RECENT ADVANCES SURVEY IN COMPUTER VISION AND ARTIFICIAL INTELLIGENCE FOR FRUIT DETECTION
/
TINJAUAN KEMAJUAN TERKINI TEKNOLOGI PENGLIHATAN KOMPUTER DAN KECERDASAN BUATAN DALAM PENGESANAN BUAH

Qi LIU[1)], Puteri Suhaiza binti SULAIMAN[1*)], Mas Rina binti MUSTAFFA[1)], Zainal bin Abdul KAHAR[1)], Lian BAI[2)] and Huicai XU[1)]

[1)] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

[2)] College of Software, Shanxi Agricultural University, Taigu, Shanxi, China

**Corresponding author: +60 39769 1717; psuhaiza@upm.edu.my*

DOI: https://doi.org/10.35633/inmateh-79-28

***Keywords**: Fruit Detection, CV, AI, DL, ML*

ABSTRACT
With the development of society and the advancement of science and technology, the object detection technology driven by Artificial Intelligence is also constantly innovating. As an important task in the field of agricultural Computer Vision, fruit detection in real environments faces many challenges. This paper reviews and analyzes the latest breakthroughs and representative studies in this field. Based on the existing research, the existing fruit detection algorithms are roughly divided into 5 categories: (1) the traditional fruit detection algorithm based on manual features; (2) the fruit detection algorithm based on two stages; (3) the fruit detection algorithm based on one stage; (4) the fruit detection algorithm based on anchor-free frame, and (5) the fruit detection algorithm based on transfer learning. This paper also discusses various application scenarios, such as multi-object detection, complex backgrounds, and edge computing. It also summarizes the classic and cutting-edge technical methods at the current time point, providing valuable insights for fruit detection in agricultural production.

ABSTRAK
Dengan perkembangan masyarakat dan kemajuan sains dan teknologi, teknologi pengesanan objek yang dipacu oleh kecerdasan buatan juga sentiasa berinovasi. Sebagai tugas penting dalam bidang penglihatan komputer pertanian, pengesanan buah-buahan dalam persekitaran sebenar menghadapi banyak cabaran. Kertas kerja ini mengulas dan menganalisis penemuan terkini dan kajian-kajian yang mewakili bidang ini. Berdasarkan kajian, algoritma pengesanan buah-buahan sedia ada secara kasarnya dibahagikan kepada lima kategori, iaitu: (1) algoritma pengesanan buah-buahan tradisional berdasarkan ciri manual; (2) algoritma pengesanan buah-buahan berdasarkan dua peringkat; (3) algoritma pengesanan buah-buahan berdasarkan satu peringkat; (4) algoritma pengesanan buah-buahan berdasarkan tanpa sauh; dan (5) algoritma pengesanan buah-buahan berdasarkan pembelajaran pemindahan. Kertas kerja ini juga membincangkan pelbagai senario aplikasi seperti pengesanan berbilang objek, latar belakang kompleks dan pengkomputeran tepi. Ia juga meringkaskan kaedah teknikal klasik dan canggih pada masa kini, memberikan pandangan berharga untuk pengesanan buah-buahan dalam pengeluaran pertanian.

INTRODUCTION
Agriculture is one of the most important sectors and plays a vital role in ensuring global food security, rural development, and economic growth. As agricultural technology advances and modernizes, several challenges need to be addressed, such as climate change, pests and diseases, labor shortages, and the need for sustainable resource management (*Devlet A., 2021*). The fusion of artificial intelligence (AI) and computer vision (CV) can alleviate these challenges and open up new opportunities to improve the efficiency, accuracy, and yield of agricultural processes. Object detection has become a key technology that helps to autonomously identify, locate, and classify objects such as leaves, fruits, pests, and diseases from images or video frames (*Gupta, S., &Tripathi, A.K., 2024*). There are a variety of technologies used for object detection in agriculture, including IoT sensors with cameras that can monitor leaves, fruits, vegetables, and crops in real time. Other technologies include drones and satellites for capturing large-scale crop images for analysis and pest detection, and robotic harvesters for automated sowing and harvesting of fruits and vegetables.

In addition, due to the advancement of deep learning (DL), object detection methods have shifted from feature-based models to data-driven methods, such as convolutional neural networks (CNNs) (*O'shea K. & Nash R., 2015*), YOLO (You Only Look Once) (*Redmon J. et al., 2016*), and region-based CNN (R-CNN) *(Girshick R. et al., 2015*), which are used to detect pests and diseases and growth stages. In addition, the recent developments of Vision Transformers (ViT) (*Dosovitskiy A. et al., 2020*), EfficientDet (*Tan M. et al., 2020*), and hybrid CNN-Transformer architectures have significantly improved the detection performance in harsh field environments.

In recent years, with the development of DL technology, the progress of AI has brought great development prospects to object detection. This paper sorts out the mainstream methods and their evolution in the current object detection field, summarizes their core features and application effects, reveals key influencing factors, and evaluates scalability and deployment feasibility, aiming to provide a reference for academia and engineering and help the digital upgrade of the industry.

Fruit Detection

As a core technology of CV, object detection aims to achieve dual perception of object location and classification in images. In the agricultural field, the goal of object detection is to use images or videos to identify and locate certain objects, such as fruits, leaves, vegetables, and pests on crops. This article focuses on the application of fruit-related aspects in agricultural planting and production. In the fruit detection scenario, its core task is to extract the feature representation of the object from complex farmland images, and then realize applications such as automatic picking, yield assessment, and quality grading. Figure 1 shows the evolution of the main technical methods for object detection.

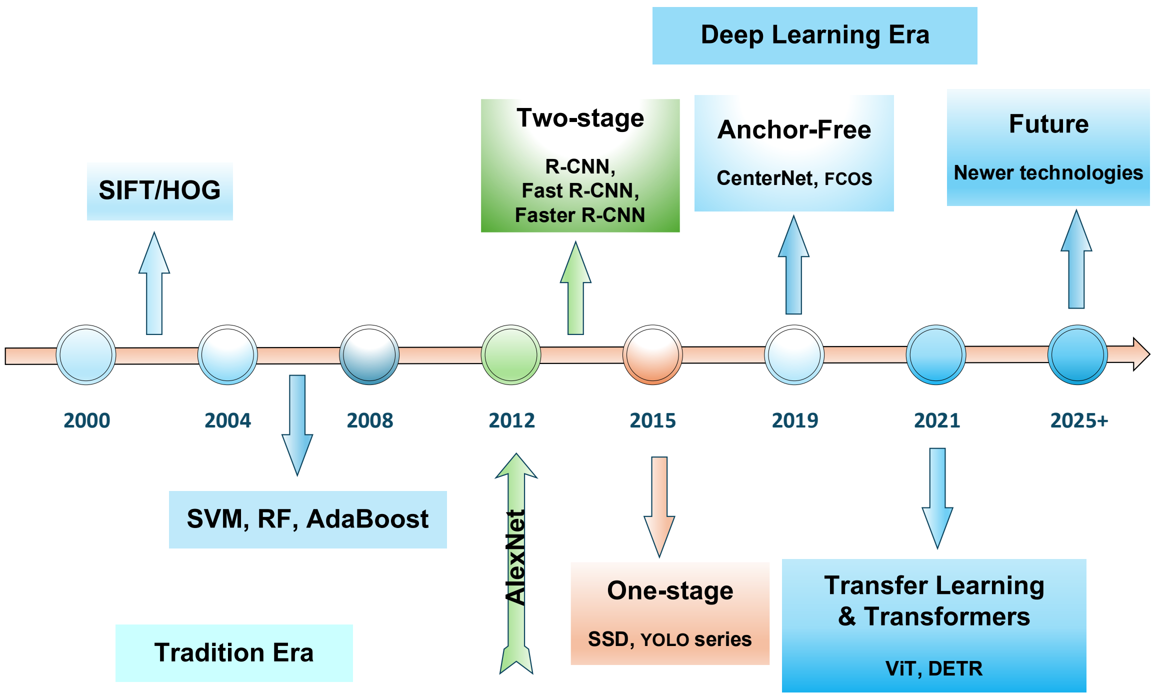

Fig. 1 - The evolution of the main technical methods for object detection

Traditional Methods

Traditional methods have been widely used in tasks such as fruit detection, counting, and grading. The core relies on the feature engineering technology route, and detection is mainly achieved through manual feature extraction and rule definition, such as color, shape, and texture. Its process usually includes four stages: image preprocessing, feature extraction, threshold setting, and classification rules. In image preprocessing, Gaussian or median filtering methods are often used for denoising, and color space conversion is used to enhance fruit features. Subsequently, the tomato area is extracted from the complex background through color segmentation (such as fixed threshold, Otsu method), and morphological operations (erosion, expansion, opening operations, and closing operations) are used to optimize the segmentation results. For the candidate area, edge detection and Hough transform are combined to fit the contour, and the real fruit is screened out using geometric features. At the same time, some fine-grained recognition tasks introduce texture analysis to improve accuracy.

The traditional algorithm is based on the image recognition model, and its detection process always consists of two parts. First, the image is feature extracted based on methods such as scale-invariant feature transform (SIFT) (*Lowe D.G., 2004*) and histogram of oriented gradient (HOG) (*Dalal N., &Triggs B., 2005*), and then the object in the image is classified using methods such as support vector machine (SVM) *(Burges C.J., 1998*) and adaptive boosting (Adaboost) *(Freund Y., & Schapire R.E., 1995*). *Ji et al. (2012)* proposed an apple recognition algorithm based on an SVM by extracting the color and shape features of the fruit in the image, and got good results on apple picking robots. However, the strategies for handling situations involving leaf shading and densely overlapping fruits are still not perfect and fall short compared to extreme conditions in real orchards. *Si et al. (2015)* proposed a fruit recognition algorithm based on color difference and its ratio, which determines the fruit by calculating the color difference and its ratio of pixels. It effectively eliminates the influence of shadows and soil in the natural environment, and achieves a recognition rate of 85%. However, the experiment used a small data scale, a single experimental scenario, and lacked systematic verification under complex conditions. *Mehra et al. (2016)* transformed images from RGB to Lab* color space, clustered pixels using k-means clustering, and then determined maturity based on color characteristics of these clusters. By integrating tomato maturity assessment with disease detection within a single CV framework, they demonstrated the practical value of multi-task recognition while validating the feasibility of image features for fruit quality monitoring and health evaluation. This approach exhibits significant systemic innovation and application significance. However, the feature extraction and classification methods employed exhibit insufficient robustness to lighting variations, fruit occlusion, and scenarios involving multiple concurrent diseases. Furthermore, the limited dataset size and diversity weaken the model's generalization capability, making it difficult to detect automated in real environments. *Al-Mashhadani et al. (2020)* used OpenCV combined with the HSV color space to estimate tomato ripeness and perform counting. They determined the colors of ripe and nearly ripe fruits through thresholding in the HSV color space. While this technique could identify fruits that were ripe and about to change color when grouped, it could not distinguish differences between them.

Traditional algorithms always rely on manually designed feature extraction methods, which are prone to false positives and false negatives when faced with complex scenarios such as changes in light, occlusion, and overlap. They also suffer from high computational complexity, low robustness, and poor adaptability. Although traditional image processing methods based on manual design have the advantages of simple implementation, low computational cost, and low hardware requirements, and have played an important role in early research, their limitations, such as slow processing speed and high computational complexity, have become increasingly apparent (*Bello R.-W. et al., 2025*).

Compared with traditional recognition methods, DL-based methods are self-learning directly from the image's own features and expression relationships. They have a strong ability to express fruit features, high generalization, and high accuracy, solving the problem that traditional methods cannot meet the real-time requirements. At the same time, with the improvement of computer computing power, more researchers have turned their attention to DL. *Girshick et al. (2014)* proposed the R-CNN algorithm, and the use of CNNs greatly improved the speed of object detection. AlexNet *(Krizhevsky A. et al., 2017*) defeated traditional object detection algorithms in the ImageNet Large-Scale Visual Recognition Challenge, and its speed and accuracy are far superior to traditional detection algorithms. So far, DL methods have begun to be widely used in various object detection and recognition algorithms. So, there are several main categories of commonly used methods for detecting fruits on plants based on DL. First is the Two-Stage-Based Algorithm, second is the One-Stage-Based Algorithm, third is the Anchor-Free-Based Algorithm, and the last is the Transformer-Based Algorithm.

Two-Stage-Based Detection Algorithm

The two-stage object detection algorithm is a classic paradigm for object detection in CV. Its core logic is to split the detection task into two stages: "candidate region generation" and "region fine detection" to achieve high-precision object positioning and classification. With their powerful feature learning and expression capabilities, CNNs have gradually replaced traditional image processing methods and become one of the mainstream technologies for detection. As an important representative of the two-stage detection framework, the R-CNN series algorithms have promoted the advancement of object detection technology.

Girshick et al. (2014) proposed the original R-CNN framework, which used a CNN to extract high-level features of candidate regions for the first time, significantly improving detection accuracy. The original R-CNN uses selective search to generate candidate regions, extracts features one by one through CNN, and combines SVM classification and bounding box regression to optimize the detection effect.

Although it improves accuracy, it still has a high computational cost and slow speed, making it difficult to meet real-time requirements. For this, Fast R-CNN (*Girshick R., 2015*) and Faster R-CNN (*Ren S. et al., 2015*) have been introduced. Fast R-CNN reduces redundant calculations and improves detection speed by sharing feature maps and ROIPooling. Faster R-CNN is later than Fast R-CNN. It introduces the region proposal network (RPN) to achieve end-to-end training of candidate region generation and detection, further improving efficiency and accuracy, realizing the integration of detection processes, and laying the foundation for a two-stage detection framework. Mask R-CNN (*He K. et al., 2017*) introduced an instance segmentation function based on Faster R-CNN, adds a mask branch, replaces ROIPooling with ROIAlign to eliminate quantization error, and realizes end-to-end training of object detection, classification, and instance segmentation. It has high accuracy and is extended to tasks such as panoramic segmentation, becoming the basic framework for multimodal vision tasks. It is particularly suitable for scenes with densely clustered fruits, and can effectively distinguish overlapping fruits and improve detection results. Figure 2 shows the basic principle of the two-stage-based detection algorithm.

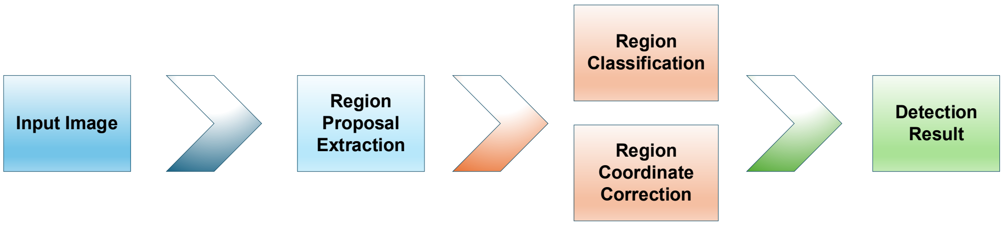

Fig. 2 - The basic principle of the two-stage-based detection algorithm

Ganesh et al. (2019) implement orange detection and instance segmentation based on Mask R-CNN. Its advantage lies in the use of a powerful two-stage segmentation framework, which enables the model to maintain high segmentation accuracy under complex backgrounds. However, Mask R-CNN itself has a large structure and slow inference speed, which makes it difficult to meet the application requirements of real-time picking or drone inspection. Furthermore, it doesn’t conduct in-depth discussions on the model's cross-scene generalization, robustness, and data diversity. *Afonso et al. (2020)* proposed a tomato detection algorithm based on the Mask R-CNN approach. This algorithm detects objects and their corresponding pixels, utilizing RealSense cameras to capture tomato images from greenhouses for identifying four distinct stages of tomato ripeness. Results demonstrate the model's excellent performance in detecting and classifying tomatoes within the challenging greenhouse environment, indicating that Mask R-CNN can implicitly learn object depth—a critical factor for background elimination. However, the model exhibits strong dependence on the dataset and lacks validation for generalization. *Wang et al. (2021)* integrated an attention mechanism into the Faster R-CNN model to address the challenge of detecting young tomato against near-color backgrounds. Results show this method achieves a mAP of 98.46% with an average detection time of just 0.084 seconds per image. It enables real-time, accurate, and efficient detection of young tomato fruits, providing a robust solution for fruit detection under near-color backgrounds. But the more complex model structure increases hardware costs and deployment complexity. *Zu et al. (2021)* built a model using Mask R-CNN with a ResNet50-FPN backbone to detect and segment green tomatoes in greenhouses. The model uses ROIAlign bilinear interpolation to compute object regions generated from training feature maps. It achieves detection and segmentation of ripe green tomatoes through parallel operations of ROI object classification, bounding box regression, and mask generation. At an IOU threshold of 0.5, the model achieves an F1 score of 92.0%. And the model remains limited as experiments were conducted solely on green tomatoes, and performance degrades when fruits overlap. *Xu et al. (2022)* used a bimodal feature map and a balanced multi-task loss model, which improved stem segmentation accuracy during cherry tomato harvesting. Nevertheless, challenges persist regarding color and shape similarity, loss of stem features, and category differences. *Yan et al. (2024)* built a comprehensive recognition framework based on an improved Faster R-CNN, integrating apple detection, quantity statistics, localization, and quality estimation. Its advantage lies in unifying multiple tasks into a single model. Through feature pyramid optimization and region proposal modules, it improves detection accuracy and stability in complex orchard environments, effectively addressing real-world challenges such as occlusion and varying lighting conditions. However, while multi-task coupling enhances functionality, it also increases training complexity. Furthermore, the mechanisms of interaction between tasks have not been systematically analyzed, and the limited dataset size means that the demonstration of generalization ability remains insufficient.

Cong et al. (2025) proposed the TQVG Model based on Mask R-CNN for tomato quality grading. Results demonstrate that the model achieves higher mAP and faster inference speeds with fewer parameters. The model exhibits high dependence on training data, sensitivity to annotation quality, and lacks performance validation for real-time and mobile deployment. Furthermore, optimization is still needed in terms of inference efficiency.

While the R-CNN series offer high detection accuracy, they still have some significant limitations for widespread application in fruit detection. Their reliance on a two-stage process of "candidate region generation + classification and regression" results in high overall computational cost, slow inference speed, large file size, and cumbersome workflow, making edge deployment difficult and hindering real-time detection. Although improvements over the original R-CNN, independent steps such as candidate region generation still exist, leading to complex structures, high training costs, high computational resource requirements, and difficulties in end-to-end optimization. They also tend to miss information about dense, small objects, making them less effective against both small and dense objects.

Therefore, although they have significantly advanced object detection technology by optimizing feature extraction and candidate region generation, and perform well in high-accuracy detection and robustness against complex backgrounds, real-time deployment on mobile terminals and edge devices remains challenging due to limitations in detection speed and model size. So, it is necessary to combine lightweight design with efficient feature extraction strategies to further improve their practicality and versatility. Figure 3 shows the mAP values of some Two-stage detectors on the COCO dataset.

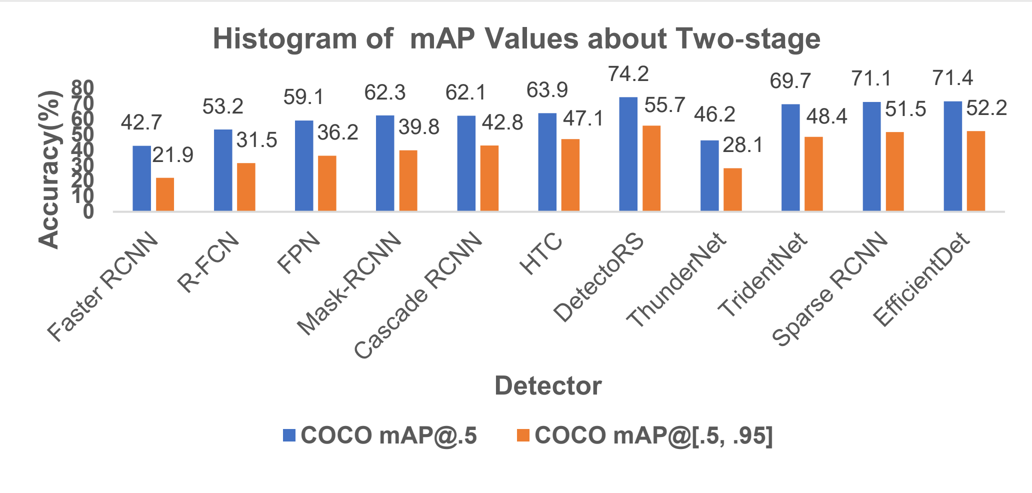

Fig. 3 - The mAP values of some Two-stage detectors on COCO dataset. There're Faster-RCNN *(Ren S. et al., 2015*), R-FCN (*Dai J. et al., 2016*), FPN (*Lin T.-Y. et al., 2017*), Mask-RCNN (*He K. et al., 2017*), Cascade RCNN *(Cai Z., Vasconcelos, N., 2018*), HTC (*Chen K. et al., 2019*), DetectoRS (*Qiao S. et al., 2021*), ThunderNet *(Qin Z. et al., 2019*), TridentNet (*Li Y. et al., 2019*), Sparse RCNN (*Sun P. et al., 2021*), EfficientDet (*Tan M. et al., 2020*).

One-Stage-Based Detection Algorithm

The regression-based one-stage fruit detection and recognition method does not need to generate candidate regions, but directly obtains the category probability and position of the object, making its network structure simpler and sacrificing some accuracy, but the detection and recognition speed of the algorithm is improved.

YOLO *(Redmon, J. et al., 2016)* series are a typical representative of one-stage detection. With its end-to-end structure, fast speed, and high accuracy, it has become the mainstream technical route for fruit detection. Since YOLO was proposed, it has been iterated on many versions and widely used in agricultural visual inspection, especially in real-time and deployment flexibility scenarios. With the iteration, the YOLO series continues to optimize detection accuracy, efficiency, and generalization ability. YOLO9000 *(Redmon J., & Farhadi A., 2017*) adopted Darknet-19, anchor box and multi-scale training to improve the performance of small object detection; YOLOv3 *(Redmon J. & Farhadi A., 2018*) adopted a deeper Darknet-53 to enhance feature expression; YOLOv4 (*Bochkovskiy A. et al., 2020*) and YOLOv5 integrated CSPNet, Mish activation function, and Mosaic data enhancement, significantly improving speed and accuracy, and are widely used; YOLOv7 (*Wang C.-Y. et al., 2023*) proposed the ELAN (Efficient Layer Aggregation Network) architecture, which improves feature utilization through grouped convolution and residual connection, and performs expansion and compound scaling, supplemented by dynamic label allocation optimization, with strong robustness, and can adapt to complex lighting and occlusion scenes.

YOLOv8 adopts a unified framework for all tasks, enhanced features, and improved loss functions, and introduces adaptive strategies to achieve real-time deployment optimization. And YOLOX (*Ge Z. et al., 2021*) in the YOLO series is also a model based on an anchor-free architecture. YOLOX divides the input image into multiple smaller grids, so that predictions can be made quickly and grid by grid. YOLOX-PAI (*Wu Z. et al., 2022*) is an improved version of YOLOX. Its core part uses an adaptive spatial feature fusion strategy for feature enhancement and an attention mechanism to organize tasks. Figure 4 shows the basic principle of the one-stage-based detection algorithm.

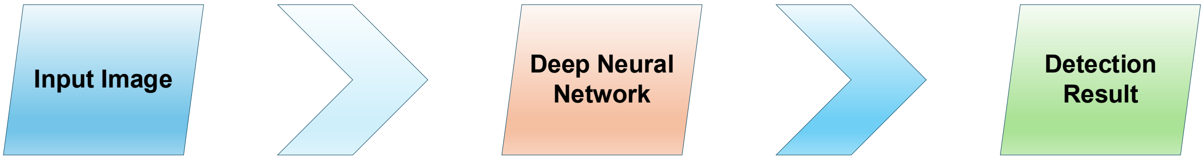

Fig. 4 - The basic principle of the one-stage-based detection algorithm

Mirhaji et al. (2021) utilize various YOLO models to detect and estimate fruit load in citrus orchards, and systematically evaluate the impact of different imaging methods and lighting conditions on detection performance. The results show that the model's precision, recall, F1 score, and mAP are 91.23%, 92.8%, 92%, and 90.8%, respectively, and the overall performance difference between nighttime and daytime imaging is not significant. But the model lacks specific optimization for orchard shading, dense fruit distribution, or strong light interference, and its robustness requires further verification.

Hernández et al. (2023) proposed a tomato ripeness detection model based on YOLOV3-tiny, achieving an F1 score of 90%. Although YOLOv3-tiny is sufficiently lightweight to significantly boost inference speed while maintaining detection accuracy, its overly simplified architecture limits feature extraction capabilities. This results in insufficient recognition accuracy under complex lighting, occlusion, and color transition conditions, necessitating further improvements in model robustness.

Zhang et al. (2023) employed YOLOv4-Tiny + FENB to detect tomatoes under occluded backgrounds during both daytime and nighttime. The nighttime mAP value reached 94.72%, with test results consistently outperforming both YOLOv4 and YOLOv4-Tiny, demonstrating the model's robust performance under nighttime and occluded conditions. However, the authors provided limited validation of the model's robustness and adaptability in complex backgrounds.

Li et al. (2023) proposed the YOLOv5s-Tomato model to identify four distinct tomato maturity stages. They enhanced small object detection performance using techniques like Mosaic data augmentation, the Focus module, and the CSPNet network architecture, while employing EIoU as the loss function. This approach yielded significant improvements in model size and detection speed. Results demonstrate that the model effectively meets greenhouse detection requirements while improving recognition accuracy for occluded and small tomatoes. Since all training images originated from a single greenhouse, further research is needed to evaluate its generalization in complex environments.

Zhang et al. (2024) proposed a cascaded DL network based on YOLOv5 for tomato detection and classification, achieving 82.4% accuracy and 90.9% recall in detecting unobstructed tomatoes. Additionally, the network achieved an average accuracy of 96.9% for tomato ripeness classification and a detection speed of 20 FPS. The validation dataset primarily relied on a single greenhouse scenario, leaving its generalization capability in complex backgrounds insufficiently validated. Furthermore, the analysis of trade-offs between accuracy, stability, and real-time performance for pose classification was inadequate.

Wang et al. (2024b) proposed the NVW-YOLOv8s model for real-time detection and segmentation of tomatoes at three distinct ripeness stages. This model addresses challenges such as sample imbalance, small object detection, and occlusion by introducing a foreground-foreground category balancing method and the C2f-N module, significantly enhancing detection and segmentation performance. The results show that both the mAP@0.5 and F1-Score outperform the baseline model YOLOv8s. However, this study lacks validation of the model's generalization ability across different geographical regions, growth conditions, or tomato varieties.

Wu et al. (2024) proposed a tomato ripeness recognition and stem detection model to address occlusion and overlap issues. The model incorporates CBAM attention and EfficientVMamba modules to enhance feature extraction from tomatoes at different maturity stages and improve small object detection. Experimental results demonstrate that the model achieves higher mAP, precision, recall, and F1 scores than the original YOLOv8n model when identifying tomatoes of varying ripeness in complex environments.

However, due to the limited diversity of scenarios, further validation of robustness and generalization capabilities is lacking.

Zhao et al. (2025) proposed a more lightweight and efficient ripeness detection algorithm, YOLO-DGS, to achieve precise detection of tomato fruit ripeness in natural environments. Experimental results demonstrate that the improved model achieves notable enhancements in inference speed, parameter count, F1 score, recall, and mAP. The study does not explore the lightweight nature or its application on edge devices.

Teng et al. (2025) propose a lightweight improved YOLOv8 model, DS-YOLO, for efficient strawberry fruit detection. By introducing depthwise separable convolutions and a lightweight feature fusion module, the model significantly reduces the number of parameters and computational cost, while maintaining or slightly improving detection accuracy and achieving higher real-time performance.

Experimental results show improvements in mAP@0.5, recall, and precision, while reducing the number of parameters and computational cost. Limited by the relatively small experimental dataset and limited scenarios, the model's generalization ability was not adequately validated. Furthermore, the model's robustness improvement was also limited for small strawberries, dense fruit clusters, and extreme lighting conditions.

The YOLO series are driving the development of detection technology towards real-time, lightweight, and industrial applications thanks to its excellent balance between speed and accuracy, flexible architecture, and complete ecosystem. However, due to its high downsampling and low feature map resolution, the YOLO series algorithms are prone to overlooking small objects or inaccurately locating them. Detection quality tends to degrade when objects are densely packed, overlapping, or when there is strong background interference. Since it relies primarily on manual resolution of class imbalances, its generalization ability is limited, and lightening the model significantly reduces its performance, making it difficult to achieve both accuracy and speed. In the future, it is hoped that solutions or approaches such as integrating Transformers, NAS optimization, and integrated detection-segmentation-tracking functions can further improve the performance of intelligent fruit detection models. Figure 5 shows the mAP of One-stage detectors on COCO dataset.

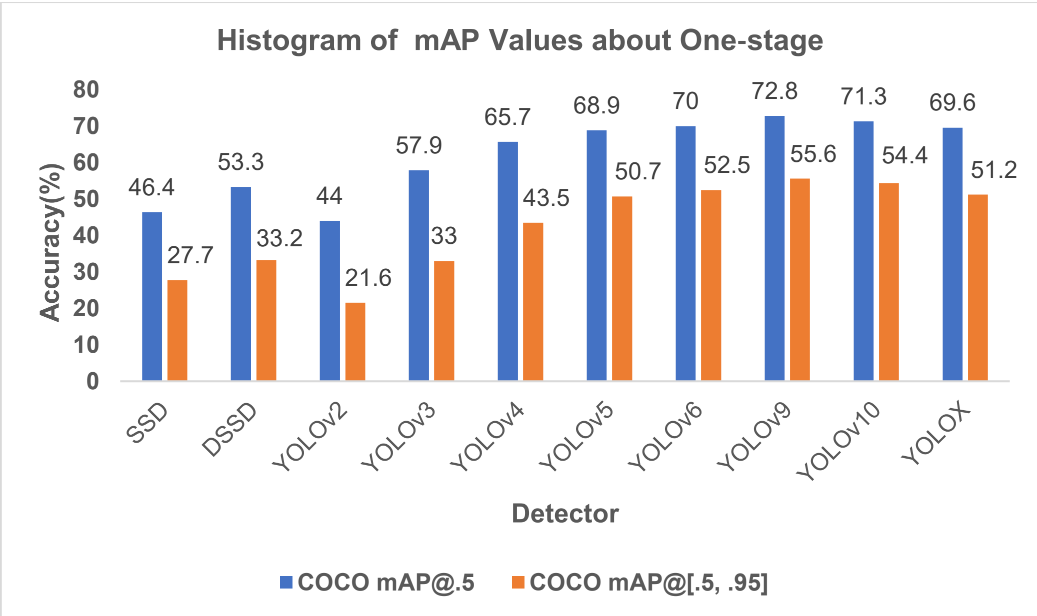

Fig. 5 - The *mAP* of One-stage detectors on *COCO* dataset.

There're SSD (*Liu W. et al., 2016*), DSSD *(Fu C.-Y. et al., 2017*), YOLOv2 (*Redmon J. & Farhadi, A., 2017*), YOLOv3 (*Redmon J. & Farhadi, A., 2018*), YOLOv4 (*Bochkovskiy A. et al., 2020*), YOLOv5 (*Jocher G. et al., 2022*), YOLOv6 (*Li C. et al., 2022*), YOLOv9 *(Wang C.-Y. et al., 2024*), YOLOv10 (*Wang A. et al., 2024a*), YOLOX (*Ge Z. et al., 2021*).

Anchor-Free-Based Detection Algorithm

Although the anchor-free-based algorithm can select positive samples for box regression, it relies on pre-defined boxes and brings some obvious limitations. When too many negative samples are generated, there will be class imbalance; the anchor box of the object class also has the problem of object positioning; the hyperparameters introduced by the anchor box can only be adjusted based on experience, which may be a huge challenge when the workload is large. Based on the idea of eliminating the pre-defined anchor box, the anchor-free-based algorithm is proposed to solve these potential problems. The algorithms can be divided into key-point-based methods and anchor-point-based (also called center-point-based) methods.

Key-point-based detection algorithms use multiple preset keypoints to detect objects and make bounding box predictions based on grouping these keypoints. The first algorithm is CornerNet *(Law H., Deng, J., 2018*), which uses keypoint pairs at the top left and bottom right corners of the bounding box to detect objects and introduces a pooling layer to better locate the bounding box corners. And the CornerNet-Lite *(Law H. et al., 2019*) improves efficiency by introducing an attention mechanism to reduce the number of pixels processed. Then, Matrix networks (xNets) (*Rashwan A. et al., 2019*) provide a lightweight architecture based on scale and aspect ratio, making detection faster. Besides, CenterNet (*Duan K. et al., 2019*) addresses the limitations of CornerNet by adding a third keypoint in the bounding box.

Zhao & Yan (2021) selected the CenterNet framework as the model for solving the problem of fruit detection in digital images. It achieves end-to-end prediction in a single-stage detection process, omitting the candidate box generation step of the traditional two-step method and improving detection speed.

Experiments demonstrate that the model can maintain high detection accuracy even in complex backgrounds, reflecting the potential of CenterNet for small and dense objects. The experiments lack comparisons with more advanced or lightweight models, making it difficult to evaluate their practical advantages in current fruit detection research. Furthermore, the detection capability for multiple categories or fruits of different ripeness levels is not verified to validate the model's generalization ability, and its robustness in occluded, low-light, or multi-view scenarios is not validated. *Wang et al. (2024)* proposed a novel model called CenterNet-LW-SE based on CenterNet. Experimental results show that the model outperforms the original CenterNet in precision, recall, model size, and detection speed, achieving a 2.1% performance improvement while reducing model size by 16.69% and shortening average detection time by 30.40%. It enhances feature representation capabilities to address the challenges of small objects, strong camouflage, and complex backgrounds of camellia fruit in natural environments, particularly excelling in fine-grained feature enhancement. However, the overall model architecture still relies on CenterNet, and it lacks sufficient comparison with more advanced lightweight models (such as YOLOv8n, RT-DETR, and MobileOne), making it difficult to demonstrate its competitive advantage among current lightweight detection frameworks. Generalization validation is insufficient, lacking cross-regional, cross-seasonal, and cross-device dataset testing; model stability requires further verification. Although it claims to be lightweight, it does not provide a complete evaluation of inference time and hardware deployment, lacking edge performance metrics to support its deployment value. *Fan et al (2024)* employed the CenterNet neural network as the detection framework, which improved keypoint prediction accuracy while maintaining low computational cost and achieving a lightweight design. They also implemented structural reinforcement and feature enhancement for complex orchard environments, significantly improving the model's robustness in unstructured environments. Experimental results show that the model achieves an AP of 96.80% on the test set, with 18.91 FPS, a size of only 17.56 MB, and a RMSE of 4.405 mm for localization accuracy. Due to its relatively limited testing scenarios, its generalization and robustness under complex occlusion conditions are an important challenge.

Anchor-point-based detection algorithms use preset anchors and perform detection pixel by pixel. DenseBox (*Huang L. et al., 2015*) is an excellent anchor-point-based detection algorithm. And some years later, the fully convolutional single-stage (FCOS) (*Tian Z. et al., 2019*) model was proposed as a representative model, which uses FPN for multi-level prediction and forces classification of positive and negative anchors, and adds a prediction branch to the classification branch to eliminate the interference of anchors. The FCOS model is not only fast in calculation speed, but also can be customized to meet more requirements.

Liu et al. (2022) propose an accurate detection and segmentation method based on the FCOS object detection model for detecting and segmenting occluded green fruits. Results show that on the apple dataset, the detection and segmentation accuracies for green fruits reach 81.2% and 85.3%, respectively. On the Apple-ape dataset, the detection accuracy and segmentation accuracy reach 77.2% and 79.7%, respectively. The model has 39.7 M parameters and 169.8 G floating-point operations. The model addresses the problem of green fruits easily blending with the background leaves in natural scenes, resulting in weak contrast. An improved model is proposed to significantly enhance the distinguishability of fruits under the "occlusion + same-color background" condition. Combined with enhanced feature extraction and segmentation mechanisms (such as attention modules or multi-scale fusion), it effectively and robustly improves the segmentation boundary quality for fruits with weak texture and partial occlusion. But the model structure is relatively complex and computationally intensive, and the dataset coverage is still limited, so the bottlenecks in lightweight design and adaptability may still exist.

Zhao et al. (2023) propose a novel fruit detection model called FCOS-LSC based on FCOS to better address the challenge of identifying green fruits. This model can directly predict the object's center point and size, avoiding the complexity and parameter tuning issues of traditional anchor-box methods. It effectively enhances the ability to identify low-contrast fruits in complex backgrounds and efficiently identifies and locates green fruit images affected by overlapping occlusion, lighting conditions, and shooting angles. Experiments are insufficient to verify the model's generalization ability under different lighting conditions, different varieties, or multiple fruit categories, indicating room for improvement in robustness. Specific analysis of deployment efficiency, model parameter count, and computational cost is lacking, limiting its guidance for practical mobile or edge applications. Figure 6 shows the mAP of Anchor-free detectors on the COCO dataset.

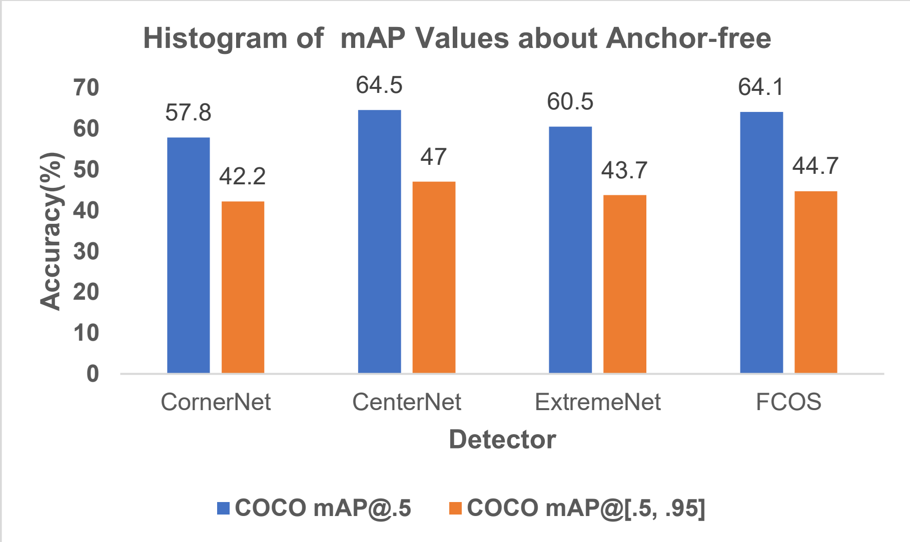

Figure 6. The mAP of Anchor-free detectors on the COCO dataset. There're CornerNet (*Law H., &Deng J., 2018*), CenterNet (*Zhou X. et al., 2019a*), ExtremeNet (*Zhou X. et al., 2019b*), FCOS (*Tian Z. et al., 2019*).

Transformer-Based Detection Algorithm

With the widespread application of the Transformer architecture in the field of CV, object detection technology has achieved a breakthrough. In 2020, the Facebook AI Research team proposed DETR (DEtection TRansformer) (*Carion N. et al., 2020*), which for the first time used a full Transformer structure to achieve end-to-end detection, eliminating traditional modules such as anchors and NMS. DETR has opened up a new path for object detection, especially in complex backgrounds and overlapping occlusions, showing good robustness and high accuracy, and has attracted attention in the field of intelligent agricultural monitoring.

DETR is the first end-to-end framework that introduces the Transformer into object detection. It eliminates NMS and anchor boxes through binary matching, directly outputs detection results, and simplifies the process. The first generation of DETR has high accuracy but slow training and limited small object detection capabilities, which affects the promotion of agricultural applications; the subsequent Deformable DETR *(Zhu X. et al., 2020*) proposed a deformable attention mechanism, introduced deformable convolution to strengthen local feature learning, and only focused on key sampling points, greatly improving small object detection performance and training efficiency; Conditional DETR *(Meng D. et al., 2021*) accelerated training convergence through the conditional cross-attention mechanism, which is faster than DETR; DN-DETR *(Li F. et al., 2022*) proposed a denoising training strategy, accelerated convergence by adding noise queries and reconstructing labels, and enhanced cross-scenario generalization capabilities through dynamic training strategies; DINO (*Zhang H. et al., 2022*) integrated DN-DETR's denoising training and conditional attention to achieve SOTA performance and support few-sample detection; RT-DETR (*Zhao Y. et al., 2024*) was improved based on DINO to achieve real-time end-to-end detection, balance accuracy and speed, and improve model performance; ViDT++ (*Song H. et al., 2021b*) combines visual Transformer and DETR to propose a sparse attention mechanism to achieve efficient detection on high-resolution images (*Zhu X. et al., 2020*). The DETR series is easy to integrate with multi-task learning, supports fruit detection, maturity grading, and size estimation, promotes the integrated development of intelligent monitoring, and is evolving towards lightweight and edge device adaptation, with good industrial application prospects. The Figure 11 shows the architecture of RT-DETR, and Figure 12 shows the mAP values of some classic transfer learning-based object detection models on the coco dataset.

Huang et al. (2024) designed a lightweight Transformer architecture-based pear detection model for natural environments, using the RT-DETR model. The improved model achieves accuracy, recall, and mean accuracy of 93.7%, 91.9%, and 98%, respectively. Compared with the original model, the number of parameters, computational cost, and weight memory are reduced by 48.47%, 56.2%, and 48.31%, respectively. This successfully combines the strong feature modeling capabilities of the Transformer with real-time requirements, alleviating the problems of large parameter count and high computational consumption in traditional ViT models. This research does not fully address the issue of insufficient training data, which may affect the model's generalization ability, and its adaptability to scenarios with extreme lighting changes and dense fruit occlusion is insufficient, requiring further strengthening of robustness. The model complexity is still higher than most lightweight CNNs, limiting its feasibility for deployment on low-power MCUs or edge devices with extremely low computing power. *Zhang (2025)* proposed PG-DETR based on DETR for detecting small, occluded early-stage pomegranate fruits in complex orchard environments. Experimental results show that PG-DETR requires only 6.7 million parameters and has a computational cost of 12.4 GFLOPs/s, achieving a mAP50 of 93.37%. Compared to the RT-DETR-R18 baseline model, our model reduces parameters by 66.2%, GFLOPs by 78.2%, and inference speed by 3.2 times, reaching 173.58 FPS. Despite its "lightweight" design, the DETR framework still incurs some computational overhead compared to YOLO-like frameworks, and its deployment advantage on truly low-power devices is not significant. Furthermore, the model's training is highly dependent on factors such as convergence speed and data size requirements, and its generalization ability may still be limited in small-sample agricultural scenarios. *Tang et al. (2025)* proposed a novel multi-scale lightweight DETR (ML-DETR) for the identification of ripe cherry tomatoes. The ML-DETR model achieves an accuracy of 91.36%, an mAP50 of 0.8875, a floating-point computation of 13.9 GFLOPs, and a frame rate of 81.2 FPS, outperforming the original RT-DETR model. This research primarily focuses on ripe fruit detection, with insufficient validation of its generalization ability to immature, occluded, and densely packed fruits, and incomplete coverage of real-world application scenarios. Furthermore, deployment experiments on real-world edge hardware are lacking, and the value of its lightweight design has not been fully validated in terms of actual device power consumption or real-time performance. *Liu et al. (2024)* proposed a green apple detection model based on an optimized deformable DETR. Experimental results show that the improved detection model significantly improves accuracy, with an overall AP of 54.1, an AP50 of 80.4, an AP75 of 58.0, an AP of 35.4 for small objects, an mAP of 60.2 for medium objects, and an AP of 85.0 for large objects. The model, based on a deformable DETR structure optimization, effectively enhances the detection capability for low-contrast, camouflaged objects like green apples, especially performing well in natural environments with leaf shading and complex lighting conditions. However, the experimental comparison range is relatively narrow, mainly focusing on green apples; cross-crop generalization ability was not tested, and scalability and universality are insufficient. Furthermore, the model's actual deployment performance is lacking, and the real-time harvesting or monitoring needs in agricultural scenarios have not been fully validated. Figure 7 shows the mAP values of some classic transfer learning-based object detection models on the coco dataset.

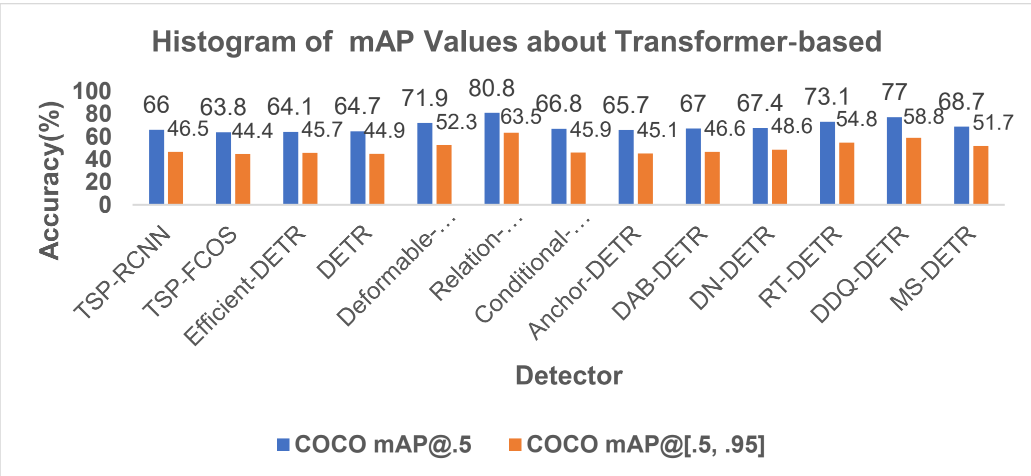

Fig. 7 - The mAP values of some classic Transformer-based detectors on the COCO dataset
They're TSP-RCNN (*Sun Z. et al., 2021*), TSP-FCOS (*Sun Z. et al., 2021*), Efficient-DETR *(Yao Z. et al., 2021*), DETR *(Carion N. et al., 2020*), Deformable-DETR (*Zhu X. et al., 2020*), Relation-DETR (*Hou X. et al., 2024*), Conditional-DETR *(Meng D. et al., 2021*), Anchor-DETR (*Wang Y. et al., 2022*), DAB-DETR (*Liu S. et al., 2022*), DN-DETR *(Li F. et al., 2022*), RT-DETR (*Zhao Y. et al., 2024*), DDQ-DETR (*Zhang S. et al., 2023*), MS-DETR (*Zhao C. et al., 2024*).

Key Challenges

Fruit detection technology can realize automatic fruit recognition, counting, and status analysis through DL algorithms, and is widely used in fields such as yield estimation, intelligent picking, and quality grading. Algorithms such as YOLOv8 and Faster R-CNN achieve a balance between detection accuracy and speed through multi-scale feature fusion and attention mechanism optimization. The Anchor-Free architecture is more adaptable to irregular fruits, and the introduction of the Transformer architecture improves the segmentation of overlapping and occluded objects. Current technology faces challenges such as missed detection of small objects, complex background interference, and insufficient real-time performance. Table 1 shows the main characteristics, advantages, and disadvantages of the backbone networks of some typical object detection models.

Table 1

The main characteristics, advantages, and disadvantages of the backbone networks of some typical object detection models.

Author	Net	Innovations	Advantages	Disadvantages
(Krizhevsky, A. et al., 2012)	AlexNet	The first employed a deep CNN, used sequential convolutional layers, and used a wide sensory field.	First pioneering DL for visual tasks; overall ease of understanding and implementation of the model; significant results for small datasets.	Although a deep network is used, the network has only 8 layers with limited depth; the computational cost of the model is high.
(Simonyan, K., & Zisserman, A., 2014)	VGGNets	Modeled with small convolutional filters (3×3); built a depth architecture; innovative use of miniature convolutional filters.	Strong model performance; performs well across datasets; good at capturing spatial hierarchies.	The model requires high computational power; the number of parameters in the model is large; the model is memory-intensive.
(He, K. et al., 2016)	ResNets	Innovative use of residual connectivity; first use of ultra-deep networks; enables batch normalization of applications.	Effectively mitigates the gradient vanishing problem; the model is highly scalable; it is faster and smaller compared to VGG.	The model requires more computational power; the model is too parameterized; the model is complex to train.
(Szegedy, C. et al., 2017)	Inception-ResNet	Used the Inception module with residual concatenation; established a multi-scale feature extraction mechanism; applied the residual Inception module	The model has more accuracy; the model is very efficient in extracting features; it effectively combines the advantages of Inception and ResNet.	Higher complexity of model architecture; high model requirements on computational resources; smaller range of applications
(Huang, G. et al., 2017)	DenseNets	The model has a compact architecture, and the input of the next layer comes from the output of the previous layer.	The model achieves efficient reuse of features, etc.; the model effectively mitigates the gradient vanishing problem.	Models are memory-intensive; internal connections are complex; memory consumption is high.
(Hu, J. et al., 2018)	SENet	Enables channel-level feature recalibration and lightweight modules	Improved channel interdependency and accuracy while enabling efficient integration with other architectures	Increases computational overhead and thus training time; slower than ResNets
(Tan, M., &Le, Q., 2019)	EfficientNets	A composite scaling method is used; depth separation of the convolution is achieved; and it is lightweight and efficient.	Higher precision and better efficiency with fewer parameters.	Requires tuning of scaling parameters; costly for computation and data movement; one-dimensional scaling only

Author	Net	Innovations	Advantages	Disadvantages
(Wang, C.-Y. et al., 2020)	CSPResNeXt	Implementing partial connectivity across stages, aggregating residual transforms; class does EfficientNet-like scaling	Reduced memory costs while not only maintaining accuracy but also obtaining better inference rates	Increased complexity, long training time, edges, and poor generalization
(Wang, C.-Y. et al., 2021)	CSPDarkNet	Partial connectivity enhancement across phases; available for YOLOv4; efficient feature extraction	Effectively balances speed and accuracy; reduces memory costs; can be used for real-time inspection programs.	Increased design complexity of the architecture leads to reduced ease of use and increased training time.
(Liu, Z. et al., 2022)	ConvNeXt	Simpler architecture with fewer activation and normalization layers	High performance; simple design; good generalization across different datasets	Requires more validation, high computational effort, and high computational power requirements
(Oquab, M. et al., 2023)	DINOv2	A self-supervised vision converter is employed to enable more efficient feature extraction.	Strong generalization ability and self-supervised learning without labeled data	It is data and computation-intensive, with high equipment requirements, while being less effective for small data sets.
(Matszangosz, K.K., &Wendt, M., 2024)	SAM	It's a base model created by the Facebook team that can be used in a wide range of applications.	Highly adaptable and can be used without adjustment or with fine-tuning	Extremely demanding on computing and storage resources, and real-time detection is not currently possible

Difficulty in Dense Detection of Small Objects

In the fruit detection task, “small object dense detection” is a key challenge, where the fruit size is often less than 2%-5% of the image resolution, and in the processing is easy to lose the features, resulting in a decline in the detection accuracy or even failure to detect the object. At the same time, the dense cluster distribution of fruits, accompanied by mutual overlap and branch and leaf occlusion, makes it difficult for the detection network to distinguish between neighboring small objects, and easy to miss detection and false detection. Researchers use multi-scale feature fusion, high-resolution input, attention mechanism, Deformable Convolution, and Transformer feature enhancement to improve the detection of dense small objects *(Bai Y. et al., 2024; Luo Q. et al., 2024; Song H. et al., 2021a*). However, the detection of dense small objects in complex environments is still a bottleneck (*Gong X., &Wu Q.*).

The Effects of Occlusion and Overlap

Fruits grow in the plant, are mostly obscured by leaves, or often grow in clusters and densely, which brings about seriousness between pairs of branches and leaves and fruits, resulting in blurred fruit outlines and making it difficult to distinguish the boundaries, which becomes a key challenge affecting the detection accuracy. Traditional detection methods and early CNN models are difficult to effectively deal with occluded or overlapping objects due to the limited local feature expression capability, and are prone to miss detection and false detection, while networks such as YOLO series and Faster R-CNN can hardly accurately distinguish overlapping objects (*Dalal M., Mittal P., 2025*).

To improve the detection performance, the researchers introduced the attention mechanism to enhance the attention of occlusion area, deformable convolution to improve the ability of local complex morphology perception, combined with instance segmentation to achieve pixel-level differentiation, and utilized the Transformer architecture to enhance the model's understanding of the complex spatial relationship between branches, leaves, and fruits (*Lin Y. et al., 2024; Singh P., Krishnamurthi R., 2024*). Despite these advancements, occlusion and object overlap remain significant challenges affecting detection robustness. Future work should focus on optimizing model architectures and data augmentation strategies to improve detection performance in complex environments.

The Impact of Light Variations

Light variations (e.g., sunlight intensity, shadows, high light reflections, backlighting, and low light) are the key challenges affecting the stability and generalization ability of the detection model, especially more prominently in the open-air planting and greenhouse transparent coverage models, where the relationship between light and shadow is complex. Under strong sunlight, high light on the fruit surface leads to color distortion; low light or a cloudy environment reduces image contrast and blurs the boundary; under partial shade, the fruit is easily overwhelmed by shadows, which seriously affects the detection accuracy *(Jiang C. et al., 2022*). At the same time, the change of sunlight angle makes the fruit morphology and shadows constantly changing, which puts forward higher requirements on model robustness.

Although the mainstream algorithms have a certain degree of light invariance, there are still performance fluctuations under extreme environments. The researcher has enhanced the light change data, introduced light enhancement broadening, adaptive exposure compensation, attention mechanism, and global modeling ability of the Transformer to alleviate the light and shadow interference *(Dritsas E., 2025*). However, there is still room for improvement in detection performance under strong dynamic lighting scenarios.

The Interference of Complex Background

Different planting environments pose great challenges to the generalization ability and robustness of detection algorithms (*Dickinson S., 2009*). The greenhouse environment can be controlled, with stable light, clean background, the algorithm is easy to achieve high accuracy; while the open-air environment has a complex background, changing light, messy plant branches and leaves, and large differences in fruit morphology, which often leads to the same model in the two scenarios, resulting in performance fluctuations. In addition, the color and maturity of fruits in the open-air environment are greatly affected by natural factors, and the differences between different batches and plots are significant, which can easily lead to misdetection and omission; different planting management methods also increase the difficulty of the model in adapting to changes in spatial layout.

The generalization of the model can be improved to a certain extent through large-scale multi-scene datasets, cross-domain migration learning, style migration, scene adaptive mechanism, and fusion Transformer architecture. However, highly robust detection across scenes, regions, and species is still a challenge that needs to be broken (*Edozie E. et al., 2025*).

DISCUSSION

The Bright Future of Lightweight Networks

In the field of fruit object detection, lightweight networks are becoming an important direction to promote the scale application of smart agriculture. The demand for automated devices (e.g., drones, mobile robots, etc.) in agricultural production has increased, and there is an urgent need for “high-performance + low-computing power demand” detection models (*Zhang P. et al., 2022*). Although traditional high-precision networks have high accuracy, the models are large and energy-consuming, which are not suitable for low-power devices. Lightweight network design is accelerating, striving to balance detection accuracy and operational efficiency. MobileNet, ShuffleNet, SqueezeNet, and other CNN lightweight backbone networks significantly reduce model complexity by means of deep separable convolution, channel rearrangement, etc. MobileNetV3, combined with the optimized design of NAS, has been widely used in tomato detection and grading (*Tan Y. et al., 2025*).

The lightweight versions of the YOLO series (e.g., YOLOv4-Tiny, YOLOv5-Nano, YOLOv7-Tiny, etc.) can be efficiently deployed through pruning optimization and sparse training, with detection speeds up to 30-60 FPS, and are suitable for both embedded and mobile applications (*Li Z. et al., 2024*). Lightweight Transformer architectures (MobileViT) are gradually applied to crop detection, combining global modeling capability and high efficiency, showing good prospects.

The Potential of Transformer Architecture

As a breakthrough DL architecture, Transformer has been extended from NLP to CV by virtue of its powerful global feature modeling capability through the self-attention mechanism, and has been widely used in object detection, yield estimation, and ripening grading, etc. Compared with CNNs, Transformers can effectively capture long-range dependencies in images and improve the robustness and accuracy of detection and grading under complex backgrounds, dense fruit distribution, occlusion, and multi-scale changes. Compared with CNNs, Transformers can effectively capture long-range dependencies in images and improve the robustness and accuracy of detection and grading in complex backgrounds, dense fruit distribution, occlusion, and multi-scale change scenes *(Jackulin C., Murugavalli S., 2022; Zhao Z. et al., 2023*).

The DETR series simplifies the detection process and enhances the detection of small and occluded objects through the query matching mechanism without an anchor, and has already outperformed the traditional YOLO architecture in dense detection scenes.

The Importance of Dataset Construction

In the application of fruit detection, the construction of high-quality datasets is the core foundation to promote technological progress and industrialized applications. Despite the rapid development of DL and Transformer architecture, the lack of standardized data still restricts algorithm performance and generalization ability. Currently, tomato datasets are generally characterized by fragmented data, non-standardized annotations, and a lack of multimodal and multitasking data (*Chen L. et al., 2020*), leading the algorithms into the strange circle of "good results in the laboratory, but difficult to promote in the field".

The variable environment of agricultural scenarios and cultivation management measures leads to high costs and difficulty in data collection and labeling. Data enhancement and simulation data technology are becoming an important means to break through the data bottleneck and improve model robustness. Datasets with high-quality and uniform standards are an important factor in enhancing the generalization ability of models across cultivars and environments (*Akbar J.U.M. et al., 2024*).

The Prospects of Edge Computing

Edge computing has become an important trend as the demand for real-time, portability, and low cost in smart agriculture increases (*Dhiman P. et al., 2023*). Compared with traditional cloud computing, edge computing processes data directly on local devices such as cell phones, drones, and robots, reducing transmission latency, enhancing system real-time and stability, and adapting to complex and changing agricultural environments (*Sathyadhas G.N. et al., 2025*).

Smartphones have become mainstream edge computing terminals due to their powerful computing power (GPU/ NPU), convenience, and low cost. With lightweight models (e.g., MobileNet, YOLOv5-Nano, MobileViT) and AI frameworks (TF Lite, N-CNN), cell phone apps can realize real-time fruit detection and grading, and support intelligent management in the field (*Zheng H. et al., 2025*). With the help of high-performance edge equipment, lightweight Transformer, and NAS-optimized network, object detection technology will be available to everyone and everywhere, promoting smart agriculture to a new stage of universality and efficiency (*Liu H.I. et al., 2024*).

CONLCUSIONS

This paper systematically reviews the applications and significant advancements of CV and DL in fruit detection. Fruit detection technology has rapidly evolved from traditional algorithms to advanced architectures such as CNNs, YOLO, and DETR. However, this analysis shows that these technologies all face specific trade-offs in practical applications. While traditional CNNs provide reliable feature extraction benchmarks, their computational redundancy often fails to meet the real-time requirements of mobile devices. Single-stage detectors like YOLO, by sacrificing some fine-grained accuracy, greatly improve inference speed, making them the preferred choice for edge computing; however, their performance tends to bottleneck when dealing with densely clustered or severely occluded small target fruits. Conversely, Transformer-based models (such as DETR), with their global attention mechanism, have demonstrated breakthrough accuracy in solving complex occlusion problems, but their large number of parameters and high computational cost remain the biggest obstacles to their widespread deployment on low-power agricultural devices. Therefore, at this juncture, finding the optimal balance between computational efficiency and the ability to understand complex scenarios is the core challenge currently facing practical applications. To overcome the limitations of the aforementioned single model, experts and scholars are exploring and attempting to integrate lightweight architecture with the Transformer mechanism to reshape the research paradigm in this field. This hybrid network design not only addresses the low-cost requirement for easy deployment at the edge but also ensures high detection accuracy in complex environments, making it an inevitable trend for the true implementation of smart agriculture.

REFERENCES

[1] Afonso, M., Fonteijn, H., Fiorentin, F.S., Lensink, D., Mooij, M., Faber, N., Polder, G., & Wehrens, R. (2020). Tomato fruit detection and counting in greenhouses using deep learning. *Frontiers in plant science*, *11*, 571299.

[2] Akbar, J.U.M., Kamarulzaman, S.F., Muzahid, A.J.M., Rahman, M.A., & Uddin, M. (2024). A comprehensive review on deep learning assisted computer vision techniques for smart greenhouse agriculture. *IEEE Access*, *12*, 4485-4522.

[3] Al-Mashhadani, Z., & Chandrasekaran, B. (2020). Autonomous ripeness detection using image processing for an agricultural robotic system. 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON),

[4] Bai, Y., Yu, J., Yang, S., & Ning, J. (2024). An improved YOLO algorithm for detecting flowers and fruits on strawberry seedlings. *Biosystems engineering*, 237.
[5] Bello, R.-W., Owolawi, P.A., van Wyk, E.A., & Tu, C. (2025). Object Detection Algorithms for Digital Imaging Applications: A Review. *Available at SSRN 5236924*.
[6] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y.M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
[7] Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.
[8] Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,
[9] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European conference on computer vision,
[10] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., & Ouyang, W. (2019). Hybrid task cascade for instance segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
[11] Chen, L., Penney, D., & Jiménez, D. (2020). AI for Computer Architecture: Principles, Practice, and Prospects. *Synthesis Lectures on Computer Architecture*, *15*(5), 1-142.
[12] Cong, P., Wang, K., Liang, J., Xu, Y., Li, T., & Xue, B. (2025). TQVGModel: Tomato Quality Visual Grading and Instance Segmentation Deep Learning Model for Complex Scenarios. *Agronomy*, *15*(6), 1273.
[13] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, *29*.
[14] Dalal, M., & Mittal, P. (2025). A Systematic Review of Deep Learning-Based Object Detection in Agriculture: Methods, Challenges, and Future Directions. *Computers, Materials & Continua*, *84*(1), 57-91. https://doi.org/10.32604/cmc.2025.066056
[15] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05),
[16] Devlet, A. (2021). Modern agriculture and challenges. *Frontiers in Life Sciences and Related Technologies*, *2*(1), 21-29.
[17] Dhiman, P., Kaur, A., Hamid, Y., Alabdulkreem, E., Elmannai, H., & Ababneh, N. (2023). Smart Disease Detection System for Citrus Fruits Using Deep Learning with Edge Computing. *Sustainability*, *15*(5), 4576. https://doi.org/10.3390/su15054576
[18] Dickinson, S. (2009). The Evolution of Object Categorization and the Challenge of Image Abstraction. *Object Categorization: Computer and Human Vision Perspectives*. https://doi.org/10.1017/CBO9780511635465.002
[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
[20] Dritsas, E. (2025). A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection. *Sensors*, *25*.
[21] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. Proceedings of the IEEE/CVF international conference on computer vision,
[22] Edozie, E., Shuaibu, A.N., John, U.K., & Sadiq, B.O. (2025). Comprehensive review of recent developments in visual object detection based on deep learning. *Artificial Intelligence Review*, *58*(9).
[23] Fan, P., Zheng, C., Sun, J., Chen, D., Lang, G., & Li, Y. (2024). Enhanced real-time target detection for picking robots using lightweight centernet in complex orchard environments. *Agriculture*, *14*(7), 1059.
[24] Freund, Y., & Schapire, R.E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. European conference on computational learning theory,
[25] Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A.C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.
[26] Ganesh, P., Volle, K., Burks, T., & Mehta, S. (2019). Deep orange: Mask R-CNN based orange detection and segmentation. *IFAC-PapersOnLine*, *52*(30), 70-75.
[27] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.
[28] Girshick, R. (2015). Fast r-cnn. Proceedings of the IEEE international conference on computer vision,
[29] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object

detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,
[30] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(1), 142-158.
[31] Gong, X., & Wu, Q. Fruit Detection Methods Based on Deep Learning in Agricultural Planting: A Systematic Literature Review. *IEEE Access*, *13*.
[32] Gupta, S., & Tripathi, A.K. (2024). Fruit and vegetable disease detection and classification: Recent trends, challenges, and future opportunities. *Engineering Applications of Artificial Intelligence*, *133*, 108260.
[33] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Proceedings of the IEEE international conference on computer vision,
[34] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,
[35] Hernández, G.A.A., Olguin, J.C., Vasquez, J.I., Uriarte, A.V., & Torres, M.C.V. (2023). Detection of tomato ripening stages using yolov3-tiny. *arXiv preprint arXiv:2302.00164*.
[36] Hou, X., Liu, M., Zhang, S., Wei, P., Chen, B., & Lan, X. (2024). Relation detr: Exploring explicit position relation prior for object detection. European Conference on Computer Vision,
[37] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
[38] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
[39] Huang, L., Yang, Y., Deng, Y., & Yu, Y. (2015). Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*.
[40] Huang, Z., Zhang, X., Wang, H., Wei, H., Zhang, Y., & Zhou, G. (2024). Pear fruit detection model in natural environment based on lightweight transformer architecture. *Agriculture*, *15*(1), 24.
[41] Jackulin, C., & Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, *24*, 100441. https://doi.org/10.1016/j.measen.2022.100441
[42] Ji, W., Zhao, D., Cheng, F., Xu, B., Zhang, Y., & Wang, J. (2012). Automatic recognition vision system guided for apple harvesting robot. *Computers & Electrical Engineering*, *38*(5), 1186-1195.
[43] Jiang, C., Ren, H., Ye, X., Zhu, J., Zeng, H., Nan, Y., Sun, M., Ren, X., & Huo, H. (2022). Object detection from UAV thermal infrared images and videos using YOLO models. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102912. https://doi.org/10.1016/j.jag.2022.102912
[44] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Wong, C., Yifu, Z., & Montes, D. (2022). ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo*.
[45] Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.
[46] Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
[47] Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV),
[48] Law, H., Teng, Y., Russakovsky, O., & Deng, J. (2019). Cornernet-lite: Efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*.
[49] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., & Nie, W. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.
[50] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., & Zhang, L. (2022). Dn-detr: Accelerate detr training by introducing query denoising. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
[51] Li, R., Ji, Z., Hu, S., Huang, X., Yang, J., & Li, W. (2023). Tomato maturity recognition model based on improved YOLOv5 in greenhouse. *Agronomy*, *13*(2), 603.
[52] Li, Y., Chen, Y., Wang, N., & Zhang, Z. (2019). Scale-aware trident networks for object detection. Proceedings of the IEEE/CVF international conference on computer vision,
[53] Li, Z., Wang, J., Gao, G., Lei, Y., Zhao, C., Wang, Y., Bai, H., Liu, Y., Guo, X., & Li, Q. (2024). SGSNet: a lightweight deep learning model for strawberry growth stage detection. *Front Plant Sci*, *15*, 1491706.

https://doi.org/10.3389/fpls.2024.1491706
[54] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,
[55] Lin, Y., Huang, Z., Liang, Y., Liu, Y., & Jiang, W. (2024). AG-YOLO: A Rapid Citrus Fruit Detection Algorithm with Global Context Fusion. *Agriculture*, *14*(1), 114. https://doi.org/10.3390/agriculture14010114
[56] Liu, H.I., Galindo, M., Xie, H., Wong, L.K., Shuai, H.H., Li, Y.H., & Cheng, W.H. (2024). Lightweight Deep Learning for Resource-Constrained Environments: A Survey. *ACM Computing Surveys*, *56*(10).
[57] Liu, M., Jia, W., Wang, Z., Niu, Y., Yang, X., & Ruan, C. (2022). An accurate detection and segmentation model of obscured green fruits. *Computers and Electronics in Agriculture*, *197*, 106984.
[58] Liu, Q., Meng, H., Zhao, R., Ma, X., Zhang, T., & Jia, W. (2024). Green apple detector based on optimized deformable detection transformer. *Agriculture*, *15*(1), 75.
[59] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., & Zhang, L. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.
[60] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A.C. (2016). Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14,
[61] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
[62] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*, 91-110.
[63] Luo, Q., Wu, C., Wu, G., & Li, W. (2024). A Small Target Strawberry Recognition Method Based on Improved YOLOv8n Model. *IEEE Access*, 12.
[64] Matszangosz, K.K., & Wendt, M. (2024). 4-torsion classes in the integral cohomology of oriented Grassmannians.
[65] Mehra, T., Kumar, V., & Gupta, P. (2016). Maturity and disease detection in tomato using computer vision. 2016 Fourth international conference on parallel, distributed and grid computing (PDGC),
[66] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., & Wang, J. (2021). Conditional detr for fast training convergence. Proceedings of the IEEE/CVF international conference on computer vision,
[67] Mirhaji, H., Soleymani, M., Asakereh, A., & Mehdizadeh, S.A. (2021). Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. *Computers and Electronics in Agriculture*, *191*, 106533.
[68] O'shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
[69] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., & El-Nouby, A. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
[70] Qiao, S., Chen, L.-C., & Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
[71] Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., & Sun, J. (2019). ThunderNet: Towards real-time generic object detection on mobile devices. Proceedings of the IEEE/CVF international conference on computer vision,
[72] Rashwan, A., Kalra, A., & Poupart, P. (2019). Matrix Nets: A new deep architecture for object detection. Proceedings of the IEEE/CVF international conference on computer vision workshops,
[73] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,
[74] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition,
[75] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
[76] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.
[77] Sathyadhas, G.N., Gladston, A., & Nehemiah, K.H. (2025). LwSANet: Light Weight Self-Attention Network Model to Recognize Fruits from Images. *Traitement du Signal*, *42*(1).
[78] Si, Y., Liu, G., & Feng, J. (2015). Location of apples in trees using stereoscopic vision. *Computers and*

Electronics in Agriculture, *112*, 68-74.

[79] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[80] Singh, P., & Krishnamurthi, R. (2024). IoT-based real-time object detection system for crop protection and agriculture field security. *Journal of Real-Time Image Processing*, *21*(4), 1-19.

[81] Song, H., Jiang, M., Wang, Y., & Song, L. (2021a). Efficient detection method for young apples based on the fusion of convolutional neural network and visual attention mechanism. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, *37*, 297-303. https://doi.org/10.11975/j.issn.1002-6819.2021.09.034

[82] Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., & Yang, M.-H. (2021b). Vidt: An efficient and effective fully transformer-based object detector. *arXiv preprint arXiv:2110.03921*.

[83] Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., & Wang, C. (2021). Sparse r-cnn: End-to-end object detection with learnable proposals. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[84] Sun, Z., Cao, S., Yang, Y., & Kitani, K.M. (2021). Rethinking transformer-based set prediction for object detection. Proceedings of the IEEE/CVF international conference on computer vision,

[85] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI conference on artificial intelligence,

[86] Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning,

[87] Tan, M., Pang, R., & Le, Q.V. (2020). Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[88] Tan, Y., Liu, X., Zhang, J., Wang, Y., & Hu, Y. (2025). A Review of Research on Fruit and Vegetable Picking Robots Based on Deep Learning. *Sensors (Basel)*, *25*(12), 3677. https://doi.org/10.3390/s25123677

[89] Tang, Z., Fang, L., Sun, S., Gong, Y., & Li, Q. (2025). ML-DETR: Multiscale-Lite Detection Transformer for identification of mature cherry tomatoes. *IEEE Transactions on Instrumentation and Measurement*.

[90] Teng, H., Sun, F., Wu, H., Lv, D., Lv, Q., Feng, F., Yang, S., & Li, X. (2025). DS-YOLO: A Lightweight Strawberry Fruit Detection Algorithm. *Agronomy*, *15*(9), 2226.

[91] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF international conference on computer vision,

[92] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024a). Yolov10: Real-time end-to-end object detection. *Advances in neural information processing systems*, *37*, 107984-108011.

[93] Wang, A., Qian, W., Li, A., Xu, Y., Hu, J., Xie, Y., & Zhang, L. (2024b). NVW-YOLOv8s: An improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages. *Computers and Electronics in Agriculture*, *219*, 108833.

[94] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y.M. (2021). Scaled-yolov4: Scaling cross stage partial network. Proceedings of the IEEE/cvf conference on computer vision and pattern recognition,

[95] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y.M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[96] Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops,

[97] Wang, C.-Y., Yeh, I.-H., & Mark Liao, H.-Y. (2024). Yolov9: Learning what you want to learn using programmable gradient information. European conference on computer vision,

[98] Wang, P., Niu, T., & He, D. (2021). Tomato young fruits detection method under near color background based on improved faster R-CNN with attention mechanism. *Agriculture*, *11*(11), 1059.

[99] Wang, Y., Deng, H., Wang, Y., Song, L., Ma, B., & Song, H. (2024). CenterNet-LW-SE net: integrating lightweight CenterNet and channel attention mechanism for the detection of Camellia oleifera fruits. *Multimedia Tools and Applications*, *83*(26), 68585-68603.

[100] Wang, Y., Zhang, X., Yang, T., & Sun, J. (2022). Anchor detr: Query design for transformer-based detector. Proceedings of the AAAI conference on artificial intelligence,

[101] Wu, Z., An, R., Zhou, L., & Che, X. (2024). Recognition of Tomato Fruit Maturity and Fruit Stems Based on Improved YOLOv8. Proceedings of the International Conference on Image Processing, Machine

Learning and Pattern Recognition,

[102] Wu, Z., Zou, X., Zhou, W., & Huang, J. (2022). YOLOX-PAI: an improved YOLOX, stronger and faster than YOLOv6. *arXiv preprint arXiv:2208.13040*.

[103] Xu, P., Fang, N., Liu, N., Lin, F., Yang, S., & Ning, J. (2022). Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Computers and Electronics in Agriculture*, *197*, 106991.

[104] Yan, Z., Zhang, H., & Li, L. (2024). An Integrated Target Recognition Method Based on Improved Faster-RCNN for Apple Detection, Counting, Localization, and Quality Estimation. 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA),

[105] Yao, Z., Ai, J., Li, B., & Zhang, C. (2021). Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*.

[106] Zhang, F., Chen, Z., Ali, S., Yang, N., Fu, S., & Zhang, Y. (2023). Multi-class detection of cherry tomatoes using improved Yolov4-tiny model. *International Journal of Agricultural and Biological Engineering*, *16*(2), 225-231.

[107] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., & Shum, H.-Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

[108] Zhang, J., Xie, J., Zhang, F., Gao, J., Yang, C., Song, C., Rao, W., & Zhang, Y. (2024). Greenhouse tomato detection and pose classification algorithm based on improved YOLOv5. *Computers and Electronics in Agriculture*, *216*, 108519.

[109] Zhang, P., Liu, X., Yuan, J., & Liu, C. (2022). YOLO5-spear: A robust and real-time spear tips locator by improving image augmentation and lightweight network for selective harvesting robot of white asparagus. *Biosystems engineering*(218-), 218.

[110] Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., Luo, P., & Chen, K. (2023). Dense distinct query for end-to-end object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[111] Zhang, X. (2025). PG-DETR: A Lightweight and Efficient Detection Transformer for Early Stage Pomegranate Fruit Detection. *IEEE Access*.

[112] Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., & Wang, J. (2024). Ms-detr: Efficient detr training with mixed supervision. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[113] Zhao, K., & Yan, W.Q. (2021). Fruit detection from digital images using CenterNet. International symposium on geometry and vision,

[114] Zhao, M., Cui, B., Yu, Y., Zhang, X., Xu, J., Shi, F., & Zhao, L. (2025). Intelligent Detection of Tomato Ripening in Natural Environments Using YOLO-DGS. *Sensors*, *25*(9), 2664.

[115] Zhao, R., Guan, Y., Lu, Y., Ji, Z., Yin, X., & Jia, W. (2023). FCOS-LSC: A novel model for green fruit detection in a complex orchard environment. *Plant Phenomics*, *5*, 0069.

[116] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[117] Zhao, Z., Hao, K., & Liu, K, Zheng, T, Xu, J, Cui, S, He, C, Zhou, J, Zhao, G. (2023). MCANet: Hierarchical cross-fusion lightweight transformer based on multi-ConvHead attention for object detection. *Image and vision computing*, *136*(Aug.), 104715.104711-104715.104714.

[118] Zheng, H., Liu, C., Zhong, L., Wang, J., Huang, J., Lin, F., Ma, X., & Tan, S. (2025). An android-smartphone application for rice panicle detection and rice growth stage recognition using a lightweight YOLO network. *Front Plant Sci*, *16*, 1561632. https://doi.org/10.3389/fpls.2025.1561632

[119] Zhou, X., Wang, D., & Krähenbühl, P. (2019a). Objects as points. *arXiv preprint arXiv:1904.07850*.

[120] Zhou, X., Zhuo, J., & Krahenbuhl, P. (2019b). Bottom-up object detection by grouping extreme and center points. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

[121] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

[122] Zu, L., Zhao, Y., Liu, J., Su, F., Zhang, Y., & Liu, P. (2021). Detection and segmentation of mature green tomatoes based on mask R-CNN with automatic image acquisition approach. *Sensors*, *21*(23), 7842.