INTEGRATED YOLO11-BASED VISION AND FOUR-DOF ROBOTIC MANIPULATOR FOR AUTOMATED POST-HARVEST CHERRY GRADING

Thanh-Lam Bui[1], Tuan-Anh Vu[1,*], Vu-Luu Hai[1], Dinh-Hieu Phan[1], Van-Nghia Le[1], Duc-Quang Nguyen[1]

DOI: https://doi.org/10.57001/huih5804.2026.004

ABSTRACT

Automated post-harvest cherry grading is implemented through an integrated vision-manipulation pipeline that couples a YOLO11s detector with a four-degree-of-freedom robotic arm for four-class ripeness sorting. A dedicated dataset is collected with 1,200 images and 2,184 annotated instances, and it is split into training (70%), validation (15%), and test (15%) subsets. On the test set, YOLO11s achieves 94.8% precision, 97.3% recall, and 95.7% mAP50 while running at 67 FPS on an RTX 4060. The mAP50 exceeds YOLOv8s and YOLOv10s by 2.3 and 1.6 points, respectively. The robotic subsystem employs an RGB-D camera and hand-eye calibration to map detections to robot targets, and fifth-order polynomial trajectories are used for pick-and-place execution. After calibration, the mean positioning error is 2.5 ± 0.9mm. Inference and planning require 43ms in total, whereas a complete pick-place cycle averages 2.8 ± 0.3s. End-to-end experiments on 200 cherries obtain an 89.5% success rate (179/200), indicating practical feasibility for small-scale post-harvest sorting.

Keywords: *YOLO11, four-DOF robot, computer vision, cherry classification, smart agriculture, deep learning.*

[1]School of Mechanical and Automotive Engineering, Hanoi University of Industry, Vietnam

[*]Email: anhvt_ck@haui.edu.vn

Received: 27/10/2025

Revised: 14/01/2026

Accepted: 28/01/2026

1. INTRODUCTION

Vietnamese agriculture is under pressure to adopt digital tools and automation across production stages. Post-harvest grading in many facilities is still carried out by hand, which raises labour costs and leads to variable productivity [1]. Artificial-intelligence methods and robotic systems are therefore being explored as practical options to ease these constraints in agricultural production [2].

Within computer vision for agriculture, many studies rely on deep neural networks to recognise fruits and to assess their quality. Badgujar et al. surveyed YOLO applications in this context and reported that one-stage detectors offer short processing times while keeping acceptable accuracy for field and post-harvest tasks [3]. Zhao et al. examined deep-learning-based object detection and noted that real-time architectures are now widely adopted in industrial environments [4]. Alif and Hussain traced the evolution from YOLOv1 to YOLOv10 and described how later versions increase network depth and refine loss functions to stabilise training [5]. Khanam and Hussain analysed the YOLO11 architecture in detail, including the C3k2 backbone block and the C2PSA attention mechanism designed to improve the handling of small or partly occluded objects [6]. Sapkota et al. compared YOLOv8 to YOLOv12 for apple detection in real orchards and showed that YOLO11 provides a balanced trade-off between accuracy and inference speed in outdoor agricultural scenes [7].

Parallel to advances in perception, robotic manipulators have been developed for harvesting and sorting tasks. Hu et al. presented a four-degree-of-freedom manipulator for apple picking and optimised joint trajectories to shorten each harvesting cycle [8]. Yoshida et al. designed a dual-arm harvesting robot that offers flexible motion but requires a mechanically complex and costly structure [9]. Dewi et al. built a fruit-sorting robot that groups products by colour and size, yet the working volume of the system remains limited for larger processing lines [10]. In the hand-eye calibration problem, Zhang et al. proposed a method based on a time-of-flight camera that achieves positioning accuracy suitable for fruit-picking robots [11], while the classical algorithm of Tsai and Lenz continues to be used as a reference solution in many robotic systems [12].

Work that focuses on cherries is still relatively rare. Gai et al. developed an improved YOLOv4-based algorithm for cherry detection and reported high recognition accuracy, but the processing speed did not yet satisfy strict real-time requirements and the dataset was collected under non-Vietnamese growing conditions, which may limit direct transfer to local farms [13]. The existing literature therefore leaves three issues open. YOLO11 has not been systematically evaluated on cherry images recorded under Vietnamese conditions. Most current solutions treat perception and manipulation as separate modules instead of an integrated system. The high price of commercial manipulators also remains a barrier for small and medium-sized enterprises.

The present study addresses these gaps by building a cherry image dataset adapted to local production, benchmarking YOLO11 against strong YOLO baselines, designing a cost-conscious four-degree-of-freedom manipulator that is tightly coupled with the vision module, and assessing the end-to-end performance of the combined system in a laboratory setting that approximates real post-harvest operations.

2. METHODS

A stratified split with random seed 42 divides the images into training, validation and test sets in a 70%-15%-15% ratio. During training, horizontal flipping, in-plane rotation, Hue-Saturation-Value color space (HSV) adjustment and mosaic augmentation are applied to increase appearance diversity and reduce overfitting. Table 1 lists the number of images and objects in each subset, and Table 2 summarises the class distribution: Mature cherries account for roughly one third of all instances, Immature and Medium each for about one quarter, and the General class for around 13%, with similar proportions across all three splits.

Table 1. Statistics of the cherry dataset by subset

Subset	Images	Objects	Share of images (%)	Subset
Training	840	1,529	70.0	Training
Validation	180	327	15.0	Validation
Test	180	328	15.0	Test
Total	1,200	2,184	100.0	**Total**

Note: object counts refer to annotated cherry instances in each subset.

Table 2. Detailed class distribution of cherries in the dataset

Cherry class	Training	Validation	Test	Total
Immature (green)	412 (26.9%)	89 (27.2%)	88 (26.8%)	589 (27.0%)
Medium (partly ripe)	381 (24.9%)	81 (24.8%)	82 (25.0%)	544 (24.9%)
Mature (ripe)	534 (34.9%)	115 (35.2%)	116 (35.4%)	765 (35.0%)
General (border-line)	202 (13.2%)	42 (12.8%)	42 (12.8%)	286 (13.1%)
Total	1,529 (100%)	327 (100%)	328 (100%)	2,184 (100%)

Note: percentages are computed within each column and rounded to one decimal place.

2.1. Detection model architecture

The main detector is the YOLO11s network, a small configuration of YOLO11 designed for real-time inference. The architecture comprises three components backbone, neck, and head [6]. The backbone incorporates the C3k2 block, which increases representational capacity while keeping the parameter count in a compact range. Several stages in the backbone integrate the C2PSA attention mechanism so that the model can focus more strongly on small cherries and partially occluded fruits.

The neck aggregates multi-scale features through a modified PANet structure. At the output, the detection head employs depthwise convolutions. This design reduces computational cost and supports high frame rates without markedly affecting accuracy.

For comparison, YOLOv8s and YOLOv10s are selected as baseline one-stage detectors. All three "small" variants (around 9 - 11 million parameters) are trained on the same dataset using exactly the same input preprocessing and augmentation pipeline. The loss function follows the default configuration of the corresponding implementations and consists of a box regression term, a classification term, and an objectness (confidence) term. High-level training settings such as optimiser type, learning rate schedule, and weight decay are reported in Section 3.1 and are shared across all models.

2.2. Design of the four-DOF manipulator

The manipulator adopts an R-R-R-R serial structure with four revolute joints. Joint 1 rotates about a vertical axis, generating planar sweeps that cover the conveyor belt and the sorting trays. Joints 2 - 4 lie in a vertical plane and form a three-link planar arm that controls reach and height of the end-effector above the workspace. Forward kinematics are modelled using the Denavit-Hartenberg (DH) convention. For each joint, four parameters are defined from the link geometry. The homogeneous transformation from frame ito frame has the standard form.

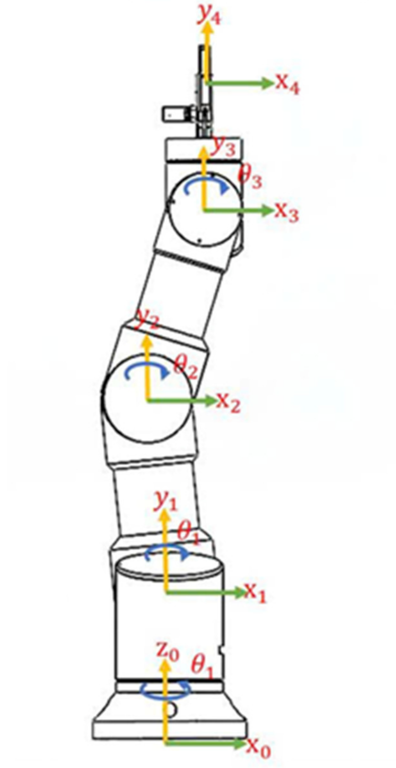


Figure 1. 3D CAD model of the four-DOF cherry-sorting robot and assignment of Denavit-Hartenberg coordinate frames

For each joint, four parameters are defined from the link geometry. The homogeneous transformation from frame ito frame has the standard form.

$$ {}^{i}T_{i+1} = \begin{bmatrix} \cos\theta_i & -\sin\theta_i\cos\alpha_i & \sin\theta_i\sin\alpha_i & a_i\cos\theta_i \\ \sin\theta_i & \cos\theta_i\cos\alpha_i & -\cos\theta_i\sin\alpha_i & a_i\sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

The overall transformation from the base frame to the end-effector frame is obtained by multiplying the four joint transformations:

$$T_4 = {}^0T_1\,{}^1T_2\,{}^2T_3\,{}^3T_4 \quad (2)$$

Table 3 lists the DH parameters of the manipulator. Joint angles θ_i are variable within their specified ranges, while the link offsets d_i, lengths a_i, and twist angles α_i are fixed by design. The base offset d_1 = 500mm sets the height of the arm above the cabinet. Links a_2 and a_4 share the same length of 1,200mm, which simplifies the inverse kinematics. The intermediate link a_3 = 980mm is chosen as a compromise between workspace and stiffness. Joints 2 and 3 are limited to $[-\pi/2, \pi/2]$ to avoid singular configurations and prevent collisions with the environment.

Table 3. Denavit-Hartenberg parameters of the four-DOF robot

Joint *i*	θ_i (rad)	d_i (mm)	a_i (mm)	α_i (rad)	Joint limits (rad)
1	θ_1	500	0	π/2	[−π, π]
2	θ_2	0	1200	0	[−π/2, π/2]
3	θ_3	0	980	0	[−π/2, π/2]
4	θ_4	0	1200	0	[−π, π]

Note: θ_i is the joint angle; d_i is the offset along z_{i-1}; a_i is the link length; α_i is the twist angle between successive z-axes.

Inverse kinematics are solved by geometric analysis. The angle of the first joint is obtained from the cherry position (x, y) in the base frame by θ_1 = atan2(y, x). The radial distance $r = \sqrt{x^2 + y^2}$ and elevation z define a triangle formed by links a_2, a_3, and the projection from the shoulder to the wrist. Angles θ_2 and θ_3 follow from the cosine rule. The wrist angle θ_4 is then chosen so that the end-effector orientation aligns with the desired approach direction. When multiple valid solutions exist, the one closest to the mid-range of each joint is selected to improve manipulability and stay away from singular poses.

Joint trajectories are generated using fifth-order polynomials:

$$\theta(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 \quad (3)$$

The six coefficients are determined from boundary conditions on position, velocity, and acceleration at the start and end of the movement. This parameterisation yields smooth motion, which helps limit vibration and reduces the load on the servo motors.

2.3. System integration and hand-eye calibration

The experimental system consists of the four-DOF robot, a conveyor belt, an Intel RealSense D435i RGB-D camera, and an industrial PC running ROS. The camera operates at 640 × 480 pixels and 30 FPS with a depth range of 0.3 - 3m. It is mounted at a height of 800mm and tilted downward by 45° towards the conveyor. A two-finger gripper with a maximum gripping force of 20N and an opening stroke of 30mm is equipped with silicone pads to limit damage to the fruit.

The robot is installed beside a 300-mm-wide conveyor. The effective workspace forms a fan-shaped region that covers the receiving area and three sorting trays. While the DH parameters in Table 3 imply a larger theoretical reach, the experimental workspace is deliberately restricted to radii of roughly 350 - 450mm from the base so that motion remains within a well-observed and mechanically safe region. The control cabinet with the PC and power electronics is located below the robot base, which minimises cable length and communication latency.

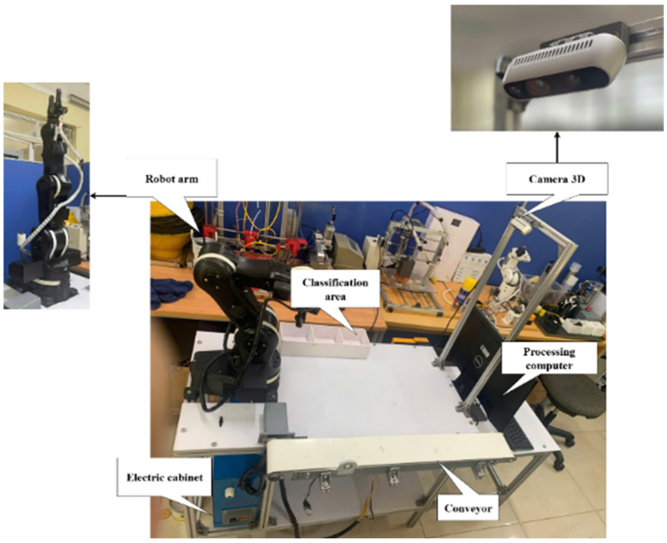


Figure 2. Laboratory setup of the cherry detection and sorting system, showing the robot, conveyor, RGB-D camera, sorting trays, and control cabinet

In software, three main ROS nodes are used.

- The vision node subscribes to RGB and depth streams and runs YOLO11s to detect cherries. For each detection with pixel coordinates *(u, v)* and camera-frame depth Z_c, the corresponding 3D point (x_c, y_c, Z_c) in the camera frame is computed by

$$x_c = \frac{(u-c_x)Z_c}{f_x},\ y_c = \frac{(v-c_y)Z_c}{f_y} \tag{4}$$

where c_x, c_y denote the principal point and f_x, f_y the focal lengths of the camera in pixel units.

- The motion-planning node transforms the camera-frame coordinates into the robot base frame, checks whether the target lies inside the admissible workspace, solves the inverse kinematics, and generates joint trajectories using the fifth-order interpolation of (3).

- The control node sends joint commands to the Dynamixel servos, reads back position and load feedback, actuates the gripper, and monitors contact force via an FSR sensor.

Hand-eye calibration estimates the rigid transformation between the camera and the robot base. During calibration, the robot successively moves the end-effector to several poses while the camera observes a fixed calibration marker. For each pose, the pose of the marker in the camera frame and the corresponding end-effector pose in the robot base frame are recorded. The unknown transformation T_c^b from camera to base is then obtained by solving the optimisation problem

$$T_c^b T_c^m \approx T_c^e \tag{5}$$

3. RESULTS AND EVALUATION

3.1. Experimental set-up

All experiments were carried out in a laboratory with ambient temperature between 22 - 25°C and relative humidity around 50 - 60%. The training and testing pipeline ran on a workstation using Ubuntu, Python and PyTorch with GPU acceleration. The three detection models (YOLO11s, YOLOv8s and YOLOv10s) were trained and evaluated on the same NVIDIA RTX 4060 GPU so that comparisons are fair. Input images were resized to 640 × 640 pixels and normalised before being fed to the networks. The training schedule was identical for all three networks. The batch size was 16 and the maximum number of epochs was 200. Early stopping on validation mAP50 was used with a patience of 50 epochs. AdamW acted as the optimiser with an initial learning rate of 0.01, gradually reduced to 0.0001 using cosine annealing. A weight-decay regularisation term was applied, and the momentum term was 0.9. Data augmentation consisted of random horizontal flip (probability 0.5), random rotation within ±15°, HSV adjustment (hue ±0.015, saturation ±0.7, value ±0.4) and mosaic augmentation with probability 0.5 during the first 150 epochs.

3.2. Detection and classification performance

Detection quality was assessed using Precision, Recall and mAP50. These metrics were computed from the numbers of true positives (TP), false positives (FP) and false negatives (FN) as

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

and

$$\text{mAP50} = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

Where *N* is the number of classes and AP_i is the average precision of class *i*.

On the test set of 180 images containing 328 annotated cherries, YOLO11s obtained the highest overall performance. Its mean precision and recall reached 94.8% and 97.3%, while the corresponding mAP50 was 95.7%. YOLOv8s and YOLOv10s achieved slightly lower values on the same data. Table 4 summarises the quantitative results, including inference speed and parameter count. All three models operated well above the 30 FPS level that is usually regarded as sufficient for real-time use on sorting lines. YOLO11s sacrificed only a small amount of speed compared with the two baselines but, in return, achieved higher mAP50 with fewer parameters.

Table 4. Performance of YOLO models on the test set (180 images, 328 objects)

Model	Precision (%)	Recall (%)	mAP50 (%)	FPS (RTX 4060)	Parameters (million)
YOLOv8s	92.8	95.1	93.4	71	11.1
YOLOv10s	93.5	96.2	94.1	69	10.2
YOLO11s	94.8	97.3	95.7	67	9.4

Note: values are averaged over three training runs with different random seeds.

To examine behaviour under challenging conditions, the test images were grouped into four subsets: a baseline subset without additional filtering, a subset with moderate occlusion, a subset of small cherries, and a subset with changed illumination. In the moderate-occlusion group, cherries had their visible area reduced by roughly 30 - 50% because of other fruits or leaves. The small-object group contained instances whose projected area was less than 50 × 50 pixels. In the illumination subset, exposure was increased or decreased by about ±2 EV relative to the reference setting.

Table 5. Performance of YOLO11s under challenging test conditions

Evaluation condition	Short description	mAP50 (%)
Normal (baseline)	Original test set without additional filtering	95.7
Moderate occlusion (30 - 50 %)	Cherries partly covered by leaves or other fruits	89.3
Small objects	Cherry area < 50 × 50 pixels	82.7
Illumination change (±2 EV)	Exposure increased or decreased by about ±2 EV	93.1

Note: each subset is drawn from the original test set based on the stated criterion.

Table 5 reports the mAP50 of YOLO11s on these four subsets. Under the baseline condition, the detector maintained an mAP50 of 95.7%. Moderate occlusion reduced this value to 89.3%, which reflects the difficulty of partially covered fruits. For small cherries, mAP50 decreased to 82.7%, indicating that recognition becomes less reliable when the objects occupy only a small number of pixels. Illumination changes within ±2 EV led to a smaller drop to 93.1%, which is consistent with the effect of the HSV-based augmentation used during training.

A class-wise analysis shows clear differences. Fully ripe cherries obtained the highest precision and recall, both above 97% and 99%, due to their distinct red colour and texture. Green cherries were also detected reliably with both indicators exceeding 92%. The most difficult class was the medium-ripe level, whose appearance lies between the green and fully ripe stages and therefore has less distinct boundaries. Even for this class, YOLO11s kept precision above 91%, slightly better than the two baselines.

The confusion matrix on the test set confirms these observations. Misclassifications were concentrated between neighbouring ripeness levels, whereas no sample was confused directly between the extreme classes "Immature" and "Mature". The diagonal entries dominated the matrix, which indicates that the model captured the colour and texture differences that separate the stages.

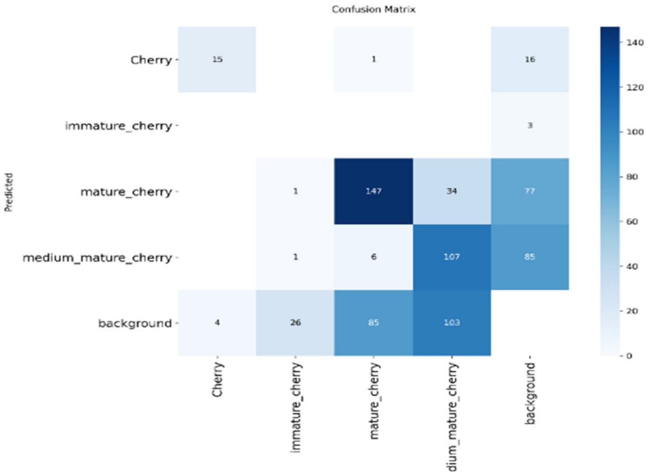
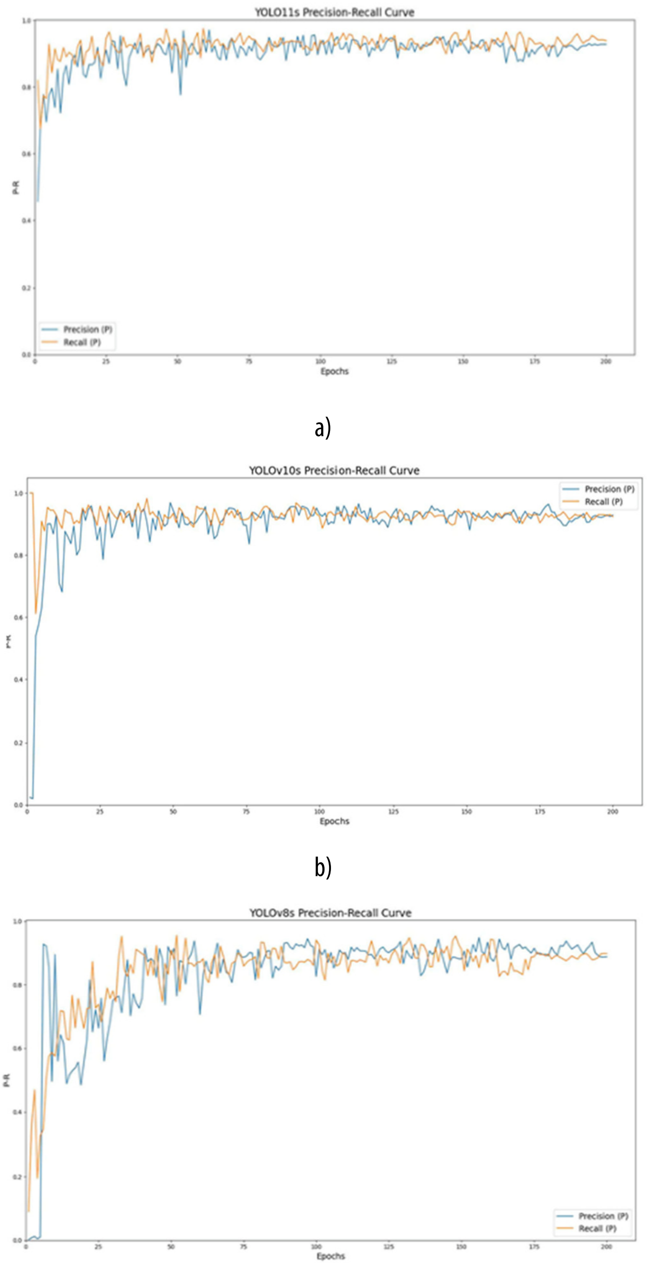


Figure 3. Confusion matrix of YOLO11s on the test set, showing correct predictions on the diagonal and errors between ripeness classes

The precision-recall (PR) curves further illustrate the relative behaviour of the three detectors. The curve of YOLO11s lies consistently above those of YOLOv8s and YOLOv10s, and the area under the PR curve reaches 0.973. When the confidence threshold decreases from 0.9 to 0.5, the precision of YOLO11s drops from about 98% to 92%, whereas YOLOv8s and YOLOv10s fall more strongly to roughly 88% and 89%. This stability at lower thresholds helps reduce false negatives in scenes where cherries have low contrast against the background. An operating confidence threshold of 0.65 offers a convenient trade-off: both precision and recall stay above 94%, which is adequate for industrial sorting lines that need to balance accuracy and detection rate.



a)

b)

c)

Figure 4. Precision-recall curves of YOLO11s (a), YOLOv8s (b) and YOLOv10s (c) on the test set

3.3. Ablation analysis

An ablation study was carried out to evaluate the contribution of the main architectural changes in YOLO11s. Starting from a baseline configuration equivalent to YOLOv8s (without C3k2 blocks, C2PSA attention or depthwise convolutions in the head), three intermediate variants were constructed. First, the standard backbone was replaced by a version using C3k2 blocks. Second, C2PSA attention was inserted. Finally, standard convolutions in the detection head were replaced by depthwise convolutions.

Table 6 shows the mAP50 for each variant. The baseline model reached an mAP50 of 93.4%. Introducing C3k2 blocks increased mAP50 to 94.1 %. Adding C2PSA attention raised it further to 95.1%. When depthwise convolutions were also used in the head, the full YOLO11s configuration achieved an mAP50 of 95.7%, an absolute gain of 2.3 percentage points compared with the baseline.

Table 6. Ablation of the main components in the YOLO11s architecture

Model variant	**C3k2 block**	**C2PSA attention**	**Depthwise head**	**mAP50 (%)**	**ΔmAP50 vs. baseline (points)**
Baseline (equivalent to YOLOv8s)	No	No	No	93.4	0.0
Baseline + C3k2	Yes	No	No	94.1	+0.7
Baseline + C3k2 + C2PSA	Yes	Yes	No	95.1	+1.7
Full YOLO11s	Yes	Yes	Yes	95.7	+2.3

Note: all models are trained and evaluated under the same settings as in Section 3.1.

Each component contributes a modest but consistent improvement, and the cumulative effect leads to a noticeable increase in mAP50 compared with the baseline. The parameter count grows only slightly, while the decrease in inference speed remains small. A paired *t* - test on mAP50 over the three training runs shows that the differences between YOLO11s and YOLOv8s are statistically significant with $p < 0.01$, and the gap between YOLO11s and YOLOv10s is significant with $p < 0.05$.

3.4. Evaluation of the integrated system

The positioning accuracy of the robotic subsystem was measured using a reference jig. After hand–eye calibration, the mean error in end-effector position was 2.5 ± 0.9mm, increasing gradually with distance from the calibration region. The dominant error sources were depth sensing, calibration residuals, mechanical backlash and encoder resolution. For cherries with an approximate diameter of 10mm, this error level remained acceptable for grasping.

The average cycle time over 150 repetitions, measured from image acquisition to placement in the tray, was 2.8 ± 0.3s. Vision inference and trajectory generation together consumed only a few tens of milliseconds; most of the time was spent on mechanical motion. Under continuous feeding, the theoretical throughput would be around 1,200 cherries per hour, while in the experiment an effective rate of about 950 cherries per hour was observed because the conveyor was not loaded continuously.

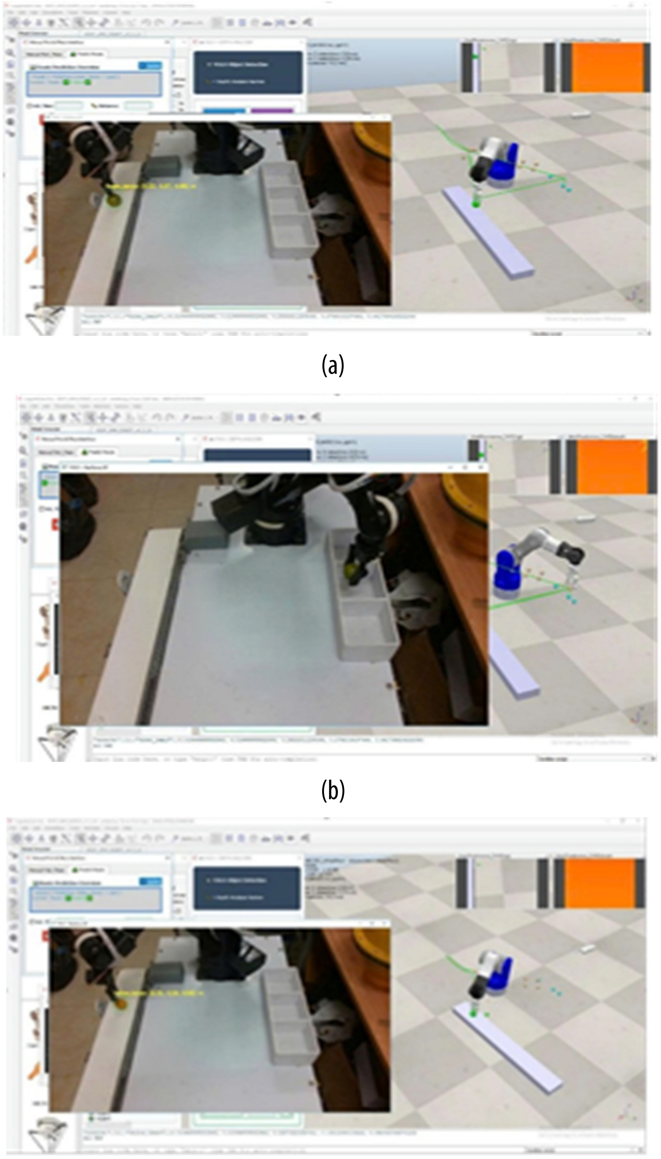
The detailed timing and success rates of individual phases are listed in Table 7. The inference step of YOLO11s on the RTX 4060 required 35 ± 4ms, and trajectory planning took 8 ± 2ms. The approach phase from the observation pose to the picking pose lasted about 1,150 ± 180ms. Grasping, including gripper opening, descent, closing and force confirmation, took approximately 400ms. Transport from the picking position to the tray required 920 ± 140ms, followed by 200ms for releasing the cherry and 780 ± 100ms for retreating to the observation pose. These values confirm that the main bottleneck is mechanical motion rather than computation.

Table 7. Summary of timing and success rates of the integrated cherry detection and sorting system

Indicator	Value
Perception + planning time per cherry	≈ 40ms
Total cycle time per cherry	2.8 ± 0.3s
Effective throughput (intermittent feed)	≈ 950 cherries/hour
Detection rate (200 cherries)	97.0% (194/200)
Grasp success rate (on detected fruits)	94.8% (184/194)
Transport success rate	98.9% (182/184)
End-to-end success rate	89.5% (179/200)

A separate test with 200 cherries was used to evaluate success rates. The detection stage identified 194 of them, corresponding to a detection rate of 97%; the six missed samples were either heavily occluded or outside the reachable workspace. Among the 194 visible cherries, 184 were grasped successfully, giving a grasp success rate of 94.8%. Of these 184, the robot transported 182 to the trays without dropping, i.e. a transport success rate of 98.9%. Classification into the correct tray was correct for 98.4% of the transported cherries, which is consistent with the detection accuracy on the test set. Overall, 179 out of 200 cherries were detected, grasped, transported and placed into the appropriate tray, resulting in an end-to-end success rate of 89.5%.

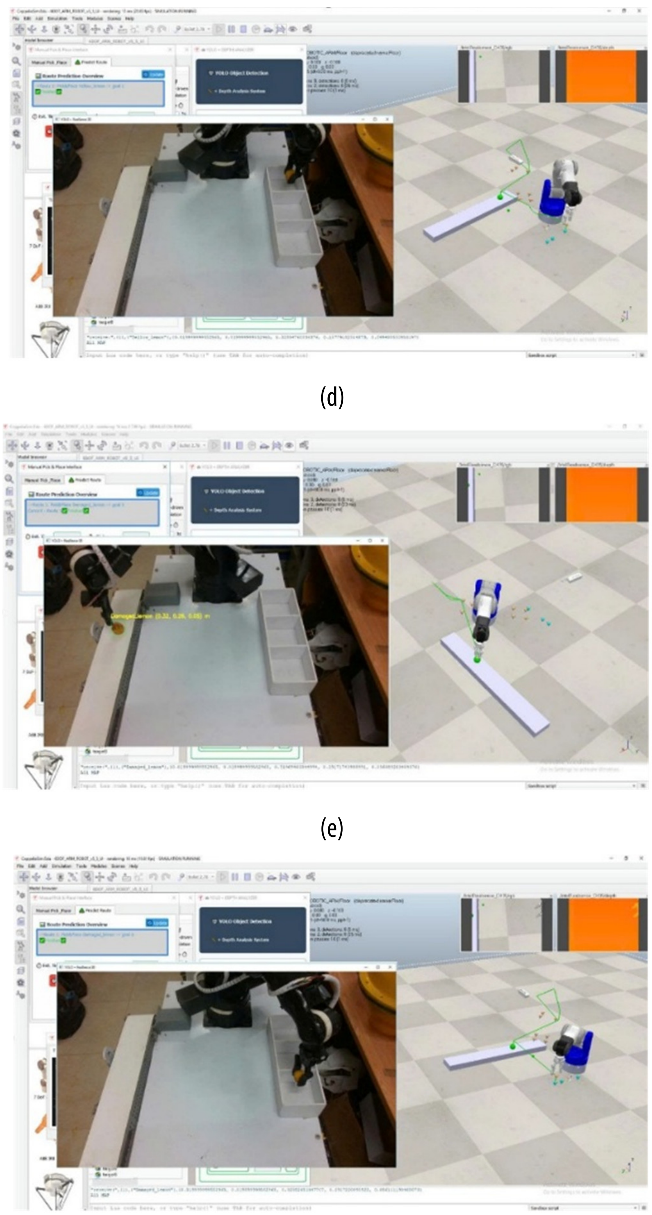
Representative qualitative results are shown in Figure 5. In Figure 5(a), a green cherry is detected with high confidence and is grasped from above, which helps the gripper hold the fruit securely without visible damage. Figure 5(b) illustrates the corresponding placement into the tray. Figures 5(c) and 5(d) present a medium-ripe cherry whose colour lies between the green and fully ripe stages; in this case the model uses the proportion of red area on the surface to assign the correct ripeness level. Figures 5(e) and 5(f) show a fully ripe cherry. The predicted bounding box closely follows the fruit contour, and the gripper force is regulated so that the soft flesh is not bruised during the pick-and-place motion. These examples are consistent with the quantitative results in Table 6 and illustrate typical successful cycles of detection, grasping and release.



(a)

(b)

(c)



(d)

(e)

(f)

Figure 5. Detection and pick-and-place results for green, medium-ripe and fully ripe cherries. Panels (a), (c) and (e) show the grasp phase; panels (b), (d) and (f) show the release phase into the corresponding trays

These results show that the integrated system meets the timing and accuracy requirements of a small-scale post-harvest sorting line, while the remaining failures can be traced to specific mechanical and perception limitations that are analysed further in the discussion section.

4. DISCUSSION

YOLO11s achieved an mAP50 of 95.7% at about 67 FPS on the RTX 4060, which meets real-time requirements while maintaining accuracy comparable with recent agricultural systems. Tang et al. reported good performance for an improved YOLOv4-tiny on Camellia oleifera, but with a frame rate only slightly above 30 FPS [14], so the present configuration operates at roughly twice the speed for a similar accuracy level. Tian et al. obtained mAP values of about 0.89 - 0.90 for apples in complex orchard scenes using YOLOv3-dense [15]; although their tests were outdoors and this system is evaluated on a post-harvest line, the comparison suggests that YOLO11s can cope with small or partially occluded fruits while narrowing the gap between laboratory settings and more demanding environments.

Li et al. used YOLOv8 with an additional segmentation branch to locate mango picking points and reached high mAP50 at the cost of higher computational load [16]. Meng et al. proposed YOLOv9-pose and YOLOv10-pose for strawberry peduncle detection, relying on more complex pose-estimation architectures [17]. In contrast, the present system deliberately keeps the network responsible only for detection and ripeness classification, and derives the grasp pose from robot geometry and depth data. This lighter design simplifies the pipeline and lowers hardware requirements, which is advantageous for small and medium-sized facilities, but it reduces flexibility in end-effector orientation relative to full pose-estimation approaches.

Several constraints remain. The image set of 1,200 samples was collected at a single farm in Sa Pa during one season, so the model has not yet been validated on cherries from other regions or varieties, and the experiments were carried out in a laboratory with controlled illumination and a stable conveyor. Industrial plants introduce dust, vibration and wear; long-term operation will require suitable protective housings and food-grade materials. The four-degree-of-freedom robot helps reduce cost but limits the reachable workspace, which explains why fruits near the conveyor edge contribute to the end-to-end success rate of 89.5%. Adding a wrist degree of freedom could improve reach and approach angles but would increase system cost and control complexity. Hand-eye calibration is also a bottleneck. As Enebuse et al. showed, calibration accuracy depends on both the algorithm and the distribution of poses and noise during data collection [18]. For durable deployment, the calibration trajectory and conditions should be redesigned and periodic recalibration scheduled to control long-term drift.

5. CONCLUSION

The study developed an integrated cherry grading system that couples a YOLO11s detector with a four-degree-of-freedom robotic manipulator for post-harvest sorting. Compared with YOLOv8s and YOLOv10s, YOLO11s improved mAP50 by about 1.6 - 2.3 percentage points with only a slight reduction in inference speed. The robot, arranged around the conveyor and sorting trays and using an RGB-D camera, hand-eye calibration and fifth-order trajectory planning, achieved a mean positioning error of 2.5 ± 0.9mm, an end-to-end success rate of 89.5%, and an average cycle time of 2.8s (≈950 cherries per hour under intermittent feeding). These results, while still constrained by workspace and gripper design, are acceptable for a laboratory system aimed at small and medium processing facilities and indicate realistic potential for industrial deployment after further standardisation of mechanical design, electrical safety and food-contact hygiene. Current limitations concern data diversity, workspace coverage, gripper compliance and calibration accuracy. The dataset comes mainly from a single growing area, some cherries near the conveyor edge remain unreachable, the gripper is relatively stiff for very delicate fruits, and hand-eye calibration can still be refined. Future work should therefore expand and diversify the image data, redesign the gripper and its synchronisation with the conveyor, and introduce semi-automatic periodic calibration procedures, followed by pilot trials in real factories to evaluate durability and long-term reliability of the complete system.

ACKNOWLEDGMENT

This study was conducted at the School of Mechanical and Automotive Engineering, Hanoi University of Industry. The authors gratefully acknowledge the financial support provided by research project No. 07-2025-RD/HĐ-ĐHCN.

REFERENCES

[1]. M. M. Poenaru, A. Cogato, M. Sozzi, A. Nikolic, E. Laroche-Pinel, "Shaping the future of horticulture: Innovative technologies, artificial intelligence, and robotic automation through a bibliometric lens," *Horticulturae*, 11, 5, art. 449, 2025.

[2]. C. Wang, Q. Han, W. Li, Z. Zhang, C. Liu, X. Tang, "A review of perception technologies for berry fruit-picking robots: Advantages, disadvantages, challenges, and prospects," *Agriculture*, 14, 8, art. 1346, 2024.

[3]. C. M. Badgujar, A. Poulose, H. Gan, "Agricultural object detection with You Only Look Once (YOLO) algorithm: A bibliometric and systematic literature review," *Computers and Electronics in Agriculture*, 223, art. 109090, 2024.

[4]. Z. Q. Zhao, P. Zheng, S. T. Xu, X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, 30, 11, 3212-3232, 2019.

[5]. M. A. R. Alif, M. Hussain, "YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain," *arXiv preprint arXiv:2406.10139,* 2024.

[6]. S. Khanam, M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[7]. R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, M. Karkee, "Comprehensive performance evaluation of YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments," *Qeios,* 2024. doi:10.32388/E9Y7XI.

[8]. G. Hu, C. Peng, Y. Xiong, Y. Grimstad, P. J. From, V. Isler, "Simplified 4-DOF manipulator for rapid robotic apple harvesting," *Computers and Electronics in Agriculture*, 199, art. 107177, 2022.

[9]. T. Yoshida, Y. Onishi, T. Kawahara, T. Fukao, "Automated harvesting by a dual-arm fruit harvesting robot," *ROBOMECH Journal*, 9, 2022. doi:10.1186/s40648-022-00233-9.

[10]. T. Dewi, P. Risma, Y. Oktarina, "Fruit sorting robot based on color and size for an agricultural product packaging system," *Bulletin of Electrical Engineering and Informatics*, 9, 4, 1438-1445, 2020.

[11]. X. Zhang, M. Yao, Q. Cheng, G. Liang, "A novel hand-eye calibration method of picking robot based on TOF camera," *Frontiers in Plant Science*, 13, art. 1099033, 2023.

[12]. R. Y. Tsai, R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, 5, 3, 345–358, 1989.

[13]. R. Gai, N. Chen, H. Yuan, "A detection algorithm for cherry fruits based on the improved YOLO-v4 model," *Neural Computing and Applications*, 35, 19, 13895-13906, 2023.

[14]. Y. Tang, H. Zhou, H. Wang, X. Wang, X. Zhao, "Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision," *Expert Systems with Applications*, 215, art. 118573, 2023.

[15]. Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO-V3 model," *Computers and Electronics in Agriculture,* 157, 417-426, 2019.

[16]. H. Li, J. Huang, Z. Gu, D. He, J. Huang, C. Wang, "Positioning of mango picking point using an improved YOLOv8 architecture with object detection and instance segmentation," *Biosystems Engineering,* 247, 202-220, 2024.

[17]. Z. Meng, X. Du, R. Sapkota, Z. Ma, H. Cheng, "YOLOv10-pose and YOLOv9-pose: Real-time strawberry stalk pose detection models," *Computers in Industry*, 165, art. 104231, 2025.

[18]. I. Enebuse, B. K. S. M. K. Ibrahim, M. Foo, R. S. Matharu, H. Ahmed, "Accuracy evaluation of hand-eye calibration techniques for vision-guided robots," *PLOS ONE*, 17, 10, art. e0273261, 2022.