2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Best-View Pedicel Localization with YOLO-DSC for Calyx-Preserving Robotic Harvesting of Cherry Tomatoes

Verianti Liana[1], Hao-Cheng Zuo[2], Yun-Chi Hsieh[2], and Ping-Lang Yen[1,2*]

***Abstract*—Robotic harvesting of cherry tomatoes remains challenging due to dense foliage, asynchronous ripening, and the strict market requirement for calyx-preserving cuts. The calyx frequently occludes the pedicel, making precise localization indispensable. In 640 × 480 images, pedicels span only 7–32 pixels, where even minor errors can lead to miscutting the calyx. To address this challenge, we apply YOLO-DSC to localize pedicels across dynamic frames as the arm-mounted camera moves during the best-view search. This strategy maximizes the visible pedicel length, exposing it perpendicularly to the camera and ensuring clear separation from the calyx, while null-data suppresses false positives from distractors such as leaves, stems, and calyces. In 15 autonomous trials along a 28 m greenhouse row, YOLO-DSC achieved the lowest pedicel localization errors, outperforming the YOLO baseline model ($p < 0.05$). This improvement directly translated into higher harvesting success, increasing from 47% with YOLO (including null data training) to 73.3% with YOLO-DSC. These results demonstrate that integrating YOLO-DSC with best-view searching enhances recall and stability under dynamic viewpoints, enabling more reliable calyx-preserving harvesting in real greenhouse conditions.**

I. Introduction

Robotic harvesting of small-sized fruits in cluttered greenhouse environments is a persistent challenge. The task requires robust perception under occlusion, uneven lighting, and visually similar distractors, combined with precise control to handle irregular fruit and pedicel growth. These challenges are especially pronounced in Taiwanese cherry tomato production, where market standards mandate calyx-retained harvesting. Retaining the calyx helps prolong freshness by covering the pedicel attachment point, reducing water loss, and delaying decay, while also meeting consumer preferences for visual aesthetics. In addition to this calyx requirement, harvesting is further complicated by the varied growth stages within a cluster. Unlike uniformly ripening crops, cherry tomatoes mature asynchronously, making selective, single-fruit harvesting essential (see Fig. 1).

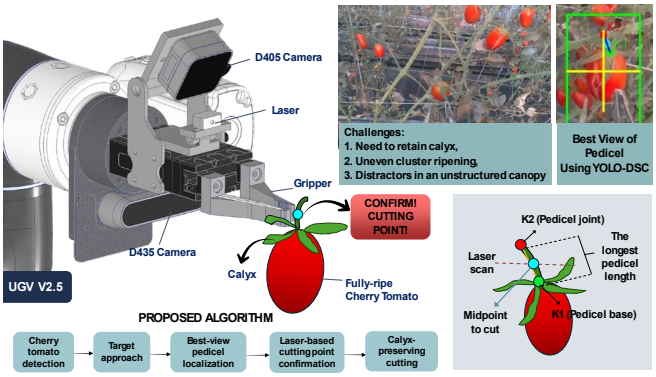

Figure 1. Conceptual illustration of an autonomous calyx-preserving cherry tomato harvesting robot

This biological asynchrony, together with the calyx-retention requirement, makes pedicel localization and precise cutting indispensable components of any viable robotic harvesting system. Despite substantial progress in agricultural robotics, most existing designs remain misaligned with these constraints. The majority of prior systems have emphasized cluster harvesting, an approach fundamentally incompatible with calyx preservation and selective picking. Moreover, their reliance on Position-Based Visual Servoing (PBVS) [1], [2], in which a static 3D pose is reconstructed before execution, has proven fragile: calibration drift, measurement errors, and occlusions often propagate through the open-loop control pipeline, resulting in frequent failures to cut at the correct pedicel location. By contrast, Image-Based Visual Servoing (IBVS) derives control directly from 2D image features, which reduces sensitivity to calibration errors. Yet, existing IBVS applications [3], [4], [5] in harvesting have rarely addressed the deeper challenge of dynamically adjusting viewpoints to resolve occluded pedicels and variable calyx orientations.

For calyx-preserving harvesting, the robot must accurately localize the pedicel and distinguish it from the calyx, even as the arm-mounted camera changes viewpoint. Because the calyx frequently grows directly in front of the pedicel, the apparent position of the pedicel in the image shifts with each angle. If localization is not continuously updated, the cut can easily be misplaced. Existing approaches, whether developed for clusters or individual fruits, often depend on segmentation or multi-keypoint detection [4], [5] from static viewpoints, and have not been validated in real-time harvesting robots.

[1] Global Agriculture Technology and Genomic Science Program, International College, National Taiwan University, Taiwan
[2] Department of Biomechatronic Engineering, National Taiwan University, Taiwan
[*] Corresponding Author: plyen@ntu.edu.tw
Research supported by the Ministry of Agriculture, Taiwan under grant 114AS-16.1.1-AS-04, and the National Science and Technology Council, Taiwan under grant NSTC 114-2218-E-002-030, and NSTC 112-2923-E-002-006-MY3

As a result, these methods are prone to failure when the pedicel is obscured or when distractors such as stems, leaves, pedicels attached to unripe fruits, supporting nets, or red trellis clips for tomato plants appear in the scene. End-effector designs based on suction, twisting, or simple cutting have also proven limited, either causing fruit damage, failing to retain the calyx, or being constrained to cluster harvesting [6], [7], [8]. In this work, we employ a cut-and-grab mechanism for detachment, with the primary focus on perception and viewpoint control rather than end-effector design.

To successfully harvest cherry tomatoes while preserving the calyx, we present a best-view searching strategy with YOLO-DSC detection supported with laser confirmation for precise cutting point (See Fig.1). Our key contributions are:

1. Best-view harvesting framework. We develop a single-fruit harvesting strategy that integrates best-view pedicel detection with Image-Based Visual Servoing (IBVS). Pedicel length in the image space is used as a proxy for viewpoint quality, guiding yaw and pitch adjustments until the pedicel is maximally visible and clearly separated from the calyx, thereby ensuring precise alignment for calyx-preserving cutting.
2. Pedicel localization with YOLO-DSC and laser confirmation. We enhance pedicel detection by integrating Dynamic Snake Convolution (DSC) into the YOLOv8l-Pose backbone to capture slender, curvilinear pedicels across dynamic viewpoints. Null-data training with distractor images is applied to suppress false positives from visually similar objects in greenhouse scenes. Once the pedicel is localized from the best-view, a compact laser module confirms the cutting point relative to the gripper, providing robustness against occlusion.
3. Real-world validation. We deploy the complete system on a harvesting robot and conduct real-time experiments in a commercial greenhouse, achieving reliable calyx-preserving tomato harvesting under natural, cluttered conditions.

II. RELATED WORKS

A. *Cherry Tomato Harvesting Robots*

Prior robotic systems for cherry tomato harvesting have mainly focused on cluster-oriented strategies, but these systems reveal clear limitations in both mobility and control. One system employed a 6-DOF manipulator mounted on a rail platform, combining stereo vision with a laser sensor to harvest entire fruit clusters [1]. Another design utilized a 7-DOF arm mounted on a lifting column with an RGB-D camera and a two-stage perception pipeline [9]. Both systems were constrained by fixed railway infrastructure, severely limiting navigational adaptability and often requiring human intervention to cross crop rows. In addition, their reliance on an open-loop "look-then-move" paradigm made them highly vulnerable to calibration drift and measurement errors, often causing the end-effector to miss the pedicel during the final approach [9]. A different approach introduced a twisting "holding-rotating" gripper to detach individual fruits, but this method damaged the calyx–pedicel junction and is therefore incompatible with markets requiring calyx preservation [10]. In contrast, our system resolves these limitations by combining the mobility of a wheeled autonomous robot with precise pedicel detection, closed-loop IBVS for continuous motion refinement, and a final laser scan for accurate positioning, enabling autonomous single-fruit harvesting with reliable calyx retention.

B. *Harvesting Control Strategies*

Harvesting control strategies for agricultural manipulators are generally classified into Position-Based (PBVS), Image-Based (IBVS), and hybrid approaches, each offering distinct trade-offs. PBVS relies on reconstructing a 3D model of the target, but it is sensitive to camera calibration errors, whereas IBVS computes control signals directly from 2D image features, offering greater robustness to such uncertainties. In cherry tomato harvesting, PBVS has been the dominant approach. This is seen in the work of [1], who used stereo vision to locate the fruit bunch, and [9] who calculated a 6D cutting pose from a fixed observation point. However, this open-loop strategy is vulnerable to accumulated errors, which can cause the end-effector to miss the pedicel during the final approach [9]. Recognizing these trade-offs, researchers have explored alternatives. A hybrid visual servoing method combined PBVS for coarse alignment with IBVS for final positioning to balance speed and precision [11]. While this acknowledges the need for high-accuracy final alignment, the initial open-loop phase is still vulnerable to positioning errors. Other advanced methods, such as the Deep Reinforcement Learning (DRL) [7] generate optimal trajectories but depend on an accurate initial 3D model and face challenges in transferring policies from simulation to reality. These limitations are more pronounced in calyx-preserving harvesting, a setting that these methods do not explicitly address and that requires precise localization of the cutting point under frequent occlusion. In real canopies, the randomly oriented calyx often blocks the pedicel, making static open-loop PBVS unreliable because a single viewpoint may be blocked. Our system addresses this challenge through a closed-loop IBVS framework with active viewpoint adjustment, enabling the manipulator to search for a clearer view of the pedicel before cutting. We further improve robustness by combining visual feedback with a laser sensor to refine depth precision near the cutting point.

C. *Locating the Harvesting Cutting Point*

Recent point localization strategies in robotic harvesting differ in both target granularity (from fruit clusters to individual fruits) and methodology, evolving from conventional vision to deep learning. Early cluster-harvesting systems used conventional vision to approximate the bunch but often collided with the main stem due to imprecision [1]. More recent deep learning methods have introduced more complex pipelines, often built around YOLO because of its speed and efficiency. For example, the "Tomatero" robot first detects ripe clusters with YOLOv5-4D and then applies instance segmentation to estimate the pedicel's 3D centroid [12]. Another system reconstructs a 3D model from seven keypoints generated by MTA-YOLACT, but remains

vulnerable to segmentation failure on thin pedicels and to depth estimation errors [5]. To enable selective harvesting, recent perception models have targeted individual pedicels. One method fits a curve from three detected key points [4], while another combines YOLOv8n-DDA with the Segment Anything Model to find pedicel centroids [13]. However, these perception-focused studies have not yet been deployed in physical harvesting robots, leaving their robustness unverified in dynamic conditions. Other approaches that target the fruit body rather than the pedicel, such as hybrid visual servoing, do not provide the precision needed for calyx-preserving harvesting [3]. A further limitation is their reliance on vision alone: once the pedicel is fully occluded, the cutting point cannot be determined, and the task fails [12]. To address this robustness gap, we present a system tailored for calyx-preserving single-fruit harvesting. The proposed two-stage framework detects fully ripe fruit and pedicels using YOLO-DSC, then confirms the cutting position with a laser before final harvesting. This laser-assisted fallback improves reliability when pedicel visibility is limited by severe occlusion.

III. Methodology

The harvesting system operates in three major steps: detection of fully ripe cherry tomatoes, search for the best view of the pedicel, and localization of the pedicel. Experiments were conducted on the UGV V2.5 mobile manipulator (102 × 58.2 × 84.75 cm) running ROS 2 on a Jetson AGX Orin, equipped with a TM5 robot arm and gripper. The differential-drive base is velocity-controlled via a real-time microcontroller unit. Two RGB-D cameras were used: an eye-in-hand RealSense D405 mounted on the gripper for close-range pedicel localization of the target tomato, and a RealSense D435 mounted below the gripper for wider-view tomato detection (horizontal camera-to-gripper tip distances: 10 cm for D405 and 18 cm for D435). Both streams were processed at 640 × 480 pixels; the depth Field of View (FoV) for both the D405 and D435 is 87° × 58°.

A. Fully ripe Cherry Tomato Detection

The initial stage of our pipeline detects fully ripe cherry tomatoes at a distance of 35-40 cm. At this range, the D435 provides a broader FoV than the D405, allowing the system to identify candidate tomatoes in the surrounding workspace (see Fig.2 (A)). The system retains detections with confidence ≥ 0.60 and selects the target as the fruit with the largest pixel area, corresponding to the fruit closest to the gripper. This fully ripe cherry tomato task was challenged by numerous red tomato trellis clips (10-15 per meter of row-plant, shown in Fig. 2(A)), leading to high False Positive (FP) rates. Since replacing these trellis clips is logistically impractical, we employed a null-dataset strategy to suppress FPs. We found that unannotated distractor images (red clips with other red objects) efficiently improved the YOLOv8l model's discriminative capability without altering its architecture. where $\lambda > 1$ is a penalty coefficient that places a greater weight on reducing the False Discovery Rate (FDR).

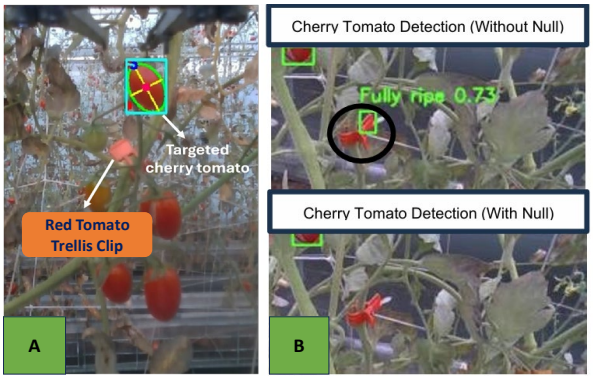

Figure 2. A.) Red clips surround the targeted cherry tomato, B.) The comparison of fully ripe cherry tomato detection with and without the null dataset addition

The optimal proportion of this null data, θ^*, is determined by maximizing an objective function that balances Precision against the FDR, thereby minimizing costly, FP-induced robot movements:

$$\theta^* = arg\max_{\theta}[\text{Precision}(\theta) - \lambda \cdot \text{FDR}(\theta)] \quad (1)$$

This optimization was performed through an iterative refinement process for both the tomato and pedicel detection models. To ensure each refinement was a significant improvement, a new null data proportion was accepted only when a one-tailed two-proportion Z-test at α=0.01 confirmed that the change led to both a significant reduction in FPs and a gain in precision. Evaluation was conducted on 570 testing images across 19 scenarios for tomato detection, and we found that a 12.3% null dataset proportion significantly increased tomato precision from 0.8958 to 0.9622 ($Z = +9.59$) and FDR decreased from 10.42% to 2.82% ($Z = -10.48$) with $p < 0.00001$. The model with a null dataset effectively avoids misidentifying a distractor as the targeted object, unlike the model without null dataset inclusion, as shown in Fig. 2 (B). Therefore, we trained a YOLOv8l model on 4,767 cherry tomato images with a 12.3% null dataset (300 epochs, batch size= 64, SGD optimizer, lr_0=0.0005, weight decay = 0.006).

B. Pedicel Localization using YOLO-DSC

Following the IBVS-controlled approach to a target cherry tomato, we perform pedicel detection at a 20-25 cm standoff distance to enlarge the target pedicel in the frame (see Fig.6) and localize the optimal cutting point under clutter, such as calyxes, leaves, peduncles, stems, and neighboring pedicels linked to unripe, semi-ripe, or absent fruits. To reliably extract the two key points defining the calyx and cutting point, the pedicel's length and angle, our work introduces a multifaceted solution. This method leverages a contextual-aware annotation, single-stage detector to identify the correct pedicel on ripe fruit directly, is robustly trained with a null-dataset strategy to reject incorrect targets actively, and integrates Dynamic Snake Convolution (DSC) [14] into the YOLOv8l-Pose backbone to fundamentally enhance sensitivity and recall for the target's slender, curvilinear geometry.

1. Contextual-Aware Annotation

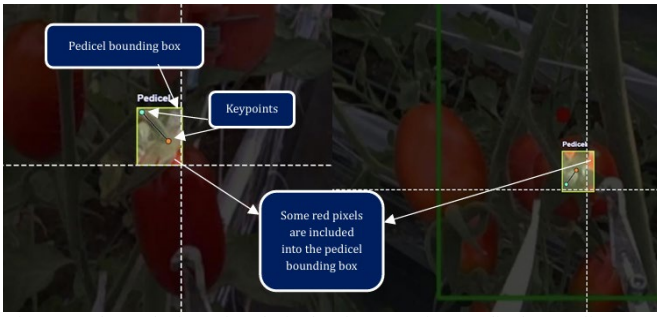

Figure 3. The Inclusion of Red Pixels of Tomato into Pedicel Bounding Box

We teach the model the direct association between a pedicel and its fully ripe cherry tomato by intentionally expanding the bounding box to include red pixels from the tomato, but the key points only connect the calyx and cutting point as shown in Fig.3. This approach enables our single-stage detector to localize pedicels on ripe cherry tomatoes directly, eliminating any need for post-processing.

2. Dataset Types with Null-Dataset Integration

In addition to the red-pixel expansion strategy for bounding boxes, our model is trained in a diverse range of dataset types, each designed to enhance specific aspects of pedicel detection. From 2,610 training images, there are seven types of datasets (see Fig.4), including a null dataset, for training our pedicel detection model. We incorporated temporal data to teach the model consistency across sequential frames with varying lighting and angles, and included isolated single tomato–pedicel images to train the model to focus on individual targets. To improve prioritization in cluttered scenes, the dataset features tomato clusters in which only the ripe target pedicel is annotated, along with images of multiple valid pedicels to increase true-positive rates. The model's robustness was further enhanced with examples of occluded or variably shaped pedicels from diverse and challenging angles. Crucially, a null dataset containing non-target objects, such as leaves and stems, was used to train the model to ignore distractors, which is essential for minimizing FPs during real-time operation.

Initial experiments with a baseline YOLOv8l-Pose model revealed that visually similar clutter, such as non-target stems, leaves, and pedicels on unripe fruit, was frequently misidentified. To address this, we first employed a null-dataset training strategy and compared the model's precision and FDR with and without a null dataset. We iteratively refined the proportion of unannotated distractor images to an optimal 8.3% and evaluated this approach on a test set of 360 images across more than 10 distinct scenarios. The impact was statistically significant; as confirmed by a one-tailed two-proportion Z-test, precision rose sharply from 0.8430 to 0.9718 ($Z = +10.47$), while the FDR plummeted from 15.70% to 2.82% ($p < 0.00001$). While this data-centric approach successfully suppressed false detections, it introduced a critical new challenge: a noticeable reduction in recall. The model, now highly precise, became overly conservative and frequently failed to identify valid pedicels.

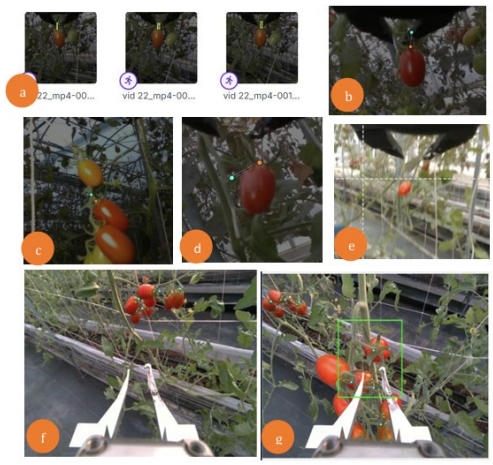

Figure 4. Annotated Dataset Types with blue and Orange Keypoints representing Pedicel Base (near calyx) and Pedicel Joint, respectively: a. Temporal information data, b. Single pedicel detection in a single tomato, c. Single pedicel detection in tomato clusters, d. Occluded pedicel detection, e. Null data, f. Multiple pedicels, g. Challenging angle of pedicels

3. Integration of YOLOv8l-Pose with Dynamic Snake Convolutional (DSC) layer

To overcome the low-recall issue, we adopted the Dynamic Snake Convolutional (DSC) layer [14] and integrated into the YOLOv8l-Pose backbone. This targeted intervention was designed to specifically enhance the model's sensitivity to the slender, curvilinear geometry of pedicels, thereby boosting the detection of true positives. We chose to modify the backbone to address detection failures at their foundational source. This upstream enhancement strategy ensures that pedicel-specific shapes are learned early, which significantly improves the quality of the feature flow to all subsequent layers. Critically, this prevents the neck from receiving and fusing noisy, low-quality inputs, a problem that cannot be rectified later in the network.

Unlike standard deformable convolutions, which allow sampling points to move arbitrarily, DSC imposes a geometric constraint that forces its kernel into a contiguous structure [14], enabling it to effectively adhere to the curvilinear geometry of pedicels. For a DSC kernel oriented along the x-axis, the coordinates of its sampling points are determined by a cumulative sum of learned vertical offsets:

$$K_{i\pm c} = (x_i \pm c, y_i + \sum_i^{i\pm c} \Delta y) \quad (2)$$

Where $K_{i\pm c}$ are the coordinates of a sampling point, (x_i, y_i) are the central coordinates of the convolution on the input feature map, c is the distance from the kernel's center along the x-axis, and Δy is the learned vertical offsets. Similarly, for a kernel oriented along the y-axis, a cumulative sum of learned horizontal offsets is applied to a vertical backbone:

$$K_{j\pm c} = (x_j + \sum_j^{j\pm c} \Delta x, y_j \pm c) \quad (3)$$

where the notations are analogous to the x-axis formulation, and Δx represents the learned horizontal offsets. Since the resulting coordinates are fractional, the output feature value at a deformed location is calculated using bilinear interpolation. The output feature map value f(K) at a deformed coordinate K is the weighted sum of the feature values at the neighboring integral grid locations:

$$f(K) = \sum_{K'} B(K', K) \cdot f(K') \quad (4)$$

where f(K) is the computed output feature value at the fractional coordinate K, K' enumerates all integral grid locations neighboring K, and B(K', K) is the bilinear interpolation weight.

Our integration inserts direction-aware DSC modules at two high-resolution stages to address the issue that standard backbones struggle with thin pedicels. Because the pedicel is slender, its weak visual cues can be suppressed during progressive downsampling. As shown in Fig. 5, we first place DSC at P2, the finest (highest-resolution) pyramid level, to preserve thin pedicel evidence and improve separation from background clutter before these cues are weakened [15]. We then extend DSC to P3, adding a second high-resolution level with stronger semantics; this multi-level design improves robustness and has been shown to benefit small-object recognition compared to relying on a single finest level [15]. At P3, our context-aware annotation further guides the model to leverage ripe-tomato cues and favor a continuous pedicel curve attached to the fruit. Overall, adding DSC increases the model capacity from 44.49M to 62.33M parameters, supporting the learning of subtle pedicel patterns.

We trained the YOLO-DSC model on 2,610 images with the following hyperparameters: epochs= 300, batch=4, image_size=512, optimizer=AdamW, learning rate= 0.0005, weight decay= 0.005, and flip= 0.5. The model's direct output is the pixel coordinates of two critical pedicel key points: K1 (pedicel base; near the calyx) and K2 (the first pedicel joint) as shown in Fig.6. We focus on the local pedicel segment (K1, K2) to support calyx-preserving, single-fruit picking in clusters where fruits ripen at different times; thus, the cut is defined near the calyx for the selected fruit, and branch level tracing is not required. From these coordinates, two essential geometric properties for closed-loop control are derived: the pedicel's apparent length and its orientation angle, which are detailed in the Best-View Searching Strategy part. The pedicel length ($L_{pedicel}$) is calculated as:

$$L_{pedicel} = \sqrt{(X_{K2} - X_{K1})^2 + (Y_{K2} - Y_{K1})^2} \quad (5)$$

C. Best-View Searching Strategy

In cluttered tomato canopies, the pedicel is often occluded by the canopy, depending on the in-hand camera's viewing angle, making it difficult to precisely and directly move the gripper for harvesting. Therefore, the harvesting strategy actively adjusts the robot's arm orientation (yaw ${}^{base}_{tool}R_z$) after detecting the tomato, so that the tool is aligned with the pedicel growth direction (See Fig. 7 (A)). This step increases the target's accessibility and prevents unsuccessful grasps due to misalignment.

First, the pedicel orientation in the image is represented by an angle $\gamma \in \mathbb{R}$ with respect to the camera frame's y-axis. The YOLO-DSC model predicts two pedicel keypoints: $P_{K1} = (X_{K1}, Y_{K1})$ and $P_{K2} = (X_{K2}, Y_{K2})$, in the image (Fig. 7(B)). The pedicel angle θ_γ in degree is estimated as:

$$\theta_\gamma = 90° - \arctan\left(\frac{Y_{K2} - Y_{K1}}{X_{K2} - X_{K1}}\right) \quad (6)$$

When $\theta_\gamma > 90°$, the smaller angle in the opposite rotational direction is used for gripper control to avoid unnecessarily large yaw motion.

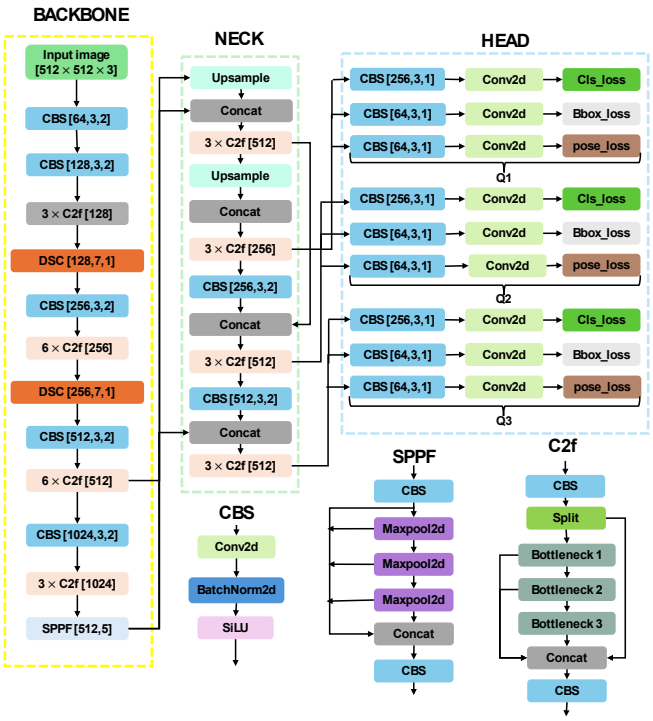

Figure 5. YOLO-DSC Architecture for Pedicel Detection

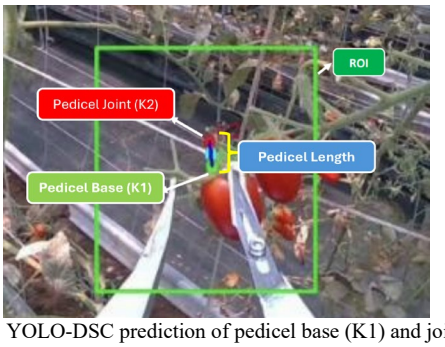

Figure 6. YOLO-DSC prediction of pedicel base (K1) and joint (K2)

In addition to aligning with the pedicel orientation, the best cutting angle requires the camera to face the pedicel directly. Geometrically, the observed pedicel length in the image depends on the camera viewing angle θ as:

$$L_{pedicel}(\theta) = L_{true} \cdot \sin(\theta) \quad (7)$$

where $L_{pedicel}(\theta)$ is the pedicel length observed in the image plane, L_{true} is the actual 3D pedicel length, and θ is the angle between the pedicel direction and the camera's optical axis. When the camera is parallel to the pedicel ($\theta = 0°$), the projection collapses to zero, while when the camera is orthogonal to the pedicel ($\theta = 90°$), the projection reaches its maximum:

$$L_{pedicel}(\theta) = L_{true} \quad (8)$$

Since the true viewing angle θ cannot be directly measured from an RGB image, we implement a search strategy to estimate it. Specifically, the pitch angle $\theta_\beta \in \mathbb{R}$ (up-down rotation) of the arm-mounted camera is adjusted incrementally (Fig. 8), and at each step, the pedicel length ($L_{pedicel}(\theta)$) is measured from the detected pedicel in the image. The pitch search performs a bidirectional sweep around the current pose within $[-10, +20]$. The sweep

follows $0 \rightarrow +20 \rightarrow -10 \rightarrow 0$. The sweeping procedure terminates when the sweep completes and returns to the starting pose, or earlier if a valid pedicel is detected. During the pitch sweep, we select the pose with the maximum observed pedicel length, and when detections are intermittent, we hold the last valid length value until the next detection (dashed segments in Fig. 9) as the final visual reference before the lateral laser sweep. Maximizing the observed pedicel length (best-view) ensures that the camera is perpendicular to the pedicel, providing the clearest separation from the calyx and enabling precise localization of the cutting point. This best-view control objective is therefore defined as:

$$\theta_\beta = \max_\theta L_{pedicel}(\theta) \tag{9}$$

After the best-view pose is selected, the detected pedicel keypoints are passed to the laser module to confirm the cutting point (P_{mid}). We define the predicted scan center (see Fig. 8 (b)) as the midpoint between the two keypoints:

$$P_{mid} = \frac{P_{K1}+P_{K2}}{2} = \left(\frac{X_{K1}+X_{K2}}{2}, \frac{Y_{K1}+Y_{K2}}{2}\right) \tag{10}$$

When the pedicel is detected, the upper-mounted laser is immediately aligned to scan around P_{mid} with a lateral sweep within ± 15 mm (Fig. 9). The narrow scan window around P_{mid} reduces the chance that the laser intersects leaves, branches, calyx, or neighboring pedicels. If the pedicel is not detected, the fallback strategy performs the same scan at a fixed position 12 mm above the targeted tomato center. By default, the system reports 100 cm when no object is detected, and the measured distance (d_{pedicel}= distance from laser to pedicel surface) decreases when the laser beam intersects plant structures. We record the distance profile $d_{\text{pedicel}}(t_y)$ along the sweep and select the lateral position with the global minimum distance as the pedicel location:

$$t_y{}^* = \min_{t_y} d_{\text{pedicel}}(t_y) \tag{11}$$

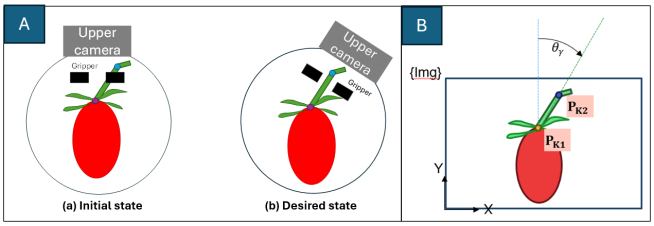

Figure 7. (A) Yaw adjustment of the arm-mounted camera from initial to desired state. (B) Pedicel orientation in the image, showing the angle derived from two detected keypoints

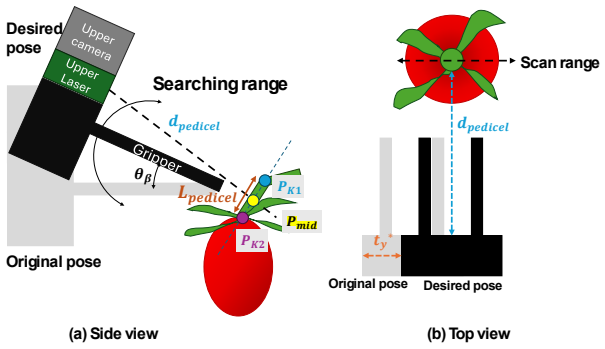

Figure 8. Searching for the best pitch angle and laser scanning for finding pedicel distance with (A) Side view and (B) Top view

D.Fine-Localization and Execution

Once t_y^* is obtained, the corresponding gripper pose ($^{base}_{gripper}T(t_y{}^*)$) is retrieved, and a point-to-point motion aligns the gripper to this pose. The gripper–cutter then advances along the stem direction and performs cutting while grasping the fruit. The laser depth (d_{pedicel}) is transformed into the gripper coordinate frame using calibration to position the cutter accurately for a reliable, calyx-preserving cut.

IV. RESULTS AND DISCUSSIONS

We evaluated our cherry tomato and pedicel detection model with the best-view searching strategy (YOLO-DSC) across 15 autonomous trials conducted along a 28-meter greenhouse row, where the robot executed the full harvesting pipeline: navigating between plants, detecting fully ripe fruits, localizing pedicels, and performing calyx-preserving cuts before moving to the next fruit. The average end-to-end cycle time was ~55 s per fruit (detection ~1 s, approach ~6 s, best-view search ~25 s (including yaw ~10 s and pitch ~15 s), laser sweep ~8 s, and cutting and placing ~15 s), demonstrating practical throughput under real greenhouse operation. Across 15 trials conducted during the daytime, the system achieved 11 successful harvests, yielding a 73.3% success rate. The YOLO-DSC model successfully detected all targeted cherry tomatoes and their pedicels. Failures occurred when tomatoes dropped after cutting or when the gripper inadvertently pushed the fruit, causing the pedicel position to shift and resulting in unsuccessful picking. These outcomes highlight that while perception and viewpoint selection were robust, further refinement of the end-effector is needed to ensure stable gripping and precise cutting.

The initial detection of fully ripe tomatoes was consistently successful across all trials, supported by null-dataset integration that suppressed FPs from red trellis clips and other distractors (Fig. 2). Operating in real time at 30 FPS, the pipeline was suitable for real-world deployment. Pedicel localization combined with YOLO-DSC with the best-view searching strategy (Equation (6)–(9)), enabling robust detection under dynamic viewpoints and maximizing pedicel length to expose the pedicel perpendicular to the camera for clear separation from the calyx. A representative case (Trial 10, Fig. 9) illustrates how pedicel detection guided yaw alignment and subsequent pitch adjustment to achieve the global best view. For comparison, the YOLO model with null-data training was deployed, as the model without null data produces high FPs, as discussed in Section III-B-2.

Although the full best-view search lasts about 25 s (10s yaw alignment and 15s pitch), Fig. 10 shows a selected 237-frame segment (frames 613-850) of the best-view search to illustrate the dynamic change in projected pedicel length. The YOLO-DSC successfully detected the pedicel in 77% of frames, even under pitch variation spanning $-10°$ to $+20°$, as indicated by the predominance of solid blue lines in Fig.9. Missed detections were bridged by reusing the last valid measurement (dashed segments).

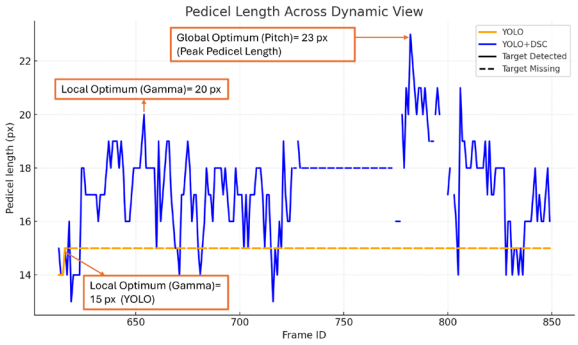

Figure 9. Comparison of YOLO with the YOLO+DSC model in detecting pedicel length across a dynamic view

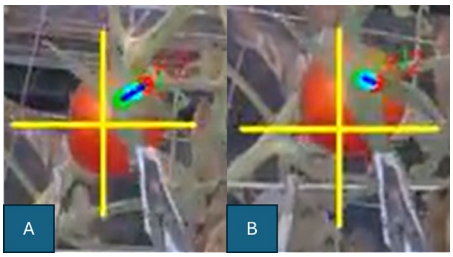

Figure 10. (A) The best view of pedicel detection. (B) non-best-view of pedicel detection

In contrast, the YOLO (trained with null data) detected the pedicel in only 2% of frames, was limited to the initial viewpoint, and was dominated by dashed lines, demonstrating poor robustness to angular changes. Importantly, the pixel-wise fluctuations observed in YOLO-DSC reflect active re-detection rather than noise, underscoring its stability in continuously updating the pedicel estimate across different angles during best-view searching. At the end of the pitch process, the system outputs the longest pedicel length and its K1–K2 coordinates to guide the cutting point. We used this frame moment before pedicel parameters passed to the laser, annotated the pedicel, and calculated its length using Equation (5) as the Ground Truth (GT). Compared with the GT of pedicel length in the selected trial (Fig.9), which is 27 pixels, YOLO-DSC achieved a global maximum of 23 pixels (error 14.8%), substantially closer than the YOLO with null data, which plateaued at 15 pixels (error 44.4%) that could not reach its global optimum. This improvement highlights how yaw alignment first establishes a local optimum (20 pixels with YOLO-DSC vs. 15 pixels with YOLO) and then enables pitch searching to reach the global best view. The effect is further illustrated in Fig. 10: under the best-view configuration (A), the pedicel is clearly separated from the calyx, whereas in a non-optimal view (B), the shortened pedicel length leads to ambiguous separation and potential cutting errors.

After identifying the longest pedicel length, the robot gripper was oriented according to the pedicel angle derived during the pitch sweep (see Equation (6)). The laser then performed a horizontal scan centered on this axis, traversing the midpoint between the detected keypoints K1 and K2 (Fig. 8). This midpoint serves as the reference for confirming the cutting location: the laser executes a ±15 mm sweep, and the global minimum distance is taken as the pedicel position. As shown in Fig.10(B), shorter detected pedicels shift the midpoint closer to the calyx, increasing the risk of missed or damaging cuts. This motivates the best-view search to maximize pedicel length before passing the midpoint to the laser.

Following the GT definition above, we benchmarked the predicted pedicel lengths (P in Table 1) from YOLO-DSC with null-data integration, YOLO with null data, and YOLO without null data against the corresponding GT values across the 15 trials. For safety reasons, the model trained without null data was not deployed for real-time harvesting, as Section III-B-2 showed it produced unstable and unsafe detections. Instead, it was tested offline on identical real-time frame sequences to YOLO-DSC, confirming that without null data, it failed in 40% of cases (Table 1), often localizing distractors such as stems or calyces and exhibiting unstable frame-to-frame switching that could dangerously disrupt arm motion.

As summarized in Table 1, null-data training successfully suppressed FPs but at the cost of reduced recall and sensitivity (Fig.9). YOLO with the null dataset avoided false targets but frequently underestimated pedicel length, with seven cases shifting the midpoint toward the calyx and one case (Trial 11) failing. By contrast, YOLO-DSC with null consistently restored recall and stability, actively updating pedicel length predictions across dynamic viewpoints (Fig. 9). It achieved the lowest error across all trials (Mean Average Error (MAE) = 5 px, Root Mean Square Error (RMSE) = 7 px), outperforming YOLO + null (MAE = 8 px, RMSE = 11 px) and YOLO without null (MAE = 13 px, RMSE = 15 px). Importantly, YOLO-DSC detected pedicels in all 15 trials, preserving the precision gains of null-data training while substantially improving sensitivity through higher true-positive detections.

After the best view was identified, the laser sweep was executed as described in Section III-C to confirm the cutting point (Fig. 11). When pedicel detection was available, the sweep was centered at the predicted midpoint, and the pedicel was confirmed immediately at the start of the scan, as indicated by the green trace. The repeated low-distance readings near the end of the trace correspond to the return pass and show consistent intersection with the same structure. This reduced the confirmation time to 7.4 s. In contrast, when pedicel detection was unavailable, the fallback scan above the targeted tomato more frequently intersected non-pedicel structures, such as calyx, leaves, or stems, resulting in a longer confirmation time of 10.6 s (orange trace). Overall, integrating YOLO-DSC pedicel detection reduced laser confirmation time by nearly one-third and reduced false targets, improving both efficiency and reliability.

Pedicel detection in a greenhouse is particularly challenging, as the visible length can range from only 7 to 32 pixels within a 640×480 frames (shown in Table 1). Despite this difficulty, our YOLO+DSC with null-data integration

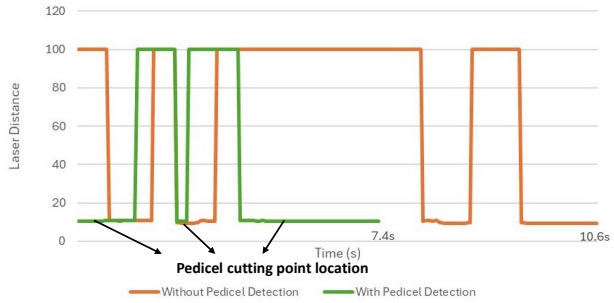

Figure 11. Comparison of laser distance during pedicel localization with detection and without detection (fallback strategy)

Table 1. Comparison of the best-view pedicel length with and without DSC

	The Longest Pedicel Length during Best-View (pixel)					
Trial (n)	**YOLO (+Null data) +DSC**		**YOLO(+null data)**		**YOLO**	
	GT	**P**	**GT**	**P**	**GT**	**P**
1	27	16	26	10	32	13
2	22	14	26	13	25	11
3	22	15	24	10	20	10
4	19	15	20	20	27	27
5	22	21	22	25	22	0 (FP)
6	21	18	24	22	21	0 (FP)
7	22	21	22	27	22	0 (FP)
8	7	24	8	15	10	18
9	17	16	17	20	15	17
10	20	23	25	15	20	0 (FP)
11	23	25	25	0 (not detected)	23	0 (FP)
12	25	23	25	25	24	28
13	23	21	25	27	24	27
14	22	30	23	14	25	23
15	16	21	16	19	16	FP
MAE	**5 px**		**8 px**		**13 px**	
RMSE	**7 px**		**11 px**		**15 px**	

achieved precise and stable localization. To rigorously assess pedicel localization performance, we applied paired t-tests and Wilcoxon signed-rank tests on per-trial errors, both appropriate for small-sample paired comparisons. As summarized in Table 1, YOLO + DSC obtained the lowest error (MAE = 5 px, RMSE = 7 px), outperforming YOLO+null (MAE = 8 px, RMSE = 11 px) and baseline YOLO without null (MAE = 13 px, RMSE = 15 px). Statistical analysis confirmed significant improvements over baseline YOLO (without null data, $p < 0.05$), while improvements over YOLO+null showed a near-significant trend ($p = 0.077$ Wilcoxon). Importantly, even a few pixels of error can result in millimeter-scale deviations at the pedicel, often shifting the midpoint toward the calyx and increasing the risk of failed or damaging cuts. This was reflected in real-world harvesting: YOLO+null achieved only 47% success, while YOLO+DSC improved to 73.3% across 15 trials. These findings validate that integrating YOLO+DSC with null-data training increases recall, enhances stability under dynamic viewpoints, and directly improves harvesting reliability for calyx-preserving cherry tomato picking in real-world environment.

REFERENCES

[1] Q. Feng *et al.*, "Design and test of robotic harvesting system for cherry tomato," *Int. J. Agric. Biol. Eng.*, vol. 11, no. 1, pp. 96–100, 2018, doi: 10.25165/j.ijabe.20181101.2853.

[2] J. Rong, P. Wang, T. Wang, L. Hu, and T. Yuan, "Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots," *Comput. Electron. Agric.*, vol. 202, p. 107430, Nov. 2022, doi: 10.1016/j.compag.2022.107430.

[3] P. Li, M. Wen, Z. Zeng, and Y. Tian, "Cherry Tomato Bunch and Picking Point Detection for Robotic Harvesting Using an RGB-D Sensor and a StarBL-YOLO Network," *Horticulturae*, vol. 11, no. 8, p. 949, Aug. 2025, doi: 10.3390/horticulturae11080949.

[4] J. Qin, Z. Chen, Y. Zhang, J. Nie, T. Yan, and B. Wan, "YOLO-CT: A method based on improved YOLOv8n-Pose for detecting multi-species mature cherry tomatoes and locating picking points in complex environments," *Measurement*, vol. 254, p. 117954, Oct. 2025, doi: 10.1016/j.measurement.2025.117954.

[5] Y. Li *et al.*, "MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting," *Eur. J. Agron.*, vol. 146, p. 126812, May 2023, doi: 10.1016/j.eja.2023.126812.

[6] Q. Pan, D. Wang, J. Lian, Y. Dong, and C. Qiu, "Development of an Automatic Sweet Pepper Harvesting Robot and Experimental Evaluation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan: IEEE, May 2024, pp. 15811–15817. doi: 10.1109/ICRA57147.2024.10610866.

[7] Y. Li *et al.*, "Peduncle collision-free grasping based on deep reinforcement learning for tomato harvesting robot," *Comput. Electron. Agric.*, vol. 216, p. 108488, Jan. 2024, doi: 10.1016/j.compag.2023.108488.

[8] J. Jun, J. Kim, J. Seol, J. Kim, and H. I. Son, "Towards an Efficient Tomato Harvesting Robot: 3D Perception, Manipulation, and End-Effector," *IEEE Access*, vol. 9, pp. 17631–17640, 2021, doi: 10.1109/ACCESS.2021.3052240.

[9] J. Rong *et al.*, "Decoupled motion planning method for 7-DOF manipulator and lifting joint in automated tomato harvesting," *Comput. Electron. Agric.*, vol. 237, p. 110693, Oct. 2025, doi: 10.1016/j.compag.2025.110693.

[10] F. Zhang *et al.*, "Research on Flexible End-Effectors with Humanoid Grasp Function for Small Spherical Fruit Picking," *Agriculture*, vol. 13, no. 1, p. 123, Jan. 2023, doi: 10.3390/agriculture13010123.

[11] H. Li *et al.*, "Image moments-based visual servoing control of bagged agricultural materials handling robot," *Int. J. Agric. Biol. Eng.*, vol. 16, no. 1, pp. 212–219, 2023, doi: 10.25165/j.ijabe.20231601.7050.

[12] J. Rong, L. Hu, H. Zhou, G. Dai, T. Yuan, and P. Wang, "A selective harvesting robot for cherry tomatoes: Design, development, field evaluation analysis," *J. Field Robot.*, vol. 41, no. 8, pp. 2564–2582, Dec. 2024, doi: 10.1002/rob.22377.

[13] G. Zhang *et al.*, "YOLOv8n-DDA-SAM: Accurate Cutting-Point Estimation for Robotic Cherry-Tomato Harvesting," *Agriculture*, vol. 14, no. 7, p. 1011, Jun. 2024, doi: 10.3390/agriculture14071011.

[14] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 6047–6056. doi: 10.1109/ICCV51070.2023.00558.

[15] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.