(Y - AM_y) \\ \frac{\partial D_y}{\partial M_y} = -2AW(Y - AM_y) \end{cases} \tag{9}$$

where D_x and D_y represent the objective functions for fitting the point-cloud set along the X and Y axes, respectively. $M_x = [a_x, b_x]^T$ and $M_y = [a_y, b_y]^T$ are the parameters of the lines fitted along the X and Y axes. $\frac{\partial D_x}{\partial M_x}$ and $\frac{\partial D_y}{\partial M_y}$ denote the partial derivatives, and the optimal values of M_x and M_y can be obtained by solving this equation when the partial derivatives are set to zero. $X = [x_1, x_2, \cdots x_n]^T$ and $Y = [y_1, y_2, \cdots y_n]^T$ represent the X and Y axis vectors of the point-cloud set P. W is the weight matrix, and A is the design matrix. Their expressions are given by the following equation:

$$W = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} \tag{10}$$

$$A = \begin{bmatrix} z_1 & 1 \\ z_2 & 1 \\ \vdots & \vdots \\ z_n & 1 \end{bmatrix} \tag{11}$$

W is a diagonal matrix, where distinct weights are assigned to each data point. Specifically, $\omega_i = e^{\alpha z_i}$, where α is the weight coefficient. This weight coefficient determines the contribution of each feature point cloud in the pose estimation process. However, due to the diverse morphological characteristics of tea leaves in natural tea garden environments, a static weight design fails to accurately capture and estimate the morphological features of tea bud leaves at different scales. To enhance the robustness of the pose estimation algorithm, this study utilizes Particle Swarm Optimization (PSO) [37] to dynamically optimize the weight coefficient α, effectively capturing and quantifying the morphological features of tea bud leaves. PSO is a population-based heuristic algorithm in which particles share information to evaluate the fitness of positions. The fitness function constructed in this study is as follows:

$$f(k) = \max_{p \in boundary(pcd)} \left(\frac{\left\| \overrightarrow{LP} \times \overrightarrow{d} \right\|}{\| \overrightarrow{d} \|} \right) \tag{12}$$

where $boundaty(pcd)$ represents the boundary point set of the preprocessed tea-bud-leaves point cloud, while P denotes the feature point cloud on this set. L and $\overrightarrow{d}$ refer to any point and direction vector along the principal axis of the tea-bud-leaves point cloud, respectively.

In the optimization iteration process of PSO, each particle continuously updates its velocity and position based on fitness information, thereby guiding the swarm towards the global optimal solution. The velocity and position updates of the particle in the search space are given by the following equations:

$$\begin{cases} v_i^{t+1} = \beta v_i^t + c_1 r_1 (pbest_i^t - s_i^t) + c_2 r_2 (gbest^t - s_i^t) \\ s_i^{t+1} = s_i^t + v_i^{t+1} \end{cases} \tag{13}$$

where v_i^t and s_i^t represent the velocity and position of the particle at time step t, respectively. $pbest_i^t$ denotes the personal best position of the i particle, while $gbest^t$ represents the global best position at time step t. β is the inertia weight, c_1 and c_2 are the cognitive and social learning factors, and r_1 and r_2 are random vectors.

2.4. Evaluation Metrics

2.4.1. Instance-Segmentation Evaluation Metrics

The performance of the multi-morphological tea-bud-leaves segmentation model is crucial for subsequent pose estimation. In this study, Average Precision (AP) and Mean Average Precision (mAP) are used to evaluate the performance of the multi-morphological tea-bud-leaves segmentation model. AP is a commonly used metric in instance segmentation, consisting of precision and recall, which reflects the overall recognition performance for a specific class. mAP is the average of the AP values across different classes, providing a comprehensive measure of the model's recognition performance for all targets. Their specific forms are given by the following equations:

$$AP = \int_0^1 P(R)dR \times 100\% \tag{14}$$

$$mAP = \frac{\sum_{n=1}^{N} AP_n}{N} \times 100\% \tag{15}$$

$$P = \frac{TP}{TP + FP} \tag{16}$$

$$R = \frac{TP}{TP + FN} \tag{17}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

2.4.2. Pose-Estimation Evaluation Metrics

In this study, the angular error θ and distance error d between the estimated pose and the ground-truth pose of tea bud leaves are used to evaluate the performance of the dynamic weight-based adaptive-pose estimation method. The specific formulations are provided below:

$$\theta = \arccos \frac{\vec{u}_1 \cdot \vec{u}_2}{\left|\vec{u}_1\right| \cdot \left|\vec{u}_2\right|} \tag{18}$$

$$d = \frac{(\vec{u}_1 \times \vec{u}_2) \cdot \vec{MN}}{\left|\vec{u}_1 \times \vec{u}_2\right|} \tag{19}$$

where $\vec{u}_1$ and $\vec{u}_2$ represent the direction vectors of the estimated and ground-truth poses, respectively, while $\vec{MN}$ represents the vector connecting any two corresponding points between the two poses.

To further comprehensively assess the performance of the dynamic weight-based adaptive-pose estimation method, several metrics are used to analyze pose estimation errors: maximum error, mean error, median error, and median absolute deviation. These are calculated as follows:

$$\begin{cases} \theta_m = \max(\theta_n) \\ \overline{\theta} = \sum_{n=1}^{N} \theta_n \\ \theta_{MEDE} = \text{median}(\theta_n) \\ \theta_{MAE} = \text{median}(|\theta_n - \theta_{MEDE}|) \end{cases} \tag{20}$$

$$\begin{cases} d_m = \max(d_n) \\ \overline{d} = \sum_{n=1}^{N} d_n \\ d_{MEDE} = \text{median}(d_n) \\ d_{MAE} = \text{median}(|d_n - d_{MEDE}|) \end{cases} \quad (21)$$

where θ_m, d_m, $\overline{\theta}$, $\overline{d}$, θ_{MEDE}, d_{MEDE}, θ_{MAE}, and d_{MAE} represent the maximum error, mean error, median error, and median absolute deviation of the angular and distance errors, respectively.

3. Results and Discussion

3.1. Tea-Bud-Leaves Instance-Segmentation-Model Performance Evaluation

3.1.1. Ablation Experiments

Table 2 shows the impact of various improvements on the instance segmentation model's performance. The results indicate that using the E-GELAN module to construct the backbone network improves the recognition performance for all three types of tea bud leaves, demonstrating the model's strong ability in feature extraction and contextual information integration. With the improvements to the Neck using DCNv2 and the adoption of the Dynamic Head, mAP@50 is further enhanced, highlighting its capacity to express diverse features and mitigate information loss during the downsampling process. Finally, by training the model with Wise-IoUv3, the mAP@50 for box and mask levels reached 92.0% and 91.9%, respectively, representing improvements of 3.2% and 3.4% over the original model.

Table 2. Experimental comparison results of different combinations.

YOLOv8s-seg	E-GELAN	DCNv2	Dynamic Head	Wise-IoUv3	AP@50/Box (%)			AP@50/Mask (%)			mAP@50/Box(%)	mAP@50/Mask(%)
					tea_Y	tea_I	tea_V	tea_Y	tea_I	tea_V		
√					92.7	88.3	85.4	92.7	88.4	84.4	88.8	88.5
√	√				93.1	90.2	90.2	92.7	89.6	89.7	91.2	90.7
√	√	√			93.4	91.8	88.6	93.2	91.8	88.3	91.3	91.1
√	√	√	√		94.1	91.9	89.3	94.0	91.4	88.4	91.8	91.3
√	√	√	√	√	94.4	91.3	90.4	94.3	91.0	90.2	92.0	91.9

3.1.2. Loss-Function Comparison Experiment

Table 3 presents the results of training the model with different loss functions. Model A refers to the model obtained by improving the YOLOv8s-seg model with E-GELAN, DCNv2, and Dynamic Head. The traditional IoU loss function lacks distance information when there is no overlap between the predicted and ground-truth boxes, leading to optimization issues during the training process. To address this, the GIoU loss function introduces the concept of the minimum enclosing box, and measures the distance by calculating the minimal enclosing rectangle between the predicted and ground-truth boxes, which better guides the model's training [38]. However, when the minimum enclosing boxes are identical, the GIoU loss function fails to distinguish the relative positioning between the predicted and ground-truth boxes. In response, the DIoU loss function incorporates the distance between the center points of the predicted and ground-truth boxes, which accelerates model convergence [39]. However, the DIoU loss function does not account for the aspect ratio of the boxes. The EIoU loss function computes the length and width of both the predicted and ground-truth boxes, providing a measure of directional loss [40]. However, these loss functions are susceptible to sample imbalance. The Wise-IoUv3 loss

function addresses this issue with a dynamic non-monotonic focusing mechanism, offering more robust guidance during model training. Experimental results demonstrate that the Wise-IoUv3 loss function is more suitable for the tea-bud-leaves instance segmentation task in unstructured tea garden environments.

Table 3. Experimental comparison results of different loss functions.

Model A	Wise-IoUv3	GIoU	DIoU	EIoU	AP@50/Box (%)			AP@50/Mask (%)			mAP@50/ Box(%)	mAP@50/ Mask(%)
					tea_Y	tea_I	tea_V	tea_Y	tea_I	tea_V		
√	√				92.7	88.3	85.4	92.7	88.4	84.4	88.8	88.5
√		√			93.7	92.4	89.2	93.7	92.0	89.0	91.8	91.6
√			√		93.8	92.3	88.1	93.5	92.0	87.8	91.4	91.1
√				√	92.6	89.5	87.8	92.6	89.6	87.8	90.0	90.0

3.1.3. Visualization of Instance-Segmentation Results

Figure 9 shows the segmentation results of the YOLOv8-seg model and the instance segmentation model proposed in this study. The magnified regions on the right display the segmentation errors. Based on the results, it can be observed that both models perform well when the target in the image is relatively large. However, when the target is small, the YOLOv8-seg model underperforms, as shown in Image 1, Image 2, and Image 4 in Figure 9. Additionally, the varying shapes of tea bud leaves pose a significant challenge for the model. For example, as shown in Image 3, when the stem is short, the YOLOv8-seg model erroneously detects the tea bud and the second leaf as one object. These results demonstrate the advanced performance of the proposed instance segmentation model in handling tea bud leaves with different scales and morphological characteristics.

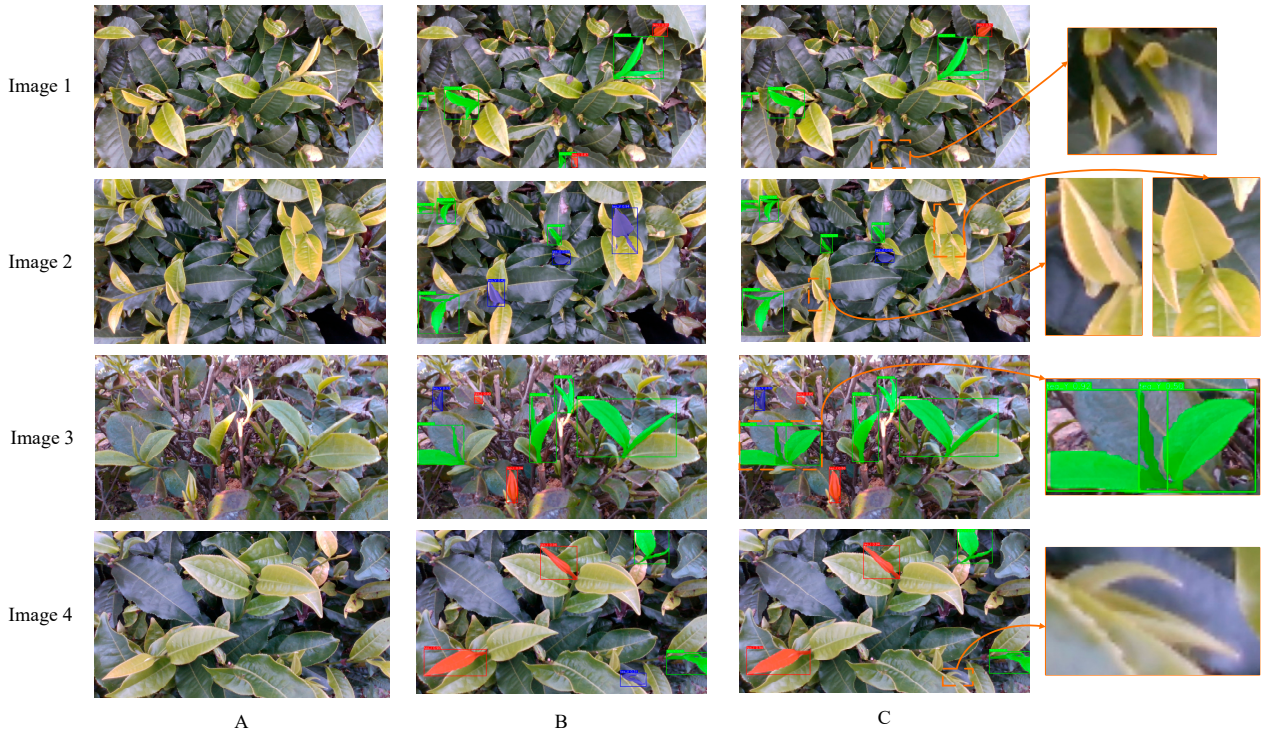

Figure 9. Tea-bud-leaves instance segmentation results. (**A**) Original Image, (**B**) proposed instance segmentation model, (**C**) YOLOv8s-seg model.

3.1.4. Comparison with Advanced Segmentation Models

To more comprehensively evaluate the performance of the multi-morphology tea-bud-leaves segmentation network model, a comparison was made between the proposed instance segmentation model and other advanced segmentation models, including Mask R-CNN [41], Cascade Mask R-CNN [42], YOLACT [43], and YOLACT++ [44]. The experimental results are presented in Table 4.

Table 4. Experimental-comparison results of different models.

Model	mAP@50 (%)		mAP@50-90 (%)	
	Box	Mask	Box	Mask
This Paper	92.0	91.9	86.0	72.4
Mask R-CNN	75.7	73.9	57.4	49.7
Cascade Mask R-CNN	80.6	78.8	64.5	52.3
YOLACT	86.0	84.4	70.0	53.9
YOLACT++	88.1	85.3	72.6	57.7

Mask R-CNN is developed based on Faster R-CNN, utilizing ResNet as the backbone network for feature extraction and FPN for fusing features at different levels. It then classifies each pixel using FCN on top of the original classification and regression tasks to complete the segmentation. Cascade Mask R-CNN, like Mask R-CNN, is a two-stage model. However, Cascade Mask R-CNN introduces a cascade structure that continuously optimizes the predicted targets. This structure uses different IoU thresholds to train the model, effectively addressing the issue of a lack of positive samples at high thresholds while avoiding the problem of poor correction performance in high IoU regions at low thresholds. The experimental results indicate that, in unstructured tea garden environments, Cascade Mask R-CNN outperforms Mask R-CNN in tea-bud-leaves recognition, though there is still potential for further improvement.

YOLACT and YOLACT++ are one-stage models that use feature pyramid networks, which effectively enhance the correlation of contextual information. In terms of performance, they are comparable to two-stage models. YOLACT and YOLACT++ construct overall image prototype masks to distinguish the foreground and background using high-resolution feature maps, and complete the instance segmentation task by applying mask coefficients. This design provides greater flexibility when handling objects with varying morphologies. Additionally, YOLACT++ introduces operations like variable convolutions to further improve the model's recognition ability. The experiments show that YOLACT and YOLACT++ outperform Mask R-CNN and Cascade Mask R-CNN in tea-bud-leaves instance segmentation tasks involving diverse morphological features.

In the unstructured tea garden environment, the proposed multi-morphology tea-bud-leaves segmentation model demonstrates significant advantages over these mainstream high-performance models. The mAP@50 for box and mask levels is 92% and 91.9%, respectively, while the mAP@50-95 is 86% and 72.4%. These improvements can be attributed to the series of optimizations specifically designed to address the morphological diversity and other characteristics of tea bud leaves in this task.

3.2. Performance Evaluation of Tea-Bud-Leaves Pose Estimation

3.2.1. Angle-Error Evaluation

In this section, we conducted experiments using the proposed dynamic weight-based adaptive estimation method and the least squares method to estimate the pose of tea bud leaves. Figure 10 and Table 5 display the comparison of the angle errors between the two methods. The dynamic weight-based adaptive estimation method proposed in

this study resulted in a maximum error of 7.76° and an average error of 3.41°, whereas the least squares method produced a maximum error of 20.97° and an average error of 10.58°. The maximum error was reduced by 67.77%, and the average error decreased by 81.53%. These reductions in both maximum and average errors effectively demonstrate the significant advantage of the dynamic-weighted adaptive estimation method in overall estimation accuracy. Moreover, this method shows better stability and reliability in the complex, unstructured tea garden environment. To minimize the influence of outlier data, median error and median absolute deviation were also used for evaluation. The dynamic weight-based adaptive estimation method yielded a median error of 3.69° and a median absolute deviation of 1.42°, while the least squares method resulted in a median error of 10.06° and a median absolute deviation of 2.90°. These results further highlight the high efficiency of the proposed method in tea-bud-leaves pose estimation tasks.

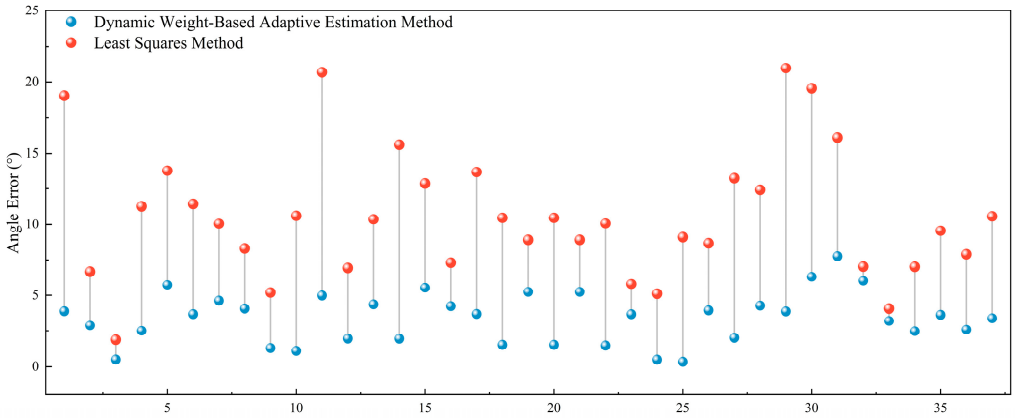

Figure 10. Angle errors in tea-bud-leaves pose estimation.

Table 5. Evaluation metrics for angle errors in tea-bud-leaves pose estimation.

Method	Maximum Error (°)	Average Error (°)	Median Error (°)	Median Absolute Deviation (°)
Dynamic Weight-Based Adaptive Estimation Method	7.76	3.41	3.69	1.42
Least Squares Method	20.97	10.58	10.06	2.90

3.2.2. Distance-Error Evaluation

The pose of tea bud leaves is represented as a line in 3D space. A single angular error does not fully reflect the deviation between the estimated and true poses; it is essential to also consider their spatial distance. Figure 11 and Table 6 display the results for the dynamic weight-based adaptive estimation method and the least squares method. The dynamic weight-based adaptive estimation method resulted in a maximum error of 8.60 mm and an average error of 2.83 mm, while the least squares method produced a maximum error of 19.75 mm and an average error of 7.15 mm. The maximum error and average error were reduced by 56.43% and 60.37%, respectively. Additionally, the dynamic weight-based adaptive estimation method yielded a median error of 2.57 mm and a median absolute deviation of 0.81 mm, compared to 6.69 mm and 1.99 mm for the least squares method. These distance-error metrics further highlight the superiority of the dynamic weight-based adaptive estimation method for tea-bud-leaves pose-estimation tasks involving multiple morphological features.

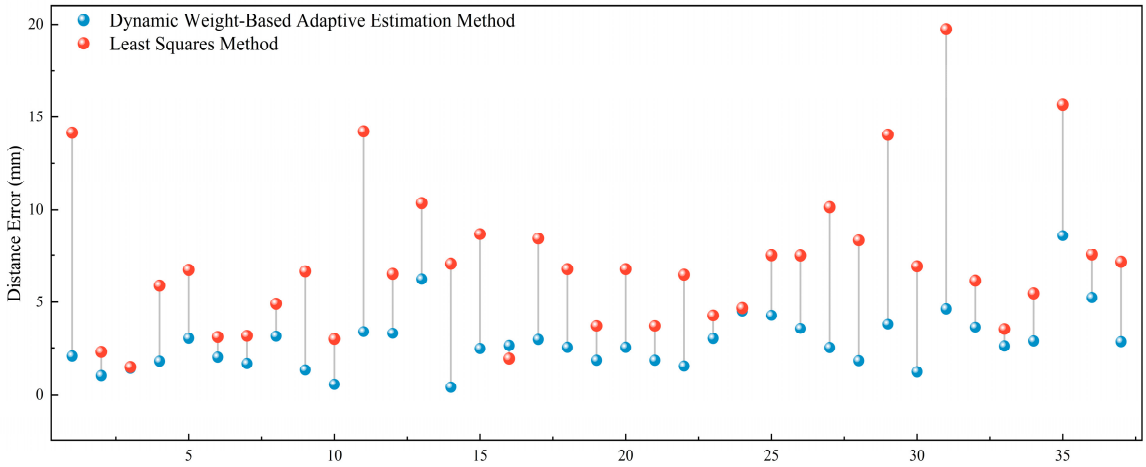

Figure 11. Distance errors in tea-bud-leaves pose estimation.

Table 6. Evaluation metrics for distance errors in tea-bud-leaves pose estimation.

Method	Maximum Error (mm)	Average Error (mm)	Median Error (mm)	Median Absolute Deviation (mm)
Dynamic Weight-Based Adaptive Estimation Method	8.60	2.83	2.57	0.81
Least Squares Method	19.75	7.15	6.69	1.99

3.2.3. Comparison with Other Pose-Estimation Methods

Existing pose-estimation methods primarily focus on relatively regular-shaped fruit objects, such as guava, grape, pitaya, and sweet pepper. Li et al. [23] estimated the pose of sweet peppers using the symmetry axis, leveraging the fruit point-cloud normals and a scoring strategy. The performance of this method depends on the quality of the point cloud; it performs well when the target is unobstructed and the point-cloud quality is high. However, its effectiveness in handling occlusions has not been fully validated. Lin et al. [27] achieved pose estimation for guava by establishing positional constraints between the fruit and branches. However, branches are smaller and share more similarities with the environment than the fruit, making pose estimation more challenging. As a result, this method introduces considerable errors in practical applications. Luo et al. [28] applied the LOWESS algorithm to fit the point cloud of grape pedicels, followed by geometric analysis for pose estimation. While this method shows some effectiveness in dealing with occluded fruit, it demands high-quality point-cloud data, specifically the depth information of the pedicels captured during the image acquisition process. Zhu et al. [29] combined the 3D bounding box of the fruit with the geometric features between the fruit and branches for pitaya-fruit pose estimation. However, in unstructured orchard environments, determining the relationship between the fruit and branches remains challenging, due to the growth characteristics of pitaya, which affects the accuracy of the pose estimation. These studies provide different approaches for fruit-pose estimation.

In contrast to the aforementioned research, the shape of tea bud leaves is more complex, and significantly influenced by environmental factors, making existing methods difficult to apply effectively for pose estimation. Table 7 presents the results of these studies. It is noteworthy that these studies only analyze the angular error in pose estimation. In comparison, the method proposed in this study offers significant advantages, primarily due to its dynamic weight-estimation strategy. This strategy allows for the precise capture

and quantitative representation of the various morphological features exhibited by tea bud leaves, thereby enhancing the robustness and generalization capability of the algorithm.

The datasets of tea bud leaves were collected in an unstructured tea garden environment. However, due to variations in outdoor lighting, mutual occlusion of the tea bud leaves, and differences in imaging scale, depth images are often disturbed by noise and contain missing depth information. These factors lead to sparsity and incompleteness of the tea-bud-leaves point cloud, which in turn affects the accuracy of pose estimation. The quality of the tea-bud-leaves point cloud is a critical factor influencing the precision of pose estimation, with excessive noise and point-cloud loss being the primary sources of estimation errors.

The pose estimation results for tea bud leaves with point clouds of varying quality are shown in Figure 12, where the red line represents the true pose of the tea bud leaves, and the blue line indicates the pose estimated using the dynamic weight-based estimation method. When part of the point cloud of the tea bud or leaf is missing, although some detailed information is lost in the missing areas, the remaining point cloud still retains the basic morphological features of the tea bud leaves, leading to relatively small estimation errors. This phenomenon is illustrated in Figure 12a,b, indicating that when the shape of the tea bud leaves is relatively complete, the impact of point cloud loss on pose estimation is minimal. However, the stem, as the supporting structure of the tea bud leaves, plays a crucial role in maintaining the overall shape. Missing point clouds of the stem, or point-cloud loss at the intersection of the stem and tea bud, significantly affects the morphological features of the target, resulting in a larger impact on pose estimation, as shown in Figure 12c,d.

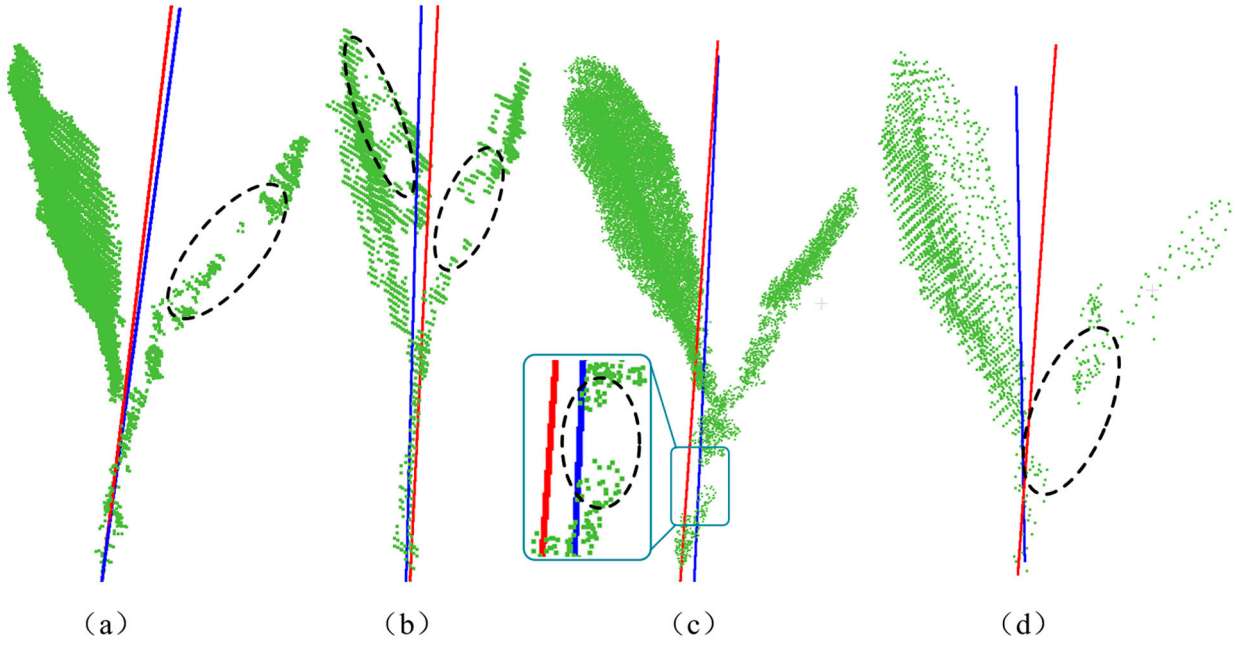

Figure 12. Pose-estimation results for tea-bud-leaves point clouds of varying quality. (**a**) Partial point-cloud loss of the tea bud; (**b**) partial point-cloud loss of both tea bud and leaf; (**c**) partial point-cloud loss of the stem; (**d**) point-cloud loss at the intersection of the stem and tea bud.

Although the quality of the point-cloud data affects the accuracy of pose estimation, the proposed adaptive estimation method based on dynamic weights adjusts the weights of valid point clouds by capturing the morphological features of the tea bud leaves. This reduces the interference of invalid data on pose estimation and keeps the error within a reasonable range.

Table 7. Comparison of pose-estimation results from different works.

Reference	Research Object	Maximum Error (°)	Average Error (°)	Median Error (°)	Median Absolute Deviation (°)
This Paper	Tea Bud Leaves	7.76	3.41	3.69	1.42
[23]	Sweet Pepper	9.34	7.37	-	-
[27]	Guava	-	-	23.43	14.18
[28]	Grape	-	22.22	-	-
[29]	Pitaya	-	8.8	-	-

3.2.4. Visualization of Pose-Estimation Results

The pose estimation of tea bud leaves is visualized in Figure 13, where the red line represents the true pose of the tea bud leaves, and the blue line indicates the pose estimated using the dynamic weight-based estimation method. The quality of pose estimation largely depends on the target's point-cloud data. When the target is larger within the camera's field of view, the RGB-D camera captures more complete depth information, reducing the estimation error, as shown in Image 1 and Image 3 of Figure 13. For the autonomous harvesting robot of tea bud leaves, pose information, within certain error limits, can still facilitate non-destructive and precise harvesting. The pose estimation method proposed in this study consistently yields results close to the true pose for tea bud leaves with varying morphologies, fulfilling the requirements of the autonomous harvesting robot.

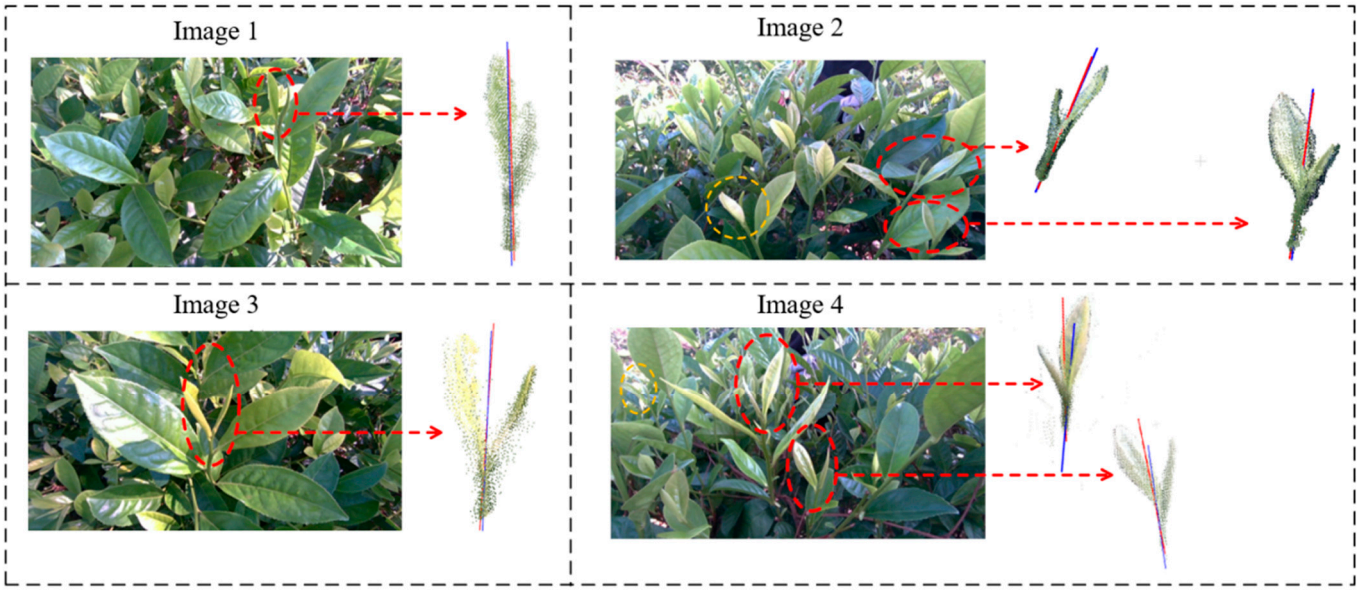

Figure 13. Tea-bud-leaves pose-estimation results.

While the dynamic weight-based estimation method for tea bud leaves demonstrates exceptional performance in unstructured tea garden environments, some limitations persist. Images 2 and 4 in Figure 13 illustrate instances where pose estimation fails, despite the camera's view. Two primary factors contribute to these failures: (i) in unstructured and complex tea garden environments, factors such as lighting conditions can impact the depth-data acquisition of tea bud leaves, especially for smaller buds, and (ii) when the tea bud leaves are small within the camera's field of view, the RGB-D camera struggles to capture complete depth information, and the visual model may fail to detect the target, resulting in a failed pose estimation. To address these challenges, future research could focus on the following aspects: (i) developing efficient point-cloud-processing algorithms that leverage tea-bud-leaves morphology prior to more effectively handling targets with

significant noise and missing point-cloud data, thereby improving the success rate of obtaining complete point clouds of tea bud leaves, and (ii) investigating adaptive coordination control algorithms where the autonomous tea-bud-leaves harvesting robot can dynamically adjust its shooting angle based on the target's scale and morphological characteristics, thus enhancing the success rate of pose estimation.

3.3. Limitations and Future Work

This study has successfully achieved pose estimation for tea bud leaves in unstructured environments. However, certain limitations remain that warrant further exploration and refinement in future research.

(i) This study was validated using the Yinghong No. 9 tea variety, and lacks a comprehensive evaluation of its applicability to other tea varieties. The tea buds and stems of Yinghong No. 9 are relatively robust, and its leaves are broader, with distinct geometric features, providing relatively clear targets for visual recognition and pose estimation. However, the morphological characteristics of different tea varieties differ significantly, such as narrower leaves and more delicate tea buds and stems. These differences may affect the performance of the visual recognition model and the accuracy of pose estimation. Therefore, future research should expand to include the recognition and pose estimation of multiple tea varieties, particularly in addressing the morphological variations brought about by different tea varieties, and focus on improving the adaptability of the algorithm to these changes.

(ii) This study highlights the fact that environmental factors, such as lighting variations and occlusions, are major sources of pose-estimation errors, due to their impact on the quality of the collected data. Although the adaptive estimation method based on dynamic weights proposed in this study can alleviate the effects of noise and point-cloud loss, to some extent, it has not fully addressed the issue of pose-estimation errors caused by a decline in point-cloud quality due to lighting changes or occlusions. Future research should focus on analyzing performance under various lighting conditions and occlusion scenarios to enhance the algorithm's adaptability to complex environmental factors. In particular, when the tea-bud-leaves point-cloud quality is compromised, it is crucial to investigate how prior knowledge of tea-bud-leaves morphology can be leveraged to reconstruct missing details, thereby improving the stability and accuracy of pose estimation. Furthermore, future work will aim to optimize the computational efficiency of the algorithm, explore its application in autonomous tea-bud-leaves harvesting robots, and assess its contribution to the efficiency of the harvesting process.

4. Conclusions

The recognition and pose estimation of tea bud leaves is crucial for the autonomous harvesting robot, as they enable the robot to perform precise and damage-free picking. In this study, we propose a method for tea-bud-leaves instance segmentation and pose estimation, using an RGB-D camera. Experimental results demonstrate the excellent performance of the proposed method. The main conclusions are as follows:

(i) The tea-bud-leaves instance segmentation model is based on the YOLOv8s-seg model. By optimizing the Backbone, Neck, Head, and loss functions, the mAP@50 for box and mask were improved to 92.0% and 91.9%, respectively, showing improvements of 3.2% and 3.4%, compared to the original model. This result demonstrates the robustness of the proposed instance segmentation model for tea-bud-leaves segmentation in unstructured environments.

(ii) This study propose a dynamic weight-based adaptive pose-estimation method for tea bud leaves, which dynamically adjusts the weight coefficients, based on the morphology

of the tea bud leaves using a PSO algorithm. This approach effectively addresses the challenge posed by the diversity in tea-bud-leaves morphology. Experimental results show that the maximum angle error, mean error, median error, and median absolute deviation are 7.76°, 3.41°, 3.69°, and 1.42°, respectively. The corresponding distance errors are 8.60 mm, 2.83 mm, 2.57 mm, and 0.81 mm.

Author Contributions: The presented work was under the supervision of Z.Z. and W.W.; H.L.: conceptualization, methodology, software, and writing—original draft; T.C. and Y.C.: validation, writing—review and editing; C.H.: methodology and writing—review; J.L.: validation; Z.Z. and W.W.: conceptualization, methodology, software and writing—review. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the 2024 Rural Revitalization Strategy Special Funds Provincial Project (2023L204), Guangdong Province (Shenzhen) Digital and Intelligent Agricultural Service Industrial Park (FNXM012022020-1), Construction of Smart Agricultural Machinery and Control Technology Research and Development, and the 2023 Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams (2023KJ120).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article.

Acknowledgments: The authors acknowledge the editors and reviewers for their constructive comments and all the support in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yu, X.L.; He, Y. Optimization of tea-leaf saponins water extraction and relationships between their contents and tea (*Camellia sinensis*) tree varieties. *Food Sci. Nutr.* **2018**, *6*, 1734–1740. [CrossRef] [PubMed]
2. Dong, Q.; Murakami, T.; Nakashima, Y. Recalculating the agricultural labor force in China. *China Econ. J.* **2018**, *11*, 151–169. [CrossRef]
3. Zhu, Y.; Wu, C.; Tong, J.; Chen, J.; He, L.; Wang, R.; Jia, J. Deviation tolerance performance evaluation and experiment of picking end effector for famous tea. *Agriculture* **2021**, *11*, 128. [CrossRef]
4. Zhang, S.; Yang, H.; Yang, C.; Yuan, W.; Li, X.; Wang, X.; Zhang, Y.; Cai, X.; Sheng, Y.; Deng, X.; et al. Edge device detection of tea leaves with one bud and two leaves based on shuffleNetv2-YOLOv5-lite-E. *Agronomy* **2023**, *13*, 577. [CrossRef]
5. Lin, Y.K.; Chen, S.F.; Kuo, Y.F.; Liu, T.L.; Lee, X.Y. Developing a guiding and growth status monitoring system for riding-type tea plucking machine using fully convolutional networks. *Comput. Electron. Agric.* **2021**, *191*, 106540. [CrossRef]
6. Zhao, C.-T.; Wang, R.-F.; Tu, Y.-H.; Pang, X.-X.; Su, W.-H. Automatic Lettuce Weed Detection and Classification Based on Optimized Convolutional Neural Networks for Robotic Weed Control. *Agronomy* **2024**, *14*, 2838. [CrossRef]
7. Hua, X.; Li, H.; Zeng, J.; Han, C.; Chen, T.; Tang, L.; Luo, Y. A review of target recognition technology for fruit picking robots: From digital image processing to deep learning. *Appl. Sci.* **2023**, *13*, 4160. [CrossRef]
8. Wu, X.; Tang, X.; Zhang, F.; Gu, J. Tea buds image identification based on lab color model and K-means clustering. *J. Chin. Agric. Mech.* **2015**, *36*, 161–164+179. [CrossRef]
9. Zhang, L.; Zhang, H.; Chen, Y.; Dai, S.; Li, X.; Kenji, L.; Liu, Z.; LI, M. Real-time monitoring of optimum timing for harvesting fresh tea leaves based on machine vision. *Int. J. Agric. Biol. Eng.* **2019**, *12*, 6–9. [CrossRef]
10. Karunasena, G.; Priyankara, H. Tea bud leaf identification by using machine learning and image processing techniques. *Int. J. Sci. Eng. Res.* **2020**, *11*, 624–628. [CrossRef]
11. Zhang, L.; Zou, L.; Wu, C.; Jia, J.; Chen, J. Method of famous tea sprout identification and segmentation based on improved watershed algorithm. *Comput. Electron. Agric.* **2021**, *184*, 106108. [CrossRef]
12. Wang, Z.; Wang, R.; Wang, M.; Lai, T.; Zhang, M. Self-supervised transformer-based pre-training method with General Plant Infection dataset. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Urumqi, China, 18–20 October 2024; pp. 189–202.
13. Wang, R.-F.; Su, W.-H. The Application of Deep Learning in the Whole Potato Production Chain: A Comprehensive Review. *Agriculture* **2024**, *14*, 1225. [CrossRef]
14. Li, J.; Li, J.; Zhao, X.; Su, X.; Wu, W. Lightweight detection networks for tea bud on complex agricultural environment via improved YOLO v4. *Comput. Electron. Agric.* **2023**, *211*, 107955. [CrossRef]

15. Chen, T.; Li, H.; Chen, J.; Zeng, Z.; Han, C.; Wu, W. Detection network for multi-size and multi-target tea bud leaves in the field of view via improved YOLOv7. *Comput. Electron. Agric.* **2024**, *218*, 108700. [CrossRef]
16. Xie, S.; Sun, H. Tea-YOLOv8s: A tea bud detection model based on deep learning and computer vision. *Sensors* **2023**, *23*, 6576. [CrossRef]
17. Xu, W.; Zhao, L.; Li, J.; Shang, S.; Ding, X.; Wang, T. Detection and classification of tea buds based on deep learning. *Comput. Electron. Agric.* **2022**, *192*, 106547. [CrossRef]
18. Chen, Y.-T.; Chen, S.-F. Localizing plucking points of tea leaves using deep convolutional neural networks. *Comput. Electron. Agric.* **2020**, *171*, 105298. [CrossRef]
19. Li, Y.; He, L.; Jia, J.; Chen, J.; Lyu, J.; Wu, C. High-efficiency tea shoot detection method via a compressed deep learning model. *Int. J. Agric. Biol. Eng.* **2022**, *15*, 159–166. [CrossRef]
20. Lu, J.; Yang, Z.; Sun, Q.; Gao, Z.; Ma, W. A machine vision-based method for tea buds segmentation and picking point location used on a cloud platform. *Agronomy* **2023**, *13*, 1537. [CrossRef]
21. Zhang, F.; Sun, H.; Xie, S.; Dong, C.; Li, Y.; Xu, Y.; Zhang, Z.; Chen, F. A tea bud segmentation, detection and picking point localization based on the MDY7-3PTB model. *Front. Plant Sci.* **2023**, *14*, 1199473. [CrossRef] [PubMed]
22. Chen, T.; Li, H.; Lv, J.; Chen, J.; Wu, W. Segmentation Network for Multi-Shape Tea Bud Leaves Based on Attention and Path Feature Aggregation. *Agriculture* **2024**, *14*, 1388. [CrossRef]
23. Li, H.; Zhu, Q.; Huang, M.; Guo, Y.; Qin, J. Pose estimation of sweet pepper through symmetry axis detection. *Sensors* **2018**, *18*, 3083. [CrossRef] [PubMed]
24. Lehnert, C.; Sa, I.; McCool, C.; Upcroft, B.; Tristan, P. Sweet pepper pose detection and grasping for automated crop harvesting. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2428–2434. [CrossRef]
25. Tao, Y.; Zhou, J. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. *Comput. Electron. Agric.* **2017**, *142*, 388–396. [CrossRef]
26. Li, T.; Feng, Q.; Qiu, Q.; Xie, F.; Zhao, C. Occluded apple fruit detection and localization with a frustum-based point-cloud-processing approach for robotic harvesting. *Remote Sens.* **2022**, *14*, 482. [CrossRef]
27. Lin, G.; Tang, Y.; Zou, X.; Xiong, J.; Li, J. Guava detection and pose estimation using a low-cost RGB-D sensor in the field. *Sensors* **2019**, *19*, 428. [CrossRef]
28. Luo, L.; Yin, W.; Ning, Z.; Wang, J.; Wei, H.; Chen, W.; Lu, Q. In-field pose estimation of grape clusters with combined point cloud segmentation and geometric analysis. *Comput. Electron. Agric.* **2022**, *200*, 107197. [CrossRef]
29. Zhu, L.; Lai, Y.; Zhang, S.; Wu, R.; Deng, W.; Guo, X. Improved U-Net Pitaya Image Segmentation and Pose Estimation Method for Picking Robot. In *Transactions of the Chinese Society for Agricultural Machinery*; Nong Ye Ji Xie Xue Bao Bian Ji Bu: Beijing, China,, 2023; pp. 1–16. Available online: http://kns.cnki.net/kcms/detail/11.1964.S.20230920.1558.002.html (accessed on 7 December 2024).
30. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
31. Wang, C.Y.; Yeh, I.H.; Mark; Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2025; pp. 1–21.
32. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
33. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Lu, Y.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
34. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
35. Yao, M.; Huo, Y.; Ran, Y.; Tian, Q.; Wang, R.; Wang, H. Neural Radiance Field-based Visual Rendering: A Comprehensive Review. *arXiv* **2024**, arXiv:2404.00714. [CrossRef]
36. Campos, C.; Elvira, R.; Rodríguez, J.J.; Jose, M.M.M.; Juan, D.T. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
37. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
38. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.

39. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
40. Zhang, Y.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
41. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
42. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
43. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9157–9166.
44. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT++: Better Real-time Instance Segmentation. *IEEE Trans Pattern Anal. Mach. Intell.* **2019**, *1912*, 06218. [CrossRef]