Computing and Artificial Intelligence 2025, 3(4), 4296.
doi: https://doi.org/10.59400/cai4296

Review

From bench to field: A systematic review of computer vision for tomato detection in precision agriculture (2018–2025)

Philippe Lyonel Mbouembe Touko[1], **Guoxu Liu**[2,*]

[1] eXsolIT Research Center, Yangsan 50611, Republic of Korea
[2] School of Computer Engineering, Weifang University, Weifang 261061, China
*** Corresponding author:** Guoxu Liu, liuguoxu@wfu.edu.cn

CITATION

Touko PLM, Liu G. From bench to field: A systematic review of computer vision for tomato detection in precision agriculture (2018–2025). Computing and Artificial Intelligence. 2025; 3(4): 4296. https://doi.org/10.59400/cai4296

ARTICLE INFO

Received: 11 October 2025
Revised: 28 November 2025
Accepted: 2 December 2025
Available online: 15 December 2025

Abstract: Accurate tomato detection enables robotic harvesting, crop yield estimation, and tomato quality control, among other agricultural tasks. Despite remarkable advances in computer vision, particularly YOLO models, a significant gap persists between laboratory research and field deployment. This PRISMA-guided review of 110 publications (2018-2025) analyzes the disparity between lab-tested model performance and reliable real-world performance in this context. To quantify current capabilities, we construct a taxonomy based on the sensing platform, task, and environment. Across the reviewed literature, the mean average precision has increased from 78.3% for YOLOv3 to 94.7% for YOLOv11, while a controlled benchmarking study on an identical dataset (LaboroTomato) reveals smaller differences. However, tomato detection performance significantly drops in deployment, with a mean cross-domain performance loss of 8.24% due to occlusion, illumination changes, and weather conditions. Our reproducibility audit shows that most research lacks protocols for model development and that 12% releases make their public code available. Finally, 73% of high-accuracy models have requirements above the popular edge-device sizes commonly used in agricultural robotics. To bridge this implementation gap, we outline: 1) reporting guidelines to promote reproducibility, 2) decision frameworks to translate pragmatic agricultural considerations into concrete technical specifications, and 3) open research directions centered on reliability, cross-domain validation, and real-world deployment. This survey will support practitioners in agriculture, robotics, and machine learning design, deployable computer vision systems for tomatoes and other crops.

Keywords: YOLO; tomato detection; precision agriculture; object detection; systematic review; edge computing; agricultural robotics; model deployment

1. Introduction

1.1. Global context and motivation

Tomato ranks among the highest-volume and most commercially valuable agricultural crops, generating ~192 million tonnes annually [1], and having an estimated worth of over USD 200 billion per year [2]. However, this position in the global marketplace is juxtaposed with growing difficulty in producing tomatoes in a reliable and profitable manner. The rising cost of labor and continued worker shortages combine with increased unpredictability of yield [3], harvest date, and crop quality due to changing climates. Labor can account for about 30–50% of total overheads in many growing operations, and harvest is often the largest labor cost when multiple pickings

are required and fruit must be handled gently to prevent bruising [4].

Automation and data-based crop management tools are being developed to address many of these challenges. A prerequisite for many forms of automated workflows is computer vision: knowing where tomatoes are in an image is critical to robot harvesting [5], counting fruit and predicting yield far in advance of harvest [6], automated sorting and quality detection, and early detection of disease or stress that manifests visually on fruit or canopy. Reliable tomato detection has proven difficult, however, in real-world agricultural settings. Tomatoes are often obscured by leaves, stems, trellising, and other fruit [7]; they can vary widely in appearance based on cultivar and stage of maturity; and lighting conditions can vary substantially throughout the day and between weather events and production type (open field, plastic house, glass greenhouse) [8]. These factors combine to create a challenging "in-the-wild" problem in which accuracy is required at the largest scales while also meeting stringent requirements at the edge for speed and power.

1.2. The role of automated detection in tomato production

Automated tomato detection is not a singular application. Automated detection serves as an enabler for a number of decision/action loops throughout the production process. For example, throughout cultivation, automated detection could provide objective, repeatable counts of fruit presence and maturity distributions to aid in harvest planning, labor needs, and supply chain management. During harvest, detection becomes a safety- and quality-critical perception step: a harvesting robot must localize fruit precisely, avoid damaging the plant, and operate at a practical speed. After harvest, detection and classification can assist with sorting and grading, improving consistency and reducing waste. In parallel, monitoring systems that run on mobile platforms or fixed cameras can identify trends (e.g., fruit set rate, color progression) and deliver alerts under the resource constraints typical of farm deployment [9].

Despite the promise, agricultural deployment imposes conditions that are markedly different from laboratory datasets and controlled imaging [10]. Real farms exhibit heterogeneous backgrounds, motion blur from mobile platforms [11], shadows cast by canopy structure, specular highlights on fruit skin, and sensor contamination from dust or moisture [12]. These realities motivate a research focus that goes beyond "high accuracy on a curated test set" toward models that remain stable across environments and are practical to deploy on edge hardware.

1.3. Why YOLO dominates agricultural computer vision?

Within object detection, the YOLO family has become the most frequently selected framework for tomato detection and, more broadly, agricultural vision tasks [13–18]. YOLO's core appeal is its single-stage, end-to-end design: predictions are produced in one forward pass, enabling real-time inference while maintaining competitive detection accuracy [19–22]. This balance is especially important for agricultural systems that require low latency (often under 100 ms) and must run on power- and memory-constrained hardware such as embedded GPUs, NPUs, or mobile processors [23].

In our survey of agricultural object detection studies published between 2018 and 2025, YOLO-based architectures account for roughly two-thirds of reported implementations, substantially exceeding adoption of two-stage approaches such as Faster R-CNN [24, 25] and other one-stage baselines such as SSD [26, 27]. The dominance is not simply based on the claim of universal supremacy in terms of accuracy. It also represents a pragmatic alignment toward deployability and ecosystem maturity, in that YOLO comes with strong open-source tool support, a strong following, strong support for transfer learning, and fast iteration between versions (e.g., from YOLOv3 to the latest versions). For researchers, these are key points in deciding on the migration from research-stage solutions to deployable solutions.

However, at the same time, large-scale adoption can obscure many details. Different versions of YOLOs have different trade-offs in terms of accuracy versus speed, and results are highly context-dependent in terms of the composition of the data set, the conventions of the annotations, evaluation metrics, and hardware configurations. A review that treats "YOLO" as a monolith risks conflating fundamentally different regimes of performance and deployability.

1.4. The implementation gap in agricultural computer vision

The transition from lab-tested to field-deployed systems faces an implementation barrier for agricultural artificial intelligence systems [28,29]. While many papers show high accuracy in controlled settings (more than 90% mAP), performance degrades under real conditions due to occlusion, illumination changes, and hardware constraints in agricultural settings [30–32]. This review quantifies this gap (Subsection 5.1.6) and its drivers (Subsection 9.2).

1.5. Novel contributions and agricultural significance

This review moves beyond cataloging algorithmic progress to address the translational challenges preventing computer vision adoption in precision agriculture. Our contributions are structured to bridge the implementation gap:

- A systems-level diagnosis: We offer a taxonomy of three dimensions (platform, task, and environment) to position technical choices in the context of agricultural workflows, understand why a certain set of methodologies succeeds in specific production contexts and fails in others.
- A reproducibility and deployment audit: To our knowledge, this is the most comprehensive quantitative audit in YOLO-based tomato detection literature, providing completeness across 110 studies and linking gaps to downstream deployment feasibility. It shows that 68% of studies lack useful methodological information for replication, and that 73% of high-accuracy exceed practical edge hardware limits.
- Actionable frameworks for practitioners: A decision framework that translates technological parameters into agricultural applications is provided in Section 6. This provides a means for engineers and agronomists to choose appropriate technology based on technological parameters.
- Interdisciplinary research roadmap: We outline the milestones (see Section 7) to

emphasize robustness, reproducibility, and validation in the field, thereby aligning research rewards with the needs in agriculture.

Through the assessment of not only successful laboratory performance but also common failures within the real-world domain, this review is critical for interdisciplinary groups aiming to apply computer vision as reliable elements within precision agricultural systems.

As illustrated in **Figure 1**, this review has used a three-dimensional taxonomy to identify the performance gap from the laboratory to the real-world domain.

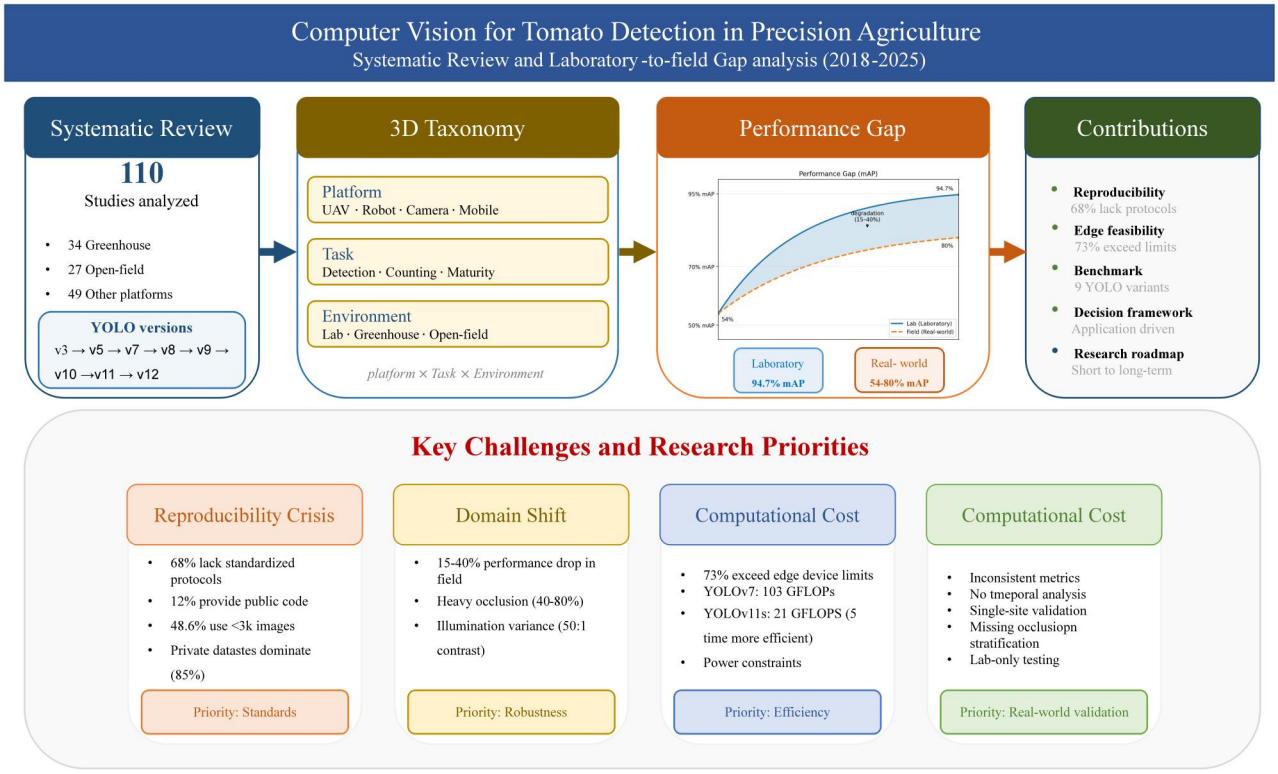

Figure 1. Overview of the systematic review: Scope, taxonomy, identified performance gap between laboratory and field conditions, and key contributions toward bridging the implementation gap in tomato detection for precision agriculture.

1.6. How this review advances the field

Many existing reviews in agricultural computer vision mainly summarize model variants or focus on a single crop or task. However, they often do not quantify the gap between laboratory results and field performance, do not check reproducibility and reporting completeness, and do not evaluate whether methods are feasible on real agricultural hardware. This review advances the field by (1) using a PRISMA-aligned, transparent screening and extraction process, (2) auditing reproducibility and reporting quality across studies, (3) analyzing field-relevant failure drivers such as occlusion, illumination changes, and domain shift, and (4) providing practical guidance through a platform–task–environment taxonomy, a decision framework, and a research roadmap for deployable tomato detection systems.

Table 1 positions this review against representative prior surveys, highlighting our systematic methodology, quantitative audit of reproducibility, and practical guidance for field implementation.

Table 1. Comparative positioning of this review against recent surveys.

Review (year)	Scope	Methodology	Reproducibility analysis	Cross-domain validation	Deployment feasibility	Practitioner guidance
Koirala et al., 2019 [33]	Fruit detection algorithms	Narrative synthesis	Not assessed	Not examined	Qualitative discussion	General recommendations
Kamilaris and Prenafeta-Boldu, 2018 [34]	Broad agricultural AI applications	Descriptive review	Not assessed	Not examined	Mentioned briefly	High-level overview
Tang et al., (2020) Vision-based fruit picking robots [35]	Fruit recognition and localization methods	Literature review	Not assessed	Not examined	Discussed conceptually	System-level considerations
Zhou et al., (2022) Intelligent robots for fruit harvesting [36]	Robotic harvesting technologies and challenges	Comprehensive literature review	Not assessed	Mentioned qualitatively	Hardware-software integration discussed	Technology readiness levels and challenges
This Review	YOLO-based tomato detection with deployment focus	PRISMA-aligned systematic review (N = 110)	Quantitative audit (68% gaps, 12% code availability)	Meta-analysis (N = 9, mean 8.1% drop)	Explicit analysis (73% exceed edge limits; 22% hardware validation)	Application-specific decision framework; stakeholder roadmap

2. Review methodology

A review protocol was defined prior to study selection, including information sources, eligibility criteria, screening procedures, data extraction items, and synthesis plans. Searches covered January 2018 to December 2025 across major bibliographic databases (e.g., Scopus, Web of Science, IEEE Xplore) and complementary sources (e.g., Google Scholar, relevant technical reports) using query blocks for YOLO/object detection, tomato/fruit, agricultural context, and deployment keywords. The PRISMA 2020 checklist and flow diagram are provided to support reporting completeness.

2.1. Eligibility and extraction criteria

Studies were eligible when they satisfied all the following criteria:

❖ **Inclusion criteria:**

- Publication type: Publications can include journal articles and conference proceedings that offer sufficient methodological detail.
- Scope: Tomato detection and localization or any vision task related to tomatoes (such as maturity stage determination, defect recognition, and yield prediction) where object detection is a key component.
- Methods: YOLO-based models (such as YOLOv3 or more advanced models) are used as the main detectors or are incorporated for comparison purposes.
- Time frame: Papers published between January 2018 and December 2025
- Language and access: Full English text is accessible.
- Transparency: sufficient description of the model architecture, dataset characteristics, and evaluation protocol to enable critical appraisal.

❖ **Exclusion criteria:**

- Classification-only studies without spatial localization (no bounding boxes or equivalent localization outputs).
- Non-vision-only modalities (e.g., LiDAR-only; thermal-only; spectral-only

without RGB components).

- Insufficient methodological or results detail (e.g., missing an evaluation protocol or core metrics).
- Non-tomato primary focus with only incidental tomato content.
- Duplicate reporting of identical experiments (the most comprehensive version retained).
- Non-YOLO-only detection frameworks without YOLO comparison (e.g., exclusive R-CNN/SSD/RetinaNet use).

We extracted the YOLO version, architectural modifications, dataset characteristics, capture conditions, annotation protocols, split strategies, evaluation metrics, and threshold settings, along with efficiency indicators (e.g., latency/FPS, parameters, and FLOPs) when reported.

2.2. Study selection and screening process

Study selection followed a PRISMA 2020-aligned, multi-stage screening pipeline as summarized in the PRISMA flow diagram in **Figure 2**:

Stage 1: Identification

A total of 312 records were obtained in the literature search. Automatic removal of duplicates using Zotero, followed by manual checking, led to the removal of 43 duplicates, reducing the total to 269 articles that underwent the screening process.

Stage 2: Title and abstract screening

This stage involved screening titles and abstracts for eligibility by two independent reviewers. Any discrepancies were resolved by consensus, with adjudication by a third party for cases that could not be resolved by consensus. This stage excluded 98 studies (e.g., for reasons of incorrect crop species, lack of detection studies, or lack of original research). The full text of 171 studies was then screened.

Stage 3: Full-text eligibility assessment

A total of 171 reports had their contents evaluated separately by two independent evaluators. A total of 62 reports were discarded for their own reasons, which included a lack of methodological details (n = 23), unavailability or inaccessibility (n = 15), a lack of quantitative performance measures (n = 12), their concentration on non-detection problems (n = 8), and duplicate reporting of experiments (n = 4).

Stage 4: Inclusion

A total of 110 studies satisfied all criteria and were retained in the final synthesis.

2.3. Data synthesis and statistical analysis

The synthesis combined a structured qualitative analysis with quantitative aggregation when outcomes were comparable.

- **Qualitative synthesis**

Studies were grouped by (i) YOLO version family, (ii) dataset/environment type (greenhouse vs. field vs. mixed), (iii) task sub-type (fruit detection vs. maturity/defect/disease-related detection), and (iv) deployment target (edge/mobile/robotics). Architectural innovations were categorized thematically

(attention, multi-scale fusion, lightweight design, loss modifications), and practical deployment constraints were summarized.

- **Quantitative synthesis and meta-analysis**

Given heterogeneity in datasets, metrics, and evaluation protocols, quantitative pooling was performed only when a minimum of three studies reported comparable outcomes under sufficiently similar conditions (e.g., same metric definition and comparable evaluation setup). When pooling was appropriate, a random-effects model was used, heterogeneity was quantified, and subgroup analyses were conducted by environment type and deployment class.

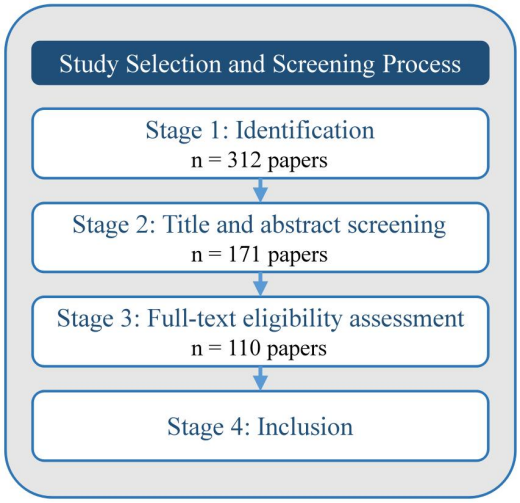

Figure 2. Multi-stage screening pipeline.

3. YOLO architecture evolution relevant to tomato detection

YOLO generations, illustrated in **Figure 3** have evolved from anchor-based multi-scale detectors [13–15] toward increasingly deployment-oriented designs that emphasize efficiency [16–18], improved feature fusion, and robustness [19–22]. Challenges in tomato detection in the context of agricultural settings include the following: small targets, far targets, heavy occlusions within the crop canopy, specular highlights on the tomatoes, and lighting variation. These challenges can be overcome using multi-scale features, careful data augmentation techniques, as well as learning models that are aware of occlusions.

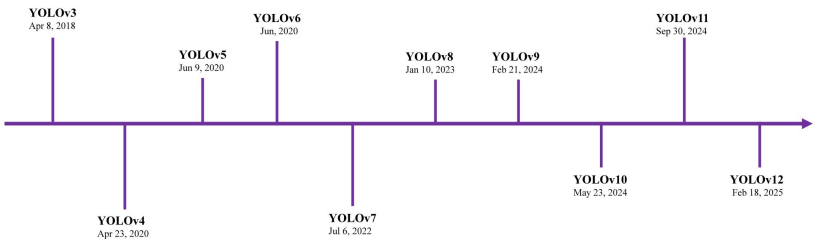

Figure 3. YOLO model timeline.

As illustrated in **Table 2**, the evolution from YOLOv3 to YOLOv12 reflects a clear trajectory toward deployment-optimized architectures. This progression is particularly relevant for tomato detection applications, where the combination of real-time constraints, edge hardware limitations, and challenging agricultural

conditions necessitates careful model selection rather than defaulting to the newest or most accurate variant.

Table 2. YOLO architecture evolution and relevance to tomato detection.

YOLO generation	Key architectural innovations	Relevance to tomato detection and deployment
YOLOv3 [13]	Darknet-53 backbone, multi-scale prediction with anchor boxes	Established baseline; it handles scale variation well but is computationally heavy for edge deployment. It often requires manual anchor tuning for tomato datasets.
YOLOv4/YOLOv5 [14,15]	CSPNet, PANet for feature fusion, advanced augmentation (Mosaic, Mix-Up)	Improved multi-scale feature fusion beneficial for occluded fruits; AutoAnchor in YOLOv5 reduces manual tuning. Streamlined training pipeline accelerated agricultural prototyping.
YOLOv6–YOLOv8 [16–18]	Reparameterization, anchor-free design, efficient decoupled heads, and model scaling	Reduced anchor-tuning dependency; YOLOv8's anchor-free approach better handles irregular fruit shapes and clustering. Enhanced export compatibility with ONNX/TensorRT facilitates deployment.
YOLOv9–YOLOv12 [19–22]	Programmable gradient information (v9), attention mechanisms, edge-first optimizations, and improved feature aggregation	Attention modules (e.g., in YOLOv12), improve detection in cluttered canopies; lightweight variants prioritize efficiency for real-time field deployment. However, fragmentation across implementations requires careful benchmarking.

3.1. From YOLOv3 to YOLOv5: Multi-scale fusion and training refinements

Early agricultural deployments widely adopted YOLOv3 [13], and its successors YOLOv4 [14] and YOLOv5 [15], which enhanced multi-scale feature fusion and training stability. In tomato detection tasks, these models provided strong baseline performance but exhibited sensitivity to dataset-specific anchor configurations and often demanded considerable computational resources for embedded systems. Advanced augmentation techniques such as Mosaic and Mix-Up, introduced in YOLOv4 and YOLOv5, increased robustness to occlusion and illumination variability commonly encountered in greenhouse environments.

3.2. YOLOv6–YOLOv8: Efficiency and improved heads for real-time systems

Later families [16–18], emphasized efficiency through architectural simplification, re-parameterization strategies, and export-friendly implementations. YOLOv8 variants [18] became common baselines due to strong accuracy-speed trade-offs and straightforward deployment workflows (e.g., ONNX/TensorRT), which are particularly valuable for agricultural robotics requiring deterministic inference timing.

3.3. YOLOv9–YOLOv12: Edge-first and robustness-driven trends

Recent work [19–22] increasingly targets edge-first design, including lightweight backbones, attention modules for cluttered scenes, and improved feature aggregation. Since the current models show the most variability across forks as well as the re-implementation models, it is important to have a benchmark before declaring an improvement. This is in line with the rising need for models that are able to effectively process agricultural hardware without compromising on accuracy.

4. Taxonomy of YOLO-based tomato detection studies

We organize the literature using a three-dimensional taxonomy as shown in **Figure 4**: (i) sensing platform (fixed cameras, mobile/edge devices, ground robots/UAVs), (ii) task type (detection, counting/yield estimation, maturity/quality/defect assessment, multi-task), and (iii) environment (controlled/lab, greenhouse, open-field, cross-domain), but not as a neutral filing system. Instead, we use this structure to expose where the field agrees and where it disagrees. A taxonomy that merely categorizes without critiquing risks can create an illusion of consensus where none exists.

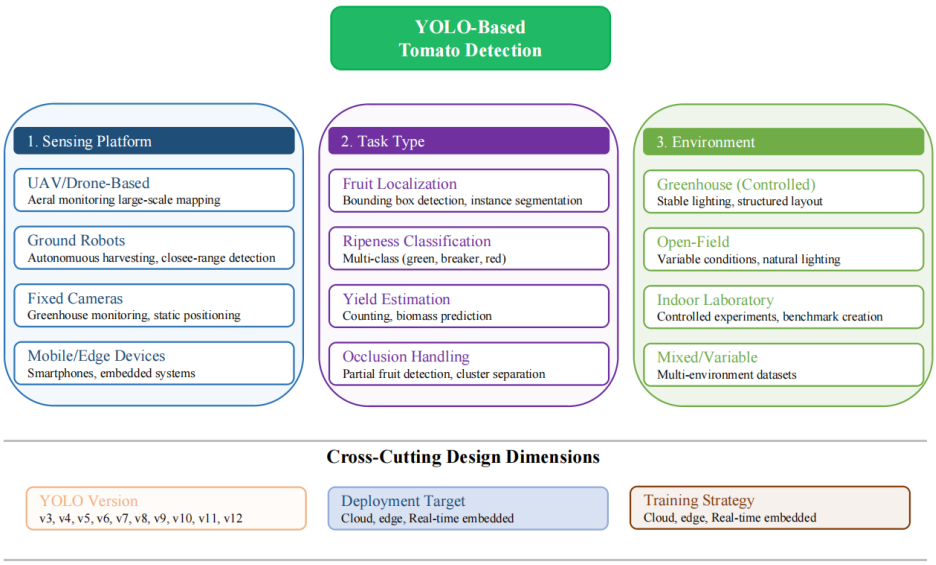

Figure 4. Taxonomy overview.

4.1. By sensing platform

The sensing platform taxonomy reveals distinct deployment strategies for YOLO-based tomato detection systems, with fixed greenhouse cameras dominating the research landscape, while ground robots and mobile devices represent growing areas of practical implementation.

4.1.1. UAV-based detection

It remains severely underexplored in the tomato area, with only one study addressing this platform [37]. This is a significant research gap in the existing body of knowledge regarding the benefits of aerial sensing for large-scale production. The use of unmanned aerial vehicles (UAVs) allows for quick field sensing in a single sweep, sensing multi-hectare areas in a single sweep (approximately 10–100 ha) and providing a bird's-eye view that overcomes the occlusions caused by foliage compared with ground sensors. The seemingly untapped potential of UAV platforms seems to stem from a set of difficulties, namely the motion blur caused by the movement of the aerial sensing platform (moving at speeds of 2–5 m/s), the varied height of the aerial sensing platform, which results in a large factor of scale variation in the sensed images of the tomatoes, and the difficulty of achieving a high-resolution sensing pattern under the strict constraints of the aerial sensing platform. The power consumption (approximately

10–15 W) and weight (approximately 100–200 g) of the sensing platforms further limit the capabilities of the computations for real-time accurate sensing on these platforms. Crucially, none of the surveyed UAV platforms have made downloadable datasets for tomato and model parameters available for cross-platform comparisons for the novel sensing modality explored in this study. Despite the difficulties, a great deal of research potential remains in the aerial sensing modality, specifically in the area of outdoor agriculture, where sensing on the ground is time-consuming and does not provide a full field view.

4.1.2. Ground robots

Ground robots dominate harvesting-oriented research, yet a central contradiction undermines the literature: while occlusion is repeatedly identified as the primary failure mode (60 to 80% are partially hidden), less than 20% of studies [4, 5, 12, 38–43] explicitly validate their models on occlusion-heavy test subsets [44–49]. Moreover, claims about RGB-D fusion improving detection are inconsistent. Rong et al. [42] report a 12% point recall gain under dense foliage using depth information, whereas Li et al. [4] find no statistically significant improvement over RGB alone when occlusion exceeds 70%. This discrepancy is never reconciled, but it likely stems from unreported differences in depth sensor quality (Intel Real-sense versus structured light), fruit cluster density [50–53], and whether the depth channel was aligned during training.

A second contradiction concerns real-time feasibility. Nearly all ground-robot papers state a latency requirement below 100 ms, yet only 22% report end-to-end latency on actual robot hardware (most use desktop GPUs without I/O or control overhead). Among those that do, reported latencies range from 45 ms to 180 ms for similar YOLO versions, a factor of four spread that suggests inconsistent measurement protocols rather than genuine architectural differences. Until the community standardizes hardware reporting (device, precision, batch size, pre-processing time) , the claim of suitability for real-time harvesting remains unverifiable.

4.1.3. Fixed greenhouse cameras

Fixed cameras in greenhouses account for the largest segment, thus establishing the suitability of the technology in the controlled environments found in greenhouses [6–10, 12, 28–30, 53–58]. The key benefits include high resolution without the distortion caused by motion, while the ability to perform temporal tasks using robotic control is a major drawback for applications that require multi-view aggregation [59–64]. Across days or weeks for yield estimation and maturity progression [65–70]. However, this platform also has persistent limitations: the fixed viewpoint can leave a substantial portion of fruits never fully visible (often around 30–50%) because of canopy structure and self-occlusion [71–76], performance can be affected by diurnal lighting variation even in greenhouses [77–82], and lens condensation under high humidity can degrade image quality and detection reliability [83–88]. Representative studies include comprehensive detection frameworks [72] maturity classification systems [60], and multi-stage recognition approaches [28], reinforcing the platform's dominance while highlighting the need for solutions that mitigate viewpoint and environmental artifacts. Despite being the

most studied platform [89–93], fewer than 15% of fixed-camera studies released their datasets publicly, contributing to benchmark fragmentation and slowing community progress.

4.1.4. Mobile/edge devices

Mobile and edge devices have emerged as a practical platform category because of ubiquitous smartphone access and the need for lightweight, easily deployable tools for farmers [6, 11, 23, 32, 94–99]. A key strength of this platform is that phones can provide contextual sensing, such as GPS/IMU data for geo-tagging observations and stabilizing capture, while enabling interactive farmer feedback (e.g., confirming detection and correcting labels) that can improve scouting workflows [100–104]. In practice, crop-scouting applications reported ~85–92% accuracy with trained users, indicating strong potential for real-world adoption. At the same time, deployment is constrained by extreme hardware heterogeneity across devices (roughly <2 GB to 8 GB RAM), tight power limits typical of mobile inference (often <2 W sustained), and variable image quality arising from user-dependent framing, distance, and motion during capture [105–109]. Accordingly, studies in this category emphasize compression and efficiency strategies for on-device inference, including Raspberry Pi deployments [103], smartphone-based systems [99], and embedded implementations [96], positioning mobile/edge platforms as a cost-effective bridge from laboratory models to field-ready tomato scouting. Mobile platform studies showed the highest variability in hardware reporting [110], with only 18% specifying exact device models and inference frameworks used for latency measurements.

4.2. By task type

Task-based taxonomy reveals the diverse applications of YOLO models in tomato production, ranging from basic fruit detection to sophisticated multi-task systems integrating detection, classification, and localization capabilities.

4.2.1. Counting and yield estimation

Counting and yield estimation systems support production forecasting, resource allocation, and harvest planning, and they typically move beyond single-image detection into three main approaches: direct fruit counting (with an average error of roughly 8 to 15%), density regression for heavily clustered scenes (with an average error of roughly 12 to 22% in the case of densely packed fruiting scenes), and weight-based yield estimation (with an average error of roughly 15 to 28%) [6,42,43,51,77,96,101, 110, 111]. A central challenge across all approaches is occlusion: in dense canopies, only 40 to 60% of fruits may be visibly observable, which makes reliable counting difficult and increases the risk of underestimation. To mitigate duplicate counts and improve robustness, studies incorporate tracking and temporal logic [43, 101] as well as spatial aggregation strategies across frames or regions [6]. Reported results also indicate that multi-view fusion can substantially reduce error, improving performance to roughly ±5–9%, highlighting a promising direction for future yield-estimation research in real field conditions.

4.2.2. Occlusion-aware detection

Occlusion handling is a specialized but critical challenge in tomato detection, especially in dense greenhouse canopies where fruits are frequently hidden by leaves, stems, or neighboring fruits [7, 60, 101, 111, 112]. Occlusion-aware studies [111, 112] typically rely on four method families: attention mechanisms to emphasize partially visible cues, part-based detection that learns discriminative sub-regions (e.g., visible fruit segments or peduncle cues), 3D depth reasoning using RGB-D or geometric priors to separate overlapping objects, and temporal fusion that aggregates evidence across consecutive frames or viewpoints. In terms of performance, reported gains are most evident under standard occlusion (about 40–60% visibility loss), where recall improves to roughly 85–91% compared with ~72–82% baselines, while extreme occlusion conditions (>~80% occluded) remain difficult, with recall commonly dropping to about 58–67%. These results indicate that treating occlusion as a general robustness issue can be insufficient, and that dedicated architectures and multi-view/temporal cues are often necessary for commercial-scale production settings.

4.2.3. Instance segmentation

Instance segmentation extends bounding box detection to pixel-level fruit delineation, enabling precise fruit boundary identification crucial for accurate size estimation, overlap resolution, and detailed phenotyping [8,39,63,70,94,113]. Studies employing YOLO-seg variants [8, 63, 70] demonstrate that pixel-wise segmentation significantly improves counting accuracy in dense scenes and enables shape-based quality assessment. The computational overhead of segmentation compared to detection has limited adoption, but improvements in efficiency (particularly in YOLOv8-seg and YOLOv11-seg) are driving increased research interest in this capability.

4.2.4. Peduncle/picking point detection

Peduncle and picking point detection addresses the specific requirements of robotic harvesting systems, which must identify precise grasp or cut locations rather than just fruit presence [4, 39, 48, 69, 86, 97, 106, 114, 115]. These studies combine object detection with keypoint localization or pose estimation to determine optimal manipulation points [4, 39, 114]. The specificity of this task requires high-accuracy positioning (usually below 5 mm error) and compatibility with robotic control systems. The results of experimental research have shown that the success of the harvesting procedure is only possible after determining the cutting point and understanding fruit pose estimation, peduncle orientation, and possible obstacles in the manipulation area; this is one of the most complex applications of YOLO models in agricultural practice.

4.2.5. Multi-task detection

Multi-task learning approaches simultaneously address multiple objectives, such as detection, classification, and localization, within unified architectures [48, 67, 89, 115]. These systems leverage shared feature representations to improve efficiency compared to separate single-task models while maintaining comparable accuracy [67, 115]. Multi-task architectures are particularly valuable for robotic harvesting applications requiring integrated detection, maturity assessment,

and grasp point identification. The limited research in this area suggests that multi-task learning remains challenging to optimize, as task-specific requirements may conflict during training, but the efficiency benefits make this a promising direction for practical deployment.

4.2.6. Disease/defect detection

Disease and defect detection [54, 88, 94, 116] addresses quality control and plant health monitoring, identifying issues such as gray mold [12], surface defects [116], and physiological disorders. These applications extend beyond fruit detection to encompass plant health assessment, critically for integrated pest management and quality assurance. From the relatively small amount of research done, it has been found that the detection of diseases has been an uncharted area in comparison to fruit detection. The reasons for this could be the complexity of the variation in the symptoms of the diseases and the fact that diseases in their early stages are difficult to identify. It's a significant area for future research due to the economic implications of crop diseases.

4.2.7. 3D localization

3D localization systems determine fruit spatial coordinates in three-dimensional space, essential for robotic manipulation and precise yield mapping. These studies [8, 11, 46] integrate YOLO detection with stereo vision, depth sensors, or monocular depth estimation to recover spatial information. The technical complexity of 3D reconstruction and the requirement for specialized sensor hardware (RGB-D cameras, stereo rigs) have limited research in this area, though it is critical for autonomous harvesting systems that must navigate complex 3D greenhouse environments.

Taken together, the task-based evidence reveals a striking imbalance: the majority of the 110 reviewed studies (>60%) frame tomato vision as a single-task bounding-box detection problem, despite the fact that real agricultural systems require simultaneous detection, maturity assessment, and spatial localization. The studies that attempt multi-task architectures [48, 68, 89, 116] consistently report efficiency gains, yet they remain a small fraction of the literature. So, the field's continued focus on single-task benchmarks is therefore partly responsible for the lab-to-field gap.

4.3. By environment

Environmental taxonomy reveals how deployment context shapes model design, with greenhouse environments dominating research due to their commercial importance, while natural field conditions present significant challenges for model robustness.

4.3.1. Greenhouse

Greenhouse environments represent the largest environmental category, reflecting the controlled conditions that facilitate high-accuracy detection and the commercial importance of protected cultivation [7, 8, 11, 39–43, 46, 48, 50, 53, 54, 57, 58, 60]. Greenhouse studies benefit from relatively stable illumination (though still challenging with variable sunlight and artificial lighting), reduced weather impacts, and structured plant arrangements that simplify detection tasks [62, 63, 67, 68, 70, 72, 84, 87, 89, 97]. Research demonstrates greenhouse-specific challenges, including reflections from plastic film [8],

condensation effects on image quality, and the need for 24/7 operation under varying artificial lighting conditions [62]. The high concentration of greenhouse research indicates strong industry interest in precision agriculture for protected cultivation, where automation ROI is highest due to intensive production systems and controlled environments enabling reliable robotic operation [101,111,113,115,117–119].

4.3.2. Open field

Natural and open-field environments present the most challenging conditions for tomato detection, featuring uncontrolled illumination [6,28,29,38,55,58,61], weather variations [65, 66, 76, 78, 81–83, 86], complex backgrounds [91, 92, 98, 105, 108], and irregular plant structures [114, 120–124]. Studies in this category [81, 98, 120] emphasize robustness to environmental variability, including direct sunlight causing glare and deep shadows, wind-induced motion blur, and variable fruit visibility due to uncontrolled plant growth. The substantial research investment in natural environments reflects the majority of global tomato production occurring in open fields, where automation could provide significant labor cost savings [125]. Nevertheless, the difficulty to keep high accuracy over various environmental conditions is still a big obstacle to the commercial use of this technology, since models frequently show a large drop in performance when moving from controlled to uncontrolled environments.

4.3.3. Controlled vs. uncontrolled lighting

Controlled environments [5, 10, 75, 77, 99], whether laboratory or indoor, provide optimal conditions for algorithm development and testing, where lighting is consistent, backgrounds are static, and test setups are carefully planned and controlled. These research papers [77, 99, 116] serve as crucial proof-of-concept and enable controlled performance evaluations, free from any variables that may be contributed by ambient conditions. The fact that systems designed and developed within controlled environments have very little practical application outside indicates that there is a research-to-practice gap, since controlled environments [116, 126, 127] are hardly representative of real-world agricultural environments. The number of research papers that are conducted in controlled environments is not very large, indicating that it is understood that performance under real-world conditions is crucial.

The environmental taxonomy exposes a fundamental research bias: greenhouse conditions dominate the literature (34/110 studies) even though open-field cultivation accounts for the majority of global tomato production. More critically, studies rarely test models across environments. They describe performance in one environment, not robustness across environments. Until cross-environment reporting becomes standard, the dominance of greenhouse-derived accuracy figures gives an inflated picture of the field's readiness for open-field deployment.

5. Datasets and evaluation protocols

5.1. Dataset landscape

Datasets are not neutral. Every choice, environment, annotation rule, and split encode assumptions about what counts as valid tomato detection. This study reveals systematic bias: the field disproportionately relies on small (fewer than 3,000 images,

48.6%), private (only 12% public), and environmentally narrow datasets (greenhouse: 34 studies; open-field: 27, despite the open-field dominating global production). This is not merely a meta-analytic inconvenience. It is a credibility gap: claims of "state-of-the-art performance".

5.1.1. Dataset scale: Four-tier distribution

This section summarizes the dataset landscape used for YOLO-based tomato perception, focusing on scale, labeling choices, and coverage. The analysis is based on the compiled reference **Table 3** (N = 110 studies).

Table 3. Dataset size tiers in the reviewed YOLO-based tomato perception studies (N = 110).

Tier	Definition (images)	Studies, n (% of N)	Median size (images)	Representative
Micro (≤1,000)	≤1,000	25 (22.9%)	867	Xu et al. [55] (n = 1,000), Liu et al. [76] (n = 966), Su et al. [53] (n = 462).
Small (1,000–3,000)	1,001–3,000	28 (25.7%)	1,860	Chen et al. [75] (n = 1,825), Rong et al. [39] (n = 1,528).
Medium (3,000–10,000)	3,001–10,000	28 (25.7%)	4,350	Ayyad et al. [85] (n = 3,797), Gao et al. [104] (n = 5,728), Zheng et al. [71] (n = 7,800).
Large (>10,000)	>10,000	5 (4.6%)	15,000	Wang et al. [86] (n = 12,880), Ali et al. [40] (n = 25,758), Zhang and Jiang [108] (n = 161,223 + 14,358).
Not reported	—	23 (21.1%)	—	Several studies omit dataset size Wu et al. [49], Qi et al. [123].

Notes: N, number of studies; n, number of studies in a tier; Dataset size refers to the number of labeled images reported by each study. Percentages are computed as n/N × 100. Median size is computed within each tier using only studies that report dataset size. Tier definitions: Micro ≤ 1,000; Small: 1,001–3,000; Medium: 3,001–10,000; Large > 10,000 images. Some studies report multiple datasets; when a combined total is explicitly stated, that value is used.

The predominance of micro and small datasets (48.6% of studies) is not merely a logical limitation; it represents a fundamental mismatch between research methodology and agricultural reality.

5.1.2. Public datasets and benchmark fragmentation

Comparability across studies is limited by dataset fragmentation: most papers evaluate on private collections, often with incomplete reporting of dataset composition and capturing conditions. In the compiled table, 23/110 studies do not report dataset size [5,42], and 15/110 do not specify the sensing platform [61,83]. Only a small subset explicitly indicates multi-environment coverage in the dataset description (9/110) , which limits evidence about cross-environment robustness [85,127].

The fragmentation described above has a compounding effect that is rarely stated directly, because most papers use private, small, single-environment datasets. When Liu et al. [66] report 98.7% mAP on their private field dataset and Chen et al. [75] report 94.2% on their industrial private datasets. This is a credibility problem for the field; without a shared benchmark with standardized splits, the claim of state-of-the-art performance is inherently local and unverifiable.

5.1.3. Annotation strategy and task formulation

Most studies formulate tomato perception as bounding-box detection. In contrast, only a small subset explicitly targets richer labels that are important for harvesting and yield pipelines. Within the compiled table, 6/110 studies explicitly include segmentation objectives [8, 63] and 9/110 explicitly report keypoint/peduncle localization [39, 69]. Counting and yield-related tasks also appear less frequently (8/110) [42, 51] but are essential for production planning.

5.1.4. Environmental and temporal coverage

Environmental coverage is dominated by controlled conditions. Greenhouse settings [11, 39, 44] are explicitly reported in 34/110 studies, while open-field settings [28,38,59] are explicitly reported in only 27/110. Many papers use broad labels (e.g., "complex" or "natural" environments) without specifying season, time-of-day, or growth-stage coverage, which complicates generalization claims. Dense canopies and self-occlusion are common in greenhouse imagery, motivating occlusion-aware labeling or difficulty tags.

5.1.5. Data augmentation strategies

Augmentation is widely used to compensate for limited coverage, but its reporting is often brief. For agricultural deployment, the most valuable augmentations are those that match real shifts observed in practice: illumination changes, scale/distance variation, motion blur (mobile platforms), and synthetic occlusion [60, 111].

5.1.6. Cross-domain YOLO-based tomato detection: Meta-analysis

To quantify the laboratory-to-field performance gap, we performed a random-effects meta-analysis on the nine studies reporting both controlled (baseline) and cross-domain accuracy. The pooled mean absolute drop in mAP was 8.24% (95% confidence interval: 6.7–9.5%), with individual study drops ranging from 6.3% to 9.2%. Heterogeneity was moderate (I^2 = 34.2%, τ^2 = 0.0012), suggesting that while most studies show a consistent degradation pattern, the exact magnitude varies with factors such as YOLO version, dataset size, and environmental shift severity (e.g., greenhouse-to-field vs. intra-greenhouse lighting shifts). The drop was statistically significant ($p < 0.001$, one-sample t-test against zero). These results confirm that domain shift consistently undermines laboratory-reported performance, and that occlusion, illumination changes, and unseen backgrounds are primary drivers (see **Table 4** and **Figure 5**).

Table 4. Study characteristics and mean cross-domain performance outcomes for YOLO-based tomato detection (N = 9).

Study	YOLO version	Environment	Dataset size	Baseline(%)	Cross-domain (%)	Drop (%)
Mbouembe et al., 2023 [78]	YOLOv4-tiny	Natural, complex	966	82.8	75.2	−7.6
Liu et al., 2020 [76]	YOLOv3	Natural, Various	966	96.4	88.1	−8.3
Mbouembe et al., 2024 [81]	YOLOv5s	Natural, complex	966	87.7	81.4	−6.3
Liu et al., 2024 [66]	YOLOv7	Natural field	4,350	98.7	89.5	−9.2
Chen et al., 2021 [75]	YOLOv3	Complex, Industrial	1,825	94.2	86.9	−7.3
Su et al., 2022 [53]	YOLOv3	Natural greenhouse, Varied	462	97.5	88.7	−8.8

Table 4. *Cont.*

Study	YOLO version	Environment	Dataset size	Baseline(%)	Cross-domain (%)	Drop (%)
Zheng et al., 2022 [59]	YOLOv4	Complex natural, Multiple	1,698	94.4	85.2	−9.2
Xu et al., 2020 [55]	YOLOv4−tiny	Natural scenes, Complex	1,000	91.9	83.5	−8.4
Wang and Liu, 2021 [12]	YOLOv3	Greenhouse, internet	15,000	96.4	87.3	−9.1
Mean cross-domain performance drop	-	-	-	-	-	8.24% (95% CI: 6.7–9.5%, N = 9)

Note: CI = Confidence Interval; Mean drop = 8.24% (95% CI: 6.7–9.5%), meaning "we are 95% confident that the true average performance drop across all possible studies lies between 6.7% and 9.5%."

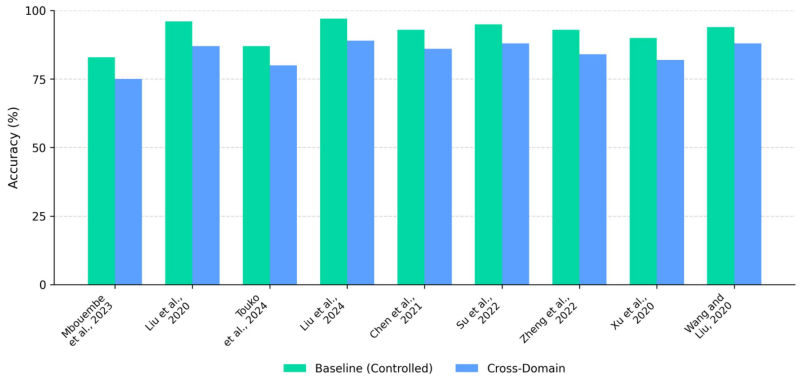

Figure 5. Comparison between the baseline (controlled) and cross-domain (field) performance across nine studies.

- The study selection and inclusion criteria: the nine studies were selected from the full corpus of 110 papers based on the following criteria: explicit reporting of both a baseline (controlled/laboratory) and a cross-domain (field/deployment) mAP using identical model weights and evaluation protocols; sufficient detail to extract or calculate the absolute performance drop; a cross-domain condition involving a meaningful environment shift (e.g., greenhouse-to-field, lab-to-greenhouse, or lighting change). Studies reporting only single-environment performance, or using different model variants across evaluation conditions, were excluded.
- Value extraction protocol: baseline and cross-domain mAP values were taken directly from each paper's primary results table, not derived or estimated. When a paper reported multiple thresholds, mAP0.5 was prioritized for comparability.
- Metric heterogeneity: We acknowledge that mixing mAP0.5 and mAP0.5:0.95 values introduces heterogeneity. All nine included studies reported mAP0.5; no studies in the cross-domain pool reported only mAP0.5:0.95.
- Sensitivity analysis: a leave-one-out sensitivity analysis was conducted; the pooled mean drop ranged from 7.9% to 8.6% when omitting any single study, confirming that no individual study disproportionately influences the result.

5.1.7. Contradictions and reporting gaps across datasets

Two contradictions are particularly instructive: (1) Occlusion claims versus test set composition; eight studies claim "robustness to heavy occlusion," yet only two report the occlusion distribution of their test sets. (2) Cross-domain drop varies even for the same YOLO versions, from the nine studies in the meta-analysis (**Table 4**). YOLOv3

shows a cross-domain drop ranging from 7.3% to 9.1%. The lower end comes from a greenhouse-to-greenhouse lighting shift, the higher end from greenhouse-to-open field with wind-induced motion blur. Surprisingly, most papers report only the average drop without stratifying by environmental shift type.

5.2. Evaluation protocols

Evaluation practices vary substantially across studies, and differences in split design and metric definition can explain a large portion of reported performance gaps. To support reproducible comparisons, we summarize recommended protocols aligned with tomato production use cases.

5.2.1. Train/validation/test splitting and leakage risks

Random image-level splits remain common in agricultural vision but can be overly optimistic when near-duplicate frames from the same row/day appear in both training and test sets. For field deployment, at least one "deployment-realistic" split should be reported (e.g., time-based split, site-based split, or cultivar-based split), in addition to any random split used for ablation.

5.2.2. Detection metrics and reporting conventions

Model performance is usually evaluated using standard classification metrics [128], including precision (P), recall (R), mean average precision (mAP), and F_1 score. Moreover, Metric settings (IoU thresholds, confidence thresholds, and class averaging) must be stated explicitly to enable comparison. These metrics are defined as follows:

$$R = \frac{TP}{TP + FN}, \tag{1}$$

$$P = \frac{TP}{TP + FP}, \tag{2}$$

where TP, FN, and FP represent true positives (correct detection), false negatives (missed detection), and false positives (incorrect detection), respectively.

The mAP metric was employed to assess overall model performance across varying confidence thresholds and is calculated as:

$$mAP = \frac{1}{N_{cls}} \sum_{a=1}^{N_{cls}} AP_a, \tag{3}$$

where N_{cls} represents the total number of classes, and the average precision (AP) for each class is computed as:

$$AP = \sum_{q}^{Q} (r_{q+1} - r_q) \max_{\tilde{r} \geq r_{q+1}} p(\tilde{r}), \tag{4}$$

where $p(\tilde{r})$ denotes the measured precision at recall $\tilde{r}$.

The F_1 score, which provides a harmonic mean of Precision and Recall, was calculated as:

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}. \tag{5}$$

5.2.3. Efficiency metrics across hardware

For practical deployment, papers should report at least one efficiency metric on the target class of hardware (FPS and/or latency, plus model size/parameters). Hardware details (GPU/CPU/NPU type, precision, batch size) should be included, because results are not transferable across platforms without this context. The practical consequence of the evaluation inconsistencies described is larger than commonly acknowledged. An analysis of the nine cross-domain studies in **Table 4** illustrates that the same YOLOv3 model produces AP0.5 values ranging from 82.8% to 96.4% across studies, a spread of 13.6% points. Reviewers and readers who do not scrutinize evaluation details may attribute performance gains to architectural innovation when the actual driver is a more favorable protocol. This calls for journals in agricultural AI to require an evaluation checklist as a condition of publication.

5.2.4. Agricultural-specific evaluation

Tomato applications often require metrics beyond detection. For counting and yield estimation, studies typically report counting error (e.g., percent error) or yield error, and may incorporate tracking to avoid double counting [43, 101]. For harvesting robotics, success should be linked to downstream outcomes such as grasp/cut success rate, peduncle localization error, and failure cases under occlusion [39, 114]. When available, multi-view or temporal fusion should be evaluated explicitly because it can reduce duplicate counting and increase visibility under occlusion.

5.2.5. Reproducibility standards and reporting gaps

Minimum reproducibility requires clear documentation of: dataset source and size, annotation rules, split policy, evaluation metrics and thresholds, training recipe (input size, augmentation, optimizer, epochs), and hardware/software. A structured reporting checklist would substantially improve comparability.

6. Practical decision frameworks for application-driven model selection

Application requirements should determine the sensing modality, model size, and evaluation protocol. **Table 5** summarizes deployment-oriented guidance across common tomato use cases.

Table 5. Practical decision framework for YOLO-based tomato detection (application to configuration).

Application needs	Platform and environment	Operational risks in practice	Recommended modality	Model choice and deployment
Robotic harvesting (precision localization, real-time action)	Ground robot; greenhouse or open field; strict real-time constraints	Occlusion by leaves/stems/trellis; motion blur; viewpoint change; domain shift	RGB baseline; RGB-D if depth supports grasp/pick-point reasoning	Prefer edge-first configurations. Benchmark YOLOv8/YOLOv11 as deployable baselines; include YOLOv12 as an emerging comparator. Report end-to-end latency, input resolution, hardware, occlusion stress cases, and split/IoU settings.

Table 5. *Cont.*

Application needs	Platform and environment	Operational risks in practice	Recommended modality	Model choice and deployment
Monitoring/scouting (presence, trends, alerts)	Fixed cameras or mobile/edge devices; greenhouses/open fields; continuous operation	Illumination variability (time of day/weather), background clutter, and lens contamination; domain shift across sites	RGB; add multispectral/thermal only with a clear field benefit; fusion only if justified	Favor lightweight models for continuous inference. Compare YOLOv8/YOLOv11; treat YOLOv12 as emerging and evaluate overhead versus robustness. Require cross-site testing and full metric-computation details.
Grading/sorting (post-harvest quality, defect screening)	Fixed camera on conveyor; controlled lighting; high throughput	Specular highlights; fast motion; small-defect sensitivity; class imbalance	RGB with controlled illumination; higher resolution if defects are small	Choose a model that meets throughput; YOLOv11 is a practical baseline; test YOLOv12 if attention-centric variants improve fine-detail robustness within compute limits. Report items/sec, exposure, and defect-size stratified performance.
Yield estimation/counting (aggregate metrics)	Fixed cameras/robots; occasional UAV studies; greenhouse/open-field	Double counting across frames; occlusion clusters; scale variance; domain shift	RGB; multi-view or temporal cues if feasible	Use the detector plus an explicit counting protocol (frame-based versus tracking). Benchmark across YOLO generations and include YOLOv12 when evaluating scale/occlusion robustness. Report counting rules and errors by occlusion density.

In addition to architectural choices, for deployment, additional optimization passes might be necessary. For edge computing systems, after-training quantization (FP16 or INT8) can result in latency reduction by 2 to 4 times with little loss of accuracy. Pruning methods to eliminate unnecessary neurons/channels can then further optimize models for ultra-low-power systems. However, such models should also be tested for agricultural scenarios to eliminate any artifacts introduced by quantization that might affect the detection of small or partially obscured fruits. The deployment-level tools TensorRT, OpenVINO, and TFLite should then be incorporated into the testing process.

Key recommendations for agricultural implementation

- For researchers and developers:
  - Validation of models on realistic occlusion (40–80% fruit visibility loss) and illumination changes.
  - Report efficiency numbers (latency, memory) on the target hardware (e.g., Jetson Nano, Raspberry Pi).
  - Share code, models, and dataset documentation to facilitate reproducibility.
- For agricultural engineers and technicians
  - For greenhouse monitoring: Fixed cameras using YOLOv8/v11 and give the best trade-off accuracy vs. deployability.
  - For harvesting by robots: Models under 25 GFLOPs with RGB-D fusion for occlusion are further preferred.
  - For scouting in the field: There are YOLO models optimized for mobile devices (e.g., YOLOv5s, YOLOv8n), which can do real-time inference on smartphones.
- For agronomists and farm managers:

- Model the cooperative high-cost sensors (multispectral) for investment to be shared among farms.
- Staged validation: controlled trials then, single-field test then, multi-location evaluation.
- Record the environmental conditions (lighting, cultivar, growth stage) when obtaining training data.

- For journal editors and reviewers:
  - Require reproducibility checklists including dataset provenance, split strategy, and hardware specifications.
  - Prioritize studies that validate across multiple environments and report failure cases.
  - Encourage publication of negative results and deployment challenges.

7. Research roadmap to improve the laboratory-to-field gap

Progress toward robust field systems requires shifting incentives from isolated, single-dataset accuracy to reproducible, cross-environment evaluation and deployment feasibility. **Table 6** outlines a structured roadmap.

Table 6. Structured research roadmap to improve the laboratory-to-field gap (short to long term).

Horizon	Core goal	Concrete research priorities	Evaluation/benchmark requirements	Minimum reporting deliverables
Short term (0–12 months)	Make results comparable and reproducible	Standardize splits and IoU thresholds; specify metric computation; report the full training recipe and augmentation; publish code/weights when possible	Use held-out test sets; document the split method (scene/plant/time); run sensitivity checks where feasible	Minimum checklist: dataset access, split definitions, training hyperparameters, metric settings, inference hardware, and latency
Medium term (1–3 years)	Robustness under domain shift becomes first-class	Cross-environment validation by default; occlusion and illumination stress tests; deployment feasibility benchmarking (latency, memory, power)	Multi-site evaluation; greenhouse-to-open-field transfer tests; accuracy-efficiency trade-off plots with consistent settings	Benchmark suite and cross-domain test protocol; deployment feasibility table with end-to-end latency and memory footprint
Long term (3–5+ years)	Scale from prototypes to reliable field systems	Generalize across cultivars, seasons, and sensors; multimodal fusion only where it improves field robustness; community benchmarks and reproducible leader-boards.	Public multi-site datasets; longitudinal evaluation; standardized reporting to prevent overstated gains without field evidence	Public benchmark datasets/protocols; a model zoo with deployment profiles; reproducible artifact releases (code, weights, configurations)

7.1. For the short term (0–12 months)

Standardize splits and IoU thresholds, specify metric computation, report full training recipes and augmentation, and publish code and model weights where possible. At minimum, each study should provide: dataset access, split definitions, training hyperparameters, metric settings, and inference hardware with latency.

7.2. For the medium term (1–3 years)

Cross-environment validation should become standard practice, with occlusion and illumination stress tests required alongside deployment feasibility benchmarking (latency, memory, power). Multi-site evaluation and greenhouse-to-field transfer tests

should be accompanied by accuracy–efficiency trade-off plots under consistent settings.

7.3. For the long term (3–5+ years)

Three priorities define this horizon. First, a Community Benchmark Initiative should establish and maintain a public, multi-environment tomato detection dataset with standardized train/test splits, annotation protocols, and evaluation metrics covering diverse conditions (greenhouse, open field, varying occlusion levels). Second, Reproducibility Certification should be adopted by agricultural vision journals and conferences, requiring authors to provide dataset access or generation code, full training configurations, model weights, or inference code, and hardware-validated efficiency metrics. Third, a Model Optimization Ecosystem should develop and share best practices for model compression specific to agricultural vision, including quantization-aware training recipes for common edge devices (Jetson, Raspberry Pi, mobile NPUs) and pruning strategies that preserve accuracy for small, occluded objects.

8. Reproducible benchmark of representative YOLO generations

To provide a transparent and reproducible performance baseline under controlled indoor conditions, we trained and evaluated nine YOLO variants on the Laboro Tomato dataset [129] under identical training settings. Specifically, all models were trained using the SGD optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}, with a batch size of 16 and an input resolution of 640 × 640. Standard YOLO data augmentation strategies were used, and the rest of the hyperparameters were kept the same as in the original implementation. All the experiments were done on a workstation with an Intel Xeon i9, 14900K CPU, an NVIDIA GeForce RTX 4090 GPU, and PyTorch 2.4.1, so the computational environment was the same for all models. We are not trying to name the best overall, but rather to offer a clear baseline under controlled settings and to demonstrate useful accuracy-efficiency trade-offs for deployment. Moreover, this benchmark does not imply field readiness; it merely isolates architecture differences under identical conditions.

From **Table 7**, we can see that YOLOv7 has outperformed other object detection models in terms of precision (83.7%), recall (79.1%), and AP@0. 5 (87. 5%), thereby clearly establishing it as a top detection performer. Unfortunately, its excellent detection capability is coupled with a high computational cost of 103. 2 GFLOPs, potentially putting it out of reach for edge devices with limited resources. When it comes to lightweight models, YOLOv11s and YOLOv12s appeared to be the best choice in terms of trade-off, recording respective AP@0. 5:0. 95 scores of 70.8% and 70.1%, while running at only ~21 GFLOPs, which is roughly a fifth of the computational demand of YOLOv3. These findings serve as a reminder that it's not just about getting the highest score but also about working efficiently: in many cases of real agriculture, the slight improvements of bigger models may not be worth the trouble, and hence optimized models such as YOLOv11s and YOLOv12s will be more appropriate.

Table 7. Baseline benchmark on the Laboro Tomato dataset under fixed training settings.

Model	P (%)	R (%)	AP@0.5 (%)	AP@0.5:0.95 (%)	Params	GFLOPS	Augmentation technique (all identical)
YOLOv3 [13]	79.8	78.1	84.3	73.1	103 M	282.2 G	Mosaic, scale/translate, HSV jitter, H-flip
YOLOv5s [15]	77.6	77	82.7	65.4	7.03 M	15.8 G	
YOLOv6s [16]	81.0	69.3	80.9	67.3	16.3 M	43.7 G	
YOLOv7 [17]	83.7	79.1	87.5	72.4	36.5 M	103.2 G	
YOLOv8s [18]	78.2	76.8	83.3	69.1	11.1 M	28.4 G	
YOLOv9s [19]	80.1	76.7	84	70.4	7.17 M	26.7 G	
YOLOv10s [20]	83.1	71.8	81.6	68.3	7.22 M	21.4 G	
YOLOv11s [21]	83.9	74.4	84.3	70.8	9.42 M	21.3 G	
YOLOv12s [22]	82.3	75.4	83.7	70.1	9.23 M	21.2 G	

What these results basically mean is that the configuration with the highest accuracy may not be the best choice in terms of deployment. Lightweight versions can deliver almost the same level of accuracy, but that comes with significantly lower parameter counts and compute, which may be a deciding factor for embedded agricultural systems.

Interpreting the controlled benchmark

The controlled benchmark presented in **Table 7** serves a specific, critical purpose within our review's argument. While we emphasize the limitations of laboratory benchmarks for predicting field performance, this standardized comparison under identical training conditions provides a transparent reference point for accuracy-efficiency trade-offs. It conclusively demonstrates that the model with the highest accuracy (YOLOv7) carries a substantial computational premium (>100 GFLOPs), while more recent, efficiency-optimized variants (YOLOv11s, YOLOv12s) deliver competitive performance at a fraction of the cost. This directly reinforces our core thesis: model selection for deployment cannot be based on accuracy alone. Practitioners must consult efficiency-aware benchmarks alongside robustness considerations (Subsection 9.2) and application-specific requirements (**Table 4**) to make informed decisions. We present this benchmark not as an endorsement of controlled testing, but as a tool to deconstruct the trade-offs that define deployability.

In addition, **Table 7** reports performance from a single train/test split under controlled conditions. This design enables a fair comparison of YOLO architectures but does not support statistical inference about model superiority. To implement 5-fold cross-validation across all nine models was constrained by computational cost: each training run required 4 to 6 h on an RTX 4090, totaling over 180 GPU h for full cross-validation, which was prohibitive within the scope of this review. The current design is not optimal; readers should therefore treat these values as descriptive point estimates, not as evidence of a statistically significant architectural advantage. For statistical claims, the confidence intervals reported in Subsection 5.1.6 should be a reference.

9. Discussion

This review of 110 studies on YOLO-based tomato detection shows fast model iteration but slower progress on what determines field success: reproducibility,

robustness to domain shift, and deployment feasibility. Reported accuracy improves markedly in controlled evaluations (e.g., 78.3% mAP for YOLOv3 to 94.7% for YOLOv11), yet real deployments often lose performance, indicating that current benchmarks and reporting practices do not adequately reflect operational conditions.

9.1. Reproducibility as a bottleneck

A substantial portion of the literature cannot be reliably re-executed or compared. Missing items frequently include dataset capture context, split design, augmentation settings, and evaluation specifics (IoU thresholds, confidence filtering, and post-processing). This limits scientific accumulation and makes practical transfer risky, especially when "mAP" is reported without clarifying the exact definition (e.g., AP@0.5 vs. AP@0.5:0.95).

Improvement requires minimum reporting standards (dataset documentation, canonical splits, a full training recipe, and fully specified metrics) and stronger incentives for releasing datasets/models. Given agricultural variability, benchmarking should use multiple reference datasets and versioned leader-boards rather than relying on a single universal benchmark.

9.2. Why lab gains do not hold in the field

As introduced in Subsection 1.4 and quantified in Subsection 5.1.6 , we identify recurring drivers of the lab-to-field gap. First, domain shift is pervasive: models trained on one site/season/cultivar often fail under new lighting, canopy structure, or background textures. This is exacerbated by limited data, where 48.6% of studies use fewer than 3,000 images, and by scarce cross-domain testing (only 9/110 evaluate across sites or seasons), where performance drops of 6–9% (mean = 8.24%) are commonly reported.

Second, occlusion is under-modeled in many benchmarks. Commercial canopies frequently include high partial occlusion (often 40–80% of fruit partially hidden), yet standard test sets tend to favor clearer views. Work explicitly targeting occlusion [7, 60, 111, 113] suggests that attention, part-based cues, or temporal fusion can recover roughly 8–15% recall under heavy occlusion, but these approaches remain relatively uncommon.

Third, illumination variability is a dominant failure source across platforms (greenhouse reflections and diurnal shifts; open-field extremes; motion-dependent exposure on mobile systems), yet only a small subset of studies (8/110) evaluates time-of-day robustness in a structured way.

Fourth, many "deployment" claims are not end-to-end. Reported latency is often measured on desktop GPUs and excludes pre/post-processing, sensor I/O, and concurrent workloads. When considering all constraints simultaneously, 73% of high-accuracy configurations go beyond typical edge limits, and only 22% confirm performance on target hardware.

Finally, different evaluation settings (IoU thresholds, checkpoint selection rules, confidence thresholds, and test time augmentation/ensembles) can change results by several mAP points, thus it becomes problematic to attribute improvements to

architecture rather than protocol.

9.3. Accuracy must be reported together with efficiency

Agricultural deployment changes the objective from "highest benchmark score" to "best accuracy under compute, power, and thermal limits." Our benchmarking illustrates this trade-off: high-accuracy models can be several times more expensive than compact variants, which often deliver slightly lower accuracy but far better feasibility for real-time systems.

Efficiency reporting should go beyond GFLOPs and parameter counts and include on-device latency, memory footprint (including activation), and sustained-power behavior. Quantization is particularly underused: while post-training quantization can deliver roughly 2–4 times speedups with small accuracy loss (often <2% mAP), only 18% of studies report quantization experiments, and fewer validate on target devices. We recommend presenting accuracy–efficiency Pareto curves rather than single "best" models.

9.4. Evaluation should match the deployment goal

Standard detection metrics remain necessary for comparability but are often insufficient for agricultural use cases. For yield estimation, counting error is typically more relevant than per-image mAP; for harvesting, spatial precision (e.g., centroid/peduncle localization) matters more than IoU thresholds; for monitoring, temporal stability and tracking reliability can outweigh single-frame recall.

We recommend a layered evaluation protocol: (1) conventional detection metrics on held-out sets; (2) task-aligned metrics (counting error, localization error, harvesting success proxies); (3) robustness reporting stratified by occlusion and environment; and (4) validated efficiency metrics on target hardware.

9.5. Limitations of this review

This synthesis is limited by publication and language restrictions (peer-reviewed English sources), a deliberate focus on YOLO-family methods (excluding alternative detector families), and heterogeneity that constrained quantitative pooling. We also rely on authors' reported hardware and deployment claims without independent replication, and we only partially analyze multimodal sensing despite its growing relevance.

Beyond these constraints, three additional limitations warrant mention. First, our controlled benchmark (**Table 7**) used only the LaboroTomato dataset under laboratory conditions; results may differ on other datasets or under field conditions with occlusion and illumination variation. Second, our exclusive focus on YOLO-family methods means we cannot evaluate whether alternative architectures (e.g., DETR, EfficientDet, Transformer-based detectors) offer better robustness or deployability. Third, we relied entirely on authors' reported claims without independent replication; the true reproducibility rate may differ from our audit findings.

9.6. Toward trustworthy agricultural AI: Building confidence for adoption

Technical performance alone cannot guarantee adoption in agricultural settings where decisions carry economic risk and safety implications. Building trust in AI systems requires addressing several non-technical dimensions:

- Economic viability: The total cost of ownership, including hardware, computation, maintenance, and technical expertise, must demonstrate a clear return on investment through labor savings, yield improvement, or quality premiums. Our cost-effectiveness analysis (**Table 5**) provides initial guidance, but farm-scale business cases remain understudied.
- Interpretability and failure transparency: Farmers and agronomists need to understand why models fail (e.g., "missed due to shadow" rather than "false negative"). Developing explainable AI techniques tailored to agricultural contexts, such as highlighting uncertainty in occluded regions, can build operational confidence.
- Integration with existing workflows: Successful deployment requires seamless integration with farm management software, irrigation systems, and harvest logistics. Modular architectures that separate detection from downstream decision logic facilitate this integration.
- Capacity building: Training programs for agricultural technicians in data collection, model validation, and system maintenance are essential for sustained operation beyond research prototypes.

Future research should expand beyond accuracy metrics to include these adoption-critical dimensions, treating agricultural AI systems as socio-technical systems rather than isolated algorithms. Journals can support this shift by publishing case studies of successful (and failed) deployments, along with economic analyses and user experience reports.

10. Conclusion

This systematic review of 110 studies reveals both the remarkable potential and persistent challenges of computer vision for tomato detection in precision horticulture. While algorithmic accuracy has improved substantially in controlled settings, translating these advances into reliable field systems requires addressing multidimensional implementation barriers: environmental variability, hardware constraints, reproducibility limitations, and interdisciplinary communication gaps.

This review provides three pathways to accelerate this translation:

1. Structured communication: The three-dimensional taxonomy and practical decision framework offer a common language and evaluation criteria for interdisciplinary teams combining agronomic knowledge with technical expertise.
2. Methodological reform: The reproducibility audit and research roadmap call for standardized reporting, cross-environment validation, and hardware-aware benchmarking, shifting incentives from isolated accuracy gains to robust, deployable performance.

3. Practitioner-centered design: By foregrounding agricultural use cases and constraints, we emphasize that successful systems must balance technical specifications with economic viability, operational practicality, and user trust.

Not merely in stepping up benchmark figures by minor margins, the future of AI in agriculture is about creating systems that keep their accuracy even when there is morning fog in the field, that can work on battery power for hours, that are able to tell farmers when they're not sure, and whose development is so clear that anyone from the global community can reproduce and enhance it.

Besides technical performance, achieving this vision requires consideration of social factors such as economic viability, model interpretability for farmers, integration with existing farm workflows, and capacity building for agricultural technicians.

Future technologies must be judged not only on detection accuracy but also on factors such as trustworthiness, user friendliness, and the overall financial benefits in practical agricultural operations. The adoption of the frameworks and priorities envisaged here will enable researchers, engineers, and agronomists to work together to overcome the implementation chasm and ensure the realization of intelligent, sustainable tomato production systems.

Funding: This research was funded by the Shandong Provincial Natural Science Foundation, grant number ZR2023QC116.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Data availability statement: Data are available online: https://github.com/laboroai/LaboroTomato.

Conflict of interest: The authors declare no conflict of interest.

AI use statement: During the preparation of this work, the authors used Grammarly for language polishing, grammar correction, and rephrasing for clarity. No AI tools were used for data analysis, interpretation, hypothesis generation, or the creation of scientific content. All technical content is the original work of the human authors. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
UAVs	Unmanned Aerial Vehicles
P	Precision
R	Recall
mAP	Mean Average Precision
NPU	Neural Processing Unit
ONNX	Open Neural Network Exchange
TensorRT	NVIDIA's deep learning inference optimizer
Faster R-CNN	Faster Region-based Convolutional Neural Network

SSD	Single Shot Multibox Detection
RGB-D	Red-Green-Blue-Depth
IoU	Intersection over Union
FPS	Frames Per Second

References

1. World Tomato Production by Country. Available online: https://ie.atlasbig.com/countries-by-tomato-production (accessed on 18 September 2025).
2. Tomato market size & share analysis—Growth trends and forecast (2026–2031). Available online: https://www.mordorintelligence.com/industry-reports/tomato-market (accessed on 18 September 2025).
3. Pathak TB, Stoddard CS. Climate change effects on the processing tomato growing season in California using growing degree day model. Modeling Earth Systems and Environment. 2018; 4(2): 765–775. doi: 10.1007/s40808-018-0460-y
4. Li P, Wen M, Zeng Z, et al. Cherry Tomato Bunch and Picking Point Detection for Robotic Harvesting Using an RGB-D Sensor and a StarBL-YOLO Network. Horticulturae. 2025; 11(8): 949. doi: 10.3390/horticulturae11080949
5. Miao Z, Yu X, Li N, et al. Efficient tomato harvesting robot based on image processing and deep learning. Precision Agriculture. 2023; 24(1): 254–287. doi: 10.1007/s11119-022-09944-w
6. Wang A, Xu Y, Hu D, et al. Tomato Yield Estimation Using an Improved Lightweight YOLO11n Network and an Optimized Region Tracking-Counting Method. Agriculture. 2025; 15(13): 1353. doi: 10.3390/agriculture15131353
7. Zhang J, Xie J, Zhang F, et al. Greenhouse tomato detection and pose classification algorithm based on improved YOLOv5. Computers and Electronics in Agriculture. 2024; 216: 108519. doi: 10.1016/j.compag.2023.108519
8. Zheng S, Liu Y, Weng W, et al. Tomato Recognition and Localization Method Based on Improved YOLOv5n-seg Model and Binocular Stereo Vision. Agronomy. 2023; 13(9): 2339. doi: 10.3390/agronomy13092339
9. Chen J, Yu R, Yang M, et al. SN-YOLO: A Rotation Detection Method for Tomato Harvest in Greenhouses. Electronics. 2025; 14(16): 3243. doi: 10.3390/electronics14163243
10. Li T, Sun M, He Q, et al. Tomato recognition and location algorithm based on improved YOLOv5. Computers and Electronics in Agriculture. 2023; 208: 107759. doi: 10.1016/j.compag.2023.107759
11. Liu X, Teng W, Yu H, et al. GAE-YOLO: a lightweight multimodal detection framework for tomato smart agriculture with edge computing. Frontiers in Plant Science. 2025; 16: 1712432. doi: 10.3389/fpls.2025.1712432
12. Wang X, Liu J. Multiscale Parallel Algorithm for Early Detection of Tomato Gray Mold in a Complex Natural Environment. Frontiers in Plant Science. 2021; 12: 620273. doi: 10.3389/fpls.2021.620273
13. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv preprint. 2018. doi: 10.48550/ARXIV.1804.02767
14. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint. 2020. doi: 10.48550/ARXIV.2004.10934
15. ultralytics/yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 22 September 2025).
16. Li C, Li L, Jiang H, et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv preprint. 2022. doi: 10.48550/ARXIV.2209.02976
17. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 17–24 June 2023; Vancouver, BC, Canada. pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721
18. ultralytics/ultralytics (YOLOv8). Available online: https://github.com/ultralytics/ultralytics (accessed on 22 September 2025).
19. Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint. 2024. doi: 10.48550/ARXIV.2402.13616
20. Wang A, Chen H, Liu L, et al. YOLOv10: Real-Time End-to-End Object Detection. arXiv preprint. 2024. doi: 10.48550/ARXIV.2405.14458
21. ultralytics/ultralytics (YOLO11). Available online: https://github.com/ultralytics/ultralytics (accessed on 22

September 2025).
22. Tian Y, Ye Q, Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint. 2025. doi: 10.48550/ARXIV.2502.12524
23. Gao X Li F, Yan J, et al. A Lightweight Greenhouse Tomato Fruit Identification Method Based on Improved YOLOv11n. Agriculture. 2025; 15(14): 1497. doi: 10.3390/agriculture15141497
24. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017; 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031
25. Gao F, Fu L, Zhang X, et al. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. Computers and Electronics in Agriculture. 2020; 176: 105634. doi: 10.1016/j.compag.2020.105634
26. Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector. In: Computer Vision—ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing; 2016. pp. 21–37. doi: 10.1007/978-3-319-46448-0_2
27. Magalhães SA, Castro L, Moreira G, et al. Evaluating the Single-Shot MultiBox Detector and YOLO Deep Learning Models for the Detection of Tomatoes in a Greenhouse. Sensors. 2021; 21(10): 3569. doi: 10.3390/s21103569
28. Fu Y, Li W, Li G, et al. Multi-stage tomato fruit recognition method based on improved YOLOv8. Frontiers in Plant Science. 2024; 15: 1447263. doi: 10.3389/fpls.2024.1447263
29. Zhao M, Cui B, Yu Y, et al. Intelligent Detection of Tomato Ripening in Natural Environments Using YOLO-DGS. Sensors. 2025; 25(9): 2664. doi: 10.3390/s25092664
30. Sun H, Zheng Q, Yao W, et al. An Improved YOLOv8 Model for Detecting Four Stages of Tomato Ripening and Its Application Deployment in a Greenhouse Environment. Agriculture. 2025; 15(9): 936. doi: 10.3390/agriculture15090936
31. Olarewaju ML. YOLOMuskmelon: Quest for Fruit Detection Speed and Accuracy Using Deep Learning. IEEE Access. 2021; 9: 15221–15227. doi: 10.1109/ACCESS.2021.3053167
32. Li X, Cai C, Yang Y, et al. YOLOV8-MR: An Improved Lightweight YOLOv8 Algorithm for Tomato Fruit Detection. IEEE Access. 2025; 13: 48120–48131. doi: 10.1109/ACCESS.2025.3533489
33. Koirala A, Walsh KB, Wang Z, et al. Deep learning—Method overview and review of use for fruit detection and yield estimation. Computers and Electronics in Agriculture. 2019; 162: 219–234. doi: 10.1016/j.compag.2019.04.017
34. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture. 2018; 147: 70–90. doi: 10.1016/j.compag.2018.02.016
35. Tang Y, Chen M, Wang C, et al. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. Frontiers in Plant Science. 2020; 11: 510. doi: 10.3389/fpls.2020.00510
36. Zhou H, Wang X, Au W, et al. Intelligent robots for fruit harvesting: Recent developments and future challenges. Precision Agriculture. 2022; 23(5): 1856–1907. doi: 10.1007/s11119-022-09913-3
37. Zhang T, Zhong L, Wu S, et al. BCP-YOLO: A Lightweight Industrial Tomato Detection Method for UAV Inspection. In: Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence; 13–15 September 2024; Shaoxing China. pp. 474–480. doi: 10.1145/3703187.3703268
38. Lawal OM. Development of tomato detection model for robotic platform using deep learning. Multimedia Tools and Applications. 2021; 80(17): 26751–26772. doi: 10.1007/s11042-021-10933-w
39. Rong J, Dai G, Wang P. A peduncle detection method of tomato for autonomous harvesting. Complex & Intelligent Systems. 2022; 8(4): 2955–2969. doi: 10.1007/s40747-021-00522-7
40. Ali M, Keller C, Huang M. Fruits Detections Using Single Shot MultiBox Detector. In: Proceedings of the 5th ACM International Symposium on Blockchain and Secure Critical Infrastructure; 10–14 July 2023; Melbourne, VIC, Australia. pp. 140–144. doi: 10.1145/3594556.3594619
41. Moreira G, Magalhães SA, Pinho T, et al. Benchmark of Deep Learning and a Proposed HSV Colour Space Models for the Detection and Classification of Greenhouse Tomato. Agronomy. 2022; 12(2): 356. doi: 10.3390/agronomy12020356
42. Rong J, Zhou H, Zhang F, et al. Tomato cluster detection and counting using improved YOLOv5 based on RGB-D fusion. Computers and Electronics in Agriculture. 2023; 207: 107741. doi: 10.1016/j.compag.2023.107741
43. Ge Y, Lin S, Zhang Y, et al. Tracking and Counting of Tomato at Different Growth Period Using an Improving YOLO-Deepsort Network for Inspection Robot. Machines. 2022; 10(6): 489. doi: 10.3390/machines10060489
44. Yang D, Ju C. Performance Comparison of Cherry Tomato Ripeness Detection Using Multiple YOLO Models. AgriEngineering. 2024; 7(1): 8. doi: 10.3390/agriengineering7010008

45. Li P, Zheng J, Li P, et al. Tomato Maturity Detection and Counting Model Based on MHSA-YOLOv8. Sensors. 2023; 23(15): 6701. doi: 10.3390/s23156701
46. Zhao J, Bao W, Mo L, et al. Design of tomato picking robot detection and localization system based on deep learning neural networks algorithm of Yolov5. Scientific Reports. 2025; 15(1): 6180. doi: 10.1038/s41598-025-90080-6
47. Chai S, Wen M, Li P, et al. DCFA-YOLO: A Dual-Channel Cross-Feature-Fusion Attention YOLO Network for Cherry Tomato Bunch Detection. Agriculture. 2025; 15(3): 271. doi: 10.3390/agriculture15030271
48. Deng L, Ma R, Chen B, et al. A detection method for synchronous recognition of string tomatoes and picking points based on keypoint detection. Frontiers in Plant Science. 2025; 16: 1614881. doi: 10.3389/fpls.2025.1614881
49. Wu X, Tian Y, Zeng Z. LEFF-YOLO: A Lightweight Cherry Tomato Detection YOLOv8 Network with Enhanced Feature Fusion. In: Advanced Intelligent Computing Technology and Applications, Lecture Notes in Computer Science. Springer Nature; 2025. pp. 474–488. doi: 10.1007/978-981-95-0006-2_40
50. Wang Y, Rong Q, Hu C. Ripe Tomato Detection Algorithm Based on Improved YOLOv9. Plants. 2024; 13(22): 3253. doi: 10.3390/plants13223253
51. Donizette AC, Rocco CD. Automated Tomato Sorting and Counting Using YOLOv11 for Industrial and Precision Agriculture Applications. Applied Artificial Intelligence. 2025; 39(1): 2576891. doi: 10.1080/08839514.2025.2576891
52. Gao G, Shuai C, Wang S, et al. Using improved YOLO V5s to recognize tomatoes in a continuous working environment. Signal, Image and Video Processing. 2024; 18(5): 4019–4028. doi: 10.1007/s11760-024-03010-w
53. Su F, Zhao Y, Wang G, et al. Tomato Maturity Classification Based on SE-YOLOv3-MobileNetV1 Network under Nature Greenhouse Environment. Agronomy. 2022; 12(7): 1638. doi: 10.3390/agronomy12071638
54. Wang X, Liu J. Tomato Anomalies Detection in Greenhouse Scenarios Based on YOLO-Dense. Frontiers in Plant Science. 2021; 12: 634103. doi: 10.3389/fpls.2021.634103
55. Xu ZF, Jia RS, Liu YB, et al. Fast Method of Detecting Tomatoes in a Complex Scene for Picking Robots. IEEE Access. 2020; 8: 55289–55299. doi: 10.1109/ACCESS.2020.2981823
56. Zhou X, Wang P, Dai G, et al. Tomato Fruit Maturity Detection Method Based on YOLOV4 and Statistical Color Model. In: Proceedings of the 2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER); 27–31 July 2021; Jiaxing, China. pp. 904–908. doi: 10.1109/CYBER53097.2021.9588129
57. Li TH, Sun M, Ding X, et al. Tomato recognition method at the ripening stage based on YOLO v4 and HSV. Transactions of the Chinese Society of Agricultural Engineering. 2021; 37(21): 183. doi: 10.11975/j.issn.1002-6819.2021.21.021 (in Chinese)
58. Zhang F, Chen Z, Ali S, et al. Multi-class detection of cherry tomatoes using improved YOLOv4-Tiny. International Journal of Agricultural and Biological Engineering. 2023; 16(2): 225–231. doi: 10.25165/j.ijabe.20231602.7744
59. Zheng T, Jiang M, Li Y, et al. Research on tomato detection in natural environment based on RC-YOLOv4. Computers and Electronics in Agriculture. 2022; 198: 107029. doi: 10.1016/j.compag.2022.107029
60. Li R, Ji Z, Hu S, et al. Tomato Maturity Recognition Model Based on Improved YOLOv5 in Greenhouse. Agronomy. 2023; 13(2): 603. doi: 10.3390/agronomy13020603
61. Cardellicchio A, Solimani F, Dimauro G, et al. Detection of tomato plant phenotyping traits using YOLOv5-based single stage detectors. Computers and Electronics in Agriculture. 2023; 207: 107757. doi: 10.1016/j.compag.2023.107757
62. He B, Zhang YB, Gong JL, et al. Fast recognition of tomato fruit in greenhouse at night based on improved YOLO v5. Transactions of the Chinese Society of Agricultural Machinery. 2022; 53(5): 201. doi: 10.6041/j.issn.1000-1298.2022.05.020 (in Chinese)
63. Liu M, Chen W, Cheng J, et al. Y-HRNet: Research on multi-category cherry tomato instance segmentation model based on improved YOLOv7 and HRNet fusion. Computers and Electronics in Agriculture. 2024; 227: 109531. doi: 10.1016/j.compag.2024.109531
64. Cai Y, Cui B, Deng H, et al. Cherry Tomato Detection for Harvesting Using Multimodal Perception and an Improved YOLOv7-Tiny Neural Network. Agronomy. 2024; 14(10): 2320. doi: 10.3390/agronomy14102320
65. Hou G, Chen H, Ma Y, et al. An occluded cherry tomato recognition model based on improved YOLOv7. Frontiers in Plant Science. 2023; 14: 1260808. doi: 10.3389/fpls.2023.1260808
66. Liu G, Zhang Y, Liu J, et al. An improved YOLOv7 model based on Swin Transformer and Trident Pyramid Networks for accurate tomato detection. Frontiers in Plant Science. 2024; 15: 1452821. doi: 10.3389/fpls.2024.1452821

67. Chen W, Liu M, Zhao C, et al. MTD-YOLO: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. Computers and Electronics in Agriculture. 2024; 216: 108533. doi: 10.1016/j.compag.2023.108533
68. Liang Z, Zhang C, Lin Z, et al. CTDA: An accurate and efficient cherry tomato detection algorithm in complex environments. Frontiers in Plant Science. 2025; 16: 1492110. doi: 10.3389/fpls.2025.1492110
69. Zhang G, Cao H, Jin Y, et al. YOLOv8n-DDA-SAM: Accurate Cutting-Point Estimation for Robotic Cherry-Tomato Harvesting. Agriculture. 2024; 14(7): 1011. doi: 10.3390/agriculture14071011
70. Wang A, Qian W, Li A, et al. NVW-YOLOv8s: An improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages. Computers and Electronics in Agriculture. 2024; 219: 108833. doi: 10.1016/j.compag.2024.108833
71. Zheng S, Jia X, He M, et al. Tomato Recognition Method Based on the YOLOv8-Tomato Model in Complex Greenhouse Environments. Agronomy. 2024; 14(8): 1764. doi: 10.3390/agronomy14081764
72. Sun X. Enhanced tomato detection in greenhouse environments: a lightweight model based on S-YOLO with high accuracy. Frontiers in Plant Science. 2024; 15: 1451018. doi: 10.3389/fpls.2024.1451018
73. Deng X, Huang T, Wang W, et al. SE-YOLO: A sobel-enhanced framework for high-accuracy, lightweight real-time tomato detection with edge deployment capability. Computers and Electronics in Agriculture. 2025; 239: 110973. doi: 10.1016/j.compag.2025.110973
74. Wang Q, Hua Y, Lou Q, et al. SWMD-YOLO: A Lightweight Model for Tomato Detection in Greenhouse Environments. Agronomy. 2025; 15(7): 1593. doi: 10.3390/agronomy15071593
75. Chen J, Wang Z, Wu J, et al. An improved Yolov3 based on dual path network for cherry tomatoes detection. Journal of Food Process Engineering. 2021; 44(10): e13803. doi: 10.1111/jfpe.13803
76. Liu G, Nouaze JC, Mbouembe PLT, et al. YOLO-Tomato: A Robust Algorithm for Tomato Detection Based on YOLOv3. Sensors. 2020; 20(7): 2145. doi: 10.3390/s20072145
77. Wang X, Vladislav Z, Viktor O, et al. Online recognition and yield estimation of tomato in plant factory based on YOLOv3. Scientific Reports. 2022; 12(1): 8686. doi: 10.1038/s41598-022-12732-1
78. Mbouembe PLT, Liu G, Sikati J, et al. An efficient tomato-detection method based on improved YOLOv4-tiny model in complex environment. Frontiers in Plant Science. 2023; 14: 1150958. doi: 10.3389/fpls.2023.1150958
79. Tsai FT, Nguyen VT, Duong TP, et al. Tomato Fruit Detection Using Modified Yolov5m Model with Convolutional Neural Networks. Plants. 2023; 12(17): 3067. doi: 10.3390/plants12173067
80. Appe SN, Arulselvi G, Balaji GN. CAM-YOLO: tomato detection and classification based on improved YOLOv5 using combining attention mechanism. PeerJ Computer Science. 2023; 9: e1463. doi: 10.7717/peerj-cs.1463
81. Mbouembe PLT, Liu G, Park S, et al. Accurate and fast detection of tomatoes based on improved YOLOv5s in natural environments. Frontiers in Plant Science. 2024; 14: 1292766. doi: 10.3389/fpls.2023.1292766
82. Wang H, Xie Z, Yang Y, et al. Fast identification of tomatoes in natural environments by improved YOLOv5s. Journal of Agricultural Engineering. 2024; 55(3). doi: 10.4081/jae.2024.1588
83. Gai R, Li M, Wang Z, et al. YOLOv5s-Cherry: Cherry Target Detection in Dense Scenes Based on Improved YOLOv5s Algorithm. Journal of Circuits, Systems and Computers. 2023; 32(12): 2350206. doi: 10.1142/S0218126623502067
84. Yang Y, Han Y, Li S, et al. Multi-Growth Period Tomato Fruit Detection Using Improved Yolov5. International Journal of Robotics and Automation Technology. 2025; 9: 44–55. doi: 10.31875/2409-9694.2022.09.06
85. Ayyad SM, Sallam NM, Gamel SA, et al. Particle swarm optimization with YOLOv8 for improved detection performance of tomato plants. Journal of Big Data. 2025; 12(1): 152. doi: 10.1186/s40537-025-01206-6
86. Wang X, Wen X, Li Y, et al. A Precise Detection Method for Tomato Fruit Ripeness and Picking Points in Complex Environments. Horticulturae. 2025; 11(6): 585. doi: 10.3390/horticulturae11060585
87. Li J, Huang Z, Xia L, et al. Tomato maturity detection based on improved YOLOv8n. INMATEH Agricultural Engineering. 2025; 619–629. doi: 10.35633/inmateh-75-53
88. Liu Z, Guo X, Zhao T, et al. YOLO-BSMamba: A YOLOv8s-Based Model for Tomato Leaf Disease Detection in Complex Backgrounds. Agronomy. 2025; 15(4): 870. doi: 10.3390/agronomy15040870
89. Bian S, Zhou J, Gao Q, et al. Simultaneous Identification on Tomato Variety and Maturity Based on Local and Global Feature Fusion. Sensors. 2025; 25(23): 7313. doi: 10.3390/s25237313
90. Huang W, Liao Y, Wang P, et al. AITP-YOLO: Improved tomato ripeness detection model based on multiple strategies. Frontiers in Plant Science. 2025; 16: 1596739. doi: 10.3389/fpls.2025.1596739
91. Li A, Wang C, Ji T, et al. D3-YOLOv10: Improved YOLOv10-Based Lightweight Tomato Detection Algorithm

Under Facility Scenario. Agriculture. 2024; 14(12): 2268. doi: 10.3390/agriculture14122268
92. Hao Y, Rao L, Fu X, et al. Tomato Ripening Detection in Complex Environments Based on Improved BiAttFPN Fusion and YOLOv11-SLBA Modeling. Agriculture. 2025; 15(12): 1310. doi: 10.3390/agriculture15121310
93. Wei J, Ni L, Luo L, et al. GFS-YOLO11: A Maturity Detection Model for Multi-Variety Tomato. Agronomy. 2024; 14(11): 2644. doi: 10.3390/agronomy14112644
94. Xue S, Li Z, Wang D, et al. YOLO-ALDS: An instance segmentation framework for tomato defect segmentation and grading based on active learning and improved YOLO11. Computers and Electronics in Agriculture. 2025; 238: 110820. doi: 10.1016/j.compag.2025.110820
95. Chen BJ, Bu JY, Xia JL, et al. AFBF-YOLO: An Improved YOLO11n Algorithm for Detecting Bunch and Maturity of Cherry Tomatoes in Greenhouse Environments. Plants. 2025; 14(16): 2587. doi: 10.3390/plants14162587
96. Ruparelia S, Jethva M, Gajjar R. Real-Time Tomato Detection, Classification, and Counting System Using Deep Learning and Embedded Systems. In: Proceedings of the International E-Conference on Intelligent Systems and Signal Processing, Advances in Intelligent Systems and Computing. Springer; 2022. pp. 511–522. doi: 10.1007/978-981-16-2123-9_39
97. Yan Y, Zhang J, Bi Z, et al. Identification and Location Method of Cherry Tomato Picking Point Based on Si-YOLO. In: Proceedings of the 2023 IEEE 13th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER); 11–14 July 2023; Qinhuangdao, China. pp. 373–378. doi: 10.1109/CYBER59472.2023.10256630
98. Wang C, Wang C, Wang L, et al. A Lightweight Cherry Tomato Maturity Real-Time Detection Algorithm Based on Improved YOLOV5n. Agronomy. 2023; 13(8): 2106. doi: 10.3390/agronomy13082106
99. Zeng T, Li S, Song Q, et al. Lightweight tomato real-time detection method based on improved YOLO and mobile deployment. Computers and Electronics in Agriculture. 2023; 205: 107625. doi: 10.1016/j.compag.2023.107625
100. Phan QH, Nguyen VT, Lien CH, et al. Classification of Tomato Fruit Using Yolov5 and Convolutional Neural Network Models. Plants. 2023; 12(4): 790. doi: 10.3390/plants12040790
101. Egi Y, Hajyzadeh M, Eyceyurt E. Drone-Computer Communication Based Tomato Generative Organ Counting Model Using YOLO V5 and Deep-Sort. Agriculture. 2022; 12(9): 1290. doi: 10.3390/agriculture12091290
102. Dong Y, Qiao J, Liu N, et al. GPC-YOLO: An Improved Lightweight YOLOv8n Network for the Detection of Tomato Maturity in Unstructured Natural Environments. Sensors. 2025; 25(5): 1502. doi: 10.3390/s25051502
103. Nahiduzzaman M, Sarmun R, Khandakar A, et al. Deep learning-based real-time detection and classification of tomato ripeness stages using YOLOv8 on raspberry Pi. Engineering Research Express. 2025; 7(1): 015219. doi: 10.1088/2631-8695/ada720
104. Gao X, Ding J, Bie M, et al. YOLOv8n-FDE: An Efficient and Lightweight Model for Tomato Maturity Detection. Agronomy. 2025; 15(8): 1899. doi: 10.3390/agronomy15081899
105. Gao X, Ding J, Zhang R, et al. YOLOv8n-CA: Improved YOLOv8n Model for Tomato Fruit Recognition at Different Stages of Ripeness. Agronomy. 2025; 15(1): 188. doi: 10.3390/agronomy15010188
106. Qin X, Cao J, Zhang Y, et al. Development of an Optimized YOLO-PP-Based Cherry Tomato Detection System for Autonomous Precision Harvesting. Processes. 2025; 13(2): 353. doi: 10.3390/pr13020353
107. Yang Z, Li Y, Han Q, et al. A Method for Tomato Ripeness Recognition and Detection Based on an Improved YOLOv8 Model. Horticulturae. 2024; 11(1): 15. doi: 10.3390/horticulturae11010015
108. Zhang W, Jiang F. AHN-YOLO: A Lightweight Tomato Detection Method for Dense Small-Sized Features Based on YOLO Architecture. Horticulturae. 2025; 11(6): 639. doi: 10.3390/horticulturae11060639
109. Wu Q, Huang H, Song D, et al. YOLO-PGC: A Tomato Maturity Detection Algorithm Based on Improved YOLOv11. Applied Sciences. 2025; 15(9): 5000. doi: 10.3390/app15095000
110. Sun H, Xi X, Wu AQ, et al. ToRLNet: A Lightweight Deep Learning Model for Tomato Detection and Quality Assessment Across Ripeness Stages. Horticulturae. 2025; 11(11): 1334. doi: 10.3390/horticulturae11111334
111. Zhao Y, Chen Y, Xu X, et al. Ta-YOLO: overcoming target blocked challenges in greenhouse tomato detection and counting. Frontiers in Plant Science. 2025; 16: 1618214. doi: 10.3389/fpls.2025.1618214
112. Vo HT, Mui KC, Thien NN, et al. Automating Tomato Ripeness Classification and Counting with YOLOv9. International Journal of Advanced Computer Science and Applications. 2024; 15(4). doi: 10.14569/IJACSA.2024.01504113
113. Wei J, Sun Y, Luo L, et al. Tomato ripeness detection and fruit segmentation based on instance segmentation. Frontiers in Plant Science. 2025; 16: 1503256. doi: 10.3389/fpls.2025.1503256
114. Yong W, Shunfa X, Konghao C. YOLOv8-LBP: multi-scale attention enhanced YOLOv8 for ripe tomato

detection and harvesting keypoint localization. Frontiers in Plant Science. 2025; 16: 1656381. doi: 10.3389/fpls.2025.1656381
115. Wu M, Lin H, Shi X, et al. MTS-YOLO: A Multi-Task Lightweight and Efficient Model for Tomato Fruit Bunch Maturity and Stem Detection. Horticulturae. 2024; 10(9): 1006. doi: 10.3390/horticulturae10091006
116. Liang Z, Zhu T, Teng G, et al. YOLO-RGDD: A Novel Method for the Online Detection of Tomato Surface Defects. Foods. 2025; 14(14): 2513. doi: 10.3390/foods14142513
117. Cui B, Zeng Z, Tian Y. A Yolov7 cherry tomato identification method that integrates depth information. In: Proceedings of the 3rd International Conference on Optics and Image Processing (ICOIP 2023); 14–16 April 2023; Hangzhou, China. doi: 10.1117/12.2689199
118. Zhu J, Xie X. EDEM-YOLO: an improved lightweight multi-scale feature fusion model for tomato variety maturity detection. In: Proceedings of the 4th International Conference on Signal Image Processing and Communication; 11 September 2024; Xi'an, China. doi: 10.1117/12.3041034
119. Gai R, Chen N, Yuan H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. Neural Computing and Applications. 2023; 35(19): 13895–13906. doi: 10.1007/s00521-021-06029-z
120. Wang S, Xiang J, Chen D, et al. A Method for Detecting Tomato Maturity Based on Deep Learning. Applied Sciences. 2024; 14(23): 11111. doi: 10.3390/app142311111
121. Tian S, Fang C, Zheng X, et al. Lightweight Detection Method for Real-Time Monitoring Tomato Growth Based on Improved YOLOv5s. IEEE Access. 2024; 12: 29891–29899. doi: 10.1109/ACCESS.2024.3368914
122. Guo J, Yang Y, Lin X, et al. Revolutionizing Agriculture: Real-Time Ripe Tomato Detection With the Enhanced Tomato-YOLOv7 System. IEEE Access. 2023; 11: 133086–133098. doi: 10.1109/ACCESS.2023.3336562
123. Qi J, Cong X, Zhang W, et al. Rapid Detection of Ripe Tomatoes in Unstructured Environments. Journal of Field Robotics. 2025; 42(6): 2920–2935. doi: 10.1002/rob.22556
124. Dong W, Zhao Y, Pei J, et al. Tomato detection in natural environment based on improved YOLOv8 network. Journal of Agricultural Engineering. 2025; 56(4). doi: 10.4081/jae.2025.1732
125. Liang X, Jia H, Wang H, et al. ASE-YOLOv8n: A Method for Cherry Tomato Ripening Detection. Agronomy. 2025; 15(5): 1088. doi: 10.3390/agronomy15051088
126. Wang X, Wu Z, Jia M, et al. Lightweight SM-YOLOv5 Tomato Fruit Detection Algorithm for Plant Factory. Sensors. 2023; 23(6): 3336. doi: 10.3390/s23063336
127. Luan F, Fan K, Xu X, et al. Cherry-YOLO: Enhanced real-time detection of Cherry ripeness and defects with optimized YOLOv8. Computing. 2025; 107(10): 198. doi: 10.1007/s00607-025-01558-0
128. Padilla R, Netto SL, Da Silva EAB. A Survey on Performance Metrics for Object-Detection Algorithms. In: Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP); 1–3 July 2020; Niterói, Brazil. pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130
129. Laboro Tomato: Instance segmentation dataset. Available online: https://github.com/laboroai/LaboroTomato (accessed on 22 September 2025).