Preprints.org

Article

Not peer-reviewed version

# Green Pepper Harvesting Robot System Based on Multi-Target Tracking with Filtering and Intelligent Scheduling

Tianyu Liu *, Zelong Liu , Jianmin Wang , Dongxin Guo , Yuxuan Tan , Ping Jiang *

Posted Date: 2 March 2026

doi: 10.20944/preprints202603.0095.v1

Keywords: green pepper; intelligent harvesting; modular system; 3D localization filtering; multi-object scheduling; hardware-software co-design

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1

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.

*Article*

# Green Pepper Harvesting Robot System Based on Multi-Target Tracking with Filtering and Intelligent Scheduling

**Tianyu Liu *, Zelong Liu, Jianmin Wang, Dongxin Guo, Yuxuan Tan and Ping Jiang ***

College of Mechanical and Electrical Engineering, Hunan Agricultural University, Changsha 410128, China

* Correspondence: liutianyu@hunau.edu.cn (T. L.); 1233032@hunau.edu.cn (P. J.)

**Abstract**

To address the challenges of unstable target localization and poor multi-module coordination in automated green pepper harvesting—caused by occlusions from branches and leaves, as well as varying lighting conditions—this paper presents the design and implementation of a modular robotic picking system. At the perception level, the system integrates a YOLOv8 detector with a RealSense D435i camera to identify and locate the calyx–ectocarp junctions of green peppers. A multi-target tracking with filtering algorithm is proposed, combining IoU-based association, Mahalanobis-distance-based matching, the Hungarian algorithm, Kalman filtering, and single exponential smoothing. This algorithm suppresses depth noise and trajectory jitter, thereby enhancing the stability and accuracy of 3D localization. At the control and execution level, a depth-first picking sequence strategy with ID freeze-state management is implemented within a multithreaded software–hardware co-design architecture. This approach avoids task conflicts and duplicate operations while supporting continuous multi-fruit harvesting. Field experiments under natural outdoor lighting and varying occlusion levels demonstrate that the proposed system achieves recognition rates of 91.57% and 80.29%, and harvesting success rates of 82.85% and 77.68% for non-occluded and lightly occluded fruits, respectively. The average picking cycle per pepper fruit is 9.8 s. This system provides an effective technical solution for addressing stability control challenges in the automated harvesting process of green peppers.

**Keywords:** green pepper; intelligent harvesting; modular system; 3D localization filtering; multi-object scheduling; hardware-software co-design

## 1. Introduction

The introduction is a beginning section of a manuscript which states the purpose of the study, overviews or summarizes previous findings and progress related to this study, and indicates its significance in this research field. It is generally followed by the body and discussion.

Pepper is a globally important economic crop, playing a critical role in agricultural production and the food industry. According to statistics from the United Nations Food and Agriculture Organization[1], China's green pepper production in 2024 reached approximately 17.3309 million tons, accounting for about 38.71% of the global total output. This solidifies China's position as the world's largest producer and consumer of green peppers. Against this backdrop, the efficient and non-destructive harvesting of mature green peppers is crucial for enhancing their economic value. Currently, harvesting operations still rely heavily on manual labor, which, amid an aging population and continuously rising labor costs[2], has become a significant bottleneck constraining the development of the pepper industry. Consequently, there is an increasingly urgent need for the development of intelligent harvesting equipment and robotic technologies tailored for green peppers[3,4].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



In recent years, substantial progress has been made in fruit and vegetable harvesting robots. Regarding perception, researchers have employed and optimized various deep learning architectures and image processing algorithms to tackle challenges in fruits and vegetables detection, localization, and picking point identification in complex agricultural environments[5–11]. In the development of harvesting robot systems, Pan et al. designed an end-effector equipped with tactile sensors, achieving a harvesting success rate of 79.17% through visual servoing and grasp posture control algorithms, with a single-fruit picking time of approximately 15 seconds[12]. Alam et al. developed a low-cost Cartesian robot integrated with a GRBL controller and a Kinect V2 sensor. By optimizing the linear movement speed to 0.83 cm/s, they achieved a harvesting efficiency of 86% and a fruit damage rate of only 5%[13]. Ravuri et al. applied Mask R-CNN and a 6-degree-of-freedom robotic arm to green pepper harvesting[14]. Although the fruit localization accuracy reached 90.7%, the harvesting success rate was only 31.4% with a damage rate of 6.97%, primarily due to the need for further improvement in the adaptability of the end-effector to dynamic environments.

Nevertheless, despite the progress in visual perception and system integration of fruit and vegetable harvesting robots, the stability and continuity of object localization and tracking in the picking task is still a problem to be further improved. Li et al. integrated YOLOv8 with ByteTrack to track falling maize kernels, achieving over 99% counting accuracy in dynamic scenes[15], yet their method primarily focused on 2D trajectory association without explicit handling of 3D coordinate noise. Paul et al. developed a comprehensive capsicum harvesting system that employed YOLOv8 for detection, segmentation, and growth stage classification, and integrated a RealSense D455 camera for 3D peduncle localization[16]. Its tracking module relied primarily on ByteTrack without dedicated filtering mechanisms to suppress depth noise or trajectory jitter in continuous harvesting sequences. Liang et al. developed a tomato-picking robotic system based on the fusion of global and hand-level vision. Through spatial asynchronous localization and vision-based servo control, it significantly improved positioning accuracy and picking success rates in complex environments[17]. However, maintaining stable servo tracking under dynamic occlusion conditions remains challenging. Rapado-Rincón et al. proposed MinkSORT, a 3D sparse convolutional network for tomato tracking in greenhouses[18], which improved tracking accuracy under occlusion but required extensive 3D data and computational resources, limiting its real-time applicability in field environments. Similarly, Arlotta et al. employed an Extended Kalman Filter with RGB-D data for grape bunch tracking on a mobile robot, demonstrating robustness in handling intermittent measurements[19], yet their framework did not incorporate trajectory smoothing or ID consistency management for continuous multi-fruit harvesting tasks.

Collectively, these methods highlight two limitations of multi-object tracking systems in agriculture: insufficient suppression of depth-sensing noise and trajectory jitter in 3D localization of fruits and vegetables and ack of tight integration between tracking consistency, sequential execution control, and system-level coordination in multi-fruit harvesting scenarios. Therefore, to tackle the aforementioned challenges in localization stability and system coordination, and to improve the perception robustness, spatial positioning accuracy, scheduling intelligence, and overall collaborative efficiency of an automated green pepper harvesting system in complex natural environments, this paper designs and implements a modular automated picking system for green pepper harvesting tasks. The core work of this study focuses on enhancing the system's overall picking performance and cooperative operational capability in challenging conditions. Specifically, at the perception level, the integration of multi-object tracking and a three-dimensional filtering algorithm significantly improves the accuracy and output stability of fruit spatial localization. At the decision-making and execution level, optimized picking sequence scheduling, ID consistency management, and a closed-loop communication mechanism between the end-effector and the control system effectively enhance the reliability, real-time performance, and operational efficiency of multi-module collaboration.

The main contributions of this paper are as follows:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



(1) A modular robotic harvesting system for green peppers was proposed and implemented, which integrates YOLOv8-based visual recognition, RealSense depth camera-based 3D localization, and a robotic arm with a cutting-gripping end-effector. The system achieves automated operation from fruit recognition and localization to precise picking, with enhanced reliability through closed-loop end-effector control.

(2) To address the problems of target loss and severe coordinate fluctuation in complex harvesting environments, a multi-object tracking and 3D coordinate smoothing algorithm is developed. The tracking employs the Hungarian algorithm for global optimal association, which integrates an IoU-based cost matrix with a Mahalanobis-distance-based gating threshold for robust matching. Subsequently, the system applies a multi-stage filtering framework comprising Kalman filtering, the 3σ criterion, and single exponential smoothing. This approach effectively suppresses outliers and jitter in the localization results, significantly improving the robustness and accuracy of the visual positioning system.

(3) A depth-first picking sequence scheduling strategy with ID freeze-state management is designed, together with a multithreaded software–hardware cooperative control architecture. This design resolves task conflicts and resource competition during multi-target harvesting and ensures continuous, stable, and efficient system operation.

(4) Extensive field experiments are conducted under natural outdoor lighting and different occlusion levels to comprehensively evaluate the system performance. The experimental results verify the effectiveness of the proposed methods in terms of recognition accuracy, harvesting success rate, and operational stability, while also revealing the current limitations of the system and providing guidance for future research and practical application.

## 2. Materials and Methods

### *2.1. Robot System Settings*

The intelligent green pepper harvesting system consists of an Intel RealSense D435i camera for RGB-D sensing, an upper computer (Ubuntu 18.04) for vision and planning algorithms, a 6-DOF industrial robotic arm with its dedicated controller, a custom-designed cutting–gripping end-effector, and an STM32-based control unit (Figure 1). The camera streams data to the upper computer via USB, which processes it to generate motion commands. These commands are then sent via Ethernet to the robot controller for execution.

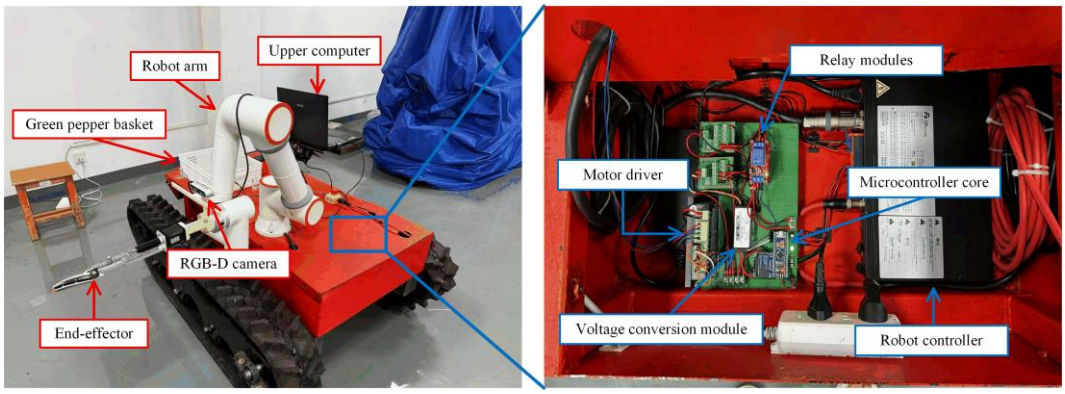



**Figure 1.** A robotic system for automated harvesting of green peppers.

The end-effector assembly features a self-designed linear actuator module (including a push-rod and mounting base) and a commercial cutting–gripping head. The actuator, driven by the STM32 control circuit, operates the head via the push-rod. The circuit itself integrates power conversion, motor drivers, and a relay module. It generates PWM signals for actuator control and enables device-status synchronization with the upper computer through I/O communication, ensuring coordinated operation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



The overall system follows a closed-loop "perception–decision–action" workflow (Figure 2).

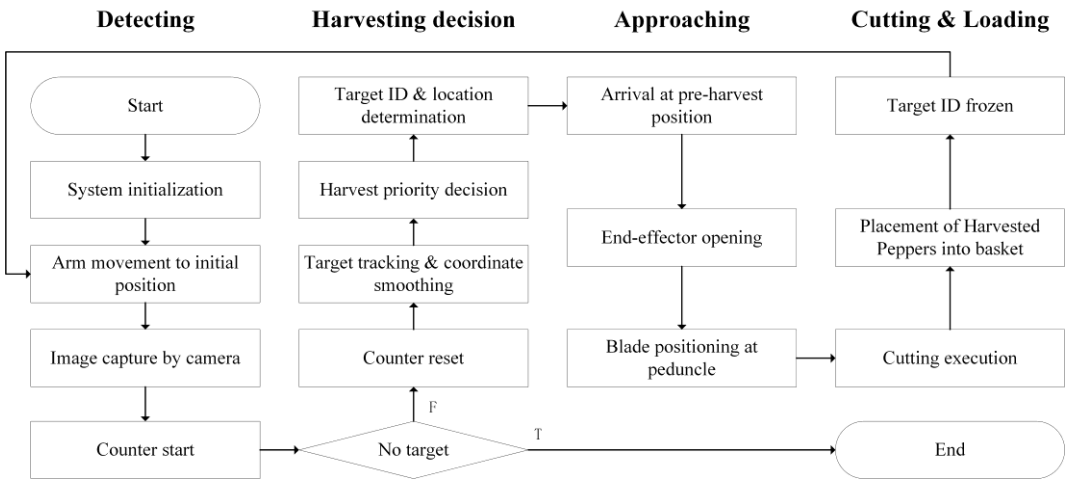



**Figure 2.** Green pepper picking system workflow diagram.

Key hardware specifications are listed in Table 1.

**Table 1.** Hardware system main parameters.

| Module name | Model/control method | Main parameters |
|---|---|---|
| Upper computer | Hasee z7-ct7na | Intel i7-9750H, 16GB RAM, GTX1660Ti, Ubuntu 18.04 |
| RGB-D camera | RealSense D435i | 1280×720@30FPS, depth accuracy ±2% |
| Robot arm | FR5-SPARK | 6-DOF, Load 5kg, Repeat positioning accuracy ±0.02mm, Fiber optic Ethernet communication |
| End-effector | Self-designed actuator module，commercial cutting–gripping head | 24V DC motor drive, actuator stroke 40mm, maximum clamping width 50mm, motor torque 2N-m |

*2.2. Image Acquisition and Dataset Annotation*

The pepper image dataset used in this study was collected in June 2025, with the acquisition objects being self-cultivated potted pepper plants under outdoor natural lighting conditions (Figure 3). The image acquisition device was an Intel RealSense D435i camera. To enhance the robustness and generalization ability of the model under different lighting conditions, image collection covered typical time periods, including morning (7:00–9:00), noon (11:00–13:00), and afternoon (15:00–17:00), encompassing diverse natural light intensities and weather conditions. The collected images had a resolution of 1280 × 720 and were saved in JPG format. The shooting distance remained stable between 0.5 and 0.8 meters. To comprehensively capture the performance of peppers under varying occlusion levels, poses, and sizes, and to enhance viewpoint and target diversity, each potted plant was rotated at 10° intervals in the horizontal direction and randomly photographed at eye-level, upward, and downward angles until full 360° coverage was achieved, before proceeding to the next plant. A total of 3,871 images were collected, and after screening, 3,450 images were retained for subsequent annotation. The screening criteria involved removing images of poor quality, such as those with blurring or ghosting caused by movement or changes in lighting during the rotation-based shooting process.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



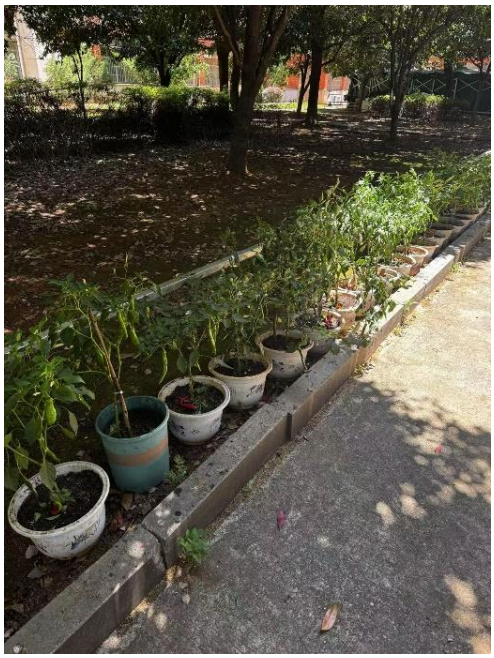


**Figure 3.** Potted green peppers.

Based on the overall structural diagram of green peppers(Figure 4), the LabelImg tool was used to manually annotate bounding boxes by precisely locating the junction between the calyx and the outer pericarp of each pepper[20], as shown in Figure 5.

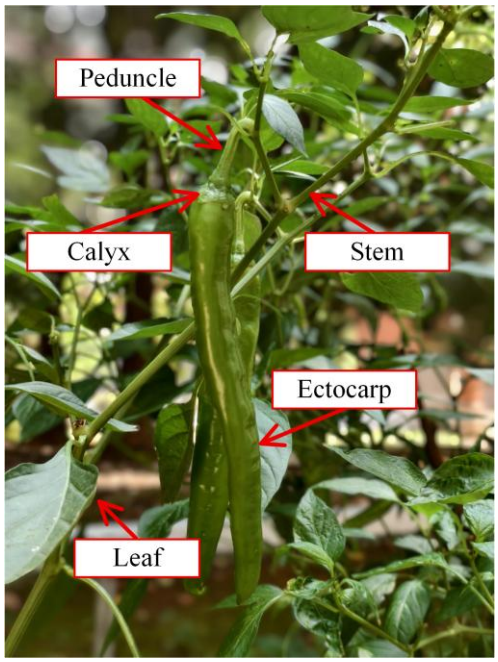



**Figure 4.** The overall structure of green peppers.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



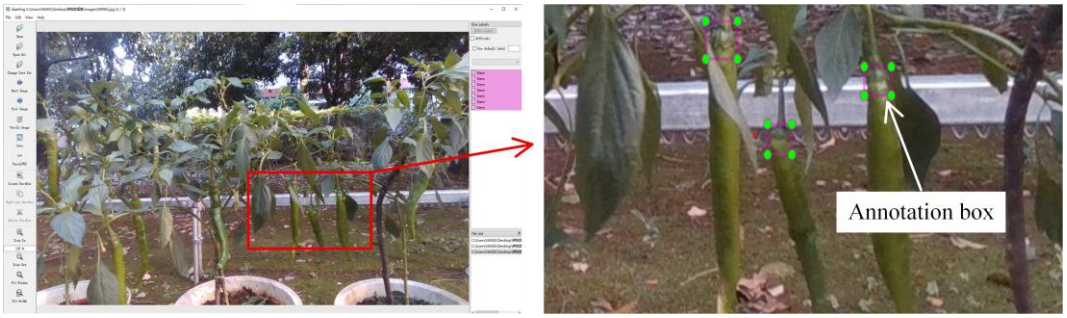



**Figure 5.** Labeling of data sets.

The annotated targets were categorized into four occlusion levels based on visual assessment of the bounding box area containing the calyx and the ectocarp junction: no occlusion, where the target region is fully visible without any obstruction; light occlusion, where a small portion of the bounding box is covered but the majority of the junction remains clearly discernible; moderate occlusion, where a significant portion is obscured, yet key characteristics of the junction are still partially visible; and heavy occlusion, where over half of the bounding box is covered, substantially limiting the visibility of the junction features. Representative examples of each occlusion level are provided in Figure 6. All annotations were saved under the unified label "Stem".


**Figure 6.** Four occlusion levels of green peppers.

### *2.3. Feature Detection and 3D Coordinate Extraction*

#### 2.3.1. YOLOv8 Target Detection Model Construction

For robust real-time detection of green peppers, the YOLOv8 object detection architecture was employed. To address the specific challenge of accurately locating small calyxes under occluded conditions, which demands high feature extraction capability, the YOLOv8l variant was selected. This model provides an optimal balance between detection precision and computational efficiency for our complex agricultural scenario.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



After partitioning the annotated dataset into training, testing, and validation sets at a ratio of 7:2:1, the model was trained on a platform equipped with an Intel i5-14600KF processor, 32 GB RAM, an NVIDIA GeForce RTX 5060TI GPU (16 GB VRAM), and a Windows 10 operating system, following the training parameters detailed in Table 2. The evaluation was performed on the test set using the saved optimal weight file. The recognition results are shown in Figure 7. The overall Precision, Recall, F1-score, and mAP reached 92.51%, 84.17%, 88.14%, and 89.65%, respectively. These metrics fully demonstrate that the model can reliably achieve high-precision recognition of small targets, meeting the practical requirements of intelligent harvesting robots in terms of detection accuracy.

**Table 2.** Model training parameters.

| Training parameter | Value |
|---|---|
| weight | Yolov8l.pt |
| batch-size | 8 |
| epochs | 300 |
| imgsz | 640×640 |
| momentum | 0.937 |
| warmup_bias_lr | 0.1 |
| lr0 | 0.01 |
| lrf | 0.01 |
| optimizer | SGD |
| weight_decay | 0.0005 |
| mosaic | 0.5 |
| cos_lr | True |
| seed | 0 |
| dropout | 0.0 |


**Figure 7.** Recognition results of green pepper.

2.3.2.3. D Coordinate Extraction

This study employs an Intel RealSense D435i RGB-D camera to achieve three-dimensional localization of harvesting points. The system simultaneously captures depth and color streams to generate spatially aligned RGB-D frames in real-time, while loading camera intrinsic parameters and distortion coefficients. Prior to depth mapping, geometric correction is first applied to RGB images using distortion coefficients to eliminate lens distortion effects on spatial alignment. Subsequently, the target pixel coordinates detected by the YOLOv8l model are matched with the corrected depth map, and the 3D spatial coordinates of targets are calculated through the camera coordinate system transformation model. This method ensures pixel-level accurate registration of color images and depth data, providing high-precision input for subsequent 3D positioning of the robotic arm.

The YOLOv8l model outputs the pixel coordinate center point (u, v) of each target. By retrieving the corresponding depth value Z (unit: mm) at this point in the depth map, the pixel coordinates can be converted to 3D coordinates (X, Y, Z) in the camera coordinate system using Equation (1)-(3), as illustrated in Figure 8:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



$$X=\frac{(u-c_x)}{f_x}*Z \quad (1)$$

$$Y=\frac{(v-c_y)}{f_y}*Z \quad (2)$$

$$Z=depth(u, v) \quad (3)$$

where (u, v) represents the target's pixel position in the RGB image coordinate system, ($f_x$, $f_y$) denotes the camera's focal length (in pixels), ($c_x$, $c_y$) indicates the principal point coordinates (i.e., the optical center's position in the pixel coordinate system), and Z represents the depth value obtained from the depth map for the corresponding pixel, measured in millimeters.


**Figure 8.** 3D Localization results based on YOLO detection.

*2.4. Multi-Target State Management and Coordinate Smoothing Filter*

The stability of 3D localization and the consistency of target identities are two critical prerequisites for continuous robotic harvesting. In outdoor environments, depth sensors such as the RealSense D435i are occasionally susceptible to transient noise caused by specular reflections from glossy surfaces such as peppers, or by missing depth values in occluded regions. These artifacts manifest as outliers or high-frequency jitter in the measured 3D coordinates. If used directly for motion planning, these unstable coordinates lead to unreliable picking points and subsequent harvesting failures. For the scheduling and decision-making system, maintaining persistent and correct identities for each pepper across frames is equally crucial. Densely clustered peppers and frequent occlusion by foliage can cause conventional IoU-based trackers to incorrectly switch identities or lose tracks between consecutive frames. Such identity fragmentation disrupts the pre-defined picking sequence and can cause the system to re-attempt already harvested or currently unreachable peppers, thereby stalling the continuous operation.

To address these cascading issues systematically, this paper proposes a multi-level filtering and state-management framework that transforms noisy, inconsistent sensor data into stable, identity-persistent picking points. The framework is designed in a layered manner, where each stage targets a specific source of instability: (1) multi-feature data association (fusing IoU and Mahalanobis distance) with Hungarian global optimization ensures robust inter-frame target matching and ID consistency; (2) a Kalman filter smooths the 2D bounding-box centers to suppress image-plane jitter

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



induced by plant motion or detection noise; (3) statistical outlier rejection based on the 3σ criterion removes depth-sensing artifacts from the historical 3D coordinate queue; and (4) single exponential smoothing further refines the inlier coordinates to produce a smooth, low-variance 3D trajectory. Once a target's coordinates stabilize, a depth-first picking sequence with ID freeze-state management is activated to schedule robotic actions while avoiding task conflicts and duplicate operations. This integrated approach ensures that coordinate smoothing and identity persistence are jointly maintained throughout the harvesting cycle. The workflow of the proposed multi-target state management and decision system is illustrated in Figure 9.

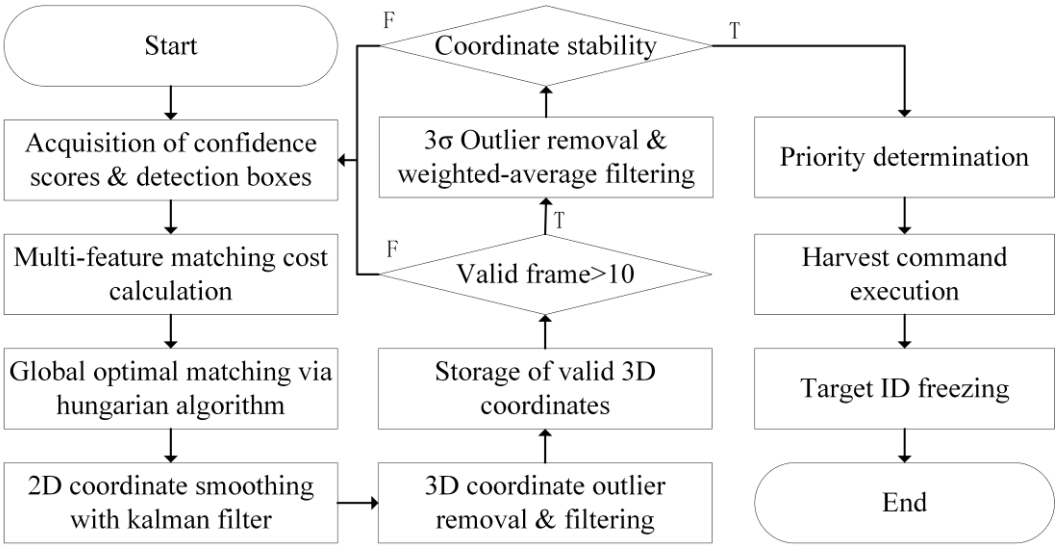



**Figure 9.** Flowchart of the harvesting decision system based on multi-level filtering.

2.4.1. Multi-Feature Fusion Data Association and Target Tracking

Based on the object bounding box information output by the YOLOv8 detection model for the current frame, spatial overlap consistency (IoU) and motion consistency (Mahalanobis distance[21]) are integrated into a unified cost matrix and globally optimized using the Hungarian algorithm[22].

The IoU between current detection $B_{current}$ and previous track $B^i_{prev}$ is defined as:

$$IOU\left(B_{current},B^i_{prev}\right)=\frac{\left|B_{current}\cap B^i_{prev}\right|}{\left|B_{current}\cup B^i_{prev}\right|} \quad (4)$$

where $\left|B_{current}\cap B^i_{prev}\right|$ represents the intersection area of the two bounding boxes, and $\left|B_{current}\cup B^i_{prev}\right|$ denotes their union area.

To determine the optimal matching threshold, this study statistically evaluated the tracking performance for multiple green pepper targets. The evaluation was conducted by applying different IoU thresholds (θ) to associate detected bounding boxes with ground-truth tracks across consecutive frames in an outdoor environment with light wind interference. The results are presented in Table 3. While a lower threshold (θ=0.1 or θ=0.3) yields a high ID matching rate with minimal switches, it increases the risk of false-positive matches due to overly permissive associations, which undermines tracking integrity. Conversely, a higher threshold (θ=0.7 or θ=0.9) leads to frequent matching failures and a sharp rise in ID switches, causing severe track fragmentation. The selection of θ=0.5 represents an optimal trade-off: it maintains a high correct matching rate (94%) to ensure track continuity, while being sufficiently stringent to significantly reduce the risk of erroneous matches compared to lower thresholds. Therefore, θ=0.5 effectively balances matching accuracy and ID stability, providing reliable tracking for subsequent harvesting operations.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



**Table 3.** Performance test comparison with different thresholds.

| θ | Total matching pairs | ID matches | ID switches | ID matching rate |
|---|---|---|---|---|
| 0.1 | 400 | 400 | 0 | 100% |
| 0.3 | 400 | 400 | 0 | 100% |
| 0.5 | 400 | 376 | 24 | 94% |
| 0.7 | 400 | 260 | 140 | 65% |
| 0.9 | 400 | 41 | 359 | 10.25% |

The Mahalanobis distance $D_M$ evaluates motion consistency between observation $z_k$ and Kalman-predicted state $\hat{x}_{k \mid k-1}$:

$$D_M^2=(z_k-H\hat{x}_{k \mid k-1})^T S_k^{-1}(z_k-H\hat{x}_{k \mid k-1}) \quad (5)$$

where $S_k=HP_{k \mid k-1}H^T+R$ is the innovation covariance matrix. A smaller $D_M$ indicates higher consistency between the observation and the predicted motion model. We set a threshold $\theta_M=9.210$ (corresponding to the 99% confidence level of the chi-squared distribution with 2 degrees of freedom) to reject improbable associations.

The final matching cost integrates both cues:

$$C_{i,j}=\lambda_{IoU}\cdot(1-IoU_{i,j})+\lambda_M\cdot \min\left(\frac{D_{M,i,j}}{\theta_M},1\right) \quad (6)$$

where $\lambda_{IoU}=0.7$ and $\lambda_M=0.3$ are empirically determined weighting coefficients that balance spatial and motion cues.

To quantitatively evaluate the effectiveness of the proposed fusion strategy, we manually annotated 300 representative image frames, covering various complex scenarios such as partial occlusion, fruit overlap, and wind-induced motion. Based on the same detection results, two tracking strategies were applied and compared, and their association performance was assessed using three metrics: identity switches, association accuracy, and miss rate.

Performance evaluation on 300 annotated frames (Table 4) shows that the fusion method reduces ID switches by 55% and improves association accuracy from 84.47% to 92.48% compared to IoU-only matching, demonstrating superior robustness in occluded and overlapping scenarios.

It is worth noting that the miss rate of the fusion method shows a slight increase, which primarily stems from the stricter matching criteria imposed by the Mahalanobis distance threshold. In complex scenarios, when the motion of a target briefly deviates from the constant velocity model—for instance, due to irregular swinging caused by light wind—its observation data may be statistically identified as an outlier, leading to tracking loss and being recorded as a missed detection. This strategy tends to reject ambiguous associations rather than forcing matches, thereby providing more reliable association results and ensuring that each green pepper can be tracked as a temporally consistent trajectory. This characteristic is particularly advantageous for robotic harvesting tasks, which depend on stable target coordinates for successful operation.

**Table 4.** Performance comparison between IoU-only and multi-feature fusion matching methods.

| Matching Method | Total Pairs | ID Switches | Association Accuracy (%) | Miss Rate (%) |
|---|---|---|---|---|
| IoU-only (θ=0.5) | 300 | 40 | 84.47 | 19.5 |
| Fusion Method | 300 | 18 | 92.48 | 22.67 |

2.4.2. Kalman Filter Modeling and State Update

2.4.2.1. State-Space Model Formulation

To smooth the target bounding box center coordinates (x, y) generated by the YOLO object detection model and estimate their motion states, a Kalman filter[23] is initialized for each newly detected target. The state vector is defined as:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



$$x_k=[x, y, v_x, v_y]^T \tag{7}$$

where x and y represent the center coordinates of the target at time k, and $v_x$,$v_y$ denote the corresponding velocity components. The state transition equation and the observation model equation are expressed as follows:

$$x_k=F*x_{K-1}+w_{k-1} \tag{8}$$

$$z_k=H*x_k+v_k \tag{9}$$

where $w_{k-1}$ and $v_k$ are zero-mean Gaussian process and observation noises with covariances Q and R. The state transition matrix F and observation matrix H are:

$$F=\begin{bmatrix}1 & 0 & \Delta t & 0\\ 0 & 1 & 0 & \Delta t\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix},H=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix} \tag{10}$$

2.4.2.2. Recursive Formulas of the Kalman Filter

The filter operates recursively in prediction and update steps:

First, prediction:

$$\hat{x}_k^-=F\hat{x}_{k-1}+w_k \tag{11}$$

$$P_k^-=FP_{k-1}F^T+Q \tag{12}$$

where $\hat{x}_k^-$ is the prior state estimate at time k, and $P_k^-$ is the prior error covariance matrix.

Subsequently, update:

$$K_k=P_k^-H^T\left(HP_k^-H^T+R\right)^{-1} \tag{13}$$

$$\hat{x}_k=\hat{x}_k^-+K(Z_k-H\hat{x}_k^-) \tag{14}$$

$$P_k=(I-KH)P_k^- \tag{15}$$

where $K_k$ is the Kalman gain, $\hat{x}_k$ is the posterior state estimate, and $P_k$ is the updated posterior error covariance matrix.

2.4.2.3. Parameter Initialization and Tuning Strategy

Upon YOLO detection of a new target, the Kalman filter state is initialized as:

$$\hat{x}_0=\begin{bmatrix}x_0\\ y_0\\ 0\\ 0\end{bmatrix} \tag{16}$$

$$P_0=\text{Diag}([10, 10, 1000, 1000]) \tag{17}$$

where $(x_0, y_0)$ is the first-frame detection coordinate. The diagonal values reflect higher initial uncertainty in velocity components.

The process noise covariance is defined as $Q=qI_4$. To determine the optimal q, we conducted tuning experiments in a natural breeze environment, using the first 60 frames of a representative video sequence. Five candidate values were evaluated: q={0.001, 0.01, 0.1, 1, 10}. The observation noise covariance was temporarily fixed at $R=\text{diag}([10, 10])$ for this stage.

Performance was assessed using two normalized metrics computed from the x-coordinate sequence:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



(1) Noise suppression ratio $R_{ns}$: This ratio serves as a core relative smoothness metric. A lower value indicates stronger noise suppression capability.

$$R_{ns}=\frac{\sigma_{kf}}{\sigma_{obs}} \tag{18}$$

(2) Normalized residual R_nr: Used to mitigate the impact of varying fluctuation levels across different observation sequences, ensuring comparability of results under different parameter settings even when initial observed data differ.

$$R_{nr}=\frac{\sigma_{res}}{\sigma_{obs}} \tag{19}$$

where $\sigma_{obs}$, $\sigma_{kf}$, and $\sigma_{res}$ are the standard deviations of the observed coordinates, Kalman-filtered estimates, and residuals, respectively.

The quantitative results are summarized in Table 5. As q increases from 0.001 to 10, $R_{ns}$ increases (smoothing decreases) while $R_{nr}$ decreases (tracking fidelity improves). A balanced trade-off is required: too small q (e.g., 0.001) yields over-smoothing with large lag ($R_{nr}$=0.9344), while too large q (e.g., 10) provides minimal filtering ($R_{ns}$=0.7728).

**Table 5.** Process noise covariance parameter q tuning results.

| q Value | Noise Reduction Ratio | Normalized Residual |
|---|---|---|
| 0.001 | 0.2437 | 0.9344 |
| 0.01 | 0.4658 | 0.7680 |
| 0.1 | 0.6198 | 0.7289 |
| 1 | 0.7271 | 0.5417 |
| 10 | 0.7728 | 0.3864 |

Among the candidates, q=0.01 and q=0.1 both offer a reasonable compromise. Visual inspection of the filtered trajectories (Figure 10) further confirms that q=0.01 provides smoother output while closely following the true motion trends. Therefore, q=0.01 is selected, yielding:

$$Q=0.01\cdot I_4 \tag{20}$$

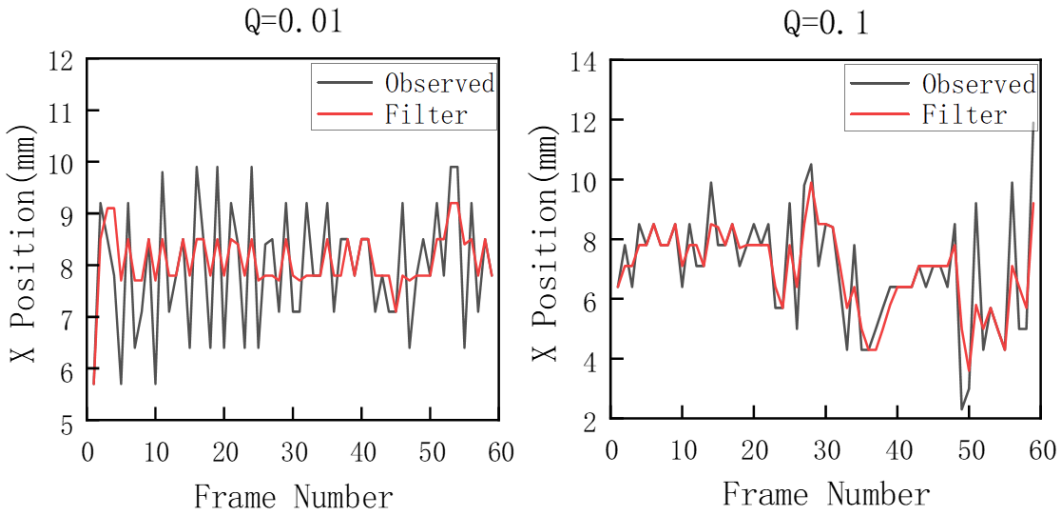



**Figure 10.** Kalman filter performance under different process noise parameters.

Following the same methodology, the observation noise covariance is defined as $R=rI_2$. Candidate values r={0.1, 1,5, 10, 50} were evaluated using the same 60-frame sequence, with Q now fixed at its optimal value $0.01\cdot I_4$.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



The performance metrics for each r are presented in Table 6. The parameter exhibits a nonlinear influence on filter behavior. While r=0.1 yields the lowest normalized residual ($R_{nr}$=0.5478), it provides insufficient smoothing ($R_{ns}$=0.7769). Conversely, r=5 leads to an abnormally high noise suppression ratio ($R_{ns}$=0.8843), indicating a mismatch between the constant-velocity model and actual target motion under breeze disturbances.

The value r=1 achieves the lowest noise suppression ratio ($R_{ns}$=0.5155) among all candidates, indicating the strongest smoothing effect, while maintaining an acceptable tracking fidelity ($R_{nr}$=0.7338). Therefore, r=1 is selected as the optimal observation noise parameter:

$$R=I_2 \tag{21}$$

**Table 6.** Process noise covariance parameter r tuning results.

| r Value | Noise Reduction Ratio | Normalized Residual |
|---|---|---|
| 0.1 | 0.7769 | 0.5478 |
| 1 | 0.5155 | 0.7338 |
| 5 | 0.8843 | 0.6074 |
| 10 | 0.7101 | 0.6972 |
| 50 | 0.5585 | 0.7842 |

The Kalman filter was applied to the complete experimental sequence using $Q=0.01\cdot I_4$ and $R=I_2$. As shown in Figure 11, the filtered trajectories in both X and Y directions exhibit significantly reduced high-frequency jitter while closely following the motion trends of the original observation data. This confirms the effectiveness of the Kalman filter in stabilizing two-dimensional target coordinates under occlusion and illumination variations.

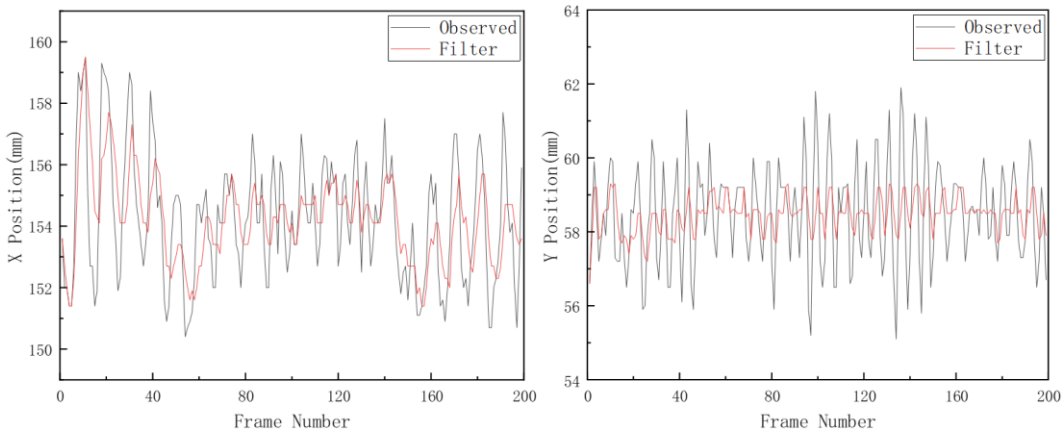



**Figure 11.** Trajectory comparison of the original observed and Kalman-filtered data in both X and Y coordinates.

The smoothed 2D coordinates are back-projected to 3D space using the RealSense depth camera. To further mitigate transient depth errors (e.g., due to leaf reflections or missing depth values, as illustrated in Figure 12), invalid depth readings are discarded, and a historical 3D position queue is maintained for each target ID, forming the input to subsequent outlier rejection and smoothing stages.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1




**Figure 12.** Examples of 3D localization anomalies.

2.4.3.3. σ Outlier Rejection and Single Exponential Smoothing

To enhance the reliability of 3D coordinate sequences, a two-step smoothing strategy is applied to the historical position queue $\{P_i\}$ of each tracked target.

First, statistical outliers are removed using the 3σ criterion[24]. The mean μ and standard deviation σ of the sequence are computed:

$$\mu=\frac{1}{N}\sum_{i=1}^{N}P_i \tag{22}$$

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i-\mu)^2} \tag{23}$$

Any point satisfying $|P_i-\mu|>3\sigma$ is discarded. The remaining inliers are averaged to obtain the fitted 3D point $P_t$ for the current frame:

$$P_t=\frac{1}{|\{P_i'\}|}\sum_{P_i'\in\{P_i'\}}P_i' \tag{24}$$

Subsequently, single exponential smoothing[25] is applied to further suppress single-frame jitter:

$$P_t^{smooth}=\alpha\times P_t+(1-\alpha)\times P_{t-1} \tag{25}$$

where $\alpha$=0.7 is the smoothing coefficient. This parameter governs the weighted fusion between the current observation and the historical trajectory, assigning predominant influence to the current data while retaining a portion of historical consistency for stability.

When the standard deviation of the target's smooth coordinates (X, Y, Z) in the last two consecutive frames is below 5 millimeters, the target is deemed spatially stable and added to the sampling queue. This provides high-precision, low-jitter spatial positioning input for subsequent robotic arm motion planning.

The overall algorithm flow for the aforementioned multi-target state management and coordinate smoothing filter is presented in the pseudocode below.

**Algorithm 1: Multi-feature fusion tracking and 3D Coordinate smoothing**

**Input:** color_image, depth_frame, depth_intri
**Output:** Smoothed 3D coordinates of tracked peppers

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



Initialize prev_boxes, current_boxes, kalman_filters, track_missing_count, xyz_histories, smoothed_history
Set IOU_TH, MAHALANOBIS_TH, LAMBDA_IOU, LAMBDA_M, max_frames_missing, window_len, $\alpha$
**while** camera stream is available **do**
    Run YOLOv8 on color_image, obtain detections D = { (box_k, conf_k) | conf_k ≥ IOU_TH } in (x, y, w, h) format
    Prune tracks whose missing count > max_frames_missing; let existing_tracks be keys of prev_boxes
    **if** existing_tracks ≠ ∅ and D ≠ ∅ **then**
        For each track_i ∈ existing_tracks and det_j ∈ D, compute IoU(track_i, det_j) in image plane
        Use Kalman prediction to get predicted center of track_i and Mahalanobis distance to det_j center
        If Mahalanobis distance > MAHALANOBIS_TH, set cost_{i,j} = +∞; else cost_{i,j} = LAMBDA_IOU·(1−IoU) + LAMBDA_M·Mahalanobis
        Apply Hungarian algorithm to the cost matrix to obtain globally optimal one-to-one matches
        For each matched pair (track_id, det_box) **do** Kalman prediction + update with det_box center, obtain (x_kf, y_kf, w, h)
            Set current_boxes[track_id] ← (x_kf, y_kf, w, h), track_missing_count[track_id] ← 0
        For unmatched tracks **do** Kalman prediction only, update current_boxes and increment track_missing_count
        For unmatched detections **do** create new track_id, initialize its Kalman filter with det_box center, add to current_boxes
    **else if** existing_tracks = ∅ and D ≠ ∅ **then**
        Initialize a new track and Kalman filter for each det_box ∈ D and fill current_boxes
    **end if**
    For each (track_id, box) ∈ current_boxes **do**
        Obtain (x_kf, y_kf) from box and query median_depth at (x_kf, y_kf) in depth_frame
        **if** median_depth is valid **then**
            Back-project (x_kf, y_kf, median_depth) via depth_intri to camera_xyz and append to xyz_histories[track_id] (kept at window_len)
            **if** |xyz_histories[track_id]| ≥ N **then**
                Compute mean μ and std σ of xyz_histories[track_id], discard points violating the 3σ rule to obtain inliers
                Let filtered_xyz be mean of inliers; apply single exponential smoothing
                    smoothed_xyz = $\alpha$·filtered_xyz + (1−$\alpha$)·previous smoothed_xyz (or filtered_xyz if no history)
                Update smoothed_history[track_id] with smoothed_xyz
            **end if**
        **end if**
    **end for**
    Set prev_boxes ← current_boxes; clear current_boxes
**end while**

*2.5. Hand-Eye Calibration*

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



To achieve vision-based grasping tasks, it is necessary to accurately transform the 3D coordinates of the target point from the camera coordinate system to the robot base coordinate system. In this experiment, the hand-eye calibration was performed in the "Eye-in-Hand" configuration to determine the transformation T_tool^camera between the camera and the robot flange. After calibrating the tool center point, multiple robot end-effector poses and their corresponding ArUco calibration plate poses were recorded (Figure 13). The classical AX=XB model was solved using both the Horaud[26] and Park[27] methods.

The accuracy of hand-eye calibration was validated by comparing the position of the vision-guided end-effector with ground truth measurements[28], where the Park method demonstrated slightly superior performance. Notably, the error along the Z-axis was consistently higher, attributable to the inherent depth measurement characteristics of the Intel RealSense D435i camera[29]. Consequently, the transformation matrix obtained via the Park method was selected for all subsequent experiments.

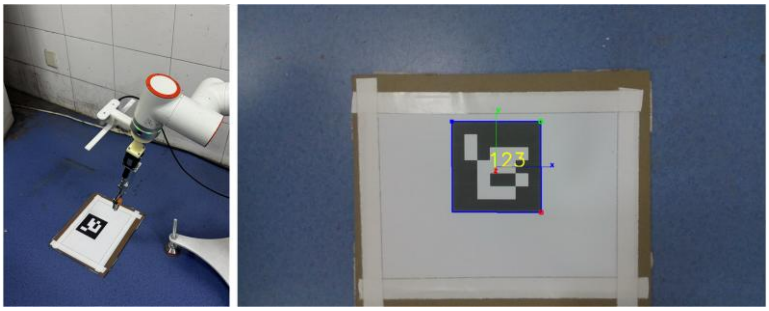



**Figure 13.** ArUco Calibration.

*2.6. Picking Task Planning and System Coordination*

After converting the smoothed target 3D coordinates into the robotic arm's base coordinate system, it is essential to further plan the picking sequence and generate safe and efficient motion trajectories. This paper proposes a depth-first-based picking sequence planning strategy that effectively reduces the risk of collisions between the robotic arm and overhead branches and foliage during movement.

To ensure overall system responsiveness and real-time performance, a multi-threaded software architecture is adopted to decouple visual perception from motion control. During robotic arm movement, the visual detection window remains active, but visual processing is temporarily suspended to avoid localization errors induced by arm motion. Furthermore, a target ID freezing mechanism is implemented to prevent repeated operations on fruits that have been successfully picked or have previously failed, thereby ensuring continuous and efficient task progression.

Precise execution is achieved through a segmented motion planning strategy integrated with coordinate compensation. This strategy defines a preset approach point P, positioned at a compensated offset (Z-axis: –100 mm, Y-axis: –20 mm) relative to the centroid of the visually estimated target bounding box, as illustrated in Figure 14. Upon reaching point P, the end-effector opens to a predefined width of 25 mm to accommodate variations in fruit stem dimensions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



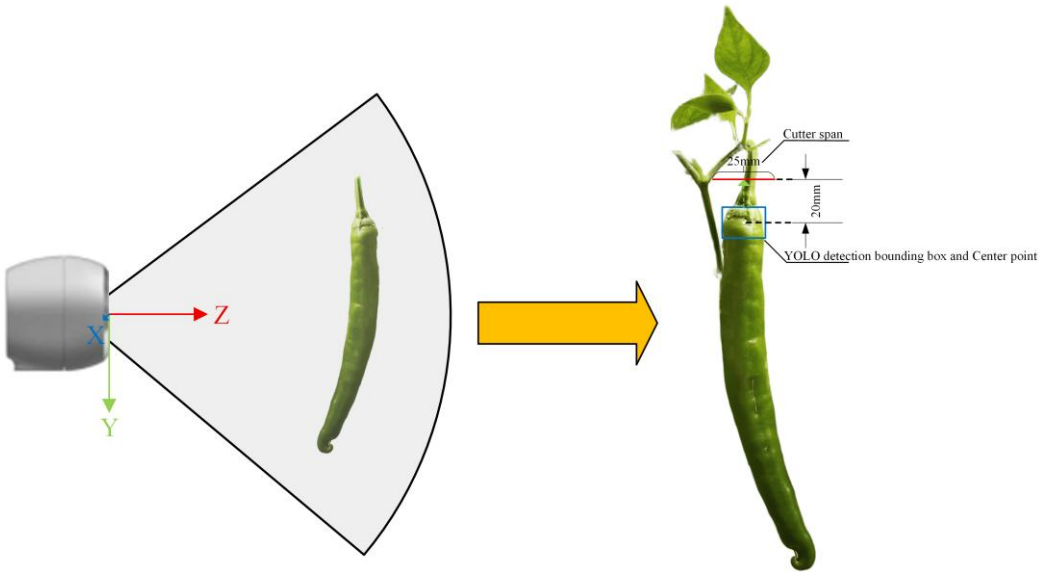



**Figure 14.** Schematic of the picking point determination strategy.

The harvesting cycle is controlled by a closed-loop event-driven protocol between the host computer and the STM32 microcontroller. During each harvesting operation cycle, the robotic arm first moves to the preset approach point P. Upon receiving the "robotic arm in position" signal, the microcontroller sends a command to drive the end effector to extend to the specified width. The robotic arm then advances along the negative Z-axis, allowing the extended end effector to wrap around the green pepper stem. After confirming the "robotic arm arrived" signal, the end effector closes to complete the grasping action. The robotic arm then retracts with the green pepper, transfers it to the unloading point for release (defined as the point 15 cm directly above the center of the green pepper basket), and finally returns to its initial position. This cycle continues until no new targets are detected for 20 consecutive frames, at which point the task is deemed complete (Figure 15).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



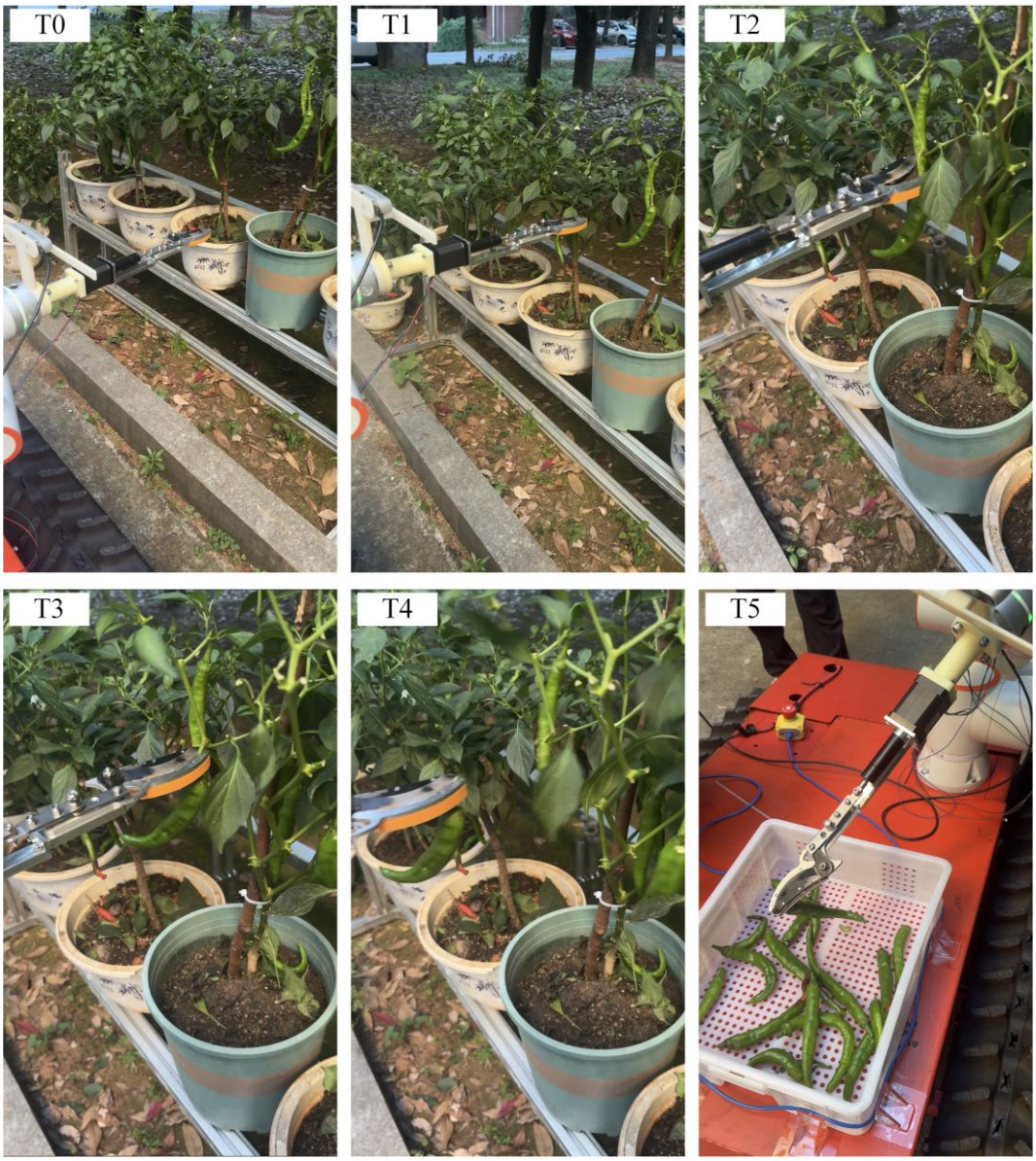



**Figure 15.** Multi-segment motion planning process for robotic systems. The arm is in the initial state (T0). The arm moves to the approach point (T1). It advances to the gripping point (T2). The end-effector cuts and grips the green pepper peduncle (T3). After gripping is complete it moves back (T4). The robotic arm moves to the unloading point and releases the green pepper (T5).

## 3. Results and Discussion

### *3.1. Experimental Setup*

The harvesting experiments were performed using a 6-DOF robotic arm equipped with the integrated cutting–gripping end-effector. The vision system consisted of an Intel RealSense D435i camera mounted in an eye-in-hand configuration. All perception and control algorithms were implemented in Python on an upper computer running Ubuntu 18.04, which communicated with the robotic arm via TCP/IP. Coordination with the STM32-based end-effector controller was achieved through relay-mediated I/O signaling for real-time status synchronization. An overview of the outdoor harvesting scenario with coordinate annotations is provided in Figure 16.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1




**Figure 16.** Harvesting test on potted pepper plants with annotated coordinate systems.

*3.2. Experimental Results*

To evaluate the overall performance of the designed pepper-harvesting robot, a total of 999 picking trials were conducted under clear and wind-free outdoor conditions using self-cultivated potted pepper plants. Before each picking operation, the system determined the occlusion level of the target pepper through visual assessment based on real-time image analysis, and the trial was then categorized into the corresponding occlusion class for subsequent statistical analysis. The experiment focused on the system's recognition and picking performance under different occlusion conditions. Detailed statistical results for each occlusion level are presented in Table 7.

Each occlusion category was evaluated according to three metrics:

Detection correct rate: the percentage of peppers correctly detected within that occlusion level.

Harvesting success rate: the percentage of successfully picked peppers among those correctly detected.

Plant damage rate: the percentage of successfully picked peppers that caused plant damage during the harvesting process.

For example, in the no occlusion case, 239 out of 261 green peppers were correctly detected (91.57%). Among these 239 green peppers, 198 were harvested successfully (82.85%), and 22 of those 198 caused plant damage (11.11%). Data for other occlusion levels were calculated using the same method.

The experimental results indicate that as the level of occlusion increases, both the recognition and picking performance of the system decline. Under no and light occlusion conditions, the detection rates were 91.57% and 80.29%, the picking success rates were 82.85% and 77.68%, and the plant damage rates were 11.11% and 34.48%, respectively. Under moderate and heavy occlusion conditions, the detection rates decreased to 72.13% and 46.51%, the picking success rates dropped to 57.39% and 40.00%, while the plant damage rates increased to 51.49% and 75.00%, respectively.

**Table 7.** Performance metrics under different occlusion levels.

| Occlusion Level | Count (pcs) | Detection Correct Rate (%) | Harvesting Success Rate (%) | Plant Damage Rate (%) |
|---|---|---|---|---|
| No occlusion | 261 | 91.57 | 82.85 | 11.11 |
| Light occlusion | 279 | 80.29 | 77.68 | 34.48 |
| Moderate occlusion | 244 | 72.13 | 57.39 | 51.49 |
| Severe occlusion | 215 | 46.51 | 40.00 | 75.00 |
| Average | 249.75 | 72.63 | 64.34 | 56.85 |

Throughout the experiment, the robotic arm consistently reached the preset picking points with precision, maintaining high-accuracy alignment between the end effector and pepper stems,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



validating the stability of the 3D positioning system. The closed-loop coordination between the vision system, robotic arm, and end effector operated without failure for over 2 hours. The average time distribution for a single harvesting cycle is shown in Figure 17: coordinate convergence (0.22 seconds), robotic arm movement to the pre-harvesting point (1.95 seconds), cutting and grasping (2.35 seconds), depositing and returning to the initial posture (5.28 seconds), ultimately achieving a total cycle time of 9.8 seconds per pepper fruit.

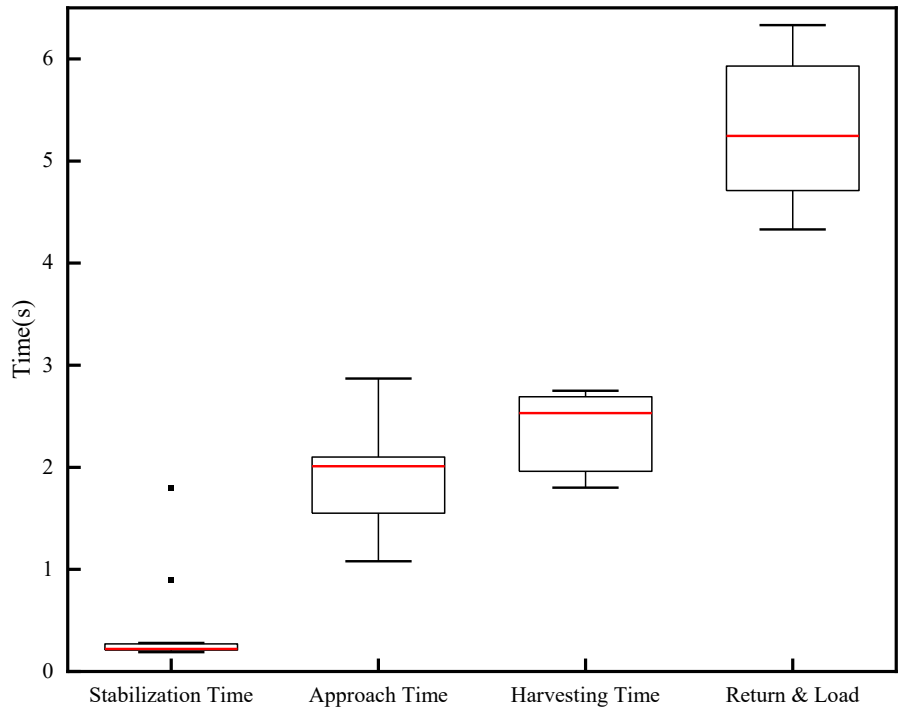



**Figure 17.** Time distribution of pepper harvesting stages.

*3.3. Discussion*

The proposed system demonstrates reliable recognition and picking performance under no occlusion conditions. However, performance degrades significantly as occlusion increases, revealing limitations of the system.

3.3.1. Perception Limitations

The perceptual challenge mainly stems from the limitation of the vision scheme. The system relies on a single view for recognition and picking. Such a restricted field of perception faces difficulties in achieving reliable identification in densely growing crops such as calyx–ectocarp junction, where heavy foliage frequently results in severe occlusion of green pepper. Moreover, the similarities in color and texture between green peppers and their surrounding stems and leaves complicate the reliable identification of key morphological features, leading to a high rate of missed detections under occluded conditions.

To overcome these limitations, future perception systems should transition from single-view perception to multi-view or multimodal global 3D sensing frameworks[30–34]. Such approaches can reconstruct a more complete three-dimensional crop scene model by collecting information from multiple perspectives, thereby enhancing feature visibility and localization accuracy under occluded conditions. Drawing on the concepts of global 3D modeling and multi-view fusion, multi-angle data acquisition and reconstruction of canopy spatial structure can enable robust perception of occluded green peppers, thereby improving the overall recognition success rate and positioning accuracy of the harvesting system.

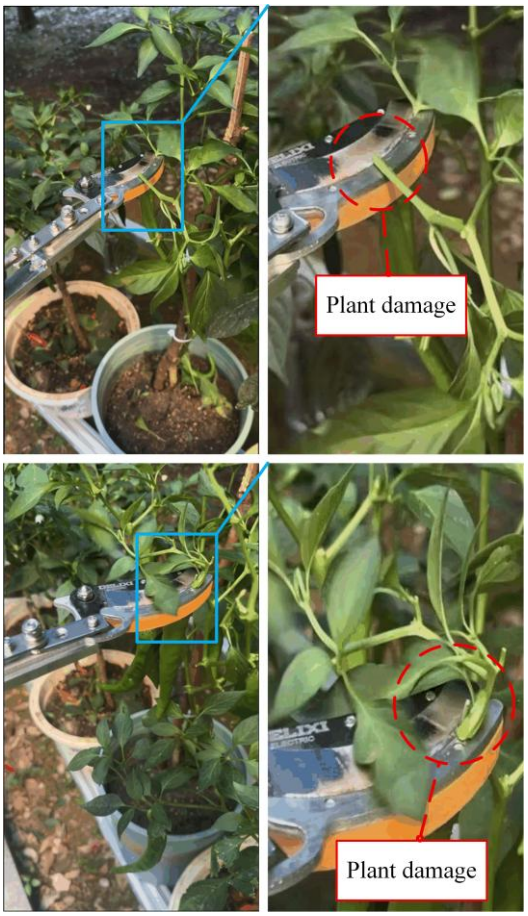
© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



3.3.2. Execution Limitations

Execution limitations under occlusion primarily stem from the current end-effector design and motion planning strategy. The relatively bulky and mechanically coarse structure of the end-effector, combined with its fixed opening width, often leads to entanglement with adjacent vegetation when multiple green peppers or stems are in close proximity (Figure 18). This not only causes collateral plant damage but also increases the risk of mis-cutting the target stem. Even under light occlusion, the damage rate remains as high as 34.48%. Improvements should be made in both the structural design of the end-effector and the optimization of operational parameters to enhance its selectivity and precision under occluded conditions. Based on the development of a miniaturized and refined end-effector, on one hand, the design of gripping units with adaptive deformation capability and real-time force sensing can be achieved by drawing on bio-inspired flexible structures and triboelectric sensing feedback technology. This would enable precise, low-damage grasping and separation in complex clustered environments[35]. On the other hand, by integrating a cutting dynamics model of multilayer composite materials with multi-objective parameter optimization methods, cutting parameters such as angle and speed of the integrated cutting-grasping mechanism can be systematically matched and dynamically adjusted. This approach helps reduce operational energy consumption and minimize plant damage[36]. Furthermore, during the approach phase to an occluded green pepper, unintended collisions with stem can displace the target stem or the green pepper itself (Figure 19), leading to harvesting failure even after successful initial localization. Future work could focus on enhancing the intelligent obstacle avoidance and path planning capabilities of picking robots in dynamic and unstructured environments[37]. On the one hand, sampling-based path planning methods could be further optimized by introducing target gravity mechanisms to improve search efficiency, combined with genetic algorithms for path post-processing[38]. On the other hand, strategies based on deep reinforcement learning could be explored, particularly by integrating recurrent neural networks to memorize and utilize historical state information, thereby achieving more efficient and robust obstacle avoidance decisions[39].




© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



**Figure 18.** Damaged harvesting.

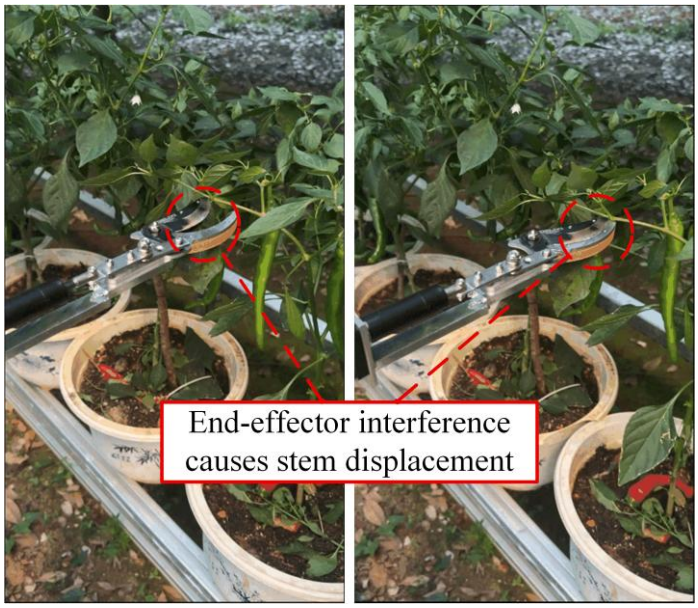



**Figure 19.** Harvesting failure.

Moreover, recent research on multi-arm coordination provides promising directions for overcoming occlusion-related challenges. Lammers et al. demonstrated successful apple harvesting using two highly coordinated robotic arms[40]. In contrast, the current single-arm system shows clear limitations in highly occluded and spatially constrained scenarios. Inspired by such dual-arm collaboration, a potential solution could involve deploying one robotic arm equipped with a compliant gripping mechanism to intelligently move obstructing branches and leaves, while another arm focuses on precise localization and harvesting of the target green pepper. Such cooperative task division is expected to significantly improve harvesting success rates and plant preservation in densely occluded environments.

## 4. Conclusions

This paper presents an integrated robotic system for automated green pepper harvesting, combining YOLOv8-based recognition, RealSense depth sensing, and a 6-DOF robotic arm with a cutting–gripping end-effector. A multi-stage filtering and tracking framework is proposed, which integrates IoU-Mahalanobis matching with the Hungarian algorithm for robust data association, Kalman filtering for trajectory smoothing, 3σ outlier rejection, and exponential smoothing for stable 3D coordinate estimation. Experimental results show that under no occlusion, the system achieves detection and picking success rates above 80%, with an average cycle time of 9.8 s per pepper fruit. While performance declines in highly occluded scenarios, the system demonstrates reliable coordination and continuous operation over extended periods. This work provides a practical modular solution for automated pepper harvesting and lays a foundation for future improvements in perception, execution, and multi-arm collaboration in complex agricultural environments.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1



**Funding:** This research was funded by the National Key R&D Program (2022YFD2002001), the Natural Science Foundation of Hunan Province (2025JJ60212), the Scientific Research Foundation of Hunan Provincial Education Department (24B0212) and the Graduate Research and Innovation Projects of Hunan Province (CX20251074).

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

**Acknowledgments:** We give special thanks to Feng Liu from the College of Horticulture, Hunan Agricultural University, for providing the test site and test materials. We thank Xuan Liu from the College of Horticulture, Hunan Agricultural University, for his assistance in the experiment.

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

## References

1. FAO. Production of Chillies and peppers, green (World and China). *FAOSTAT. Retrieved from http://www.fao.org/faostat* **2024**.
2. Lin, Z.; Chen, H. Labor Structure, Wage and Efficiency of Economic Growth. In Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, Nanchang, China, 26-27 August 2012; pp. 167-170.
3. Chen, Y.; Ren, T.; Li, Y.; Jiang, G.; Liu, Q.; Chen, Y.; Yang, S.X. AI-empowered intelligence in industrial robotics: technologies, challenges, and emerging trends. *Intell. Robot* **2026**, *6*, 1-18, doi:10.20517/ir.2026.01.
4. Tang, Y.C.; Chen, M.Y.; Wang, C.L.; Luo, L.F.; Li, J.H.; Lian, G.P.; Zou, X.J. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. *Frontiers in Plant Science* **2020**, *11*, doi:10.3389/fpls.2020.00510.
5. Yuan, J.; Fan, J.; Liu, H.; Yan, W.; Li, D.; Sun, Z.; Liu, H.; Huang, D. RT-DETR Optimization with Efficiency-Oriented Backbone and Adaptive Scale Fusion for Precise Pomegranate Detection. *Horticulturae* **2026**, *12*, 42, doi:10.3390/horticulturae12010042.
6. Du, P.C.; Chen, S.; Li, X.; Hu, W.W.; Lan, N.; Lei, X.M.; Xiang, Y. Green pepper fruits counting based on improved DeepSort and optimized Yolov5s. *Frontiers in Plant Science* **2024**, *15*, doi:10.3389/fpls.2024.1417682.
7. Huang, Y.K.; Zhong, Y.L.; Zhong, D.C.; Yang, C.C.; Wei, L.F.; Zou, Z.P.; Chen, R.Q. Pepper-YOLO: an lightweight model for green pepper detection and picking point localization in complex environments. *Frontiers in Plant Science* **2024**, *15*, doi:10.3389/fpls.2024.1508258.
8. Ji, W.; Gao, X.X.; Xu, B.; Chen, G.Y.; Zhao, D. Target recognition method of green pepper harvesting robot based on manifold ranking. *Computers and Electronics in Agriculture* **2020**, *177*, doi:10.1016/j.compag.2020.105663.
9. Jiang, H.K.; Liu, J.Z.; Lei, X.J.; Xu, B.C.; Jin, Y.C. Multi-stage fusion of dual attention mask R-CNN and geometric filtering for fast and accurate localization of occluded apples. *Artificial Intelligence in Agriculture* **2026**, *16*, 187-205, doi:10.1016/j.aiia.2025.10.005.
10. Jin, Y.C.; Liu, J.Z.; Wang, J.; Xu, Z.J.; Yuan, Y. Far-near combined positioning of picking-point based on depth data features for horizontal-trellis cultivated grape. *Computers and Electronics in Agriculture* **2022**, *194*, doi:10.1016/j.compag.2022.106791.
11. Yaojun, G.; Ping, M.; Wenmin, L.; Yuanshuang, M.; Changfei, G.; Runyu, L.; Lyuwen, H. Multi-label recognition of ripen persimmons varieties and phenotypic characteristics based on improved YOLOv8m [J] *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)* **2025**, *41*, 143-152, doi:10.11975/j.issn.1002-6819.202503128.
12. Pan, Q.; Wang, D.; Lian, J.; Dong, Y.; Qiu, C. Development of an Automatic Sweet Pepper Harvesting Robot and Experimental Evaluation. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13-17 May, 2024; pp. 15811-15817.
13. Alam, U.K.; Garcia, L.; Grajeda, J.; Haghshenas-Jaryani, M.; Boucheron, L.E. Automated Harvesting of Green Chile Peppers with a Deep Learning-based Vision-enabled Robotic Arm. In Proceedings of the 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Boston, MA, USA, 15-19 July, 2024; pp. 805-811.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1




14. Ravuri, S.P.; Allimuthu, S.; Ramasamy, K.; K, N.; K, V.; R, R. Automation in Agriculture: Capsicum Harvesting Using GRBL and Arduino-Driven Cartesian Robot. In Proceedings of the 2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi Campus, Centurion University of Technology and Management, Odisha., India, 19-21 December 2024; pp. 1-6.
15. Li, R.; Liu, Q.; Wang, M.; Su, Y.; Li, C.; Ou, M.; Liu, L. Maize Kernel Batch Counting System Based on YOLOv8-ByteTrack. *Sensors (Basel)* **2025**, *25*, doi:10.3390/s25175584.
16. Paul, A.; Machavaram, R.; Ambuj; Kumar, D.; Nagar, H. Smart solutions for capsicum Harvesting: Unleashing the power of YOLO for Detection, Segmentation, growth stage Classification, Counting, and real-time mobile identification. *Computers and Electronics in Agriculture* **2024**, *219*, doi:10.1016/j.compag.2024.108832.
17. Liang, Z.; Li, X.; Wang, G.; Wu, F.; Zou, X. Palm vision and servo control strategy of tomato picking robot based on global positioning. *Computers and Electronics in Agriculture* **2025**, *237*, doi:10.1016/j.compag.2025.110668.
18. Rapado-Rincón, D.; van Henten, E.J.; Kootstra, G. MinkSORT: A 3D deep feature extractor using sparse convolutions to improve 3D multi-object tracking in greenhouse tomato plants. *Biosystems Engineering* **2023**, *236*, 193-200, doi:10.1016/j.biosystemseng.2023.11.003.
19. Arlotta, A.; Lippi, M.; Gasparri, A. An EKF-Based Multi-Object Tracking Framework for a Mobile Robot in a Precision Agriculture Scenario. In Proceedings of the 2023 European Conference on Mobile Robots (ECMR), Coimbra, Portugal, 04-07 September, 2023; pp. 1-6.
20. Dai, N.; Fang, J.; Yuan, J.; Liu, X. 3MSP2: Sequential picking planning for multi-fruit congregated tomato harvesting in multi-clusters environment based on multi-views. *Computers and Electronics in Agriculture* **2024**, *225*, doi:10.1016/j.compag.2024.109303.
21. Mahalanobis, P.C. On the Generalized Distance in Statistics. *proceedings of the national institute of sciences* **1936**.
22. Kuhn, H.W. The Hungarian method for the assignment problem. *Naval Research Logistics* **1955**, *2*, 83-97, doi:10.1002/nav.3800020109.
23. Michaelis, S.S.P.A.A.-H.B. Intelligent feature-guided multi-object tracking using Kalman filter. *2009 2nd International Conference on Computer, Control and Communication* **2009**, doi:10.1109/IC4.2009.4909260.
24. Cui, W.T.; Yan, X.F. Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. *Chemometrics and Intelligent Laboratory Systems* **2009**, *98*, 130-135, doi:10.1016/j.chemolab.2009.05.008.
25. Karahasan, O.; Bas, E.; Egrioglu, E. A hybrid deep recurrent artificial neural network with a simple exponential smoothing feedback mechanism. *Information Sciences* **2025**, *686*, doi:10.1016/j.ins.2024.121356.
26. Horaud, F.D.R. Simultaneous robot-world and hand-eye calibration. *IEEE Transactions on Robotics and Automation ( Volume: 14, Issue: 4, August 1998)* **1998**, 617 - 622, doi:10.1109/70.704233.
27. Park, F.C.; Martin, B.J. Robot sensor calibration solving AX=XB on the Euclidean group. *IEEE Transactions on Robotics and Automation ( Volume: 10, Issue: 5, October 1994)* **1994**, 717 - 721, doi:10.1109/70.326576.
28. Wang, Q.; Wu, D.; Sun, Z.; Zhou, M.; Cui, D.; Xie, L.; Hu, D.; Rao, X.; Jiang, H.; Ying, Y. Design, integration, and evaluation of a robotic peach packaging system based on deep learning. *Computers and Electronics in Agriculture* **2023**, *211*, doi:10.1016/j.compag.2023.108013.
29. Condotta, I.; Brown-Brandl, T.M.; Pitla, S.K.; Stinn, J.P.; Silva-Miranda, K.O. Evaluation of low-cost depth cameras for agricultural applications. *Computers and Electronics in Agriculture* **2020**, *173*, doi:10.1016/j.compag.2020.105394.
30. Pan, Y.; Han, Y.; Wang, L.; Chen, J.; Meng, H.; Wang, G.; Zhang, Z.; Wang, S. 3D Reconstruction of Ground Crops Based on Airborne LiDAR Technology. *IFAC-PapersOnLine* **2019**, *52*, 35-40, doi:10.1016/j.ifacol.2019.12.376.
31. Li, X.; Liu, B.; Shi, Y.; Xiong, M.; Ren, D.; Wu, L.; Zou, X. Efficient three-dimensional reconstruction and skeleton extraction for intelligent pruning of fruit trees. *Computers and Electronics in Agriculture* **2024**, *227*, doi:10.1016/j.compag.2024.109554.


© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2026 doi:10.20944/preprints202603.0095.v1




32. Jiang, Y.; Zhao, S.; Liu, J.; Wu, S.; Jiang, Y.; Jin, Y. Review of dual-arm parallel and collaborative motion: methods, progress and applications in agriculture. *Computers and Electronics in Agriculture* **2025**, *239*, doi:10.1016/j.compag.2025.111081.
33. Chen, M.; Tang, Y.; Zou, X.; Huang, Z.; Zhou, H.; Chen, S. 3D global mapping of large-scale unstructured orchard integrating eye-in-hand stereo vision and SLAM. *Computers and Electronics in Agriculture* **2021**, *187*, doi:10.1016/j.compag.2021.106237.
34. Yang, D.; Cui, D.; Ying, Y. Object perception in sparse 3D point cloud scenes for floor-rearing chicken farming robots using an improved PointNet++ algorithm. *Computers and Electronics in Agriculture* **2025**, *237*, doi:10.1016/j.compag.2025.110773.
35. Sun, J.; Sun, L.; Zhao, G.; Liu, J.; Chen, Z.; Jing, L.; Cao, X.; Zhang, H.; Tang, W.; Wang, J. Triboelectric force feedback-based fully actuated adaptive apple-picking gripper for optimized stability and non-destructive harvesting. *Computers and Electronics in Agriculture* **2025**, *237*, doi:10.1016/j.compag.2025.110725.
36. Zhao, H.; Chen, Y.; Guo, J.; Liu, J.; Zhang, Z.; Zhang, X. Cutting dynamics modeling and parameter configuration optimization of rubber tree multi bark-layer composite system based on dynamic finite element method and quasi-static mechanical testing. *Computers and Electronics in Agriculture* **2026**, *240*, doi:10.1016/j.compag.2025.111165.
37. Zuoxun, W.; Chuanzhe, P.; Jinxue, S.; Guojian, Z.; Wangyao, W.; Liteng, X. Time-optimal trajectory planning for a six-degree-of-freedom manipulator: a method integrating RRT and chaotic PSO. *Intell. Robot* **2024**, *4*, 479-502, doi:10.20517/ir.2024.28.
38. Cao, X.; Zou, X.; Jia, C.; Chen, M.; Zeng, Z. RRT-based path planning for an intelligent litchi-picking manipulator. *Computers and Electronics in Agriculture* **2019**, *156*, 105-118, doi:10.1016/j.compag.2018.10.031.
39. Lin, G.; Zhu, L.; Li, J.; Zou, X.; Tang, Y. Collision-free path planning for a guava-harvesting robot based on recurrent deep reinforcement learning. *Computers and Electronics in Agriculture* **2021**, *188*, doi:10.1016/j.compag.2021.106350.
40. Lammers, K.; Zhang, K.; Zhu, K.; Chu, P.; Li, Z.; Lu, R. Development and evaluation of a dual-arm robotic apple harvesting system. *Computers and Electronics in Agriculture* **2024**, *227*, doi:10.1016/j.compag.2024.109586.


**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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.