agriculture

MDPI

Review

Visual Recognition and Three-Dimensional Localization of Picking Points for Strawberry Harvesting Robots

Lili Jiang, Haohao Yan, Jiyuan Wang, Yingzi Yin and Chong Wu *

Shandong Institute of Pomology, Taian 271000, China; gssjianglili@shandong.cn (L.J.); 17863801132@163.com (H.Y.); wangjy11_2026@163.com (J.W.); yinyingzi555@126.com (Y.Y.)
* Correspondence: wuchongge@163.com

Abstract

Reliable robotic strawberry harvesting depends on accurate mature-fruit recognition and the estimation of picking poses that satisfy the geometric constraints imposed by the detachment mechanism and surrounding canopy. This review critically examines the relationship between visual recognition and three-dimensional (3D) picking-point localization in strawberry-harvesting robots. Using a narrative technical synthesis, it compares methods for maturity and pickability assessment, fruit-instance segmentation, anatomical landmark representation, depth acquisition, geometric reconstruction, coordinate transformation, and perception-guided manipulation. Cross-crop evidence is considered without assuming equivalent performance in strawberries. The analysis highlights distinctions among image-level recognition, metric localization, tool-pose estimation, and physical harvesting success. Occluded peduncles, invalid or mixed depth, color–depth registration errors, hand–eye calibration uncertainty, and sensing-to-actuation delays may compromise the accuracy and executability of a predicted picking pose. Furthermore, heterogeneous cultivars, cultivation systems, target materials, end effectors, and success criteria limit direct comparison of reported results. By relating visual representations to geometric observability, error propagation, tool-specific accessibility, and execution feedback, this review establishes an action-oriented framework for evaluating the evidence. It further identifies priorities for uncertainty-aware perception, explicit anatomical landmark definitions, independently validated 3D localization, and real-fruit experiments that jointly report positioning error, detachment success, fruit damage, and cycle time.

Keywords: selective harvesting; peduncle perception; keypoint estimation; RGB-D sensing; instance segmentation; active vision; hand–eye calibration; visual servoing

Check for updates

Academic Editor: Aichen Wang

Received: 8 October 2026
Revised: 7 November 2026
Accepted: 16 November 2026
Published: 18 November 2026

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

1. Introduction

Selective strawberry harvesting requires the identification of fruit at an appropriate maturity stage and its removal without damaging the fruit or adjacent plant tissues. Unlike bulk harvesting, robotic systems must assess individual targets and determine a feasible interaction strategy within a spatially constrained canopy. This task therefore involves more than visual detection: reliable operation requires the coordination of semantic perception, metric localization, manipulator motion, end-effector interaction, and fruit handling. Previous reviews have emphasized the interdependence of these capabilities in practical fruit-harvesting systems [1–5]. For strawberries, the critical visual output is consequently an anatomically meaningful picking pose that the selected end effector can reach and execute, rather than the location of the fruit body alone.

Early strawberry-harvesting robots demonstrated automated operation in tabletop and elevated cultivation systems, while exposing limitations associated with target visibility, access to the peduncle, and harvesting cycle time [6–9]. Subsequent platforms combined visual sensing with cable-driven gripping mechanisms and autonomous manipulation, including evaluations under cultivation conditions [10,11]. Collectively, these developments underscore the distinction between fruit recognition and actionable target localization. A detector may classify a fruit as mature without establishing the location of its attachment structure, an appropriate approach orientation, or sufficient clearance for the intended detachment mechanism.

Although closely coupled, visual recognition and three-dimensional (3D) localization address different estimation problems. Recognition provides semantic and image-space information about a candidate fruit, including its maturity, instance boundary, and observable anatomical landmarks. Localization assigns metric coordinates to interaction-relevant features through depth sensing, stereo reconstruction, or other geometric measurements. To guide manipulation, the resulting point or pose must subsequently be expressed in the robot and tool coordinate frames. High image-level recognition accuracy therefore does not, by itself, establish reliable 3D localization or precise placement of the end effector.

The transition from perception to action can be described as a sequence of interdependent operations: candidate qualification, instance separation, identification of the relevant anatomical structure, depth association, 3D reconstruction, coordinate transformation, and tool-accessibility assessment. Errors arising at one stage may propagate to those that follow. Peduncle occlusion can prevent direct landmark observation; invalid or mixed depth can displace an otherwise correctly detected feature; and calibration or synchronization errors can further alter its position in the robot frame. Conversely, active viewpoint adjustment and visual feedback can introduce additional observations and refine the pose estimate during approach. Accordingly, the reliability of a picking pose must be assessed over the entire sensing-to-execution chain rather than inferred from an isolated perception metric.

Existing reviews provide complementary perspectives on visual perception and robotic fruit harvesting. Huang et al. [1] surveyed visual perception technologies for fruit-harvesting robots, including recognition, depth sensing, active vision, and visual servoing. Jia et al. [12] reviewed information technologies supporting apple-harvesting robots, whereas Li et al. [13] examined the relationship between end-effector design and recognition–localization methods for fruit and vegetable harvesting. Oliveira et al. [14] discussed broader developments and challenges in agricultural robotics. Cui et al. [15] provided a strawberry-specific synthesis covering cultivation systems, harvesting mechanisms, robotic integration, and post-harvest operations. Although these reviews establish an important foundation, their broader perspectives do not specifically organize the available evidence around the complete transformation from strawberry recognition to an executable, tool-dependent picking pose. The present review therefore focuses on the connections among anatomical landmark representation, metric three-dimensional reconstruction, coordinate transformation, geometric uncertainty, manipulator accessibility, and execution feedback. Its contribution lies in critically examining the interfaces between these stages and distinguishing the performance of individual perception modules from localization accuracy and physical harvesting outcomes, rather than treating the technologies as independent components.

These developments motivate four questions central to this review: how biological maturity differs from robot- and tool-dependent pickability; how anatomical landmarks can be represented and assigned metric 3D coordinates; how visual, depth, calibration, and timing errors affect the final tool pose; and how reported perception performance relates to physical harvesting outcomes. These questions guide the discussion of visual information

requirements, localization methods, robotic applications, and unresolved challenges in Sections 2–5. Strawberry-specific investigations provide the direct empirical basis, whereas cross-crop studies are used to examine transferable sensing and estimation principles. Differences in crop anatomy, cultivation, and detachment mechanisms preclude direct extrapolation of numerical performance to strawberries.

This article adopts a narrative technical synthesis rather than a systematic review or meta-analysis. The discussion draws on foundational studies and research publications available up to 4 November 2026. To update the coverage and check the interpretation of key evidence during revision, a targeted supplementary literature check was conducted on 4 November 2026 using publicly accessible scholarly publisher records and topic-focused searches, including material available through MDPI, Springer Nature, ScienceDirect, and Frontiers. Search concepts combined terms such as "strawberry", "harvesting robot", "visual recognition", "ripeness", "pickability", "peduncle", "picking point", "keypoint", "RGB-D", "stereo vision", "3D localization", "hand–eye calibration", "visual servoing", and "end effector". Strawberry-specific primary studies were prioritized when they reported evidence concerning target recognition, interaction-point localization, or robotic harvesting execution. Cross-crop studies and general robotics research were considered for transferable sensing, geometry-estimation, calibration, and control principles, without treating their crop-specific performance as direct evidence for strawberries. Studies focused on grading, yield estimation, or other tasks without a harvesting interaction were not treated as direct evidence of picking performance. Comparisons consider target material, cultivation and image-acquisition conditions, sensing configuration, anatomical representation, coordinate frame, reference measurement, end-effector design, sample size or denominator where reported, validation level, and outcome definition. Missing information is identified as unreported, and numerical outcomes obtained under incompatible experimental conditions are not pooled or directly ranked. Because complete historical database-search logs and screening counts from initial manuscript preparation were not retained, the present work does not claim exhaustive retrieval or a reproducible systematic screening process. These qualifications define the scope and evidential limits of the narrative synthesis.

2. Visual Information Required for Strawberry Picking-Point Localization

Reliable picking-point localization requires three complementary categories of information: target eligibility, interaction geometry, and local scene context. Ripeness and instance identity identify candidate fruit; peduncle, calyx, or contact-surface geometry determines the intended location and orientation of tool interaction; and surrounding 3D geometry constrains visibility and physical access. These outputs are interdependent rather than interchangeable. Instance segmentation provides a pixel-level representation for depth association, whereas the depth of an occluded peduncle and the collision-free feasibility of a tool approach require additional geometric evidence. A task-oriented representation should therefore preserve fruit identity, landmark visibility, the provenance and reliability of depth measurements, nearby obstacle geometry, and the reference frame used to describe each quantity.

2.1. Ripeness and Pickable-Target Information

Ripeness recognition addresses biological eligibility under a specified harvesting standard. Its output depends on the maturity-stage definition, cultivar, fruit appearance, illumination, and viewing geometry. A mature-stage label does not establish that the fruit–peduncle junction is visible, that valid depth is available at the intended interaction point, or that the end effector can reach it. The categories illustrated in Figure 1 therefore describe

the biological state used to qualify a candidate; they do not constitute a complete robotic picking decision.

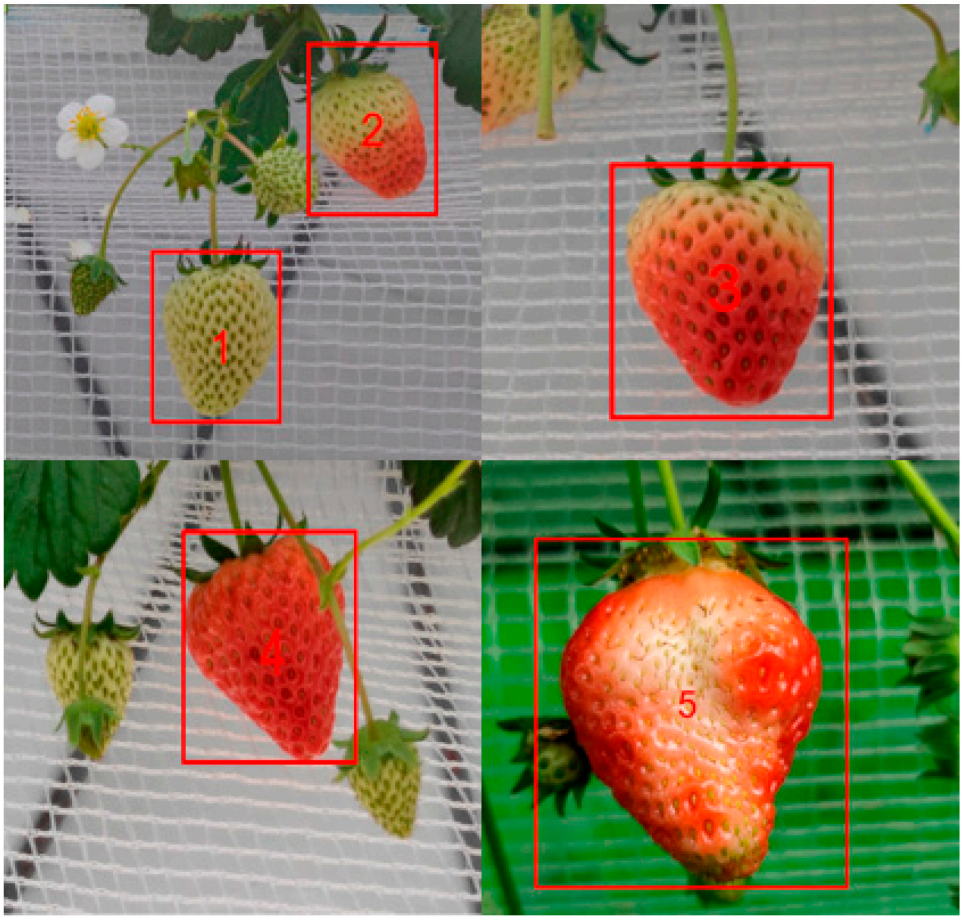


Figure 1. Strawberry maturity categories used for selective-harvesting recognition. (Reprinted from Ref. [16] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

RGB and spectral approaches differ in the maturity evidence they measure and in their acquisition requirements. Image enhancement can improve the presentation of color and texture to a detector [17], whereas hyperspectral estimation uses wavelength-dependent information and requires suitable illumination, sensor calibration, and acquisition time [18]. Recognition in complex backgrounds further tests whether maturity cues remain separable from surrounding vegetation [19]. Comparisons should therefore state the maturity classes, cultivar and imaging conditions where reported, and should not rank methods solely by accuracy values obtained under different acquisition protocols. Table 1 separates the decision supported by each information type from the requirements left unresolved.

Candidate assessment should distinguish four successive requirements: appropriate biological maturity, sufficient observability of the fruit and attachment region, reliable geometric measurements, and feasible physical interaction. Tafuro et al. [20] combined maturity, picking-point, and weight estimation to describe candidate fruit, whereas He et al. [21] treated mature-fruit detection and pickability classification as separate stages. The latter formulation captures an operational distinction, although pickability labels necessarily depend on the robot, end-effector geometry, and cultivation context used to define them. Neither predicted fruit attributes nor pickability classes alone establish valid depth at the attachment point, adequate tool clearance, or a kinematically feasible approach.

Table 1. Decision scope and unresolved requirements of maturity and pickability information.

Information and Sources	Decision Supported	Unresolved Requirement	Ref.
RGB maturity class	Select fruit at the defined visual maturity stage	Class definitions and illumination sensitivity; no direct evidence of tool access	[16,17,19]
Spectral maturity estimate	Characterize ripeness using spectral information	Additional sensing and calibration; relationship to the harvesting threshold	[18]
Joint fruit attributes	Combine ripeness, picking point and weight for candidate assessment	Predicted attributes still require valid depth and feasible tool interaction	[20]
Explicit pickability class	Screen detected candidates for operational suitability	Class labels depend on tool geometry and cultivation conditions	[21]

A practical decision process should distinguish mandatory feasibility constraints from ranking preferences. Workspace limits, collision avoidance, tool-specific clearance, and a minimum level of localization confidence determine whether a target is eligible for execution. Among targets that satisfy these constraints, maturity, travel cost, and expected success may determine priority. An unconstrained weighted sum is inadequate when a favorable appearance score could numerically compensate for an unsafe approach or an invalid depth estimate. Any ranking model should therefore be calibrated for the robot, end effector, and cultivation setting, while hard safety and feasibility constraints remain explicit.

Pickability is intrinsically tool-dependent rather than a universal property of a mature fruit. A fruit with an exposed body may be accessible to a compliant gripper but unsuitable for peduncle cutting; conversely, a visible peduncle may remain inaccessible to a cutter whose swept envelope intersects neighboring fruit or foliage. The annotation protocol should consequently distinguish biological maturity, optical visibility, depth validity, geometric reachability, and tool-specific interaction feasibility. Recording the end-effector type and the operational criterion used to assign a pickable label is necessary for interpreting results and for transferring a classifier to another harvesting mechanism.

2.2. Peduncle, Calyx, Picking Point and Fruit-Pose Information

The physical definition of a picking point follows from the detachment mechanism and should be specified in relation to the selected tool. A cutting tool requires an admissible peduncle segment and cutting orientation; suction requires a suitable contact surface and approach normal; and clamping requires adequate clearance around the intended contact region. Twisting or pulling also depends on fruit orientation and attachment direction. Thus, the fruit centroid, bounding-box center, anatomical landmark, and final tool interaction point are not interchangeable quantities. A study should state which of these is annotated and how it is converted into the commanded tool pose.

The suitability of a picking-point representation depends on the extent to which attachment geometry is directly observed rather than inferred. Contour- and attention-based methods use visible fruit and stem structure to constrain the predicted interaction

point [22,23], although incomplete boundaries can bias the estimate. Anatomical landmarks at the calyx and harvesting region provide more explicit structural references, as illustrated in Figure 2 [24]. Joint maturity–keypoint prediction, multiple-keypoint representations, and orientation estimation extend these outputs toward tool alignment [25–28]. A further challenge is ensuring that a detected peduncle belongs to the intended fruit [29,30]. Qing et al. [29] addressed this fruit–stem association by adapting the localization strategy to whether the corresponding stem could be detected. Their approach also illustrates why points inferred from fruit shape under occlusion should be distinguished from those obtained from visible attachment structures.

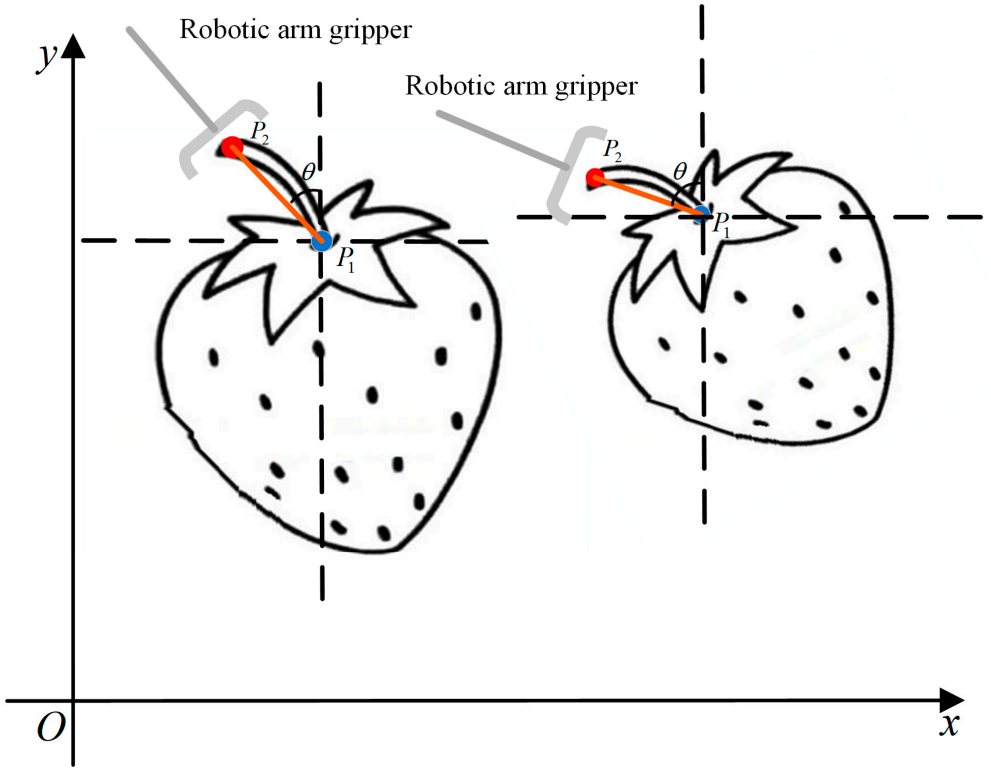


Figure 2. Definition of the calyx keypoint P_1 and harvesting point P_2 used to describe strawberry orientation. (Reprinted from Ref. [24] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

The transferability of a representation depends on the interaction constraint it encodes and on the anatomical level represented. Grape methods may associate a cluster with its detachment structure [31,32]; mango studies may estimate an individual fruit–peduncle relationship [33,34]; and litchi methods may identify a fruit-bearing branch [35]. These examples support the general principle of associating a target with the structure through which it is detached. They do not establish interchangeable landmark definitions: cluster architecture, branch hierarchy, stem thickness, fruit mass, attachment geometry, and harvesting tools differ from those of an individual strawberry. Cross-crop results are therefore methodological context, not direct evidence of strawberry localization accuracy or harvesting performance (Table 2).

For manipulation, the estimated target should include a feasible interaction region and, where required, its orientation. A short peduncle segment is more informative for cutting than a single unconstrained pixel, while an exposed surface patch with an estimated normal is more useful for suction than a fruit centroid. Evaluation should measure position and orientation error relative to the admissible tool region and its tolerance, rather than

relying exclusively on generic keypoint or detection metrics. The geometric definition should remain consistent between annotation, localization, and the controller.

Table 2. Picking-point representations compared by geometric content and tool requirements.

Representation and Sources	Geometric Content	Useful Interaction	Principal Ambiguity	Ref.
Contour or axis geometry	Fruit axis and a geometrically inferred point near the peduncle	Initial approach or picking-point estimate	Partial contours can bias the axis; the attachment point is inferred rather than directly detected in [36]	[22,23,36]
Calyx and harvesting landmarks	Discrete anatomical/action points; Ref. [24] defines P1 at the calyx and P2 at the harvesting point	Tool placement linked to visible anatomy	Image keypoints still require valid depth and a calibrated tool offset; Ref. [24] evaluates camera-frame 3D coordinates on a simulated platform	[24,25]
Multiple anatomical keypoints	Relative arrangement of fruit and attachment landmarks	Pose description and structural consistency checks	Occluded points are inferred; visibility and uncertainty remain necessary	[26]
Orientation and pose	Position together with an estimated direction	Directional approach and tool alignment	Orientation convention and available degrees of freedom must match the tool	[27,28]
Fruit–stem association	Attachment geometry assigned to a fruit instance	Select the corresponding detachment structure	Wrong instance association can invalidate an otherwise precise point	[29,30]

Observed and inferred landmarks require separate treatment. A network may estimate an occluded attachment point from visible fruit shape, but that estimate is a prediction of hidden geometry rather than a direct measurement of the attachment structure. The output should retain visibility status, confidence, and spatial uncertainty, and should identify whether the point was obtained from a visible landmark, a geometric offset, or a learned inference. If uncertainty exceeds the tolerance of the selected tool, additional observation or target deferral is preferable to executing an inadequately supported pose.

2.3. Occlusion, Spatial Context and Manipulator Reachability

Occlusion affects both measurement availability and physical accessibility, but these are distinct failure modes. Leaves, neighboring fruit, flowers, and support structures may obstruct the camera view, the tool corridor, or both. A partially visible strawberry may therefore be recognized accurately while its attachment point remains unobservable or its approach remains physically blocked. Figure 3 illustrates how canopy management and the original growth configuration alter target exposure in ridge cultivation.

RGB-D perception can provide fruit positions together with local scene geometry, as demonstrated in the tabletop strawberry-harvesting system developed by Ge et al. [37]. Three-dimensional fruit-shape sensing offers a complementary description of visible surfaces [38], but the workspace and occlusion patterns of tabletop cultivation are not necessarily representative of ridge-grown greenhouse or open-field systems. Mask-constrained

depth association and geometric fitting may improve localization when neighboring depth samples correspond to the same physical structure. Attention-guided viewpoint planning [39] and shape-completion methods [40] can also address incomplete observations; nevertheless, model-completed surfaces remain inferred geometry and should not be interpreted as direct measurements of a hidden peduncle.



Figure 3. Ridge-grown strawberry scenes after thinning and in the original growth state, illustrating differences in exposure and occlusion. (Reprinted from Ref. [41] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Reachability must be evaluated in the robot coordinate system rather than in image space. The target position must lie within the manipulator workspace, the required orientation must be attainable, and the planned motion must maintain clearance for the complete tool envelope, not only its nominal tip. Probabilistic occupancy mapping and semantic reconstruction provide representations for this assessment [42,43]. A target pose should therefore be accompanied by local obstacle geometry, uncertainty, and the frame in which the pose is expressed before inverse kinematics and collision checking are performed.

Three constraints are useful to distinguish analytically: insufficient optical visibility, obstruction of the tool corridor, and unacceptable contact with vulnerable plant organs. Their remedies differ. Additional sensing may resolve incomplete observations; trajectory modification may address geometric obstruction; and target deferral may be necessary when contact risk remains unacceptable. A single occlusion percentage cannot capture these mechanisms, and a visibility improvement should not be used as a surrogate for collision-free access or successful detachment.

Occlusion management should distinguish improvements in visual observability from improvements in physical access. In a field evaluation of a ridge-grown dual-arm strawberry harvester, Yu et al. [41] examined operating performance under thinned and untrimmed canopy conditions, highlighting the influence of plant configuration on both target exposure and manipulation. Because these conditions were assessed at different sites and dates, their reported performance difference cannot be attributed exclusively to thinning. He et al. [44] investigated a complementary intervention by incorporating a fan-assisted end effector to disperse local occlusion before picking. Together, these studies distinguish agronomic canopy management from robot-mediated occlusion handling. Additional observation can reveal an attachment point without clearing the tool corridor, whereas a physically reachable fruit may still require another view to obtain valid depth. Table 3 therefore links distinct failure mechanisms to re-observation, trajectory adjustment, local occlusion removal, or target deferral.

Table 3. Distinct causes of perception and access failure and their implications for intervention.

Barrier and Illustrative Sources	Information Needed	Operational Implication	Ref.
Optical occlusion	Visibility of the attachment region and expected benefit of another view	Acquire a targeted observation when it can resolve the missing structure	[39,45]
Incomplete scene geometry	Valid depth coverage and uncertainty of inferred surfaces	Distinguish measured free space from regions completed by a model	[37,38,40]
Obstructed tool corridor	Clearance around the full tool envelope	Change approach or defer; improved image visibility alone does not create access	[41,44]
Motion during observation and approach	Measurement timestamps and target–camera motion	Update stale coordinates and assess filtering delay	[46]
Uncertain physical transfer	Real-fruit compliance, plant motion and contact response	Verify successful simulation or synthetic-target operation on real plants	[47]

3. Key Visual Technologies Supporting Picking-Point Recognition and 3D Localization

Visual recognition methods differ principally in the information they provide to downstream geometric estimation and manipulation. Conventional processing characterizes color, texture, and contour geometry; object detection locates candidate regions; instance segmentation separates individual visible fruit; and landmark or pose estimation represents interaction-relevant structures. These predictions acquire metric meaning only when they are associated with valid measurements of the intended physical surface. The resulting information forms a sequence of complementary outputs rather than interchangeable performance measures: detection accuracy does not quantify segmentation reliability, landmark observability, 3D positioning error, or harvesting success. Appropriate method selection therefore depends on resolution, measurement uncertainty, computational latency, and calibration requirements relative to the harvesting task.

3.1. Conventional Machine Vision and Color/Geometric Processing

Conventional strawberry vision uses color-space transformations, thresholding, morphology, connected components, and contour analysis to estimate mature-fruit regions and principal axes [7]. Related fruit-vision research evaluates texture classification, edge extraction, and color or contour processing [48–51]. These methods offer interpretable intermediate results and low computational requirements, making them useful as baselines or components of hybrid systems.

The main limitation is sensitivity to departures from assumed appearance and shape. Shadows and reflections alter color distributions, while overlapping fruit can merge into one connected component and bias the estimated axis. Partial-shape matching and Hough-based methods address some incomplete contours, but their validity depends on the remaining

geometric evidence [52–57]. Errors near the calyx region are especially consequential when fruit orientation is used to infer a peduncle location.

Preprocessing changes the evidence presented to a learned model and adds its own computational cost. Image enhancement within maturity detection [17], illustrated in Figure 4, may improve contrast while also changing color statistics relevant to class assignment. Its contribution therefore requires an ablation using the same model, data partition and latency boundary. Table 4 organizes conventional components by their function in a hybrid pipeline and the assumptions that each component introduces.

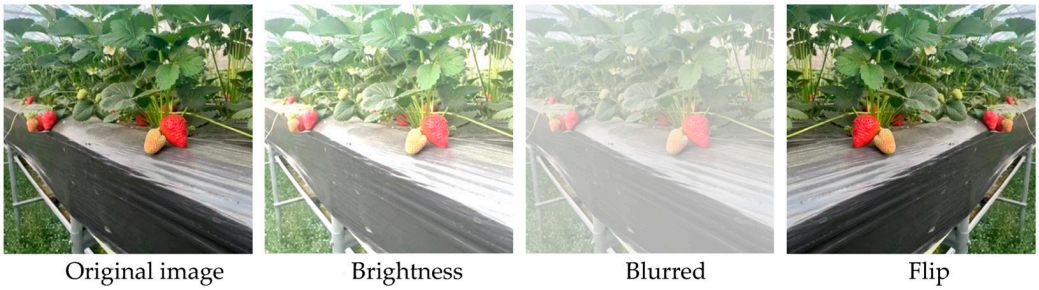


Figure 4. Image-enhancement examples in a strawberry maturity-detection pipeline. (Reprinted from Ref. [17] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Table 4. Roles and testable assumptions of conventional components in hybrid perception.

Processing Component and Sources	Contribution to a Hybrid Pipeline	Assumption to Test	Ref.
Color and region extraction	Provide maturity cues and candidate regions at low computational cost	Appearance remains separable under the tested illumination and cultivar	[7,16]
Contour and axis reasoning	Constrain orientation or infer an attachment region from a detected fruit	The visible contour adequately represents the fruit axis	[22,36]
Image enhancement	Modify contrast before learned detection	Any accuracy gain persists on independent data and exceeds processing cost	[17]
Training-image augmentation	Expose learned recognition to additional appearance variation	Artificial variation represents deployment conditions without corrupting labels	[19]

Hybrid processing is most defensible when geometry constrains a learned output that remains ambiguous. A mask can define the image region used for depth sampling, and a sufficiently complete contour can constrain the fruit axis. Such operations reduce the set of admissible estimates but may introduce bias if the segmentation or shape assumption is wrong. Table 4 consequently separates a component's intended contribution from the condition needed for that contribution to hold. Computational savings should be demonstrated for the complete pipeline, including post-processing and rejected estimates.

A hybrid system should report failure indicators as well as estimated coordinates. Contour completeness, geometric residuals, and agreement between independent cues can

identify cases in which axis-based inference is unreliable. Thresholds should be evaluated across illumination, cultivar, and occlusion conditions. Returning an explicit invalid estimate is appropriate when the available image evidence cannot support the requested action.

3.2. Deep-Learning Detection, Instance Segmentation and Keypoint/Pose Estimation

Deep-learning approaches differ in the spatial and anatomical information available after inference. Object detection provides candidate regions, whereas instance segmentation delineates the visible support of individual fruit against neighboring instances and background. Strawberry-perception systems based on Mask R-CNN and computationally efficient segmentation architectures demonstrate the latter capability [58,59]. Although an instance mask can restrict depth sampling to the target fruit, it cannot directly reveal a peduncle hidden by foliage or positioned outside the visible mask. Segmentation should therefore be regarded as an intermediate representation that supports, but does not replace, interaction-point localization.

Compact mask prediction and contour representation, exemplified by StrawSeg and StrawSnake [60,61], address computational cost and boundary description, respectively. Coupling maturity recognition with registered RGB-D measurements addresses a different transition—from candidate eligibility to metric position [62]. These approaches solve different parts of the perception chain. Their comparison should establish whether a more detailed output reduces downstream depth contamination, pose error, or failed approaches at a comparable computational budget. Table 5 relates each output to the action information still missing, rather than treating image metrics as interchangeable measures of harvesting performance.

Table 5. Information supplied by deep-learning outputs and evaluation needed at the perception-to-action interface. Assessments are proposed criteria, not reported benchmark scores.

Output	Available Information	Remaining Action Requirement	Assessment	Ref.
Detection and maturity	Candidate region and eligibility class	Instance boundary, attachment geometry and valid depth	Class-wise detection and false target selection	[36,62]
Instance mask or contour	Separation of neighboring fruit and visible boundary	Tool interaction point and hidden geometry	Mask error; depth bias	[58–61]
Keypoint detection	Anatomical/action points; Ref. [24] uses calyx and harvesting-point keypoints	Metric position and tool-specific offset	97.3% keypoint mAP; separate 3D: 1.06 mm ($n = 10$) [24]	[24,25,30]
Pose estimation	Relative landmark geometry or interaction direction	Clearance and kinematic feasibility	Position–orientation error relative to tool tolerance	[26,28]
Fruit-axis estimation	Contour-derived axis and an estimated point near the peduncle	Verify the estimated point against the visible stem and define the end-effector pose	90.8% image-plane; 76 ms; occlusion failures; n unreported	[36]

Model architectures offer different compromises among spatial detail, computational cost, and robustness. Region-based detection, instance-mask extensions, and multi-scale feature pyramids support candidate refinement and spatial representation [63–66]. YOLO architectures emphasize efficient dense prediction [67–69], while transformer-based net-

works offer alternative mechanisms for feature association and context modeling [70,71]. Studies involving longan, Zanthoxylum, pepper, tomato, coconut, grape, and apple provide transferable examples of small-object detection, maturity assessment, and anatomical-part recognition [72–84]. However, differences in target scale, peduncle morphology, canopy organization, fruit density, and detachment mechanism preclude a direct ranking of strawberry picking-point localization methods using those cross-crop scores. Their principal relevance lies in motivating strawberry-specific evaluation of feature resolution, anatomical annotations, and instance–structure associations.

Fine structures remain difficult because their image width may approach the effective spatial resolution of the network. Multi-scale features, high-resolution representations, and deformable attention offer relevant mechanisms [85–87], but their benefit must be assessed against memory and latency costs. Generalization tests should separate acquisition sequences, plants, greenhouse rows, sites, and seasons where the dataset permits; random image-level splits may place near-duplicate frames in both training and test sets and overstate performance. Cultivar, maturity distribution, camera distance, illumination, and occlusion should be reported where available, while unreported descriptors should remain explicitly unreported.

Boundary errors matter most where they change the geometric measurements selected for the robot. Background pixels admitted by an instance mask can contaminate depth, while omitted calyx or peduncle pixels can remove evidence needed for an attachment estimate. Figure 5 presents qualitative segmentation comparisons that illustrate such error patterns. In addition to aggregate mask overlap, evaluation should consider boundary error near the interaction region, the proportion of valid depth samples assigned to the correct instance, and the resulting change in 3D point or tool-pose error. The relevant sensitivity depends on whether the tool uses the fruit surface or a narrow attachment region.

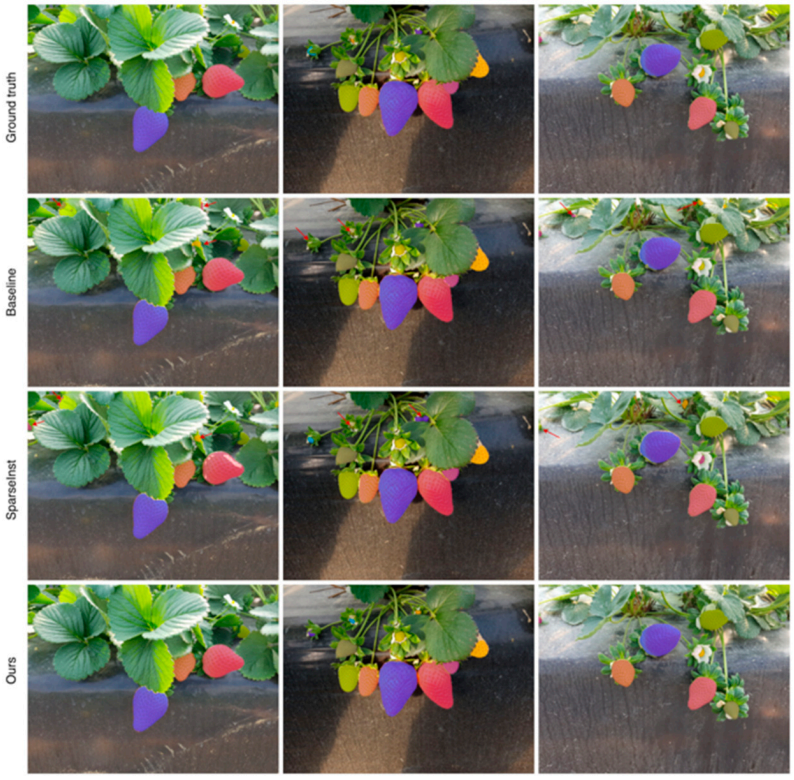


Figure 5. Qualitative comparison of strawberry instance segmentation on StrawDI_Db1 test images; the source uses red arrows to identify inaccurate predictions. (Reprinted from Ref. [60] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Multi-task learning can share features between maturity, segmentation, and landmark prediction, but annotation cost and coupled errors require attention. A maturity error can select the wrong fruit; an instance-association error can assign a valid stem to the wrong fruit; and a landmark error can displace the interaction point. Temporal behavior is equally relevant: a landmark with acceptable average pixel accuracy may fluctuate enough to disturb the final approach. Useful additional measures include visibility classification, frame-to-frame coordinate variation, confidence calibration, and action failure under defined occlusion levels. These measures connect network outputs to manipulator tolerance more directly than a single aggregate accuracy value.

3.3. RGB-D, Stereo Vision and Point-Cloud-Based 3D Localization

The reliability of metric localization depends on depth observability and correct assignment of measurements to the anatomical structure of interest. Registered RGB-D sensors associate color-image targets with measured depth [88,89]; calibrated stereo systems reconstruct depth from disparity and are consequently sensitive to correspondence and baseline geometry [90–93]. Multi-view reconstruction and SLAM combine observations acquired from different camera poses, introducing additional requirements for pose estimation, temporal synchronization, and consistent association [94,95]. In a comparative study of strawberry localization, Ge et al. [88] demonstrated that performance varies with the algorithm, camera, and operating conditions. Sensor selection should therefore be task-dependent; no modality eliminates the need for a clear physical definition of the interaction point.

For the pinhole model, Equations (1) and (2) express back-projection and transformation into robot coordinates. Here, u and v are image coordinates; c_x and c_γ are the principal-point coordinates; f_x and f_γ are focal lengths in pixels; and Z is optical-axis depth. R and t denote the rotation and translation from the camera frame to the robot base frame. Using the convention that ${}^{A}\mathbf{T}_{B}$ maps coordinates from frame B into frame A, a fixed external camera (eye-to-hand) uses the base-from-camera transform ${}^{B}\mathbf{T}_{C}$. For an eye-in-hand camera, the target is transformed through the camera-to-end-effector and time-dependent end-effector-to-base transforms, ${}^{B}\mathbf{T}_{E}(t){}^{E}\mathbf{T}_{C}$. The resulting target point must then be related to the tool center point through the calibrated tool-to-interaction-point offset before a commanded end-effector pose is generated. The transformations assume corrected lens distortion, registered color and depth, consistent units, synchronized robot and camera timestamps, and calibration valid for the acquisition configuration. Figure 6 illustrates the image-space estimation stage in Li et al. [36], in which fruit cropping, binary segmentation, contour extraction, and geometric reasoning yield a candidate point near the stem. Depth association, calibration, and coordinate transformation are still required to obtain a robot-frame target; the image-based estimate alone does not establish final end-effector accuracy.

$$X = \frac{(u - c_x)Z}{f_x},\ Y = \frac{(v - c_y)Z}{f_y} \tag{1}$$

$$p_B = R_{p_C} + t \tag{2}$$

The principal practical difficulty is selecting depth measurements that belong to the required structure. Fruit boundaries can mix foreground and background depth, and thin peduncles may contain no valid samples. A valid sensor return at the target pixel is a direct depth observation; a value interpolated from neighboring pixels, fitted from a local surface, or predicted by a completion network is inferred geometry and should be identified as such. Fruit-shape estimation illustrates the role of segmentation in defining the reconstructed surface [38]. RGB-D saliency methods combine appearance and depth to distinguish foreground objects [96], but saliency does not itself provide a metric picking

point. A neighboring fruit-surface depth should not be assigned automatically to a hidden stem, because image proximity does not establish physical co-location.

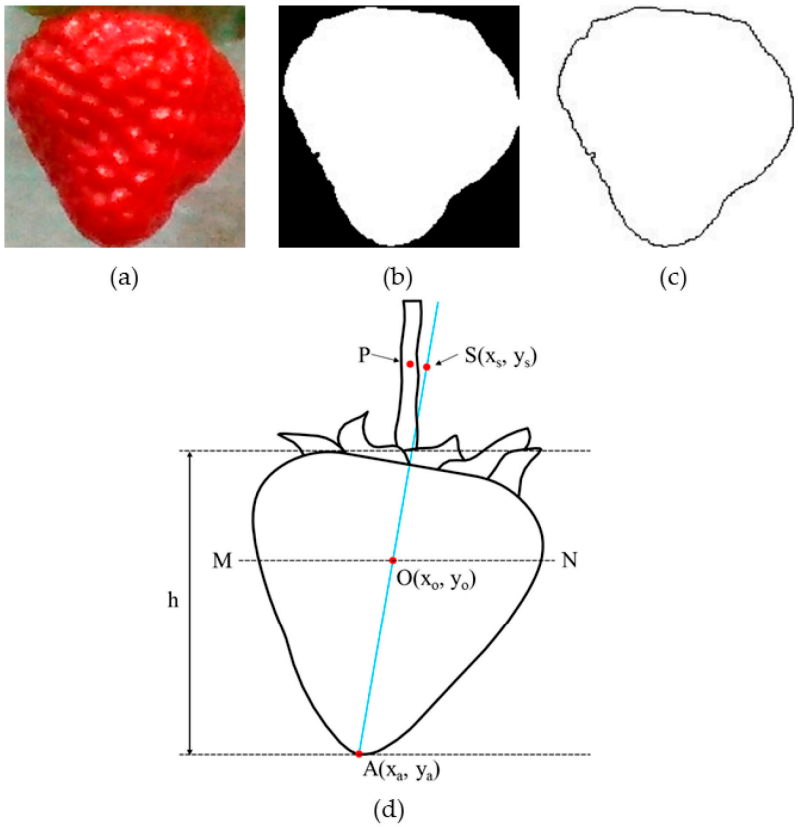


Figure 6. Image-based picking-point localization: (**a**) cropped fruit image; (**b**) binary image; (**c**) edge contour; and (**d**) geometric picking-point estimate. (Reprinted from Ref. [36] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Table 6 contrasts the principal 3D localization routes according to their measurement assumptions and sources of uncertainty. RGB-D back-projection requires valid depth at the relevant landmark, whereas stereo and multi-view reconstruction additionally depend on correspondence and viewing geometry. These considerations extend to coordinate calibration, for which inconsistent reporting can compromise the interpretation of accuracy. For example, Li et al. [36] reported Z-axis calibration-error bounds of 2.7 mm in the abstract and 2.9 mm in the Results section, while one printed coordinate pair implies a 29.5 mm difference. Because the source does not resolve this discrepancy, its summary error values should not be treated as a verified bound on end-effector placement accuracy.

Localization uncertainty is generally direction dependent. Pixel-coordinate error produces larger lateral position error as working distance increases, whereas depth noise and registration bias can dominate along the viewing direction. The uncertainty of the back-projected point should be propagated through the camera-to-robot transform; calibration uncertainty, robot-pose uncertainty, and tool-offset uncertainty should then be considered separately. Random measurement variation should be distinguished from systematic bias, which filtering alone cannot remove. Monocular depth learning and uncertainty-aware visual odometry address estimation from image sequences [97–99], while camera comparisons, multi-view depth observation, and viewpoint selection address complementary measurement conditions [100–102].

Table 6. Comparison of geometric localization routes. Stereo and multi-view examples include other crops; their numerical performance is not transferred to strawberries.

Localization Route and Sources	Metric Information	Suitable Condition	Dominant Limitation	Ref.
Registered RGB-D	Back-projected 3D point when valid depth belongs to the intended landmark; Ref. [24] reports a camera-frame 10-point test	Valid depth at the required keypoint, registered RGB/depth and known intrinsics	Missing/mixed depth on thin peduncles; camera-frame error does not establish robot-frame or tool accuracy	[24,36,37,62,88]
Calibrated stereo	Depth recovered from disparity	Resolvable correspondence and sufficient image detail	Matching ambiguity, occlusion and baseline calibration	[90–93]
Multi-view reconstruction	Geometry fused across observations	Adequate view overlap and consistent scene association	Motion, accumulated pose error and additional acquisition time	[94,95]
Local surface or shape model	Surface geometry and inferred orientation	Sufficient observed points on the fruit	Partial visibility and model assumptions near attachment; inferred surfaces must be distinguished from measured depth	[38]
Visual–inertial estimation	Motion-constrained temporal localization	Synchronized visual and inertial measurements	Inertial constraints on camera motion do not directly measure fruit motion	[46]

These measurement differences imply that sensor selection should follow cultivation geometry, working range, target scale, and the required interaction tolerance. Comparisons should report color–depth registration and valid-depth coverage for RGB-D systems, disparity and calibration conditions for stereo systems, and motion or viewpoint assumptions for multi-view estimation. For all modalities, validation should separate image landmark error, camera-frame 3D error, robot-frame error, and final tool-to-target error. Error distributions and failure rates should be reported across working distances and visibility conditions rather than inferring tool accuracy from a camera specification or a single mean error.

3.4. Edge Deployment and Multimodal Fusion

Lightweight perception models address only part of the computational demand of robotic harvesting. Depth registration, scene interpretation, motion planning, communication, and control also contribute to end-to-end latency. Pruned strawberry landmark detectors and compact pose-estimation models seek to reduce this burden [24,26], while MobileViT illustrates a related design direction in mobile vision [103]. StrawPose-Lite, for example, reports 0.73 million parameters and 79.2% pose mAP@0.5:0.95 on a specified public validation split [26]. These measures quantify network complexity and image-level estimation, rather than complete sensing-to-command delay or closed-loop harvesting reliability. Deployment comparisons must therefore specify the test data, computing hardware, numerical precision, and boundaries of the timed processing pipeline.

Sensor fusion can address complementary measurement limitations when the role and reliability of each measurement are explicit. RGB supplies appearance, depth provides metric geometry where valid returns are available, and an inertial measurement unit (IMU) constrains sensor motion rather than directly measuring fruit motion. Local gripper sensing can support correction when the visual target becomes obscured near contact [46,104]. Multimodal tomato perception and sensorized raspberry harvesting provide cross-crop examples of complementary sensing [105,106], but their performance should not be transferred directly to strawberry systems with different target geometry and contact conditions.

A hierarchical sensing design can allocate computation according to the current operation. Scene perception identifies and ranks candidates; local depth and geometric estimation determine the selected target pose; and near-field sensing supports final alignment. Figure 7 presents the edge-deployment workflow of StrawPose-Lite. Its relevance is the integration of inference with runtime processing, rather than model compactness considered in isolation.

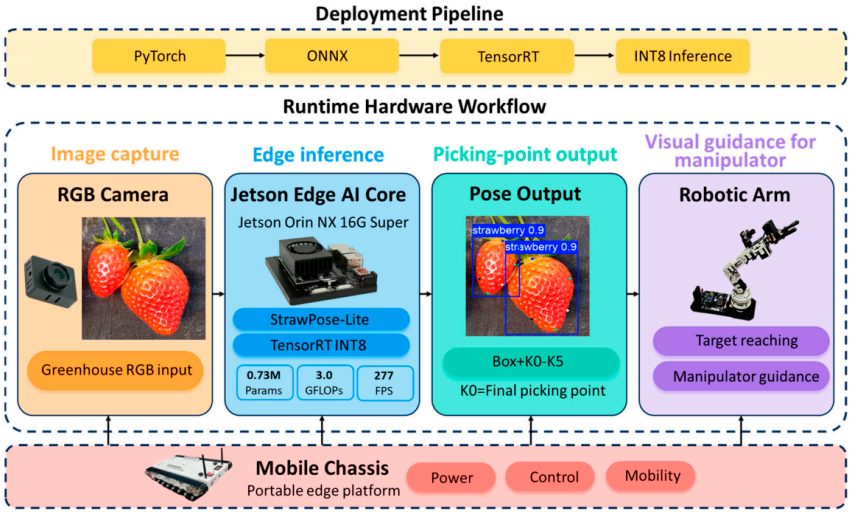


Figure 7. Edge-deployment architecture and runtime workflow of StrawPose-Lite. (Reprinted from Ref. [26] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Deployment evaluation should report acquisition, preprocessing, inference, depth association, coordinate transformation, communication, and controller-update delays. Mean latency, high-percentile latency, and jitter characterize different aspects of performance. Sustained measurements under representative thermal, power, and illumination conditions are necessary because short inference benchmarks may not represent continuous operation. Compression should also be assessed for disproportionate errors on thin peduncles and partly occluded landmarks, and the field evaluation set should be independent of training and model-selection data.

Fusion should account for the reliability and temporal relevance of each measurement rather than assign equal confidence to every modality. IMU data can constrain camera motion without directly observing fruit motion; local tactile or proximity measurements apply only within their operating range. Event-based vision provides low-latency temporal information and has been studied for detection and reconstruction [107–109], but its benefit for strawberry picking requires direct validation against the additional sensing, calibration, and processing requirements.

3.5. Study Context and Validation Boundaries

Comparative interpretation requires attention to the environment and the level at which performance is validated. For instance, Dai et al. [24] evaluated 3D keypoint localization on real strawberries in a simulated ridge-growing environment, separately from image-based keypoint recognition. Li et al. [36] separately assessed geometric positioning and robotic stem grasping using plastic strawberries under controlled conditions, whereas Tufail et al. [47] evaluated an integrated robot on indoor synthetic targets. By comparison, Yu et al. [41] reported stage-specific results from natural-fruit greenhouse harvesting. These experiments characterize distinct levels of technical maturity but differ in target material, cultivation geometry, coordinate frame, and physical interaction. Moreover, the two greenhouse conditions evaluated by Yu et al. [41] differed in location and test date as well as canopy management; their comparison therefore does not isolate the effect of thinning.

Accordingly, quantitative synthesis must distinguish the estimated quantity, reference measurement, experimental setting, and definition of success. Keypoint mAP, camera-frame 3D error, geometric positioning success, stem-grasp success, and completed fruit detachment are different outcomes. The reported measurements for keypoint localization [24], plastic-fruit positioning and grasping [36], natural-fruit greenhouse harvesting [41], and synthetic-target manipulation [47] are thus informative at different stages of the perception-to-action process. Table 7 retains their evaluation conditions and denominators, where reported, to support evidence-aware interpretation rather than an artificial overall ranking.

Table 7. Selected quantitative results for perception, 3D localization, calibration-related assessment and robotic harvesting. The endpoints, denominators, target materials, experimental settings and timing boundaries differ; values are not pooled or ranked.

Study/Setting	Reported Outcome	Target and Sensing Conditions	Timing/Evidence Boundary	Ref.
Simulated ridge setup; real strawberries	Keypoint mAP: 97.3% (image test); separate 3D test (n = 10): 0.5–1.6 mm, mean 1.06 mm	Real strawberries; front-mounted D435i (>200 mm); manual reference coordinates	Camera-frame test only; no robot-frame or harvesting outcome	[24]
Simulated elevated-substrate scene	90.8% image-plane positioning success; denominator not reported	23-pixel segment around point intersects stem; no/slight occlusion; severe failures	76 ms positioning time; not complete robot-cycle time	[36]
Plastic-fruit robot trial	89/98 ripe targets = 90.8% stem-grasp success	Plastic strawberries; fixed end-effector posture; stem gripped, not cut	7.5 s/fruit; source states 124 total, but 98 + 25 + 14 = 137	[36]
Ridge greenhouse robot	105/213 = 49.30% thinned; 65/215 = 30.23% untrimmed	A: Changping, Jan 2024; B: Xuzhou, Mar 2024; DF500 (640 × 480) + D435i; cultivar NR	7.3 s single-arm; 4.7 s dual-arm; damage 12.68% vs. 19.53%; sites/dates differ	[41]
Indoor synthetic-fruit trial	86.5% peduncle-side region identification (200 images); 36/50 = 72% synthetic-fruit harvests	YOLOv11-small; image-based point test; separate indoor xArm 6 trial on synthetic fruit	14 failures: 6 detection and 8 unstable grasp; no real-fruit field test	[47]

For the selected primary studies [24,36,41,47], the comparison consistently distinguishes target material, sensing and cultivation conditions, the estimated quantity, the coordinate frame, sample denominator where available, and level of robotic validation. The 1.06 mm camera-frame error in [24], the positioning and plastic-stem grasp results in [36], the natural-fruit greenhouse outcomes in [41], and the indoor synthetic-target trials in [47] thus represent different types of evidence. This selected comparison does not imply recovery of unreported variables or a complete re-extraction of every primary study.

4. Perception-to-Action Applications in Strawberry Harvesting

Visual measurements acquire operational meaning when they determine target eligibility, tool pose, approach constraints, or feedback corrections. The reliability of these transformations is a central criterion for evaluating harvesting perception. A system can fail at any interface even when its individual components perform well under isolated tests. A useful execution sequence comprises candidate detection, qualification, metric pose estimation, approach validation, and feedback during interaction. Later stages can reject or refine earlier estimates. A mature target can be deferred after a reachability check, and an uncertain pose can be corrected using a new observation. Retaining intermediate variables makes such decisions traceable and supports diagnosis of repeated failures.

4.1. Pickability Assessment and Picking-Sequence Decision

Pickability screening changes the trade-off between attempting difficult fruit and spending time on unsuccessful motion. Explicit detection and pickability classification [21] provide an early decision point, but their benefit depends on the cost of false rejection as well as failed attempts. Workspace, clearance and localization confidence are mandatory constraints; maturity and travel distance can rank the candidates that satisfy them. Evaluating only the success rate of attempted picks would favor conservative rejection without revealing how many harvestable fruit were missed. Target coverage and additional observation time are therefore necessary companion measures.

Picking order couples motion efficiency to changes in the scene. Removing an exposed fruit may reveal an inner target, while reaching into a cluster may disturb neighboring organs. Dual-arm conflict avoidance and scheduling address interference and workload distribution [110,111]; sequence planning and multi-arm allocation address related travel and coordination costs [112–115]. For strawberry clusters, these costs must be evaluated together with the observations made obsolete by each action. A short nominal path can become inefficient if plant disturbance repeatedly forces localization to restart.

Integrated strawberry-harvesting systems connect target selection with manipulation and scene updates [116–118], but successful module integration does not establish robust operation across cultivation settings. The evidence summarized in Table 7 spans camera-frame localization, controlled plastic-fruit grasping, natural-fruit greenhouse harvesting, and indoor synthetic-target trials. Yu et al. [41] reported overall successful-harvest rates of 49.30% and 30.23% in their thinned and untrimmed greenhouse conditions, respectively. As the two trials also differed in location and date, this contrast should not be interpreted as the isolated effect of canopy management. More broadly, differences in target material, endpoint definition, and timing prevent the percentages reported across these studies from supporting a universal ranking of algorithms or systems.

Pickability is conditional on the target, surrounding scene, robot, and tool. Its estimate can change after another fruit is removed or after the camera obtains a clearer view. A decision system should retain the specific reason for rejection, such as missing peduncle depth, insufficient clearance, or an unattainable wrist orientation. This allows revisiting policies to distinguish potentially recoverable targets from persistent geometric constraints.

Fruit removal and plant disturbance invalidate parts of the previous scene estimate. The next target should therefore be selected from an updated local representation, rather than a fixed list created before harvesting. Figure 8 illustrates the sequence of an integrated ridge-grown system. Sequence-level outcomes include newly exposed targets, success on revisited fruit and completed picks per operating time. These outcomes reveal whether observation and recovery costs offset an apparently short duration for successful individual picks.



Figure 8. Execution sequence of the ridge-grown strawberry harvesting robot. (Reprinted from Ref. [41] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

4.2. 3D Picking-Point Localization and End-Effector Alignment

End-effector alignment requires a task-specific pose. Peduncle cutting combines position, blade orientation, and clearance around the cutting region. Fruit-contact strategies additionally require a suitable surface normal or gripping direction. Tool offsets and admissible position–orientation tolerances should therefore be defined before assessing the adequacy of visual localization.

The relevant error budget includes landmark localization, depth measurement, color–depth registration, camera intrinsics, hand–eye calibration, robot positioning, structural compliance, and timing. These contributions need not be independent and can produce systematic as well as random deviations. A compliant or locally sensorized gripper can accommodate some positioning error [104], but mechanical tolerance does not remove the need to characterize the visual estimate.

The evaluation of strawberry picking-point localization should distinguish recognition, geometric estimation, and physical execution. Dai et al. [24] reported 97.3% keypoint mAP and, in a separate ten-point test on real strawberries in a simulated ridge setup, a mean camera-frame Euclidean error of 1.06 mm; neither metric directly measures final tool placement. Li et al. [36] reported 90.8% image-plane positioning success (23-pixel stem-intersection criterion) with a mean positioning time of 76 ms. Their subsequent robotic trial achieved 89 successful stem grasps among 98 ripe plastic targets, without cutting the stem. Although the two reported percentages are similar, they represent distinct experimental endpoints. Related simulated localization and harvesting studies also require explicit validation boundaries [29]. Accordingly, localization assessment should connect visual estimates to tool-frame error distributions and, ultimately, to real-fruit detachment, damage, and cycle-time outcomes.

Calibration bias and temporal variation call for different corrections. Recalibration can reduce a persistent coordinate offset, whereas repeated observation or motion estimation addresses an estimate that becomes stale during approach. Multi-frame filtering may suppress random variation at the expense of lag. The hand–eye arrangement in Figure 9 forms part of a visual–inertial localization system, but calibration accuracy alone does not quantify its dynamic alignment performance. The two contributions should be evaluated separately before their combined effect is measured at the tool.

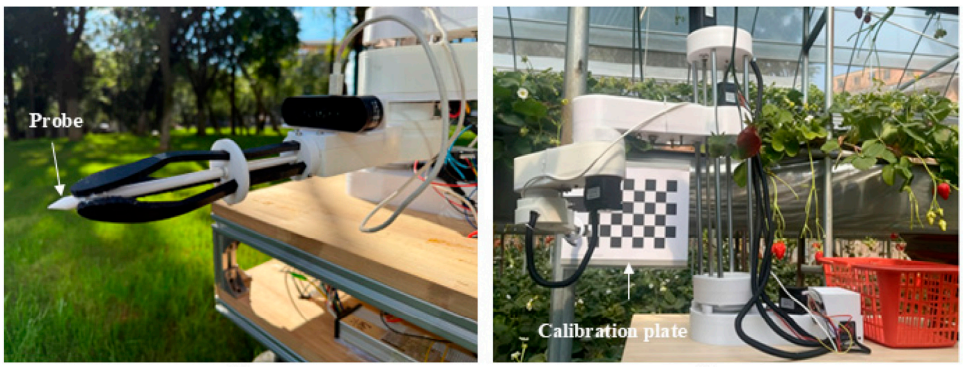


Figure 9. Hand–eye calibration in a visual–inertial strawberry localization system: (**a**) end-effector probe; and (**b**) calibration arrangement. (Reprinted from Ref. [46] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

The usable localization tolerance is set by the interaction geometry. A cutter must remain within an admissible stem segment while excluding the fruit body and adjacent peduncles; a compliant gripper may accept larger position errors but still requires controlled force and orientation. Thus, an identical spatial uncertainty can produce different action risks for different tools. Table 8 decomposes the error sources so that a failure can be assigned to image estimation, depth, calibration, timing or actuation. This decomposition supports a targeted correction rather than assuming that a larger perception model will improve every localization failure.

Table 8. Proposed error-budget audit for localization and alignment. Source-reported numerical examples retain their original validation boundaries; the listed additional measurements are review recommendations rather than reported experimental results.

Error Component and Examples	Effect on Tool Placement	Measurement Needed to Isolate the Effect	Ref.
2D keypoint detection [24]	Pixel error propagates through depth association; mAP is not 3D accuracy	Visible/occluded landmark error plus a separate camera-frame 3D point error	[24]
Depth sensing and registration [100]	Depth bias or wrong-surface assignment shifts the recovered point	Valid-depth rate and error by range, surface and illumination	[100]
Camera-to-end-effector calibration [36]	Table 4 test 13 implies a 29.5 mm Z discrepancy; source summaries conflict	Reconcile all 25 points; report per-axis MAE/RMSE/max and independent robot-frame validation	[36]
Motion and synchronization [46]	A valid estimate becomes stale during approach	Timestamp alignment, temporal residuals and dynamic tool-to-target error	[46]
Learned depth and pose uncertainty [99]	Scale or pose uncertainty affects geometric confidence	Uncertainty calibration and metric-scale verification	[99]
Integrated actuation [119]	Kinematic, compliance and control errors remain after visual localization	Final tool-to-target position and orientation over the working envelope	[119]
Camera-frame RGB-D localization [24]	Depth association and keypoint error jointly determine 3D deviation	10 manually referenced points; repeat by range/occlusion and state the coordinate frame	[24]

A meaningful localization evaluation should distinguish image-space landmark error, camera-frame 3D error, robot-frame positioning error, and final tool-to-target deviation. Repeated tests across working distances, camera poses, and manipulator configurations can identify systematic biases and long-tailed failures. Reprojection accuracy and mean depth error alone cannot characterize the effects of robot kinematics, structural compliance, and temporal misalignment. The unresolved coordinate inconsistency in Li et al. [36] further illustrates the importance of reporting raw measurements together with summary statistics and clearly defined outlier handling. Such a decomposed error budget supports targeted improvements in sensing, model estimation, calibration, and actuation.

4.3. Occlusion Handling, Active Perception, Obstacle Avoidance and Closed-Loop Visual Servoing

Active perception is useful when uncertainty can be reduced by a feasible change in observation. Learned viewpoint selection, active viewing of occluded grape targets and graph-based view-motion planning provide different ways to select such observations [120–122]. The shared objective is to acquire information relevant to a pending decision. For strawberries, the limiting quantity may be attachment depth or tool clearance rather than the visible area of the fruit. An additional view is advantageous only if its improvement in localization or decision reliability justifies the motion time and collision exposure. This condition should be tested across cluster density and visibility levels.

Collision checking requires local scene geometry and a model of the tool envelope. Semantic labels can distinguish rigid supports, vulnerable fruit, and flexible foliage, but class membership alone does not justify contact. Any permitted interaction with leaves should be constrained and experimentally validated. Cross-crop continuous-picking and path-planning work offers relevant geometric methods [123,124]; its action costs and contact assumptions require adaptation to strawberry plants.

Feedback reduces residual misalignment only while the relevant error remains observable and correctable. Local gripper sensing can provide near-contact information [104], while vision-based fault diagnosis can identify a need for recovery after unsuccessful interaction [125]. Position-based visual servoing is also integrated in the platform illustrated in Figure 10 [47]. These functions differ from improving the initial target estimate: they respond to error during or after execution. Synthetic-target trials establish a limited part of this capability because compliant real fruit and plant motion can alter both the sensed residual and the response to corrective contact.

An interactive perception policy should identify when current evidence is inadequate, select a new observation, and reassess the intended action. This requires explicit confidence and failure states. A maximum-likelihood coordinate without an indication of visibility or depth validity cannot support a principled decision to continue, defer, or reacquire the target.

A control policy should distinguish information shortage from physical infeasibility. Missing attachment depth can justify another observation; a blocked approach may require replanning; a detected failed grasp may justify recovery. These actions have different costs and should not be triggered by a single generic confidence threshold. Table 9 links the limiting condition to a candidate response and an outcome measure. Persistent target identity is needed across these transitions so that newly exposed fruit, repeated observations and revisited targets remain distinguishable.

Integration determines whether these corrections remain useful in the complete harvesting sequence. Combining ordering, visual servoing and end-effector design in cucumber harvesting [126] illustrates the coupling between target selection and local control. Integrated sweet-pepper, apple and lettuce systems provide related evidence about coordinating perception with planning and execution [119,127–129]. Their common implication for straw-

berries is architectural: feedback must act on a pose that remains reachable and compatible with the tool. Crop-specific detachment forces, contact tolerance and plant response still require direct validation.

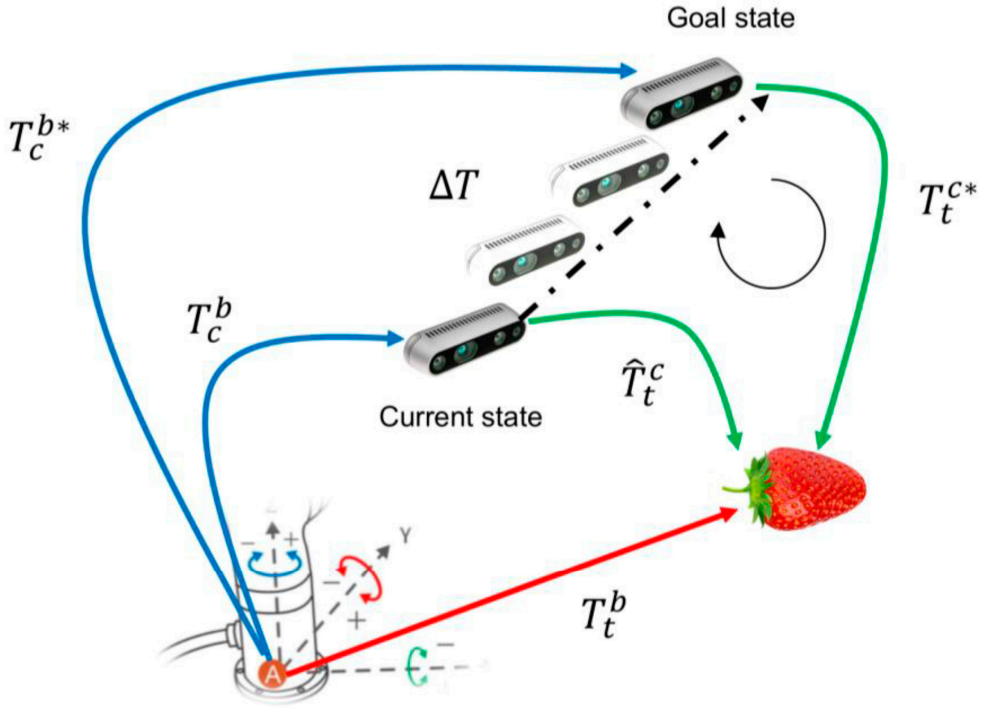


Figure 10. Position-based visual servoing for manipulator approach to a detected strawberry. (Reprinted from Ref. [47] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

Table 9. Decision matrix for additional observation, replanning and recovery. Candidate responses and outcome measures synthesize the cited approaches rather than reproduce study-specific protocols.

Trigger and Sources	Candidate Response	Cost or Residual Constraint	Outcome to Assess	Ref.
Hidden attachment or incomplete view	Select an informative reachable viewpoint	Extra travel and sensing time	Resolved landmarks and completed picks per operating time	[45,122]
Blocked approach corridor	Change tool approach or defer the target	Observation alone cannot remove physical obstruction	Collision-free access, damage and deferral rate	[44]
Scene change after a pick	Re-observe and update target identity and pose	Risk of duplicate targets or stale geometry	Reassociation accuracy and success on revisited targets	[41]
Detected harvesting fault	Diagnose failure and execute a bounded recovery action	Recovery can add delay or disturb the plant	Recovery success, added time and secondary damage	[125]
Competing dual-arm tasks	Reassign or reschedule targets with conflict checks	Coordination overhead and shared workspace constraints	Completed picks, interference and arm idle time	[110,111]

4.4. Integrated Perception–Information–Actuation Framework

A perception-to-action architecture should associate each visual output with a specific robotic decision. Ripeness and identity support candidate qualification; instance masks delimit image regions used for depth sampling; anatomical landmarks define the intended tool interaction; and metric reconstruction provides a spatial target in the robot coordinate frame. Reachability analysis, collision checking, and feedback then determine whether and how the pose can be executed. Table 10 presents this evidence-informed mapping. Its evaluation metrics are proposed synthesis criteria and should not be mistaken for results reported by every cited study.

Table 10. Evidence-informed mapping between perception information, robotic decisions, and evaluation measures. The mapping is synthesized by the authors; proposed measures are not attributed to individual studies unless explicitly reported.

Perception Input	Representation	Robotic Decision	Evaluation Criteria
Maturity and identity	RGB/spectral maturity class and confidence [16–21]	Qualify candidate; rank after feasibility	Class-wise metrics; false-pickable/rejection; maturity ≠ access
Boundary/visibility	Instance mask; visible/occluded state [58–62]	Select depth pixels; re-observe if needed	Mask/landmark metrics by occlusion; valid-depth coverage
Picking geometry	Peduncle/calyx, contact region and fruit–stem association [24–30]	Set tool interaction point, orientation and offset	3D pose error vs. tool tolerance; observed vs. inferred
3D localization	Fixed camera: C → B; eye-in-hand: C → E → B; apply tool offset [24,36,37,88–95]	Generate robot-frame target pose	Per-axis/Euclidean error by frame; valid depth; calibration
Accessibility	Obstacles, occupancy and tool swept volume [37,41–45]	Reachability/collision; approach, re-plan or defer	Clearance, collisions/aborts, damage and deferrals
Feedback	Time-stamped target ID, tool residual and local sensing [46,47,104,125]	Update pose; re-observe, correct or recover	Stale-pose error, recovery success and cycle time

Consistency among representations is essential when transferring perception estimates to a robot controller. A keypoint annotated at the calyx center may differ from the cutting segment required by the tool, while a numerically valid depth sample may belong to an adjacent fruit. Figure 11 reproduces the control architecture of a specific ridge-grown strawberry harvester developed by Yu et al. [41]; it is retained as an example of an implemented system, not as the conceptual framework proposed here. Figure 12 instead presents the author-derived synthesis of visual information, 3D localization, uncertainty, access constraints, and execution feedback. A complete evaluation should verify the anatomical definitions, coordinate-frame transforms, tool offsets, and timestamps linking these stages.

Figure 12 summarizes the evidence-aware perception-to-action chain developed from this review: candidate qualification, instance separation, action-structure representation, depth-based 3D reconstruction, coordinate transformation, uncertainty propagation, execution feasibility, and action feedback. The framework also places experimental context and cross-stage evaluation alongside the processing chain. Table 10 translates these information layers into robotic decisions and suitable assessment measures. Image metrics characterize recognition and landmark prediction; geometric metrics quantify position and orientation error in the stated coordinate frame; execution metrics include successful detachment, fruit

damage, collision or abort frequency, and recovery; and deployment metrics describe latency and sustained operation. The denominator and inclusion of failed attempts must be stated for every success or throughput measure.

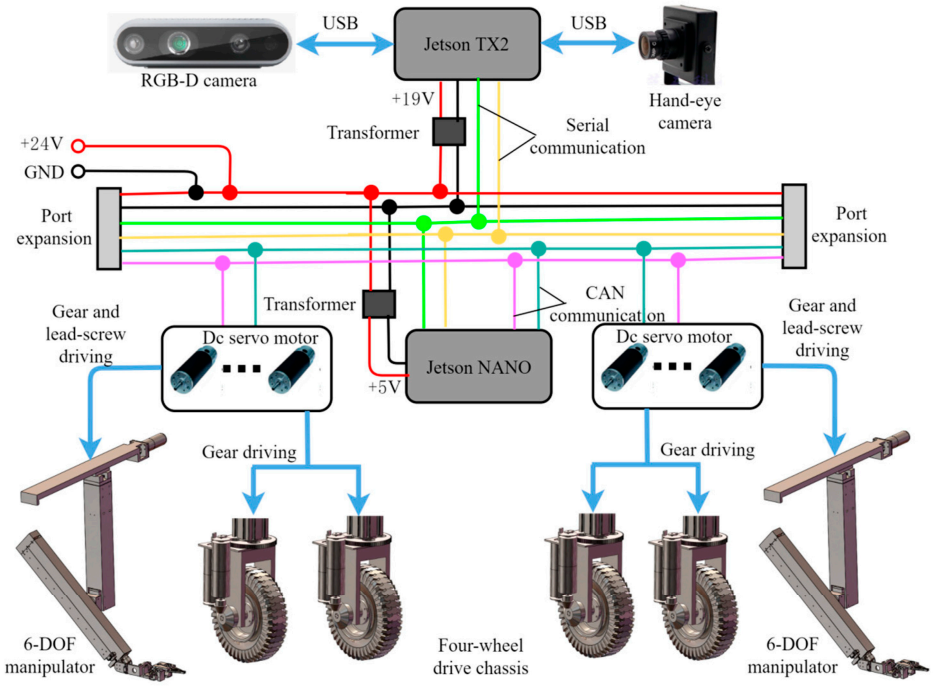


Figure 11. Source-specific integrated control architecture of the ridge-grown strawberry harvesting robot reported in Ref. [41]. (Reprinted from Ref. [41] under the Creative Commons Attribution 4.0 International license (CC BY 4.0).).

The greenhouse evaluations by Yu et al. [41] illustrate how the mapping in Table 10 can be used to interpret stage-specific system outcomes. Target detection, cutting–clamping, and final harvesting success all differed between the reported thinned and untrimmed conditions. The documented failures involved missed targets, occluded stems, inappropriate triggering, insufficient gripping force, and fruit loss during transport, implicating perception, geometric estimation, tool interaction, and post-detachment handling, respectively. These observations indicate that overall harvesting reliability reflects several coupled bottlenecks rather than a single vision-model limitation. Since location and test date differed alongside canopy treatment, the reported contrast should not be assigned exclusively to thinning.

Global scene processing and local feedback can operate at different update rates. Candidate ranking and broader mapping can be refreshed as the scene changes, while a lightweight local estimator updates the selected target during approach. This architecture can reduce computational contention, provided that timestamps and stale-data handling are explicit. A fast controller operating on outdated perception does not provide accurate feedback solely because its update rate is high.

A comparable benchmark should specify cultivation mode, cultivar, sensing geometry, tool type, picking-point definition, and test difficulty. Perception and robotic outcomes should be measured on corresponding targets where possible, allowing errors to be traced across the pipeline. Continuous sequences should include failures, target deferrals, and recovery time. These conditions permit assessment of whether an image-level improvement remains beneficial at the point of physical interaction.



Figure 12. Evidence-aware perception-to-action framework for strawberry picking-point localization and robotic harvesting. This conceptual synthesis was developed by the authors from the studies reviewed in this article; the evaluation layer distinguishes reported outcomes from proposed assessment criteria.

5. Current Challenges and Future Development Directions

5.1. Main Challenges

The challenges of strawberry harvesting perception arise from the interdependence of biological eligibility, anatomical observability, and tool-specific accessibility. A maturity label indicates biological suitability under a defined harvesting standard, but does not ensure that the attachment region is visible or that the intended tool can reach it safely. Peduncles occupy a limited image area and are frequently obscured by foliage, neighboring fruit, or support structures [20,21,24,29]. Moreover, fruit centroids, calyx keypoints, stem centerlines, and tool-specific cutting or gripping regions impose different geometric constraints. Without consistent anatomical definitions across annotation, localization, and control, strong recognition performance may not produce an executable picking pose.

Metric localization is vulnerable to uncertainty introduced throughout the sensing and transformation chain. Occluded or slender peduncles can yield invalid or mixed depth, while registration, intrinsic calibration, hand–eye transformation, robot pose, tool offset, and sensing-to-actuation delay can each influence the final interaction pose [37,38,46,88]. Their contributions depend on viewing geometry, working distance, and manipulator configura-

tion, and may be systematic rather than purely random. Results from limited camera-frame tests [24] and inconsistently reported calibration measurements [36] emphasize the need to identify the reference frame, retain raw measurements, and describe uncertainty beyond a single mean error. The error categories in Table 8 support a distinction between camera-level accuracy and the reliability of final tool placement.

Generalization remains an unresolved issue because both visual appearance and manipulation geometry vary among cultivation conditions. Differences in cultivar, maturity stage, fruit morphology, and peduncle presentation can alter recognition cues and landmark visibility [16–21]. Ridge versus elevated-substrate cultivation, canopy density, and camera placement also affect occlusion and accessible workspace, while sunlight, cast shadows, reflections, and exposure variation modify the image distribution [17–19]. The existing evidence includes a simulated ridge setup with real strawberries, plastic or synthetic targets, and natural-fruit greenhouse operation [24,36,41,47], but these settings are not interchangeable tests of cross-cultivar or cross-site robustness. Claims of general applicability therefore require independent evaluation across acquisition sessions, cultivars, cultivation systems, and environmental conditions.

A further limitation is the incomplete connection between visual performance and harvesting outcomes. Detection accuracy, mask overlap, landmark metrics, 3D localization error, stem-grasp success, and actual fruit detachment describe different stages. As illustrated in Table 7, their meaning depends on the target material, denominator, and definition of success. Random image-level partitions may also overestimate robustness when frames from the same plants or collection sequences appear in both training and testing. This is a general methodological risk rather than evidence that every reviewed study has such leakage. Future comparisons should state the partitioning unit and evaluate perception on matched targets alongside tool positioning, successful detachment, damage, deferred attempts, and recovery time.

5.2. Future Directions

Future progress should prioritize task-oriented visual representations that reflect the geometry and constraints of strawberry detachment. Joint prediction of maturity, fruit instances, fruit–peduncle association, structural landmarks, pose, and confidence could improve the consistency of information supplied to the manipulator. However, any advantage over modular pipelines must be demonstrated using comparable datasets, annotation definitions, and computational resources. Crucially, inferred landmarks and completed geometry should be distinguished from direct observations, and uncertainty should inform whether a pose can be executed or further sensing is required.

Complementary sensing and active observation may help resolve targets whose attachment geometry is incompletely observed. Registered RGB-D provides metric measurements where valid depth is available; stereo and multi-view methods can contribute additional geometric constraints when correspondence and pose estimation are reliable [37–40,45]. Local proximity or tactile feedback may further support alignment during contact [104]. These modalities should be fused according to measurement validity, temporal relevance, and uncertainty rather than treated as equally reliable. Viewpoint changes should be justified by their expected benefit to attachment visibility or approach clearance, accounting for additional motion and sensing time.

Calibration and localization should be validated across the full camera-to-tool transformation chain. Independent reference targets positioned throughout the workspace, repeated measurements at different distances and robot configurations, and explicit characterization of camera intrinsics, hand–eye transformation, robot pose, and tool offsets would help distinguish systematic and random errors. Validation should report accuracy in the

relevant camera, robot, and tool frames rather than equating calibration residuals with harvesting accuracy. Time synchronization is equally important because a geometrically accurate estimate may become outdated before contact.

Real-time claims should be supported by measurements of the complete perception-and-control pipeline. Lightweight networks and model compression reduce inference cost, but acquisition, depth registration, coordinate transformation, motion planning, communication, and controller updates also contribute to latency [26,60]. Future evaluations should document the hardware, precision mode, processing boundaries, latency distribution, temporal stability, and sustained performance under representative environmental conditions. Network frame rate alone is insufficient to characterize the responsiveness of a harvesting robot.

Robust generalization and meaningful cross-study comparisons require standardized yet task-relevant evaluation protocols. Where images are temporally or spatially correlated, training, validation, and testing should be separated by plant, row, plot, acquisition session, or site rather than by random frames alone. External validation should consider held-out cultivars, cultivation systems, seasons, and illumination conditions. Studies should report maturity criteria, target material, camera geometry, working distance, occlusion, end-effector design, picking-point definition, and success criteria. Whenever possible, detection and localization should be assessed on the same fruit instances used for harvesting, with 3D position and orientation error, successful detachment, fruit damage, retries, deferred targets, and cycle time recorded together. Such experiments are necessary to establish whether improvements demonstrated in controlled settings translate to reliable performance on real fruit.

6. Conclusions

Strawberry picking-point localization is best understood as an integrated inference and manipulation process rather than a stand-alone visual detection problem. The reviewed methods contribute complementary information about fruit maturity, instance identity, anatomical attachment structures, metric geometry, and action constraints. Their operational value depends on whether these outputs can be combined to estimate a physically valid interaction pose and determine when that pose is sufficiently reliable for execution. Accurate recognition is therefore necessary but not sufficient for robust robotic harvesting.

The literature demonstrates advances in fruit recognition, anatomical keypoint estimation, and controlled 3D localization, but the extent of physical validation remains heterogeneous. Results obtained from simulated coordinate tests, plastic or synthetic targets, and natural-fruit greenhouse operation address different research questions and cannot be combined into a universal performance estimate. In particular, image-level accuracy does not directly quantify robot-frame positioning, actual fruit detachment, or damage risk. Interpretation of reported performance must therefore preserve the experimental context, reference frame, target material, and success definition.

The principal unresolved challenges occur at the interfaces linking visual evidence to physical action. Occlusion and ambiguous landmark definitions limit attachment-structure observability; invalid or mixed depth and calibration uncertainty affect metric reconstruction; and tool clearance, plant motion, and sensing-to-actuation delays determine whether a predicted pose remains executable. These problems are particularly consequential in dense canopies and for harvesting mechanisms with narrow position or orientation tolerances. Their evaluation requires explicit uncertainty characterization, consistent anatomical and coordinate-frame definitions, and measurements that extend from image perception to final tool interaction.

This review contributes an action-oriented synthesis that links biological eligibility, anatomical representation, metric localization, uncertainty propagation, tool-dependent feasibility, and execution feedback within a common analytical framework. Beyond comparing individual algorithms, it identifies where seemingly accurate perception outputs may fail to support reliable harvesting. Priorities for future investigation include independent cross-cultivar and cross-environment validation, uncertainty-aware depth and landmark estimation, coordinated camera-to-tool calibration, and matched-target assessment of localization error, detachment success, fruit damage, and cycle time. These developments are essential for demonstrating that improvements in visual perception yield measurable benefits during complete strawberry-harvesting operations.

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

Funding: This research was funded by the Central Government-Guided Local Science and Technology Development Fund Project, grant number YDZX2025015; the Shandong Provincial Rural Revitalization Science and Technology Innovation Promotion Action Project, grant numbers 2025TZXD034 and 2023TZXD057; and the Key Research and Development Program of Shandong Province, grant number 2022CXPT017.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article. The information analyzed is derived from the published studies cited in the manuscript.

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek V3.2 for language translation and polishing. The authors reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Huang, Y.; Xu, S.; Chen, H.; Li, G.; Dong, H.; Yu, J.; Zhang, X.; Chen, R. A review of visual perception technology for intelligent fruit harvesting robots. *Front. Plant Sci.* **2025**, *16*, 1646871. [CrossRef] [PubMed]
2. Zhang, J.; Kang, N.; Qu, Q.; Zhou, L.; Zhang, H. Automatic fruit picking technology: A comprehensive review of research advances. *Artif. Intell. Rev.* **2024**, *57*, 54. [CrossRef]
3. Zhou, H.; Wang, X.; Au, W.; Kang, H.; Chen, C. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precis. Agric.* **2022**, *23*, 1856–1907. [CrossRef]
4. Suresh Kumar, M.; Mohan, S. Selective fruit harvesting: Research, trends and developments towards fruit detection and localization—A review. *Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci.* **2022**, *237*, 1405–1444. [CrossRef]
5. Xiao, X.; Wang, Y.; Jiang, Y. Review of research advances in fruit and vegetable harvesting robots. *J. Electr. Eng. Technol.* **2024**, *19*, 773–789. [CrossRef]
6. Hayashi, S.; Shigematsu, K.; Yamamoto, S.; Kobayashi, K.; Kohno, Y.; Kamata, J.; Kurita, M. Evaluation of a strawberry-harvesting robot in a field test. *Biosyst. Eng.* **2010**, *105*, 160–171. [CrossRef]
7. Rajendra, P.; Kondo, N.; Ninomiya, K.; Kamata, J.; Kurita, M.; Shiigi, T.; Hayashi, S.; Yoshida, H.; Kohno, Y. Machine vision algorithm for robots to harvest strawberries in tabletop culture greenhouses. *Eng. Agric. Environ. Food* **2009**, *2*, 24–30. [CrossRef]
8. Feng, Q.; Wang, X.; Zheng, W.; Qiu, Q.; Jiang, K. New strawberry harvesting robot for elevated-trough culture. *Int. J. Agric. Biol. Eng.* **2012**, *5*, 1–8. [CrossRef]
9. Yamamoto, S.; Hayashi, S.; Yoshida, H.; Kobayashi, K. Development of a stationary robotic strawberry harvester with a picking mechanism that approaches the target fruit from below. *JARQ* **2014**, *48*, 261–269. [CrossRef]
10. Xiong, Y.; Peng, C.; Grimstad, L.; From, P.J.; Isler, V. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Comput. Electron. Agric.* **2019**, *157*, 392–402. [CrossRef]

11. Xiong, Y.; Ge, Y.; Grimstad, L.; From, P.J. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. *J. Field Robot.* **2020**, *37*, 202–224. [CrossRef]
12. Jia, W.; Zhang, Y.; Lian, J.; Zheng, Y.; Zhao, D.; Li, C. Apple harvesting robot under information technology: A review. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420925310. [CrossRef]
13. Li, Z.; Yuan, X.; Wang, C. A review on structural development and recognition–localization methods for end-effector of fruit–vegetable picking robots. *Int. J. Adv. Robot. Syst.* **2022**, *19*, 17298806221104906. [CrossRef]
14. Oliveira, L.F.P.; Moreira, A.P.; Silva, M.F. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics* **2021**, *10*, 52. [CrossRef]
15. Cui, T.; Dong, Y.; Zhang, R.; Wu, Y.; Zhang, Y.; Lu, X.; Tang, Z. From field to market: Evolution of strawberry-harvesting techniques and research progress in intelligent robotic systems. *Appl. Sci.* **2026**, *16*, 7522. [CrossRef]
16. Tao, Z.; Li, K.; Rao, Y.; Li, W.; Zhu, J. Strawberry maturity recognition based on improved YOLOv5. *Agronomy* **2024**, *14*, 460. [CrossRef]
17. Zhao, S.; Fang, C.; Hua, T.; Jiang, Y. Detecting the maturity of red strawberries using improved YOLOv8s model. *Agriculture* **2025**, *15*, 2263. [CrossRef]
18. Gao, Z.; Shao, Y.; Xuan, G.; Wang, Y.; Liu, Y.; Han, X. Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. *Artif. Intell. Agric.* **2020**, *4*, 31–38. [CrossRef]
19. Yue, Y.; Xu, S.; Wu, H. A strawberry ripeness detection method based on improved YOLOv8. *Appl. Sci.* **2025**, *15*, 6324. [CrossRef]
20. Tafuro, A.; Adewumi, A.; Parsa, S.; Amir, G.E.; Debnath, B. Strawberry picking point localization, ripeness and weight estimation. In Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 2295–2302. [CrossRef]
21. He, Z.; Karkee, M.; Zhang, Q. Enhanced machine vision system for field-based detection of pickable strawberries: Integrating an advanced two-step deep learning model merging improved YOLOv8 and YOLOv5-cls. *Comput. Electron. Agric.* **2025**, *234*, 110173. [CrossRef]
22. Yu, Y.; Zhang, K.; Liu, H.; Yang, L.; Zhang, D. Real-time visual localization of the picking points for a ridge-planting strawberry harvesting robot. *IEEE Access* **2020**, *8*, 116556–116568. [CrossRef]
23. Xie, H.; Zhang, Z.; Zhang, K.; Yang, L.; Zhang, D.; Yu, Y. Research on the visual location method for strawberry picking points under complex conditions based on composite models. *J. Sci. Food Agric.* **2024**, *104*, 8566–8579. [CrossRef] [PubMed]
24. Dai, S.; Bai, T.; Zhao, Y. Keypoint detection and 3D localization method for ridge-cultivated strawberry harvesting robots. *Agriculture* **2025**, *15*, 372. [CrossRef]
25. Hao, H.; Xi, J.; Dai, J.; Wang, G.; Liu, D.; Zhu, L. DHN-YOLO: A joint detection algorithm for strawberries at different maturity stages and key harvesting points. *Plants* **2025**, *14*, 3439. [CrossRef] [PubMed]
26. Liu, H.; Liang, Y.; He, Q.; Li, B.; Wang, W.; He, H.; Xu, Y.; Yao, Y.; Cao, X.; Yin, Y.; et al. StrawPose-Lite: A lightweight pose network for strawberry picking point prediction on edge devices. *Agriculture* **2026**, *16*, 1185. [CrossRef]
27. Le Louëdec, J.; Cielniak, G. Key point-based orientation estimation of strawberries for robotic fruit picking. In *Computer Vision Systems, ICVS 2023*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 14253, pp. 148–158. [CrossRef]
28. Zhu, T.; Miao, Z.; Zhao, C.; Chen, L.; Xiong, Y. Five-degree-of-freedom strawberry pose estimation using key points and oriented bounding box detection. *Comput. Electron. Agric.* **2025**, *239*, 110974. [CrossRef]
29. Qing, J.; Zeng, Z.; Chen, J.; Liang, X.; Mo, B.; Liu, Y.; Zhou, A.; Wu, S. Fruit-stem structural visual perception for picking point localization in strawberry harvesting robots. *Front. Plant Sci.* **2026**, *17*, 1894265. [CrossRef] [PubMed]
30. Ma, Z.; Dong, N.; Gu, J.; Cheng, H.; Meng, Z.; Du, X. STRAW-YOLO: A detection method for strawberry fruits targets and key points. *Comput. Electron. Agric.* **2025**, *230*, 109853. [CrossRef]
31. Chen, J.; Ma, A.; Huang, L.; Li, H.; Zhang, H.; Huang, Y.; Zhu, T. Efficient and lightweight grape and picking point synchronous detection model based on key point detection. *Comput. Electron. Agric.* **2024**, *217*, 108612. [CrossRef]
32. Wang, J.; Lin, X.; Luo, L.; Chen, M.; Wei, H.; Xu, L.; Luo, S. Cognition of grape cluster picking point based on visual knowledge distillation in complex vineyard environment. *Comput. Electron. Agric.* **2024**, *225*, 109216. [CrossRef]
33. Li, H.; Huang, J.; Gu, Z.; He, D.; Huang, J.; Wang, C. Positioning of mango picking point using an improved YOLOv8 architecture with object detection and instance segmentation. *Biosyst. Eng.* **2024**, *247*, 202–220. [CrossRef]
34. Zheng, C.; Chen, P.; Pang, J.; Yang, X.; Chen, C.; Tu, S.; Xue, Y. A mango picking vision algorithm on instance segmentation and key point detection from RGB images in an open orchard. *Biosyst. Eng.* **2021**, *206*, 32–54. [CrossRef]
35. Zhong, Z.; Xiong, J.; Zheng, Z.; Liu, B.; Liao, S.; Huo, Z.; Yang, Z. A method for litchi picking points calculation in natural environment based on main fruit bearing branch detection. *Comput. Electron. Agric.* **2021**, *189*, 106398. [CrossRef]
36. Li, Y.; Wang, W.; Guo, X.; Wang, X.; Liu, Y.; Wang, D. Recognition and positioning of strawberries based on improved YOLOv7 and RGB-D sensing. *Agriculture* **2024**, *14*, 624. [CrossRef]

37. Ge, Y.; Xiong, Y.; Tenorio, G.L.; From, P.J. Fruit localization and environment perception for strawberry harvesting robots. *IEEE Access* **2019**, *7*, 147642–147652. [CrossRef]
38. Le Louëdec, J.; Cielniak, G. 3D shape sensing and deep learning-based segmentation of strawberries. *Comput. Electron. Agric.* **2021**, *190*, 106374. [CrossRef]
39. Burusa, A.K.; van Henten, E.J.; Kootstra, G. Attention-driven next-best-view planning for efficient reconstruction of plants and targeted plant parts. *Biosyst. Eng.* **2024**, *246*, 248–262. [CrossRef]
40. Menon, R.; Zaenker, T.; Dengler, N.; Bennewitz, M. NBV-SC: Next best view planning based on shape completion for fruit mapping and reconstruction. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, Detroit, MI, USA, 1–5 October 2023; pp. 4197–4203. [CrossRef]
41. Yu, Y.; Xie, H.; Zhang, K.; Wang, Y.; Li, Y.; Zhou, J.; Xu, L. Design, development, integration, and field evaluation of a ridge-planting strawberry harvesting robot. *Agriculture* **2024**, *14*, 2126. [CrossRef]
42. Hornung, A.; Wurm, K.M.; Bennewitz, M.; Stachniss, C.; Burgard, W. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Auton. Robots* **2013**, *34*, 189–206. [CrossRef]
43. Wang, P.; Luo, W.; Liu, J.; Zhou, Y.; Li, X.; Zhao, S.; Zhang, G.; Zhao, Y. Real-time semantic SLAM-based 3D reconstruction robot for greenhouse vegetables. *Comput. Electron. Agric.* **2025**, *237*, 110582. [CrossRef]
44. He, Z.; Liu, Z.; Zhou, Z.; Karkee, M.; Zhang, Q. Improving picking efficiency under occlusion: Design, development, and field evaluation of an innovative robotic strawberry harvester. *Comput. Electron. Agric.* **2025**, *237*, 110684. [CrossRef]
45. Sun, T.; Zhang, W.; Gao, X.; Zhang, W.; Li, N.; Miao, Z. Efficient occlusion avoidance based on active deep sensing for harvesting robots. *Comput. Electron. Agric.* **2024**, *225*, 109360. [CrossRef]
46. Liu, B.; Chen, C.; Li, J.; Zhang, Q.; Meng, Y. Research on three-dimensional positioning method for automatic strawberry fruit picking based on Vision–IMU fusion. *Agriculture* **2026**, *16*, 893. [CrossRef]
47. Tufail, M.; Iqbal, J.; Ahmad, R. Development of a general-purpose AI-powered robotic platform for strawberry harvesting. *Agriculture* **2026**, *16*, 769. [CrossRef]
48. Rabby, M.K.M.; Chowdhury, B.; Kim, J.H. A modified Canny edge detection algorithm for fruit detection and classification. In Proceedings of the 2018 10th International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 20–22 December 2018; pp. 237–240. [CrossRef]
49. Tan, S.H.; Lam, C.K.; Kamarudin, K.; Ismail, A.H.; Rahim, N.A.; Azmi, M.S.M.; Yahya, W.M.N.W.; Sneah, G.K.; Seng, M.L.; Hai, T.P.; et al. Vision-based edge detection system for fruit recognition. *J. Phys. Conf. Ser.* **2021**, *2107*, 012066. [CrossRef]
50. Wei, X.; Jia, K.; Lan, J.; Li, Y.; Zeng, Y.; Wang, C. Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot. *Optik* **2014**, *125*, 5684–5689. [CrossRef]
51. Zhang, Y.; Wu, L. Classification of fruits using computer vision and a multiclass support vector machine. *Sensors* **2012**, *12*, 12489–12505. [CrossRef] [PubMed]
52. Guo, Q.; Chen, Y.; Tang, Y.; Zhuang, J.; He, Y.; Hou, C.; Chu, X.; Zhong, Z.; Luo, S. Lychee fruit detection based on monocular machine vision in orchard environment. *Sensors* **2019**, *19*, 4091. [CrossRef] [PubMed]
53. Lin, G.; Tang, Y.; Zou, X.; Cheng, J.; Xiong, J. Fruit detection in natural environment using partial shape matching and probabilistic Hough transform. *Precis. Agric.* **2019**, *21*, 160–177. [CrossRef]
54. Luo, L.; Tang, Y.; Zou, X.; Ye, M.; Feng, W.; Li, G. Vision-based extraction of spatial information in grape clusters for harvesting robots. *Biosyst. Eng.* **2016**, *151*, 90–104. [CrossRef]
55. Lv, J.; Wang, F.; Ma, Z. Peach fruit recognition method under natural environment. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*; SPIE: Bellingham, WA, USA, 2016; Volume 10033, pp. 232–236. [CrossRef]
56. Lv, J.; Wang, F.; Ma, Z.; Rong, H. Yellow apple recognition method under natural environment. In Proceedings of the 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2015; Volume 1, pp. 46–49. [CrossRef]
57. Wang, D.; Song, H.; Tie, Z.; Zhang, W.; He, D. Recognition and localization of occluded apples using K-means clustering algorithm and convex hull theory: A comparison. *Multimed. Tools Appl.* **2015**, *75*, 3177–3198. [CrossRef]
58. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
59. Pérez-Borrero, I.; Marín-Santos, D.; Gegúndez-Arias, M.E.; Cortés-Ancos, E. A fast and accurate deep learning method for strawberry instance segmentation. *Comput. Electron. Agric.* **2020**, *178*, 105736. [CrossRef]
60. Cao, L.; Chen, Y.; Jin, Q. Lightweight strawberry instance segmentation on low-power devices for picking robots. *Electronics* **2023**, *12*, 3145. [CrossRef]
61. Guo, Z.; Hu, X.; Zhao, B.; Wang, H.; Ma, X. StrawSnake: A real-time strawberry instance segmentation network based on the contour learning approach. *Electronics* **2024**, *13*, 3103. [CrossRef]
62. Li, K.; Wei, X.; Wang, Q.; Zhang, W. Research on strawberry visual recognition and 3D localization based on lightweight RAFS-YOLO and RGB-D camera. *Agriculture* **2025**, *15*, 2212. [CrossRef]

63. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
64. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
65. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
66. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
67. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
68. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*. [CrossRef]
69. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; Volume 15089, pp. 1–21. [CrossRef]
70. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
71. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
72. Chen, H.; Wu, K.; Lin, H.; Zhou, H.; Zhou, Z.; Mai, Y.; Shi, L.; Zhang, M.; Ma, Z.; Lin, P.; et al. A real-time vision guidance method for autonomous longan picking by the UAV. *Comput. Electron. Agric.* **2025**, *229*, 109735. [CrossRef]
73. Guo, J.; Xiao, X.; Miao, J.; Tian, B.; Zhao, J.; Lan, Y. Design and experiment of a visual detection system for Zanthoxylum-harvesting robot based on improved YOLOv5 model. *Agriculture* **2023**, *13*, 821. [CrossRef]
74. Huang, Y.; Zhong, Y.; Zhong, D.; Yang, C.; Wei, L.; Zou, Z.; Chen, R. Pepper-YOLO: A lightweight model for green pepper detection and picking point localization in complex environments. *Front. Plant Sci.* **2024**, *15*, 1508258. [CrossRef] [PubMed]
75. Lawal, M.O. Tomato detection based on modified YOLOv3 framework. *Sci. Rep.* **2021**, *11*, 1447. [CrossRef] [PubMed]
76. Lawal, O.M. Development of tomato detection model for robotic platform using deep learning. *Multimed. Tools Appl.* **2021**, *80*, 26751–26772. [CrossRef]
77. Liu, G.; Nouaze, J.C.; Touko Mbouembe, P.L.; Kim, J.H. YOLO-Tomato: A robust algorithm for tomato detection based on YOLOv3. *Sensors* **2020**, *20*, 2145. [CrossRef] [PubMed]
78. López-Barrios, J.D.; Escobedo Cabello, J.A.; Gómez-Espinosa, A.; Montoya-Cavero, L.-E. Green sweet pepper fruit and peduncle detection using Mask R-CNN in greenhouses. *Appl. Sci.* **2023**, *13*, 6296. [CrossRef]
79. Parvathi, S.; Tamil Selvi, S. Detection of maturity stages of coconuts in complex background using Faster R-CNN model. *Biosyst. Eng.* **2021**, *202*, 119–132. [CrossRef]
80. Wang, Y.; Ouyang, C.; Peng, H.; Deng, J.; Yang, L.; Chen, H.; Luo, Y.; Jiang, P. YOLO-ALW: An enhanced high-precision model for chili maturity detection. *Sensors* **2025**, *25*, 1405. [CrossRef] [PubMed]
81. Wang, F.; Sun, Z.; Chen, Y.; Zheng, H.; Jiang, J. Xiaomila green pepper target detection method under complex environment based on improved YOLOv5s. *Agronomy* **2022**, *12*, 1477. [CrossRef]
82. Sozzi, M.; Cantalamessa, S.; Cogato, A.; Kayad, A.; Marinello, F. Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms. *Agronomy* **2022**, *12*, 319. [CrossRef]
83. Wan, S.; Goudos, S. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* **2020**, *168*, 107036. [CrossRef]
84. Wang, Z.; Jin, L.; Wang, S.; Xu, H. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.* **2022**, *185*, 111808. [CrossRef]
85. Guo, C.; Zhu, C.; Liu, Y.; Huang, R.; Cao, B.; Zhu, Q.; Zhang, R.; Zhang, B. End-to-end lightweight transformer-based neural network for grasp detection towards fruit robotic handling. *Comput. Electron. Agric.* **2024**, *221*, 109014. [CrossRef]
86. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12009–12019. [CrossRef]
87. Mia, M.S.; Voban, A.A.B.; Arnob, A.B.H.; Naim, A.; Ahmed, M.K.; Islam, M.S. DANet: Enhancing small object detection through an efficient deformable attention network. In Proceedings of the 2023 International Conference on Cognitive Computing and Complex Data, Dhaka, Bangladesh, 21–23 December 2023; pp. 51–62. [CrossRef]
88. Ge, Y.; Xiong, Y.; From, P.J. Three-dimensional location methods for the vision system of strawberry-harvesting robots: Development and comparison. *Precis. Agric.* **2023**, *24*, 764–782. [CrossRef]

89. Yoshida, T.; Kawahara, T.; Fukao, T. Fruit recognition method for a harvesting robot with RGB-D cameras. *ROBOMECH J.* **2022**, *9*, 15. [CrossRef]
90. Hou, C.; Zhang, X.; Tang, Y.; Zhuang, J.; Tan, Z.; Huang, H.; Chen, W.; Wei, S.; He, Y.; Luo, S. Detection and localization of citrus fruit based on improved YOLOv5s and binocular vision in the orchard. *Front. Plant Sci.* **2022**, *13*, 972445. [CrossRef] [PubMed]
91. Jun, J.; Kim, J.; Seol, J.; Kim, J.; Son, H.I. Towards an efficient tomato harvesting robot: 3D perception, manipulation, and end-effector. *IEEE Access* **2021**, *9*, 17631–17640. [CrossRef]
92. Ling, X.; Zhao, Y.; Gong, L.; Liu, C.; Wang, T. Dual-arm cooperation and implementing for robotic harvesting tomato using binocular vision. *Robot. Auton. Syst.* **2019**, *114*, 134–143. [CrossRef]
93. Wu, Y.; Qiu, C.; Liu, S.; Zou, X.; Li, X. Tomato harvesting robot system based on binocular vision. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems, Beijing, China, 15–17 October 2021; pp. 757–761. [CrossRef]
94. 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. *Comput. Electron. Agric.* **2021**, *187*, 106237. [CrossRef]
95. Rapado-Rincón, D.; van Henten, E.J.; Kootstra, G. Development and evaluation of automated localisation and reconstruction of all fruits on tomato plants in a greenhouse based on multi-view perception and 3D multi-object tracking. *Biosyst. Eng.* **2023**, *231*, 78–91. [CrossRef]
96. Zhou, T.; Fan, D.-P.; Cheng, M.-M.; Shen, J.; Shao, L. RGB-D salient object detection: A survey. *Comput. Vis. Media* **2021**, *7*, 37–69. [CrossRef] [PubMed]
97. Cheng, Z.; Zhang, Y.; Tang, C. Swin-Depth: Using transformers and multi-scale fusion for monocular-based depth estimation. *IEEE Sens. J.* **2021**, *21*, 26912–26920. [CrossRef]
98. Khan, F.; Salahuddin, S.; Javidnia, H. Deep learning-based monocular depth estimation methods—A state-of-the-art review. *Sensors* **2020**, *20*, 2272. [CrossRef] [PubMed]
99. Yang, N.; von Stumberg, L.; Wang, R.; Cremers, D. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1278–1289. [CrossRef]
100. Liu, X.; Jing, X.; Jiang, H.; Younas, S.; Wei, R.; Dang, H.; Wu, Z.; Fu, L. Performance evaluation of newly released cameras for fruit detection and localization in complex kiwifruit orchard environments. *J. Field Robot.* **2024**, *41*, 881–894. [CrossRef]
101. Lu, A.; Lv, X.; Ma, L.; Li, Q.; Ma, Q. An effective picking point localization method for multi-posture lotus pods based on three-view depth vision observation. *Comput. Electron. Agric.* **2024**, *227*, 109492. [CrossRef]
102. Zaenker, T.; Smitt, C.; McCool, C.; Bennewitz, M. Viewpoint planning for fruit size and position estimation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 27 September–1 October 2021; pp. 3271–3277. [CrossRef]
103. Mehta, S.; Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
104. Xiong, Y.; From, P.J.; Isler, V. Design and evaluation of a novel cable-driven gripper with perception capabilities for strawberry picking robots. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 7384–7391. [CrossRef]
105. Chen, W.; Rao, Y.; Wang, F.; Zhang, Y.; Wang, T.; Jin, X.; Hou, W.; Jiang, Z.; Zhang, W. MLP-based multimodal tomato detection in complex scenarios: Insights from task-specific analysis of feature fusion architectures. *Comput. Electron. Agric.* **2024**, *221*, 108951. [CrossRef]
106. Junge, K.; Pires, C.; Hughes, J. Lab2field transfer of a robotic raspberry harvester enabled by a soft sensorized physical twin. *Commun. Eng.* **2023**, *2*, 40. [CrossRef]
107. Gallego, G.; Delbruck, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.J.; Conradt, J.; Daniilidis, K.; et al. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 154–180. [CrossRef] [PubMed]
108. Gehrig, M.; Scaramuzza, D. Recurrent vision transformers for object detection with event cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13884–13893. [CrossRef]
109. Rebecq, H.; Gallego, G.; Mueggler, E.; Scaramuzza, D. EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time. *Int. J. Comput. Vis.* **2017**, *126*, 1394–1414. [CrossRef]
110. Liu, X.; Song, Z.; Tan, Y.; Yang, S.; Ma, Y. Conflict avoidance task planning strategy for dual-arm cooperative strawberry harvesting robots. *Comput. Electron. Agric.* **2026**, *241*, 111067. [CrossRef]
111. Zhu, Y.; Lu, W.; Ren, G.; Gao, H.; Vougioukas, S.; Ying, Y.; Peng, C. Optimal scheduling of a dual-arm robot for efficient strawberry harvesting in plant factories. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19–25 October 2025; pp. 1231–1238. [CrossRef]
112. Li, C.; Ma, W.; Liu, F.; Fang, B.; Lu, H.; Sun, Y. Recognition of citrus fruit and planning the robotic picking sequence in orchards. *Signal Image Video Process.* **2023**, *17*, 4425–4434. [CrossRef]

113. Li, T.; Xie, F.; Qiu, Q.; Feng, Q. Multi-arm robot task planning for fruit harvesting using multi-agent reinforcement learning. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, Detroit, MI, USA, 1–5 October 2023; pp. 4176–4183. [CrossRef]
114. Li, T.; Xie, F.; Zhao, Z.; Zhao, H.; Guo, X.; Feng, Q. A multi-arm robot system for efficient apple harvesting: Perception, task plan and control. *Comput. Electron. Agric.* **2023**, *211*, 107979. [CrossRef]
115. Ning, Z.; Luo, L.; Ding, X.; Dong, Z.; Yang, B.; Cai, J.; Chen, W.; Lu, Q. Recognition of sweet peppers and planning the robotic picking sequence in high-density orchards. *Comput. Electron. Agric.* **2022**, *196*, 106878. [CrossRef]
116. Chang, C.-L.; Huang, C.-C. Design and implementation of an AI-based robotic arm for strawberry harvesting. *Agriculture* **2024**, *14*, 2057. [CrossRef]
117. Parsa, S.; Debnath, B.; Khan, M.A.; E., A.G. Modular autonomous strawberry picking robotic system. *J. Field Robot.* **2024**, *41*, 2226–2246. [CrossRef]
118. Zhang, Y.; Xie, H.; Zhang, D.; Yang, L.; Lu, Z.; Cui, T.; He, X.; Zhang, K. Design and experimental study of ridge-grown strawberry automatic harvesting robot. *Comput. Electron. Agric.* **2026**, *240*, 111112. [CrossRef]
119. Zhang, K.; Lammers, K.; Chu, P.; Dickinson, N.; Li, Z.; Lu, R. Algorithm design and integration for a robotic apple harvesting system. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, Kyoto, Japan, 23–27 October 2022; pp. 9217–9224. [CrossRef]
120. Rehman, H.U.; Miura, J. Viewpoint planning for automated fruit harvesting using deep learning. In Proceedings of the 2021 IEEE/SICE International Symposium on System Integration, Fukushima, Japan, 11–14 January 2021; pp. 409–414. [CrossRef]
121. Yi, T.; Zhang, D.; Luo, L.; Luo, J. View planning for grape harvesting based on active vision strategy under occlusion. *IEEE Robot. Autom. Lett.* **2024**, *9*, 2535–2542. [CrossRef]
122. Zaenker, T.; Rückin, J.; Menon, R.; Popović, M.; Bennewitz, M. Graph-based view motion planning for fruit detection. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, Detroit, MI, USA, 1–5 October 2023; pp. 4219–4225. [CrossRef]
123. Wang, L.; Wang, Z.; Liu, M.; Ying, Z.; Xu, N.; Meng, Q. Full coverage path planning methods of harvesting robot with multi-objective constraints. *J. Intell. Robot. Syst.* **2022**, *106*, 17. [CrossRef]
124. Wang, Y.; Wu, H.; Zhu, Z.; Ye, Y.; Qian, M. Continuous picking of yellow peaches with recognition and collision-free path. *Comput. Electron. Agric.* **2023**, *214*, 108273. [CrossRef]
125. Sun, M.; Zhao, C.; Yang, L.; Liu, H.; Hu, S.; Xiong, Y. Vision-based early fault diagnosis and self-recovery for strawberry harvesting robots. *Artif. Intell. Agric.* **2026**, *in press*. [CrossRef]
126. Park, Y.; Seol, J.; Pak, J.; Jo, Y.; Kim, C.; Son, H.I. Human-centered approach for an efficient cucumber harvesting robot system: Harvest ordering, visual servoing, and end-effector. *Comput. Electron. Agric.* **2023**, *212*, 108116. [CrossRef]
127. Arad, B.; Balendonck, J.; Barth, R.; Ben-Shahar, O.; Edan, Y.; Hellström, T.; Hemming, J.; Kurtser, P.; Ringdahl, O.; Tielen, T.; et al. Development of a sweet pepper harvesting robot. *J. Field Robot.* **2020**, *37*, 1027–1039. [CrossRef]
128. Baeten, J.; Donné, K.; Boedrij, S.; Beckers, W.; Claesen, E. Autonomous fruit picking machine: A robotic apple harvester. In *Field and Service Robotics*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 531–539. [CrossRef]
129. Birrell, S.; Hughes, J.; Cai, J.Y.; Iida, F. A field-tested robotic harvesting system for iceberg lettuce. *J. Field Robot.* **2020**, *37*, 225–245. [CrossRef] [PubMed]

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.