agriculture

MDPI

Article

Reliable Visual Guidance for Robotic Safflower Harvesting Using Corolla–Bud Topology and Selective Dual-Camera Correction

Hui Guo [1,2,*], Dailin Zhu [1,2], Tianlun Wu [3], Yuande Dong [1,2], Guangyuan Zhang [1,2], Wenhao Bo [1,2] and Jianglin Liu [1,2]

[1] College of Mechanical and Electrical Engineering, Xinjiang Agricultural University, Urumqi 830052, China
[2] Xinjiang Intelligent Agricultural Machinery Equipment Engineering Technology Research Center, Urumqi 830052, China
[3] College of Mechanical and Electrical Engineering, Xinjiang Institute of Engineering, Urumqi 830052, China
* Correspondence: gh97026@xjau.edu.cn

Abstract

Field safflower harvesting is challenged by target-scale variation, occlusion, and random growth orientations, which can degrade visual perception and end-effector alignment. This study proposes a robotic visual-guidance framework integrating FAPB-Net segmentation, corolla–bud topology, and selective main-auxiliary dual-camera correction. Inclination is estimated from the fitted corolla ellipse, while in-plane orientation is obtained from the directed bud-to-corolla vector. Auxiliary inclination estimates are introduced only when three-dimensional target matching and angular consistency are satisfied. FAPB-Net achieved corolla and bud intersection over union (IOU) values of 90.28% and 70.85% at 28 FPS. On 138 synchronized dual-camera test pairs, inclination mean absolute error (MAE) decreased from 7.5° to 6.3°, while Acc@10°, defined as the proportion of estimates with absolute error below 10°, increased from 86.2% to 90.6%. In field tests with 515 targets, recognition/localization and end-effector arrival rates reached 95.92% and 94.95%, respectively. For 447 targets with complete reference measurements, inclination and orientation MAEs were 6.0° and 4.72°. Pose-guided harvesting achieved a mean harvesting completeness rate (HCR) of 94.63%, compared with 91.40% for position-only guidance. These results indicate the practical value of pose information for reliable safflower harvesting.

Keywords: safflower harvesting robot; semantic segmentation; corolla–bud topology; dual-camera vision; pose estimation

1. Introduction

Safflower (*Carthamus tinctorius* L.) is an important medicinal and economic crop whose filament harvesting still relies heavily on manual labor. With the development of machine vision and robotic harvesting, visual systems are increasingly required to provide not only target positions but also geometric information for end-effector alignment and autonomous execution [1–4].

A large number of studies have focused on crop target detection and picking point localization. Bai et al. [5] investigated clustered tomato detection and picking point localization. Li et al. [6] combined object detection and instance segmentation for mango picking point positioning. For occluded targets, Zhang et al. [7] localized grape picking points on partially visible fruit axes, and Li et al. [8] used point cloud processing for occluded apple localization. Active sensing has also been introduced to improve localization reliability under complex agricultural conditions [9]. Similar progress has been achieved for safflower. Lightweight detection methods have been developed for small and densely

Check for updates

Academic Editor: Wei Ji

Received: 20 September 2026
Revised: 30 October 2026
Accepted: 2 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

distributed safflower targets [10], and improved YOLO-based approaches have been used for filament recognition and picking-point localization [11]. Geometric analysis has been applied to determine filament-necking positions [12], while segmentation combined with structural information has been used to localize picking points throughout the harvesting period [13]. Depth information has also been incorporated for three-dimensional safflower localization [14]. These studies have substantially improved safflower target detection and spatial localization; however, their outputs are still dominated by positional information and provide limited guidance for end-effector orientation adjustment.

For robotic harvesting, target position alone may be insufficient when the end effector must approach the crop along an appropriate direction. Consequently, crop pose estimation has received increasing attention. Binocular vision and deep neural networks have been used for grape-cluster pose estimation [15], while point-cloud segmentation combined with geometric analysis has enabled in-field grape pose estimation [16]. For tomatoes, structural features have been exploited for two-dimensional pose estimation [17], and three-dimensional keypoints have been introduced for pose inference [18]. Anatomical keypoints combined with point clouds have also been used to estimate tomato peduncle-node pose and investigate viewpoint effects [19]. Subsequent studies developed keypoint-based six-dimensional pose estimation [20] and semantic-segmentation-based observation-pose estimation methods [21]. Similar RGB-D, instance-segmentation, and three-dimensional geometric approaches have been applied to citrus and tea bud leaves [22,23], while detection, keypoints, and instance segmentation have been integrated for selective fruit harvesting and six-dimensional pose estimation [24]. Pose information has also been directly incorporated into robotic execution, including fruit-growth-direction-based grasping [25], adaptive end-effector pose control [26], and coordinated perception and manipulation for selective harvesting [27]. These studies demonstrate that orientation information can serve as an important link between visual localization and end-effector alignment.

But direct application of existing crop pose-estimation strategies to safflower remains challenging because the structural cues available for pose reasoning differ from those of larger fruits. Previous safflower studies have reported difficulties associated with small and densely distributed targets, complex backgrounds, occlusion, and structural variation [10,13]. In the field images considered in this study, the safflower corolla presents an approximately planar and compact structure, whereas the associated bud occupies a relatively small image region. Variations in plant height and imaging distance further cause substantial changes in target scale. Under these conditions, the projected corolla shape can provide a geometric cue for inclination, but incomplete contours may destabilize ellipse fitting. Moreover, ellipse geometry alone cannot reliably distinguish the actual growth direction in the image plane. The relative spatial relationship between the bud and corolla can therefore provide an additional directional cue. Viewpoint-dependent pose-estimation performance has also been reported for plant structures [19]. Recent harvesting studies have shown that multi-camera observations can improve cutting-point perception by providing complementary views [28], while dual-view localization can alleviate the information limitations of a single viewpoint under occlusion [29]. In addition, the reliability of robotic visual guidance depends not only on localization within the camera coordinate system but also on accurate hand–eye pose acquisition for transferring visual information to the robot coordinate system [30]. Nevertheless, auxiliary observations may themselves become unreliable because of occlusion, segmentation errors, or incorrect correspondence; therefore, unconditional multi-view fusion does not necessarily improve pose-estimation reliability.

To address the limitations of position-oriented safflower perception and the uncertainty of pose estimation under field conditions, this study develops a visual-guidance framework that integrates semantic segmentation, corolla–bud topology, and selective dual-camera cor-

rection. The main contributions are summarized as follows: (1) a structural pose-estimation method is developed to estimate corolla inclination from projected ellipse geometry and determine directed in-plane orientation using the corolla–bud topological relationship; (2) a selective main-auxiliary camera correction strategy is introduced to improve inclination estimation through spatial correspondence and angular-consistency assessment; and (3) the proposed framework is evaluated through segmentation experiments, pose-estimation analysis, and field robotic harvesting tests to examine its accuracy and practical applicability.

2. Materials and Methods

The overall workflow of the proposed visual-guidance framework is illustrated in Figure 1. The system comprises three main stages: corolla and bud segmentation, topology-based pose estimation with selective dual-camera correction, and robotic guidance and harvesting evaluation.

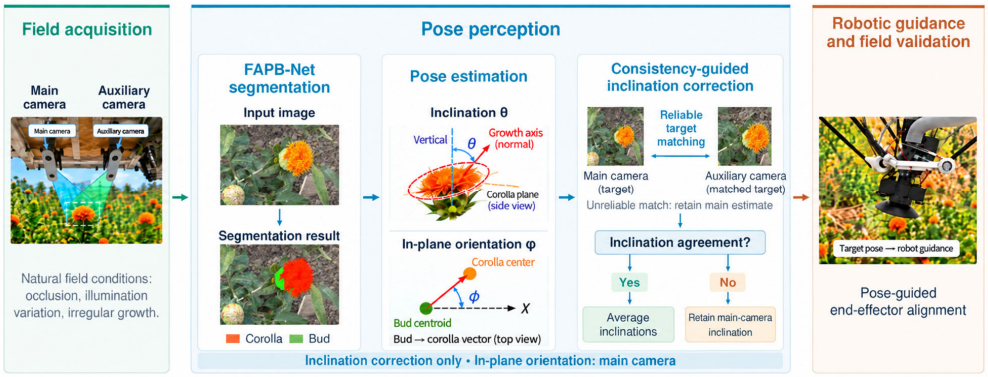


Figure 1. Overview of the proposed dual-camera framework for safflower pose estimation and robotic guidance.

2.1. Image Acquisition and Dataset Construction

2.1.1. Image Acquisition

The experiments were conducted in 2025 at Hongqi Farm, Sixth Division of the Xinjiang Production and Construction Corps, Changji Hui Autonomous Prefecture, Xinjiang, China. Safflower images were collected in July 2025 during the flowering stage. The safflower variety used in this study was Jihong No.3. The plants were cultivated with an approximate row spacing of 0.35 m and plant spacing of 0.15 m. A ZED2i stereo camera(Stereolabs, San Francisco, CA, USA) was used for image acquisition at an RGB resolution of 1280 × 720 pixels and a frame rate of 15 frames s^{-1}. The camera was fixed at a height of approximately 1.5 m and oriented downward to cover the same safflower row from a harvesting viewpoint. Images were acquired under representative field illumination conditions, including sunny and cloudy weather in both the morning and afternoon. In total, 2235 raw images were collected, examples of which are shown in Figure 2.

During dual-camera validation, two ZED2i cameras were rigidly mounted on the harvesting platform and synchronized to acquire RGB and depth data. The main camera served as the reference for robot coordinates and pose output, while the auxiliary camera was mounted side by side with the main camera. Both cameras were oriented vertically downward with approximately parallel optical axes, and the ground vertical direction was used as the common reference for inclination. The distance between the camera centers was approximately 220 mm, and the centers of their fields of view covered the same safflower row. Network inputs were uniformly cropped to 640 × 640 pixels. The camera arrangement is shown in Figure 3.



(**a**) (**b**) (**c**) (**d**)

Figure 2. Examples of the safflower image dataset under different field illumination conditions: (**a**) Sunny morning. (**b**) Sunny afternoon. (**c**) Cloudy morning. (**d**) Cloudy afternoon.

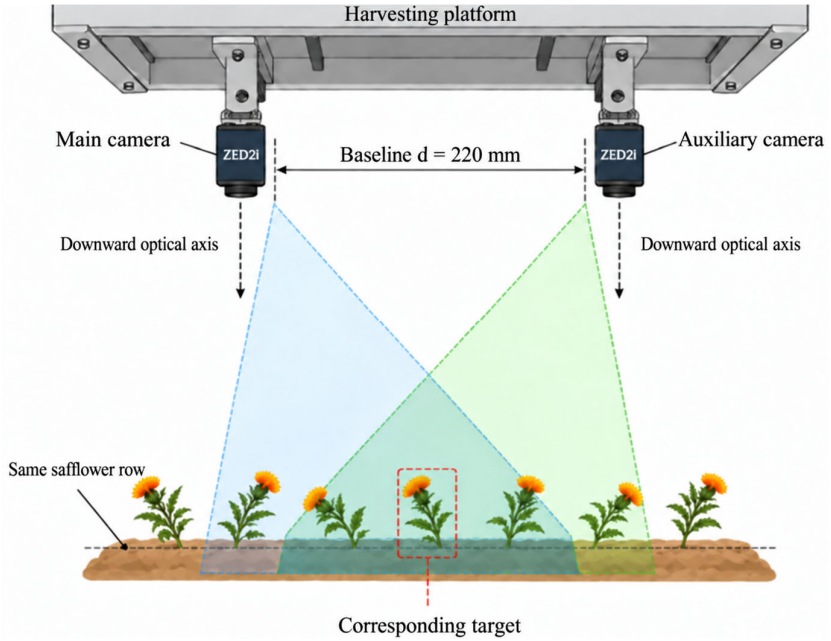


Figure 3. Schematic of the main-auxiliary dual-camera installation.

2.1.2. Pixel-Level Annotation and Dataset Split

After removing images with motion blur, severe overexposure, or unrecognizable target structures, 2034 valid images were retained. Pixel-level annotation was performed using Labelme, with each image divided into three classes: background, corolla, and bud. The corolla and its corresponding bud regions were used as structural features for subsequent pose estimation. An example of the pixel-level annotation is shown in Figure 4. The dataset was divided into training, validation, and test sets at a ratio of 8:1:1. Images from the same acquisition sequence or closely related plant region were assigned to only one subset to reduce scene-level information leakage. The training set was used for network optimization, the validation set was used for model selection and determination of geometric parameters such as the bud–corolla matching threshold, and the test set was reserved exclusively for final performance evaluation.

A total of 154 images containing 534 safflower instances were randomly selected from the predefined test set for pose-estimation evaluation. Following the image-level screening used for dataset construction, image quality was assessed separately for each target instance. An instance was classified as image-quality-invalid when motion blur or overexposure prevented reliable identification of the structures required for pose estimation. This assessment identified 43 image-quality-invalid instances. Among the remaining 491 image-quality-valid instances, 90 failed to produce complete pose estimates, including 20 instances with

incomplete corolla contours, 42 with missing structural information, and 28 with unreliable corolla–bud correspondence. A sequential exclusion strategy was adopted, whereby each target was excluded once a failure criterion was met, ensuring that the three failure categories were mutually exclusive. Consequently, 401 instances produced valid inclination and in-plane orientation estimates, corresponding to a valid-pose output rate of 75.09%. Angular-error metrics were calculated using these 401 valid outputs, whereas all 534 instances were included in the assessment of pose-output availability. During inference, connected-component analysis was performed on the corolla and bud masks to obtain the instance regions required for subsequent ellipse fitting and centroid calculation.

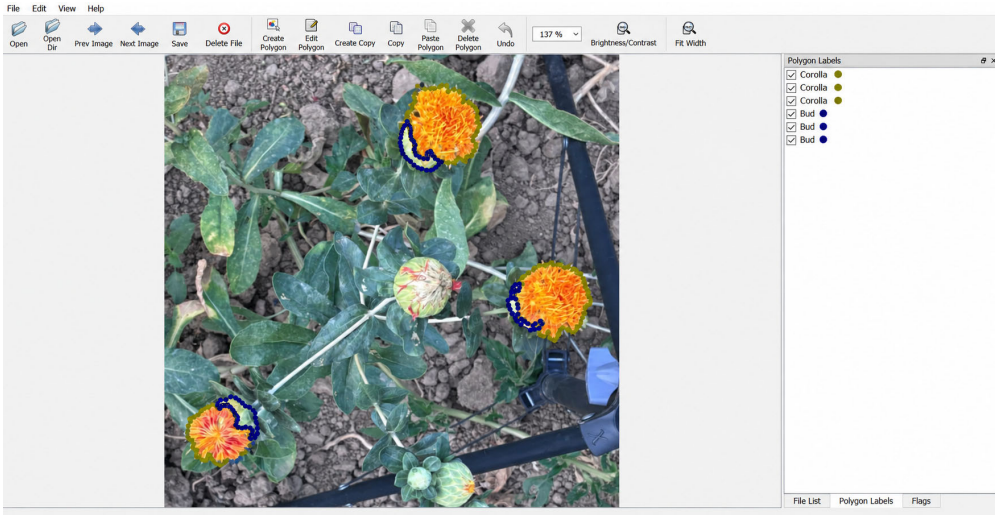


Figure 4. Example of pixel-level annotation for safflower images.

A separate synchronized dual-view dataset was constructed for cooperative correction. It comprised 193 validation target pairs and 138 independent test target pairs. The validation pairs were used only to select the spatial-matching and angular-consistency thresholds. The 138 test pairs were used as a common test set for main-camera-only estimation, direct main-auxiliary averaging, and the proposed consistency-guided correction. The validation and test subsets did not overlap.

2.1.3. Ground-Truth Pose Measurement

External physical measurements were used to obtain reference ground truth for the inclination and in-plane orientation angles. An H-D651G five-line laser level (SNDWAY, Dongguan, China) was used in the test area to project mutually perpendicular horizontal and vertical reference lines and establish a ground coordinate system, with the X-axis along the crop row, the Y-axis perpendicular to the row, and the Z-axis vertically upward.

During image acquisition, the main camera was rigidly mounted in a downward-looking configuration, with its optical axis approximately aligned with the ground vertical. The horizontal image axis was aligned with the X-axis of the ground reference system. For each safflower target, the center of the corolla top, $P_{top} = (x_t, y_t, z_t)$, and the basal point at the bud–corolla junction, $P_{base} = (x_b, y_b, z_b)$, were selected as the endpoints of the pose axis. Adhesive markers were used to assist localization, and plumb lines projected the two points onto the ground reference plane. Directed perpendicular distances from the projected points to the Y- and X-axes were used as the X and Y coordinates, respectively, and the vertical height relative to the reference plane was used as the Z coordinate. Distances were measured using an SNDWAY SW-100G laser distance meter (0.05–100 m measuring range; ±2 mm ranging accuracy). Each of the six coordinate components defining $P_{top}(x_t, y_t, z_t)$

and $P_{base}(x_b, y_b, z_b)$ was measured three times, and the coordinate means were used to calculate the reference pose. For each target, the sample standard deviation of the three readings was calculated separately for each coordinate component. The resulting pose vector was calculated as follows:

$$\Delta V = (x_t - x_b, y_t - y_b, z_t - z_b) \tag{1}$$

The ground-truth inclination and in-plane orientation angles were then calculated as follows:

$$\theta = \arccos\left(\frac{|z_t - z_b|}{\|\Delta V\|}\right) \tag{2}$$

$$\varphi = \text{atan2}(y_t - y_b, x_t - x_b) \tag{3}$$

Coordinate repeatability was propagated to reference-angle uncertainty through Equations (2) and (3). Let p contain the six coordinate means and Σ_p their covariance matrix. Assuming independent coordinate errors, Σ_p was diagonal, with each entry equal to the corresponding sample variance of the three readings divided by three. The resulting angular variances were

$$u_\theta^2 = J_\theta \Sigma_P J_\theta^T \tag{4}$$

$$u_\varphi^2 = J_\varphi \Sigma_P J_\varphi^T \tag{5}$$

where J_θ and J_φ are the 1 × 6 Jacobian row vectors of the reference inclination and in-plane orientation functions, respectively, evaluated at the measured coordinate means. Both angle functions were expressed in degrees when calculating the Jacobians. Therefore, the propagated standard uncertainties u_θ and u_φ are expressed in degrees. A geometric screening condition was applied when summarizing the first-order propagation results. Specifically, the horizontal length of the reference pose vector was required to exceed the root-sum-of-squares standard uncertainty of its two horizontal displacement components. These displacement uncertainties were calculated from the coordinate-mean uncertainties at the two endpoints under the independent-error assumption. Targets that did not satisfy this condition were flagged and reported separately. Coordinate measurement repeatability was evaluated for all 401 targets. Its first-order contributions to reference-angle uncertainty were evaluated for the 387 targets satisfying the horizontal-projection screening condition. Landmark placement, plumb-line alignment, and camera-axis alignment may introduce additional systematic contributions that are not captured by this repeatability analysis. The sensitivity to axis misalignment is discussed in Section 2.3.2.

The measured reference poses were used to calculate MAE and Acc@10°. Reference inclination was defined by the bud–corolla growth axis relative to the ground vertical, whereas ellipse-based inclination represented the corolla normal relative to the approximately vertical viewing axis. Their correspondence relies on the approximate alignment between the corolla normal and the growth axis. The reference-pose measurement principle is illustrated in Figure 5.

2.1.4. Camera Calibration and Coordinate Transformation

In the dual-camera system, the main camera served as the reference for final target coordinates and pose output, whereas the auxiliary camera was used only to assist inclination correction. Before the experiments, an 8 × 11-square checkerboard calibration board with a square size of 35 mm × 35 mm, corresponding to 10 × 7 inner corners, was imaged at different positions and orientations within the common field of view of both cameras. A total of 31 valid checkerboard image pairs were obtained. The root-mean-square (RMS) calibration errors of the main and auxiliary cameras were 0.0952 and 0.0896 pixels, respectively, and the RMS error of the joint stereo calibration was 0.1086 pixels.

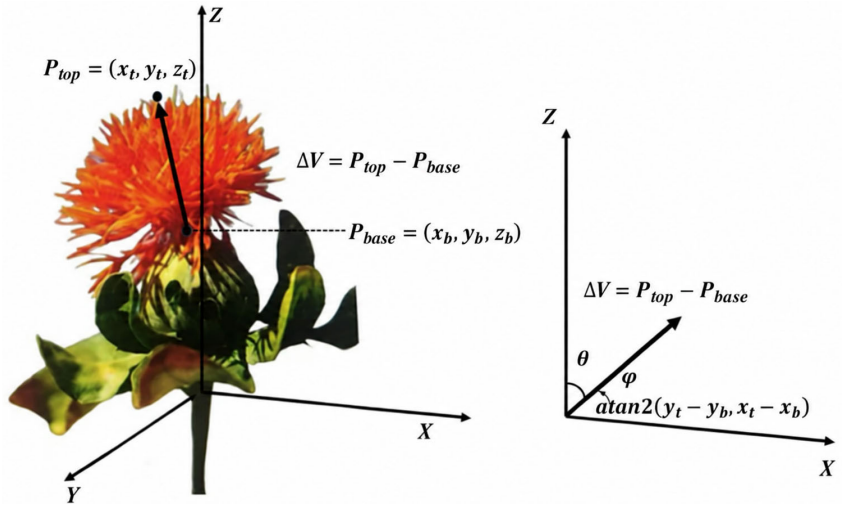


Figure 5. Schematic of pose ground-truth measurement.

A target point expressed in the auxiliary-camera coordinate system was transformed into the main-camera coordinate system as follows:

$$P_m = R_{am}P_a + t_{am} \tag{6}$$

where R_{am} and t_{am} are the rotation matrix and translation vector from the auxiliary-camera coordinate system to the main-camera coordinate system, respectively. After coordinate unification, targets observed by the two cameras could be spatially matched in the same coordinate frame, providing the basis for subsequent inclination-consistency correction.

2.1.5. Robot Coordinate Calibration

A calibration device carrying an ArUco marker was fixed to the robot end-effector. The manipulator was moved to different three-dimensional positions within the main-camera field of view. For each position, the robot controller recorded the end-effector reference point in the robot base coordinate system, while the main camera obtained the three-dimensional coordinates of the ArUco marker center in the camera coordinate system. The calibration procedure is shown in Figure 6, where Figure 6a illustrates the ArUco device installation and Figure 6b shows the acquisition of corresponding coordinate points. A total of 60 point pairs were collected. Let the ith corresponding point in the main-camera and robot-base coordinate systems be ${}^{C}P_i$ and ${}^{B}P_i$, respectively. The relationship between the two coordinate systems is expressed as

$${}^{B}P_i = R_{BC}{}^{C}P_i + t_{BC} \tag{7}$$

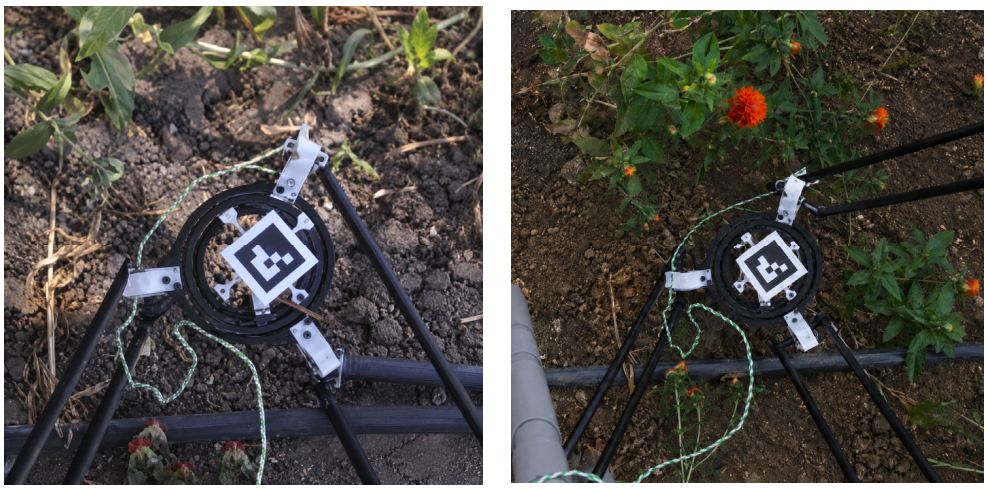
where R_{BC} and t_{BC} denote the rotation matrix and translation vector from the main-camera coordinate system to the robot-base coordinate system, respectively. The transformation parameters were estimated by least-squares rigid registration using all 60 point pairs. The mean Euclidean residual on these calibration points was 4.58 mm, representing the fitting residual rather than an independent assessment of positioning accuracy.

2.2. FAPB-Net Segmentation Network

2.2.1. Network Structure

To address large target-scale variation, complex background textures, and small bud regions in field safflower images, several commonly used segmentation methods were considered. YOLO-Seg combines object detection and instance segmentation but focuses more strongly on object localization and instance-level recognition. DeepLabv3 has strong

multi-scale feature extraction ability through atrous convolution and ASPP, but its relatively simplified decoder leaves room for improvement in small-object and fine-boundary recovery. SegFormer strengthens global-context modeling with a Transformer architecture, but its network structure and training procedure are relatively complex and place greater demands on data volume and computational resources. In contrast, U-Net has a simple and stable encoder–decoder architecture, and its skip connections effectively fuse shallow spatial details with deep semantic information, which is advantageous for preserving corolla contours and small bud regions and for integrating task-specific multi-scale and attention modules. Therefore, FAPB-Net was developed based on a U-Net encoder–decoder framework. ResNet34 was adopted as the encoder to enhance hierarchical feature extraction, while an ASPP module was introduced at the end of the encoder to aggregate multi-scale contextual information. A Parallel-CBAM module was embedded in the skip connections to refine the encoder features before decoder fusion. The overall network architecture is shown in Figure 7.



(**a**) (**b**)

Figure 6. Calibration between the main-camera and robot-base coordinate systems: (**a**) ArUco calibration device installation. (**b**) Corresponding-point acquisition.

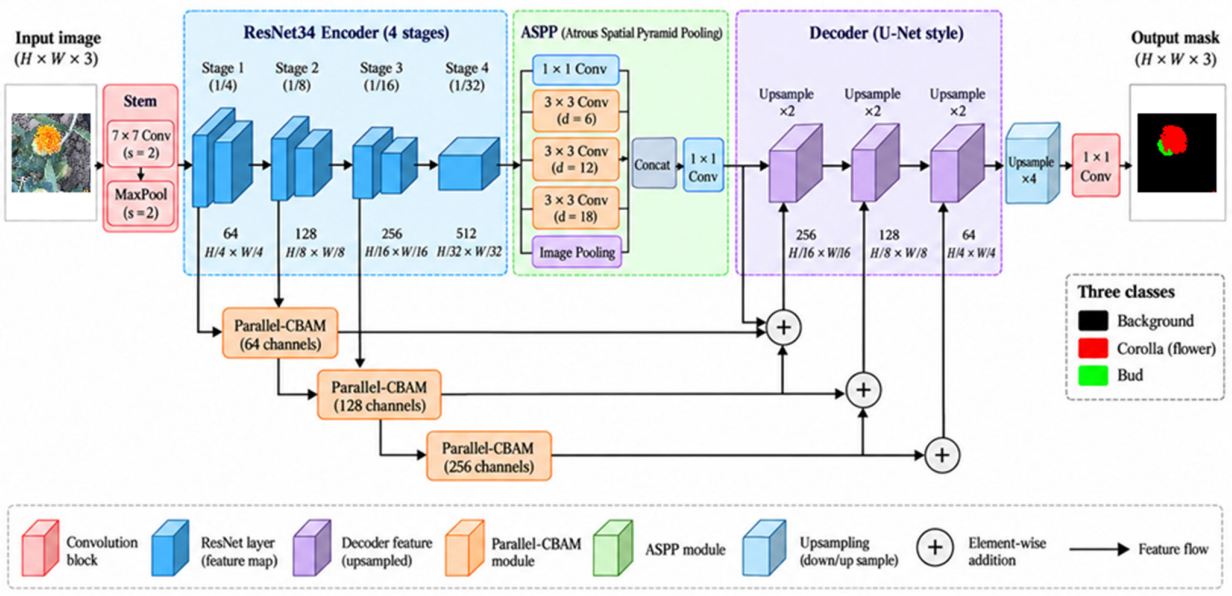


Figure 7. Overall architecture of FAPB-Net.

In Parallel-CBAM, channel and spatial attention are computed in parallel and fused by element-wise addition (Figure 8):

$$F' = (F \otimes M_c(F)) \oplus (F \otimes M_s(F)) \tag{8}$$

where F denotes the input feature map, and M_c and M_s are the channel- and spatial-attention weights, respectively. Channel and spatial attention are computed in parallel and their refined features are fused by element-wise addition. During decoding, the refined encoder features are channel-aligned with the upsampled decoder features and fused by addition before convolutional refinement. The final layer outputs pixel-wise probabilities for the background, corolla, and bud classes.

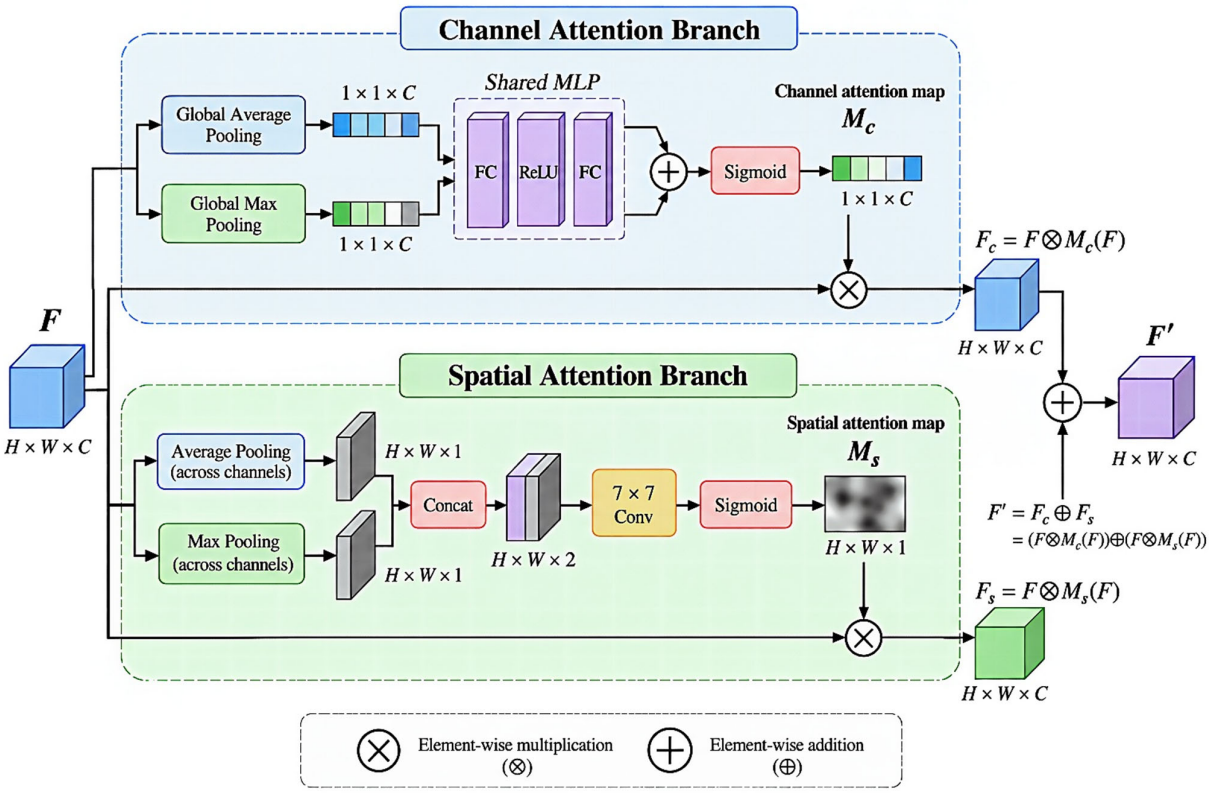


Figure 8. Structure of the Parallel CBAM module.

2.2.2. Loss Function

Because pose estimation depends on the geometric quality of corolla contours, the training objective combines class-weighted cross-entropy with a boundary-enhanced term:

$$L = L_{wCE} + \lambda L_{bd} \tag{9}$$

The class-weighted cross-entropy loss is defined as

$$L_{wCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c y_{i,c} \log p_{i,c} \tag{10}$$

Here, N denotes the number of pixels; $C = 3$ is the number of classes, namely background, corolla, and bud; c is the class index; and $y_{i,c}$ and $p_{i,c}$ denote the one-hot label and predicted probability, respectively. The class weight w_c was set inversely proportional to the pixel proportion of class c in the training set to alleviate class imbalance. The boundary loss is defined as

$$L_{bd} = -\frac{1}{N}\sum_{i=1}^{N} w_{bd}(i) \sum_{c=1}^{C} y_{i,c} \log p_{i,c} \tag{11}$$

The boundary weight map w_{bd} was obtained from the difference between morphological dilation and erosion of the ground-truth mask:

$$W_{bd} = Dilate(Y,r) - Erode(Y,r) \quad (12)$$

where Y denotes the ground-truth label and r is the radius of the morphological operation, which was set to 3 pixels in this study. By increasing the loss weight of boundary-region pixels, the network is encouraged to focus more on the corolla and bud contours during training, thereby reducing the influence of boundary misclassification on subsequent geometric feature extraction and pose estimation. λ denotes the weighting coefficient of the boundary term.

To determine an appropriate value of λ, a grid search was performed on the validation set while keeping all other training parameters unchanged. The segmentation performance under different λ values was compared, and the corresponding validation results are presented in Table 1.

Table 1. Effect of different boundary-loss weights on foreground segmentation performance on the validation set.

λ	Corolla IOU%	Bud IOU%	Mean Foreground IOU%
0	86.92	70.83	78.88
0.1	87.48	70.12	78.80
0.3	88.21	71.45	79.83
0.5	88.70	73.69	81.20
0.7	88.39	72.33	80.36
1.0	87.81	72.67	80.24

The highest mean foreground IOU of 81.20% was obtained at $\lambda = 0.5$; therefore, this value was used in subsequent experiments. Table 1 reports validation-set results used only for hyperparameter selection.

2.2.3. Training Configuration

Training followed Table 2, with random rotation, scaling, brightness perturbation, and mild Gaussian noise applied only to the training set. The original dataset split was unchanged; unaugmented validation and test sets were used for model selection and independent evaluation, respectively.

Table 2. Network Training Configuration.

Parameter	Value
GPU	NVIDIA GeForce RTX 3090 (24 GB)
Optimizer	Adam
CPU	Intel Xeon Silver 4310 @ 2.10 GHz
Operating system	Windows 11
Environment	CUDA 11.7; cuDNN 8.5.0
Learning rate	0.0001
Batch size	8
Training epochs	200
Number of workers	8
Framework	PyTorch 2.0.1

FAPB-Net used a ResNet34 encoder. The DeepLabV3 and FCN baselines used ResNet50 backbones, while U-Net and U-Net++ used convolutional encoder–decoder architectures.

All compared models were configured with three output classes corresponding to the background, corolla, and bud. All compared models were randomly initialized and trained without loading pretrained weights.

After training, images from both the main and auxiliary cameras were processed using the same trained FAPB-Net weights, thereby ensuring a consistent corolla and bud segmentation procedure across the two views.

2.3. Safflower Pose Estimation with Corolla–Bud Topology and Dual-Camera Correction

2.3.1. Structural Feature Extraction

Images acquired by the main and auxiliary cameras were separately fed into the trained FAPB-Net to obtain pixel-level predictions for the background, corolla, and bud classes. During inference, the class with the highest predicted probability at each pixel was selected to generate a discrete label map, from which the corolla and bud regions were extracted. For each corolla region, connected-component filtering and contour tracing were performed to obtain the boundary point set, followed by ellipse fitting for subsequent geometric analysis. For each bud region, the centroid coordinates were calculated from the segmented region. The overall procedure for semantic segmentation and geometric feature extraction is illustrated in Figure 9.

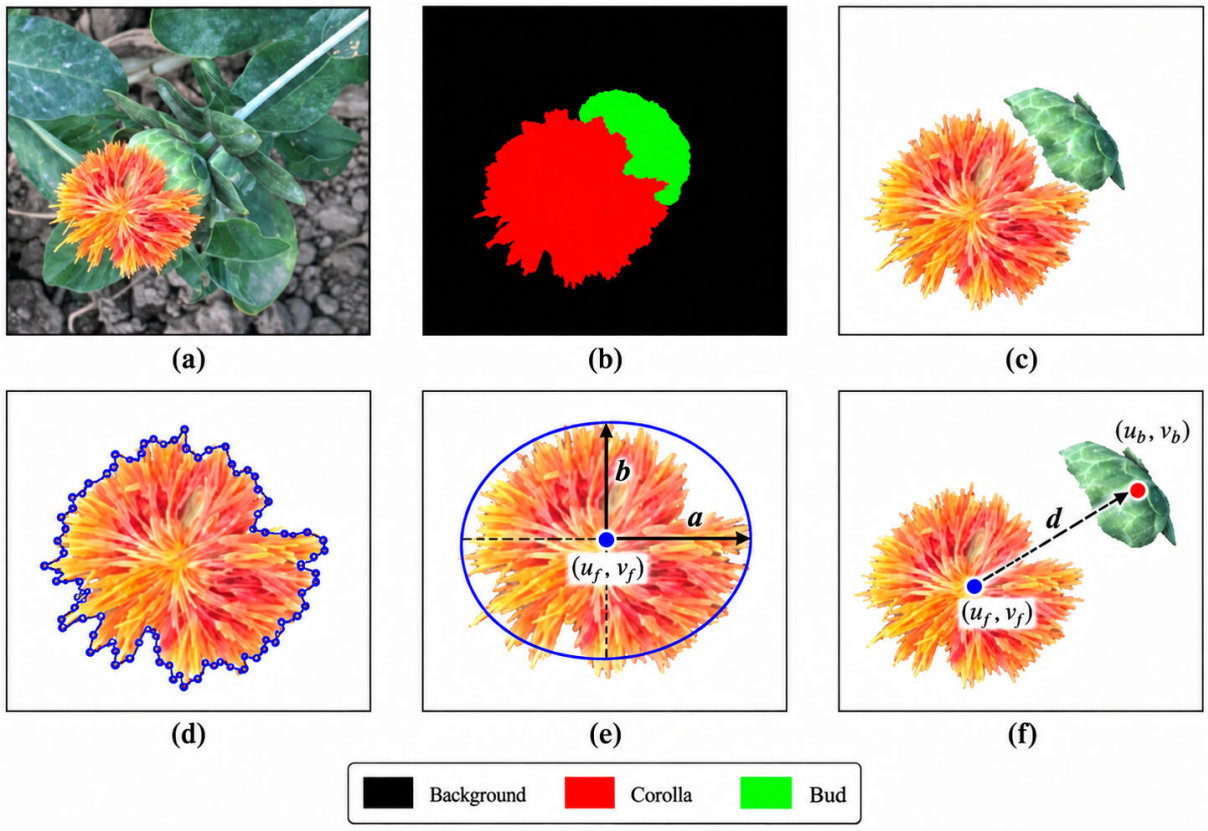


Figure 9. Workflow of semantic segmentation and geometric feature extraction for safflower images. (**a**) Original image. (**b**) Semantic segmentation. (**c**) Corolla and bud extraction. (**d**) Corolla contour extraction. (**e**) Corolla ellipse fitting. (**f**) Corolla and bud centroid extraction.

For each corolla instance, the contour point set $\mathcal{D}_f$ was represented by N boundary points (u_i, v_i). A least-squares ellipse was fitted to estimate the ellipse center (u_f, v_f) and its major and minor axis lengths a and b. For each bud instance, region moments were used to

determine a robust centroid. Let m_{00} denote the zeroth-order moment and m_{10} and m_{01} the first-order moments. The bud centroid (u_b,v_b) is calculated as

$$u_b = \frac{m_{10}}{m_{00}}, v_b = \frac{m_{01}}{m_{00}} \tag{13}$$

2.3.2. Main-Camera Inclination Estimation

The corolla contour was extracted from the FAPB-Net segmentation result and used for ellipse fitting. Inclination estimation assumes an approximately planar and circular corolla under weak-perspective projection. Under these conditions, the ellipse axis ratio provides an approximation of the corolla inclination. However, irregular petal shapes, incomplete contours, and deviations between the corolla normal and growth axis may introduce estimation errors. Consequently, the method is most applicable to targets with sufficiently complete and regular corolla contours. The geometric relationship is illustrated in Figure 10.

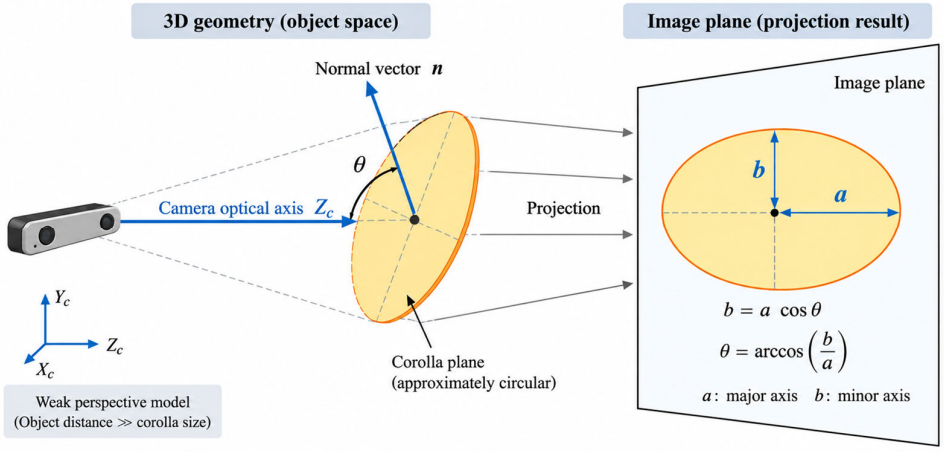


Figure 10. Geometric relationship between corolla inclination and its projected ellipse.

Assuming that the actual corolla plane is approximately circular, its projection after tilting can be represented by an ellipse with major axis a and minor axis b. When the inclination angle is θ, the dimension along the major axis remains approximately unchanged, whereas that along the minor axis is shortened by a factor of $\cos\theta$. Therefore,

$$\theta = arccos\left(\frac{b}{a}\right) \tag{14}$$

where $a \geq b > 0$, and a and b denote the major and minor axis lengths of the fitted ellipse, respectively. The inclination angle θ is defined as the angle between the corolla normal and the camera optical axis. Letting the axis ratio be $r = b/a$, the inclination angle can be expressed as $\theta = arccos(r)$. The first-order inclination change caused by an axis-ratio perturbation is expressed in degrees in Equation (15):

$$|\Delta\theta|_{deg} \approx \frac{180}{\pi} \frac{|\Delta r|}{\sqrt{1-r^2}} \tag{15}$$

For an absolute axis-ratio perturbation of 0.001, the first-order inclination changes are approximately 0.33°, 0.11°, and 0.07° at inclinations of 10°, 30°, and 60°, respectively. Axis-ratio deviations caused by corolla shape or segmentation therefore have a greater angular effect near the upright configuration. If the corolla normal and the reference growth axis differ by δ, the difference between their vertical inclination angles is bounded by δ. A camera-axis installation error of β similarly contributes up to β to the inclination

discrepancy. In-plane orientation is more sensitive to centroid displacement when the projected bud-to-corolla separation is small.

Partially occluded corollas can be processed when sufficient boundary information is available for ellipse fitting. At inclinations approaching 90°, the projected corolla becomes nearly edge-on, reducing the stability of ellipse fitting. Targets with severely incomplete contours or unsuccessful ellipse fitting were classified as invalid pose outputs and excluded from subsequent inclination correction. These cases were included in the pose-output availability assessment in Section 3.2.

2.3.3. Main-Camera In-Plane Orientation Estimation

In addition to the axis ratio used for inclination estimation, the fitted ellipse also provides an image-plane orientation cue. Under the approximately circular-corolla and weak-perspective assumptions, the minor axis of the projected ellipse approximately corresponds to the image-plane projection of the corolla normal and was therefore regarded as a geometric growth-axis cue. However, an ellipse axis is intrinsically undirected; the two directions separated by 180° represent the same geometric axis. Consequently, ellipse geometry alone can provide an orientation axis but cannot uniquely determine the directed growth orientation over 0°−360°.

Although the minor axis of the fitted corolla ellipse provides a geometric orientation axis, it cannot distinguish the two opposite directions along that axis because of its intrinsic 180° ambiguity. To obtain a directed in-plane growth orientation, the structural relationship between the corolla and its associated bud was therefore introduced. After obtaining the corolla ellipse parameters from the main-camera image, corolla–bud correspondence was established for each valid target. Let $C = \{C_1, C_2, \ldots\ldots, C_{N_c}\}$ denote the set of detected corolla centers and $B = \{B_1, B_2, \ldots\ldots, B_{N_b}\}$ denote the set of bud centroids. A Euclidean distance matrix $D \in \mathbb{R}^{N_c \times N_b}$ was constructed, where each element D_{ij} represents the distance between the i-th corolla center $C_i(x_{ci}, y_{ci})$ and the j-th bud centroid $B_j(x_{bj}, y_{bj})$:

$$D_{ij} = \sqrt{\left(x_{ci} - x_{bj}\right)^2 + \left(y_{ci} - y_{bj}\right)^2} \tag{16}$$

Corolla–bud association used distance-constrained greedy matching with an adaptive threshold $T_{cb} = ka$, where a is the major-axis length of the corresponding corolla ellipse and k is a dimensionless coefficient. Candidate pairs within this threshold were ranked by Euclidean distance. The shortest eligible pair was assigned first, and its corolla and bud were removed from further assignment to enforce one-to-one correspondence. This process continued until no eligible pairs remained. Unassigned corollas were excluded from orientation estimation. To determine k, values of 1.0, 1.5, 2.0, and 2.5 were evaluated on valid pose instances in the validation set using matching output rate, mismatch rate, and orientation Acc@10° (Table 3).

Table 3. Bud–corolla matching performance under different matching thresholds on the validation set.

k	Matching Output Rate (%)	Mismatch Rate (%)	Orientation Acc@10° (%)
1.0	78.27	2.14	91.24
1.5	89.70	3.83	94.51
2.0	93.24	8.92	88.32
2.5	95.77	16.38	79.15

The matching output rate was defined as the number of corollas assigned a bud divided by the total number of corollas evaluated for association. The mismatch rate was

defined as the number of incorrect assignments divided by the total number of assigned pairs. Orientation Acc@10° was calculated on all assigned pairs with valid orientation estimates and reference angles, including incorrect assignments; unmatched cases were excluded. At $k = 1.5$, the matching output rate was 89.70%, the mismatch rate was 3.83%, and orientation Acc@10° was highest at 94.51%, as shown in Table 3. This threshold was adopted for subsequent corolla–bud association. Instances with unreliable correspondences were included in the invalid pose-output statistics reported in Section 3.2.

After establishing a valid corolla–bud correspondence, the in-plane growth orientation of the safflower was represented by the directed vector from the bud centroid to the corolla center. Let $B_j = (x_{bj}, y_{bj})$ and $C_i = (x_{ci}, y_{ci})$ denote the centroid of the j-th bud and the center of the i-th corolla, respectively. Because the image coordinate system has its positive y-axis pointing downward, the vertical component was sign-reversed to conform to the Cartesian coordinate convention. The direction vector was therefore defined as

$$v_{ij} = (x_{ci} - x_{bj}, y_{bj} - y_{ci}) \tag{17}$$

The initial in-plane orientation angle was calculated using the four-quadrant arctangent function:

$$\varphi_{geo} = atan2\left(y_{bj} - y_{ci}, x_{ci} - x_{bj}\right)\frac{180}{\pi} \tag{18}$$

The angle was then normalized to the range $[0°, 360°)$:

$$\varphi_m = mod(\varphi_{geo} + 360°, 360°) \tag{19}$$

Accordingly, the positive horizontal direction was defined as 0°, and the angle increased counterclockwise, with 90°, 180°, and 270° corresponding to the upward, leftward, and downward directions, respectively. This directed corolla–bud vector resolves the inherent 180° directional ambiguity of the ellipse minor axis. Because the auxiliary camera has a different installation position, its image-plane orientation representation may not be fully consistent with that of the main camera. Therefore, the in-plane orientation was not fused across the two views, and the main-camera orientation φ_m was used as the directional guidance parameter for the robot end-effector.

2.3.4. Cross-View Target Matching

Because the main and auxiliary cameras are spatially separated, the same safflower appears at different pixel locations in the two images. To establish cross-view correspondence, ZED2i depth information was used to reconstruct target coordinates in three dimensions. The external calibration parameters obtained in Section 2.1.4 were then used to transform auxiliary-camera targets into the main-camera coordinate system for spatial matching.

For a target detected by the main camera, let (u_m, v_m) be the corolla-center pixel coordinates, Z_m the corresponding depth, and $(f_{xm}, f_{ym}, c_{xm}, c_{ym})$ the intrinsic parameters of the main camera. Its three-dimensional point $P_m = (X_m, Y_m, Z_m)$ in the main-camera coordinate system is calculated as

$$X_m = \frac{(u_m - c_{xm})Z_m}{f_{xm}} \tag{20}$$

$$Y_m = \frac{\left(v_m - c_{ym}\right)Z_m}{f_{ym}} \tag{21}$$

$$P_m = (X_m, Y_m, Z_m) \tag{22}$$

For an auxiliary-camera target, the three-dimensional point $P_a = (X_a, Y_a, Z_a)$ was calculated from its pixel coordinates, depth, and intrinsic parameters in the same manner, and then transformed into the main-camera coordinate system using Equation (6).

For each main-camera target, the nearest candidate in the transformed auxiliary-camera target set was searched using the Euclidean distance:

$$D_i = \|P_m - P_{a,i}^m\| \tag{23}$$

The spatial matching threshold was determined using 193 manually verified synchronized target pairs. At thresholds of 0.03, 0.05, and 0.07 m, the correct matching rates were 85.49%, 93.78%, and 94.30%, while the mismatch rates were 1.04%, 2.59%, and 5.18%, respectively. Increasing the threshold from 0.05 to 0.07 m improved the correct matching rate by only 0.52 percentage points but doubled the mismatch rate. Since incorrect cross-view associations may introduce erroneous inclination corrections, 0.05 m was selected to prioritize matching reliability.

If the nearest distance d_1 was smaller than the matching threshold $T_{3D} = 0.05$ m and the difference between the second-nearest and nearest distances satisfied $d_2 - d_1 > 0.5T_{3D}$, the two observations were considered to correspond to the same physical safflower. Otherwise, the cross-view match was regarded as unreliable, and the auxiliary-camera result was not used for inclination correction.

2.3.5. Consistency-Guided Inclination Correction

After cross-view target matching, inclination was independently estimated from each camera using the method in Section 2.3.2. The angular-consistency threshold T_1 was selected by comparing 3°, 5°, and 7° on the 193-pair validation set. At $T_1 = 5°$, the correction trigger rate was 63.21%, the inclination MAE was lowest at 6.5°, and Acc@10° reached 90.16% (Table 4). Accordingly, the final inclination was calculated as the mean of the two estimates when both the spatial-matching criteria and the angular-difference criterion $|\theta_m - \theta_a| \leq 5°$ were satisfied. Otherwise, the main-camera inclination was retained. In-plane orientation was always obtained from the main camera.

Table 4. Inclination correction under different angular-consistency thresholds.

Threshold (°)	Trigger Rate (%)	MAE (°)	Acc@10° (%)
3	43.01	6.9	89.12
5	63.21	6.5	90.16
7	79.79	6.7	89.64

2.4. Implementation and Evaluation Metrics

Model training and testing were conducted under the experimental environment summarized in Table 2. To evaluate segmentation and pose estimation performance, pixel-level segmentation metrics, angular-error metrics, and field-operation performance metrics were employed.

Segmentation performance was evaluated using intersection over union (IOU), Dice coefficient, and recall. Pose-estimation performance was evaluated using the mean absolute error (MAE) of the inclination and in-plane orientation angles and Acc@10°, defined as the percentage of samples with an absolute angular error below 10°. All orientation angles followed the Cartesian convention defined in Section 2.3.3. For the geometric baselines, each extracted axis angle was reduced modulo 180° to $[0°, 180°)$, and this canonical angle was used directly as the orientation prediction without further direction disambiguation. The proposed method retained its directed bud-to-corolla orientation in $[0°, 360°)$. Reference orientations were also normalized to $[0°, 360°)$ and were used only for error evaluation. For all four methods, orientation errors were calculated using the minimum circular angular difference with a periodicity of 360°. For Table 3, orientation Acc@10° followed the association evaluation defined in Section 2.3.3. For each angle, the number of valid estimates with

errors below 10° was also reported relative to all evaluated targets. Recognition/localization and arrival rates used all presented field targets. Single-target operation time extended from the start of manipulator motion to completion of cutting and suction. Angular errors required valid outputs and complete reference measurements.

To further quantify the practical contribution of pose estimation to safflower harvesting, a harvesting-completeness experiment was conducted using two robotic guidance strategies. In the position-only strategy, the same target recognition, three-dimensional localization, and coordinate-transformation procedures were retained, while the end-effector orientation was maintained at a predefined fixed harvesting orientation. In the proposed pose-guided strategy, both the target position and the estimated inclination and in-plane orientation were used to adjust the end-effector pose. Among the 515 field targets, 50 were selected for the harvesting-completeness comparison and allocated to the position-only and pose-guided strategies, with 25 targets per group. Target inclination, in-plane orientation, and maturity were considered during allocation to obtain comparable target conditions. The experimental scale reflected the requirement to collect and weigh harvested and residual filaments separately for each target. Sample-size sensitivity was evaluated under a two-sided independent-samples *t*-test framework.

After each harvesting operation, the safflower filaments removed by the robotic end effector and the residual harvestable filaments remaining on the same flower head were collected separately and weighed using the same electronic balance. For the *i*-th target, the harvesting completeness rate (HCR) was calculated as

$$HCR_i = \frac{m_{h,i}}{m_{h,i} + m_{r,i}} \times 100\% \quad (24)$$

where $m_{h,i}$ is the mass of filaments harvested by the robot and $m_{r,i}$ is the mass of residual harvestable filaments manually collected from the same flower head after robotic harvesting. The mean HCR of all targets in each strategy was used to characterize harvesting completeness. HCR values are reported as mean ± standard deviation (SD). Differences between the position-only and pose-guided strategies were evaluated using an independent-samples *t*-test, with $p < 0.05$ considered statistically significant.

3. Results

3.1. Semantic Segmentation Performance

3.1.1. Comparison with Different Segmentation Models

FAPB-Net was compared with U-Net, U-Net++, DeepLabV3, FCN, and SegFormer-B2 using the same training, validation, and test splits. All models used the same input resolution, data augmentation strategy, and optimization settings described in Section 2.2.3 and Table 2. Each model was configured with three output classes corresponding to background, corolla, and bud, and segmentation performance was evaluated using a common evaluation protocol. The comparative results for segmentation accuracy, model size, and inference speed are summarized in Table 5.

FAPB-Net achieved corolla and bud IOUs of 90.28% and 70.85%, respectively, with corresponding recall values of 97.05% and 88.70% and Dice coefficients of 94.89% and 82.94%. The model contained 32 M parameters and achieved a segmentation inference speed of 28 FPS on an NVIDIA GeForce RTX 3090 GPU (24 GB). The reported FPS measures network inference only, excluding image preprocessing, segmentation postprocessing, and subsequent pose estimation. Compared with SegFormer-B2, FAPB-Net increased the corolla and bud IOUs by 0.60 and 0.73 percentage points, respectively. Qualitative segmentation results under different illumination conditions are shown in Figure 11.

Table 5. Comparison of different semantic segmentation models.

Model	Corolla IOU (%)	Corolla Recall (%)	Corolla Dice (%)	Bud IOU (%)	Bud Recall (%)	Bud Dice (%)	Params (M)	FPS
U-Net	86.58	95.00	92.81	67.35	83.00	80.49	31	42
U-Net++	88.08	95.87	93.67	68.69	85.63	81.44	36	31
DeepLabV3	88.32	96.05	93.80	68.91	86.20	81.59	39	28
FCN	87.56	95.58	93.37	68.17	84.86	81.07	33	34
SegFormer-B2	89.68	96.60	94.56	70.12	87.85	82.44	25	26
FAPB-Net	90.28	97.05	94.89	70.85	88.70	82.94	32	28

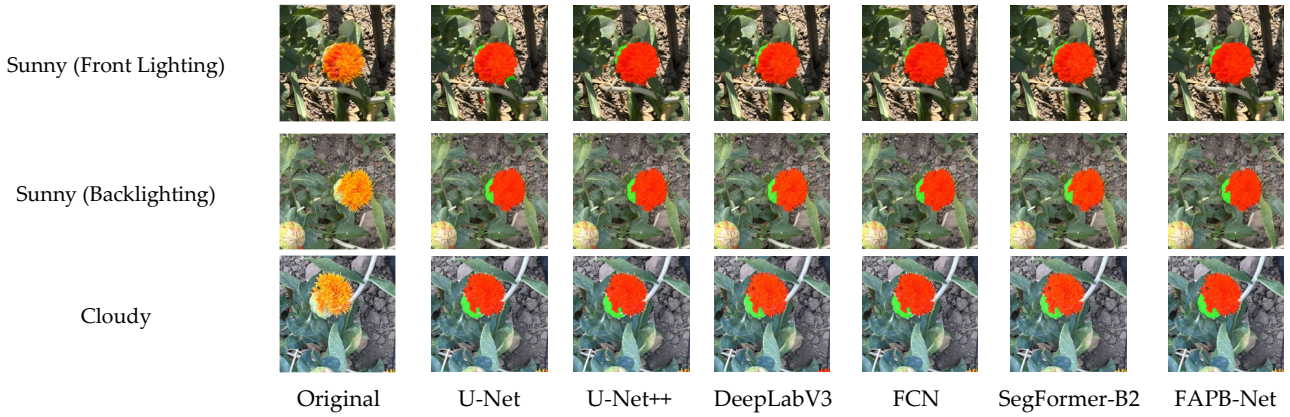


Figure 11. Qualitative comparison of segmentation results under different field conditions.

3.1.2. Ablation Study of Key Components

U-Net was used as the baseline to evaluate the contributions of the main components of FAPB-Net. Representative configurations involving the encoder architecture, attention mechanism, ASPP-based multi-scale feature aggregation, and boundary-enhanced supervision were compared under the same dataset split, training configuration, and evaluation protocol. A standard sequential CBAM was also included for comparison with Parallel-CBAM. The results are presented in Table 6.

Table 6. Ablation results of key components in FAPB-Net.

Configuration	Corolla IOU (%)	Corolla Recall (%)	Corolla Dice (%)	Bud IOU (%)	Bud Recall (%)	Bud Dice (%)
U-Net	86.58	95.00	92.81	67.35	83.00	80.49
U-Net + CBAM	87.03	95.29	93.07	68.15	84.63	81.06
U-Net + Parallel-CBAM	87.56	95.58	93.37	68.37	84.86	81.21
U-Net + ASPP	87.86	95.81	93.54	68.49	85.29	81.30
U-Net + Parallel-CBAM + ASPP	88.35	96.12	93.81	68.97	86.25	81.64
ResNet34-U-Net + Parallel-CBAM + ASPP	89.23	96.62	94.35	69.42	87.37	81.95
FAPB-Net	90.28	97.05	94.89	70.85	88.70	82.94

Replacing the standard CBAM with Parallel-CBAM increased the corolla IOU from 87.03% to 87.56% and the bud IOU from 68.15% to 68.37%. Combining Parallel-CBAM with ASPP further increased the corresponding IOUs to 88.35% and 68.97%. After the ResNet34 encoder was introduced, the configuration without boundary-enhanced supervision achieved corolla and bud IOUs of 89.23% and 69.42%, respectively. With boundary-

enhanced supervision, the complete FAPB-Net achieved the highest corolla and bud IOUs of 90.28% and 70.85%, respectively.

3.2. Single-View Safflower Pose Estimation Performance

The 154-image evaluation contained 534 targets. Forty-three were image-quality-invalid; of the remaining 491, 90 produced no valid pose: 20 had incomplete corolla contours, 42 lacked structural information, and 28 had unreliable corolla–bud correspondence. The 401 valid outputs gave an output rate of 75.09%. All 133 invalid targets, representing 24.91%, were included in availability assessment but excluded from angular-error calculations. Table 7 summarizes the mean within-target coordinate SDs from three readings per component; the overall mean coordinate SD was 1.00 mm.

Table 7. Coordinate measurement repeatability for 401 safflower targets.

Measurement Point	X-Axis SD (mm)	Y-Axis SD (mm)	Z-Axis SD (mm)
Corolla top	0.84	0.91	0.72
Bud base	1.26	1.33	0.95

Under the independent-coordinate-error assumption, the repeatability contributions to reference-angle uncertainty were calculated by first-order propagation using the coordinate means. Fourteen targets did not satisfy the horizontal-projection screening condition described in Section 2.1.3 and were excluded from this uncertainty summary. For the remaining 387 targets, the mean standard uncertainty contributions were 1.14° for inclination and 3.01° for in-plane orientation, with corresponding medians of 1.07° and 2.17° and 95th percentiles of 1.87° and 7.03°, respectively (Table 8).

Table 8. First-order contributions of coordinate measurement repeatability to reference-angle uncertainty.

Reference Angle	N	Mean (°)	Median (°)	95th Percentile (°)
Inclination	387	1.14	1.07	1.87
In-plane orientation	387	3.01	2.17	7.03

For the 401 valid pose instances, the proposed method achieved an inclination MAE of 7.38° with an Acc@10° of 87.28%. The overall in-plane orientation MAE was 4.65°, with an Acc@10° of 94.76%. Among all 534 evaluated targets, 350 produced valid pose outputs with inclination errors below 10°, and 380 produced valid pose outputs with in-plane orientation errors below 10°.

The minimum bounding rectangle (MBR), principal component analysis (PCA), fitted-ellipse, and proposed corolla–bud representations were compared on the same set of 401 valid pose instances. The geometric baselines used their canonical axis angles directly as orientation predictions, while the proposed method obtained directed orientation from the bud-to-corolla vector. All orientation predictions were evaluated against the full-range reference orientations using the same 360° circular-error criterion. The comparative results are summarized in Table 9.

To examine performance across different directional ranges, the 401 valid pose instances were divided into two subsets according to their reference in-plane orientations. The corresponding results are shown in Table 10.

Table 9. Comparison of geometric axis and topology-based pose representations for safflower.

Method	Orientation Representation	Incl. MAE (°)	Incl. Acc@10° (%)	Orient. MAE (°)	Orient. Acc@10° (%)
Minimum bounding rectangle (MBR)	Geometric axis	11.32	72.32	13.78	62.09
PCA principal axis	Geometric axis	9.74	80.55	10.24	74.31
Ellipse minor axis (without bud topology)	Undirected geometric axis	7.38	87.28	9.36	80.30
Proposed method	Directed corolla–bud vector	7.38	87.28	4.65	94.76

Table 10. Orientation-estimation performance of the proposed method over different directional ranges.

Orientation Range	Samples	Acc@10° (%)	Orient. MAE (°)
0°−180°	249	95.18	4.48
180°−360°	152	94.08	4.92
0°−360°	401	94.76	4.65

For the first directional subset containing 249 samples, the orientation MAE and Acc@10° were 4.48° and 95.18%, respectively. For the second subset containing 152 samples, the corresponding values were 4.92° and 94.08%. Across all 401 valid pose instances, the overall orientation MAE was 4.65° and the Acc@10° was 94.76%.

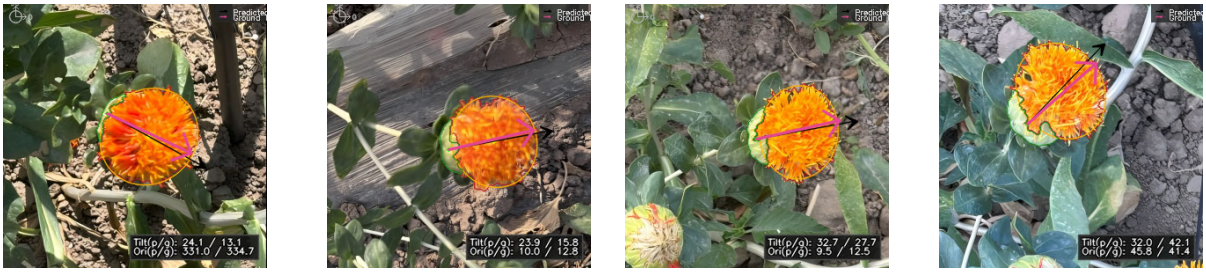
The stratified inclination errors are summarized in Table 11. Acc@10° was higher in the intermediate inclination ranges, whereas nearly circular targets at small inclinations and heavily occluded targets at large inclinations showed larger errors. These behaviors were mainly associated with limited sensitivity of the ellipse axis ratio near unity and contour incompleteness at large inclinations, respectively.

Table 11. Stratified inclination-estimation errors.

Range (°)	Samples	MAE (°)	Acc@10° (%)
<20°	102	8.45	79.41
20°−40°	158	6.98	91.77
40°−60°	98	6.85	93.88
>60°	43	7.50	74.42
Total	401	7.38	87.28

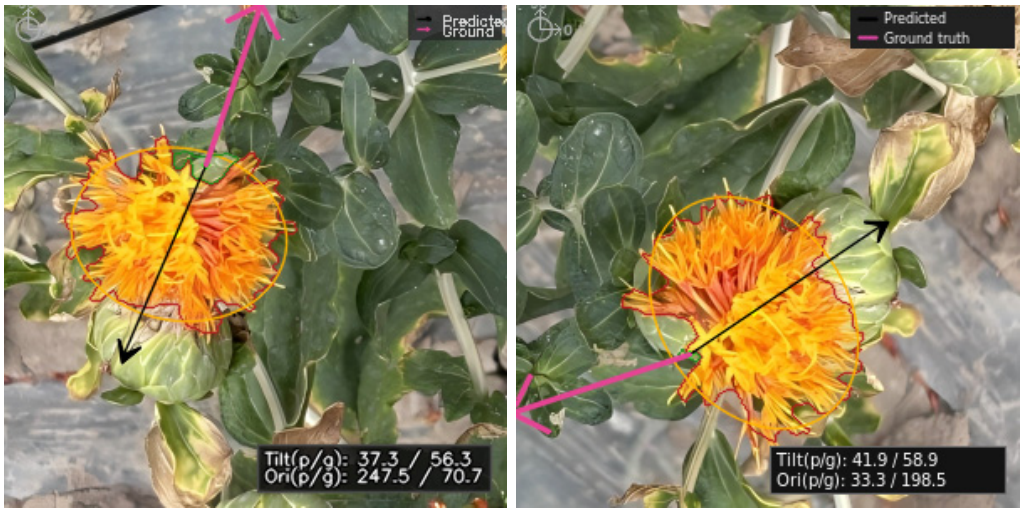
Representative pose-estimation results are shown in Figure 12. The figure displays the fitted corolla ellipse, corolla centroid, bud centroid, predicted in-plane orientation, and ground-truth direction. Black arrows indicate predictions, magenta arrows indicate ground truth, green points indicate bud centroids, and yellow points indicate corolla centroids. Predicted and ground-truth inclination and orientation values are also shown. When the corolla contour is complete and the bud is visible, the method provides stable estimates of both inclination and in-plane orientation.

Figure 13 shows pose-estimation failures. In Figure 13a, the target bud was missed and a neighboring smaller bud was incorrectly associated with the corolla. Figure 13b illustrates incorrect correspondence among adjacent targets. These structural and association errors can prevent pose output or produce erroneous orientation estimates.



(**a**) (**b**) (**c**) (**d**)

Figure 12. Representative safflower pose-estimation results: (**a**) Sunny morning. (**b**) Sunny afternoon. (**c**) Cloudy morning. (**d**) Cloudy afternoon.



(**a**) (**b**)

Figure 13. Representative failure cases of safflower pose estimation: (**a**) omission of the target bud leading to incorrect association with a neighboring bud; (**b**) incorrect corolla–bud correspondence among adjacent targets.

Figure 14 presents the cumulative error distributions of inclination and in-plane orientation. At an absolute-error threshold of 10°, the cumulative proportions were 87.28% and 94.76%, respectively.

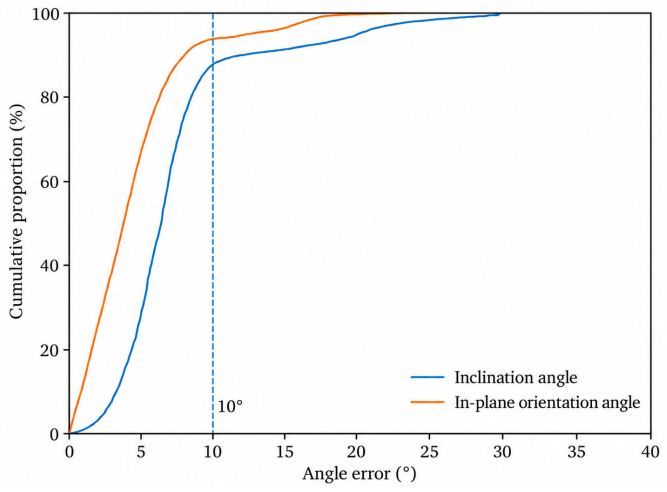


Figure 14. Cumulative distributions of inclination and in-plane orientation errors.

3.3. Dual-Camera Correction and Field Robotic Validation

3.3.1. Inclination Correction Evaluation

A total of 138 independent synchronized dual-camera target pairs were used to evaluate the proposed consistency-guided inclination-correction strategy. Main-camera-only estimation, direct main-auxiliary averaging, and the proposed selective correction strategy were compared on the same test set. Cross-view correspondence was established using the three-dimensional coordinate transformation and spatial matching procedure described in Section 2.3.4. The inclination results are summarized in Table 12.

Table 12. Inclination-estimation performance of different dual-camera fusion strategies.

Method	MAE (°)	Acc@10° (%)	Targets with Error <10°
Main camera only	7.5	86.2	119
Direct main-auxiliary averaging	7.3	87.7	121
Proposed dual-camera correction	6.3	90.6	125

Main-camera-only estimation produced an inclination MAE of 7.5° and an Acc@10° of 86.2%. Direct averaging reduced the MAE to 7.3° and increased Acc@10° to 87.7%. The proposed consistency-guided correction further reduced the MAE to 6.3° and increased Acc@10° to 90.6%. The stratified results for different inclination ranges are presented in Table 13.

Table 13. Stratified performance before and after dual-camera correction.

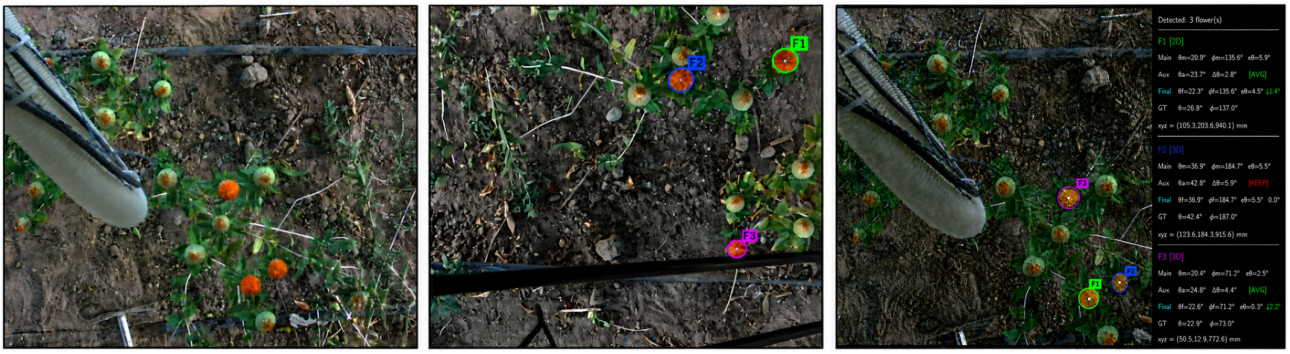
Inclination Range	Samples	Main-Camera MAE (°)	Corrected MAE (°)	Main-Camera Acc@10° (%)	Corrected Acc@10° (%)
<20°	22	8.7	8.1	77.3	81.8
20–40°	52	7.2	5.8	90.4	94.2
40–60°	43	7.0	5.6	93.0	97.7
>60°	21	7.8	7.1	71.4	76.2
Total	138	7.5	6.3	86.2	90.6

For inclinations below 20°, the MAE decreased from 8.7° to 8.1°, while Acc@10° increased from 77.3% to 81.8%. In the 20–40° range, the MAE decreased from 7.2° to 5.8° and Acc@10° increased from 90.4% to 94.2%. For the 40–60° range, the corresponding MAE decreased from 7.0° to 5.6°, while Acc@10° increased from 93.0% to 97.7%. For inclinations above 60°, the MAE decreased from 7.8° to 7.1° and Acc@10° increased from 71.4% to 76.2%. Across all 138 test pairs, the corrected MAE and Acc@10° were 6.3° and 90.6%, respectively.

A representative example of the main-auxiliary dual-camera inclination correction is shown in Figure 15.

3.3.2. Field Validation of Robot Visual Guidance and Harvesting Completeness

Field robotic guidance experiments were conducted at Hongqi Farm in Changji, Xinjiang, China. The experimental platform consisted of a wheeled mobile chassis, two ZED2i cameras, a six-degree-of-freedom manipulator, and a negative-pressure-assisted cutting-and-suction end effector. The cameras and manipulator base were rigidly mounted on the same platform to maintain a fixed spatial relationship during operation. Using the calibrated transformation between the main-camera coordinate system and the robot-base coordinate system, the estimated target positions and poses were transformed into the robot coordinate system for subsequent manipulator planning and end-effector alignment. The field experimental platform is shown in Figure 16.



(**a**) (**b**) (**c**)

Figure 15. Example of consistency-guided inclination decision and correction using the main-auxiliary dual-camera system: (**a**) Main-camera view. (**b**) Auxiliary-camera view. (**c**) Final inclination after consistency assessment.



Figure 16. Safflower harvesting experimental platform.

The robotic visual-guidance workflow is shown in Figure 17. After the mobile platform reached the designated working position, the system sequentially performed safflower segmentation, corolla–bud association, main-camera pose estimation, cross-view target matching, and consistency-guided inclination correction. Valid target positions and poses were then transformed into the robot coordinate system and transmitted to the manipulator for end-effector alignment. According to the recognized target coordinates and pose, the manipulator moved the end effector to cover the flower head, after which the filament was cut and then aspirated by the negative-pressure fan. Targets for which a reliable corolla contour or corolla–bud topological relationship could not be obtained were classified as invalid pose outputs and excluded from subsequent robotic guidance.

Of 515 field targets, 494 were recognized and localized, giving a rate of 95.92%, and the end effector reached the commanded pose for 489, giving 94.95%. The 26 arrival failures comprised 21 recognition/localization and five robotic execution failures. For 447 targets with complete reference measurements, inclination and orientation MAEs were 6.0° and 4.72°. Mean visual-processing time was 0.12 s per target; one observed harvesting operation took approximately 8.2 s. The separate 50-target HCR comparison is reported in Table 14. Figure 18 shows single-target harvesting with the end-effector cover removed to reveal cutting and suction.

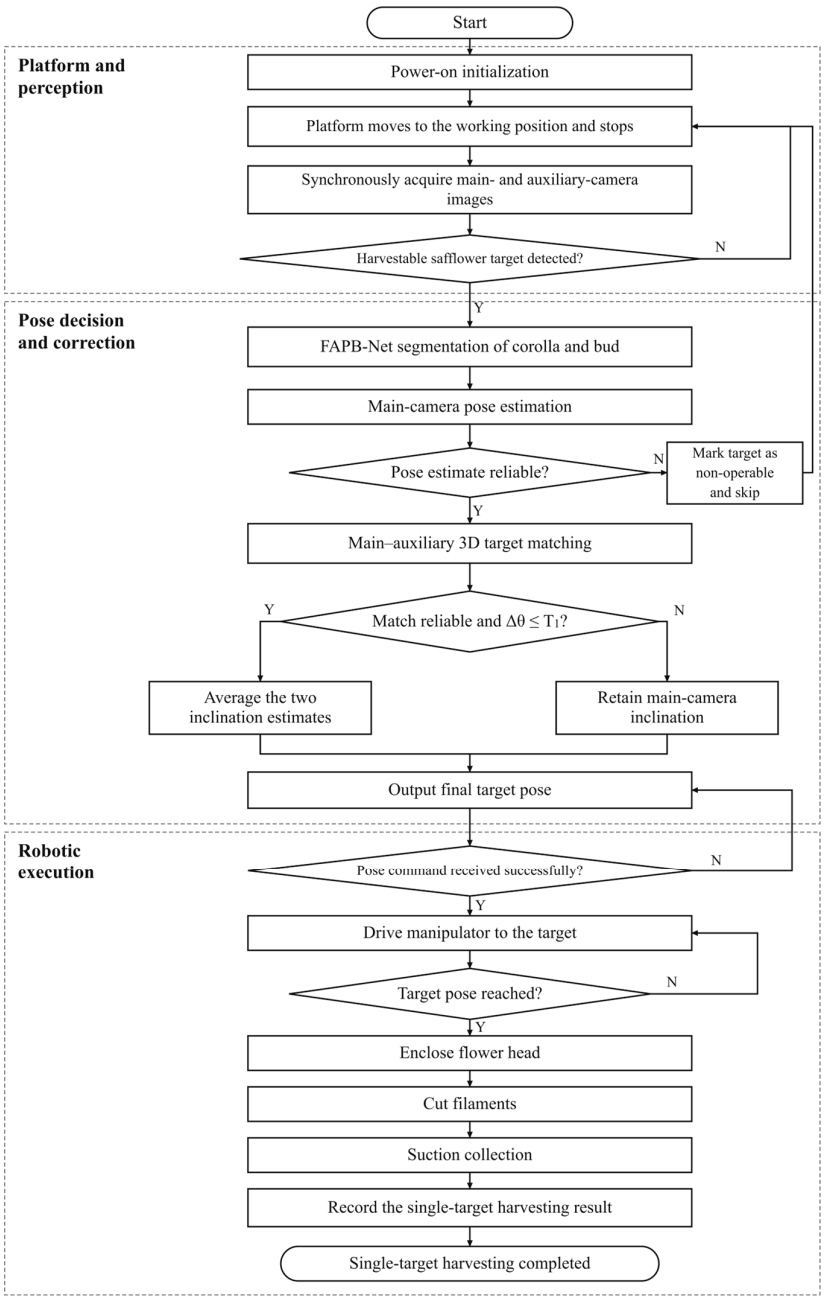


Figure 17. Workflow of robotic safflower harvesting with main-auxiliary dual-camera pose guidance.

Table 14. Comparison of harvesting completeness under different robotic guidance strategies.

Guidance Strategy	Samples	Mean HCR ± SD (%)	*p*-Value
Position-only guidance	25	91.40 ± 4.96	<0.05
Proposed pose-guided strategy	25	94.63 ± 4.09	

Pose guidance increased mean HCR from 91.40 ± 4.96% to 94.63 ± 4.09%, as shown in Table 14. The difference was 3.23 percentage points, with a 95% CI of 0.64–5.82 percentage points, $p < 0.05$, and Cohen's d of 0.71. Under the observed variability, 25 targets per group provided 80% power to detect a difference of approximately 3.68 percentage points at a two-sided significance level of 0.05.



(**a**) (**b**) (**c**) (**d**)

Figure 18. Representative field robotic harvesting process: (**a**) approaching the target flower head; (**b**) aligning the end effector with the target flower head; (**c**) filament cutting followed by suction collection; (**d**) completion of single-target harvesting.

Representative field-operation failures are shown in Figure 19, mainly including visual-perception failure and mechanical-execution failure. In Figure 19a, F1–F5 denote individual safflower targets, and different colors are used to distinguish their marked locations.



(**a**) (**b**)

Figure 19. Representative field-operation failure cases during robotic harvesting: (**a**) visual-perception failure leading to inaccurate target localization; (**b**) end-effector execution failure during the target-covering, cutting, and suction process.

4. Discussion

4.1. Segmentation Performance and Structural Pose Representation

FAPB-Net provided the structural information required for subsequent safflower pose estimation. Compared with SegFormer-B2, FAPB-Net improved the corolla and bud IOUs by 0.60 and 0.73 percentage points, respectively, while maintaining an inference speed of 28 FPS. Although the absolute improvements in segmentation accuracy were moderate, accurate preservation of corolla boundaries and small bud regions is particularly important for the proposed pipeline because inclination estimation depends on corolla ellipse fitting, whereas in-plane orientation estimation relies on the spatial relationship between the bud and corolla. The ablation results also showed progressive improvements after introducing Parallel-CBAM, ASPP, the ResNet34 encoder, and boundary-enhanced supervision, indicating that these components jointly improved the extraction of structural features required for subsequent geometric analysis.

The geometric baselines represent corolla shape through axes and use a fixed angular convention to produce reproducible orientation predictions. This convention does not distinguish opposite biological growth directions, so their full-range orientation errors include the effects of unresolved directional ambiguity. The directed vector from the bud

centroid to the corolla center provides a biological directional cue, enabling the proposed method to represent growth orientation over the full 0–360° range.

4.2. Inclination Error Characteristics and Selective Dual-Camera Correction

The inclination-estimation results showed a clear dependence on corolla projection geometry. Estimation accuracy was higher in the intermediate inclination ranges than at the two ends of the angular range. When the inclination was below 20°, the projected corolla approached a circular shape and the ellipse axis ratio became close to unity. Under this condition, relatively small contour perturbations could cause comparatively large angular variations. At large inclinations, occlusion and incomplete corolla boundaries increasingly affected ellipse fitting. These characteristics are consistent with the lower Acc@10° values observed below 20° and above 60°, as well as the long-tail component in the cumulative inclination-error distribution.

Selective correction reduced inclination MAE from 7.5° to 6.3°, a 16.0% reduction, while direct averaging yielded 7.3°. Acc@10° increased from 86.2% to 90.6%, with qualifying targets increasing from 119 to 125 (Table 12). Gains were greater at 20–60°, while near-circular projections and incomplete contours limited improvement at extreme inclinations (Table 13). The harvesting benefit of pose guidance was evaluated separately through the HCR comparison in Table 14. Dual-camera correction requires an additional camera and mounting space, inter-camera calibration, synchronized acquisition, segmentation of both views, and cross-view matching. The main-camera inclination is retained when matching or angular consistency fails; orientation always comes from the main camera.

4.3. Field Robotic Performance and Practical Implications

The field experiments evaluated the integrated visual-guidance framework across perception, pose estimation, and robotic execution. Among 515 field targets, recognition/localization and end-effector arrival rates reached 95.92% and 94.95%, respectively. For the 447 targets with complete reference measurements, the inclination and in-plane orientation MAEs were 6.0° and 4.72°, respectively. These results indicate that the proposed framework can provide position and pose information to support robotic harvesting under the tested field conditions. Related tomato and mango harvesting systems have also linked visual localization to occlusion-aware motion planning [31] and coordinated grasping and stem cutting [32], respectively.

In the harvesting-completeness experiment, pose-guided harvesting achieved a mean HCR of 94.63%, compared with 91.40% for position-only guidance, representing an improvement of 3.23 percentage points (Table 14). Unlike position-only guidance, the pose-guided strategy incorporates inclination and in-plane orientation to provide directional information for end-effector approach and alignment. The observed improvement supports the practical value of incorporating pose information into robotic harvesting.

4.4. Limitations and Future Work

The reliability of safflower pose estimation is closely related to the visibility and structural integrity of the corolla and bud. In the single-view evaluation, 401 of 534 targets produced valid pose estimates, corresponding to a valid-pose output rate of 75.09%. Invalid estimates were mainly associated with occlusion, incomplete segmentation, and unreliable corolla–bud correspondence, while mechanical interference affected robotic execution (Figure 19). Although selective dual-camera correction improved inclination estimation, its effectiveness remained limited when both viewpoints were affected by occlusion or reliable cross-view correspondence could not be established. Active occlusion handling explored

in strawberry harvesting [33] may offer a complementary approach to improving target visibility and access.

The field evaluation covered single-target operations on Jihong No. 3 safflower at Hongqi Farm during the flowering period under the reported sunny and cloudy illumination conditions. The harvesting-completeness results also require further validation with larger samples and more diverse field conditions. During consecutive harvesting, wind-induced flower movement, changing occlusion, and removal of neighboring targets may invalidate previously estimated target positions and correspondences, requiring renewed detection and matching before subsequent execution. Dynamic tracking, target scheduling, and throughput during consecutive operations were not quantified in the present evaluation. Future work will integrate repeated image acquisition, target updating, and scheduling into consecutive harvesting and evaluate overall harvesting success, operation-time variability, and throughput across additional sites, growth stages, and varieties. Hybrid visual servoing investigated in apple picking [34] also provides a reference for coupling updated visual observations with manipulator control during target approach. Application to other flower crops will require crop-specific segmentation and verification of the approximately planar corolla and directed bud–corolla structural assumptions.

5. Conclusions

This study developed a visual-guidance framework for robotic safflower harvesting by integrating FAPB-Net segmentation, corolla–bud topology, and selective main-auxiliary dual-camera inclination correction. FAPB-Net achieved corolla and bud IOUs of 90.28% and 70.85%, respectively, at 28 FPS, providing structural information for subsequent geometric pose estimation.

Using corolla ellipse geometry and the directed corolla–bud relationship, the proposed method estimated inclination and 0–360° in-plane orientation with MAEs of 7.38° and 4.65°, respectively, on 401 valid pose instances. The corresponding Acc@10° values were 87.28% and 94.76%. On 138 synchronized dual-camera test pairs, the consistency-guided correction reduced inclination MAE from 7.5° to 6.3° and increased Acc@10° from 86.2% to 90.6%.

Field experiments with 515 targets achieved recognition/localization and end-effector arrival rates of 95.92% and 94.95%, respectively. For 447 targets with complete reference measurements, the inclination and in-plane orientation MAEs were 6.0° and 4.72°, respectively. Pose-guided harvesting increased the mean HCR from 91.40 ± 4.96% to 94.63 ± 4.09% ($p < 0.05$), demonstrating the practical value of target pose information for end-effector alignment and harvesting completeness under the tested field conditions.

Author Contributions: Conceptualization, H.G. and D.Z.; methodology, D.Z.; software, D.Z.; validation, H.G., D.Z. and G.Z.; formal analysis, T.W.; investigation, J.L.; resources, H.G.; data curation, W.B.; writing—original draft preparation, D.Z.; writing—review and editing, H.G.; visualization, D.Z.; supervision, Y.D.; project administration, H.G.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the earmarked fund for XJARS, grant number (XARS-05-202610).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available due to ongoing research but are available from the corresponding author upon reasonable request.

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

References

1. Huang, Y.; Xu, S.; Chen, H.; Li, G.; Dong, H.; Yu, J.; Zhang, X.; Chen, R. A review of visual perception technology for intelligent fruit harvesting robots. *Front. Plant Sci.* **2025**, *16*, 1646871. [CrossRef] [PubMed]
2. Zhou, H.; Wang, X.; Au, W.; Kang, H.; Chen, C. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precis. Agric.* **2022**, *23*, 1856–1907. [CrossRef]
3. Zhang, K.; Lammers, K.; Chu, P.; Li, Z.; Lu, R. An automated apple harvesting robot—From system design to field evaluation. *J. Field Robot.* **2024**, *41*, 2384–2400. [CrossRef]
4. Yin, C.; Huang, J.; Xia, Y.; Zheng, H.; Fu, W.; Zhang, B. Design, development, integration and field evaluation of a dual robotic arm mango harvesting robot. *J. Field Robot.* **2025**, *42*, 3705–3725. [CrossRef]
5. Bai, Y.; Mao, S.; Zhou, J.; Zhang, B. Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting. *Precis. Agric.* **2023**, *24*, 727–743. [CrossRef]
6. Li, H.; Huang, J.; Gu, Z.; He, D.; Huang, J.; Wang, C. Positioning of mango picking point using an improved YOLOv8 architecture with object detection and instance segmentation. *Biosyst. Eng.* **2024**, *247*, 202–220. [CrossRef]
7. Zhang, T.; Wu, F.; Wang, M.; Chen, Z.; Li, L.; Zou, X. Grape-bunch identification and location of picking points on occluded fruit axis based on YOLOv5-GAP. *Horticulturae* **2023**, *9*, 498. [CrossRef]
8. 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]
9. Sun, T.; Zhang, W.; Miao, Z.; Zhang, Z.; Li, N. Object localization methodology in occluded agricultural environments through deep learning and active sensing. *Comput. Electron. Agric.* **2023**, *212*, 108141. [CrossRef]
10. Guo, H.; Wu, T.; Gao, G.; Qiu, Z.; Chen, H. Lightweight safflower cluster detection based on YOLOv5. *Sci. Rep.* **2024**, *14*, 18579. [CrossRef] [PubMed]
11. Wang, X.; Zhou, J.; Xu, Y.; Cui, C.; Liu, Z.; Chen, J. Location of safflower filaments picking points in complex environment based on improved YOLOv5 algorithm. *Comput. Electron. Agric.* **2024**, *227*, 109463. [CrossRef]
12. Xing, Z.; Zhang, Z.; Shi, R.; Guo, Q.; Zeng, C. Filament-necking localization method via combining improved PSO with rotated rectangle algorithm for safflower-picking robots. *Comput. Electron. Agric.* **2023**, *215*, 108464. [CrossRef]
13. Zhang, H.; Ge, Y.; Xia, H.; Sun, C. Safflower picking points localization method during the full harvest period based on SBP-YOLOv8s-seg network. *Comput. Electron. Agric.* **2024**, *227*, 109646. [CrossRef]
14. Chen, B.; Ding, F.; Ma, B.; Yao, Q.; Ning, S. Research on detection and location method of safflower filament picking points during the blooming period in unstructured environments. *Sci. Rep.* **2025**, *15*, 10851. [CrossRef] [PubMed]
15. Yin, W.; Wen, H.; Ning, Z.; Ye, J.; Dong, Z.; Luo, L. Fruit detection and pose estimation for grape cluster-harvesting robot using binocular imagery based on deep neural networks. *Front. Robot. AI* **2021**, *8*, 626989. [CrossRef] [PubMed]
16. 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]
17. Kim, T.; Lee, D.H.; Kim, K.C.; Kim, Y.J. 2D pose estimation of multiple tomato fruit-bearing systems for robotic harvesting. *Comput. Electron. Agric.* **2023**, *211*, 108004. [CrossRef]
18. Zhang, F.; Gao, J.; Song, C.; Zhou, H.; Zou, K.; Xie, J.; Yuan, T.; Zhang, J. TPMv2: An end-to-end tomato pose method based on 3D key points detection. *Comput. Electron. Agric.* **2023**, *210*, 107878. [CrossRef]
19. Ci, J.; Wang, X.; Rapado-Rincón, D.; Burusa, A.K.; Kootstra, G. 3D pose estimation of tomato peduncle nodes using deep keypoint detection and point cloud. *Biosyst. Eng.* **2024**, *243*, 57–69. [CrossRef]
20. Ni, J.; Zhu, L.; Dong, L.; Wang, R.; Chen, K.; Gao, J.; Wang, W.; Zhou, L.; Zhao, B.; Rong, J.; et al. TomatoPoseNet: An efficient keypoint-based 6D pose estimation model for non-destructive tomato harvesting. *Agronomy* **2024**, *14*, 3027. [CrossRef]
21. Dong, L.; Zhu, L.; Zhao, B.; Wang, R.; Ni, J.; Liu, S.; Chen, K.; Cui, X.; Zhou, L. Semantic segmentation-based observation pose estimation method for tomato harvesting robots. *Comput. Electron. Agric.* **2025**, *230*, 109895. [CrossRef]
22. Zhang, G.; Li, L.; Zhang, Y.; Liang, J.; Chun, C. Citrus pose estimation under complex orchard environment for robotic harvesting. *Eur. J. Agron.* **2025**, *162*, 127418. [CrossRef]
23. 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. [CrossRef]
24. Zhao, G.; Dong, S.; Wen, J.; Ban, Y.; Zhang, X. Selective fruit harvesting prediction and 6D pose estimation based on YOLOv7 multi-parameter recognition. *Comput. Electron. Agric.* **2025**, *229*, 109815. [CrossRef]
25. Zheng, X.; Rong, J.; Zhang, Z.; Yang, Y.; Li, W.; Yuan, T. Fruit growing direction recognition and nesting grasping strategies for tomato harvesting robots. *J. Field Robot.* **2024**, *41*, 300–313. [CrossRef]
26. Wang, D.; Dong, Y.; Lian, J.; Gu, D. Adaptive end-effector pose control for tomato harvesting robots. *J. Field Robot.* **2023**, *40*, 535–551. [CrossRef]
27. Rong, J.; Hu, L.; Zhou, H.; Dai, G.; Yuan, T.; Wang, P. A selective harvesting robot for cherry tomatoes: Design, development, field evaluation analysis. *J. Field Robot.* **2024**, *41*, 2564–2582. [CrossRef]

28. Yang, L.; Noguchi, T.; Hoshino, Y. Development of a grape cut point detection system using multi-cameras for a grape-harvesting robot. *Sensors* **2024**, *24*, 8035. [CrossRef] [PubMed]
29. Liu, Z.; Abeyrathna, R.M.R.D.; Sampurno, R.M.; Nakaguchi, V.M.; Ahamed, T. A new occlusion-avoidance and dual-view fruit localization method with a 6-DoF manipulator for orchard harvesting. *Comput. Electron. Agric.* **2025**, *237*, 110634. [CrossRef]
30. Jiang, Z.; Zhou, J.; Han, H.; Wang, Y. An apple fruit localization system based on accurate and flexible hand-eye pose acquisition for robotic harvesting. *Comput. Electron. Agric.* **2025**, *229*, 109705. [CrossRef]
31. Matache, M.G.; Marin, F.B.; Persu, C.I.; Cristea, R.D.; Nenciu, F.; Atanasov, A.Z. Autonomous tomato harvesting system integrating AI-controlled robotics in greenhouses. *Agriculture* **2026**, *16*, 847. [CrossRef]
32. Liu, Q.; Lu, Z. A vision-based robot system with grasping-cutting strategy for mango harvesting. *Agriculture* **2026**, *16*, 132. [CrossRef]
33. He, Z.; Liu, Z.; Zhou, Z.; Karkee, M.; Zhang, Q. Improving picking efficiency under occlusion: Design, development, and field evaluation of an innovative robotic strawberry harvester. *Comput. Electron. Agric.* **2025**, *237*, 110684. [CrossRef]
34. Kan, J.; Wu, Y.; Dong, R.; Yao, S.; Zhao, X.; Zou, T.; Kang, B.; Li, J. A progressive hybrid automatic switching visual servoing method for apple-picking robots. *Agriculture* **2026**, *16*, 620. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.