UC Merced

UC Merced Previously Published Works

Title

Machine vision for fruit detection and localization in intelligent harvesting robots: A comprehensive review

Permalink

https://escholarship.org/uc/item/8b06g6jx

Journal

Smart Agricultural Technology, 15

ISSN

2772-3755

Authors

Yu, Liujing
Pu, Yingjun
Lin, Ju
et al.

Publication Date

2026-12-01

DOI

10.1016/j.atech.2026.102673

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Smart Agricultural Technology 15 (2026) 102673

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Machine vision for fruit detection and localization in intelligent harvesting robots: A comprehensive review

Liujing Yu [a], Yingjun Pu [a,*], Ju Lin [a], Mustafa Mhamed [b], Reza Ehsani [c], Cheng Shen [d], Fuzeng Yang [e]

[a] College of Engineering and Technology, Southwest University, Chongqing Key Laboratory of Intelligent Agricultural Machinery Equipment for Hilly and Mountainous Regions, Chongqing, 400715, China
[b] College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China
[c] School of Engineering, University of California, Merced, CA, 95343, United States
[d] Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing, 210014, China
[e] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China

ARTICLE INFO

Keywords:
Fruit-harvesting robots
Machine vision
Fruit detection
Pose estimation
Picking-point localization
Hand-eye coordination

ABSTRACT

Autonomous fruit harvesting requires not only accurate fruit detection but also the reliable transformation of uncertain visual observations into safe, efficient, and low-damage robotic harvesting actions. Achieving this objective remains challenging in commercial orchards because variable illumination, severe fruit occlusion, plant motion, and constrained workspaces introduce substantial uncertainty into both visual perception and robotic execution. In this review, we systematically synthesize recent advances in machine vision for fruit detection, segmentation, three-dimensional localization, pose estimation, and picking-point localization, and examine their integration with end-effector operation and hand–eye coordination in intelligent harvesting robots. We compare eye-in-hand, eye-to-hand, and hybrid vision configurations; two-dimensional and three-dimensional sensing modalities; machine-learning-, deep-learning-, and Transformer-based perception methods; and representative harvesting-system configurations. This review establishes an integrated perspective on the interactions among sensing, perception, motion planning, grasping, detachment, and feedback within a perception–action–interaction cycle. It highlights the transition from object-oriented detection to operation-oriented perception, which involves not only understanding the fruit position but also the attachment structure, approach direction, obstacles, interaction risk, and harvestability. There are several challenges to deployment in complex field conditions, including calibration drift, incomplete spatial information, limited end-effector adaptability, and accumulated perception-execution errors. Future harvesting systems will need to integrate multimodal sensing, active view planning, uncertainty-aware decision-making, visual servoing, and tactile-force feedback to enable autonomous fruit harvesting to be scalable and damage-aware.

1. Introduction

Fruit harvesting is a labor-intensive and time-sensitive operation in horticultural production, and its efficiency directly affects production costs and the economic sustainability of the fruit industry [1–3]. In recent years, rising labor costs and instability in the availability of seasonal workers have further increased the demand for automated harvesting technologies [4,5]. Compared with structured industrial environments, however, orchards and protected cultivation systems are characterized by variable illumination, dense foliage, irregular fruit distribution, occlusion, and restricted operating spaces, making autonomous harvesting particularly challenging [6]. To reduce dependence on manual labor, extensive research has therefore been devoted to harvesting robots capable of autonomously perceiving crop environments, locating target fruits, planning manipulator motions, and performing physical harvesting operations [7]. With advances in sensing, artificial intelligence, robotic manipulation, and intelligent control, fruit harvesting robots have gradually evolved from systems focusing primarily on mechanized picking toward integrated robotic systems incorporating environmental perception, robotic manipulation, end-effector operation, and autonomous control [8]. Nevertheless, the complexity and uncertainty of natural agricultural environments

* Corresponding author.
E-mail address: puyingjun@swu.edu.cn (Y. Pu).

https://doi.org/10.1016/j.atech.2026.102673
Received 15 September 2026; Received in revised form 21 October 2026; Accepted 25 October 2026
Available online 26 October 2026
2772-3755/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

continue to constrain harvesting success, efficiency, and robustness, preventing many robotic systems from achieving reliable large-scale deployment.

Environmental perception is a prerequisite for autonomous robotic harvesting because robots must not only distinguish harvestable fruits from complex backgrounds but also obtain spatial information required for subsequent manipulation. Early machine-vision methods mainly relied on manually designed color, shape, texture, and geometric features, whereas recent advances in deep learning have substantially improved fruit detection and segmentation under complex agricultural environments. In particular, deep-learning-based detectors have demonstrated improved recognition performance for fruits subjected to different levels of occlusion, indicating their potential for perception in cluttered crop environments [9]. Meanwhile, perception requirements have gradually expanded from two-dimensional fruit detection toward three-dimensional localization and the extraction of operation-oriented information, including depth, fruit attachment structures, picking points, and target pose. Stereo vision, red-green-blue-depth (RGB-D) sensing, and multi-view vision have therefore been increasingly introduced to obtain more complete spatial information. For example, multi-view perception can compensate for the limited field of view of individual cameras and enhance the completeness and stability of three-dimensional target reconstruction in complex orchard environments [10]. Nevertheless, illumination variation, fruit clustering, leaf and branch occlusion, similarity between targets and backgrounds, and incomplete observations continue to degrade perception reliability. More importantly, accurate fruit detection alone does not necessarily provide sufficient information for robotic manipulation. Harvesting robots increasingly require perception systems to determine not only where a fruit is located, but also its spatial relationship with surrounding plant structures and the appropriate harvesting position and approach direction. Therefore, robust and operation-oriented perception in unstructured environments remains a fundamental challenge for autonomous harvesting robots.

The end-effector directly influences harvesting success and fruit damage and needs to be able to account for fruit size, shape, firmness, and fruit attachment characteristics. Current designs combine gripping, suction, cutting, twisting, or pulling and more and more rely on compliant, adaptive, and biomimetic mechanisms for reliable, low-damage harvesting. Studies of manual harvesting have shown that appropriate picking motions can reduce fruit detachment forces and provide a biomechanical basis for robotic harvesting strategies; for example, bending combined with pulling can reduce the force required to detach apples compared with direct pulling [11]. Human grasping studies further indicate that grasp type varies with fruit size and hand characteristics, providing useful guidance for the dimensional design and grasp planning of multi-finger robotic end-effectors [12]. Accordingly, compliant constant-force mechanisms have been developed to maintain relatively stable grasping forces and reduce picking-induced bruising [13], while different fruit detachment patterns, such as pressing–breaking, pulling, pulling–rotating, and twisting, have been systematically evaluated to guide end-effector design [14]. Nevertheless, end-effectors must still accommodate fruit variability, confined approach directions, surrounding obstacles, uncertain fruit poses, and differences in fruit–stem attachment while maintaining sufficient harvesting efficiency. Therefore, achieving adaptive, low-damage, and efficient physical interaction in cluttered crop environments remains a major barrier to practical robotic harvesting.

Successful harvesting requires tight hand–eye coordination that connects perception to manipulation through eye-to-hand, eye-in-hand or hybrid vision configurations. Hybrid systems combine global localization with local refinement, improving target and cutting point accuracy [15,16]. But sequential perception–planning–execution pipelines are vulnerable to target motion, plant deformation and accumulated localization and control errors. Therefore, harvesting robots are moving towards visual servoing, active perception, multimodal feedback and closed-loop perception–action coordination for robust operation in dynamic orchard environments.

To improve the transparency and reproducibility of this review, a structured literature search and screening procedure was adopted. Relevant studies were primarily retrieved from the Web of Science Core Collection, Scopus, ScienceDirect, IEEE Xplore, and Google Scholar, with the main search period covering publications from 2015 to 2026. Earlier studies were additionally considered when they represented seminal developments in fruit harvesting robots, machine vision, end-effector design, or hand–eye coordination. Search terms included combinations of “fruit harvesting robot”, “robotic harvesting”, “fruit detection”, “fruit segmentation”, “pose estimation”, “picking point localization”, “end-effector”, “visual servoing”, and “hand–eye coordination”. After removing duplicate and weakly relevant records, studies were selected according to their relevance to robotic fruit harvesting, methodological representativeness, experimental validation, and contribution to the development of perception, manipulation, or perception–action coordination technologies. Particular emphasis was placed on studies reporting quantitative experimental results or proposing representative system architectures and technical methods. A total of 233 publications were finally included in this review.

This study provides a comprehensive review of research progress in the fields of fruit detection, end-effector design, and hand-eye coordination control for harvesting robots, offering a reliable theoretical and technical reference for the future iterative development and technical research of harvesting robots. The structure of this paper is as follows: Section 2 reviews representative fruit-harvesting robots, system configurations, and hand-eye coordination. Sections 3 and 4 examine visual sensing technologies and machine vision methods for fruit detection, segmentation, pose estimation, and picking-point localization. Section 5 discusses vision-guided harvesting execution and end-effector configurations. Section 6 presents the challenges and future prospects, and Section 7 concludes the review.

2. Representative fruit harvesting robots and system configurations

Smart technology has rapidly advanced over time leading to mechanical harvesting of fruits and vegetables from a simple, mechanically assisted harvesting technique to the intelligent harvesting robot system which is equipped with environmental sensing, target identification and localization, operational management and autonomous navigation. Unlike industrial robots operating in stable, orderly environments, orchards carry out tasks that are challenging and chaotic. There are different growth characteristics for different planting sites and types of fruit in orchard ecosystems, for example fruit size, cultivation techniques, operational elevation, and fruit drop. Due to the complex nature of orchard ecosystems and different types of crops, no harvesting robot that can harvest various types of crops has yet to be developed. Most harvesting robots are customized to a specific task, and only for one kind of fruit or a few species of fruits. As shown in Fig. 1, researchers both domestically and internationally have developed multiple specialized harvesting robots tailored to various fruit and vegetable crops.

In the present framework for enhancing harvesting robots, beyond modifying navigation and mobility systems for various orchard layouts, the majority of technical advancements concentrate on two primary components: the visual perception system and the end-effector. Simultaneously, investigators develop hand-eye system arrangements customized to the morphological traits and harvesting attributes of various fruits, thereby augmenting the robot's adaptation to certain crops and boosting operational accuracy. Hu et al. [17] developed a compact mobile harvesting robot for dwarf, high-density apple orchards by integrating stereo vision, a robotic manipulator with four degrees of freedom (DOF), a vacuum-based harvesting mechanism, and an Ackermann-steering mobile platform, enabling target localization, manipulator motion, and fruit detachment to be performed within a



Fig. 1. Representative robotic systems for fruit harvesting: (A) apple; (B) litchi; (C) apple; (D) pear; (E) tomato; (F) apple; (G) strawberry; (H) grape; (I) strawberry; (J) grape; (K) banana; and (L) pumpkin.

unified robotic system (Fig. 1A). Building on this basic configuration, robotic structures began to adapt more closely to the natural growth characteristics of fruits and the spatial constraints of harvesting environments. Yao et al. [18] developed a harvesting system for cluster-growing lychees by integrating visual localization, a robotic manipulator, and an end-effector with combing and cutting functions. This design requires the robot not only to reach the target fruit but also to provide sufficient operational space for fruit access and pedicel cutting (Fig. 1B). Wang et al. [19] developed an RGB-D vision-guided bionic harvesting system for apples, achieving fruit localization and detachment through the integration of machine vision, a robotic manipulator, and a suction-based harvesting mechanism (Fig. 1C). Li and Liu [20], considering the fragility and size variability of pears, incorporated an adaptive human-hand-inspired mechanism together with force and posture sensing into the harvesting end-effector (Fig. 1D). Although these systems target different fruits and employ distinct harvesting methods, they have established a fundamental robotic architecture in which vision provides target information, the robotic manipulator performs spatial approach and positioning, and the end-effector executes fruit manipulation and detachment.

The integration of individual harvesting mechanisms is giving way to coordinated operation among vision systems, robotic manipulators, end-effectors, and mobile platforms as harvesting robots operate in dense plant growth, foliage occlusion, and continuous harvesting. Rong et al. [21] addressed densely growing tomato plants by integrating RGB-D vision, a robotic manipulator, a mobile platform, and a flexible auxiliary gripping mechanism. During operation, the system integrated information on both the target fruit and neighboring fruits, enabling visual perception to contribute to the determination of the robot's operational strategy (Fig. 1E). Zhou et al. [22] designed a robotic harvesting system for unstructured apple orchards that included an RGB-D camera, collaborative robotic manipulator, end-effector, fruit collection module, and mobile chassis. This system displays how robotic operation has developed from "reaching a target position" to spatial manipulation that compensates for surrounding branches and foliage by using three-dimensional environmental information to guide the manipulator. The strawberry harvesting robot developed by Xiong et al. [15] further integrates RGB-D vision, a 5-DOF robotic manipulator, a harvesting end-effector with integrated sensing capabilities, and a modular Thorvald II mobile platform, enabling continuous execution of tasks such as target searching, localization, approach, and harvesting in elevated cultivation environments (Fig. 1G). For trellis-grown grapes, Xu et al. [23] constructed a dedicated harvesting platform comprising a fixed vision system, a selective compliance assembly robot arm, a self-propelled mobile chassis, a continuous cutting mechanism, and a fruit-receiving device, enabling visual information to be used not only for target localization but also for organizing the harvesting sequence of multiple targets and supporting continuous harvesting operations (Fig. 1H). The operational scenarios become increasingly complex, the interactions among the functional modules of harvesting robots become progressively stronger, while the role of visual perception expands from providing positional information for individual targets to serving as a key source of information for continuous robotic operation and task organization.

In recent years, fruit-harvesting robots have increasingly exhibited trends toward multi-actuator coordination, active perception, and highly integrated system configurations tailored to complex operational requirements. Yu et al. [3] developed a dual-arm collaborative harvesting system for ridge-grown strawberries, integrating visual perception, dual-manipulator operation, and mobile platform control within a unified system framework, thereby enabling the robot to cover different operational areas through coordinated multi-actuator operation (Fig. 1I). In contrast to traditional fixed-view observation techniques [24], the system tackled the obstruction of grape stems by vegetation by integrating the camera observation process with manipulator movement. When the primary perspective does not yield adequate target information, the system reestablishes target data by modifying the robot's observation location, thus transitioning visual perception from static observation at a fixed angle to dynamic perception through robotic movement (Fig. 1J). For large-scale crops such as bananas, Chen et al. [25] further integrated visual localization, manipulator operation, gripping and cutting mechanisms, and a mobile system into an autonomous harvesting platform. The system must simultaneously coordinate target localization, the manipulator workspace, and the fruit-detachment process (Fig. 1K). Yang et al. [26] integrated RGB-D vision, a UR10 collaborative robotic manipulator, a specialized robotic gripper, and a mobile work platform for ground-grown pumpkins, enabling the robot to continuously perform operations such as target

detection, spatial localization, grasping, and transportation (Fig. 1L). These systems demonstrate that harvesting robots are evolving from serial architectures composed of relatively independent functional modules toward integrated robotic systems in which perception, manipulation, and platform motion are closely interconnected.

A review of the representative harvesting robot studies shown in Fig. 1 reveals that, although the harvesting success rate of existing harvesting robots has gradually improved, they still struggle to achieve large-scale, long-term autonomous operation in real orchard environments, and their overall harvesting performance has not yet surpassed that of manual harvesting. The core bottleneck contributing to this situation lies in fruit recognition and pose estimation within orchard environments; simultaneously, the end-effector's ability to harvest fruit without damage and the ability to perform harvesting operations when fruit is occluded are also key challenges limiting the practical deployment of harvesting robots. Consequently, most current research in this field focuses on visual perception systems and actuators.

2.1. *Vision–manipulator configurations and hand–eye coordination*

Hand-eye coordination technology is a core component of fruit-harvesting robots and directly determines their picking accuracy. Fruit-harvesting robots integrate visual sensing, target perception, spatial localization, hand-eye coordination, and end-effector operation into a complete perception-to-action system. As illustrated in Fig. 2, the harvesting process generally begins with environmental information acquisition using two-dimensional or three-dimensional visual sensors, followed by fruit detection and the extraction of visual features required for target identification. The detected fruit is subsequently localized and its picking-related pose is estimated from red-green-blue (RGB) images, depth information, or three-dimensional point-cloud data. Based on these perception results, hand–eye coordination establishes the spatial relationship between the visual sensor and the manipulator, transforms target information from the camera coordinate system into robot-operational coordinates, and guides the manipulator toward the desired harvesting pose. The end-effector then performs the final grasping, cutting, detachment, or enclosing operation according to the physical characteristics of the target fruit. Within this functional chain, the spatial arrangement between the camera and the manipulator is particularly important because it affects the field of view, target observability, coordinate transformation, and visual feedback available during harvesting. Accordingly, fruit-harvesting robots can be further classified into eye-in-hand, eye-to-hand, and hybrid hand–eye configurations, which are discussed in the following sections.

2.1.1. *Eye-in-hand fruit harvesting robots*

In an eye-in-hand system, the camera is affixed to the manipulator or positioned close to the end-effector, moving in conjunction with the robotic arm. This configuration facilitates near-field scrutiny of the objective and enables the robot to alter its perspective while approaching the target, making it very effective for localized enhancement and obstruction management. In dense crops like bell peppers, adjusting the perspective dynamically may enhance the visibility of targets when fruits are obscured by foliage or branches [27]. However, because the camera moves with the manipulator, its instantaneous field of view is relatively limited, and a single observation may provide insufficient global information about the harvesting workspace. To compensate for this limitation, simultaneous localization and mapping and multi-view reconstruction have been introduced to integrate sequential local observations into a more complete representation of the environment [28].

For dense citrus harvesting environments, Choi et al. [29] developed a robotic system driven by artificial intelligence (AI) equipped with a depth camera mounted near the end-effector. The system estimated the

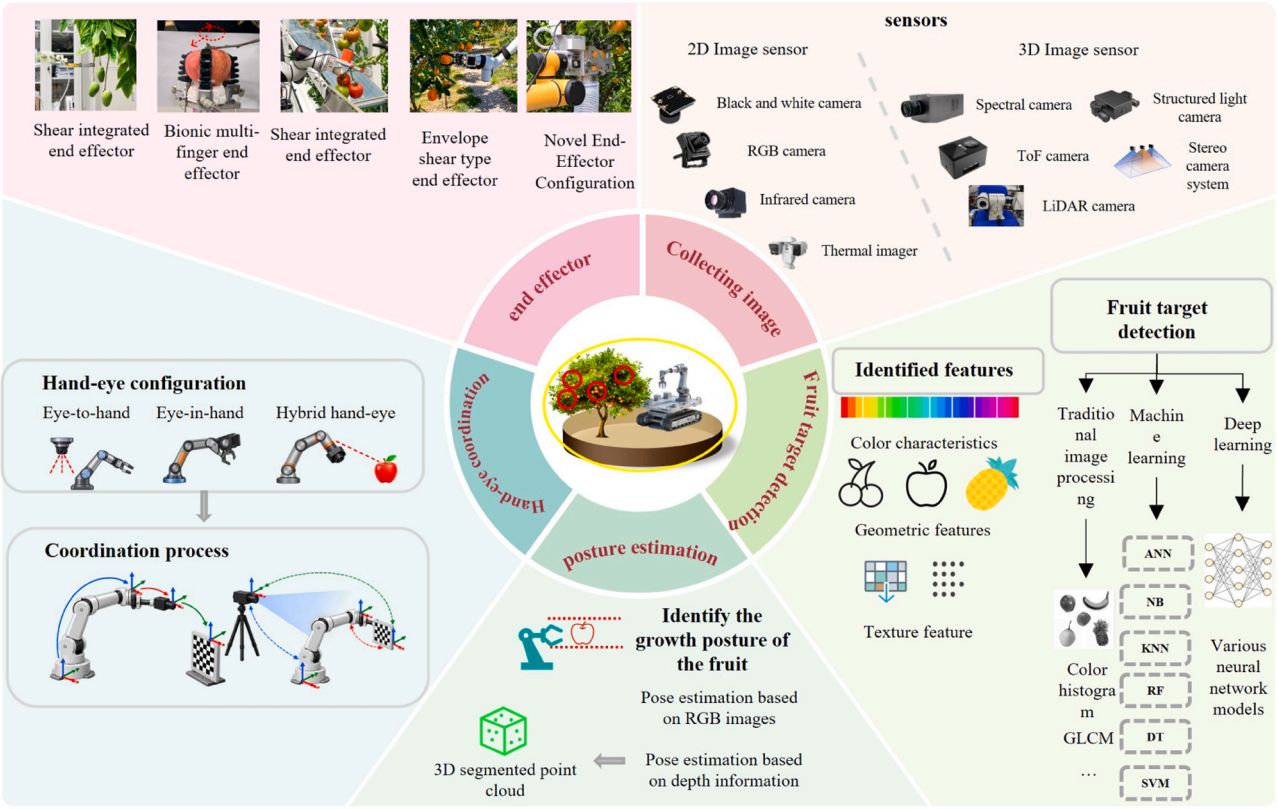


Fig. 2. Overall Framework for Reviewing Advances in Key Hand–Eye Technologies for Harvesting Robots in Standardized Orchards. Abbreviations: two-dimensional (2D); three-dimensional (3D); time-of-flight (ToF); light detection and ranging (LiDAR); artificial neural network (ANN); naive Bayes (NB); k-nearest neighbors (KNN); random forest (RF); decision tree (DT); support vector machine (SVM); gray-level co-occurrence matrix (GLCM).

six-dimensional (6D) pose of fruits in real time and generated corresponding grasping and harvesting actions based on the perceived target information. Such systems illustrate the principal advantage of the eye-in-hand configuration: visual perception can be continuously updated as the manipulator approaches the target, providing a basis for close-range perception and perception–action coordination. Nevertheless, the moving camera introduces additional requirements for dynamic coordinate transformation and calibration accuracy, while its restricted field of view may limit global target organization and task planning.

2.1.2. Eye-to-hand fruit harvesting robots

Within an eye-to-hand setup, the visual sensor is secured to the robot base or installed separately from the manipulator, permitting the camera to survey a considerable expanse of the harvesting workspace from a holistic viewpoint. As the camera remains fixed with respect to the manipulator, this setup promotes global target recognition, multi-target positioning, job assignment, and motion planning. Thus, it is uniquely advantageous for harvesting systems that need simultaneous surveillance of many fruits or cooperation between several manipulators. Williams et al. [30] created a four-arm kiwifruit harvesting system that employs a global vision setup, allowing for the sharing of visual information across numerous manipulators to enhance coordinated harvesting and shorten the harvesting cycle to 5.5 s. Nevertheless, the fixed vantage point restricts the visual system's proficiency in actively managing occlusion. When the manipulator gets closer to a target, elements like fruits, leaves, branches, or the robotic arm may obstruct the view, impairing the dependability of close-range localization. Relative to eye-in-hand configurations, eye-to-hand solutions provide superior worldwide observation proficiency but inferior viewpoint versatility in heavily obstructed settings [31]. These interdependent attributes have driven the development of hybrid global-local vision systems.

2.1.3. Hybrid hand–eye fruit harvesting robots

Hybrid hand-eye configurations integrate the expansive environmental perception of eye-to-hand vision with the high local localization accuracy and adaptable perspective of eye-in-hand vision. Global vision facilitates fruit identification, target determination, and motion strategizing, and local sensing enhances fruit localization, attachment-structure alignment, picking coordinates, and obstacle recognition, hence allowing for more dependable grabbing or severing in densely populated crop settings. For example, strawberry harvesting systems have incorporated fruit localization together with the perception of surrounding structures such as cultivation tables and support straps, allowing candidate targets to be filtered according to a safe manipulation region before execution [32]. At the system level, the integration of fruit detection, depth information, growth-scene categorization, and motion planning further enables the robot to determine whether a detected fruit is actually suitable for harvesting rather than directly executing a picking action based solely on fruit recognition [33]. A major technical characteristic of hybrid hand–eye perception is therefore the transition from single-stage localization to hierarchical coarse-to-fine localization. At a relatively long distance, the system primarily requires robust fruit detection and approximate three-dimensional localization, whereas fine plant structures such as peduncles, bifurcated stems, and cutting points become more important after the manipulator approaches the target. This principle was explicitly implemented in litchi harvesting by separating perception into long-distance fruit-cluster localization and close-distance picking-point localization: the You Only Look Once (YOLO) detector, specifically YOLOv5, and point-cloud clustering were first used to separate fruit clusters and determine the harvesting sequence, after which a region-based convolutional neural network (R-CNN), specifically Mask R-CNN, was applied at close range to segment the fruit-bearing branches and determine the final picking position [34]. Similar coarse-to-fine behavior can also be achieved through manipulator-guided visual servoing. In cherry-tomato harvesting, image-space deviation between the fruit center and camera center was used to adjust the manipulator posture, followed by laser ranging and coordinate transformation for more accurate three-dimensional localization [35]. For sweet peppers, multi-view RGB-D observations acquired during manipulator scanning can be registered into a unified three-dimensional model, reducing point-cloud noise and alleviating partial occlusion before grasp-pose calculation [36]. Once the target has been observed at a sufficiently close distance, perception can further shift from fruit-level localization to operation-oriented structural estimation; for example, fruit and stem segmentation can be used to estimate their relative positions and angular relationship, thereby determining a stem-dependent end-effector pose for grasping and cutting [37]. More recently, hybrid hand–eye coordination has evolved from predetermined global–local observation toward active perception, in which the robot dynamically determines whether the current visual information is sufficient for harvesting and actively changes the observation viewpoint when necessary. This is especially useful for fruit clusters whose peduncles or cutting points may be completely hidden from the initial view. Instead of treating occlusion only as an image-recognition problem, active sensing links target visibility directly with manipulator motion. In truss-tomato harvesting, the occlusion ratio and relative position between the fruit and its occluders were used to guide camera movement; the manipulator repeatedly acquired observations from new viewpoints until the peduncle reached an acceptable detection confidence, after which the estimated picking point and peduncle inclination were used to adjust the end-effector orientation [38]. This strategy has been further developed into explicit view-planning and next-best-view selection. For grape harvesting, candidate viewpoints can be evaluated according to the spatial coverage of the occluded picking region and the associated motion cost, enabling the manipulator-mounted depth camera to search for a viewpoint from which the stem becomes visible [24]. More recent approaches replace repeated candidate-view search and three-dimensional map evaluation with learned viewpoint policies; self-supervised deep reinforcement learning can directly generate camera-motion actions from visual and robot-state information, thereby reducing the decision time required to expose occluded grape stems [39].

In summary, hybrid hand–eye fruit harvesting is progressing from a basic amalgamation of global and local cameras to a sophisticated hierarchical perception framework that encompasses global scene comprehension, approximate target identification, detailed close-range structural enhancement, multi-perspective information integration, and task-oriented active viewpoint modification. In contrast to solely eye-to-hand systems, this arrangement enhances the visibility of intricate picking structures and offers greater adaptability to occlusion; in comparison to exclusively eye-in-hand systems, global or long-range information can influence target selection, approach trajectory, and manipulator movement prior to close-range engagement. Nonetheless, these benefits are coupled with heightened demands for multi-sensor calibration, coordinate-frame translation, cross-view target association, perception-stage transition, and real-time information amalgamation. As a result, the primary technological challenge has shifted from just the camera's placement to the coordination of visual data across various spatial scales and perspectives, in accordance with the demands of detection, approach, localization, and harvesting execution.

2.2. Key influencing factors constraining hand-eye coordination performance

The aforementioned hand-eye coordination technology operates as an operational closed-loop under ideal conditions. However, when deployed in complex and highly variable orchard environments rather than controlled laboratory settings, the system is affected by multiple environmental disturbances. Therefore, this section will systematically explore the constraints of hand-eye coordination technology in practical applications from the dimensions of orchard environments, visual perception, and executive mechanisms.

In contrast to industrial operating settings characterized by steady circumstances and consistent structures, orchard environments display considerable dynamism and complexity, which directly influence the dependability of fruit identification and localization (as seen in Fig. 3). Environmental issues mostly pertain to external disruptions inside orchard settings, including fluctuations in light and weather, backlighting and shading influences, fog, and picture deterioration caused by mechanical vibrations. Variations in lighting conditions may result in overexposure or underexposure, diminished contrast, and a loss of texture and edge details [40–43], whereas vibrations from wind or robotic movement frequently lead to motion blur and inconsistent exposure. Fruit growth–related factors are associated with the biological characteristics and spatial distribution of fruits, such as occlusion by leaves and branches, partial visibility of fruit surfaces, fruit vibration, and overlapping or clustering among adjacent fruits. These intrinsic factors prevent complete observation of individual fruits and increase ambiguity in segmentation and pose estimation, further amplifying localization uncertainty [44]. Errors accumulated in the visual perception stage are subsequently propagated to the executive body, affecting path planning accuracy, positioning error tolerance, and grasp execution. Mechanical arm vibrations, time delays, and differences in end-effector design and fault tolerance capabilities further influence the stability and success rate of picking actions. As a result, variations in picking strategies and system robustness are required to cope with different orchard conditions and fruit arrangements [45].

Among these factors, reliable acquisition of visual and spatial information constitutes the first step in limiting error propagation through the hand–eye coordination chain. The characteristics of the visual sensor determine what information can be observed from the harvesting environment and directly affect subsequent fruit recognition, three-dimensional localization, and manipulation-oriented perception. Different sensing modalities vary considerably in their imaging mechanisms, spatial information acquisition capabilities, and robustness to illumination, texture, and environmental interference. Therefore, the following section reviews the principal visual sensors used in fruit harvesting robots and examines their suitability for fruit recognition and localization.

3. Visual sensing technologies for fruit-harvesting robots

Reliable fruit recognition and localization begin with the acquisition of visual information, and the sensing modality directly determines the appearance and spatial information available for subsequent perception and manipulation. As illustrated in Fig. 4, fruit-harvesting robots employ a wide range of visual sensors, including monochrome, RGB, infrared, spectral, and thermal cameras for appearance-related information acquisition, as well as stereo cameras, ToF cameras, LiDAR sensors, and structured-light cameras for depth and geometric information acquisition. These sensing modalities differ in their sensitivity to illumination, target texture, occlusion, sensing distance, and workspace scale, resulting in distinct advantages and limitations in complex orchard environments. From the perspective of the dimensionality and type of perceptual information provided, visual sensing technologies for fruit-harvesting robots can therefore be broadly categorized into 2D imaging sensors and 3D sensing systems. The sensing principles, representative applications, advantages, and limitations of these two categories are reviewed in the following sections.

3.1. Two-dimensional visual sensors

As illustrated in Figs. 5A, 2D visual sensing represents orchard scenes in the image plane and provides appearance- and image-domain information for fruit perception, including intensity, color, texture, contour, spectral and thermal responses, semantic categories, and pixel coordinates. Subsequently, color cameras equipped with charge-coupled device or complementary metal-oxide-semiconductor sensors have been widely used in fruit recognition research. In addition to geometric and texture features, color cameras can acquire color information, which has greatly improved the accuracy of fruit recognition [46]. However, color cameras struggle to distinguish fruits in scenarios where the fruit color is similar to the background. Spectral cameras can compensate for this limitation of color cameras by capturing both spectral and color information simultaneously. Due to the differences in reflectivity wavelengths between the background and fruits, spectral cameras can detect fruits even when their color is similar to the background [47]. Thermal imaging can capture the temperature characteristics of fruits. Studies have shown that fruits absorb more heat and

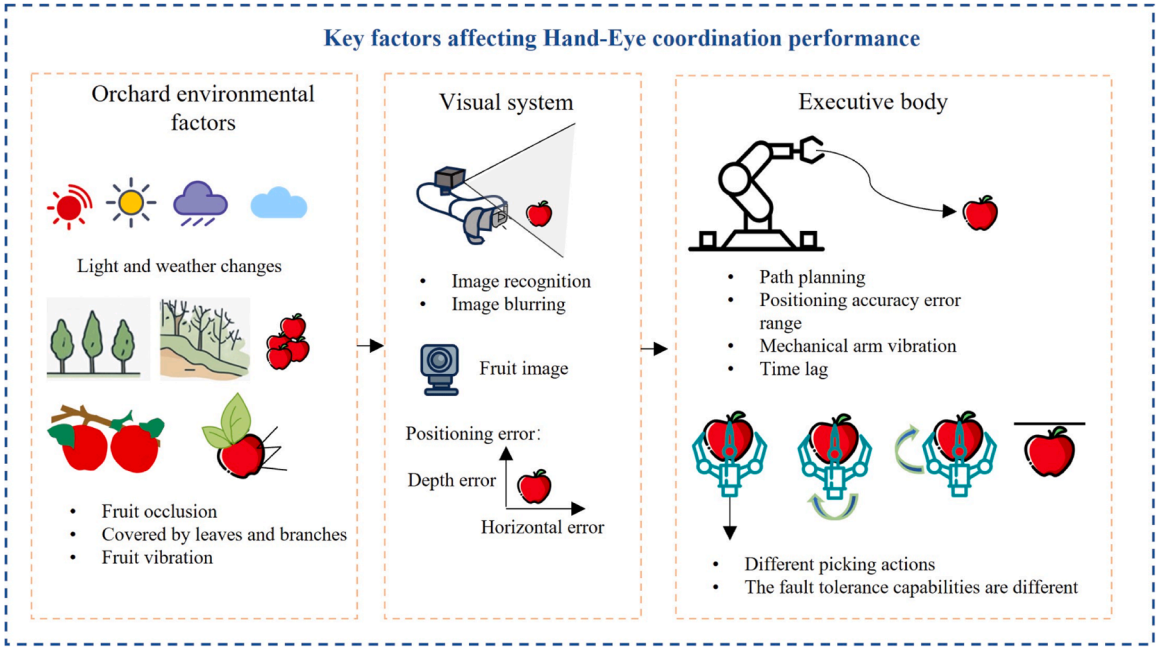


Fig. 3. Factors Constraining the Successful Picking via Hand-Eye Coordination Technology.



Fig. 4. Typical visual sensing modalities used in fruit-harvesting robots.

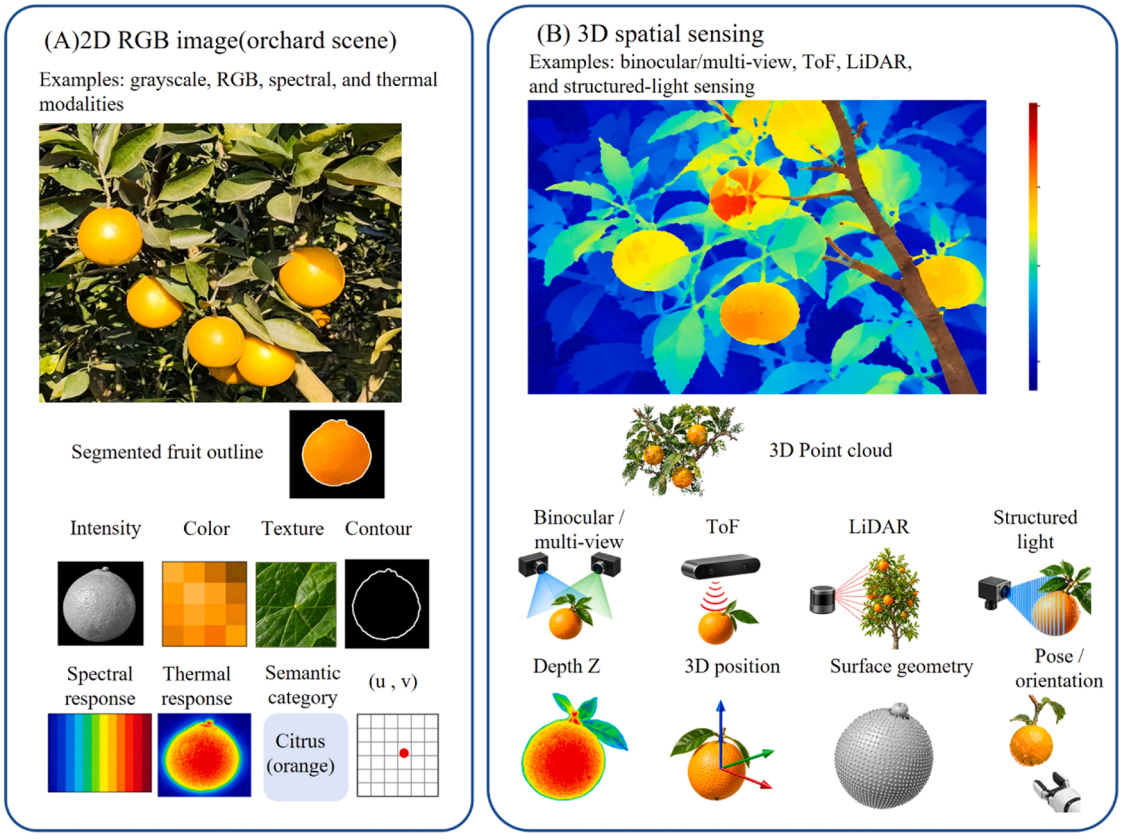


Fig. 5. Comparison of 2D and 3D visual sensing modalities and their perceptual outputs for robotic fruit harvesting.

radiate more heat than the background, but this method performs better under daytime conditions and is mainly used to distinguish plant materials in practical applications [48].

The exploration of 2D sensors began with basic grayscale imaging. Early studies utilized black-and-white cameras to detect fruits by analyzing geometric and textural features of images [49]. However, this approach overlooked color, a crucial feature for fruit recognition. While RGB cameras provide rich texture information enabling high-precision instance segmentation [50–52], they lack robustness in unstructured orchard environments. Specifically, fluctuating natural illumination causes significant drift in color feature distribution, while similarity between green fruits and foliage leads to frequent false negatives.

Given the identical shades shown by various fruits alongside their backgrounds, which hinders identification using color cameras, spectral and thermal imaging have been proposed to alleviate color confusion and illumination difficulties by exploiting information that remains unseen by human eyesight. On the other hand, the hyperspectral imaging technique capitalizes on the contrast in reflectance between fruits and the backdrop in certain spectral ranges. Van Beers et al. [53] and Wang et al. [54] demonstrated that, by analyzing reflectance at specific wavelengths, hyperspectral imaging can not only distinguish fruit targets but also non-destructively assess the maturity and internal quality of apples. However, its high cost and complex data processing requirements limit its widespread application. In contrast, thermal imaging technology detects fruits by exploiting the temperature differences between fruits and leaves. Gan et al. [55] fused color and thermal imaging for immature citrus detection, improving recall from 78.1% to 90.4% and precision from 86.6% to 95.5%. However, because 2D imaging lacks metric depth, 3D sensing is required for accurate fruit localization and robotic manipulation.

3.2. Three-dimensional visual sensors

As illustrated in Figs. 5B, 2D visual information alone is insufficient for robotic harvesting because the manipulator requires metric spatial information, including target depth, 3D position, surface geometry, and pose or orientation, to perform accurate approach, grasping, and detachment. Accordingly, 3D visual sensing extends image-plane perception into spatial perception by recovering depth information and reconstructing the geometric structure of fruits and their surrounding environment. Existing 3D sensing approaches can generally be divided into two categories. The first is passive vision, represented by binocular and multi-view systems, which estimates depth from disparities or geometric correspondences between images acquired from different viewpoints [56]. The other type directly performs distance measurement using light to acquire depth maps of images [57], including ToF cameras, LiDAR, and structured light cameras.

Binocular vision calculates depth by simulating human eye parallax, and has attracted considerable attention due to its low cost and simple structure. Hou et al. [58] proposed a method that successfully achieves the localization of citrus picking points by combining a binocular vision system with Mask R-CNN. Lei et al. [59] adopted a binocular vision system for avocado image acquisition, trained based on texture matching using YOLOv7, with a maximum error of 4.279 mm and a standard deviation of 1.142 mm. However, its accuracy is highly dependent on the correct matching of image features, and it is prone to failure in low-texture or repetitive texture regions. To improve the field of view and accuracy, researchers have further explored multi-camera solutions such as quad-camera systems, which reduce errors through point cloud stitching, but the complexity of system calibration and data synchronization also increases accordingly [10]. ToF cameras and LiDAR sensors both calculate distance by measuring the flight time of light. ToF cameras, with their small size, are often integrated into end effectors for close-range precise localization or 3D reconstruction [60–62]. However, they have low resolution and limited effective distance. In contrast, LiDAR offers a wider scanning range and higher accuracy, demonstrating significant advantages in orchard navigation [63] and large-scale point cloud map construction [64]. Their common advantage is strong resistance to ambient light interference, but high cost, massive data volume, and dependence on powerful algorithms are the main obstacles to their popularization and application.

Structured light cameras achieve high-precision 3D reconstruction at close range by projecting light with a known pattern onto objects and decoding the deformed pattern into a depth map [65]. For instance, Lu et al. [66] utilized a structured light system to realize real-time detection of tiny scratches on apple surfaces. Its core advantages lie in high precision and insensitivity to the surface texture of objects, but its drawbacks are equally prominent: the projected light is susceptible to interference from strong ambient light (especially sunlight), which greatly restricts its application in outdoor open scenarios.

4. Machine vision in fruit-harvesting robots

As summarized in Fig. 6, machine vision for robotic fruit harvesting has undergone a clear technological evolution from traditional geometric vision and shallow machine learning toward deep learning, three-dimensional perception, and increasingly operation-oriented visual understanding. Early studies mainly relied on handcrafted color, shape, texture, and geometric features, whereas convolutional neural network (CNN)-based methods gradually became dominant with the rapid development of two-stage and, particularly, one-stage detection frameworks. More recently, Transformer-based architectures have received increasing attention, while RGB-D/stereo vision, point-cloud-based 3D perception, and picking-point localization have expanded visual perception from two-dimensional object recognition toward spatial and structural understanding. Meanwhile, active vision, visual servoing, and emerging foundation models and vision-language models indicate a further shift toward adaptive perception and more generalizable semantic understanding. These trends demonstrate that machine vision in harvesting robots is no longer limited to determining whether and where a fruit is present, but is progressively required to provide information that is directly relevant to robotic manipulation, including target structure, three-dimensional position, orientation, and executable grasping or cutting locations.

In unstructured orchards, visual recognition is impeded by fluctuations in lighting, obstruction, aggregation of fruit, movement of targets, and morphological diversity. Effective harvesting requires dependable fruit identification and segmentation, posture assessment, and picking-point localization, assessed not only by recognition precision but also by the spatial dependability and practical applicability of perception results.

4.1. Fruit detection and segmentation

Early fruit detection primarily relied on traditional image processing and shallow machine-learning methods, in which manually designed color, geometric, and texture features were used to distinguish fruits from complex backgrounds. As summarized in Table 1, these methods were applied to a wide range of crops using combinations of color-space transformation, thresholding, shape analysis, handcrafted descriptors, and conventional classifiers. Although satisfactory recognition performance was reported for particular fruits and experimental conditions, these methods generally depended strongly on predefined visual features, making them susceptible to illumination variation, occlusion, fruit overlap, and morphological variability. With the introduction of deep learning into agricultural vision, convolutional neural networks enabled hierarchical and discriminative features to be learned directly from image data, reducing the dependence on manually designed features. As further summarized in Table 1, Faster R-CNN, Mask R-CNN, and the YOLO series have been increasingly applied to different fruit species and sensing modalities, demonstrating the broader applicability of data-driven feature learning in complex harvesting environments. It should

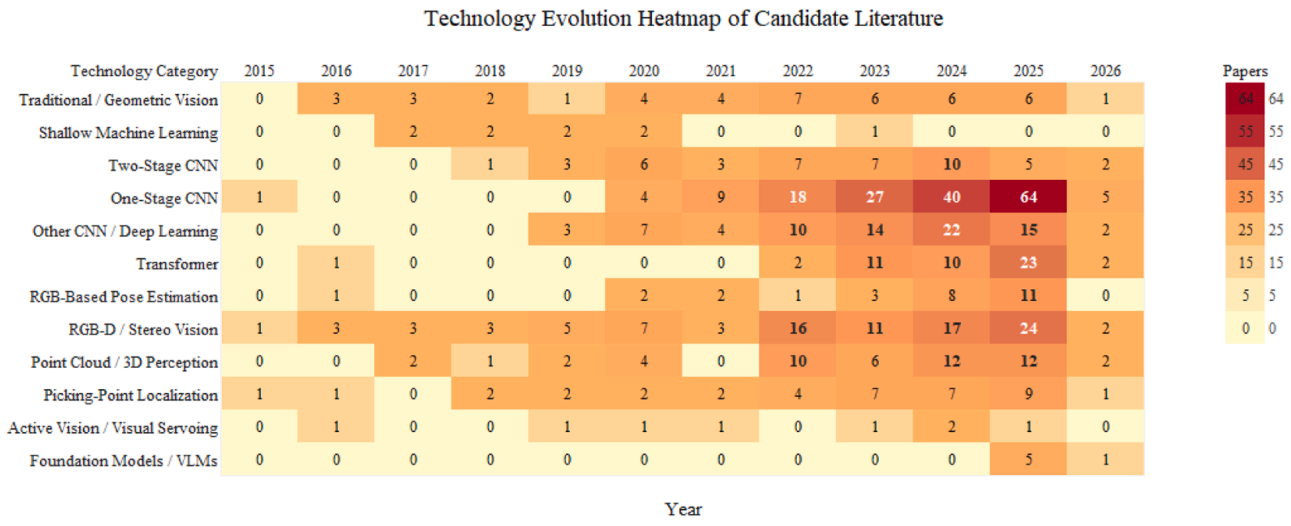


Fig. 6. Temporal evolution of machine vision technologies for robotic fruit harvesting from 2015 to 2026. Abbreviation: vision-language models (VLMs).

Table 1
Representative handcrafted feature-based and deep-learning approaches for fruit and vegetable detection.

Machine Vision	Fruit categories	Input modality	Algorithms	Reported computational speed	Reported performance / %	Reference
Traditional image processing algorithms	Tomato	RGB image	adaptive boosting classifier + color analysis	0.0667 f/s	96.5(P)	[69]
	Mango	RGB image	Variance + Hessian + luminance/chrominance (YCbCr) color filter	0.05 f/s	78.3(P)	[70]
	green citrus	RGB image	SVM + Hough line + scale-invariant feature transform	Not given	80.4(P)	[71]
	Apple	RGB image	Hue-intensity-saturation (HIS) color + SVM	2.84 f/s	89.0(P)	[72]
	Strawberry	RGB image	Hue-saturation-value color + histogram of oriented gradients (HOG) + SVM	0.50 f/s	87.0(P)	[73]
	Blueberry	RGB image	HOG + linear SVM + L*a*b* color + KNN/template matching with weighted Euclidean distance (TMWE)	0.15 f/s	96.1(P)	[74]
Deep learning	Tomato	RGB image	Faster R-CNN	11.84 f/s	97(P)	[75]
	Cucumber	RGB image	Mask R-CNN	2.89 f/s	90.68(P)	[76]
	Pecan	RGB and depth images	Faster R-CNN	11.90 f/s	95.9 (mAP)	[77]
	Apple	RGB image	Mask R-CNN	6.67 f/s	94.6(P)	[78]
	Green persimmon	RGB image	Faster R-CNN	1.39 f/s	98.4 (mAP)	[79]
	Dragon fruit	RGB and depth images	YOLOv3	34.6 f/s	94.6(P)	[80]
	Litchi	RGB image	YOLOv8n	5.65 f/s	98.7(P)	[81]
	Apple	RGB and depth images	YOLOv5	58 f/s	92.4(P)	[82]
	Citrus	RGB image	YOLOv8	1.52 f/s	91.79(P)	[83]

Note: The quantitative results listed in this table are those reported in the original studies. Because these studies employed different datasets, evaluation metrics, image resolutions, experimental environments, and computing platforms, the reported performance and computational speeds are not directly comparable across studies. The values are provided only to illustrate the reported performance characteristics of representative methods under their respective experimental settings. "Not reported" indicates that the corresponding information was not provided in the original study.

be noted, however, that the quantitative results reported in different studies were obtained using different datasets, evaluation metrics, image resolutions, experimental conditions, and computing platforms. Therefore, the reported detection performance and computational speeds in Table 1 should be interpreted as study-specific results rather than as directly comparable benchmarks. Overall, this technological transition reflects a shift from handcrafted feature engineering toward data-driven representation learning and has established deep-learning-based detection and segmentation as an important paradigm in robotic harvesting vision [67,68].

4.1.1. Traditional image processing and shallow machine learning

As illustrated in Fig. 7, traditional fruit recognition generally follows an image acquisition–preprocessing–feature extraction–feature selection–classification–post-processing pipeline, in which manually designed color, shape, and texture features are extracted to distinguish fruits from backgrounds. For example, Mohammadi et al. [84] combined differences among RGB channels with linear discriminant analysis and quadratic discriminant analysis classifiers for persimmon grading and achieved an accuracy of 90%. However, the effectiveness of color-based methods decreases substantially when fruits and backgrounds exhibit similar spectral characteristics or when illumination changes alter their apparent colors. Under such conditions, shape and texture features provide complementary information for target recognition. Zhang et al. [85], for instance, employed the Hough Circle Transform for apple recognition based on geometric characteristics. Although satisfactory recognition performance was obtained, the method relied strongly on the regularity and completeness of fruit contours, making it susceptible to backlighting, occlusion, and fruit overlap. To improve the adaptability of handcrafted-feature-based approaches, subsequent studies

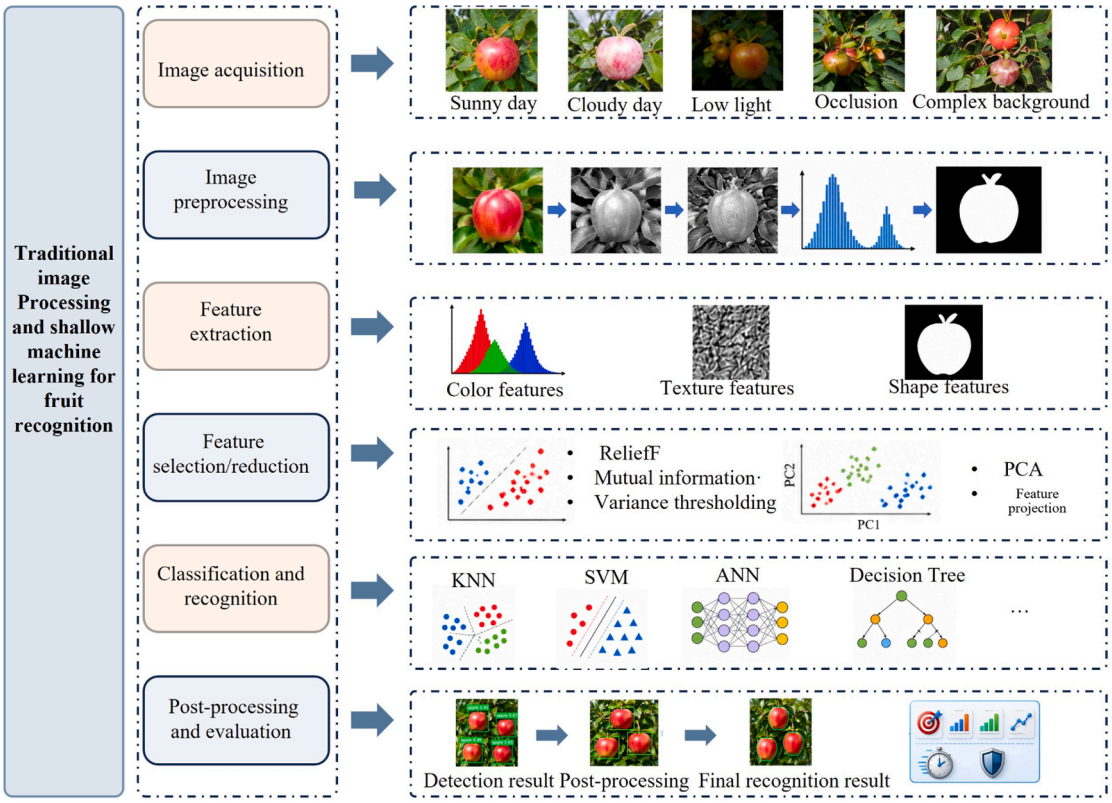


Fig. 7. Statistical machine learning-based fruit recognition process. Abbreviation: principal component analysis (PCA). PC1 and PC2 denote the first and second principal components, respectively.

mainly focused on classifier optimization and multi-feature fusion. Li et al. [86] evaluated classifiers such as KNN, SVM, and shallow neural networks for blueberry and mango maturity recognition, respectively, demonstrating that recognition performance depends not only on the extracted features but also on the suitability of the classifier for a specific task. In parallel, combining complementary visual information became an important strategy for improving recognition under complex backgrounds. Syazwani et al. [87] integrated color, shape, and texture information to improve fruit recognition performance, showing that multi-feature fusion can compensate, to some extent, for the limitations of individual handcrafted features. Nevertheless, the performance of these methods remains strongly dependent on predefined feature representations and scene-specific thresholds or geometric assumptions. Their robustness therefore deteriorates under severe occlusion, fruit clustering, illumination variation, and large morphological differences. These limitations have driven fruit detection toward deep-learning-based approaches, in which hierarchical visual features can be learned directly from data rather than being manually designed.

4.1.2. *Deep learning-based methods*

Deep-learning-based fruit perception has gradually replaced handcrafted feature extraction with automatic hierarchical feature learning, substantially improving the robustness of fruit detection and segmentation in complex agricultural environments. Existing methods are primarily based on CNN architectures and can generally be divided into two-stage and one-stage frameworks. Two-stage detectors, represented by the R-CNN series, first generate candidate regions and subsequently perform classification and bounding-box regression, generally achieving high localization accuracy at the cost of greater computational complexity [88,89]. In contrast, one-stage detectors, represented by YOLO and the Single Shot MultiBox Detector (SSD), directly predict object categories and locations within a unified network, providing higher inference efficiency and stronger suitability for real-time harvesting applications [90,91]. More recently, Transformer-based architectures have been introduced to enhance long-range dependency modeling and global contextual representation, with Vision Transformer (ViT) and Swin Transformer representing typical examples [92,93]. Rather than exhaustively describing the generic evolution of these architectures, the following sections focus on how they have been adapted to harvesting-specific perception tasks, including fruit detection and segmentation, fine attachment-structure perception, picking-point localization, and three-dimensional robotic manipulation.

4.1.2.1. Two-stage models. The R-CNN series represents a typical two-stage object detection framework whose early architectural evolution primarily focused on reducing redundant feature computation and integrating region proposal generation into the detection network. As shown in Fig. 8, early R-CNN generated candidate regions using Selective Search and independently extracted CNN features from each proposal, resulting in substantial computational redundancy. Subsequent developments progressively shared convolutional features across proposals, with SPPNet and Fast R-CNN improving feature reuse and region-level prediction efficiency [94]. Faster R-CNN further introduced a region proposal network (RPN) that operates on shared backbone features, replacing external proposal generation with a learnable module and establishing an end-to-end two-stage detection framework [95]. Accordingly, agricultural adaptations of Faster R-CNN have mainly focused on enhancing robustness to occlusion and background complexity, strengthening multi-scale feature representation, and reducing computational cost for robotic deployment. For harvesting-oriented perception, occlusion-aware category modeling has been introduced to distinguish different fruit visibility states, allowing detection results to provide information related not only to fruit presence but also to harvesting accessibility [44]. Meanwhile, data augmentation and stronger backbone networks have been employed to improve robustness to variations in object appearance and complex

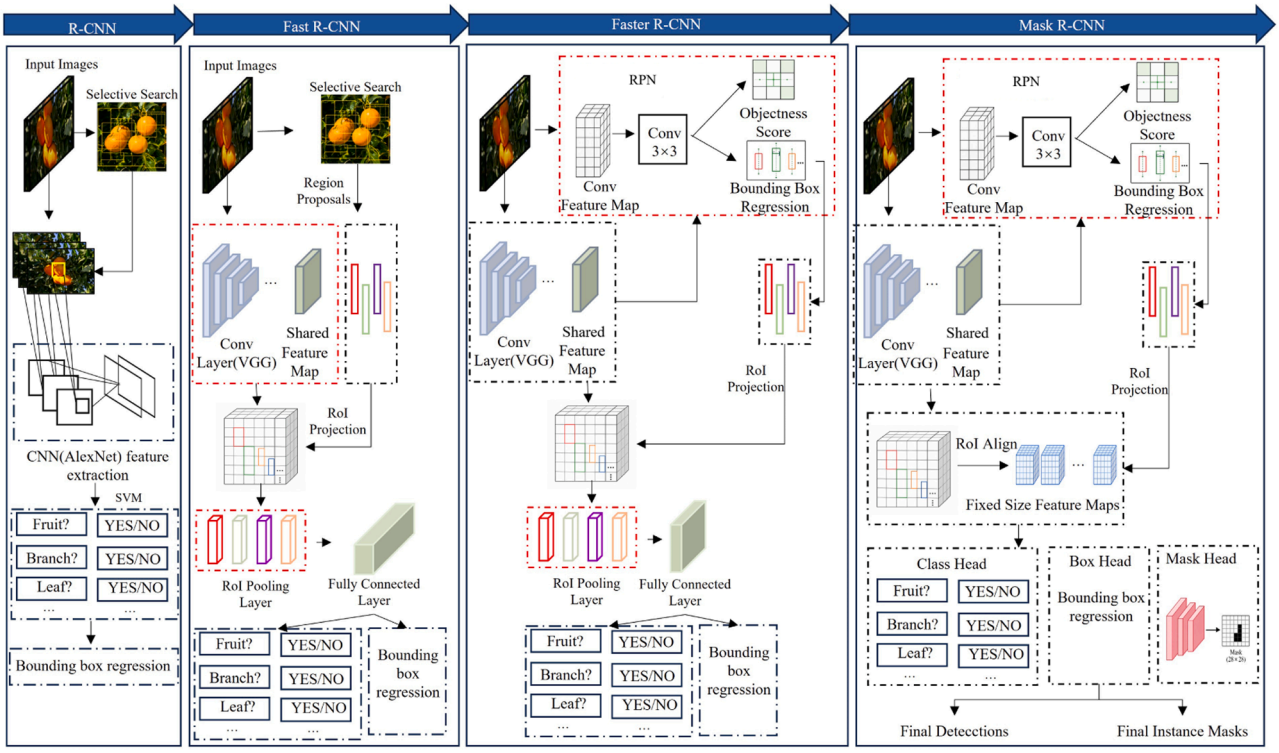


Fig. 8. R-CNN series schematics. Abbreviation: convolution (Conv).

backgrounds, as demonstrated in coconut maturity detection using a 50-layer residual network (ResNet-50) combined with rotation- and color-based augmentation [96]. As attention shifted toward small and scale-varying agricultural targets, feature pyramid structures and attention mechanisms were incorporated to enhance multi-scale representation and suppress irrelevant background information; for example, the integration of ResNet, a feature pyramid network (FPN), and a convolutional block attention module (CBAM) improved the representation of small-scale targets in complex scenes [97]. In parallel, lightweight backbone networks have been explored to meet the computational constraints of on-robot deployment. Replacing the conventional 16-layer Visual Geometry Group (VGG) model VGG16 with MobileNetV3 reduces model parameters and computational complexity while maintaining competitive detection accuracy, thereby improving the suitability of Faster R-CNN for embedded harvesting platforms [98].

Bounding-box-level detection alone is insufficient for robotic harvesting tasks that require fine-grained information such as fruit contours, fruit stems, and manipulable regions. Mask R-CNN extends Faster R-CNN from object detection to instance segmentation by introducing RolAlign and an instance-specific mask prediction branch, thereby preserving spatial correspondence and enabling pixel-level representation of object boundaries and fine structures [99]. In agricultural applications, improvements to Mask R-CNN have progressively shifted from enhancing general fruit segmentation toward recognizing fine structures directly related to harvesting operations. Attention mechanisms have been incorporated into both backbone feature extraction and mask prediction to improve robustness under illumination variation, fruit overlap, and foliage occlusion; for example, the combination of efficient channel attention (ECA) and a dual attention network enhances feature representation for grape detection and segmentation in complex scenes [100]. As the perception target extends from fruits to slender structures such as stalks and pedicels, subsequent modifications have increasingly emphasized multi-scale feature representation and proposal preservation. The integration of CBAM with a ResNet-50-FPN backbone and the replacement of conventional non-maximum suppression (NMS) with Soft-NMS improve the detection of small citrus fruit stalks by enhancing relevant features and reducing erroneous suppression of overlapping proposals [58]. More importantly, fruit-peduncle spatial relationships can be further exploited for peduncle matching and picking-point localization, directly transforming instance-level perception into manipulation-oriented information. These developments reflect a broader transition of Mask R-CNN-based agricultural perception from general fruit segmentation toward operation-oriented representation of peduncles, boundaries, and fruit–stem connecting regions.

Recent improvements to Mask R-CNN have increasingly exploited the biological and spatial relationships between fruits and their associated stalks. Because fruits and pedicels differ substantially in scale and may be easily confused with surrounding branches, treating them as independent visual targets can limit recognition stability. To address this issue, structural priors have been incorporated into both region proposal generation and feature extraction. For example, Xu et al. [101] integrated RGB-D information into Mask R-CNN and proposed an instance-feature region proposal network, in which a fruit and its corresponding pedicel are modeled within a unified proposal. Combined with multi-task loss balancing and adaptive feature pooling, this strategy improves feature representation across the markedly different scales of fruits and pedicels. Such developments indicate a shift from generic visual feature enhancement toward structure-constrained perception, in which fruit–stalk relationships are explicitly exploited to support recognition. Nevertheless, severe overlap among instances of the same class and incomplete structural observations can still lead to segmentation leakage and inaccurate instance separation. In parallel, R-CNN-based perception has progressively expanded from two-dimensional detection and instance segmentation toward three-dimensional spatial perception for robotic harvesting. In these systems, Faster R-CNN or Mask R-CNN generally serves as a reliable 2D perception front end, while stereo vision, RGB-D sensing, or point-cloud processing is subsequently used to recover depth and geometric information. This combination has enabled the estimation of fruit position and size from stereo segmentation results [102,103], the transformation of RGB-D instance masks into structured point-cloud representations for

the three-dimensional analysis of grape clusters, stems, and leaves [104], and the recovery of apple 3D coordinates through instance contours, stereo matching, and triangulation [105]. Accordingly, the role of R-CNN-based models in harvesting robots has gradually evolved from standalone 2D object recognition toward providing structured visual priors for subsequent 3D localization and robotic manipulation.

Overall, the evolution of R-CNN-based methods reflects not only improvements in detection architecture but, more importantly, a progressive expansion of perception capabilities toward robotic harvesting. Through feature sharing, learnable region proposal generation, and instance-level segmentation, the R-CNN framework has evolved from generic object detection toward increasingly fine-grained visual perception. In agricultural applications, further integration of multi-scale feature fusion, attention mechanisms, lightweight backbones, fruit–stem structural constraints, and RGB-D/stereo sensing has extended the available information from fruit presence and 2D location to fruit contours, peduncle structures, picking-related regions, and 3D spatial information. This evolution demonstrates a broader transition from object-centered recognition toward operation-oriented perception for robotic manipulation. However, severe occlusion, slender fruit stalks, overlapping instances, and cross-scene variations continue to limit recognition stability, while two-stage proposal generation and multimodal processing increase computational demands. Therefore, balancing fine-grained perception accuracy, environmental robustness, and real-time on-robot deployment remains a major challenge for R-CNN-based fruit perception.

4.1.2.2. One-stage models. As illustrated in Fig. 9, the YOLO series represents a typical one-stage object detection framework that has progressively evolved from basic unified detection toward improved multi-scale representation, lightweight deployment, and increasingly end-to-end prediction. YOLOv1 established the basic end-to-end detection paradigm by jointly performing object classification and bounding-box regression within a single convolutional network, but its fixed-grid prediction mechanism limited its performance for small and densely distributed targets [106]. Subsequent versions progressively addressed these limitations improved scale adaptation, feature representation, and computational efficiency. YOLOv2 improved scale adaptation through anchor-based and multi-scale prediction strategies [107], while YOLOv3 combined Darknet-53 with multi-scale prediction to improve the representation of objects at different spatial resolutions [108]. YOLOv4 further combined enhanced feature extraction, cross-scale feature fusion, and optimized training strategies, reflecting a shift from increasing backbone complexity toward coordinated optimization of detection performance. Later YOLO variants increasingly emphasized computational efficiency and deployment-oriented design. YOLOv5 strengthened model scalability and engineering deployment, whereas YOLOv6 further emphasized hardware-efficient feature extraction and decoupled prediction [109]. YOLOv7 further improved feature aggregation, gradient propagation, and parameter utilization [110]. YOLOv8 subsequently adopted anchor-free and decoupled prediction to improve detection flexibility and efficiency [111]. YOLOv9 subsequently focused on reducing information loss and improving gradient utilization [112], while YOLOv10 further advanced end-to-end detection by eliminating conventional NMS during inference and improving the balance between accuracy and computational efficiency [113]. YOLO11 continues this trend toward more efficient feature extraction and enhanced feature representation. For fruit-harvesting applications, the practical significance of these architectural advances lies primarily in improving the detection of small and occluded fruits, enhancing the perception of fine attachment structures, supporting real-time deployment, and extending conventional fruit detection toward keypoint and picking-point localization.

In fruit and vegetable computer vision, adaptations of YOLO have mainly focused on small and occluded fruit detection, multi-scale feature representation, and computational efficiency for real-time robotic deployment. For fine-scale targets and subtle visual differences, attention mechanisms and feature fusion have been introduced to improve the discrimination of fruits and harvesting-related structures. For example, Lu et al. [114] enhanced YOLOv5 with attention and bidirectional feature fusion to improve minor citrus defect detection, whereas Liu et al. [115] combined lightweight feature extraction with global attention to improve apple ripeness recognition while reducing model complexity. As fruit detection increasingly involves small and densely distributed targets, subsequent improvements have emphasized high-resolution features and scale-specific detection to preserve fine target information. RDE-YOLOv7 integrates RepGhost, a decoupled head, and ECA to improve dragon fruit detection in natural environments [116], while the introduction of a P2 detection layer, P2BiFPN, and an ultra-lightweight subspace attention module improves the

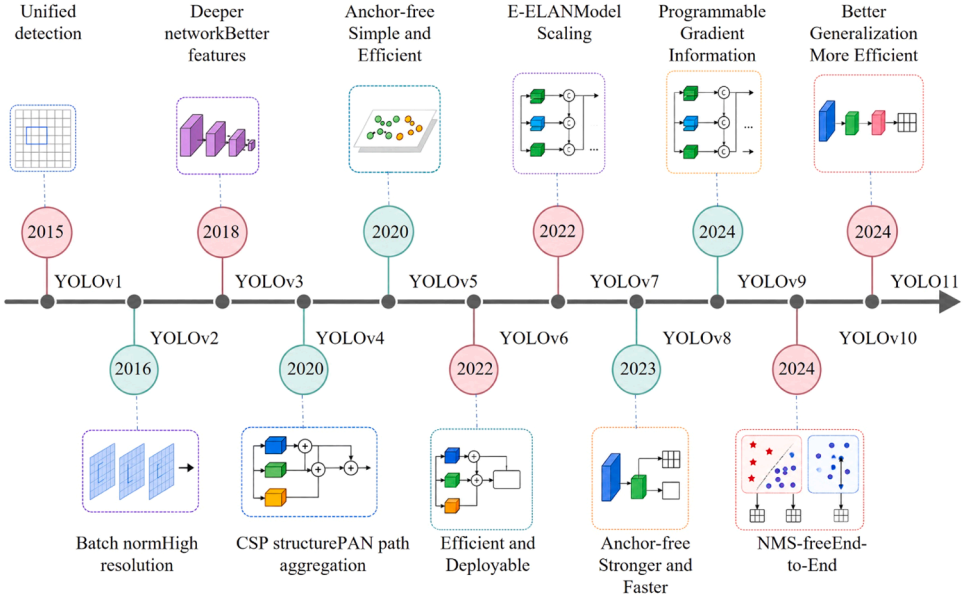


Fig. 9. Evolution of the YOLO series from YOLOv1 to YOLO11 and representative architectural improvements. C denotes concatenation, + denotes element-wise addition, and Batch norm denotes batch normalization.

representation of small apples at long distances [117]. Similar trends have continued in YOLOv8-based models, where modifications to feature extraction, loss functions, and regression strategies have been used to improve tomato detection and segmentation across different ripeness stages [118]. Overall, these adaptations indicate a shift from generic feature enhancement toward scale-aware, lightweight, and deployment-oriented perception for complex agricultural scenes.

As visual perception becomes increasingly coupled with robotic manipulation, YOLO-based methods have expanded beyond conventional 2D fruit detection toward keypoint estimation, picking-point localization, pose perception, and three-dimensional spatial reconstruction. Keypoint-based extensions enable harvesting-related structures to be predicted directly; for example, STRAW-YOLO employs YOLOv8-Pose to jointly detect strawberry and stem keypoints for subsequent picking-point localization and stem pose estimation [119]. Meanwhile, high-resolution features and small-object detection heads have been introduced to improve cherry-tomato and picking-point localization while maintaining real-time performance on embedded platforms such as Jetson Orin NX [120]. Beyond 2D perception, YOLO outputs are increasingly integrated with stereo vision and RGB-D sensing to recover 3D picking-related spatial information. YOLO-based detections have been combined with stereo point clouds for 3D grasping-position estimation of camellia fruits [121], with RGB-D point clouds, oriented bounding boxes (OBBs), and PCA for citrus orientation and picking-point estimation [122], and with the RAFT-Stereo matching model for the 3D localization of occluded pears [123]. These developments demonstrate a clear transition from object-centered fruit detection toward operation-oriented perception that directly provides structural, positional, and pose information for robotic harvesting.

Overall, the evolution of YOLO-based agricultural perception reflects a shift from accuracy-oriented fruit detection toward efficient and operation-oriented visual perception. Improvements in multi-scale feature representation, lightweight architectures, decoupled prediction, and end-to-end inference have enhanced robustness to small objects, occlusion, and complex backgrounds, while key point estimation and depth sensing have further extended YOLO from 2D detection to picking-point localization and 3D spatial perception. Nevertheless, most existing methods still rely primarily on single-frame observations and therefore remain sensitive to target displacement caused by branch and foliage disturbance, fruit oscillation, and robot motion. Future development should increasingly integrate temporal tracking, fruit–stem structural constraints, and RGB-D information to enable continuous target perception, stable pose estimation, and real-time updating of picking points in dynamic orchard environments.

As another representative one-stage detector, SSD adopts a multi-scale prediction strategy distinct from the unified prediction paradigm of early YOLO models, as shown in Fig. 10. Although both frameworks eliminate the region-proposal stage, SSD performs detection across multiple feature scales using predefined default boxes, improving its adaptability to objects of different sizes. The original SSD employs VGG16 as the backbone and adds convolutional layers to construct multi-scale prediction features. Predictions are generated from feature maps at different resolutions, allowing objects of different sizes to be detected at appropriate feature scales. This design improves multi-scale detection while retaining the computational efficiency of one-stage detection. However, the different SSD prediction layers operate relatively independently: shallow layers retain precise spatial information but contain weak semantics, whereas deeper layers provide stronger semantics at the expense of fine spatial details. Consequently, small fruits, heavily occluded targets, and dense agricultural objects can still be poorly represented [124]. With the transition from generic fruit detection to robotic harvesting, the computational burden of the VGG16 backbone became increasingly important. Although SSD retained a speed advantage over two-stage detectors, early implementations still relied heavily on high-performance GPUs. In sweet-pepper harvesting, SSD showed robustness to illumination variation and clustered fruits, shifting the main engineering concern toward real-time deployment on resource-constrained robotic platforms [125]. To reduce this burden, subsequent SSD variants increasingly adopted lightweight backbone replacement rather than increasing network complexity. Lightweight backbones such as MobileNet and Inception V2 reduce feature-extraction cost while maintaining effective visual representation. Increasing the input resolution from SSD300 to SSD512 further preserves fine-scale information for small-target detection. These modifications indicate that SSD can balance detection performance and computational cost through coordinated optimization of the backbone and input resolution. For cherry-tomato detection, SSD-Inception V2 achieved an AP of 98.85% and showed strong adaptability to illumination variation and occlusion [126]. This replacement-based strategy further promoted the development of lightweight SSD. Combining SSD with MobileNet reduced feature-extraction complexity and improved inference efficiency. In fruit detection and counting, SSD-MobileNet required approximately 60 ms per frame, compared with about 220 ms for Faster R-CNN with Inception V2, demonstrating its advantage when real-time processing is prioritized [127]. Model compression was subsequently introduced to further improve deployment efficiency without altering the basic SSD detection framework. An 8-bit quantized MobileNetV2 reduced model size from 17.5 MB to 4.5 MB and inference time from 163 to 103 ms per image, while the true detection rate decreased only slightly from 90.8% to 89.7%. Thus, SSD progressively evolved from VGG16-based detection toward lightweight and compressed models suitable for embedded deployment. Nevertheless, overlapping kiwifruits were still occasionally merged into a single detection, indicating that lightweighting improves efficiency but does not fundamentally resolve weak feature representation under dense occlusion [128]. As deployment efficiency gradually improved, later SSD research shifted its emphasis toward a more fundamental weakness of the architecture: insufficient interaction among multiscale feature layers.

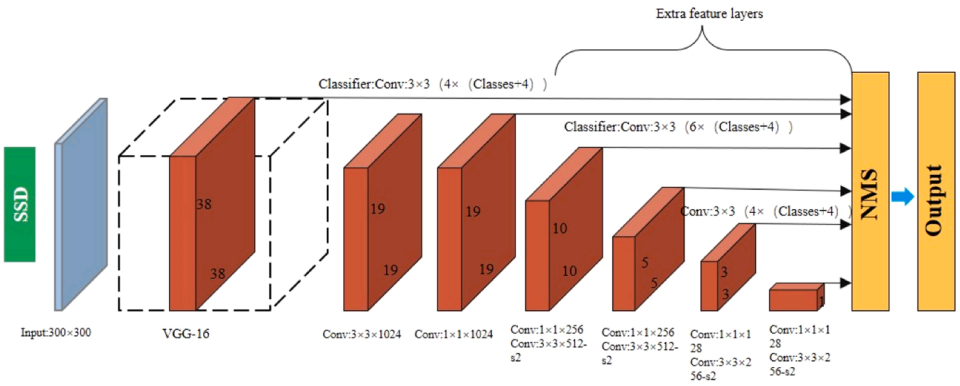


Fig. 10. SSD Network Architecture Diagram.

Instead of merely replacing VGG16, new structures began to add feature-enhancement branches and explicitly fuse shallow spatial details with deep semantic information. For citrus pest detection, the original single-backbone SSD was expanded into a VGG16–ResNet-50 dual-backbone architecture, while a deep–shallow feature fusion module was introduced to strengthen small-target representation. At the detection level, prior-box sizes were redesigned for small pests, and Repulsion Loss was added to suppress interference from adjacent objects and similar backgrounds. This represents an important change in SSD optimization: from reducing computation through module replacement to adding cross-layer feature interaction and loss constraints to recover information lost during downsampling, thereby improving small-object detection while maintaining real-time performance [129]. More recent SSD variants have further reconstructed both the feature-extraction operator and the neck. Fixed-grid convolution is inherently limited when agricultural targets exhibit irregular shapes, weak boundaries, or large geometric variations; accordingly, a deformable convolutional network (DCN) can adapt its sampling positions to target geometry. Meanwhile, the weak communication among the original SSD prediction layers can be compensated for by adding a path aggregation feature pyramid network (PAFPN), which establishes top-down semantic transfer and bottom-up localization-information propagation, and adding a balanced feature pyramid (BFP), which aggregates and redistributes features across different scales to reduce feature imbalance. In JujubeSSD, the progressive integration of pretrained feature transfer, DCN, PAFPN, and BFP increased mAP@0.5 from 91.3% to 97.1%, with BFP contributing the largest individual gain. This indicates that SSD improvement has progressed from modifying external parameters such as default boxes and input resolution toward reorganizing internal feature flow, adaptive sampling, and balanced multiscale representation [130].

Overall, SSD has largely retained its default-box-based multiscale prediction framework, with agricultural improvements focusing on default-box adaptation, backbone replacement, lightweighting, and multiscale feature fusion. Its development has gradually shifted from adapting to crop-specific target scales, to reducing computational cost for robotic deployment, and then to strengthening small-object and cross-scale feature representation. However, severe occlusion, dense overlap, and extremely small targets remain challenging, requiring more efficient feature interaction without sacrificing real-time performance.

4.1.2.3. Transformer-based models. Despite the success of CNN-based detectors, their predominantly local feature extraction limits the modeling of global context and long-range dependencies. To address this limitation, Transformer architectures have been introduced into computer vision to enhance global feature modeling. As illustrated in Fig. 11, the self-attention mechanism enables global interactions among image features, while representative architectures such as ViT and Swin Transformer further extend this capability through patch-based representation and hierarchical feature modeling, respectively [131]. Self-attention enables visual features to interact globally, allowing contextual information to be modeled within a unified representation. Single-head attention captures limited information. Multi-head attention further enables the model to capture complementary visual cues, such as texture and shape [132]. This global modeling capability is particularly valuable for fruit perception under severe occlusion in complex orchard environments. For example, global contextual information can help identify partially occluded fruits and infer their spatial relationships with surrounding leaves and branches. Transformer-based detectors, including ViT-derived and DETR-based architectures, further exploit global contextual information for fruit detection and structural perception. Such global modeling is particularly beneficial in orchards where fruits are frequently embedded in visually complex foliage and exhibit ambiguous local textures. Nevertheless, the quadratic computational cost of attention and the demand for large-scale annotated data limit their practical deployment on mobile agricultural robots. Current research focuses on hybrid CNN-Transformer designs and query-efficient DETR variants to reduce training cost and inference latency, but widespread field-level application remains nascent.

ViT represents a pioneering work that directly applies the Transformer architecture to image classification [133]. ViT processes images as sequences of embedded patches with positional information, which are subsequently encoded by a Transformer. With sufficient pre-training data, ViT has demonstrated competitive or superior performance to CNN-based models, as shown in Fig. 11B.

However, global self-attention in ViT introduces substantial computational cost for high-resolution detection and segmentation tasks. To reduce this computational burden, Swin Transformer introduces a hierarchical visual representation [134]. Its window-based and shifted-window attention mechanisms restrict attention computation to local regions while enabling cross-window information

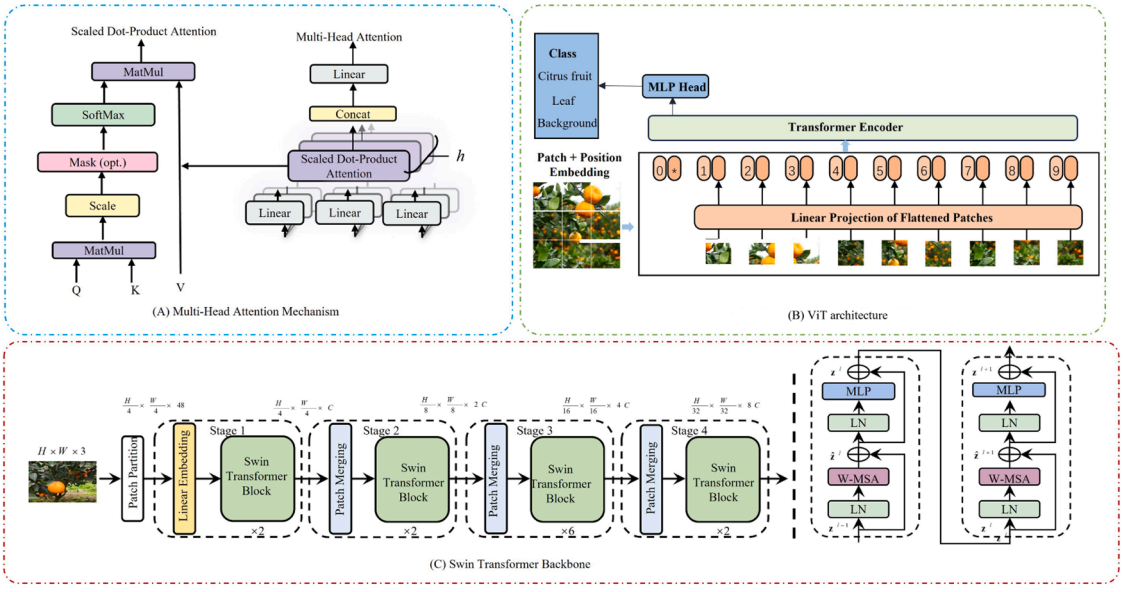


Fig. 11. Core attention mechanism and representative Transformer architectures for visual feature extraction. Abbreviations: multilayer perceptron (MLP); layer normalization (LN); window-based multi-head self-attention (W-MSA); shifted-window multi-head self-attention (SW-MSA). Q, K, and V denote query, key, and value, respectively; MatMul denotes matrix multiplication, and Concat denotes concatenation.

exchange. This design reduces computational complexity while preserving hierarchical feature representation, making Swin Transformer well suited as a backbone for detection and segmentation tasks, as shown in Fig. 11C.

In fruit perception, Transformer-based models have mainly developed along two application directions: serving as global-context backbones and forming hybrid architectures with CNN-based detectors. The first is the use of Transformer models as backbones for fruit recognition and segmentation. Their global contextual modeling is particularly useful for identifying heavily occluded fruits that cannot be reliably characterized from local appearance alone. To address this, architectures like Swin Transformer have been integrated as backbones [135], as shown in Fig. 11C. Their self-attention mechanism enables the model to capture long-range dependencies, effectively inferring the complete shape of a fruit even when only a fraction is visible—a capability with which traditional CNNs often struggle. Zheng et al. [136] also verified the superiority of the classifier based on Swin Transformer in strawberry appearance quality recognition. Although Transformers excel in capturing global dependencies critical for handling severe occlusion, their computational cost remains a bottleneck. A prevailing trend is the hybrid CNN-Transformer architecture, which combines the local feature extraction efficiency of CNNs with the global receptive field of Transformers. For instance, Gu et al. [137] and Neji et al. [138] embedded attention blocks into YOLO backbones, effectively reducing background interference while maintaining inference speeds acceptable for robotic hardware. Rong et al. [139] designed a semantic segmentation model with Swin Transformer as the backbone network, which was combined with the Unified Perceptual Parsing network (UPerNet) encoder to realize the recognition of mature tomato picking points. The accuracy reached 97.43%, which improved the semantic segmentation performance of fruit stems in complex backgrounds. In addition to these two types of applications, Culman et al. [140] combined self-supervised learning with Transformer models to automatically label pear data, reducing manual annotation effort and accelerating model training. In scenarios with insufficient data, automatic labeling improved detection accuracy; however, in cases with large amounts of data, classification errors may occur. For accurate training, subsequent manual fine-tuning is still necessary.

Collectively, these studies show that Transformer-based models can improve fruit and picking-related structure perception under occlusion, scale variation, and complex backgrounds. However, their high computational cost remains a key concern and optimization direction for deployment on resource-constrained picking robots. Consequently, lightweight Transformer architectures and model compression remain important directions for balancing perception accuracy and real-time robotic deployment [141].

4.2. Fruit pose estimation

Fruit pose estimation determines the position and orientation of fruits and harvesting-related structures to support end-effector approaching, grasping, and cutting. A complete spatial pose generally includes three translational and three rotational DOF, forming a 6D pose. As summarized in Table 2, existing methods have evolved from 2D geometric estimation and pose classification to keypoint-based estimation, 3D reconstruction, structure-constrained estimation, multi-information fusion, and active vision. Early RGB-based methods are simple but provide limited depth and coarse orientation information, whereas stereo vision, RGB-D sensing, and point clouds enable direct 3D pose reconstruction. More recent approaches further integrate structural constraints and multi-view observations to improve robustness under occlusion and incomplete visual information. Overall, fruit pose estimation is progressing toward structure-aware, operation-oriented, and active closed-loop perception.

4.2.1. RGB image-based pose estimation

RGB-image-based methods primarily infer fruit poses from external morphological and structural features extracted from two-dimensional images, typically without explicitly reconstructing a three-dimensional model of the target. Depending on the pose representation strategy, existing studies can be broadly categorized into overall-shape-based orientation estimation, discrete pose representation based on pose categories, and pose modeling based on key structural features.

Methods based on overall shape primarily use fruit contours, principal axes, or rotated bounding boxes to characterize the orientation of the target in the image plane. For fruits with distinct longitudinal axes, such as strawberries, a two-dimensional principal axis can provide intuitive information about the picking direction. Yu et al. [148] used rotated bounding boxes to simultaneously determine the position and longitudinal axis of strawberries and further calculated the picking point along the principal axis; Xiao et al. [149], on the other hand, combined semantic segmentation with rotated object detection to determine the harvesting pose of citrus fruits based on fruit contours and the orientation information of the grasping region. Compared with conventional horizontal bounding boxes, rotated bounding boxes and contour geometry can incorporate directional information into object detection without requiring additional depth sensors, thereby reducing computational complexity. However, such methods primarily reflect the projected orientation of the target in the image plane, and their effectiveness depends on the integrity of the fruit contour and its morphological features. When the fruit is nearly spherical, severely occluded, or observed from substantially different viewing angles, a noticeable deviation may occur between the 2D principal axis and the

Table 2
Development and characteristics of fruit pose estimation methods.

Method stage	Input information	Representative output	Reference	Advantages	Limitations	Development trend
2D geometric estimation	RGB, contours, rotated boxes	Rotation vector / principal axis	[142]	Simple	No depth; sensitive to occlusion	2D orientation perception
Pose classification	RGB images	Discrete pose category	[143]	Simple annotation; low computation	Coarse pose information	Task-oriented representation
Keypoint-based estimation	RGB, structural keypoints	Structural keypoints / 2D orientation	[144]	Clear structural information	Keypoints are easily occluded	Biological structure modeling
3D geometric estimation	Stereo, RGB-D, point clouds	3D position and orientation / 3D pose	[145]	Provides real 3D pose	High computation; depends on point-cloud quality	3D pose reconstruction
Local 3D estimation	RGB-D, local point clouds	Local 3D position / structural orientation	[58]	Balances accuracy and efficiency	Depends on detection and depth quality	Local structural perception
Structure-constrained 3D estimation	Keypoints, depth information	Structure-constrained 3D pose	[146]	Closely related to harvesting operations	Performance degrades with missing structures	Operation-oriented pose
Learning–geometry fusion	RGB-D, depth features	Refined 3D pose / pose adjustment	[147]	Combines semantic and geometric information	Higher model complexity	Multi-information fusion
Active vision	Multi-view, eye-in-hand vision	Refined pose from updated viewpoints	[24]	Improves perception under severe occlusion	More complex motion and control	Active closed-loop perception

actual growth direction.

Another approach discretizes the continuous pose space into several operationally meaningful categories according to the pose accuracy requirements of the picking task. Zhang et al. [150] classified fruits into pose categories such as Front-Right, Front-Middle, Front-Left, and Back based on the spatial distribution of the tomato longitudinal axis and used the classification results to guide the end-effector approach. Compared with continuous angle regression, pose classification reduces the complexity of data annotation and model learning and is well suited to grasping tasks that allow a certain degree of pose tolerance. However, discrete pose categories can provide only coarse-grained orientation information. When the end-effector needs to approach precisely along the fruit stem direction or perform directional cutting, a limited number of pose categories makes it difficult to fully characterize the actual spatial orientation of the fruit.

With advances in deep-learning-based key point detection, the representation of fruit orientation has gradually shifted from overall shape to local biological structures and their topological relationships. Such methods typically define the fruit center, pedicel attachment point, sepals, abscission layer, fruit navel, or branching points as key points and determine the fruit orientation based on the connections or spatial relationships among these key points. Kim et al. [151] represented the tomato fruit support system using key points such as the fruit center, sepals, abscission layer, and branching points and employed Part Affinity Fields to establish associations among these key points, thereby reconstructing the fruit axis and pedicel structure. Liu et al. [152], on the other hand, used multiple keypoints to characterize the main body of the strawberry and the pedicel attachment region, directly extracting structural information closely related to picking operations within a lightweight network. Compared with dense segmentation of the entire object, keypoint-based methods concentrate computational resources on local regions that determine fruit pose and picking location, thereby establishing a more direct link between visual perception outputs and robotic operations.

Expanding upon this methodology, structural keypoints of specific biological relevance are furthermore used to establish fruit orientation axes [144]. The pedicel and calyx extremities of the dragon fruit are identified as essential keypoints, whilst the yaw and roll angles are calculated using the fruit's elliptical projection and the spatial relationships of these keypoints. In contrast to random principal axes obtained from fruit outlines, structural connections like “pedicel end–calyx end” and “fruit center–sepals” more accurately represent the fruit's natural growth orientation, thus transitioning the representation of orientation from solely image-based geometric attributes to task-specific characteristics linked to the fruit's biological structure and harvesting processes. This organized depiction allows for the use of geometric restrictions among several keypoints to mitigate the effects of absent data during partial occlusion.

In general, RGB pictures may provide fruit orientation data with comparatively little sensing and computing expenses. Among these techniques, rotated bounding boxes and contour-based strategies are effective for swift assessment of two-dimensional orientation; pose classification is relevant for tasks that can accommodate a greater degree of pose inaccuracies; and keypoint-based approaches along with biological structure modeling can yield more detailed and practically significant pose data. Nonetheless, each RGB picture does not provide clear depth information; hence, spatial pose reconstruction depends on projection geometry, previous knowledge of object morphology, or data-driven methodologies. In the meantime, local formations like fruit stalks, sepals, and calyxes are prone to obstruction in their native habitats. Consequently, whereas RGB-image-based techniques provide notable benefits for estimating two-dimensional orientation, they possess inherent limits in harvesting applications that need correct three-dimensional spatial orientation and perfect alignment of the end effector.

4.2.2. *Depth information-based pose estimation*

Depth-information-based methods utilize stereo vision, RGB-D cameras, or 3D point clouds to acquire the spatial geometric information of the target, thereby extending pose representation from two-dimensional projected orientations to three-dimensional positional and rotational relationships at true scale. Compared with RGB-image-based methods, these approaches can directly establish three-dimensional spatial relationships among the fruit, peduncle, and harvesting mechanism, making them more suitable for precise grasping and cutting tasks. Based on the manner in which three-dimensional pose information is acquired, existing methods can be broadly categorized into global-geometric-model-based pose reconstruction, 2D-vision-guided local 3D analysis, and key-structure-based 3D pose estimation.

Early research on 3D vision primarily focused on the spatial localization of targets and picking points. Stereo vision recovers target depth through feature matching and triangulation, whereas multi-view and RGB-D vision can provide more complete three-dimensional surface information. Xiong et al. [153] employed stereo vision to achieve spatial localization of dynamic lychee clusters and picking points, while Gené-Mola et al. [154] projected 2D fruit detection results onto reconstructed 3D point clouds to determine the three-dimensional positions of apples. Although these methods primarily address the question of “where” a target is located, the reconstruction of three-dimensional position and structural information provides a foundation for further estimating the fruit principal axis, pedicel orientation, and spatial orientation from multiple 3D points. Existing studies also indicate that stereo and RGB-D vision can directly provide depth information and three-dimensional point clouds of agricultural targets, making them commonly used sensing modalities for 3D perception in harvesting robots.

In addition to the direct use of RGB-D information, multi-view vision, stereo matching, and point cloud semantic analysis also provide a three-dimensional structural basis for fruit pose estimation. Chen et al. [10] improved the completeness of 3D information acquired in complex agricultural environments through adaptive stereo matching and multi-view point cloud registration; Zheng et al. [155] combined instance segmentation with stereo vision, using object masks to constrain disparity estimation and point cloud regions to determine the three-dimensional positions of tomatoes; Hou et al. [58] further performed stereo matching and three-dimensional localization of picking points based on the spatial relationship between citrus fruits and their stems. Meanwhile, Luo et al. [104] employed PointResNet to perform semantic segmentation of 3D point clouds containing targets such as grapes, fruit stalks, and leaves, providing a more structured three-dimensional representation for subsequent fruit stalk orientation extraction, picking point determination, and pose estimation. Although some of the aforementioned studies primarily focus on three-dimensional localization or point cloud analysis, collectively, they enhance the spatial observability of harvesting-related structures, thereby laying the foundation for extending visual perception from target localization to spatial pose estimation.

After acquiring the target point cloud, its spatial orientation can be directly estimated from the global geometric features of the fruit. Such methods typically approximate the fruit using geometric models with regular shape characteristics and determine its spatial orientation based on symmetry axes, principal axes, surface models, or template models. Li et al. [145] exploited the approximate symmetry of the 3D point cloud of sweet peppers to extract the fruit symmetry axis, thereby enabling target orientation estimation even when the peduncle was not visible. Guo et al. [156] employed sample consensus initial alignment and iterative closest point registration to register the observed point cloud with a pre-established target template and determined the spatial position and orientation of the fruit or vegetable by estimating the rigid-body transformation. Compared with two-dimensional methods, geometric modeling can directly recover three-dimensional poses and provides better physical interpretability; however, its performance depends on

the validity of the target geometric model and the completeness of the point cloud. Particularly under severe occlusion, missing local point cloud data may lead to principal-axis estimation errors or registration failure, while global point cloud registration typically incurs substantial computational overhead.

To reduce the computational complexity associated with processing complete point clouds, 2D object detection and segmentation have increasingly been used to guide the extraction of local 3D information. These methods first identify fruits, stems, or regions of interest in RGB images, then generate local point clouds from the corresponding depth data, and finally estimate the target pose through geometric fitting. Luo et al. [157] mapped grape clusters and stems obtained through instance segmentation into 3D space and determined stem orientation at the cutting points through local curve fitting. Du et al. [158] used tomato detection results and sepal keypoints to define the point cloud processing region, estimated the 3D fruit center through spherical fitting, and subsequently represented the fruit pose using the spatial vector between the fruit center and the sepal center. Compared with direct processing of the full-scene point cloud, this approach reduces interference from background and foliage points while concentrating computational resources on harvesting-relevant target regions, thereby achieving a favorable balance between semantic recognition capability and 3D geometric accuracy. With advances in 3D semantic perception, structures such as fruits, stems, and leaves within point clouds can be further differentiated. Methods represented by PointResNet perform 3D semantic segmentation of grapes, stems, and leaves by integrating the spatial coordinates, color information, and local geometric features of point clouds. Although these methods do not directly output complete poses, they can provide more structured 3D scene information, establishing a reliable data foundation for subsequent tasks such as stem fitting, picking point determination, and collision-region identification. This indicates that the role of 3D vision is gradually expanding from simply acquiring object depth to analyzing the internal structures of harvesting targets and their spatial relationships with the surrounding environment.

In recent years, increasing attention has been devoted to establishing 3D pose constraints using key structures with explicit biological significance. The fundamental idea is to avoid relying on the entire fruit surface and instead characterize the fruit growth orientation using spatial vectors between the fruit center and key structures, such as the calyx, fruit navel, or pedicel attachment point. Niu et al. [159] separately determined the 3D centers of the apple body and calyx cavity and used the spatial vector between them to characterize the apple orientation. Zhuang et al. [160] determined the fruit growth orientation and grasping position based on the spatial relationship between the tomato center and calyx center. For fruits with more complex morphologies, such as strawberries, Zhu et al. [161] further integrated multiple structural features, including the rotated bounding box, fruit tip, left and right boundaries, and fruit stalk, and combined them with RGB-D point cloud information to achieve 5-DOF pose estimation. Compared with methods relying solely on the principal axes of the overall point cloud, the use of multiple structural features introduces additional geometric constraints, offering greater potential for pose recovery when local point cloud data are missing or key structures are partially occluded.

In addition, the integration of 2D keypoints with local depth information further improves the real-time performance of 3D pose estimation. Wang et al. [162] employed a lightweight keypoint detection network to extract structural features of strawberries and their stems and combined these features with depth images and coordinate transformations to reconstruct 3D picking points and stem orientations. Rather than pursuing complete 3D reconstruction of the fruit and its surrounding scene, this approach focuses exclusively on recovering local spatial information that directly determines robotic motion, thereby reducing the computational burden associated with point cloud processing and making it more suitable for resource-constrained real-time harvesting platforms.

There is also a growing trend toward the integration of deep learning and geometric methods. Direct pose regression from RGB-D data can reduce the number of conventional geometric processing steps, but its performance is influenced by the training data and the model's generalization capability across unseen scenes. In contrast, geometric refinement using point cloud information after deep learning provides an initial pose estimate can further improve spatial consistency. Alterani et al. [163] compared robotic grasping performance between direct RGB-D-based pose estimation and approaches combining deep learning with geometric optimization; the results demonstrated that geometric refinement improved grasping stability across different fruit sizes and poses. To address incomplete fruit contours under occlusion, Chen et al. [164] further integrated the global fruit orientation obtained through rotated object detection with local depth-gradient information. The fruit stem orientation was determined based on both the two-dimensional contour and depth variations, and the resulting visual information was converted into the pose adjustments required by the robotic end-effector. Compared with complete point cloud modeling and registration, this method primarily extracts local depth features directly relevant to pose estimation, thereby reducing the computational requirements for 3D data processing while enhancing the representation of spatial orientation under occlusion.

In complex environments with severe occlusion, even depth-based vision from a single fixed viewpoint may fail to capture the complete fruit stalk or harvesting region. Therefore, active vision has increasingly been employed to improve the observability of pose-related information. Dong et al. [165] calculated the optimal camera observation pose based on the three-dimensional spatial relationship between the tomato cluster and the main stem, thereby guiding the eye-in-hand vision system to adjust its viewing direction before performing precise local perception. Unlike conventional methods that attempt to recover the target pose from incomplete single-frame observations, active vision reduces information loss caused by occlusion by actively modifying the observation conditions, providing a new approach to robust pose perception in complex natural environments.

Overall, depth-based fruit pose estimation can provide harvesting robots with more comprehensive three-dimensional position and orientation information, offering significant advantages for tasks such as precise grasping, stem cutting, and collision avoidance. However, such methods are highly sensitive to the quality of depth data. Complex illumination conditions, surface reflections, occlusions, and variations in sensing distance can introduce depth noise and local point cloud incompleteness. In addition, point cloud processing and three-dimensional geometric computation increase the computational burden of the system. Therefore, improving robustness, real-time performance, and computational efficiency in complex environments while maintaining spatial pose accuracy remains a key challenge for fruit pose estimation in practical harvesting applications. It should be noted that the position and orientation of the fruit alone cannot directly determine the final operational pose of the end-effector. After perceiving the spatial state of the target, the specific picking point must be further determined by considering the fruit–peduncle structure, harvesting strategy, and operational constraints of the end-effector. This represents a critical step in advancing the visual perception of harvesting robots from "pose understanding" to "operational decision-making."

4.3. Picking point detection and localization

The aforementioned fruit pose estimation methods primarily determine where the fruit and harvesting-related structures are located and how they are oriented. However, robotic harvesting further requires these perception results to be converted into actionable picking targets. As illustrated in Fig. 12, existing picking-point localization methods can be broadly divided into four categories: (A) geometry-based inference, in which picking points are derived from fruit contours, centers, axes, or other geometric priors; (B) attachment-structure-segmentation-based

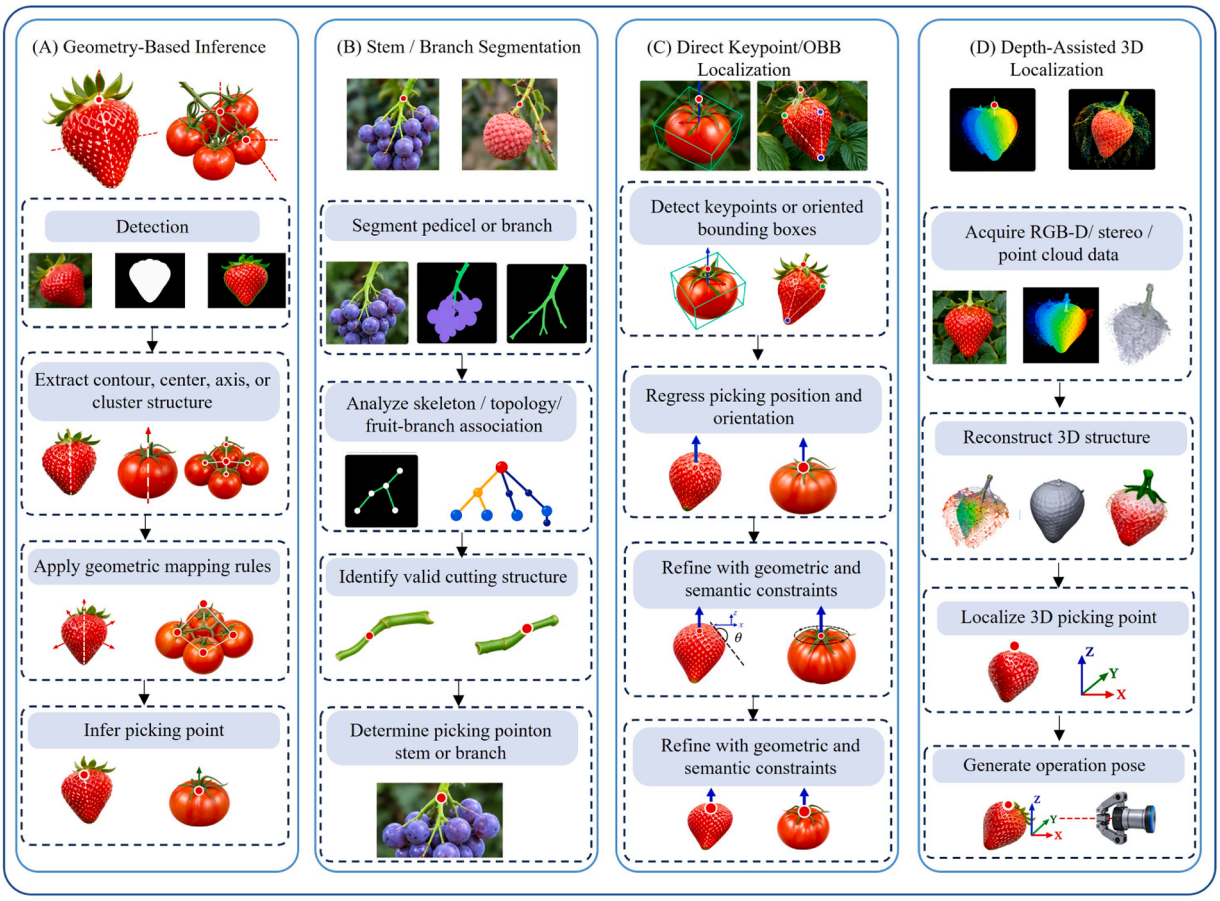


Fig. 12. Representative workflows for picking-point detection and localization in fruit harvesting robots: (A) geometry-based inference, (B) stem/branch segmentation-based localization, (C) direct keypoint/OBB-based localization, and (D) depth-assisted 3D localization.

localization, which identifies valid cutting structures through attachment-structure extraction and structural analysis;(C) direct keypoint- or oriented-bounding-box-based localization, which directly regresses the picking position and orientation from images; and (D) depth-assisted 3D localization, which combines RGB-D, stereo vision, or point-cloud information to reconstruct the target structure and determine the 3D picking point and operation pose. For suction- or gripping-based harvesting, the operational target is often defined on the fruit surface or near its geometric center, whereas for crops requiring cutting, such as grapes, lychees, tomatoes, and strawberries, the picking point is typically associated with the fruit attachment structure or fruit-attachment junction. Therefore, picking-point localization serves as the key link between pose perception and the generation of executable robotic manipulation targets.

4.3.1. *Geometry-based picking point inference*

Methods based on geometric priors do not directly detect the picking point. Instead, they first identify the fruit or fruit cluster and subsequently establish a geometric mapping relationship between the picking point and the fruit based on the target contour, principal axis, center position, and growth morphology. These methods are based on the fundamental assumption that a relatively stable spatial relationship exists between the external morphology of the fruit and its actual attachment structure. Therefore, even when the attachment structure is not fully observable, the picking position can still be inferred from the fruit itself. In early research, Yu et al. [166] employed Mask R-CNN to obtain instance masks of strawberries, constructed an approximate axis of symmetry from the fruit contour, and extended it toward the pedicel to determine the picking location; the average localization error for 573 picking points on ripe strawberries was $\pm$1.2 mm. The key advancement of this method lies not in modifying the geometric localization rules, but in replacing conventional color-based segmentation, which is susceptible to illumination variations, with instance segmentation, thereby enabling geometric inference to be based on more robust fruit contours. For tomato clusters with distinct growth patterns, picking points can also be indirectly reconstructed from the structural relationships among multiple fruits. Bai et al. [167] addressed the characteristic fishbone-like arrangement of tomato clusters by first identifying fruits through the fusion of color, texture, and shape features. They then employed the Hough circle transform to determine the center and radius of each fruit, followed by cubic spline interpolation to reconstruct the cluster contour and branch orientation, and finally determined the branch and picking positions based on the spatial symmetry between the two sides of the cluster. Compared with extrapolation based on the contour of a single fruit, this method effectively exploits the structural constraints imposed by the collective arrangement of fruits on occluded branches. Thus, even when some branches lack distinct visual features, their approximate locations can still be inferred from the spatial distribution of the fruits.

These methods offer several advantages, including a clear computational procedure, no requirement for additional picking-point annotations, and ease of integration with agronomic rules. However, their reliability fundamentally depends on the stability of the mapping relationship between “visual morphology” and the “actual picking structure.” Variations in fruit orientation, asymmetric growth of fruit clusters, and occlusion by branches and leaves can alter the two-dimensional projected structure, resulting in systematic errors when fixed geometric rules are applied. Therefore, as visual models have become increasingly capable of perceiving slender fruit attachment structures, the research focus has gradually shifted from “estimating picking points from the fruit” to “first perceiving the actual harvesting

structure and then determining the picking point within that structure."

4.3.2. Fruit attachment-structure segmentation and structure-based localization

Methods based on attachment-structure segmentation first explicitly extract the pedicel, peduncle, or main fruit-bearing branch and subsequently convert these pixel-level regions into manipulation targets using skeletons, endpoints, centers of mass, or predefined cross-sections. This technical approach enables picking points to be established directly on the actual attachment structure, rather than relying entirely on prior knowledge of fruit shape. Qi et al. [168] established a typical "object detection–semantic segmentation–skeleton extraction–geometric localization" workflow. YOLOv5 was used to detect the RoI of the main lychee stem, followed by the Pyramid Scene Parsing Network (PSPNet) to segment the main stem at the pixel level. After image post-processing, the stem skeleton was extracted, and candidate picking positions were determined from the intersections between the skeleton and cross-sections at different heights, ultimately achieving a precision of 92.50% and a recall of 76.29%. In comparison, Chen et al. [169] addressed the challenges posed by the slender morphology, substantial pose variations, and color similarity of prune stems to surrounding leaves and branches by incorporating MobileNetV2, CBAM, and dense atrous spatial pyramid pooling into DeepLabv3+ to improve the completeness of stem segmentation. By directly defining the endpoints of the refined skeleton as picking points, the localization strategy shifted from manually predefined fixed cross-sections to locations determined by the intrinsic topological structure of the stem, achieving a localization success rate of 88.8%. As application scenarios expanded from individual fruits to fruit clusters, obtaining only the fruit stalk mask became insufficient for accurate localization because visual scenes often contain multiple fruit stalks, branches, and main stems simultaneously. Consequently, structural determination methods began to incorporate target association. Zhang et al. [170] integrated mango and fruit stalk detection with fruit stalk semantic segmentation into the YOLOMS multi-task network. After obtaining multiple candidate fruit stalks, the method further exploited the spatial relationship between the mangoes and pedicels to identify the main pedicel associated with the entire fruit cluster. The picking point was then determined through morphological processing and skeleton analysis, transforming "picking point localization" from a purely local geometric problem into one of "first identifying the correct operational structure and then localizing the picking point within that structure." Ultimately, a picking point localization success rate of 89.84% was achieved. For targets with limited color contrast between the fruit and pedicel, such associative constraints become even more critical. Li et al. [171] addressed challenges including the small size of kiwifruit pedicels on trellised plants, limited color contrast between fruits and pedicels, and occlusion caused by clustered fruits by using DS-UNet to segment fruits and pedicels. The association between each fruit and its corresponding pedicel was then determined based on the spatial relationship between the downward extension of the pedicel and the fruit region. Ultimately, the center of mass of the minimum bounding rectangle of the target pedicel was defined as the valid picking point, achieving a prediction probability of 91.65%. In PLYS-Longan, Liao et al. [172] used object detection to identify longan clusters and semantic segmentation to extract the main fruiting branches. Morphological erosion was subsequently applied to the branch masks, and the center of mass of the resulting stable region was calculated as the picking location. In this case, morphological erosion serves not merely as an image post-processing operation; by removing irregular boundaries and fine branches from the mask, it helps ensure that the picking point is located within the main body of the branch, where greater operational tolerance is available.

Consequently, segmentation-based picking point localization has gradually evolved from the early paradigm of "fruit stalk segmentation–centerline extraction" toward a more comprehensive workflow of "object recognition–fruit–branch association–valid branch and stalk selection–safe region determination." The primary advantage of this approach is that the resulting picking points have a clear structural basis, facilitating the incorporation of prior constraints such as the required stem retention length, branch boundaries, and end-effector dimensions. However, such multi-stage processing also introduces significant cascading dependencies. Once a slender fruit stalk is missed during segmentation, its continuity is disrupted by occlusion, or it is incorrectly associated with a neighboring fruit, subsequent skeleton extraction and center-of-mass calculations, even if performed correctly, cannot yield a valid picking point. This issue is particularly pronounced under severe occlusion. Consequently, research has gradually shifted from simply improving segmentation accuracy toward reasoning over incomplete semantic information. Lin et al. [173] addressed scenarios involving short grape stems, partial occlusion, and stem misidentification by no longer assuming that segmentation results are always reliable. Instead, building on the scene-level semantic information extracted by SegFormer, they developed a Semantic Reasoning Module, an RoI threshold adjustment strategy, and a Picking Point Optimization Module to infer and refine candidate picking locations under different anomalous conditions, ultimately achieving a localization success rate of 94.96%.

4.3.3. Direct keypoint regression and oriented representation

Segmentation-based methods offer strong structural interpretability, but the complete workflow typically involves object detection, segmentation, morphological processing, and geometric computation. To reduce the computational overhead and error propagation associated with intermediate representations, recent studies have increasingly defined picking points themselves as key features to be learned by the network, enabling the model to directly establish a mapping between image features and picking locations. Zhang et al. [7] combined YOLO-based object detection with point regression to develop EYOLO-R, which detects grape stems while simultaneously regressing three candidate points: the midpoint represents the optimal picking position, while the points on either side define the safe cutting region. The model further enhances feature representation for slender stems and complex occlusion through space-to-depth convolution and a C3-Transformer. It achieved an AP@0.5 of 88.7% and an average Euclidean localization error of 24.512 pixels, demonstrating that the picking point can transition from a "derivative of segmentation results" to an output variable directly optimized by the network. Pose-based frameworks further advance this end-to-end localization paradigm. Su et al. [174] directly represented zucchini picking locations as keypoints in YOLOv8n-Pose, enhanced high-resolution feature recovery using DySample, and reduced the computational burden through Slim-Neck. The model achieved a picking-point mAP@50 of 95.6%, with localization errors primarily concentrated within 0–4 pixels. Huang et al. [175] addressed the challenge that eggplant picking points correspond to small, millimeter-scale regions whose features can be easily overwhelmed by background information during multi-scale feature fusion. They incorporated region-, channel-, and spatial-level multidimensional attention mechanisms, together with dynamic upsampling, into YOLOv10n-Pose, enabling the network to focus more effectively on local key regions such as the upper part of the fruit stalk. The model achieved a keypoint mAP@50 of 94.7%, with an average Euclidean localization error of 19.41 pixels. These studies indicate that research on direct keypoint-based methods has shifted its focus from "whether a picking point can be regressed" toward how to preserve harvesting-relevant local features under conditions of small target size, occlusion, and complex backgrounds. However, a single keypoint represents only positional information and cannot explicitly characterize the orientation of the attachment structure, whereas cutting operations are often constrained by both position and orientation. Therefore, OBBs have emerged as another structured representation that bridges the gap between dense segmentation and single-point regression. Li et al. [176] employed YOLO11-GS to perform rotated detection of grape stems,

directly determining the picking point from the OBB center while using the rotation angle to estimate stem orientation and further calculate the cutting pose. In this way, information regarding both "where to cut" and "in which direction to cut" can be derived from a single detection result. When combined with ByteTrack and BoT-SORT, the system can also continuously track picking points under vibration, with picking-point localization errors primarily within 20 pixels and an average pose error of 5.08°. The additional value of OBBs lies in their ability to explicitly characterize the directional relationship between the fruit and the peduncle. Qing et al. [177] simultaneously performed rotated object detection on both strawberry fruits and peduncles and established fruit–peduncle correspondences based on the position, orientation, and projection relationships of their OBBs. When the peduncle was visible, the cutting point was determined by shifting 18 mm upward along the peduncle centerline from its lower end; when the peduncle was missed, the picking point was compensated for by shifting 20 mm upward along the fruit principal axis from its upper end. Compared with approaches in which the picking point can be determined only when the peduncle is successfully detected, this dual-strategy approach enables the vision system to retain a certain degree of picking-point recovery capability even when local structures are missing, achieving a fruit–peduncle association success rate of 97.99%.

Consequently, end-to-end keypoint-based and OBB-based methods represent two distinct strategies for simplifying picking point localization. The former eliminates intermediate geometric processing through direct learning, offering the advantages of a shorter processing pipeline and improved real-time performance; the latter retains explicit structural information, such as the position, scale, and orientation of the attachment structure, facilitating integration with the cutting constraints of the end-effector. A common limitation of both approaches is their continued strong dependence on the observability of key structures within a single frame. When the attachment structure is completely occluded, relying solely on local image features makes it difficult to ensure that the predicted results are consistent with the actual fruit–-branch topology.

4.3.4. *Depth-assisted and 3D picking point localization*

Picking points identified in 2D images ultimately need to be transformed into 3D coordinates within the robotic manipulator workspace. Therefore, picking point localization exhibits a similar expansion in perceptual dimensionality to the pose estimation discussed in Section 4.2, namely, a gradual transition from 2D image features to spatial information acquired through stereo vision, RGB-D sensing, and point clouds. Stereo vision provides an early approach to spatial reconstruction through stereo correspondence and triangulation. Li et al. [178] employed Faster R-CNN to detect targets in stereo images of apples, subsequently refined the target regions based on color differences, and established feature correspondences between the left and right images using template matching constrained by epipolar geometry. Finally, triangulation was employed to recover the 3D coordinates of fruit feature points, achieving an average depth localization accuracy of 99.64%. For small attachment structures that are difficult to reliably identify using vision alone, active sensing can also provide near-field compensation for 2D visual perception. First, visual segmentation is used to determine the bounding region of the grape cluster. Multiple sets of infrared transmitter–receiver pairs are then mounted on the end-effector, and the upper region of the cluster is scanned through mechanical motion. The stem picking position is subsequently determined based on infrared beam interruption. Compared with methods that rely entirely on RGB images for stem recognition, this approach decomposes the task into "visual coarse localization + close-range active sensing," thereby reducing the influence of stem color, texture, and complex backgrounds, although it also introduces additional sensing hardware and search motions. RGB-D vision enables a more direct transformation of 2D keypoints into 3D space; however, slender stems are prone to missing data in depth maps, making it difficult to obtain reliable depth information by simply querying a single pixel. Song et al. [179] employed an improved YOLOv8 model to detect the stems of vine tomatoes and combined K-means++, morphological erosion, and Zhang thinning to extract a 2D skeleton and determine the picking points. To address missing depth values for short and slender fruit stalks, a secondary depth extraction method was developed to recover valid depth values and 3D coordinates, achieving a picking point identification success rate of 91.3%. Liang et al. [180] employed the CPPL framework, in which KD-YOLOP was first used for rapid citrus fruit detection and picking-region definition, followed by RG-YOLO-seg for branch segmentation and picking point determination based on branch structure. Unlike methods that directly query depth at the fruit center, this approach first identifies the branch region on which the cutting operation is actually performed and then conducts spatial mapping, thereby providing the resulting 3D coordinates with clearer operational semantics. Multimodal fusion further integrates depth information from the localization backend into the visual perception frontend. Li et al. [181] employed an RGB-D sensor to acquire images of cherry tomato clusters and, after fusing color and depth information, fed the fused data into a lightweight StarBL-YOLO model to simultaneously identify harvestable clusters, unharvestable clusters, and key picking points. The key innovation lies not merely in the incorporation of depth information, but in the integration of harvestability assessment with picking point generation. For clusters lacking reliable picking structures due to occlusion or other factors, picking points are no longer generated by default, thereby preventing the vision system from incorrectly equating "detectable fruit" with "safely harvestable fruit." As keypoint-based methods are increasingly extended to real-world robotic harvesting, two-dimensional accuracy metrics are gradually being complemented or replaced by three-dimensional physical localization errors. Huang et al. [182] proposed Pepper-Det, which detects chili peppers and their harvesting-related keypoints within a unified framework and subsequently uses RGB-D depth information and camera intrinsic parameters to back-project the 2D keypoints into 3D space. The method achieved a 2D Euclidean localization error of 11.13 pixels and a median 3D Euclidean localization error of 9.51 mm, with a 95% success rate in simulated harvesting experiments. The study by Qing et al. [177] on strawberries similarly demonstrated that, when 2D fruit–pedicel associations are combined with depth information, the mean absolute errors of the picking point along the x-, y-, and z-axes can reach 2.2, 2.6, and 2.3 mm, respectively. These results indicate that the evaluation criteria for picking point localization are gradually shifting from network-level metrics, such as precision, mAP, and pixel-level localization error, toward millimeter-level spatial localization error and final picking success rate, which are more directly relevant to end-effector operation. Compared with the "2D picking point + depth back-projection" approach, directly identifying the cutting region within a 3D point cloud enables further exploitation of local three-dimensional geometric structures. Wang et al. [183] employed an eye-in-hand RGB-D camera to acquire lychee point clouds from multiple robotic manipulator viewpoints. Multi-view point cloud fusion was performed through coordinate transformations to compensate for occlusion from individual viewpoints, after which an improved Fcaf3d-lychee model was utilized to directly detect 3D cutting regions on the main fruiting branches. The key advancement of this approach lies in transforming picking points from coordinates derived from 2D images into target regions directly perceived within 3D point clouds, thereby reducing errors associated with pixel-wise correspondence between 2D images and depth maps. The detection accuracy under unobstructed, partially occluded, and severely occluded conditions reached 0.932, 0.824, and 0.765, respectively, with localization errors along all three axes controlled within ±1.5 cm. Zhao et al. [184] employed YOLOv7-hv to simultaneously output cucumber bounding boxes, instance masks, and picking keypoints. Based on this multi-parameter visual information, the fruit growth orientation parameters and spatial position were calculated, after which the 6D operational pose required by the cutting tip was

generated through coordinate transformation, with average positional and angular errors of 6.18 mm and 6.25°, respectively. These results indicate that, for cutting-based harvesting, a single 3D point is insufficient to fully characterize the operational target. Instead, the representation of the picking target is evolving from a single spatial coordinate toward operational-level parameters that simultaneously incorporate position, peduncle orientation, and end-effector approach pose.

Overall, picking point detection and localization are evolving from indirect estimation based on fruit geometric priors toward attachment-structure analysis, direct keypoint regression, and 3D operational localization. Geometric and segmentation-based methods offer strong interpretability and can incorporate fruit–branch relationships and agronomic constraints, but they are susceptible to occlusion, structural incompleteness, and cascading errors. Keypoint- and OBB-based methods simplify the localization pipeline and can further characterize attachment-structure orientation, but they remain highly dependent on the observability of fine-scale picking structures. With advances in RGB-D sensing, stereo vision, and point cloud technologies, picking points have gradually evolved from two-dimensional pixel locations into three-dimensional operational information that incorporates both spatial position and orientation. However, severe occlusion, missing depth data, and target disturbances still limit their practical applicability, while two-dimensional detection metrics cannot fully reflect robotic operational performance. Therefore, future research should further integrate prior knowledge of fruit–attachment relationships, three-dimensional spatial information, temporal observations, and end-effector constraints to advance picking point localization from static, single-frame perception toward dynamic three-dimensional operational target estimation for real-world closed-loop harvesting.

4.4. *Frontier trends*

As summarized in Fig. 13, machine vision for fruit harvesting has progressively evolved from handcrafted image processing and shallow machine learning toward deep learning-based automatic feature extraction and, more recently, large-scale foundation models. The methods reviewed in the preceding sections demonstrate that deep learning has substantially improved fruit detection, segmentation, pose estimation, and picking-point localization. Nevertheless, most existing models remain task-specific and strongly dependent on annotated datasets collected for particular crops, environments, and perception tasks. Consequently, improving cross-crop generalization, reducing data dependence, and enabling more flexible semantic understanding are emerging as important directions for the next generation of harvesting-robot perception.

VLMs provide a promising pathway toward more generalizable agricultural perception by combining visual representations with large-scale semantic knowledge. Compared with conventional supervised models trained for predefined categories and tasks, VLMs offer stronger potential for zero-shot and few-shot transfer, thereby reducing the need to construct and annotate new datasets for every crop or orchard environment. Duan et al. [185], for example, applied Contrastive Language–Image Pretraining (CLIP) to quality grading across seven fruit varieties and demonstrated its potential under limited-data conditions. Huang et al. [186] further combined YOLO11 with a fine-tuned VLM to evaluate fruit characteristics from multiple dimensions, including quantity, ripeness, and surface condition. However, current VLMs still exhibit limitations in fine-grained visual discrimination, accurate counting in cluttered scenes, domain generalization, and computational efficiency. More importantly for robotic harvesting, semantic understanding must ultimately be converted into geometrically accurate and physically executable information, such as fruit position, pose, stem structure, picking point, and end-effector approach direction. Therefore, integrating foundation models with task-specific visual models and geometric perception may be more practical in the near term than directly replacing existing harvesting perception pipelines.

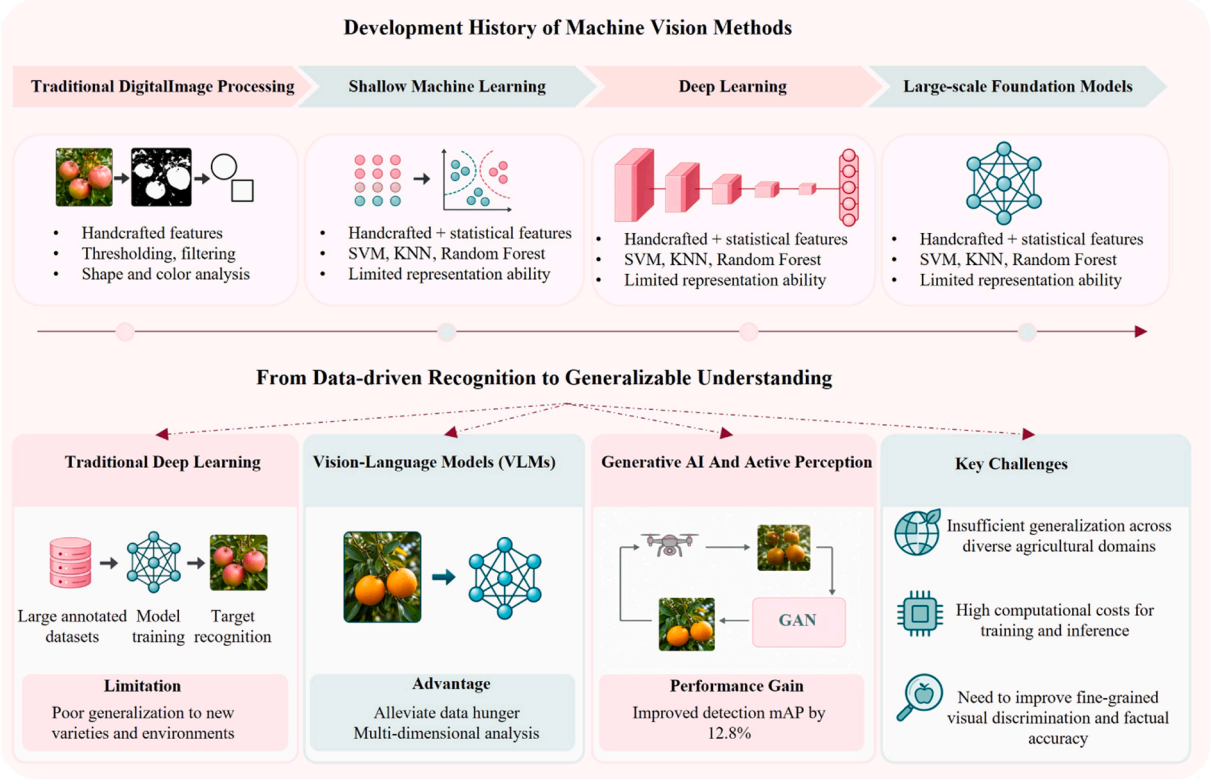


Fig. 13. Development of machine vision from traditional image processing to foundation-model-based perception in fruit-harvesting robots. Abbreviation: generative adversarial network (GAN).

Another important trend is the transition from passive single-view perception toward active and closed-loop perception. Most existing fruit perception methods estimate target states from images acquired at predetermined viewpoints, making their performance highly dependent on initial target visibility. In heavily occluded environments, however, robotic systems can actively adjust the camera or manipulator viewpoint according to perception uncertainty and reacquire more informative observations. Lacerda et al. [187], for example, combined active camera viewpoint control with generative image enhancement to improve agricultural target perception. Although such approaches remain at an early stage and are not necessarily designed specifically for fruit harvesting manipulators, they illustrate a broader transition from processing passively acquired images toward actively controlling sensing according to task requirements. For harvesting robots, this concept can be further integrated with eye-in-hand sensing, visual servoing, and manipulator motion to establish a closed perception–action loop in which the robot actively acquires additional information when the target, stem, or picking region cannot be reliably observed.

The forthcoming advancement of machine vision for fruit harvesting is expected to go beyond singular enhancements in detection precision toward a more universal, multimodal, dynamic, and task-focused perception. Foundation models could enhance semantic generalization and lessen reliance on task-specific annotations; multimodal sensing has the potential to integrate appearance, depth, and temporal data to improve spatial comprehension; and active perception allows robots to modify their observations based on target visibility and manipulation needs. Simultaneously, lightweight architectures and edge deployment are crucial for meeting the real-time demands of harvesting activities. Ultimately, these advancements are anticipated to evolve machine vision from a static target-recognition component into a dynamic perception framework that consistently delivers dependable and actionable data for robotic handling.

5. Vision-guided harvesting execution and end effectors

Following the coordinate transformation between the visual target's pixel coordinates and the robotic arm's coordinates, and the verification of the positional data of the target and obstacles, the robotic arm and end effector must execute the fruit-picking task. The execution system is essential for the effective gathering of fruits, since successful collection depends significantly on the synergy between the vision unit and the execution unit. Consequently, this chapter provides an analysis of the execution unit.

5.1. Vision-guided harvesting execution

The ultimate objective of machine vision in robotic harvesting is to convert environmental observations into operation-oriented information that can directly support fruit detachment. As discussed in the preceding sections, fruit detection and segmentation determine the target region, while three-dimensional localization, pose estimation, and picking-point detection further provide spatial position, orientation, and interaction-related information. However, obtaining these visual features does not itself guarantee successful harvesting. They must be transformed into executable constraints, such as the target position, approach direction, end-effector orientation, and allowable alignment error, and continuously associated with the motion state of the harvesting system. Vision-guided execution therefore constitutes the interface between visual perception and physical manipulation, through which machine vision has gradually evolved from a front-end target localization module into an integral component of the harvesting control loop.

Early vision-guided harvesting generally followed a relatively sequential perception–execution architecture, in which fruit position was first estimated and subsequently transformed into the robot coordinate system to guide target approach. This paradigm established the basic connection between fruit localization and robotic motion and enabled visual information to participate directly in end-effector positioning rather than being used solely for target recognition. Visual servoing further strengthened this connection by introducing image feedback during the approach process, allowing deviations between the observed and desired target states to be used for motion correction [188]. Consequently, the role of vision began to shift from providing a fixed spatial target toward continuously constraining the relative relationship between the harvesting mechanism and the fruit.

Nevertheless, a sequential localization-and-execution strategy implicitly assumes that the target remains approximately stationary after localization, an assumption that is difficult to satisfy in unstructured orchards. Fruits and supporting branches may be displaced by wind, manipulator motion, branch contact, or end-effector interaction, while positioning and coordinate-transformation errors can further accumulate during target approach. Under these conditions, even an initially accurate picking position may become invalid before physical interaction occurs. Continuous visual feedback provides a means of repeatedly estimating the relative target state and correcting execution errors, thereby transforming fruit position from a static perception result into a dynamically updated operational variable. Robust visual-servo strategies have therefore been developed to compensate for unknown fruit motion [189]. These developments indicate that vision-guided harvesting is concerned not only with where the robot should operate, but increasingly with how accurately and how rapidly the required operational state can be reached under dynamic disturbances.

The incorporation of depth sensing and changing observation viewpoints has further expanded visual guidance from image-plane correction toward spatially informed perception–action coordination. Combining RGB observations with depth information allows target displacement to be represented directly in three-dimensional space, thereby providing more explicit geometric constraints for end-effector approach [190]. Meanwhile, the information required by the harvesting system changes as the robot approaches the target. Relatively distant observations are more suitable for target discovery and coarse localization, whereas close-range observations provide more detailed information for target refinement and final alignment. Accordingly, dynamic multi-distance and multi-view visual servoing has been introduced to reorganize visual observations according to different operational stages [191]. This progression reduces dependence on a single viewpoint and establishes a closer relationship between sensing strategy and robotic motion, which is particularly important in dense canopies where target visibility, depth quality, and occlusion conditions vary continuously during approach.

As visual perception provides increasingly rich spatial information, the objective of vision-guided execution has also expanded from positional convergence toward manipulation-oriented pose alignment. For harvesting mechanisms that must grasp, twist, or cut a specific fruit structure, reaching the fruit center alone is insufficient; the end-effector must also satisfy constraints imposed by fruit orientation, peduncle direction, picking-point geometry, and local obstacle distribution. Position-based visual servoing (PBVS) can exploit reconstructed spatial information for large-range pose regulation, whereas image-based visual servoing (IBVS) directly uses image features and can reduce sensitivity to residual geometric errors near the target. Hybrid strategies that progressively switch between PBVS and IBVS therefore exploit their complementary characteristics during different stages of target approach [192]. From the perspective of machine vision, this development also increases the operational significance of the pose and picking-point information reviewed in Sections 4.2 and 4.3: these outputs are no longer merely geometric descriptions of the fruit but become constraints that determine how the end-effector should approach and interact with the target.

More recently, the perception–execution relationship has begun to extend beyond predefined geometric servoing toward action generation based on sequential visual observations. Instead of completing

perception before initiating execution, the harvesting system can repeatedly observe the target, generate an intermediate action, evaluate the resulting visual state, and then produce a corrective action. Learning-based action policies can be incorporated after coarse geometric positioning to compensate for residual localization and execution errors during the final approach [193]. This changes the conventional "perceive–plan–execute" pipeline into a recursive "perceive–act–reobserve–correct" process, in which perception and action continuously constrain each other. Such a framework is particularly suitable for harvesting environments because target motion, partial occlusion, model uncertainty, and fruit–robot interaction make it difficult for a single precomputed target pose to remain reliable throughout the entire operation.

The vision-guided harvesting implementation has progressed from static localization-centric guiding to dynamic, pose-sensitive, and closed-loop perception-action integration. The function of machine vision has progressively evolved from identifying an initial fruit location to perpetually refining operational objectives, regulating approach trajectory and end-effector orientation, and supplying input for execution adjustments. This development also demonstrates a progressively robust connection between visual perception and the physical harvesting technique: the visual data necessary for effective execution is contingent upon the manner in which the fruit is eventually gripped, severed, twisted, suctioned, or contained. Cutting-based mechanisms often demand more precision in peduncle localization and tool orientation, whereas multi-fingered, vacuum-assisted, and enveloping systems have varying tolerances for positional and orientational inaccuracies. Consequently, the efficacy of vision-guided execution must be assessed alongside the interaction attributes and operational demands of the end-effector, which will be examined in the subsequent section.

5.2. End effector

The end effector is the core component that directly contacts fruits, and its design must adapt to the physical properties (size, hardness, attachment methods) and growth environments of different fruits. For fruits with smooth surfaces and high damage resistance (e.g., apples, pears, tomatoes), fixed-contact and pulling-based harvesting methods can be adopted. For fragile fruits (e.g., strawberries, bananas, grapes), cutting-based harvesting is preferred, as it minimizes fruit contact and mitigates damage. Fruit-harvesting robots are evolving toward higher success rates, lower damage rates, stronger adaptability, and higher efficiency. Future research directions should focus on improving positioning accuracy in complex environments (e.g., dense occlusion, extreme lighting conditions), enhancing the adaptive capabilities of end effectors to diverse fruit morphologies, and reducing costs to enable large-scale applications.

5.2.1. Integrated clamping-cutting end effector

The integrated clamping-cutting end effector accomplishes harvesting through a "grasp-and-cut" workflow: for fruits with hard, extrusion-resistant peels (e.g., apples, citrus fruits, tomatoes), direct fruit clamping can be adopted; for cluster-grown fruits or those with damage-sensitive peels (e.g., strawberries, grapes), only the pedicel is clamped to avoid contact between the gripper fingers and the fruit surface, thereby preventing mechanical damage and pathogen invasion. As illustrated in Fig. 14A, Yin et al. [194] developed a cylinder-driven device that clamps mango pedicels, followed by scissor closure to achieve fruit detachment. However, pedicel cutting failure occurred primarily due to misalignment caused by inaccurate positioning and insufficient cutting force, requiring the pedicel to fall within the optimal harvesting region during

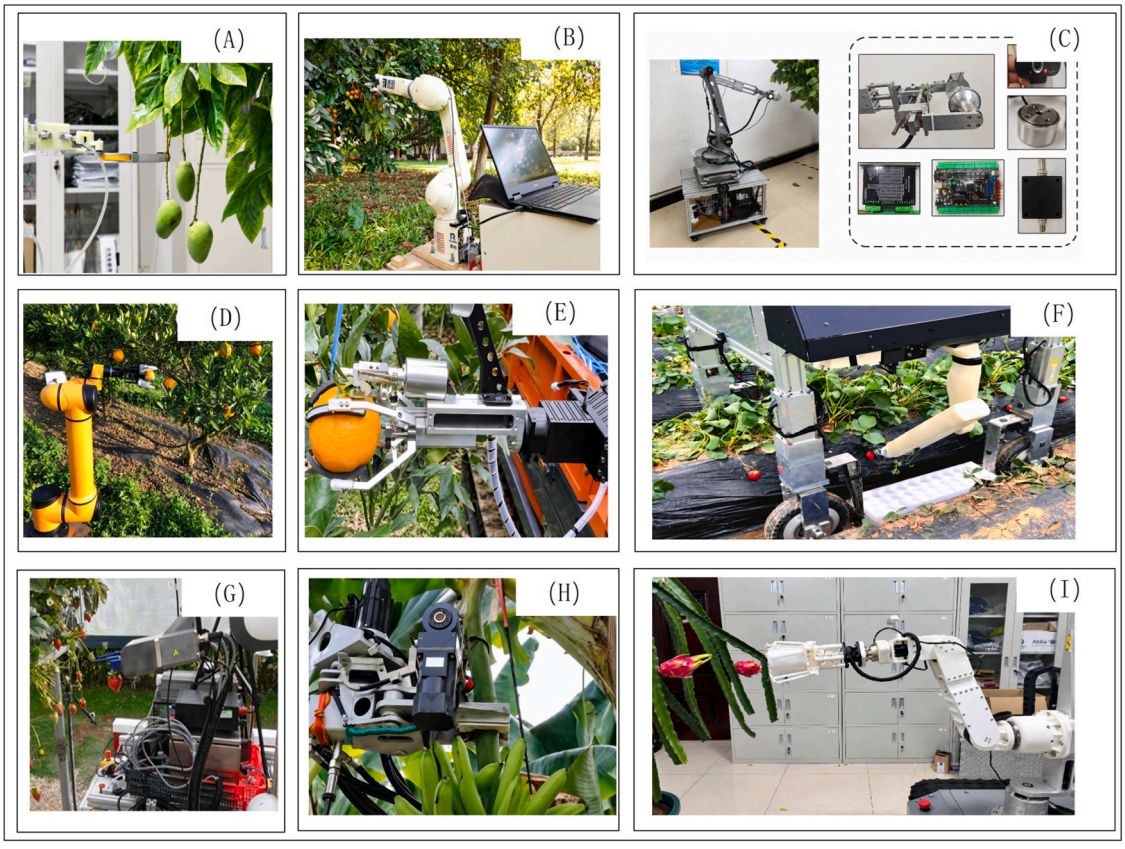


Fig. 14. Integrated gripper-cutter end effectors: (A) Mango gripping and cutting integrated picking end effector, (B) Lychee stem gripping end effector, (C) Apple end effector with spoon-shaped gripper and integrated cutting, (D) Citrus end effector with three-finger gripping and integrated cutting, (E) Citrus picking end effector with claw gripping and cutting, (F) Ridge-cultivated strawberry gripper-cutter end effector, (G) Detachable obstacle gripper integrated strawberry end effector, (H) Autonomous banana gripper-cutter integrated end effector, (I) Leaf damage-reducing gripper integrated dragon fruit end effector.

operation. Wang et al. [195] integrated a stem-clamping-and-cutting end-effector with a 6-DOF robotic arm for harvesting cluster-grown litchis. The end-effector was guided toward the fruit-bearing stem using close-range visual localization, while the depth camera was mounted above the end-effector to minimize visual occlusion. However, the harvesting operation remained sensitive to the recognition of fine bifurcate stems and depth acquisition errors, requiring sufficient end-effector tolerance to compensate for localization deviations (Fig. 14B). Zhao et al. [196] employed pneumatic grippers combined with a spoon-shaped structure to clamp and cut apples during approach. Leaf interference and fruit sway during harvesting led to pedicel offset, resulting in cutting failure—thus, vibration effects must be considered in the design (Fig. 14C). To prevent fruit offset during clamping, Xiao et al. [197] developed fin-inspired flexible fingers that adapt to citrus sizes and envelop the fruit surface. A blade was added to the end effector in the horizontal direction, driven by the robotic arm axis to rotate and complete cutting. While this design effectively prevented fruit offset, it required precise identification of fruit posture: the end effector must grasp perpendicular to the fruit's growth orientation to ensure successful cutting (Fig. 14D). Choi et al. [29] proposed a four-bar clamping mechanism paired with a cutting disc. Compared with traditional three-bar mechanisms, this design features a more simplified structure, reduced repositioning requirements, and higher harvesting speed. Its adaptive clamping capability also applies to round fruits such as apples, pears, and persimmons (Fig. 14E). Zhang et al. [198] developed an end effector for ridge-cultivated strawberries, which clamps and cuts from above the pedicel to adapt to ridge-shaped environments (Fig. 14F). Parsa et al. [199] designed a device with separating fingers at the top to move aside surrounding fruits and a clamping-cutting module at the bottom to grip and cut the pedicel. A short-range RGB camera was integrated to compensate for limited field of view. Harvesting point errors primarily originated from camera calibration, visual detection inaccuracies, and robotic arm motion errors (Fig. 14G). The banana-harvesting robot developed by Chen et al. [25] performs harvesting by clamping and cutting banana stalks; however, some stalks cannot be reliably identified under strong illumination. Successful harvesting relies on accurate localization of the optimal cutting point and stable clamping force, particularly for heavier banana bunches (Fig. 14H). Wang et al. [200] developed an end effector for pitayas growing close to leaf surfaces, which can envelop the fruit, clamp the pedicel, and complete cutting-rotation detachment to minimize leaf damage. However, the vision system failed to accurately determine the occlusion range, affecting harvesting success rate (Fig. 14I).

As can be inferred from Fig. 14 and the aforementioned descriptions, the integrated clamping-cutting end effector imposes extremely high requirements on recognition and positioning accuracy—particularly for pedicel-clamping configurations. It necessitates the simultaneous assurance of precise clamping of the pedicel (or fruit) and that the pedicel lies within the cutting range of the cutting mechanism while maintaining an appropriate posture. Once the cutting position deviates, issues such as failure to clamp the pedicel or incomplete cutting may arise. In real orchard environments, factors including occlusion, fruit sway, pedicel offset, and slippage significantly increase the failure probability.

5.2.2. *Bionic multi-fingered end effector*

Bionic multi-fingered end-effectors mimic human-hand grasping, combining multi-finger clamping with rotation, pulling, or cutting for fruit detachment. Three-fingered and flexible designs improve grasping stability and fruit conformity, such as silicone-coated grippers for enveloping pear grasping. Nevertheless, as shown in Fig. 15A, its real-orchard operation remains dependent on the collaborative control of the vision system and robotic arm [20]. As shown in Fig. 15B, the optimized flexible finger structure enhances grasping posture and downward-pulling actions, enabling reliable apple detachment [201]. Fig. 15C presents a 9-DOF bionic hand that excels in non-destructive harvesting: precise control of grasping force and contact positions improves tomato harvesting reliability, but imposes stringent demands on fruit centroid detection and force control systems [202]. For densely clustered tomatoes Fig. 15D, a rigid-flexible hybrid end effector reduces fruit interference and achieves a 1.9% damage rate through dynamic modeling and auxiliary support structures, demonstrating efficacy in clustered fruit scenarios [203]. As shown in Fig. 15E, Zhang et al. [204] developed a three-finger end-effector for cherry tomato harvesting, which provides a practical reference for the picking of small-sized spherical fruits. Huang et al. [205] designed an adaptive gripper

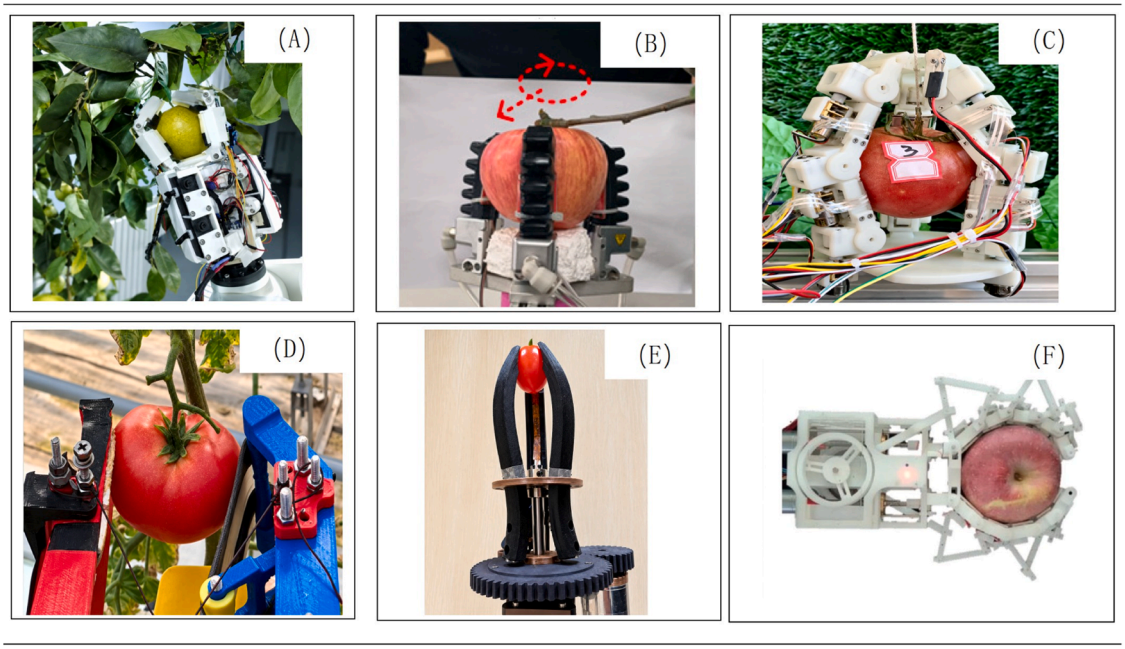


Fig. 15. Bionic Multi-Finger End Effectors: (A) Three-fingered end effector for picking pears, (B) Three-fingered flexible end effector for apple picking, (C) Three-fingered bionic flexible end effector for tomato picking, (D) Hybrid rigid-flexible bionic end effector for tomato picking, (E) Three-fingered compact end effector for cherry tomato picking, (F) Adaptive gripper end effector for apple picking.

featuring five articulated joints per finger, enabling each joint to establish contact with the apple surface. This end-effector can accommodate various spherical fruits and ensures a uniform distribution of grasping forces across the joints, thereby reducing the risk of fruit damage. However, the gripper requires a relatively large operational space during harvesting, which limits its applicability in cluttered environments. Therefore, more compact end-effector designs are needed for fruit harvesting under multi-obstacle conditions, as illustrated in Fig. 15F.

The bionic multi-fingered end effectors offer distinct advantages in flexibility, adaptability, and non-destructive grasping. However, their reliance on recognition accuracy, grasping force control, and environmental accessibility means their stability and versatility in natural orchard environments require further improvement.

5.2.3. *Vacuum-assisted end effector*

In modern agricultural automation, vacuum end effectors have been widely adopted in fruit-harvesting robots by virtue of their negative pressure adsorption capability. Compared with traditional clamping structures, vacuum adsorption reduces contact pressure on the fruit surface during grasping, thereby effectively minimizing mechanical damage. Meanwhile, the rapid response and continuous adsorption capacity of pneumatic actuation make it suitable for high-throughput harvesting operations and mitigate interference between the gripper and branches during rapid movement. As illustrated in Fig. 16A, Park et al. [206] proposed an integrated tomato-harvesting end effector incorporating cutting, adsorption, and transportation functions. It relies on a flexible suction cup to adhere to the fruit surface and achieve adsorption, followed by fruit detachment via a cutting module. The harvesting success rate reached 80.63%, with most failures attributed to obstructed traction caused by obstacles entering the cutting zone. Vacuum-based end-effectors (Fig. 16B) offer inherent compliance, reducing the risk of fruit damage [17]. From a visual control perspective, these effectors relax the stringent positioning accuracy requirements imposed by rigid grippers. However, successful adsorption relies heavily on the vision system's ability to detect the optimal planar surface on the fruit (e.g., the fruit equator), which remains challenging under varying fruit poses [207], as shown in Fig. 16C. Equipped with a 33 mm suction cup, a rotation-pulling detachment mode, and a −65 kPa negative pressure, it effectively reduced the damage rate. Nevertheless, since the optimal adsorption point for apples is located at the center of the equatorial circle, adsorption stability is highly dependent on the precise positioning of keypoints and fruit posture via image recognition. Gao et al. [14] conducted multi-mode tests (twisting, pulling, pull-rotation, breaking) on vacuum adsorption and rotary end effectors (Fig. 16D). Results indicated that the vacuum adsorber achieved a harvesting cycle of only 5.3 s, but its success rate was lower than that of the rotary mechanism, primarily due to detachment failure. In contrast, failures of the rotary mechanism were mostly caused by inaccurate positioning and collisions. To address fruit drop issues arising from unstable adsorption, Vu and Ronzhin [208] proposed an adsorption-based end effector with mechanical fingers. The suction cup extends between the fingers to adsorb the fruit, which is then cut by coordinating finger closure and robotic arm retraction. This structure reduces collisions with adjacent fruits or branches in dense environments, improving harvesting stability (Fig. 16E). Wang et al. [19] adopted a structure with four uniformly distributed suction cups to detach apples via rotation-pulling, achieving excellent non-destructive performance. However, when apples have a large curvature or the surrounding environment is complex, the suction cups cannot fully adhere to the fruit surface, easily resulting in adsorption failure (Fig. 16F).

In summary, vacuum end effectors offer significant advantages such as high speed, efficiency, and low damage. However, their performance is largely influenced by adsorption force stability and adsorption point positioning accuracy. In practical orchard environments, adsorption failures are mostly associated with fruit slippage, insufficient suction cup adherence, and collisions between the end effector and surrounding obstacles. Therefore, future research should focus on enhancing adsorption stability, optimizing negative pressure control strategies, and leveraging more precise image recognition algorithms to determine fruit adsorption points and postures. Additionally, while vacuum adsorption can compensate for positioning errors to a certain extent-resulting in lower positioning accuracy requirements compared to clamping-type end effectors-its robustness in complex environments still needs substantial improvement.

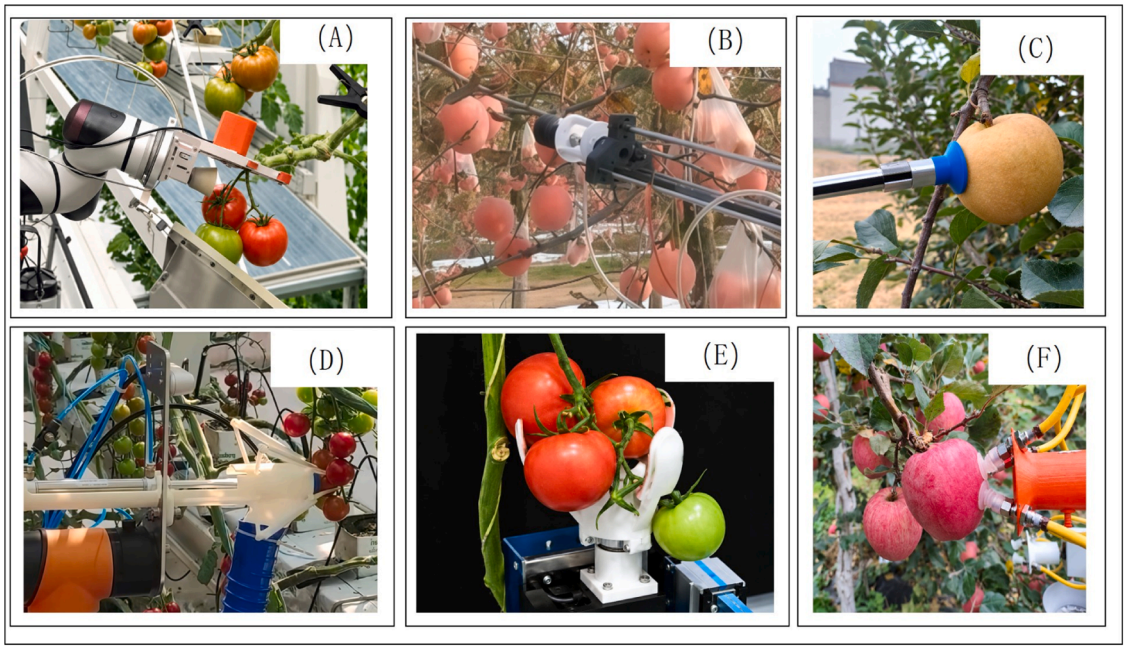


Fig. 16. Negative Pressure End Effectors: (A) Vacuum Gripping Cutting Tomato Picking End Effector, (B) 4-DOF Robotic Arm Vacuum Apple Picking End Effector, (C) Apple Picking Vacuum End Effector, (D) Claw Vacuum End Effector, (E) Mechanical Finger Obstacle-Avoiding Vacuum End Effector, (F) Four-Suction Cup Vacuum Apple Picking End Effector.

5.2.4. Enveloping end effector

Enveloping end effectors achieve full-enclosure clamping of fruits through circular or cylindrical structures and realize pedicel detachment via multi-point force application. Thus, they exhibit high adaptability in harvesting fruits with irregular surfaces, soft textures, or damage-sensitive characteristics (e.g., strawberries, pineapples, and kiwifruits). Their full-enclosure feature reduces the requirement for target positioning accuracy; however, for clustered fruits (particularly clustered strawberries), significant inter-fruit interference results in the highest difficulty in positioning and detachment. As illustrated in Fig. 17A, the kiwifruit harvesting system developed by Fu et al. [209] employs a collaborative workflow between a bottom-up enclosing structure and a top rotary cutting mechanism. This design maintains uniform force distribution on the fruit during pedicel cutting, avoids damage to immature fruits caused by high pedicel toughness, and significantly improves harvesting efficiency. Fig. 17B shows the cable-driven enveloping end effector developed by Xiong et al. [15] for strawberry harvesting. The mechanism approaches the fruit from below and opens six fingers to form a closed ring that encloses the target strawberry. After enclosure, the fingers guide the peduncle toward the internal cutter while separating neighboring berries, thereby reducing the dependence on precise peduncle localization. Internal infrared sensors provide gripper-level positional correction, improving tolerance to visual localization errors, while the integrated storage chamber enables continuous harvesting of multiple strawberries. As illustrated in Fig. 17C, Wang et al. [210] developed a bioinspired pneumatic soft enveloping gripper inspired by the structure and feeding behavior of the Asian elephant trunk. The gripper consists of a fingertip-like process and an enveloping structure with independent pneumatic actuation. The fingertip-like process first contacts the fruit for tactile confirmation, after which the enveloping structure conforms to the fruit surface to establish large-area contact. This design reduces localized contact force while maintaining sufficient holding capacity, thereby improving the reliability and damage resistance of fragile-fruit manipulation. Combined with visual positioning, tactile feedback, and closed-loop force control, the system further enhances its adaptability to fruits with different sizes and mechanical properties.

To address challenges such as short/concealed pedicels, high toughness, and damage susceptibility during pulling, Yin et al. [211] proposed a sleeve-type design for citrus harvesting. By avoiding direct grasping and adopting terminal cutting, the design improved success rates. Nevertheless, its performance is still affected by fruit size variations and posture deviations induced by vibration (Fig. 17D). Li et al. [212] developed an end effector for clustered tomato harvesting, which encloses target fruits with a net from bottom to top and cuts pedicels using a blade. This design effectively avoids clamping clustered fruits, reduces harvesting damage, and is applicable to various cluster-grown fruits (e.g., grapes, litchis). However, the end effector exhibits low motion degrees of freedom in natural environments, as shown in Fig. 17E.

The encompassing end effectors provide benefits including the reduction of gripping damage, the mitigation of failure rates due to pressure variations in adsorption-based systems, and the alleviation of computing demands on visual algorithms. Nonetheless, they are susceptible to collisions and issues with decoupling during the removal of clustered fruits, with a somewhat narrow spectrum of suitable crops. Moreover, their adaptability in intricate obstacle-laden situations is subpar compared to bionic multi-fingered configurations, limiting their extensive use in natural orchards. At present, investigations on clustered and multi-fruit harvesting are markedly inadequate. Future research needs to concentrate on advancements in adaptive enveloping architectures, collision management, and detachment systems for densely populated growing settings.

5.2.5. A novel end-effector configuration

Recent advances in novel harvesting end-effectors have shown parallel trends toward rigid-flexible coupling, biomimetic manipulation, functional integration, and closed-loop sensing. As shown in Fig. 18A, An et al. [213] proposed a linkage-integrated Fin Ray tomato gripper that couples the adaptive deformation of flexible fingers with a four-bar peduncle-cutting mechanism. It achieves compliant enveloping and peduncle severing without an additional cutting motor, delivering a cutting force of 13.9 N and accommodating tomatoes up to 73 mm in diameter. As shown in Fig. 18B, Pu et al. [214] developed a three-finger fruit-displacement and dual V-shaped passive cutting configuration to address branch and leaf occlusion and variations in citrus peduncle inclination. A crank-slider-rocker mechanism enables fruit displacement and peduncle cutting; orchard trials achieved a harvesting success rate

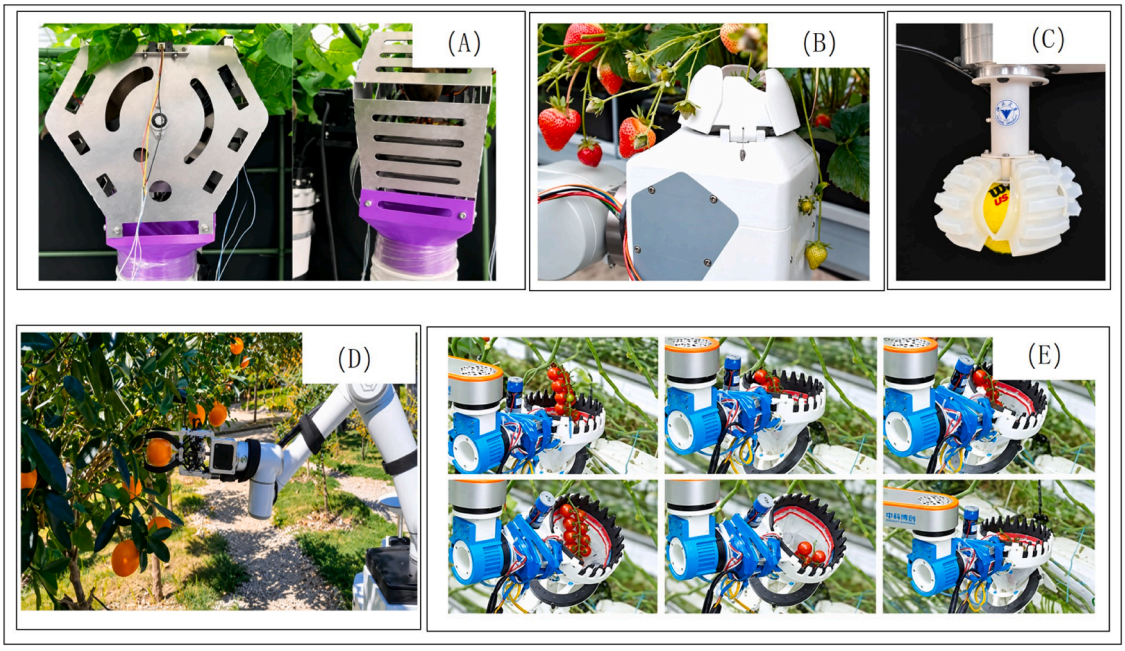


Fig. 17. Enveloping End Effectors: (A) Enveloping end effector for multiple kiwifruit, (B) Enveloping end effector for strawberries, (C) Bioinspired pneumatic soft enveloping gripper, (D) Sleeve-type enveloping end effector for citrus harvesting, (E) Enveloping end effector for cluster tomato harvesting.

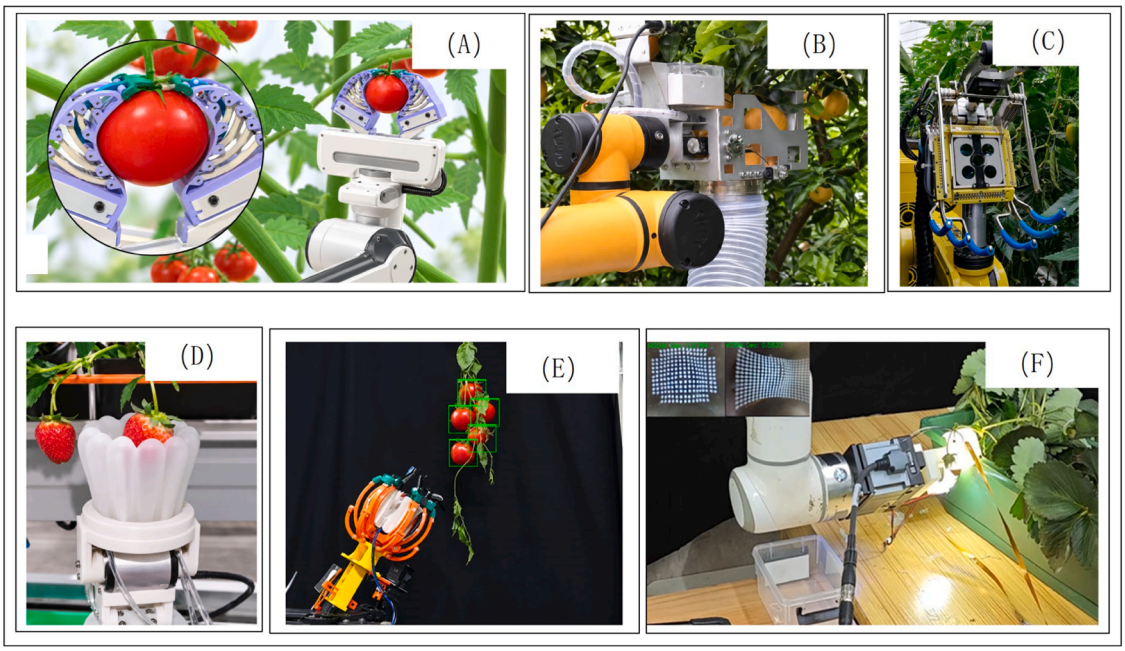


Fig. 18. Novel Integrated Harvesting End Effectors: (A) Linkage-integrated Fin Ray gripper for tomato harvesting, (B) Three-finger fruit-displacement and dual V-shaped passive cutting end effector for citrus harvesting, (C) Integrated end effector of the SWEEPER sweet pepper harvesting robot, (D) Rigid-flexible coupled gripper for strawberry harvesting, (E) Hybrid negative-Poisson's-ratio gripper for cluster tomato harvesting, (F) Biomimetic tactile end effector for strawberry harvesting.

of 95.38%, zero fruit damage, and an efficiency of 6.50 s per fruit. In the SWEEPER sweet pepper harvesting robot, Arad et al. [215] integrated the end-effector with visual localization, robotic-arm servo control, and a greenhouse mobile platform, and validated the system in commercial greenhouses. Their results indicated that foliage occlusion and fruit clustering remain key limitations to harvesting performance, while improved cultivation conditions substantially increased the harvesting rate, suggesting that end-effector configurations should be co-optimized with crop canopy architecture, as shown in Fig. 18C. As shown in Fig. 18D, Yang et al. [216] developed a rigid-flexible strawberry gripper incorporating annular pneumatic chambers and rigid supports. Through engulfing-style harvesting, air-jet obstacle clearing, and SVM-based state recognition, the gripper improved tolerance to positioning errors and obstacles, achieving a harvesting success rate of 93.4% and a damage rate of only 2.9%. Ansari et al. [217] designed a hybrid tomato gripper comprising six soft negative-Poisson's-ratio structures and rigid linkages to enable enveloping grasping. The system integrates force-sensing resistors, an infrared-triggered cutting blade, and depth vision to improve selective harvesting in clustered tomato environments, as shown in Fig. 18E. As shown in Fig. 18F, Guo et al. [218] drew inspiration from the thumb-finger pinch pattern of the human hand and designed a TacTip tactile thumb with an integrated "fingernail" peduncle-cutting function, together with a supporting tactile finger. A hybrid model combining structural similarity index measure (SSIM) with CNN features, termed SSIM-CNN, was used to recognize tactile deformation states and enable closed-loop force regulation during strawberry harvesting.

These studies demonstrate that novel end-effectors are evolving from single gripping or cutting mechanisms toward integrated designs that combine compliant contact, biomimetic motion, active obstacle clearing, peduncle-cutting mechanisms, and multi-modal sensing. Such advances improve adaptability, non-destructive handling, and harvesting reliability in complex orchard environments.

6. Challenges and prospects

Despite substantial advances in artificial intelligence, sensing technologies, robotic manipulation, and end-effector design, the performance of fruit harvesting robots in natural environments still differs considerably from that achieved under controlled experimental conditions. The remaining limitations are increasingly associated with insufficient coordination among environmental perception, physical interaction, and robotic execution, rather than the performance of individual modules. Accordingly, the major challenges can be categorized into three interrelated aspects: visual perception and localization, end-effector adaptability and fruit–robot interaction, and hand–eye coordination.

6.1. Challenges

6.1.1. Visual perception and localization

Although fruit recognition has progressed from handcrafted visual features to deep neural networks, robust perception in natural harvesting environments remains difficult. Variations in illumination, shadows, visual similarity between fruits and backgrounds, changes in target scale, dense fruit distribution, and varying degrees of occlusion affect target appearance and visibility. These challenges are particularly pronounced when detecting small or visually inconspicuous structures relevant to harvesting operations, such as pedicels and cutting regions, which are often less distinguishable and more severely occluded than the fruit body. Meanwhile, lightweight deployment introduces an additional trade-off between feature representation, recognition accuracy, and computational efficiency. Consequently, improvements in image-level detection performance have not fully resolved missed detections and false positives under complex field conditions, particularly when targets exhibit visual characteristics similar to those of surrounding vegetation or are partially occluded [219–221].

More importantly, accurate fruit recognition alone does not guarantee sufficient perceptual information for successful harvesting. Robotic harvesting requires information beyond fruit category and position, including the fruit–pedicel relationships, attachment structure, picking or cutting point, and feasible approach region. Existing methods have therefore gradually evolved from fruit-only detection toward joint perception and structural reasoning of fruits, pedicels, stems, and fruit-

bearing branches. However, this information is commonly obtained through multi-stage pipelines involving segmentation, structural association, morphological processing, skeleton extraction, or geometric fitting. Errors in any intermediate stage can propagate to subsequent picking-point calculation, while severe occlusion may make the actual fruit–stem connection completely unobservable. Thus, a fruit may be accurately detected while its operational structure remains unavailable, revealing a fundamental gap between object-level perception and manipulation-level perception [222–226].

The transition from 2D recognition to 3D operational localization introduces further uncertainty. RGB-D, structured-light, and binocular vision can map image-space detections into physical coordinates, but depth noise, missing measurements, outdoor illumination interference, stereo correspondence errors, and calibration uncertainty continue to affect spatial accuracy. More importantly, many systems still use the fruit centroid or bounding-box center as a geometric proxy for localization. Although this strategy simplifies 3D reconstruction, such points do not necessarily correspond to the actual grasping or cutting location and provide limited information about the orientation required for end-effector interaction. Therefore, the challenge is gradually shifting from simply determining whether a fruit can be localized in 3D to whether its operation-relevant position and pose can be reconstructed accurately and consistently [227,228].

In addition, target visibility does not necessarily imply target harvestability. A correctly detected and localized fruit may remain inaccessible because of stems, branches, supporting structures, neighboring fruits, or insufficient free space for the end effector. Perception must therefore support not only target recognition but also maturity assessment, obstacle understanding, accessibility evaluation, and target selection. This suggests that conventional detection and segmentation metrics alone cannot fully characterize the practical value of a harvesting vision system. The central challenge is consequently shifting from recognizing and locating fruits toward operation-oriented scene understanding, in which perception must answer not only what and where the target is, but also where to interact, how to approach, and whether it can be harvested under the current environmental configuration [229]. As a complementary temporal component, Mhamed et al. [230] developed a vision transformer-based approach for apple variety and growth-stage classification and used time-series prediction to inform harvesting schedules and decisions.

6.1.2. End-effector adaptability and fruit–robot interaction

The end effector directly determines whether perceptual and planning results can be converted into successful physical harvesting. Unlike standardized industrial objects, fruits exhibit substantial variations in size, shape, orientation, stiffness, surface properties, maturity, and attachment characteristics. The end effector must therefore generate sufficient force for stable grasping and detachment while simultaneously limiting compression, impact, slip, and tissue damage. Flexible and bio-inspired structures have improved passive adaptability by conforming to fruit geometry and increasing contact compliance, but such adaptability is still largely mechanical and is generally optimized for a predefined range of crop morphology. Passive deformation can compensate for moderate geometric variation and localization errors, yet it provides limited information about whether contact is stable, excessive force is being applied, or detachment has been successfully completed [231].

More importantly, fruit harvesting is not simply a grasping problem but a constrained physical interaction process within a cluttered biological environment. Before contacting the target, the end effector may need to pass through leaves, branches, neighboring fruits, stems, and other obstacles, while the available operating space can be comparable to the physical dimensions of the end effector. Consequently, accurate picking-point localization alone cannot guarantee successful execution. A visually correct target may remain physically inaccessible, and an unsuitable approach direction may cause collision, fruit compression, or disturbance of surrounding branches before the intended interaction occurs. Existing attempts to combine obstacle perception, free-space assessment, and active obstruction removal demonstrate that the accessibility of a fruit is jointly determined by environmental geometry, obstacle characteristics, end-effector dimensions, and approach posture, thereby exposing a persistent gap between visual localization accuracy and actual harvesting success [195].

The deeper challenge is therefore the transition from passive mechanical adaptability toward active interaction adaptability. Current end effectors can often tolerate a certain degree of uncertainty through structural compliance, but their ability to perceive and regulate the fruit–robot interaction state remains limited. Fruit damage is consequently often assessed after harvesting rather than controlled continuously during contact and detachment. Improving adaptability will require the end effector to move beyond simply accommodating fruit geometry and toward understanding whether contact is safe, stable, and sufficient for the intended harvesting action.

6.1.3. Hand–eye coordination

Hand–eye coordination forms the critical bridge between perception and physical execution. Most current harvesting systems still follow a sequential process in which targets are detected and localized in the camera coordinate system, transformed into the robot coordinate system through hand–eye calibration, and subsequently approached according to a planned trajectory. This architecture is effective when the target and surrounding environment remain approximately static, but natural harvesting is inherently dynamic. Wind, branch elasticity, manipulator motion, and end-effector contact can alter fruit position and orientation after the initial observation, meaning that the target pose used for motion planning may no longer represent the actual state when the end effector reaches the interaction region.

At the same time, operational errors accumulate throughout the perception–transformation–planning–execution chain. Recognition uncertainty, depth error, camera calibration, hand–eye transformation, robot kinematics, and end-effector positioning all contribute to the final interaction error. Even when individual modules achieve relatively high accuracy, their accumulated deviation can become significant relative to the dimensions of a thin pedicel or narrow cutting region. This explains why high recognition and localization accuracy does not always translate into an equivalent harvesting success rate. The problem becomes more evident during the final approach, where millimeter-level deviations can determine whether the end effector successfully grasps or cuts the intended structure [232].

Coordination complexity increases further in multi-arm harvesting systems. Although multiple manipulators can theoretically increase harvesting throughput through parallel operation, they introduce additional requirements for target allocation, workspace partitioning, collision avoidance, trajectory synchronization, and shared environmental perception. The action of one manipulator can disturb branches or neighboring fruits and thereby invalidate the perception or trajectory of another. Therefore, increasing mechanical parallelism also increases spatial and temporal coupling among perception, planning, and execution, making system-level coordination increasingly important [233].

The central challenge of hand–eye coordination is therefore no longer merely obtaining an accurate static transformation between the camera and robot coordinate systems. Instead, it is maintaining continuous spatial and temporal consistency among visual perception, manipulator motion, end-effector state, and environmental changes throughout the harvesting process. The transition from open-loop or weakly closed-loop execution toward continuous perception–action coupling is essential for improving harvesting reliability under dynamic natural conditions.

6.2. Future outlook

6.2.1. Operation-oriented visual perception

Future harvesting vision should evolve from object-oriented

perception toward operation-oriented perception. Rather than optimizing fruit detection and segmentation as isolated tasks, perception systems should directly provide information required for manipulation, including target harvestability, attachment structure, picking or cutting points, 3D pose, feasible approach direction, obstacle distribution, and available operating space. Accordingly, the evaluation of visual perception should also move beyond conventional image-level metrics toward task-level indicators that reflect whether perception results can successfully support robotic execution.

An important direction will be the development of unified scene representations that explicitly model the structural and functional relationships among fruits, pedicels, branches, leaves, and environmental obstacles. Instead of independently recognizing individual objects and reconstructing their relationships through multiple post-processing stages, future perception systems should directly reason about how a fruit is attached, where separation should occur, and which approach directions are physically feasible. Such representations would reduce the propagation of errors between independent perception stages and provide information that is more directly aligned with manipulation requirements.

Multi-view and multimodal perception will be particularly important for resolving incomplete observations. RGB images provide rich semantic information, while depth, stereo, and point clouds provide spatial geometry, and temporal observations can supplement information lost in individual frames. These modalities should therefore be integrated according to their complementary strengths. When critical operation-related structures remain invisible, the robot should actively adjust the sensor or manipulator viewpoint to acquire additional observations rather than relying solely on inference from incomplete information. This would transform visual perception from passive observation toward active perception driven by harvesting requirements.

Future perception systems should also explicitly model temporal dynamics and uncertainty. Fruit pose, picking points, and obstacle configurations should be continuously estimated and updated rather than treated as fixed outputs from a single image. When perceptual uncertainty exceeds an acceptable level, the robot should be able to acquire another viewpoint, modify the approach direction, or postpone the target instead of immediately executing an unreliable action. In this way, visual perception can become an active component of robotic decision-making rather than merely the front end of the harvesting pipeline.

6.2.2. Adaptive and perception-integrated end effectors

Future end effectors should progress from passive compliant mechanisms toward adaptive and perception-integrated interaction systems. Mechanical flexibility will remain important for accommodating fruit morphology and reducing contact damage, but it should increasingly be combined with force, tactile, pressure, proximity, slip, and deformation sensing. Such multimodal information can provide direct observations of the interaction state after physical contact, where conventional external vision may become partially or completely unavailable.

With interaction feedback, grasping force, finger displacement, suction pressure, cutting motion, and detachment strategy could be adjusted online according to fruit size, stiffness, maturity, surface condition, and attachment state. Rather than executing fixed manipulation parameters, the end effector could progressively establish contact, evaluate grasp stability, regulate interaction force, detect detachment, and terminate the action when harvesting is completed. This would extend adaptability from geometric accommodation to active regulation of the entire interaction process.

Visual perception and end-effector design should also become more tightly integrated. The geometry of fingers, cutters, suction devices, and protective structures should be designed together with camera placement and sensing coverage to minimize self-occlusion during the final approach. Local cameras and proximity sensors near the end effector could provide fine-scale information when global cameras lose visibility in dense foliage, enabling continuous refinement of picking-point position and approach posture.

Ultimately, the end effector should evolve from a passive mechanical tool into an intelligent interaction unit capable of sensing, adapting, and regulating harvesting actions. Its performance should therefore be evaluated not only by grasping success, but also by accessibility, contact stability, detachment reliability, fruit and plant damage, cycle time, and adaptability across different morphological and environmental conditions.

6.2.3. Closed-loop hand–eye coordination

Future harvesting systems should move from the conventional detect–locate–move–harvest sequence toward continuous closed-loop hand–eye coordination. Rather than relying on a target pose calculated before manipulator motion, the relative relationship between the target and end effector should be continuously updated during approach, alignment, contact, and detachment, allowing robot trajectories and end-effector posture to be corrected in real time.

A hierarchical coordination framework is particularly suitable for this purpose. Global vision can first perform scene understanding, target selection, and coarse 3D localization, allowing the manipulator to rapidly enter the target region. Local eye-in-hand vision can subsequently provide fine relative localization during the final approach, while force and tactile sensing can regulate physical interaction after contact. This would establish a progressive closed loop of global perception → coarse approach → local perception → fine alignment → contact feedback → harvesting verification, reducing dependence on the accuracy of any single sensing or calibration stage.

Dynamic prediction should also become an integral component of hand–eye coordination. For fruits moving because of wind, branch elasticity, or robot-induced disturbance, the robot should estimate motion states and predict future target positions rather than simply reacting to instantaneous measurements. Combining target tracking, motion prediction, visual servoing, and compliant control would help compensate for delays introduced by image processing, communication, trajectory generation, and mechanical response.

For multi-arm harvesting systems, closed-loop coordination should further extend from individual manipulators to the entire robotic system. Target allocation, collision avoidance, trajectory replanning, and harvesting status should be continuously updated using shared environmental perception. When one manipulator disturbs a branch or changes the configuration of neighboring fruits, other manipulators should update their perception and motion plans accordingly rather than continuing to execute trajectories generated from outdated scene information.

Ultimately, fruit harvesting robots should move from separately optimized vision, manipulators, and end effectors toward embodied closed-loop harvesting. In such a framework, perception continuously guides action, robot actions actively alter sensing conditions, and physical interaction feedback continuously updates subsequent decisions. Establishing this tightly coupled perception–action–interaction loop will be essential for improving the robustness, efficiency, and autonomy of harvesting robots in complex and dynamic natural environments.

7. Conclusion

This review synthesized the progress in machine vision, fruit localization, end-effector design, and hand-eye coordination in a perception-action-interaction context. It observed that better detection accuracy does not guarantee reliable harvesting, as occlusion, plant motion, calibration errors, and limited interaction feedback cause uncertainty to be present from the beginning. Operation-oriented perception became the keypoint of view and linked fruit position and attachment with accessibility, approach direction, and harvestability. The work

identified active perception, adaptive end-effectors and visual-force-tactile feedback in closed-loop control as keypoints of improvement in the field of orchard management. The research also focused on harvesting success, cycle time, fruit damage, and performance across orchard conditions. The review indicated that coordinated perception and physical interaction are the top research priority in order to bridge the gap between accurate fruit recognition and reliable autonomous harvesting.

CRediT authorship contribution statement

Liujing Yu: Writing – review & editing, Writing – original draft. **Yingjun Pu:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition. **Ju Lin:** Writing – review & editing. **Mustafa Mhamed:** Writing – review & editing. **Reza Ehsani:** Writing – review & editing. **Cheng Shen:** Writing – review & editing. **Fuzeng Yang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank all researchers for their achievements and all the materials that have been referred to in this study. The authors greatly appreciate the funding from the National Key Research and Development Program of China (No. 2025YFE0209300) and the National Natural Science Foundation of China (No. 5227120922, 52205284). This work was also supported by the Fundamental Research Funds for the Central Universities (No. SWU-KT26029). Meanwhile, we are grateful for the critical comments and suggestions from the anonymous reviewers for improving the manuscript.

Data availability

Data will be made available on request.

References

[1] Z. Zhang, C. Igathinathane, J. Li, H. Cen, Y. Lu, P. Flores, Technology progress in mechanical harvest of fresh market apples, Comput. Electron. Agric. 175 (2020) 105606, https://doi.org/10.1016/j.compag.2020.105606.
[2] M. Mhamed, Z. Zhang, J. Yu, Y. Li, M. Zhang, Advances in apple's automated orchard equipment: a comprehensive research, Comput. Electron. Agric. 221 (2024) 108926, https://doi.org/10.1016/j.compag.2024.108926.
[3] Y. Yu, H. Xie, K. Zhang, Y. Wang, Y. Li, J. Zhou, L. Xu, Design, development, integration, and field evaluation of a ridge-planting strawberry harvesting robot, Agriculture 14 (2024) 2126, https://doi.org/10.3390/agriculture14122126.
[4] S.L. Wang, N.R. Loduca, The changing values of the US farm workers' legal status and labor quality adjustment: a hedonic price analysis, Appl. Econ. Perspect. Policy 47 (2025) 1791–1810, https://doi.org/10.1002/aepp.13530.
[5] Y. Zhang, K. Zhang, L. Yang, D. Zhang, T. Cui, Y. Yu, H. Liu, Design and simulation experiment of ridge planting strawberry picking manipulator, Comput. Electron. Agric. 208 (2023) 107690, https://doi.org/10.1016/j.compag.2023.107690.
[6] Y. Zhao, L. Gong, Y. Huang, C. Liu, A review of key techniques of vision-based control for harvesting robot, Comput. Electron. Agric. 127 (2016) 311–323, https://doi.org/10.1016/j.compag.2016.06.022.
[7] S. Zhang, K. Hu, W. Sha, Q. Chen, Z. Hou, S. Weng, Efficient one-stage location method for grape picking points in natural scene by combining detection network and point regression, Comput. Electron. Agric. 230 (2025) 109725, https://doi.org/10.1016/j.compag.2024.109725.
[8] L. He, Y. Sun, L. Chen, Q. Feng, Y. Li, J. Lin, Y. Qiao, C. Zhao, Advance on agricultural robot hand–Eye coordination for agronomic task: a review, Engineering 51 (2025) 263–279, https://doi.org/10.1016/j.eng.2025.01.022.
[9] X. Du, H. Cheng, Z. Ma, W. Lu, M. Wang, Z. Meng, C. Jiang, F. Hong, DSW-YOLO: a detection method for ground-planted strawberry fruits under different occlusion levels, Comput. Electron. Agric. 214 (2023) 108304, https://doi.org/10.1016/j.compag.2023.108304.
[10] M. Chen, Y. Tang, X. Zou, K. Huang, Z. Huang, H. Zhou, C. Wang, G. Lian, Three-dimensional perception of orchard banana central stock enhanced by adaptive multi-vision technology, Comput. Electron. Agric. 174 (2020) 105508, https://doi.org/10.1016/j.compag.2020.105508.
[11] J. Li, M. Karkee, Q. Zhang, K. Xiao, T. Feng, Characterizing apple picking patterns for robotic harvesting, Comput. Electron. Agric. 127 (2016) 633–640, https://doi.org/10.1016/j.compag.2016.07.024.
[12] Z. Li, F. Miao, Z. Yang, P. Chai, S. Yang, Factors affecting human hand grasp type in tomato fruit-picking: a statistical investigation for ergonomic development of harvesting robot, Comput. Electron. Agric. 157 (2019) 90–97, https://doi.org/10.1016/j.compag.2018.12.047.
[13] Y. Miao, J. Zheng, Optimization design of compliant constant-force mechanism for apple picking actuator, Comput. Electron. Agric. 170 (2020) 105232, https://doi.org/10.1016/j.compag.2020.105232.
[14] J. Gao, F. Zhang, J. Zhang, H. Guo, J. Gao, Picking patterns evaluation for cherry tomato robotic harvesting end-effector design, Biosyst. Eng. 239 (2024) 1–12, https://doi.org/10.1016/j.biosystemseng.2024.01.009.
[15] Y. Xiong, C. Peng, L. Grimstad, P.J. From, V. Isler, Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper, Comput. Electron. Agric. 157 (2019) 392–402, https://doi.org/10.1016/j.compag.2019.01.009.
[16] Y. Jin, C. Yu, J. Yin, S.X. Yang, Detection method for table grape ears and stems based on a far-close-range combined vision system and hand-eye-coordinated picking test, Comput. Electron. Agric. 202 (2022) 107364, https://doi.org/10.1016/j.compag.2022.107364.
[17] G. Hu, C. Chen, J. Chen, L. Sun, A. Sugirbay, Y. Chen, H. Jin, S. Zhang, L. Bu, Simplified 4-DOF manipulator for rapid robotic apple harvesting, Comput. Electron. Agric. 199 (2022) 107177, https://doi.org/10.1016/j.compag.2022.107177.
[18] Z. Yao, J. Xiong, J. Yang, X. Wang, Z. Li, Y. Huang, Y. Li, Design and verification of a litchi combing and cutting end-effector based on visual-tactile fusion, Comput. Electron. Agric. 232 (2025) 110077, https://doi.org/10.1016/j.compag.2025.110077.
[19] M. Wang, B. Yan, S. Zhang, P. Fan, P. Zeng, S. Shi, F. Yang, Development of a novel biomimetic mechanical hand based on physical characteristics of apples, Agriculture 12 (2022) 1871, https://doi.org/10.3390/agriculture12111871.
[20] M. Li, P. Liu, A bionic adaptive end-effector with rope-driven fingers for pear fruit harvesting, Comput. Electron. Agric. 211 (2023) 107952, https://doi.org/10.1016/j.compag.2023.107952.
[21] J. Rong, P. Wang, T. Wang, L. Hu, T. Yuan, Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots, Comput. Electron. Agric. 202 (2022) 107430, https://doi.org/10.1016/j.compag.2022.107430.
[22] Q. Zhou, R. Gao, Z. Ma, G. Cao, J. Chen, Research on system integration and control methods of an apple-picking robot in unstructured environment, Ind. Robot: Int. J. Robot. Res. Appl. 51 (2024) 369–379, https://doi.org/10.1108/ir-11-2023-0282.
[23] Z. Xu, J. Liu, J. Wang, L. Cai, Y. Jin, S. Zhao, B. Xie, Realtime picking point decision algorithm of trellis grape for high-speed robotic cut-and-catch harvesting, Agronomy 13 (2023) 1618, https://doi.org/10.3390/agronomy13061618.
[24] T. Yi, D. Zhang, L. Luo, J. Luo, View planning for grape harvesting based on active vision strategy under occlusion, IEEE Robot. Autom. Lett. 9 (2024) 2535–2542, https://doi.org/10.1109/lra.2024.3357397.
[25] T. Chen, S. Zhang, J. Chen, G. Fu, Y. Chen, L. Zhu, Development, integration, and field experiment optimization of an autonomous banana-picking robot, Agriculture 14 (2024) 1389, https://doi.org/10.3390/agriculture14081389.
[26] L. Yang, T. Noguchi, Y. Hoshino, Development of a pumpkin fruits pick-and-place robot using an RGB-D camera and a YOLO based object detection AI model, Comput. Electron. Agric. 227 (2024) 109625, https://doi.org/10.1016/j.compag.2024.109625.
[27] R. Barth, J. Hemming, E.J. van Henten, Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation, Biosyst. Eng. 146 (2016) 71–84, https://doi.org/10.1016/j.biosystemseng.2015.12.001.
[28] M. Chen, Y. Tang, X. Zou, Z. Huang, H. Zhou, S. Chen, 3D global mapping of large-scale unstructured orchard integrating eye-in-hand stereo vision and SLAM, Comput. Electron. Agric. 187 (2021) 106237, https://doi.org/10.1016/j.compag.2021.106237.
[29] D.W. Choi, J.H. Park, J.-H. Yoo, K. Ko, AI-driven adaptive grasping and precise detaching robot for efficient citrus harvesting, Comput. Electron. Agric. 232 (2025) 110131, https://doi.org/10.1016/j.compag.2025.110131.
[30] H.A. Williams, M.H. Jones, M. Nejati, M.J. Seabright, J. Bell, N.D. Penhall, J. J. Barnett, M.D. Duke, A.J. Scarfe, H.S. Ahn, J. Lim, B.A. MacDonald, Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms, Biosyst. Eng. 181 (2019) 140–156, https://doi.org/10.1016/j.biosystemseng.2019.03.007.
[31] Z. Jiang, J. Zhou, H. Han, Y. Wang, An apple fruit localization system based on accurate and flexible hand-eye pose acquisition for robotic harvesting, Comput. Electron. Agric. 229 (2025) 109705, https://doi.org/10.1016/j.compag.2024.109705.
[32] Y. Ge, Y. Xiong, G.L. Tenorio, P.J. From, Fruit localization and environment perception for strawberry harvesting robots, IEEE Access 7 (2019) 147642–147652, https://doi.org/10.1109/access.2019.2946369.
[33] G. Ren, T. Wu, T. Lin, L. Yang, G. Chowdhary, K.C. Ting, Y. Ying, Mobile robotics platform for strawberry sensing and harvesting within precision indoor farming systems, J. Field Robot. 41 (2024) 2047–2065, https://doi.org/10.1002/rob.22207.

[34] H. Wang, Y. Lin, X. Xu, Z. Chen, Z. Wu, Y. Tang, A study on long-close distance coordination control strategy for Litchi picking, Agronomy 12 (2022) 1520, https://doi.org/10.3390/agronomy12071520.
[35] Q. Feng, W. Zou, P. Fan, C. Zhang, X. Wang, Design and test of robotic harvesting system for cherry tomato, Int. J. Agric. Biol. Eng. 11 (2018) 96–100, https://doi.org/10.25165/j.ijabe.20181101.2853.
[36] C. Lehnert, A. English, C. McCool, A.W. Tow, T. Perez, Autonomous sweet pepper harvesting for protected cropping systems, IEEE Robot. Autom. Lett. 2 (2017) 872–879, https://doi.org/10.1109/lra.2017.2655622.
[37] R. Barth, J. Hemming, E.J. Van Henten, Angle estimation between plant parts for grasp optimisation in harvest robots, Biosyst. Eng. 183 (2019) 26–46, https://doi.org/10.1016/j.biosystemseng.2019.04.006.
[38] T. Sun, W. Zhang, Z. Miao, Z. Zhang, N. Li, Object localization methodology in occluded agricultural environments through deep learning and active sensing, Comput. Electron. Agric. 212 (2023) 108141, https://doi.org/10.1016/j.compag.2023.108141.
[39] T. Yi, D. Zhang, L. Luo, Y. Wang, B. Liu, View planning for grape harvesting based on self-supervised deep reinforcement learning under occlusion, Comput. Electron. Agric. 239 (2025) 110913, https://doi.org/10.1016/j.compag.2025.110913.
[40] Z.-F. Xu, R.-S. Jia, H.-M. Sun, Q.-M. Liu, Z. Cui, Light-YOLOv3: fast method for detecting green mangoes in complex scenes using picking robots, Appl. Intell. 50 (2020) 4670–4687, https://doi.org/10.1007/s10489-020-01818-w.
[41] J. Xiong, Z. Liu, R. Lin, R. Bu, Z. He, Z. Yang, C. Liang, Green grape detection and picking-point calculation in a night-time natural environment using a charge-coupled device (CCD) vision sensor with artificial illumination, Sensors 18 (2018) 969, https://doi.org/10.3390/s18040969.
[42] C. Yang, L. Xiong, Z. Wang, Y. Wang, G. Shi, T. Kuremot, W. Zhao, Y. Yang, Integrated detection of citrus fruits and branches using a convolutional neural network, Comput. Electron. Agric. 174 (2020) 105469, https://doi.org/10.1016/j.compag.2020.105469.
[43] K. Zhao, Y. Li, Z. Liu, Three-dimensional spatial perception of blueberry fruits based on improved YOLOv11 network, Agronomy 15 (2025) 535, https://doi.org/10.3390/agronomy15030535.
[44] F. Gao, L. Fu, X. Zhang, Y. Majeed, R. Li, M. Karkee, Q. Zhang, Multi-class fruit-on-plant detection for apple in SNAP system using faster R-CNN, Comput. Electron. Agric. 176 (2020) 105634, https://doi.org/10.1016/j.compag.2020.105634.
[45] X. Jing, Y. Wang, D. Li, W. Pan, Melon ripeness detection by an improved object detection algorithm for resource constrained environments, Plant Methods 20 (2024) 127, https://doi.org/10.1186/s13007-024-01259-3.
[46] Z. He, Z. Liu, Z. Zhou, M. Karkee, Q. Zhang, Improving picking efficiency under occlusion: design, development, and field evaluation of an innovative robotic strawberry harvester, Comput. Electron. Agric. 237 (2025) 110684, https://doi.org/10.1016/j.compag.2025.110684.
[47] H. Okamoto, W.S. Lee, Green citrus detection using hyperspectral imaging, Comput. Electron. Agric. 66 (2009) 201–208, https://doi.org/10.1016/j.compag.2009.02.004.
[48] J.P. Wachs, H.I. Stern, T. Burks, V. Alchanatis, Low and high-level visual feature-based apple detection from multi-modal images, Precis. Agric. 11 (2010) 717–735, https://doi.org/10.1007/s11119-010-9198-x.
[49] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, K. Lewis, Sensors and systems for fruit detection and localization: a review, Comput. Electron. Agric. 116 (2015) 8–19, https://doi.org/10.1016/j.compag.2015.05.021.
[50] R. Linker, O. Cohen, A. Naor, Determination of the number of green apples in RGB images recorded in orchards, Comput. Electron. Agric. 81 (2012) 45–57, https://doi.org/10.1016/j.compag.2011.11.007.
[51] M. Teixidó, D. Font, T. Pallejà, M. Tresanchez, M. Nogués, J. Palacín, An embedded real-time red peach detection system based on an OV7670 camera, ARM Cortex-M4 processor and 3D look-up tables, Sensors 12 (2012) 14129–14143, https://doi.org/10.3390/s121014129.
[52] C. Zheng, P. Chen, J. Pang, X. Yang, C. Chen, S. Tu, Y. Xue, A mango picking vision algorithm on instance segmentation and key point detection from RGB images in an open orchard, Biosyst. Eng. 206 (2021) 32–54, https://doi.org/10.1016/j.biosystemseng.2021.03.012.
[53] R. Van Beers, B. Aernouts, L. León Gutiérrez, C. Erkinbaev, K. Rutten, A. Schenk, B. Nicolaï, W. Saeys, Optimal illumination-detection distance and detector size for predicting braeburn apple maturity from vis/NIR laser reflectance measurements, Food Bioprocess Tech. 8 (2015) 2123–2136, https://doi.org/10.1007/s11947-015-1562-4.
[54] F. Wang, C. Zhao, H. Yang, H. Jiang, L. Li, G. Yang, Non-destructive and in-site estimation of apple quality and maturity by hyperspectral imaging, Comput. Electron. Agric. 195 (2022) 106843, https://doi.org/10.1016/j.compag.2022.106843.
[55] H. Gan, W. Lee, V. Alchanatis, R. Ehsani, J. Schueller, Immature green citrus fruit detection using color and thermal images, Comput. Electron. Agric. 152 (2018) 117–125, https://doi.org/10.1016/j.compag.2018.07.011.
[56] S. Hayashi, K. Shigematsu, S. Yamamoto, K. Kobayashi, Y. Kohno, J. Kamata, M. Kurita, Evaluation of a strawberry-harvesting robot in a field test, Biosyst. Eng. 105 (2010) 160–171, https://doi.org/10.1016/j.biosystemseng.2009.09.011.
[57] S. Mehta, C. Ton, S. Asundi, T. Burks, Multiple camera fruit localization using a particle filter, Comput. Electron. Agric. 142 (2017) 139–154, https://doi.org/10.1016/j.compag.2017.08.007.
[58] C. Hou, J. Xu, Y. Tang, J. Zhuang, Z. Tan, W. Chen, S. Wei, H. Huang, M. Fang, Detection and localization of citrus picking points based on binocular vision, Precis. Agric. 25 (2024) 2321–2355, https://doi.org/10.1007/s11119-024-10169-2.
[59] X. Lei, M. Wu, Y. Li, A. Liu, Z. Tang, S. Chen, Y. Xiang, Detection and positioning of camellia oleifera fruit based on LBP image texture matching and binocular stereo vision, Agronomy 13 (2023) 2153, https://doi.org/10.3390/agronomy13082153.
[60] J. Wang, Z. Xie, P. Mao, M. Sun, J. Guo, Fruit modeling and application based on 3D imaging technology: a review, J. Food Meas. Charact. 18 (2024) 4120–4136, https://doi.org/10.1007/s11694-024-02480-3.
[61] X. Zhang, M. Yao, Q. Cheng, G. Liang, F. Fan, A novel hand-eye calibration method of picking robot based on TOF camera, Front. Plant Sci. 13 (2023) 1099033, https://doi.org/10.3389/fpls.2022.1099033.
[62] Z. Zheng, L. Chen, L. Wei, W. Huang, D. Du, G. Qin, Z. Yang, S. Wang, An efficient and lightweight banana detection and localization system based on deep CNNs for agricultural robots, Smart Agric. Technol. 9 (2024) 100550, https://doi.org/10.1016/j.atech.2024.100550.
[63] J. Li, H. Zhou, Y. Mai, Y. Jia, Z. Zhou, K. Wu, H. Chen, H. Lin, M. Luo, L. Shi, An autonomous obstacle avoidance and path planning method for fruit-picking UAV in orchard environments, Smart Agric. Technol. 10 (2025) 100752, https://doi.org/10.1016/j.atech.2024.100752.
[64] J. Gené-Mola, E. Gregorio, J. Guevara, F. Auat, R. Sanz-Cortiella, A. Escolà, J. Llorens, J.-R. Morros, J. Ruiz-Hidalgo, V. Vilaplana, J.R. Rosell-Polo, Fruit detection in an apple orchard using a mobile terrestrial laser scanner, Biosyst. Eng. 187 (2019) 171–184, https://doi.org/10.1016/j.biosystemseng.2019.08.017.
[65] C. Zhang, C. Zhao, W. Huang, Q. Wang, S. Liu, J. Li, Z. Guo, Automatic detection of defective apples using NIR coded structured light and fast lightness correction, J. Food Eng. 203 (2017) 69–82, https://doi.org/10.1016/j.jfoodeng.2017.02.008.
[66] Y. Lu, R. Li, R. Lu, Structured-illumination reflectance imaging (SIRI) for enhanced detection of fresh bruises in apples, Postharvest Biol. Technol. 117 (2016) 89–93, https://doi.org/10.1016/j.postharvbio.2016.02.005.
[67] H. Kang, C. Chen, Fast implementation of real-time fruit detection in apple orchards using deep learning, Comput. Electron. Agric. 168 (2020) 105108, https://doi.org/10.1016/j.compag.2019.105108.
[68] Z. Zheng, J. Xiong, H. Lin, Y. Han, B. Sun, Z. Xie, Z. Yang, C. Wang, A method of green citrus detection in natural environments using a deep convolutional neural network, Front. Plant Sci. 12 (2021) 705737, https://doi.org/10.3389/fpls.2021.705737.
[69] Y. Zhao, L. Gong, B. Zhou, Y. Huang, C. Liu, Detecting tomatoes in greenhouse scenes by combining AdaBoost classifier and colour analysis, Biosyst. Eng. 148 (2016) 127–137, https://doi.org/10.1016/j.biosystemseng.2016.05.001.
[70] A. Payne, K. Walsh, P. Subedi, D. Jarvis, Estimating mango crop yield using image analysis using fruit at 'stone hardening' stage and night time imaging, Comput. Electron. Agric. 100 (2014) 160–167, https://doi.org/10.1016/j.compag.2013.11.011.
[71] S. Sengupta, W.S. Lee, Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions, Biosyst. Eng. 117 (2014) 51–61, https://doi.org/10.1016/j.biosystemseng.2013.07.007.
[72] W. Ji, D. Zhao, F. Cheng, B. Xu, Y. Zhang, J. Wang, Automatic recognition vision system guided for apple harvesting robot, Comput. Electr. Eng. 38 (2012) 1186–1195, https://doi.org/10.1016/j.compeleceng.2011.11.005.
[73] Y. Xu, K. Imou, Y. Kaizu, K. Saga, Two-stage approach for detecting slightly overlapping strawberries using HOG descriptor, Biosyst. Eng. 115 (2013) 144–153, https://doi.org/10.1016/j.biosystemseng.2013.03.011.
[74] K. Tan, W.S. Lee, H. Gan, S. Wang, Recognising blueberry fruit of different maturity using histogram oriented gradients and colour features in outdoor scenes, Biosyst. Eng. 176 (2018) 59–72, https://doi.org/10.1016/j.biosystemseng.2018.08.011.
[75] P. Wang, T. Niu, D. He, Tomato Young fruits detection method under near color background based on improved faster R-CNN with attention mechanism, Agriculture 11 (2021) 1059, https://doi.org/10.3390/agriculture11111059.
[76] X. Liu, D. Zhao, W. Jia, W. Ji, C. Ruan, Y. Sun, Cucumber fruits detection in greenhouses based on instance segmentation, IEEE Access 7 (2019) 139635–139642, https://doi.org/10.1109/access.2019.2942144.
[77] C. Hu, Z. Shi, H. Wei, X. Hu, Y. Xie, P. Li, Automatic detection of pecan fruits based on faster RCNN with FPN in orchard, Int. J. Agric. Biol. Eng. 15 (2022) 189–196, https://doi.org/10.25165/j.ijabe.20221506.7241.
[78] D. Wang, D. He, Fusion of mask RCNN and attention mechanism for instance segmentation of apples under complex background, Comput. Electron. Agric. 196 (2022) 106864, https://doi.org/10.1016/j.compag.2022.106864.
[79] Y. Liu, H. Ren, Z. Zhang, F. Men, P. Zhang, D. Wu, R. Feng, Research on multi-cluster green persimmon detection method based on improved faster RCNN, Front. Plant Sci. 14 (2023) 1177114, https://doi.org/10.3389/fpls.2023.1177114.
[80] Y. Nan, H. Zhang, Y. Zeng, J. Zheng, Y. Ge, Intelligent detection of Multi-class pitaya fruits in target picking row based on WGB-YOLO network, Comput. Electron. Agric. 208 (2023) 107780, https://doi.org/10.1016/j.compag.2023.107780.
[81] T. Huang, J. Guo, L. Yu, H. Chen, Z. Su, Y. Xue, An automatic tracking method for fruit abscission of litchi using convolutional networks, Comput. Electron. Agric. 224 (2024) 109213, https://doi.org/10.1016/j.compag.2024.109213.
[82] S. Kaukab, Komal, B.M. Ghodki, H. Ray, Y.B. Kalnar, K. Narsaiah, J.S. Brar, Improving real-time apple fruit detection: multi-modal data and depth fusion

with non-targeted background removal, Ecol. Inf. 82 (2024) 102691, https://doi.org/10.1016/j.ecoinf.2024.102691.
[83] A. Gao, Z. Tian, W. Ma, Y. Song, L. Ren, Y. Feng, J. Qian, L. Xu, Fruits hidden by green: an improved YOLOV8n for detection of young citrus in lush citrus trees, Front. Plant Sci. 15 (2024) 1375118, https://doi.org/10.3389/fpls.2024.1375118.
[84] V. Mohammadi, K. Kheiralipour, M. Ghasemi-Varnamkhasti, Detecting maturity of persimmon fruit based on image processing technique, Sci. Hortic. 184 (2015) 123–128, https://doi.org/10.1016/j.scienta.2014.12.037.
[85] Y. Zhang, M. Liu, J. Gong, Y. Lan, Apple recognition based on two-level segmentation and region-marked gradient Hough circle transform, Trans. Chin. Soc. Agric. Eng. (Trans. CSAE) 38 (2022) 110–121, https://doi.org/10.11975/j.issn.1002-6819.2022.19.013.
[86] H. Li, W.S. Lee, K. Wang, Identifying blueberry fruit of different growth stages using natural outdoor color images, Comput. Electron. Agric. 106 (2014) 91–101, https://doi.org/10.1016/j.compag.2014.05.015.
[87] R. Wan Nurazwin Syazwani, H. Muhammad Asraf, M. Megat Syahirul Amin, K. Nur Dalila, Automated image identification, detection and fruit counting of top-view pineapple crown using machine learning, Alex. Eng. J. 61 (2022) 1265–1276, https://doi.org/10.1016/j.aej.2021.06.053.
[88] Y. Tang, H. Zhou, H. Wang, Y. Zhang, Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision, Expert Syst. Appl. 211 (2023) 118573, https://doi.org/10.1016/j.eswa.2022.118573.
[89] F. Xiao, H. Wang, Y. Xu, R. Zhang, Fruit detection and recognition based on deep learning for automatic harvesting: an overview and review, Agronomy 13 (2023) 1625, https://doi.org/10.3390/agronomy13061625.
[90] Z. Liang, X. Li, G. Wang, F. Wu, X. Zou, Palm vision and servo control strategy of tomato picking robot based on global positioning, Comput. Electron. Agric. 237 (2025) 110668, https://doi.org/10.1016/j.compag.2025.110668.
[91] F. Meng, J. Li, Y. Zhang, S. Qi, Y. Tang, Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks, Comput. Electron. Agric. 214 (2023) 108298, https://doi.org/10.1016/j.compag.2023.108298.
[92] H. Si, M. Li, W. Li, G. Zhang, M. Wang, F. Li, Y. Li, A dual-branch model integrating CNN and swin transformer for efficient Apple leaf disease classification, Agriculture 14 (2024) 142, https://doi.org/10.3390/agriculture14010142.
[93] Z. Zhang, Z. Lei, M. Omura, H. Hasegawa, S. Gao, Dendritic Learning-incorporated vision Transformer for image recognition, IEEE/CAA J. Autom. Sin. 11 (2024) 539–541, https://doi.org/10.1109/jas.2023.123978.
[94] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587, https://doi.org/10.1109/cvpr.2014.81.
[95] S. Ren, K. He, R. Girshick, J. Sun, R-CNN Faster, Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, https://doi.org/10.1109/tpami.2016.2577031.
[96] S. Parvathi, S.T. Selvi, Detection of maturity stages of coconuts in complex background using faster R-CNN model, Biosyst. Eng. 202 (2021) 119–132, https://doi.org/10.1016/j.biosystemseng.2020.12.002.
[97] S. Zhao, J. Liu, S. Wu, Multiple disease detection method for greenhouse-cultivated strawberry based on multiscale feature fusion faster R_CNN, Comput. Electron. Agric. 199 (2022) 107176, https://doi.org/10.1016/j.compag.2022.107176.
[98] Y. Nan, H. Zhang, J. Zheng, K. Yang, Pitaya detection using an improved lightweight faster R-CNN based on MobileNetV3 in densely planted pitaya orchards, J. Agric. Eng. 56 (2025) 1886, https://doi.org/10.4081/jae.2025.1886.
[99] K. He, G. Gkioxari, P. Dollar, R. Girshick, R-CNN Mask, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 386–397, https://doi.org/10.1109/tpami.2018.2844175.
[100] X. Huang, D. Peng, H. Qi, L. Zhou, C. Zhang, Detection and instance segmentation of grape clusters in orchard environments using an improved mask R-CNN model, Agriculture 14 (2024) 918, https://doi.org/10.3390/agriculture14060918.
[101] P. Xu, N. Fang, N. Liu, F. Lin, S. Yang, J. Ning, Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation, Comput. Electron. Agric. 197 (2022) 106991, https://doi.org/10.1016/j.compag.2022.106991.
[102] K.-W. Hsieh, B.-Y. Huang, K.-Z. Hsiao, Y.-H. Tuan, F.-P. Shih, L.-C. Hsieh, S. Chen, I.-C. Yang, Fruit maturity and location identification of beef tomato using R-CNN and binocular imaging technology, J. Food Meas. Charact. 15 (2021) 5170–5180, https://doi.org/10.1007/s11694-021-01074-7.
[103] O. Mirbod, D. Choi, P.H. Heinemann, R.P. Marini, L. He, On-tree apple fruit size estimation using stereo vision with deep learning-based occlusion handling, Biosyst. Eng. 226 (2023) 27–42, https://doi.org/10.1016/j.biosystemseng.2022.12.008.
[104] J. Luo, D. Zhang, L. Luo, T. Yi, PointResNet: a grape bunches point cloud semantic segmentation model based on feature enhancement and improved PointNet++, Comput. Electron. Agric. 224 (2024) 109132 https://doi.org/10.1016/j.compag.2024.109132.
[105] T. Fang, W. Chen, L. Han, Location research and picking experiment of an apple-picking robot based on improved mask R-CNN and binocular vision, Horticulturae 11 (2025) 801, https://doi.org/10.3390/horticulturae11070801.
[106] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788, https://doi.org/10.1109/cvpr.2016.91.
[107] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517–6525, https://doi.org/10.1109/cvpr.2017.690.
[108] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, arXiv 1804 (2018) 02767. https://arxiv.org/abs/1804.02767.
[109] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, X. Wei, YOLOv6: a single-stage object detection framework for industrial applications, arXiv 2209 (2022) 02976. https://arxiv.org/abs/2209.02976.
[110] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: trainable bag-of-freebies sets new State-of-the-art for real-time object detectors, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475, https://doi.org/10.1109/cvpr52729.2023.00721.
[111] Ultralytics, Ultralytics YOLOv8 (2023). https://docs.ultralytics.com/models/yolov8/.
[112] C.-Y. Wang, I.-H. Yeh, H.-Y.M. Liao, YOLOv9: learning what you want to learn using programmable gradient information, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, Springer Nature Switzerland, 2025, pp. 1–21, https://doi.org/10.1007/978-3-031-72751-1_1.
[113] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, YOLOv10: real-time end-to-end object detection, Adv. Neural Inf. Process Syst. 37 (2024) 107984–108011, https://doi.org/10.52202/079017-3429.
[114] J. Lu, W. Chen, Y. Lan, X. Qiu, J. Huang, H. Luo, Design of citrus peel defect and fruit morphology detection method based on machine vision, Comput. Electron. Agric. 219 (2024) 108721, https://doi.org/10.1016/j.compag.2024.108721.
[115] J. Liu, G. Zhao, S. Liu, Y. Liu, H. Yang, J. Sun, Y. Yan, G. Fan, J. Wang, H. Zhang, New progress in intelligent picking: online detection of apple maturity and fruit diameter based on machine vision, Agronomy 14 (2024) 721, https://doi.org/10.3390/agronomy14040721.
[116] J. Zhou, Y. Zhang, J. Wang, RDE-YOLOv7: an improved model based on YOLOv7 for better performance in detecting dragon fruits, Agronomy 13 (2023) 1042, https://doi.org/10.3390/agronomy13041042.
[117] L. Ma, L. Zhao, Z. Wang, J. Zhang, G. Chen, Detection and counting of small target apples under complicated environments by using improved YOLOv7-tiny, Agronomy 13 (2023) 1419, https://doi.org/10.3390/agronomy13051419.
[118] A. Wang, W. Qian, A. Li, Y. Xu, J. Hu, Y. Xie, L. Zhang, NVW-YOLOv8s: an improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages, Comput. Electron. Agric. 219 (2024) 108833, https://doi.org/10.1016/j.compag.2024.108833.
[119] Z. Ma, N. Dong, J. Gu, H. Cheng, Z. Meng, X. Du, STRAW-YOLO: a detection method for strawberry fruits targets and key points, Comput. Electron. Agric. 230 (2025) 109853, https://doi.org/10.1016/j.compag.2024.109853.
[120] J. Qin, Z. Chen, Y. Zhang, J. Nie, T. Yan, B. Wan, YOLO-CT: a method based on improved YOLOv8n-pose for detecting multi-species mature cherry tomatoes and locating picking points in complex environments, Measurement 254 (2025) 117954, https://doi.org/10.1016/j.measurement.2025.117954.
[121] L. Zhou, S. Jin, J. Wang, H. Zhang, M. Shi, H. Zhou, 3D positioning of camellia oleifera fruit-grabbing points for robotic harvesting, Biosyst. Eng. 246 (2024) 110–121, https://doi.org/10.1016/j.biosystemseng.2024.07.019.
[122] L. Ye, J. Ma, Y. Lv, Z. Guo, Z. Lai, C. Ou, J. Li, F. Wu, The YOLO-OBB-based approach for citrus fruit stem pose estimation and robot picking, Agriculture 15 (2025) 2330, https://doi.org/10.3390/agriculture15222330.
[123] S. Liu, T. Zhao, J. Xue, P. Lv, R. Liu, Y. Zhang, W. Gao, H. Sun, T. Zhang, Binocular localization method for pear-picking robots based on YOLO-CDS and RSIQR module collaborative optimization, Comput. Electron. Agric. 239 (2025) 110962, https://doi.org/10.1016/j.compag.2025.110962.
[124] Q. Liang, W. Zhu, J. Long, Y. Wang, W. Sun, W. Wu, A real-time detection framework for on-tree mango based on SSD network, in: Z. Chen, A. Mendes, Y. Yan, S. Chen (Eds.), Intelligent Robotics and Applications, Springer International Publishing, 2018, pp. 423–436, https://doi.org/10.1007/978-3-319-97589-4_36.
[125] B. Arad, P. Kurtser, E. Barnea, B. Harel, Y. Edan, O. Ben-Shahar, Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. The case study of sweet pepper robotic harvesting, Sensors 19 (2019) 1390, https://doi.org/10.3390/s19061390.
[126] T. Yuan, L. Lv, F. Zhang, J. Fu, J. Gao, J. Zhang, W. Li, C. Zhang, W. Zhang, Robust cherry tomatoes detection algorithm in Greenhouse scene based on SSD, Agriculture 10 (2020) 160, https://doi.org/10.3390/agriculture10050160.
[127] J. Vasconez, J. Delpiano, S. Vougioukas, F. Auat Cheein, Comparison of convolutional neural networks in fruit detection and counting: a comprehensive evaluation, Comput. Electron. Agric. 173 (2020) 105348, https://doi.org/10.1016/j.compag.2020.105348.
[128] Z. Zhou, Z. Song, L. Fu, F. Gao, R. Li, Y. Cui, Real-time kiwifruit detection in orchard using deep learning on Android™ smartphones for yield estimation, Comput. Electron. Agric. 179 (2020) 105856, https://doi.org/10.1016/j.compag.2020.105856.
[129] J. Qiang, W. Liu, X. Li, P. Guan, Y. Du, B. Liu, G. Xiao, Detection of citrus pests in double backbone network based on single shot multibox detector, Comput. Electron. Agric. 212 (2023) 108158, https://doi.org/10.1016/j.compag.2023.108158.
[130] Z.-B. Yin, F.-Y. Liu, H. Geng, Y.-J. Xi, D.-B. Zeng, C.-J. Si, M.-D. Shi, A high-precision jujube disease spot detection based on SSD during the sorting process, PLoS One 19 (2024) e0296314, https://doi.org/10.1371/journal.pone.0296314.
[131] K. Zhu, J. Li, K. Zhang, C. Arunachalam, S. Bhattacharya, R. Lu, Z. Li, Foundation model-based apple ripeness and size estimation for selective harvesting, Comput.

Electron. Agric. 236 (2025) 110407, https://doi.org/10.1016/j.compag.2025.110407.
[132] Y. Chen, Y. Niu, W. Cheng, L. Zheng, D. Sun, Apple detection method in the natural environment based on improved YOLOv5, INMATEH Agric. Eng. 72 (2024) 183–192, https://doi.org/10.35633/inmateh-72-17.
[133] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Qin, Y. Zheng, X. Qiu, X. Huang, Q. Zhang, T. Gui, The rise and potential of large language model based agents: a survey, Sci. China Inf. Sci. 68 (2025) 121101, https://doi.org/10.1007/s11432-024-4222-0.
[134] Y. Xu, X. Wang, H. Zhang, H. Lin, SE-Swin: an improved Swin-Transfomer network of self-ensemble feature extraction framework for image retrieval, IET Image Process. 18 (2024) 13–21, https://doi.org/10.1049/ipr2.12929.
[135] D. Seo, S.K. Lee, J.G. Kim, I.-S. Oh, High-precision peach fruit segmentation under adverse conditions using Swin transformer, Agriculture 14 (2024) 903, https://doi.org/10.3390/agriculture14060903.
[136] H. Zheng, G. Wang, X. Li, Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron, J. Food Meas. Charact. 16 (2022) 2789–2800, https://doi.org/10.1007/s11694-022-01396-0.
[137] Z. Gu, D. He, J. Huang, J. Chen, X. Wu, B. Huang, T. Dong, Q. Yang, H. Li, Simultaneous detection of fruits and fruiting stems in mango using improved YOLOv8 model deployed by edge device, Comput. Electron. Agric. 227 (2024) 109512, https://doi.org/10.1016/j.compag.2024.109512.
[138] I. Neji, N. Ben Aoun, N. Boujnah, R. Ejbali, DenseViT-XGB: a hybrid approach for dates varieties identification, Neurocomputing 596 (2024) 127976, https://doi.org/10.1016/j.neucom.2024.127976.
[139] Q. Rong, C. Hu, X. Hu, M. Xu, Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing, Comput. Electron. Agric. 210 (2023) 107923, https://doi.org/10.1016/j.compag.2023.107923.
[140] M. Culman, S. Delalieux, B. Beusen, B. Somers, Automatic labeling to overcome the limitations of deep learning in applications with insufficient training data: a case study on fruit detection in pear orchards, Comput. Electron. Agric. 213 (2023) 108196, https://doi.org/10.1016/j.compag.2023.108196.
[141] C. Liu, W. Min, J. Song, Y. Yang, G. Sheng, T. Yao, L. Wang, S. Jiang, Channel grouping vision transformer for lightweight fruit and vegetable recognition, Expert Syst. Appl. 292 (2025) 128636, https://doi.org/10.1016/j.eswa.2025.128636.
[142] Q. Sun, M. Zhong, X. Chai, Z. Zeng, H. Yin, G. Zhou, T. Sun, Citrus pose estimation from an RGB image for automated harvesting, Comput. Electron. Agric. 211 (2023) 108022, https://doi.org/10.1016/j.compag.2023.108022.
[143] S. Jin, L. Zhou, H. Zhou, CO-YOLO: a lightweight and efficient model for Camellia oleifera fruit object detection and posture determination, Comput. Electron. Agric. 235 (2025) 110394, https://doi.org/10.1016/j.compag.2025.110394.
[144] X. Yang, L. Bai, T. Zhang, R. Wu, A monocular pose estimation framework for automatic dragon fruit harvesting using navel and stem keypoints, Horticulturae 12 (2026) 505, https://doi.org/10.3390/horticulturae12040505.
[145] H. Li, Q. Zhu, M. Huang, Y. Guo, J. Qin, Pose estimation of sweet pepper through symmetry axis detection, Sensors 18 (2018) 3083, https://doi.org/10.3390/s18093083.
[146] J. Chen, T. Yu, M. Lou, X. Xu, H. Jiang, G. Ren, Tomato harvesting pose estimation based on fruit-stem spatial relationship, Comput. Electron. Agric. 250 (2026) 111913, https://doi.org/10.1016/j.compag.2026.111913.
[147] J. Kim, H. Pyo, I. Jang, J. Kang, B. Ju, K. Ko, Tomato harvesting robotic system based on Deep-ToMaToS: deep learning network using transformation loss for 6D pose estimation of maturity classified tomatoes with side-stem, Comput. Electron. Agric. 201 (2022) 107300, https://doi.org/10.1016/j.compag.2022.107300.
[148] Y. Yu, K. Zhang, H. Liu, L. Yang, D. Zhang, Real-time visual localization of the picking points for a ridge-planting strawberry harvesting robot, IEEE Access 8 (2020) 116556–116568, https://doi.org/10.1109/access.2020.3003034.
[149] X. Xiao, Y. Wang, Y. Jiang, H. Wu, B. Zhou, Monocular pose estimation method for automatic citrus harvesting using semantic segmentation and rotating target detection, Foods 13 (2024) 2208, https://doi.org/10.3390/foods13142208.
[150] J. Zhang, J. Xie, F. Zhang, J. Gao, C. Yang, C. Song, W. Rao, Y. Zhang, Greenhouse tomato detection and pose classification algorithm based on improved YOLOv5, Comput. Electron. Agric. 216 (2024) 108519, https://doi.org/10.1016/j.compag.2023.108519.
[151] T. Kim, D.-H. Lee, K.-C. Kim, Y.-J. Kim, 2D pose estimation of multiple tomato fruit-bearing systems for robotic harvesting, Comput. Electron. Agric. 211 (2023) 108004, https://doi.org/10.1016/j.compag.2023.108004.
[152] H. Liu, Y. Liang, Q. He, B. Li, W. Wang, H. He, Y. Xu, Y. Yao, X. Cao, Y. Yin, X. Duan, T. Pang, StrawPose-lite: a lightweight pose network for strawberry picking point prediction on edge devices, Agriculture 16 (2026) 1185, https://doi.org/10.3390/agriculture16111185.
[153] J. Xiong, Z. He, R. Lin, Z. Liu, R. Bu, Z. Yang, H. Peng, X. Zou, Visual positioning technology of picking robots for dynamic litchi clusters with disturbance, Comput. Electron. Agric. 151 (2018) 226–237, https://doi.org/10.1016/j.compag.2018.06.007.
[154] J. Gené-Mola, R. Sanz-Cortiella, J.R. Rosell-Polo, J.-R. Morros, J. Ruiz-Hidalgo, V. Vilaplana, E. Gregorio, Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry, Comput. Electron. Agric. 169 (2020) 105165, https://doi.org/10.1016/j.compag.2019.105165.
[155] S. Zheng, Y. Liu, W. Weng, X. Jia, S. Yu, Z. Wu, Tomato recognition and localization method based on improved YOLOv5n-seg model and Binocular stereo vision, Agronomy 13 (2023) 2339, https://doi.org/10.3390/agronomy13092339.
[156] N. Guo, B. Zhang, J. Zhou, K. Zhan, S. Lai, Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning, Comput. Electron. Agric. 179 (2020) 105818, https://doi.org/10.1016/j.compag.2020.105818.
[157] L. Luo, W. Yin, Z. Ning, J. Wang, H. Wei, W. Chen, Q. Lu, In-field pose estimation of grape clusters with combined point cloud segmentation and geometric analysis, Comput. Electron. Agric. 200 (2022) 107197, https://doi.org/10.1016/j.compag.2022.107197.
[158] X. Du, Z. Meng, Z. Ma, W. Lu, H. Cheng, Tomato 3D pose detection algorithm based on keypoint detection and point cloud processing, Comput. Electron. Agric. 212 (2023) 108056, https://doi.org/10.1016/j.compag.2023.108056.
[159] J. Niu, M. Bi, Q. Yu, Apple pose estimation based on SCH-YOLO11s segmentation, Agronomy 15 (2025) 900, https://doi.org/10.3390/agronomy15040900.
[160] Y. Zhuang, Y. Wang, L. Zheng, J. Dai, H. Liu, J. Zhu, Z. Cui, A detection method for tomato pose estimation and grasping point localization in robotic harvesting based on YOLOv8s-ECC, Horticulturae 12 (2026) 369, https://doi.org/10.3390/horticulturae12030369.
[161] T. Zhu, Z. Miao, C. Zhao, L. Chen, Y. Xiong, Five-degree-of-freedom strawberry pose estimation using key points and oriented bounding box detection, Comput. Electron. Agric. 239 (2025) 110974, https://doi.org/10.1016/j.compag.2025.110974.
[162] D. Wang, X. Wang, Y. Shi, X. Zhang, Y. Chen, J. Zheng, N. Liu, A lightweight keypoint detection model-based method for strawberry recognition and picking point localization in multi-occlusion scenes, Artif. Intell. Agric. 16 (2026) 316–341, https://doi.org/10.1016/j.aiia.2025.10.009.
[163] A.B. Alterani, M. Costanzo, M. De Simone, S. Federico, C. Natale, Experimental comparison of two 6D pose estimation algorithms in robotic fruit-picking tasks, Robotics 13 (2024) 127, https://doi.org/10.3390/robotics13090127.
[164] C. Chen, Z. Zhou, R. Zhang, F. Yang, Z. Wang, Spatial pose estimation algorithm for orchard obscured apple for picking robots, Expert Syst. Appl. 333 (2027) 133911, https://doi.org/10.1016/j.eswa.2026.133911.
[165] L. Dong, L. Zhu, B. Zhao, R. Wang, J. Ni, S. Liu, K. Chen, X. Cui, L. Zhou, Semantic segmentation-based observation pose estimation method for tomato harvesting robots, Comput. Electron. Agric. 230 (2025) 109895, https://doi.org/10.1016/j.compag.2025.109895.
[166] Y. Yu, K. Zhang, L. Yang, D. Zhang, Fruit detection for strawberry harvesting robot in non-structural environment based on mask-RCNN, Comput. Electron. Agric. 163 (2019) 104846, https://doi.org/10.1016/j.compag.2019.06.001.
[167] Y. Bai, S. Mao, J. Zhou, B. Zhang, Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting, Precis. Agric. 24 (2023) 727–743, https://doi.org/10.1007/s11119-022-09972-6.
[168] X. Qi, J. Dong, Y. Lan, H. Zhu, Method for identifying Litchi picking position based on YOLOv5 and PSPNet, Remote Sens. (Basel) 14 (2022) (2004) 2004, https://doi.org/10.3390/rs14092004.
[169] X. Chen, G. Dong, X. Fan, Y. Xu, T. Liu, J. Zhou, H. Jiang, Fruit stalk recognition and picking point localization of new plums based on improved DeepLabv3+, Agriculture 14 (2024) 2120, https://doi.org/10.3390/agriculture14122120.
[170] B. Zhang, Y. Xia, R. Wang, Y. Wang, C. Yin, M. Fu, W. Fu, Recognition of mango and location of picking point on stem based on a multi-task CNN model named YOLOMS, Precis. Agric. 25 (2024) 1454–1476, https://doi.org/10.1007/s11119-024-10119-y.
[171] L. Li, K. Li, Z. He, H. Li, Y. Cui, Kiwifruit segmentation and identification of picking point on its stem in orchards, Comput. Electron. Agric. 229 (2025) 109748, https://doi.org/10.1016/j.compag.2024.109748.
[172] Y. Liao, G. Huang, J. Li, X. Zhou, C. Wu, C. Liu, PLYS-Longan: a picking point localization model for Longan in natural environments, Agriculture 16 (2026) 789, https://doi.org/10.3390/agriculture16070789.
[173] X. Lin, J. Wang, J. Wang, H. Wei, M. Chen, L. Luo, Picking point localization method based on semantic reasoning for complex picking scenarios in vineyards, Artif. Intell. Agric. 15 (2025) 744–756, https://doi.org/10.1016/j.aiia.2025.05.004.
[174] H. Su, S. Wang, H. Su, F. Ma, Y. Li, J. Li, YOLOv8n-Pose-DSW: a precision picking point localization model for zucchini in complex greenhouse environments, Agriculture 15 (2025) 1954, https://doi.org/10.3390/agriculture15181954.
[175] Y. Huang, G. Li, J. Li, H. Chen, H. Lin, C. Yang, R. Chen, Accurate localization of fruit targets and picking points with multi-dimensional attention and dynamic upsampling, Comput. Electron. Agric. 240 (2026) 111211, https://doi.org/10.1016/j.compag.2025.111211.
[176] P. Li, J. Chen, Q. Chen, L. Huang, Z. Jiang, W. Hua, Y. Li, Detection and picking point localization of grape bunches and stems based on oriented bounding box, Comput. Electron. Agric. 233 (2025) 110168, https://doi.org/10.1016/j.compag.2025.110168.
[177] J. Qing, Z. Zeng, J. Chen, X. Liang, B. Mo, Y. Liu, A. Zhou, S. Wu, Fruit-stem structural visual perception for picking point localization in strawberry harvesting robots, Front. Plant Sci. 17 (2026) 1894265, https://doi.org/10.3389/fpls.2026.1894265.
[178] T. Li, W. Fang, G. Zhao, F. Gao, Z. Wu, R. Li, L. Fu, J. Dhupia, An improved binocular localization method for apple based on fruit detection using deep learning, Inf. Process. Agric. 10 (2023) 276–287, https://doi.org/10.1016/j.inpa.2021.12.003.
[179] G. Song, J. Wang, R. Ma, Y. Shi, Y. Wang, Study on the fusion of improved YOLOv8 and depth camera for bunch tomato stem picking point recognition and localization, Front. Plant Sci. 15 (2024) 1447855, https://doi.org/10.3389/fpls.2024.1447855.

[180] Y. Liang, W. Jiang, Y. Liu, Z. Wu, R. Zheng, Picking-point localization algorithm for citrus fruits based on improved YOLOv8 model, Agriculture 15 (2025) 237, https://doi.org/10.3390/agriculture15030237.
[181] P. Li, M. Wen, Z. Zeng, Y. Tian, Cherry tomato bunch and picking point detection for robotic harvesting using an RGB-D sensor and a StarBL-YOLO network, Horticulturae 11 (2025) 949, https://doi.org/10.3390/horticulturae11080949.
[182] Y. Huang, Y. Zhong, J. Li, H. Chen, G. Li, R. Chen, Pepper-Det: a real-time framework for automated chili pepper detection and harvesting point localization, Smart Agric. Technol. 13 (2026) 101860, https://doi.org/10.1016/j.atech.2026.101860.
[183] H. Wang, G. Zhang, H. Cao, K. Hu, Q. Wang, Y. Deng, J. Gao, Y. Tang, Geometry-aware 3D point cloud learning for precise cutting-point detection in unstructured field environments, J. Field Robot. 42 (2025) 3063–3076, https://doi.org/10.1002/rob.22567.
[184] G. Zhao, S. Dong, J. Wen, Y. Ban, X. Zhang, Selective fruit harvesting prediction and 6D pose estimation based on YOLOv7 multi-parameter recognition, Comput. Electron. Agric. 229 (2025) 109815, https://doi.org/10.1016/j.compag.2024.109815.
[185] J.-L. Duan, L.-Q. Lai, Z. Yang, Z.-J. Luo, H.-T. Yuan, Multi-feature language-image model for fruit quality image classification, Comput. Electron. Agric. 227 (2024) 109462, https://doi.org/10.1016/j.compag.2024.109462.
[186] G. Huang, Z. Xu, L. Liu, J. Chen, Z. He, F. Liu, Evaluating and deploying large vision-language models for fruit quality assessment in smart agriculture systems, Comput. Electron. Agric. 238 (2025) 110806, https://doi.org/10.1016/j.compag.2025.110806.
[187] C.F. Lacerda, Y. Ampatzidis, A. de Oliveira Costa Neto, V. Partel, Cost-efficient high-resolution monitoring for specialty crops using AgI-GAN and AI-driven analytics, Comput. Electron. Agric. 237 (2025) 110678, https://doi.org/10.1016/j.compag.2025.110678.
[188] S. Mehta, T. Burks, Vision-based control of robotic manipulator for citrus harvesting, Comput. Electron. Agric. 102 (2014) 146–158, https://doi.org/10.1016/j.compag.2014.01.003.
[189] S. Mehta, W. MacKunis, T. Burks, Robust visual servo control in the presence of fruit motion for robotic citrus harvesting, Comput. Electron. Agric. 123 (2016) 362–375, https://doi.org/10.1016/j.compag.2016.03.007.
[190] D.M. Bulanon, C. Burr, M. DeVlieg, T. Braddock, B. Allen, Development of a visual servo system for robotic fruit harvesting, AgriEngineering 3 (2021) 840–852, https://doi.org/10.3390/agriengineering3040053.
[191] M. Chen, Z. Chen, L. Luo, Y. Tang, J. Cheng, H. Wei, J. Wang, Dynamic visual servo control methods for continuous operation of a fruit harvesting robot working throughout an orchard, Comput. Electron. Agric. 219 (2024) 108774, https://doi.org/10.1016/j.compag.2024.108774.
[192] J. Kan, Y. Wu, R. Dong, S. Yao, X. Zhao, T. Zou, B. Kang, J. Li, A progressive hybrid automatic switching visual servoing method for apple-picking robots, Agriculture 16 (2026) 620, https://doi.org/10.3390/agriculture16050620.
[193] J. Rong, Z. Xiong, C. Xue, J. Yuan, C. Zhang, P. Wang, T. Yuan, W. Li, Field validation of vision-guided action policy generation for robotic tomato harvesting, Inf. Process. Agric. (2026), https://doi.org/10.1016/j.inpa.2026.05.008.
[194] C. Yin, J. Huang, Y. Xia, H. Zheng, W. Fu, B. Zhang, Design, development, integration and field evaluation of a dual robotic arm mango harvesting robot, J. Field Robot. 42 (2025) 3705–3725, https://doi.org/10.1002/rob.22580.
[195] C. Wang, C. Li, Q. Han, F. Wu, X. Zou, A performance analysis of a Litchi picking robot system for actively removing obstructions, using an artificial intelligence algorithm, Agronomy 13 (2023) 2795, https://doi.org/10.3390/agronomy13112795.
[196] D.-A. Zhao, J. Lv, W. Ji, Y. Zhang, Y. Chen, Design and control of an apple harvesting robot, Biosyst. Eng. 110 (2011) 112–122, https://doi.org/10.1016/j.biosystemseng.2011.07.005.
[197] X. Xiao, Y. Wang, B. Zhou, Y. Jiang, Flexible hand claw picking method for citrus-picking robot based on target fruit recognition, Agriculture 14 (2024) 1227, https://doi.org/10.3390/agriculture14081227.
[198] Y. Zhang, H. Xie, D. Zhang, L. Yang, Z. Lu, T. Cui, X. He, K. Zhang, Design and experimental study of ridge-grown strawberry automatic harvesting robot, Comput. Electron. Agric. 240 (2026) 111112, https://doi.org/10.1016/j.compag.2025.111112.
[199] S. Parsa, B. Debnath, M.A. Khan, A.E. Ghalamzan, Modular autonomous strawberry picking robotic system, J. Field Robot. 41 (2024) 2226–2246, https://doi.org/10.1002/rob.22229.
[200] Z. Wang, K. Lou, B. Zhang, Y. Gu, Q. Xu, W. Fu, Compliant picking control of dragon fruit picking robot based on adaptive variable impedance, Biosyst. Eng. 252 (2025) 126–143, https://doi.org/10.1016/j.biosystemseng.2025.02.014.
[201] W. Ji, G. He, B. Xu, H. Zhang, X. Yu, A new picking pattern of a flexible three-fingered end-effector for apple harvesting robot, Agriculture 14 (2024) 102, https://doi.org/10.3390/agriculture14010102.
[202] T. Wang, W. Du, L. Zeng, L. Su, Y. Zhao, F. Gu, L. Liu, Q. Chi, Design and testing of an end-effector for tomato picking, Agronomy 13 (2023) 947, https://doi.org/10.3390/agronomy13030947.
[203] J. Fang, N. Dai, Z. Xin, J. Yuan, X. Liu, Design and interaction dynamics analysis of a novel hybrid bending-twisting-pulling end-effector for robotic tomato picking, Comput. Electron. Agric. 231 (2025) 110011, https://doi.org/10.1016/j.compag.2025.110011.
[204] F. Zhang, Z. Chen, Y. Wang, R. Bao, X. Chen, S. Fu, M. Tian, Y. Zhang, Research on flexible end-effectors with humanoid grasp function for small spherical fruit picking, Agriculture 13 (2023) 123, https://doi.org/10.3390/agriculture13010123.
[205] H. Huang, R. Wang, F. Huang, J. Chen, Analysis and realization of a self-adaptive grasper grasping for non-destructive picking of fruits and vegetables, Comput. Electron. Agric. 232 (2025) 110119, https://doi.org/10.1016/j.compag.2025.110119.
[206] Y. Park, J. Seol, J. Pak, Y. Jo, J. Jun, H.I. Son, A novel end-effector for a fruit and vegetable harvesting robot: mechanism and field experiment, Precis. Agric. 24 (2023) 948–970, https://doi.org/10.1007/s11119-022-09981-5.
[207] W. Hua, W. Zhang, Z. Zhang, X. Liu, M. Huang, C. Igathinathane, S. Vougioukas, C.K. Saha, N. Mustafa, D.S. Salama, Y. Zhang, M. Zhang, Vacuum suction end-effector development for robotic harvesters of fresh market apples, Biosyst. Eng. 249 (2025) 28–40, https://doi.org/10.1016/j.biosystemseng.2024.11.011.
[208] Q. Vu, A. Ronzhin, A model of four-finger gripper with a built-in vacuum suction nozzle for harvesting tomatoes, in: Proceedings of 14th International Conference on Electromechanics and Robotics "Zavalishin's Readings", Springer, Singapore, 2020, pp. 149–160, https://doi.org/10.1007/978-981-13-9267-2_13.
[209] M. Fu, S. Guo, A. Chen, R. Cheng, X. Cui, Design and experimentation of multi-fruit envelope-cutting kiwifruit picking robot, Front. Plant Sci. 15 (2024) 1338050, https://doi.org/10.3389/fpls.2024.1338050.
[210] Q. Wang, K. Bai, L. Zhang, Z. Sun, T. Jia, D. Hu, Q. Li, J. Zhang, A. Knoll, H. Jiang, M. Zhou, Y. Ying, Towards reliable and damage-less robotic fragile fruit grasping: an enveloping gripper with multimodal strategy inspired by Asian elephant trunk, Comput. Electron. Agric. 234 (2025) 110198, https://doi.org/10.1016/j.compag.2025.110198.
[211] H. Yin, Q. Sun, X. Ren, J. Guo, Y. Yang, Y. Wei, B. Huang, X. Chai, M. Zhong, Development, integration, and field evaluation of an autonomous citrus-harvesting robot, J. Field Robot. 40 (2023) 1363–1387, https://doi.org/10.1002/rob.22178.
[212] X. Li, Y. Han, N. Ma, Y. Liu, J. Pan, S. Yang, S. Zheng, Autonomous tomato harvesting with top–Down fusion network for limited data, IEEE Trans. Robot. 41 (2025) 3609–3628, https://doi.org/10.1109/tro.2025.3567544.
[213] B. An, T. Choi, U. Kim, Linkage integrated fin ray gripper capable of safe adaptive grasping for tomato harvesting, Comput. Electron. Agric. 232 (2025) 110118, https://doi.org/10.1016/j.compag.2025.110118.
[214] Y. Pu, J. Luo, G. Zhang, M. Deng, S. Ran, S. Xie, Design and experiment of the end effector for citrus picking imitating finger plucking and cutting, Trans. Chin. Soc. Agric. Eng. (Trans. CSAE) 41 (2025) 63–73, https://doi.org/10.11975/j.issn.1002-6819.202506022.
[215] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, J. Hemming, P. Kurtser, O. Ringdahl, T. Tielen, B. van Tuijl, Development of a sweet pepper harvesting robot, J. Field Robot. 37 (2020) 1027–1039, https://doi.org/10.1002/rob.21937.
[216] L. Yang, H. Li, S. Lin, Y. Xiong, High-tolerance soft-rigid gripper for low-damage robotic strawberry harvesting, IEEE Robot. Autom. Lett. 11 (2026) 121–128, https://doi.org/10.1109/lra.2025.3630528.
[217] S. Ansari, M.K. Gohil, Y. Maeda, B. Bhattacharya, Design and development of a new hybrid auxetic gripper system for adaptive precision tomato harvesting, Int. J. Intell. Robot. Appl. 10 (2026) 60–83, https://doi.org/10.1007/s41315-025-00488-6.
[218] K. Guo, H. Chen, Y. Feng, J. Li, Z. Lu, C. Yang, A humanoid TacTip gripper with SSIM-CNN recognition for strawberry harvesting, IEEE Trans. Autom. Sci. Eng. 23 (2026) 10400–10412, https://doi.org/10.1109/tase.2026.3697535.
[219] Z.-L. He, J.-T. Xiong, R. Lin, X. Zou, L.-Y. Tang, Z.-G. Yang, Z. Liu, G. Song, A method of green litchi recognition in natural environment based on improved LDA classifier, Comput. Electron. Agric. 140 (2017) 159–167, https://doi.org/10.1016/j.compag.2017.05.029.
[220] C. Wang, H. Wang, Q. Han, Z. Zhang, D. Kong, X. Zou, Strawberry detection and ripeness classification using YOLOv8+ model and image processing method, Agriculture 14 (2024) 751, https://doi.org/10.3390/agriculture14050751.
[221] J. Wu, J. Zhang, J. Zhu, Y. Duan, Y. Fang, J. Zhu, L. Yin, J. Jiang, Z. He, Y. Huang, Y. Meng, Multi-scale convolution and dynamic task interaction detection head for efficient lightweight plum detection, Food Bioprod. Process. 149 (2025) 353–367, https://doi.org/10.1016/j.fbp.2024.12.007.
[222] H. Li, Z. Yin, Z. Zuo, L. Pan, J. Zhang, Precision citrus segmentation and stem picking point localization using improved YOLOv8n-seg algorithm, Front. Plant Sci. 16 (2025) 1655093, https://doi.org/10.3389/fpls.2025.1655093.
[223] Y. Li, Q. Feng, C. Liu, Z. Xiong, Y. Sun, F. Xie, T. Li, C. Zhao, MTA-YOLACT: multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting, Eur. J. Agron. 146 (2023) 126812, https://doi.org/10.1016/j.eja.2023.126812.
[224] P. Xiang, F. Pan, T. Liu, X. Zhao, M. Hu, D. He, B. Zhang, MixSegNext: a CNN-Transformer hybrid model for semantic segmentation and picking point localization algorithm of Sichuan pepper in natural environments, Comput. Electron. Agric. 237 (2025) 110564, https://doi.org/10.1016/j.compag.2025.110564.
[225] T. Zhang, F. Wu, M. Wang, Z. Chen, L. Li, X. Zou, Grape-bunch identification and location of picking points on occluded fruit axis based on YOLOv5-GAP, Horticulturae 9 (2023) 498, https://doi.org/10.3390/horticulturae9040498.
[226] Z. Zhong, J. Xiong, Z. Zheng, B. Liu, S. Liao, Z. Huo, Z. Yang, A method for litchi picking points calculation in natural environment based on main fruit bearing branch detection, Comput. Electron. Agric. 189 (2021) 106398, https://doi.org/10.1016/j.compag.2021.106398.
[227] H. Yi, K. Song, X. Song, Watermelon detection and Localization technology based on GTR-net and binocular vision, IEEE Sens. J. 24 (2024) 19873–19881, https://doi.org/10.1109/jsen.2024.3393916.

[228] Q. Zhang, W.-H. Su, Real-time recognition and localization of apples for robotic picking based on structural light and deep learning, Smart Cities 6 (2023) 3393–3410, https://doi.org/10.3390/smartcities6060150.
[229] L. Zhou, A. Hu, Y. Cheng, W. Zhang, B. Zhang, X. Lu, Q. Wu, N. Ren, Barrier-free tomato fruit selection and location based on optimized semantic segmentation and obstacle perception algorithm, Front. Plant Sci. 15 (2024) 1460060, https://doi.org/10.3389/fpls.2024.1460060.
[230] M. Mhamed, Z. Zhang, W. Hua, L. Yang, M. Huang, X. Li, T. Bai, H. Li, M. Zhang, Apple varieties and growth prediction with time series classification based on deep learning to impact the harvesting decisions, Comput. Ind. 164 (2025) 104191, https://doi.org/10.1016/j.compind.2024.104191.
[231] J. Pi, J. Liu, K. Zhou, M. Qian, An octopus-inspired bionic flexible gripper for apple grasping, Agriculture 11 (2021) 1014, https://doi.org/10.3390/agriculture11101014.
[232] Z. Mei, Y. Li, R. Zhu, S. Wang, Intelligent fruit localization and grasping method based on YOLO VX model and 3D vision, Agriculture 15 (2025) 1508, https://doi.org/10.3390/agriculture15141508.
[233] Z. He, L. Ma, Y. Wang, Y. Wei, X. Ding, K. Li, Y. Cui, Double-arm cooperation and implementing for harvesting kiwifruit, Agriculture 12 (2022) 1763, https://doi.org/10.3390/agriculture12111763.