

# RF-DETR Object Detection vs YOLOv12 : A Study of Transformer-based and CNN-based Architectures for Single-Class and Multi-Class Greenfruit Detection in Complex Orchard Environments Under Label Ambiguity

Ranjan Sapkota<sup>a,\*</sup>, Rahul Harsha Cheppally<sup>b</sup>, Ajay Sharda<sup>b</sup>, Manoj Karkee<sup>a,\*</sup>

<sup>a</sup>*Biological & Environmental Engineering, Cornell University, Ithaca, 14850, NY, USA*

<sup>b</sup>*Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, 66502, KS, USA*

## Abstract

This study presents a comprehensive comparison between RF-DETR object detection and YOLOv12 object detection models for greenfruit recognition in complex orchard environments characterized by label ambiguity, occlusion, and background camouflage. A custom dataset was developed featuring both single-class (greenfruit) and multi-class (occluded and non-occluded greenfruits) annotations to assess model performance under real-world conditions. The RF-DETR object detection model, leveraging a DINOv2 backbone with deformable attention mechanisms, excelled in global context modeling, which proved particularly effective for identifying partially occluded or visually ambiguous greenfruits. Conversely, the YOLOv12 model employed CNN-based attention mechanisms to enhance local feature extraction, optimizing it for computational efficiency and edge deployment suitability. In the single-class detection scenarios, RF-DETR achieved the highest mean Average Precision (mAP@50) of 0.9464, showcasing its robust capability to accurately localize greenfruits within cluttered scenes. Despite YOLOv12N achieving the highest mAP@50:95 of 0.7620, RF-DETR object detection model consistently outperformed in managing complex spatial scenarios. In multi-class detection, RF-DETR again led with an mAP@50 of 0.8298, demonstrating its effectiveness in distinguishing between occluded and non-occluded fruits, whereas YOLOv12L topped the mAP@50:95 metric with 0.6622, indicating superior classification under detailed occlusion conditions. The analysis of model training dynamics revealed RF-DETR's rapid convergence, particularly in single-class scenarios where it plateaued at fewer than 10 epochs, underscoring the efficiency and adaptability of transformer-based architectures to dynamic visual data. These results confirm RF-DETR's suitability for accuracy-critical agricultural tasks, while YOLOv12 remains ideal for speed-sensitive deployments.

**Keywords:** Object Detection, RF-DETR object detection model (Roboflow Detection Transformer), YOLOv12 object detection model, You Only Look Once, Transformers, CNNs, Greenfruit Detection, Deep Learning , Machine-Vision

## 1. Introduction

As illustrated in Figure 1, the field of object detection over the past decade, propelled by breakthroughs in deep learning, have shifted from basic pattern recognition to sophisticated systems capable of complex image understanding. The object detection approaches can be branched into six primary methodologies as illustrated in Figure 1, each with unique strengths and applications in various domains of technology and automation. This evolution is vital for overcoming common visual recognition challenges in fields requiring high precision and adaptability, such as autonomous driving [1, 2], healthcare [3], security surveillance [4], and notably in agriculture [5, 6], where accurate and efficient object detection supports advancements like automated field monitoring [7] and robotic harvesting [8].

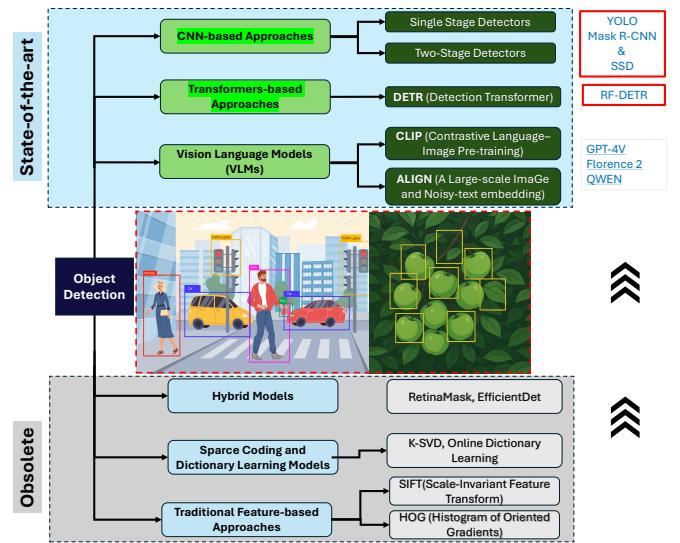


Figure 1: Classification of object detection methodologies: Top features state-of-the-art CNN-based and Transformer-based methods, widely adopted; Vision Language Models are emerging. Also includes Hybrid, Sparse Coding, and Traditional Feature-based approaches.

\*Corresponding Authors: Manoj Karkee and Ranjan Sapkota

Email address: mk2684@cornell.edu, rs2672@cornell.edu

(Manoj Karkee )

# RF-DETR目标检测与YOLOv12对比研究：基于Transformer与CNN架构在标签模糊条件下复杂果园环境中单类及多类青果检测的性能分析

拉詹·萨普科塔<sup>a,\*</sup>, 拉胡尔·哈沙·切帕利<sup>b</sup>, 阿贾伊·沙尔达<sup>b</sup>, 马诺吉·卡尔基<sup>a,\*</sup>

<sup>a</sup>Biological & Environmental Engineering, Cornell University, Ithaca, 14850, NY, USA

<sup>b</sup>Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, 66502, KS, USA

## 摘要

本研究针对标签模糊、遮挡及背景伪装等复杂果园环境下的青果识别，全面对比了RF-DETR与YOLOv12目标检测模型的性能。通过构建包含单类别（青果）与多类别（遮挡/非遮挡青果）标注的自定义数据集，评估模型在真实场景中的表现。采用DINOv2骨架与可变形注意力机制的RF-DETR模型在全局上下文建模方面表现卓越，尤其擅长识别部分遮挡或视觉模糊的青果；而基于CNN注意力机制的YOLOv12模型则通过增强局部特征提取能力，在计算效率与边缘部署适用性上更具优势。单类别检测RF-DETR以0.9464的mAP@50最高值展现了其在杂乱场景中精确定位青果的强劲能力。尽管YOLOv12N在mAP@50:95指标上以0.7620领先，但RF-DETR在处理复杂空间场景时始终更胜一筹。多类别检测中，RF-DETR再次以0.8298的mAP@50领跑，验证其区分遮挡/非遮挡果实的有效性；而YOLOv12L则以0.6622的mAP@50:95值，表明其在细节遮挡条件下的分类优势。模型训练动态分析显示RF-DETR在单类别场景中仅需不足10个epoch即可收敛，凸显了基于Transformer架构对动态视觉数据的高效适应能力。这些结果证实RF-DETR适用于精度优先的农业任务，而YOLOv12仍是速度敏感型部署的理想选择。

关键词：目标检测，RF-DETR目标检测模型（Roboflow Detection Transformer），YOLOv12目标检测模型，You Only Look Once，Transformers，卷积神经网络，绿果检测，深度学习，机器视觉

## 1. 引言

如图1所示，过去十年间，在深度学习突破的推动下，目标检测领域已从基础模式识别发展为能够实现复杂图像理解的精密系统。目标检测方法可划分为六大主要技术路线（如图1所示），每种方法在技术与自动化的不同应用领域均具有独特优势。这一演进对于克服自动驾驶[1, 2]、医疗健康[3]、安防监控[4]等需要高精度与适应性的领域中的常见视觉识别挑战至关重要，尤其在农业领域[5, 6]——精准高效的目标检测技术支撑着农田自动化监测[7]与机器人采收[8]等进步。

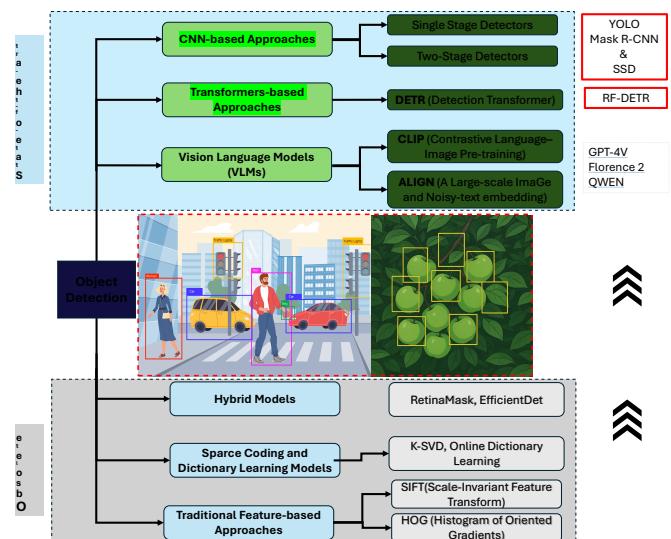


图1：目标检测方法分类：顶部展示了当前最先进的基于CNN和Transformer的方法，这些方法被广泛采用；视觉语言模型正在兴起。同时还包括混合方法、稀疏编码以及基于传统特征的方法。

\*Corresponding Authors: Manoj Karkee and Ranjan Sapkota  
Email address: mk2684@cornell.edu, rs2672@cornell.edu  
(Manoj Karkee)

The six approaches depicted in Figure 1 are Convolutional Neural Networks (CNNs) [9], Transformer-based models [10, 11], Vision Language Model-based approaches [12, 13], Hybrid models such as RetinaMask and EfficientDet [14, 15], Sparse Coding and Dictionary Learning models, and Traditional Feature-based approaches. Among them, CNNs, including the YOLO (You Only Look Once) series [16, 17] and R-CNN (Region-based CNN) family such as Mask R-CNN [18], have become staples in practical deployments due to their proficient handling of spatial hierarchies. Transformer-based models like DETR (Detection Transformer) such as dynamic DETR [19] and deformable DETR [20] utilize self-attention mechanisms to treat images as sequences of patches, which helps in integrating a global context and eliminates the need for Non-Maximum Suppression (NMS) [21], streamlining post-processing [22]. Vision Language Models (VLMs) such as CLIP (Contrastive Language-Image Pre-training) represent an emerging area that integrates textual and visual data, aiming to enhance robustness through multimodal learning, though their application in real-world scenarios, particularly in robotics and automation, is still developing. On the other hand, Hybrid models such as RetinaMask, Sparse Coding models like Online Dictionary Learning, and Traditional Feature-based methods such as Histogram of Oriented Gradients (HOG) are increasingly considered obsolete [23, 24]. These have been superseded by more advanced systems that provide not only greater accuracy but also the capability to perform in real-time, a critical requirement for latency-sensitive operations such as those found in modern agricultural settings. As object detection continues to evolve, the focus remains on technologies that combine high precision with efficient processing, positioning CNN and Transformer-based models as the current state-of-the-art in the field.

Among the six primary object detection approaches illustrated in Figure 1, CNN-based and Transformer-based models have emerged as the most widely adopted and actively developed over the past five years. These two paradigms now dominate both research and real-world applications due to their scalability, accuracy, and adaptability. This ongoing dominance has sparked a competitive evolution between the two approaches, particularly with the release of powerful Transformer-based models such as RF-DETR, developed by Roboflow. RF-DETR integrates architectural innovations from Deformable DETR and LW-DETR, and utilizes a DINOv2 backbone, offering superior global context modeling and domain adaptability. The model eliminates reliance on anchor boxes and Non-Maximum Suppression (NMS), supporting end-to-end training and real-time inference. With two variants, Base (29M) and Large (128M), RF-DETR offers scalability from edge deployment to high-performance scenarios. The model has been shown to outperform YOLOv11 and is the only model to surpass 60% mAP on the COCO dataset to date. Its performance on COCO and RF100-VL benchmarks is visualized in Figure 2a and 2b. However, despite its promise, RF-DETR has not yet been officially benchmarked against YOLOv12, the latest and most advanced model in the YOLO family. A comparative evaluation is warranted, particularly since YOLOv12 builds upon the strengths

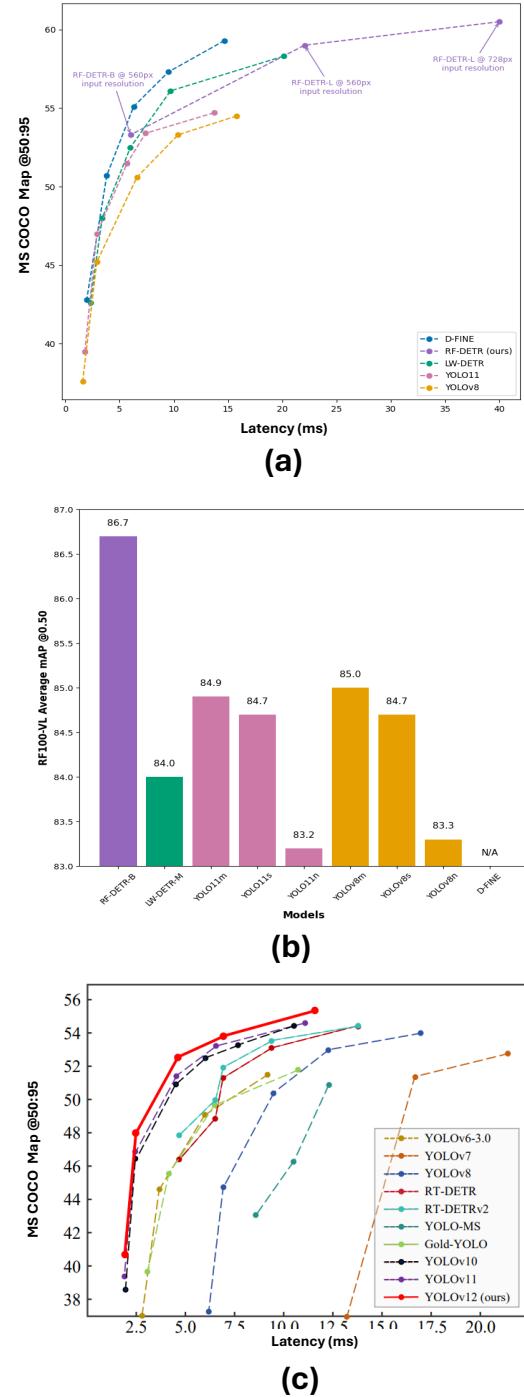


Figure 2: CNN vs Transformer-based model performance comparison focusing on YOLOv12 (CNN-based) and RF-DETR (Transformer-based) architectures: (a) RF-DETR object detection model benchmark evaluation with YOLOv11, YOLOv8 and other DETR-based object detection models ; (b)RF-DETR evaluation on the RF100-VL dataset, highlighting domain adaptability and edge deployment potential. ; and (c) Performance overview of recent CNN-based models, including YOLOv6 through YOLOv12, Gold-YOLO, RT-DETR, RT-DETRv2, and YOLO-MS. b) RF-DETR benchmark results on the MS COCO dataset, surpassing 60% mAP

of YOLOv11, YOLOv10, and Gold-YOLO RT-DETR, as illustrated in Figure 2c.

图1所示的六种方法包括卷积神经网络（CNNs）[9]、基于Transformer的模型[10,11]、视觉语言模型方法[12,13]、混合模型（如RetinaMask和EfficientDet）[14,15]、稀疏编码与字典学习模型以及传统基于特征的方法。其中，CNN系列（如YOLO（You Only Look Once）[16,17]和R-CNN家族（如Mask R-CNN[18]）凭借其出色的空间层次处理能力，已成为实际部署的主流选择。基于Transformer的模型（如动态DETR[19]和可变形DETR[20]）利用自注意力机制将图像视为补丁序列，有助于整合全局上下文并省去非极大值抑制（NMS）[21]，从而简化后处理流程[22]。视觉语言模型（如对比语言-图像预训练模型CLIP）代表了融合文本与视觉数据的新兴领域，旨在通过多模态学习提升鲁棒性，尽管其在机器人学和自动化等现实场景的应用仍处于发展阶段。另一方面，RetinaMask等混合模型、在线字典学习等稀疏编码模型，以及方向梯度直方图（HOG）等传统特征方法正逐渐被视为过时技术[23,24]。这些方法已被更先进的系统所取代，后者不仅能提供更高精度，还具备实时处理能力——这对现代农业等延迟敏感场景至关重要。随着目标检测技术的持续演进，高精度与高效处理相结合的技术仍是关注焦点，使得CNN和Transformer模型成为当前领域的最先进解决方案。

在图1展示的六种主要目标检测方法中，基于CNN和基于Transformer的模型已成为过去五年间最广泛采用且积极发展的技术。这两种范式凭借其可扩展性、准确性和适应性，如今主导着研究领域与实际应用。这种持续的统治地位激发了两大方法间的竞争性演进，尤其是随着Roboflow开发的RF-DETR等强大Transformer模型的发布。RF-DETR融合了可变形DETR与LW-DETR的架构创新，并采用DINOv2骨干网络，提供卓越的全局上下文建模与领域适应能力。该模型摒弃了锚框和非极大值抑制（NMS）的依赖，支持端到端训练与实时推理。通过Base（29M）和Large（128M）两个变体，RF-DETR实现了从边缘部署到高性能场景的可扩展性。实验表明，其性能超越YOLOv11，并成为迄今唯一在COCO数据集上mAP突破60%的模型。图2a与2b直观呈现了该模型在COCO和RF100-VL基准测试中的表现。然而尽管前景广阔，RF-DETR尚未与YOLO家族最新旗舰模型YOLOv12进行官方基准对比。鉴于YOLOv12在原有优势基础上的全面升级，开展比较性评估尤为必要。

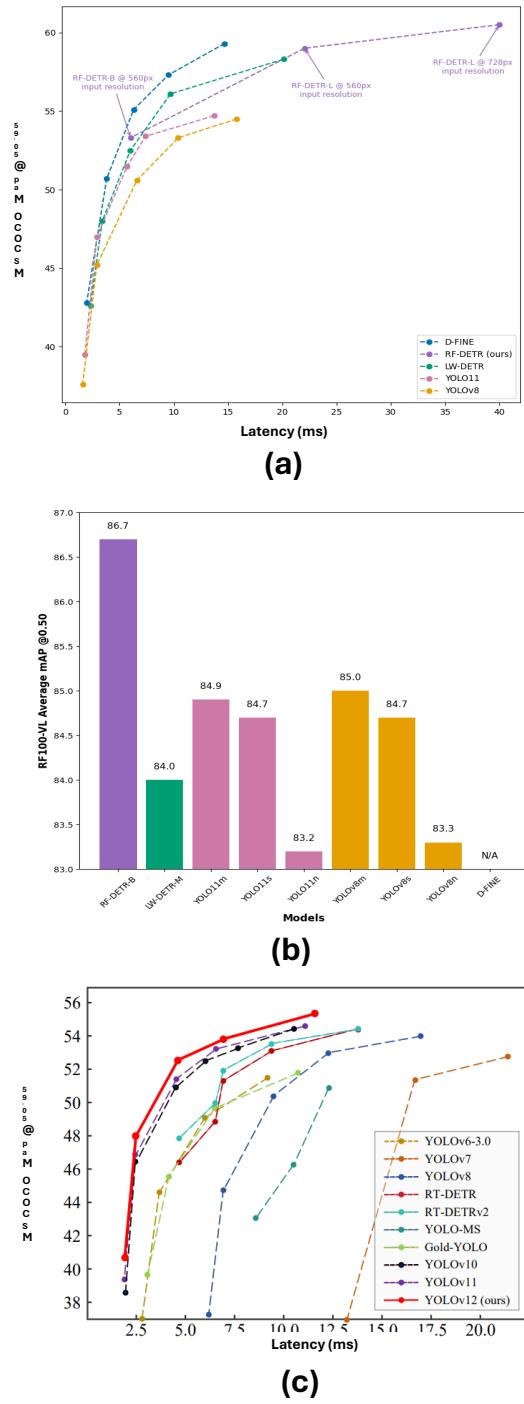


图2：基于CNN与Transformer的模型性能对比，聚焦于YOLOv12（基于CNN）和RF-DETR（基于Transformer）架构：(a) RF-DETR目标检测模型与YOLOv11、YOLOv8及其他基于DETR的目标检测模型的基准评估；(b) RF-DETR在RF100-VL数据集上的评估，突出其领域适应性和边缘部署潜力；(c) 近期基于CNN模型的性能概览，包括YOLOv6至YOLOv12、Gold-YOLO、RT-DETR、RT-DETRv2及YOLO-MS。b) RF-DETR在MS COCO数据集上的基准测试结果，mAP突破60%

YOLOv11、YOLOv10和Gold-YOLO RT-DETR的性能对比，如图2c所示。

### *1.1. CNN-based Object Detection Approaches*

CNNs have been pivotal in advancing object detection, significantly since AlexNet catalyzed the field in 2012 [25]. These networks utilize hierarchical layers of convolutions, pooling, and nonlinear activations to effectively learn feature representations from images [26]. Unlike transformers that handle global relationships via attention mechanisms, CNNs are structured to excel at extracting local features, which is facilitated by their inherent inductive biases such as translation equivariance and the establishment of spatial hierarchies [27]. This fundamental architectural distinction makes CNNs particularly well-suited for scenarios demanding real-time processing and edge computing deployments, albeit with a noted limitation in their ability to model global contextual information comprehensively [28, 29].

The progress in CNN architectures for object detection has been marked by several significant innovations as:

- R-CNN Family: This series began with R-CNN in 2014 [30], which utilized selective search to generate region proposals that were then processed by CNNs to extract features, achieving a 53.3% mAP on the PASCAL VOC dataset but with a high computational cost. Subsequent iterations, Fast R-CNN and Faster R-CNN, introduced ROI pooling and Region Proposal Networks (RPNs), respectively, enhancing the efficiency and speed of these models dramatically.
- Mask R-CNN: An extension of Faster R-CNN that includes a branch for predicting segmentation masks on each Region of Interest (ROI), effectively handling instance segmentation with a high precision level [18, 31].
- YOLO Series: Starting with YOLOv1 [16], which reframed object detection as a single regression problem from image pixels to bounding box coordinates and class probabilities, through to YOLOv12 [32], which introduced improvements like anchor-free detection and dynamic label assignment for enhanced accuracy and efficiency [33, 34, 35, 31].
- SSD: This model combined multi-scale feature maps with default bounding boxes to perform detections, facilitating direct classification and localization from feature maps without needing separate region proposals [36].
- RetinaNet: Known for tackling class imbalance with the focal loss function, which helps focus the model on hard-to-classify examples by down-weighting the loss assigned to well-classified examples [37].
- EfficientDet: This model utilized a scaling method that systematically adjusts the depth, width, and resolution of the network, integrated with a BiFPN for feature fusion across different scales, achieving high efficiency and accuracy [15].

### *1.2. Transformers-based Object Detection Approaches*

DETR have revolutionized object detection by integrating transformer architectures, traditionally used in natural language

processing, into visual recognition tasks [22]. Introduced by Facebook AI in 2020, DETR presents a novel approach by treating object detection as a direct set prediction problem, eliminating the need for traditional components like anchor boxes and complex post-processing steps such as Non-Maximum Suppression (NMS) [22]. At its core, DETR uses a standard CNN backbone, typically ResNet-50, for initial feature extraction. This is followed by a transformer that consists of an encoder and a decoder where the encoder processes the spatial features across the image and the decoder uses learned object queries to predict the presence of objects along with their categories and bounding boxes, all in parallel.

The key architectural variants of DETR have addressed its initial shortcomings such as slow convergence and high computational demands:

- Deformable DETR: Introduced to tackle the inefficiencies of the standard transformer attention mechanism, it employs deformable attention which focuses on a small set of key sampling points around each reference point, significantly reducing the computational load and improving detection of small objects [20]. This variant leverages iterative bounding box refinement and multi-scale features to enhance accuracy and speed up training .
- RT-DETR: Developed for real-time applications, this variant from Baidu features a hybrid encoder that merges CNN and transformer features to optimize both intra-scale interaction and cross-scale fusion, achieving impressive speeds on standard hardware. It introduces IoU-aware query selection, dynamically adjusting the decoding process based on predicted objectness scores [38, 39].
- Co-DETR: Enhances training stability and performance by implementing a dual supervision strategy that combines traditional one-to-many (like Faster R-CNN) and one-to-one (like DETR) label matching [40]. This approach, supported by hierarchical attention mechanisms, significantly improves feature representation, especially in challenging conditions such as occlusions [41].
- YOLOS: Stands out by adapting Vision Transformers (ViTs) directly for object detection without any CNNs [42]. It uses a sequence of image patches (tokens) along with a set of learnable detection tokens, demonstrating that transformers can effectively encode spatial relationships inherent in detection tasks [43].
- OWL-ViT: Expands the applicability of transformers to open-vocabulary detection by integrating vision and language, using a transformer decoder to align image features with text queries [44]. This model facilitates zero-shot detection, where the system can recognize objects it has never seen during training, described only by text [45, 46].
- DINO (DETR with Improved Denoising Anchor boxes): Focuses on enhancing small object detection through a novel training strategy that involves adding noise to

### 1.1. CNN-based Object Detection Approaches

自2012年AlexNet推动该领域发展以来[25]，卷积神经网络(CNN)在目标检测领域的进步中发挥了关键作用。这些网络通过卷积层、池化层和非线性激活函数的层级结构，高效学习图像特征表示[26]。与依赖注意力机制处理全局关系的Transformer不同，CNN凭借其固有的归纳偏置（如平移等变性和空间层级结构的建立）[27]，在提取局部特征方面表现卓越。这种根本性的架构差异使CNN特别适合需要实时处理和边缘计算部署的场景，尽管其在全面建模全局上下文信息方面存在明显局限[28, 29]。

CNN在目标检测架构上的进展，以以下几项重大创新为标志：

- R-CNN系列：该系列始于2014年的R-CNN[30]，它利用选择性搜索生成候选区域，随后通过CNN提取特征，在PASCAL VOC数据集上实现了53.3%的mAP，但计算成本较高。后续迭代版本Fast R-CNN和Faster R-CNN分别引入了ROI池化与区域提议网络(RPN)，显著提升了模型的效率和速度。
- Mask R-CNN：Faster R-CNN的扩展版本，它在每个感兴趣区域（ROI）上增加了一个预测分割掩码的分支，从而以高精度水平有效处理实例分割任务[18, 31]。
- YOLO系列：从YOLOv1[16]开始，它将目标检测重新定义为从图像像素到边界框坐标及类别概率的单一回归问题，直至YOLOv12[32]，后者引入了无锚点检测和动态标签分配等改进，以提升精度与效率[33, 34, 35, 31]。
- SSD：该模型将多尺度特征图与默认边界框相结合进行检测，无需单独的区域提议即可直接从特征图完成分类与定位[36]。
- RetinaNet：以采用焦点损失函数解决类别不平衡问题而闻名，该函数通过降低分配给易分类样本的损失权重，帮助模型集中关注难以分类的样本[37]。
- EfficientDet：该模型采用了一种系统性调整网络深度、宽度和分辨率的缩放方法，并结合BiFPN实现跨尺度特征融合，从而在保证高效率的同时达到了高精度[15]。

### 1.2. Transformers-based Object Detection Approaches

DETR通过整合传统上用于自然语言的Transformer架构，彻底革新了目标检测领域。

处理，转化为视觉识别任务[22]。由Facebook AI于2020年提出的DETR，通过将目标检测视为一个直接的集合预测问题，提出了一种新颖的方法，消除了对传统组件（如锚框）和复杂后处理步骤（如非极大值抑制NMS）的需求[22]。DETR的核心在于使用标准的CNN主干网络（通常是ResNet-50）进行初始特征提取。随后是一个由编码器和解码器组成的Transformer结构：编码器处理图像的空间特征，解码器则利用学习到的对象查询 $\{v^*\}$ ，并行预测物体的存在及其类别与边界框。

DETR的关键架构变体已解决了其初始缺陷，如收敛速度慢和计算需求高的问题：

- 可变形DETR：为解决标准Transformer注意力机制效率低下的问题而提出，它采用可变形注意力机制，该机制仅聚焦于每个参考点周围的一小组关键采样点，显著降低了计算负担并提升了小物体检测性能[20]。该变体通过迭代式边界框优化和多尺度特征融合来提升精度并加速训练过程。
- RT-DETR：专为实时应用设计，此百度推出的变体采用混合编码器架构，巧妙融合CNN与Transformer特征，优化了尺度内交互与跨尺度融合，在标准硬件上实现了卓越的运行速度。其创新性引入IoU感知查询选择机制，能根据预测的目标性分数动态调整解码流程[38, 39]。
- Co-DETR：通过实施一种结合传统一对多（如Faster R-CNN）和一对一（如DETR）标签匹配的双重监督策略[40]，显著提升了训练稳定性与性能。该方法在层次化注意力机制的支持下，大幅优化了特征表示能力，尤其在遮挡等挑战性场景中表现突出[41]。
- YOLOS：其独特之处在于直接采用视觉变换器（ViTs）进行目标检测，无需任何卷积神经网络（CNNs）[42]。该方法通过将图像分块（标记）序列与一组可学习的检测标记相结合，证明了变换器能够有效编码检测任务中固有的空间关系[43]。
- OWL-ViT：通过整合视觉与语言，利用Transformer解码器将图像特征与文本查询对齐，扩展了Transformer在开放词汇检测中的适用性[44]。该模型支持零样本检测，即系统能够识别训练期间从未见过的物体，仅通过文本描述即可实现[45, 46]。
- DINO（带改进去噪锚框的DETR）：通过一种新颖的训练策略专注于提升小物体检测能力，该策略涉及向 $\{v^*\}$ 添加噪声

ground truth boxes and learning to predict corrective offsets, improving precision and robustness [47].

- RF-DETR: Released by Roboflow, RF-DETR is a real-time transformer-based object detection model that achieves 60.5 mAP at 25 FPS on an NVIDIA T4 GPU, outperforming models like YOLOv11 and LW-DETR on benchmarks such as COCO and RF100-VL [48]. Its architecture is designed for high-speed edge deployment and domain adaptability, with two variants: RF-DETR-Base (29M parameters) and RF-DETR-Large (128M parameters).

### 1.3. Objectives

Despite significant advances in object detection, the performance of state-of-the-art models in complex, label-ambiguous agricultural environments remains underexplored. The recently released RF-DETR, a Transformer-based real-time object detection model developed by Roboflow, has demonstrated remarkable performance by surpassing 60% mAP on the MS COCO dataset, the highest recorded for any Transformer-based detector to date. However, RF-DETR's performance has only been benchmarked against earlier versions of YOLO, including YOLOv11, and a few other models like LW-DETR, leaving a notable gap in comparative evaluation with YOLOv12, the most recent and advanced CNN-based detector from the YOLO family. This lack of direct comparison has created uncertainty regarding which model, RF-DETR or YOLOv12, offers better detection capabilities in real-world conditions, particularly under occlusion, camouflage, and ambiguous labeling.

This study addresses this gap by conducting a detailed evaluation of RF-DETR and YOLOv12 for the task of greenfruit detection in commercial apple orchards. Immature green apple fruitlets are critical for early yield estimation and thinning but are notoriously difficult to detect due to their small size, color similarity with the background, and frequent occlusion by foliage or other fruits. This visual complexity results in label ambiguity, making it difficult to determine whether fruitlets are fully visible, partially visible, or completely hidden conditions that challenge both manual annotation and automated detection.

To assess the robustness of these two architectures, we developed a custom dataset and evaluated both models using identical training protocols and hyperparameters. Performance was assessed in both single-class and multi-class detection tasks using key metrics: Precision, Recall, F1-Score, mAP@50, and mAP@50:95. We also measured inference speed and processing efficiency, aiming to provide a clear, evidence-driven comparison of CNN-based versus Transformer-based object detection in precision agriculture.

## 2. Methods

This experiment is performed in four steps, as depicted in Figure 3a. Initially, real field images were collected from a commercial orchard under complex conditions, characterized by immature greenfruits camouflaged against a green canopy,

presenting significant challenges for machine vision due to occlusions. Subsequently, these images were captured using a robotic platform and a machine vision camera, followed by pre-processing and preparation. In the third step, two deep learning models, RF-DETR and YOLOv12, were implemented using the same dataset, hyperparameters, and number of epochs. Finally, the performance of these models was evaluated in terms of their ability to detect single-class and multi-class greenfruit objects in this challenging orchard environment.

### 2.1. Study Site and Data Acquisition

Data acquisition for this study was conducted in a commercial orchard situated in Prosser, Washington State, USA, as illustrated in Figure 3b. The orchard was densely planted with 'Scifresh' apple trees, commonly known as Jazz apples. The selection of this specific orchard was due to its complex environmental conditions characterized by the green color of immature fruitlets blending with the green canopy background, as depicted in Figure 3c. This similarity in color created a challenging scenario for accurate image detection due to significant occlusions and visual confusion, typical in complex orchard scenes.

Image collection was executed using a sophisticated robotic platform that incorporated an Intel RGB-D camera, which was mounted on a UR5e robotic arm, as depicted in Figure 3d. This setup enabled the precise capture of RGB images, specifically focusing on the immature 'Scifresh' apple fruitlets. The imagery was collected in May 2024, just prior to the commencement of fruitlet thinning activities. The timing for the collection was carefully chosen based on continuous monitoring of the orchard's developmental stages (before thinning, exactly during the fruitlet thinning week) and in consultation with local growers and orchard workers to ensure optimal data relevance for the study.

The orchard, established in 2008, was methodically laid out with tree rows spaced 3 meters apart and an intra-row spacing of 1 meter. Throughout the course of this study, a total of 857 images were captured utilizing an Intel RealSense D435i camera, as shown in Figure 3d. The selected camera is equipped with a depth-sensing system that operates on active infrared (IR) stereo vision, complemented by an inertial measurement unit (IMU). This camera's depth sensor employs structured light technology, which utilizes a pattern projector to induce disparities between stereo images captured by two IR cameras.

The camera's 3D sensor boasts a resolution of  $1280 \times 720$  pixels, capable of capturing depth information up to a distance of 10 meters. It supports a frame rate of up to 90 frames per second (fps), and features a horizontal field of view (HFOV) of  $69.4^\circ$  and a vertical field of view (VFOV) of  $42.5^\circ$ . Additionally, the integrated 6-axis IMU provides critical orientation data, significantly enhancing the alignment of depth data with the actual scene, thus improving the overall understanding and analysis of the captured images. This detailed and methodical approach to data collection was fundamental in addressing the visual complexities presented by the orchard environment.

真实标注框并学习预测校正偏移量，从而提高精度和鲁棒性[47]。

- RF-DETR：由Roboflow发布的RF-DETR是一款基于Transformer的实时目标检测模型，在NVIDIA T4 GPU上以25 FPS的速度实现了60.5 mAP，在COCO和RF100-VL等基准测试中超越了YOLOv11和LW-DETR等模型[48]。其架构专为高速边缘部署和领域适应性设计，提供两种变体：RF-DETR-Base（2900万参数）和RF-DETR-Large（1.28亿参数）。

### 1.3. Objectives

尽管目标检测领域取得了显著进展，但在复杂、标签模糊的农业环境中，最先进模型的性能仍未得到充分探索。Roboflow最新发布的RF-DETR——基于Transformer的实时目标检测模型，在MS COCO数据集上实现了超过60%的m AP，创下了迄今为止所有基于Transformer的检测器中的最高记录，展现出卓越性能。然而，RF-DETR的基准测试仅针对早期版本的YOLO（包括YOLOv11）以及LW-DETR等少数模型进行，与YOLO家族最新、最先进的基于CNN的检测器YOLOv12的对比评估存在明显空白。这种直接比较的缺失，使得在现实条件下（尤其是存在遮挡、伪装和模糊标签时）RF-DETR与YOLOv12哪个模型具备更优检测能力的问题仍存不确定性。

本研究通过详细评估RF-DETR和YOLOv12在商业苹果园绿果检测任务中的表现，填补了这一空白。未成熟的绿色小苹果对早期产量预估和疏果至关重要，但由于其体积小、与背景颜色相近、常被枝叶或其他果实遮挡，检测难度极大。这种视觉复杂性导致标签模糊，难以判断小果是完全可见、部分可见还是完全被遮挡的状态，这对人工标注和自动检测都构成了挑战。

为了评估这两种架构的鲁棒性，我们开发了一个定制数据集，并采用相同的训练协议和超参数对两个模型进行了评估。性能评估涵盖单类别和多类别检测任务，关键指标包括：精确率（Precision）、召回率（Recall）、F1分数（F1-Score）、mAP@50以及mAP@50:95。此外，我们还测量了推理速度和处理效率，旨在为精准农业中基于CNN与基于Transformer的目标检测方法提供清晰、数据驱动的对比分析。

## 2. 方法

本实验分四个步骤进行，如图3a所示。首先，在复杂条件下从商业果园采集了实地图像，其特征是未成熟的绿色果实与绿色树冠形成伪装，

由于遮挡问题，这对机器视觉提出了重大挑战。随后，这些图像通过机器人平台和机器视觉相机采集，并进行了预处理和准备工作。第三步中，使用相同的数据集、超参数和训练周期数，实现了两种深度学习模型——RF-DETR与YOLOv12。最后，评估了这些模型在具有挑战性的果园环境下检测单类别与多类别绿色果实目标的性能表现。

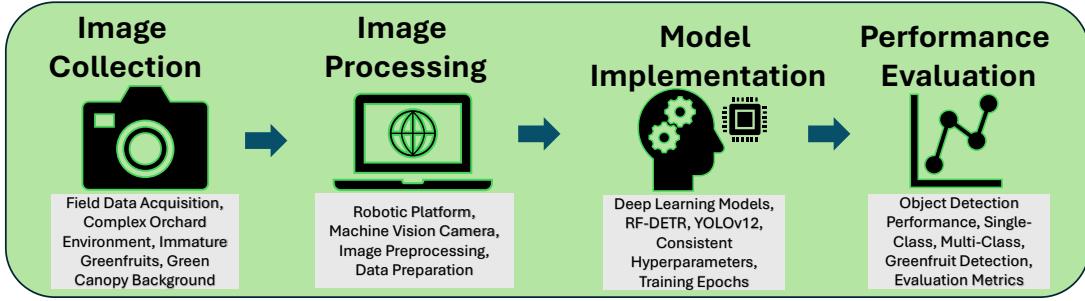
### 2.1. Study Site and Data Acquisition

本研究的数据采集工作在美国华盛顿州普罗斯瑟市的一处商业果园中进行，如图3b所示。该果园密集种植着俗称爵士苹果的‘Scifresh’苹果树。选择这一特定果园的原因在于其复杂的环境条件——未成熟小果的绿色与树冠背景的绿色相互交融（如图3c所示）。这种颜色相似性造成了显著的遮挡和视觉混淆，为精准图像检测带来了挑战，这也是复杂果园场景中的典型情况。

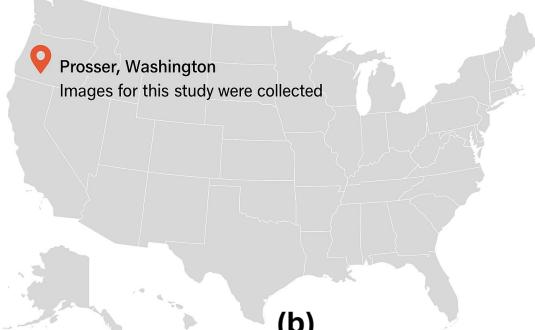
图像采集采用了一款集成Intel RGB-D相机的精密机器人平台完成，该相机安装在UR5e机械臂上，如图3d所示。这一配置实现了对未成熟‘Scifresh’苹果幼果RGB图像的精准捕捉。影像采集于2024年5月进行，恰好在幼果疏除作业开始之前。采集时机的选择基于对果园发育阶段的持续监测（疏果前，正值疏果周期间），并与当地种植者和果园工作人员协商确定，以确保研究数据的最佳相关性。

该果园始建于2008年，采用系统化布局，行距3米，株距1米。本研究期间，共使用英特尔实感D435i相机采集了857张图像（如图3d所示）。所选相机搭载基于主动红外立体视觉的深度感知系统，并配有惯性测量单元（IMU）。该相机的深度传感器采用结构光技术，通过图案投影器在两台红外相机拍摄的立体图像间制造视差。

该相机的3D传感器拥有 $1280 \times 720$ 像素的分辨率，能够捕捉最远10米距离的深度信息。它支持高达每秒90帧（fps）的帧率，并具备69.4°的水平视场角（HFOV）和42.5°的垂直视场角（VFOV）。此外，集成的6轴IMU提供了关键的姿态数据，显著提升了深度数据与实际场景的对齐度，从而增强了对捕获图像的整体理解和分析能力。这种细致且系统化的数据采集方法，对于应对果园环境带来的视觉复杂性起到了基础性作用。



(a)



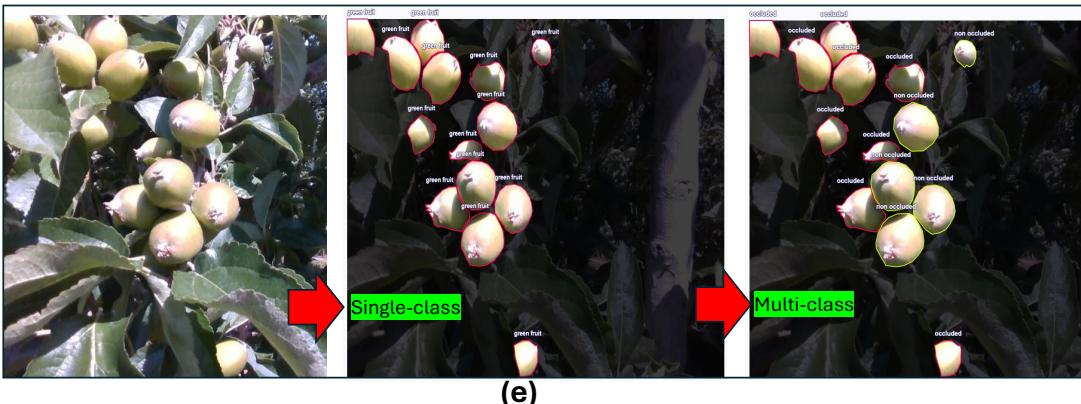
(b)



(c)



(d)



(e)

Figure 3: Overview of data collection setup and environment: a) Flow diagram showing the methodology of RF-DETR vs YOLOv12 comparison ; b) Map highlighting the study location in Prosser, Washington, USA ; c) of 'Scifresh' apple trees, known as Jazz apples; d) The robotic platform used for image acquisition, featuring an Intel RGB-D camera mounted on a UR5e robotic arm, capturing images of immature greenfruits in complex orchard environment.

## 2.2. Data Preprocessing and Preparation

Following data collection, the acquired RGB images underwent a systematic preprocessing and annotation pipeline to prepare them for deep learning model training and evaluation. Im-

age annotation was performed manually using the Roboflow platform (Roboflow, Des Moines, Iowa), a widely used tool for custom dataset generation in computer vision workflows. The dataset construction involved two labeling schemes: (i) a

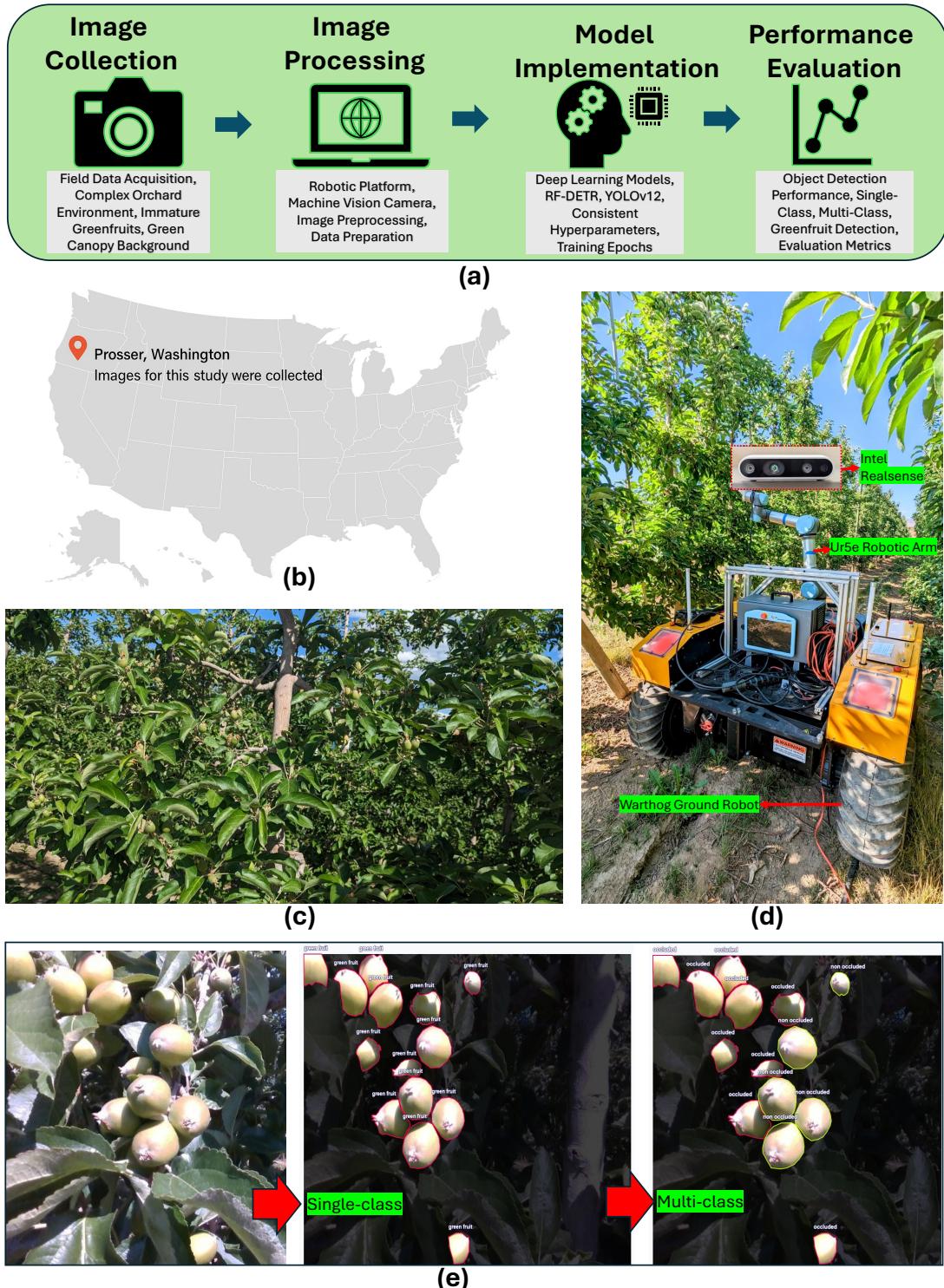


图3：数据收集设置与环境概览：a) 展示RF-DETR与YOLOv12对比方法的流程图；b) 标注美国华盛顿州普罗瑟研究地点的地图；c) 被称为爵士苹果的‘Scifresh’ 苹果树；d) 用于图像采集的机器人平台，配备安装在UR5e机械臂上的英特尔RGB-D相机，在复杂果园环境中捕捉未成熟绿色果实的图像。

## 2.2. Data Preprocessing and Preparation

在数据收集之后，所获取的RGB图像经过系统的预处理和标注流程，为深度学习模型的训练与评估做好准备。

年龄标注是使用Roboflow平台（Roboflow，爱荷华州得梅因）手动完成的，该平台是计算机视觉工作流中广泛使用的自定义数据集生成工具。数据集构建涉及两种标注方案：(i) 一种

single-class dataset and (ii) a multi-class dataset, both of which aimed to capture the inherent complexity of greenfruit detection under real-world orchard conditions.

In the first scheme, all visible immature apples were annotated under a single class labeled as “greenfruit,” regardless of their degree of visibility or occlusion. A total of 857 high-resolution orchard images were uploaded and processed in Roboflow for this purpose. As illustrated in the center image of Figure 3e, this dataset captured a wide range of greenfruit appearances, resulting in 4,125 individual object labels. The uniform labeling in this scheme was suitable for establishing baseline detection performance but did not capture the dynamics of visual challenges like partial occlusions or background blending. To explore these complexities more explicitly, a second labeling scheme was developed to create a multi-class dataset. In this case, each greenfruit was categorized into one of two classes: occluded greenfruit and non-occluded greenfruit. The classification criteria were based on the degree of visibility. Greenfruits with at least 90% of their surface area clearly visible, unobstructed by foliage, branches, or other fruits, were labeled as non-occluded. Conversely, any fruit partially hidden, whether by overlapping apples, intersecting leaves, or obstructing branches was labeled as occluded. This dynamic annotation approach is depicted in the rightmost image of Figure 3e.

However, the labeling process was complicated by label ambiguity, a critical issue in computer vision tasks, especially in natural environments. Label ambiguity refers to the uncertainty or subjectivity in assigning a label due to unclear visual boundaries, overlapping objects, or inconsistent visibility. In this study, several practical instances of label ambiguity emerged. First, in cases where multiple greenfruits clustered tightly, it was often unclear whether one was partially occluding the other or if they were side-by-side. Second, some fruits appeared occluded due to lighting and shadows rather than actual physical obstruction, leading to inconsistent labeling across images. Third, foliage sometimes mimicked the texture and color of immature fruits, making it difficult to distinguish between the object of interest and the background. Fourth, partial occlusions at the edges of images often left annotators uncertain whether to classify the object as occluded or simply truncated due to the field of view. These examples underscore why greenfruit detection in real orchard environments is particularly prone to labeling inconsistencies. Although classification guidelines were applied rigorously, the complex interplay between object geometry, environmental texture, and visibility made complete objectivity difficult to achieve. Thus, the term label ambiguity is used to describe the dataset’s inherent subjectivity and the potential variability it introduces during model training and evaluation.

### 2.3. Training Object Detection Models

#### 2.3.1. Training RF-DETR Object Detection Model

RF-DETR is a real-time, transformer-based object detection architecture optimized for both accuracy and efficiency<sup>1 2</sup>. As illustrated in Figure 4a, RF-DETR builds upon the

foundations of Deformable DETR and LW-DETR, integrating a pre-trained DINOv2 vision transformer as its backbone. This backbone enhances cross-domain generalization through self-supervised learning, making the model highly adaptable to domain-specific challenges like greenfruit detection in agricultural environments.

A key innovation of RF-DETR is its ability to eliminate traditional object detection components such as anchor boxes and NMS. Instead, it uses a transformer-based encoder-decoder architecture with deformable cross-attention to selectively attend to spatially relevant features, improving detection under occlusion, clutter, and camouflage. Unlike traditional DETR variants, RF-DETR employs a single-scale feature extraction strategy to reduce computational overhead, thus enabling faster inference without compromising accuracy.

Two variants of the model are available: RF-DETR-Base (29 million parameters) and RF-DETR-Large (128 million parameters). In this study, the **\*\*RF-DETR-Base\*\*** model was selected due to its balance between computational efficiency and high detection accuracy, making it suitable for real-time processing in field robotics. The RF-DETR-Base model achieves a mAP of 53.3 on the COCO benchmark and 86.7 mAP@50 on the RF100-VL dataset, positioning it as one of the few models to exceed 60% mAP@50:95 in real-time settings.

Training followed the official Roboflow implementation. The model was initialized with DINOv2-pretrained weights and trained using the AdamW optimizer with a learning rate of 1e-4 and batch size of 8 for 300 epochs. The training leveraged hybrid encoder optimizations inspired by RT-DETR and deformable attention mechanisms. Loss functions included cross-entropy for classification and a combination of L1 and GIoU losses for bounding box regression.

Additionally, contrastive denoising training was employed to improve detection robustness for partially visible and small objects. RF-DETR also adopted collaborative label assignments for stability in ambiguous annotation conditions, as well as multi-resolution input support (640–1280 px), allowing latency-accuracy trade-offs without retraining. This configuration made RF-DETR-Base a powerful and efficient model for detecting occluded and camouflaged immature greenfruits in complex orchard environments.

#### 2.3.2. Training YOLOv12 Object Detection Model

YOLOv12 represents a transformative leap in CNN-based object detection, merging the efficiency of traditional convolutional architectures with attention-inspired mechanisms to address modern computer vision demands [32]. Departing from previous YOLO iterations, the model introduces R-ELAN (Residual Efficient Layer Aggregation Network) as its core backbone as depicted in Figure 4b, combining residual connections with multi-scale feature fusion to resolve gradient bottlenecks while enhancing feature reuse across network depths. A novel 7×7 separable convolution layer replaces standard 3×3 kernels, preserving spatial context with 60% fewer parameters than conventional large-kernel convolutions while implicitly encoding positional relationships, effectively circumventing the need for explicit positional embeddings used in

<sup>1</sup><https://blog.roboflow.com/rf-detr/>

<sup>2</sup><https://github.com/roboflow/rf-detr>

单类别数据集和(ii)多类别数据集，两者均旨在捕捉现实果园条件下绿果检测的固有复杂性。

在第一种方案中，所有可见的未成熟苹果均被统一标注为“greenfruit”类别，无论其可见程度或遮挡情况如何。为此，我们共上传并处理了857张高分辨率果园图像至Roboflow平台。如图3e中间图像所示，该数据集涵盖了绿色果实的多种外观形态，最终生成4,125个独立对象标签。这种统一标注方案适用于建立基准检测性能，但未能捕捉部分遮挡或背景融合等视觉挑战的动态特征。<sup>1</sup> 为更深入探究这些复杂性，我们开发了第二种标注方案以创建多类别数据集。该方案将每个绿色果实划分为两类：被遮挡绿色果实（occluded greenfruit）和未遮挡绿色果实（non-occluded greenfruit）。分类标准基于可见程度：表面区域至少90%清晰可见、未被枝叶或其他果实遮挡的绿色果实标注为未遮挡；而任何被重叠苹果、交叉叶片或枝条部分遮挡的果实则标注为被遮挡。这种动态标注方法如图3e最右侧图像所示<sup>2</sup>。

然而，标注过程因标签模糊性而变得复杂，这是计算机视觉任务尤其是自然环境中面临的关键问题。标签模糊性指的是由于视觉边界不清晰、物体重叠或可见性不一致导致的标注不确定性或主观性。本研究中出现了几种标签模糊性的实际案例。首先，当多个绿色果实紧密聚集时，往往难以判断是一个果实部分遮挡了另一个，还是它们并排生长。其次，某些果实因光照和阴影看似被遮挡，而非实际物理阻碍，导致不同图像间的标注不一致。第三，枝叶有时会模仿未成熟果实的纹理和颜色，使得难以区分目标对象与背景。第四，图像边缘的部分遮挡常让标注者犹豫应将其归类为被遮挡还是单纯因视野范围被截断。这些例子凸显了为何真实果园环境中的绿色果实检测特别容易出现标注不一致。尽管分类指南被严格执行，但物体几何形状、环境纹理与可见性之间复杂的相互作用，使得完全客观的标注难以实现。因此，标签模糊性这一术语被用来描述数据集固有的主观性及其在模型训练与评估过程中可能引入的变异性。

### 2.3. Training Object Detection Models

#### 2.3.1. Training RF-DETR Object Detection Model

RF-DETR是一种基于Transformer的实时目标检测架构，针对准确性和效率<sup>1,2</sup>进行了优化。如图4a所示，RF-DETR构建于

Deformable DETR与LW-DETR的基础框架，集成了预训练的DINOv2视觉变换器作为其主干网络。该主干通过自监督学习增强了跨领域泛化能力，使模型能高度适应农业环境中诸如青果检测等特定领域的挑战。

RF-DETR的一项关键创新在于其能够摒弃传统目标检测组件，如锚框和非极大值抑制（NMS）。取而代之的是，它采用基于Transformer的编码器-解码器架构，结合可变形交叉注意力机制，选择性地聚焦于空间相关特征，从而提升在遮挡、杂乱和伪装场景下的检测性能。与传统DETR变体不同，RF-DETR采用单尺度特征提取策略以降低计算开销，由此在不牺牲精度的情况下实现更快的推理速度。

该模型有两种变体可供选择：RF-DETR-Base（2900万参数）和RF-DETR-Large（1.28亿参数）。本研究选择了\*\*RF-DETR-Base\*\*模型，因其在计算效率与高检测精度之间取得了平衡，适合野外机器人技术的实时处理。RF-DETR-Base模型在COCO基准测试中实现了53.3的mAP，在RF100-VL数据集上达到86.7的mAP@50，使其成为少数能在实时设置下mAP@50:95超过60%的模型之一。

训练遵循了Roboflow官方实现方案。模型采用DINOv2预训练权重进行初始化，使用AdamW优化器以1e-4的学习率和8的批量大小训练300个周期。训练过程融合了受RT-DETR启发的混合编码器优化技术以及可变形注意力机制。损失函数包含分类任务的交叉熵，以及边界框回归中L1损失与GIoU损失的组合。

此外，该研究采用了对比去噪训练以增强对部分可见及小尺寸目标的检测鲁棒性。RF-DETR还通过协作式标签分配策略应对标注模糊场景下的稳定性问题，并支持多分辨率输入（640–1280像素），无需重新训练即可实现延迟与精度的灵活权衡。这一配置使RF-DETR-Base成为在复杂果园环境中检测遮挡与伪装状态未成熟绿色果实的强效高效模型。

#### 2.3.2. Training YOLOv12 Object Detection Model

YOLOv12代表了基于CNN的目标检测领域的一次变革性飞跃，它将传统卷积架构的高效性与受注意力启发的机制相融合，以应对现代计算机视觉的需求[32]。不同于以往YOLO版本，该模型采用R-ELAN（残差高效层聚合网络）作为核心骨干结构（如图4b所示），通过结合残差连接与多尺度特征融合，既解决了梯度瓶颈问题，又提升了网络深度间的特征复用效率。创新的7×7可分离卷积层取代了标准3×3卷积核，在比传统大核卷积减少60%参数量的同时保留了空间上下文信息，并隐式编码了位置关系，从而有效规避了显式位置嵌入的需求。

<sup>1</sup><https://blog.roboflow.com/rf-detr/>

<sup>2</sup><https://github.com/roboflow/rf-detr>

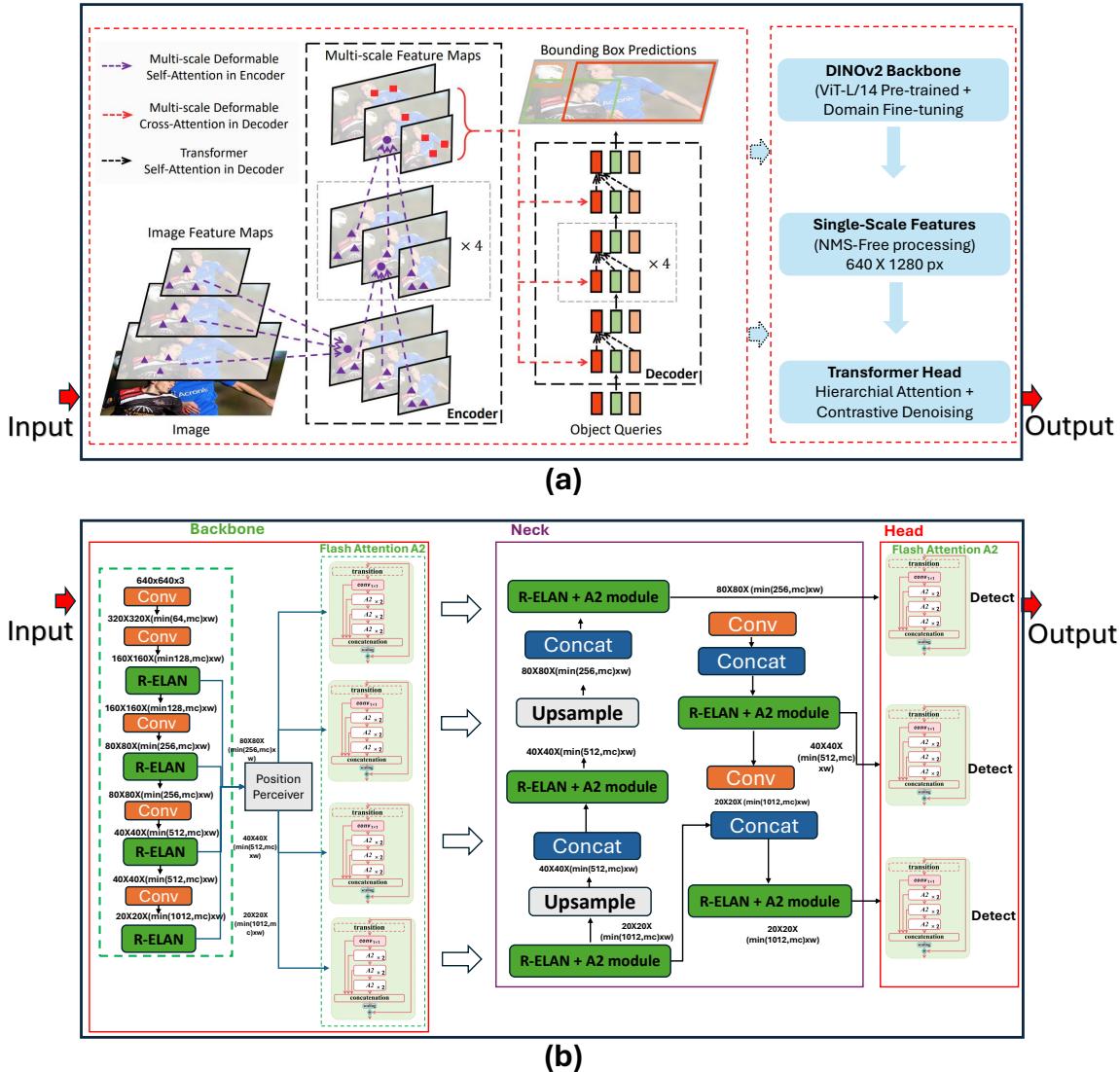


Figure 4: (a) RF-DETR Architecture diagram for object detection ; (b) YOLOv12 Architecture Diagram for object detection

transformer-based detectors. The neck architecture integrates FlashAttention-optimized area attention, dividing feature maps into four horizontal/vertical regions for localized processing without sacrificing global context, achieving 40% reduced memory overhead compared to standard self-attention implementations. These innovations enable state-of-the-art accuracy while maintaining real-time performance, with the YOLOv12-S variant outperforming RT-DETR-R18 in both speed (1.2× faster) and precision (62.1 vs 59.3 COCO mAP). The architecture further supports multi-task learning through unified prediction pathways, allowing simultaneous object detection, oriented bounding box (OBB) estimation, and instance segmentation via specialized heads—a first for the YOLO series. Hardware-aware optimizations ensure sub-10ms inference on edge devices, with the 12n variant (2.1M parameters) achieving 9.8ms latency while maintaining robust detection of sub-50px objects through lightweight MLP ratios (1.2-2.0 vs traditional 4.0) in task-specific heads.

YOLOv12’s architectural refinements optimize convolutional operations for contemporary hardware while introducing transformer-like capabilities through innovative attention hybrids. The area attention mechanism processes feature map segments independently through FlashAttention’s memory-efficient algorithms, enabling precise region-specific focus without the computational burden of full self-attention. This design philosophy extends to the model’s scalability, offering four configurations (12n/12s/12m/12x) ranging from 2.1M to 42M parameters to accommodate edge deployments (Jetson Nano) to cloud clusters (A100 GPUs). Unlike previous YOLO versions limited to axis-aligned detection, YOLOv12 introduces an OBB head with angle prediction capabilities, critical for aerial imagery and document analysis. Training stability is enhanced through block-level residual scaling in R-ELAN, which prevents feature degradation in deep networks while maintaining the single-pass efficiency that defines the YOLO series. Benchmark results demonstrate 4-8% higher mAP than YOLOv11

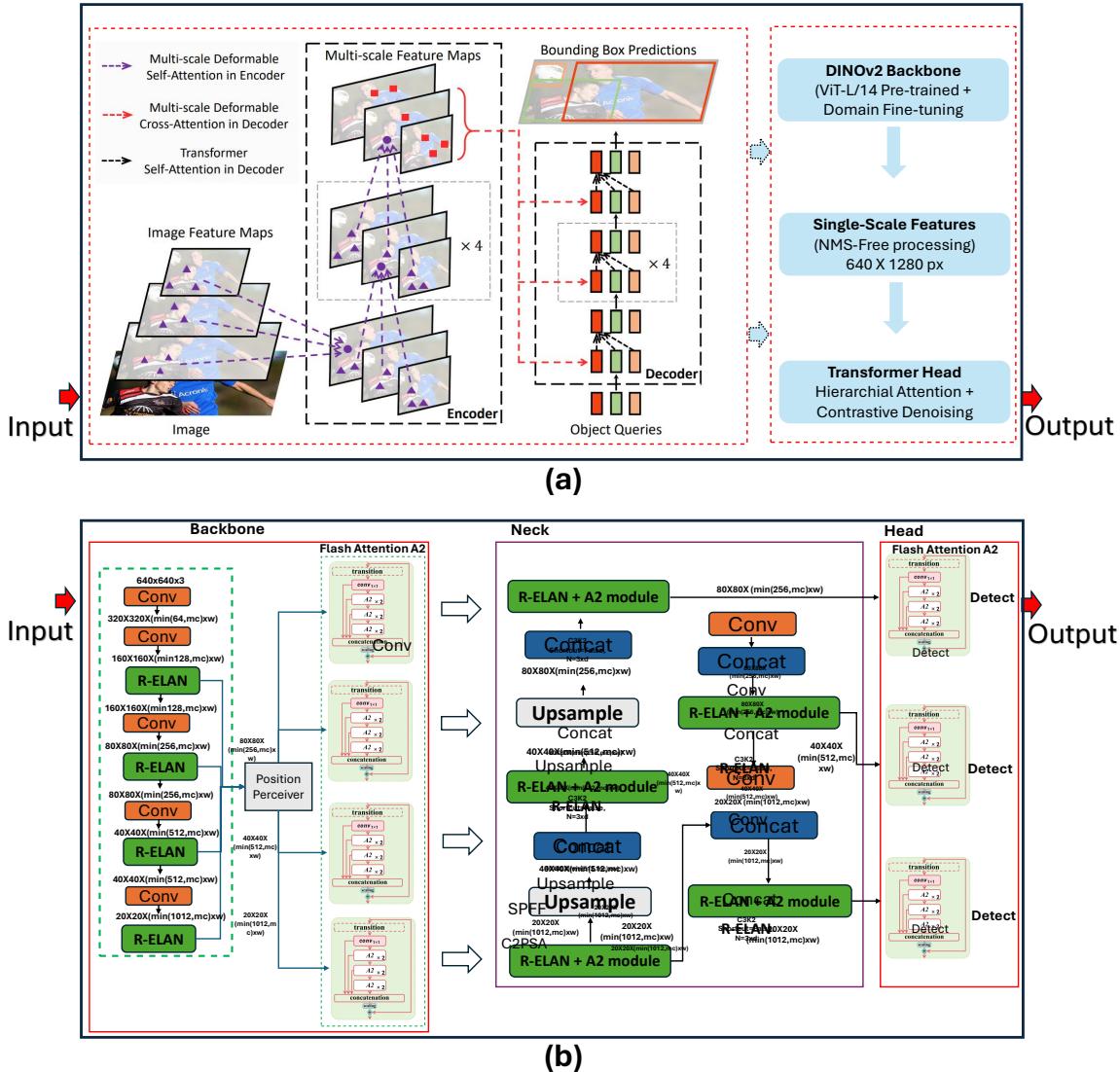


图4: (a) RF-DETR架构 命物体检测的架构示意图; (b) YOLOv12架构图

用于目标检测的m

基于Transformer的检测器。该颈部架构集成了经FlashAttention优化的区域注意力机制，将特征图划分为四个水平/垂直区域进行局部处理，同时不牺牲全局上下文，相比标准自注意力实现减少了40%的内存开销。这些创新在保持实时性能的同时实现了最先进的精度，其中YOLOv12-S变体在速度（快1.2×）和精度（62.1 vs 59.3 COCO mAP）上均优于RT-DETR-R18。该架构通过统一预测路径进一步支持多任务学习，可同时进行目标检测、定向边界框（OBB）估计和通过专用头实现的实例分割——这是YOLO系列的首创。硬件感知优化确保在边缘设备上实现低于10ms的推理延迟，其中12n变体（210万参数）在任务专用头中采用轻量级MLP比例（1.2-2.0 vs 传统4.0）保持对50像素以下物体的鲁棒检测，同时实现9.8ms延迟。

YOLOv12的架构优化针对现代硬件改进了卷积运算，同时通过创新的注意力混合机制引入了类Transformer的能力。其区域注意力机制借助FlashAttention的内存高效算法独立处理特征图片段，实现了精准的区域聚焦，而无需承担完整自注意力的计算负担。这一设计理念延伸至模型的可扩展性，提供四种配置（12n/12s/12m/12x），参数量从210万到4200万不等，可适配从边缘设备（Jetson Nano）到云端集群（A100 GPU）的部署需求。不同于以往仅支持轴对齐检测的YOLO版本，YOLOv12新增了具备角度预测功能的OBB检测头，这对航拍图像和文档分析至关重要。通过R-ELAN中的块级残差缩放技术增强了训练稳定性，在防止深度网络特征退化的同时，保持了YOLO系列标志性的单次推理效率。基准测试显示其mAP较YOLOv11提升4-8%。

across all variants, with the 12x model achieving 68.9 mAP on COCO, surpassing similarly sized transformer hybrids like DINO-DETR in small object detection tasks. The architecture's separation of feature extraction (backbone) and attention-driven refinement (neck) allows targeted optimization, enabling the 12s variant to process 4K video streams at 45 FPS on an NVIDIA T4 GPU. By integrating the parameter efficiency of CNNs with the contextual awareness of attention mechanisms, YOLOv12 establishes a new standard for real-time vision systems, particularly in industrial applications requiring simultaneous detection, segmentation, and geometric prediction under strict latency constraints.

#### 2.4. Training Methodology

The training procedures for both deep learning models RF-DETR and YOLOv12 were carried out under identical experimental settings to ensure a fair and rigorous comparison. All training was conducted on a workstation equipped with an Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz (10 cores, 20 threads), running Ubuntu 24.04.1, and supported by an NVIDIA RTX A5000 GPU with 24 GB VRAM. This high-performance hardware configuration ensured sufficient computational resources for training large-scale object detection models. The RF-DETR object detection model, specifically the Base variant, was trained for 50 epochs on the single-class greenfruit dataset and 100 epochs on the multi-class dataset.

Notably, RF-DETR demonstrated rapid convergence in the single-class setting, with performance plateauing at under 20 epochs, highlighting the model's efficient learning dynamics and its suitability for low-epoch training regimes. YOLOv12 models, including YOLOv12X, YOLOv12L, and YOLOv12N, were each trained for 100 epochs for both single-class and multi-class datasets to ensure convergence and optimal generalization. RF-DETR was implemented in PyTorch using Roboflow's `rf-detr` framework, which integrates a Deformable DETR architecture with a pre-trained DINOV2 backbone to enhance global context modeling and cross-domain adaptability. The YOLOv12 models were trained using the official Ultralytics PyTorch framework, optimized for fast detection and efficient edge deployment. For both models, the input image resolution was standardized to 640×640 pixels, a resolution commonly adopted in orchard-based object detection tasks.

The models training was performed using FP32 precision with a batch size of approximately 16 images per iteration. The software environment included CUDA 11.7+ and cuDNN 8.4+, ensuring full compatibility with GPU acceleration and deep learning libraries. This standardized setup enabled a reliable comparative evaluation of model accuracy, convergence behavior, and training efficiency across both transformer- and CNN-based architectures.

#### 2.5. Performance Evaluation

To rigorously assess the capabilities of RF-DETR and YOLOv12 in identifying greenfruits in a complex orchard environment, a comprehensive evaluation was conducted using

standardized metrics. Both models underwent training and testing under uniform conditions utilizing the same datasets, number of training epochs, learning rates, optimizers, and batch sizes to ensure an equitable comparison between the CNN-based YOLOv12 and Transformer-based RF-DETR architectures.

#### Detection Evaluation Metrics

The evaluation metrics employed were Precision, Recall, F1-Score, mean Average Precision (mAP@50 and mAP@50:95), and mean Intersection over Union (mIoU). These metrics quantify the performance based on the interaction between predicted bounding boxes and ground truth annotations:

- **True Positive (TP):** A predicted bounding box correctly identifies a ground truth fruit with an Intersection over Union (IoU)  $\geq$  the defined threshold (commonly 0.50).
- **False Positive (FP):** A predicted bounding box either insufficiently overlaps with any ground truth box (IoU  $<$  0.50) or erroneously marks a non-existent object.
- **False Negative (FN):** A real fruit is overlooked by the detection model, with no corresponding predicted box sufficiently overlapping it.

The metrics were calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Precision assesses the accuracy of the detected greenfruits, Recall gauges the completeness of the detection, and F1-Score balances both aspects to provide a single measure of model efficacy.

#### Intersection over Union (IoU) and Mean IoU (mIoU)

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN} \quad (4)$$

IoU quantifies the exactness of the overlap between the predicted and actual bounding boxes, crucial in scenarios with densely packed and overlapping fruits. mIoU averages the IoU across all detections to give a holistic measure of spatial accuracy.

#### mAP@50 and mAP@50:95

$$\text{mAP}@50 = \frac{1}{N} \sum_{i=1}^N \text{AP}_i(\text{IoU} \geq 0.50) \quad (5)$$

$$\text{mAP}@50:95 = \frac{1}{10} \sum_{t=0.50}^{0.95} \text{mAP}_t \quad (6)$$

在所有变体中，12倍模型在COCO数据集上实现了68.9 mA P，在小物体检测任务中超越了类似规模的Transformer混合架构（如DINO-DETR）。该架构将特征提取（主干网络）与注意力驱动优化（颈部网络）分离，实现了针对性优化，使得12s变体能在NVIDIA T4 GPU上以45 FPS处理4K视频流。通过融合CNN的参数效率与注意力机制的上下文感知能力，YOLOv12为实时视觉系统树立了新标杆，尤其适用于需要同时进行检测、分割和几何预测且对延迟有严格要求的工业应用场景。

#### 2.4. Training Methodology

深度学习模型RF-DETR与YOLOv12的训练流程均在完全一致的实验环境下进行，以确保对比的公平性与严谨性。所有训练任务均搭载英特尔酷睿i9-10900K处理器（3.70 GHz主频，10核20线程）、Ubuntu 24.04.1操作系统及配备24GB显存的NVIDIA RTX A5000显卡的工作站完成。该高性能硬件配置为大规模目标检测模型的训练提供了充足算力支撑。其中RF-DETR目标检测模型（Base版本）在单类绿果数据集上训练50个周期，在多类数据集上训练100个周期。

值得注意的是，RF-DETR在单类别设置中展现出快速收敛特性，性能在20个训练周期内即趋于稳定，凸显了该模型高效的学习动态及其对低周期训练机制的良好适应性。YOLOv12系列模型（包括YOLOv12X、YOLOv12L和YOLOv12N）在单类别与多类别数据集上均训练了100个周期，以确保充分收敛和最优泛化能力。RF-DETR基于PyTorch框架通过Roboflow的rf-detector工具实现，该方案将可变形DETR架构与预训练的DINOv2骨干网络相结合，增强了全局上下文建模与跨域适应能力。YOLOv12模型则采用官方Ultralytics PyTorch框架进行训练，该框架专为快速检测与高效边缘部署优化。所有模型的输入图像分辨率均统一为640×640像素——这一分辨率在果园目标检测任务中被广泛采用。

模型训练采用FP32精度进行，每批次约16张图像。软件环境包含CUDA 11.7+和cuDNN 8.4+，确保与GPU加速及深度学习库的完全兼容。这一标准化配置使得我们能够对基于Transformer和CNN架构的模型在准确性、收敛行为及训练效率方面进行可靠的对比评估。

#### 2.5. Performance Evaluation

为严格评估RF-DETR与YOLOv12在复杂果园环境中识别绿色果实的性能，我们采用{v\*}进行了全面测试

标准化指标。两种模型在相同数据集、训练周期数、学习率、优化器和批量大小的统一条件下进行了训练和测试，以确保基于CNN的YOLOv12与基于Transformer的RF-DETR架构之间公平的比较。

#### 检测评估指标

采用的评估指标包括精确率（Precision）、召回率（Recall）、F1分数（F1-Score）、平均精度均值（mAP@50与mAP@50:95）以及平均交并比（mIoU）。这些指标通过预测边界框与真实标注框之间的交互关系量化模型性能：

- **真正例（TP）**：预测边界框正确识别出一个真实水果，并且其交并比（IoU） $\geq$ 超过设定的阈值（通常为0.50）。
- **假阳性（FP）**：预测的边界框与任何真实标注框的重叠不足（IoU  $\{v^*\} \leq 0.50$ ），或错误标记了不存在的物体。
- **假阴性（FN）**：真实的水果未被检测模型识别，且没有足够重叠的预测框与之对应。

指标计算方式如下：

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

精确度评估检测到的绿色水果的准确性，召回率衡量检测的完整性，而F1-Score则平衡这两个方面，提供一个衡量模型效能的单一指标。

交并比（IoU）与平均交并比（mIoU）

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN} \quad (4)$$

IoU量化了预测边界框与实际边界框之间重叠的精确程度，在果实密集且相互重叠的场景中至关重要。mIoU通过计算所有检测结果的IoU平均值，提供了空间准确性的整体衡量指标。

mAP@50 和 mAP@50:95

$$\text{mAP}@50 = \frac{1}{N} \sum_{i=1}^N \text{AP}_i(\text{IoU} \geq 0.50) \quad (5)$$

$$\text{mAP}@50:95 = \frac{1}{10} \sum_{t=0.50}^{0.95} \text{mAP}_t \quad (6)$$

mAP@50 measures the mean Average Precision at an IoU threshold of 0.50, commonly used to evaluate detection effectiveness. mAP@50:95 averages the AP at ten IoU thresholds from 0.50 to 0.95 (in increments of 0.05), offering a rigorous assessment of a model’s precision across a range of criteria from loose to strict overlaps, reflecting the model’s robustness in both precise and approximate detection scenarios.

### Application to Our Dataset

For the single-class detection task, all greenfruits were uniformly considered, while the multi-class task also evaluated the accuracy of classifying fruits as occluded or non-occluded. Misclassifications of occlusion status were considered false positives, and undetected fruits, particularly those obscured by occlusion, were counted as false negatives.

## 3. Results

The results for single-class and multi-class greenfruit detection using RF-DETR and YOLOv12 are presented to evaluate their performance in detecting green apples in complex orchard environments. Figure 5 displays three examples that highlight how each model performs in single-class detection scenarios, illustrating their efficacy in challenging conditions characterized by dense foliage and partial occlusions. Similarly, Figure 6 showcases three examples for multi-class detection, focusing on the models’ ability to handle label ambiguity effectively.

Each example includes the original RGB image collected in the orchard (left), detection output from RF-DETR (middle), and detection output from YOLOv12 (right). Key regions of interest are highlighted with yellow dotted circles, focusing on areas where fruitlets were either clustered, camouflaged, or heavily occluded. In Figure 5a, three immature green apples appeared closely clustered within a dense canopy, with significant partial occlusions caused by overlapping leaves. The original image, shown on the left, presented a challenging scenario due to low fruit-background contrast and complex foliage structure. As illustrated in the middle image of Figure 5a, RF-DETR successfully detected all three greenfruit instances, correctly bounding each fruit despite their partial visibility. In contrast, YOLOv12, shown on the right, detected only two out of the three apples and failed to identify the third fruit, which was most heavily occluded. This result highlighted RF-DETR’s superior capability in handling complex spatial relationships and occlusions. Figure 5b provided another challenging condition where a single green apple within the yellow dotted circle was camouflaged due to its visual similarity with the surrounding canopy. Despite the low contrast between the fruit and the background, RF-DETR accurately identified the greenfruit, as shown in the middle figure. Conversely, YOLOv12 failed to detect this fruit, indicating its limitations in distinguishing camouflaged targets in homogeneous backgrounds.

In Figure 5c, a different scenario was examined where only a small portion (approximately 10%) of the fruit’s calyx was visible due to heavy occlusion by a leaf and low ambient lighting. The original RGB image (left) demonstrated minimal visible

surface area of the fruit. Remarkably, RF-DETR still managed to detect the partially exposed fruit in the middle image, whereas YOLOv12 again failed to register the object in its detection output. These examples consistently demonstrated RF-DETR’s higher sensitivity and robustness in single-class greenfruit detection, especially under conditions of occlusion, camouflage, and low visibility, compared to the CNN-based YOLOv12 model.

Figure 6 presents qualitative comparisons between RF-DETR and YOLOv12 for multi-class greenfruit detection, where fruitlets were classified as either occluded or non-occluded. This evaluation highlights the models’ performance in handling label ambiguity—situations where visibility is unclear due to clustering, occlusion, or edge truncation.

Likewise, in Figure 6a, a dense cluster of greenfruits appears near the image edge, creating a highly ambiguous scene. As shown in the rightmost image, YOLOv12 detected 7 greenfruit instances in this region. However, ground truth annotation confirmed that only 5 greenfruits were actually present. YOLOv12 misclassified background textures or overlapping canopy features as non-occluded apples, resulting in false positives. In contrast, RF-DETR, shown in the middle image, correctly detected the 5 actual greenfruits but missed classifying them into occluded/non-occluded categories with high certainty. In this example, YOLOv12 appeared visually more active but less accurate, while RF-DETR provided precise detection with lower misclassifications.

Furthermore, in Figure 6 b, yellow circles in the original image (left) highlight true greenfruits. RF-DETR detected 12 apples, including an occluded apple at the bottom of the frame, which was correctly labeled occluded (middle). YOLOv12 detected 11 apples but incorrectly labeled the bottom occluded apple as non-occluded (right), suggesting that RF-DETR object detection model performed better in differentiating occlusion classes, likely due to its global attention modeling.

Likewise, Figure 6c presents a challenging low-visibility case, where only 10% of a greenfruit is visible beneath leaf cover (indicated by a blue arrow). RF-DETR successfully detected and classified it as occluded (middle), while YOLOv12 failed to detect the fruit at all (right). This reinforces RF-DETR’s strength in handling extreme occlusion.

### 3.1. Evaluation of Precision, Recall and F1-Score

Among all the models evaluated, YOLOv12N achieved the highest performance in terms of recall (0.8901) and the F1 score (0.8784) for single-class greenfruit detection, indicating its strong ability to detect almost all greenfruit instances while maintaining balanced precision. However, in terms of precision, YOLOv12L outperformed all other configurations of YOLOv12 and RF-DETR object detection model, achieving a top value of 0.8892 in single-class detection. This demonstrates the superior ability of YOLOv12L in reducing false positives and making accurate predictions. Detailed precision, recall, and F1 metrics for all models and detection types are presented in Table 1.

mAP@50衡量的是在IoU阈值为0.50时的平均精度均值，常用于评估检测效果的有效性。mAP@50:95则在0.50至0.95（以0.05为步长）的十个IoU阈值上取AP的平均值，通过对从宽松到严格重叠的一系列标准进行综合考量，提供了对模型精度的严格评估，反映了模型在精确和近似检测场景下的鲁棒性。

### 应用于我们的数据集

在单类别检测任务中，所有绿色果实被统一考量；而多类别任务还需评估将果实正确分类为被遮挡或非遮挡状态的准确率。遮挡状态的误判被视为假阳性，而未检测到的果实——尤其是因遮挡而难以辨识的——则计入假阴性。

## 3. 结果

为评估RF-DETR与YOLOv12在复杂果园环境中检测青苹果的性能，本文展示了单类别与多类别青果检测的结果。图5呈现了三个单类别检测场景的示例，凸显各模型在枝叶茂密、部分遮挡等挑战性条件下的检测效能。图6则通过三个多类别检测实例，重点展示模型有效处理标签模糊性的能力。

每个示例均包含果园中采集的原始RGB图像（左）、RF-DETR的检测输出（中）以及YOLOv12的检测输出（右）。关键关注区域用黄色虚线圆圈标出，聚焦于果实簇集、伪装或严重遮挡的区域。在图5a中，三个未成熟的青苹果密集簇拥在茂密树冠内，因叶片重叠造成显著部分遮挡。左侧原始图像因果实与背景对比度低、枝叶结构复杂，呈现了极具挑战性的场景。如图5a中间图像所示，RF-DETR成功检测到全部三个青苹果实例，即使果实部分可见仍正确框定每个目标。相比之下，右侧的YOLOv12仅检测到三个苹果中的两个，未能识别遮挡最严重的第三个果实。这一结果凸显了RF-DETR在处理复杂空间关系及遮挡方面的卓越能力。图5b展示了另一种挑战性场景：黄色虚线圆圈内单个青苹果因与周围树冠视觉相似而伪装。尽管果实与背景对比度极低，RF-DETR仍准确识别出该青苹果（中间图所示）；而YOLOv12未能检测到此果实，表明其在同质背景中辨识伪装目标的局限性。

在图5c中，研究了一个不同的场景：由于叶片严重遮挡及环境光线不足，果实花萼仅有小部分（约10%）可见。原始RGB图像（左）显示可见部分极少

果实的表面积。值得注意的是，RF-DETR仍成功检测到中间图像中部分暴露的果实，而YOLOv12再次未能在其检测输出中识别该目标。这些实例一致表明，与基于CNN的YOLOv12模型相比，RF-DETR在单类青果检测中具有更高的敏感性和鲁棒性，特别是在遮挡、伪装和低可见度条件下。

图6展示了RF-DETR与YOLOv12在多类青果检测中的定性对比，其中幼果被分类为遮挡或非遮挡状态。该评估重点突出了模型在处理标签模糊性——即因簇集、遮挡或边缘截断导致可见性不明确的情况——时的性能表现。

同样地，在图6a中，图像边缘附近出现了一簇密集的青果，形成了高度模糊的场景。如最右侧图像所示，YOLOv12在该区域检测到了7个青果实例。然而，真实标注确认实际仅存在5个青果。YOLOv12将背景纹理或重叠的树冠特征误分类为非遮挡苹果，导致了误报。相比之下，中间图像展示的RF-DETR正确检测到了5个真实青果，但在高置信度下未能将其分类为遮挡/非遮挡类别。此例中，YOLOv12在视觉上显得更为活跃但准确性较低，而RF-DETR则以更低的误分类率提供了精确检测。

此外，在图6b中，原始图像（左）的黄色圆圈标出了真实的绿色果实。RF-DETR检测到12个苹果，包括画面底部一个被遮挡的苹果，该苹果被正确标记为遮挡状态（中）。YOLOv12检测到11个苹果，但错误地将底部被遮挡的苹果标记为非遮挡状态（右），这表明RF-DETR目标检测模型在区分遮挡类别方面表现更优，这很可能得益于其全局注意机制。

同样地，图6c展示了一个极具挑战性的低可见度案例，其中仅有10%的青果在叶片覆盖下可见（由蓝色箭头标示）。RF-DETR成功检测并将其分类为被遮挡状态（中图），而YOLOv12则完全未能检测到该果实（右图）。这一结果进一步印证了RF-DETR在应对极端遮挡情况时的优势。

### 3.1. Evaluation of Precision, Recall and F1-Score

在评估的所有模型中，YOLOv12N在单类青果检测的召回率（0.8901）和F1分数（0.8784）方面表现最佳，这表明其在检测几乎所有青果实例的同时保持了精确度的平衡能力。然而，就精确度而言，YOLOv12L超越了YOLOv12的所有其他配置以及RF-DETR目标检测模型，在单类检测中达到了0.8892的最高值，这证明了YOLOv12L在减少误报和做出准确预测方面的卓越能力。表1详细列出了所有模型和检测类型的精确度、召回率及F1指标。

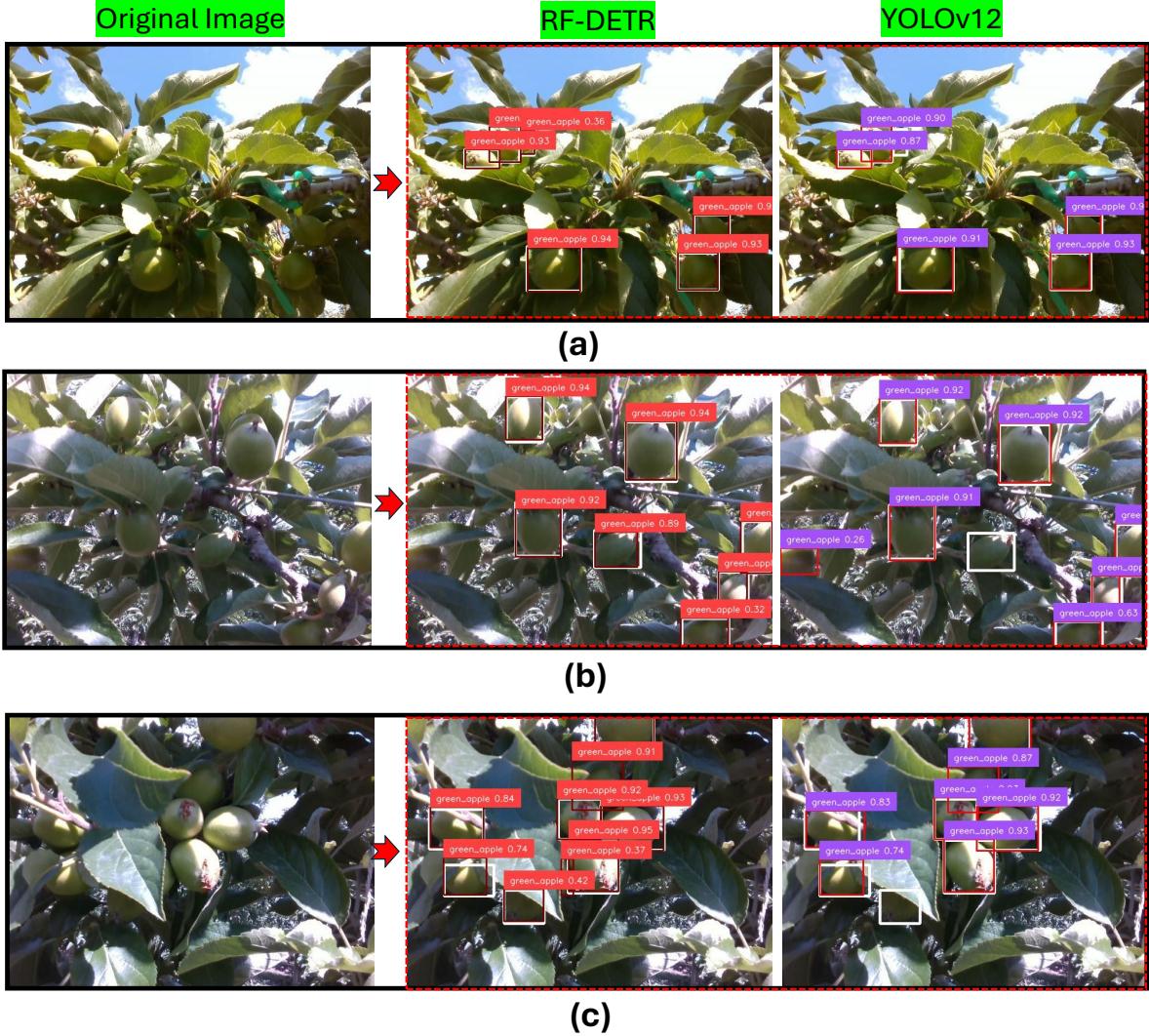


Figure 5: Visual comparison of single-class greenfruit detection using RF-DETR and YOLOv12 in complex orchard scenes. a) Three clustered greenfruits partially occluded by dense canopy; RF-DETR detected all, YOLOv12 missed one. b) A camouflaged greenfruit blending into the canopy; RF-DETR correctly detected it, YOLOv12 failed. c) A heavily occluded greenfruit with only the calyx visible under low light; RF-DETR identified it, YOLOv12 missed detection.

Table 1: Comparative analysis of Precision, Recall, and F1 Score for single-class and multi-class greenfruit detection using RF-DETR (Transformer-based) and YOLOv12 (CNN-based) object detection algorithms. The table presents model performance across different YOLOv12 configurations (X, L, N) and highlights their effectiveness in detecting greenfruits under varying complexity and class conditions in orchard environments.

Models	Single-Class			Multi-Class		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
RF-DETR	0.8663	0.8828	0.8744	0.7652	0.8109	0.7874
YOLOv12X	0.8797	0.8595	0.8694	0.6986	0.8261	0.7570
YOLOv12L	0.8892	0.8631	0.8759	0.7692	0.7827	0.7759
YOLOv12N	0.8671	0.8901	0.8784	0.7569	0.7406	0.7487

### 3.2. Analysis of Mean Average Precision (mAP)

For single-class greenfruit detection, RF-DETR outperformed all other models with the highest mAP@50 of 0.9464, indicating its superior ability to accurately detect and localize greenfruits with sufficient overlap. Furthermore, RF-DETR achieved a mAP@50:95 of 0.7433, the second-highest among

the tested models. Although YOLOv12N achieved a slightly higher mAP@50:95 of 0.7620, RF-DETR's consistently higher mAP@50 suggests more reliable performance in practical orchard detection scenarios, especially when bounding box precision at the 50% threshold is critical. In the multi-class detection scenario, where greenfruits were labeled as either occluded or

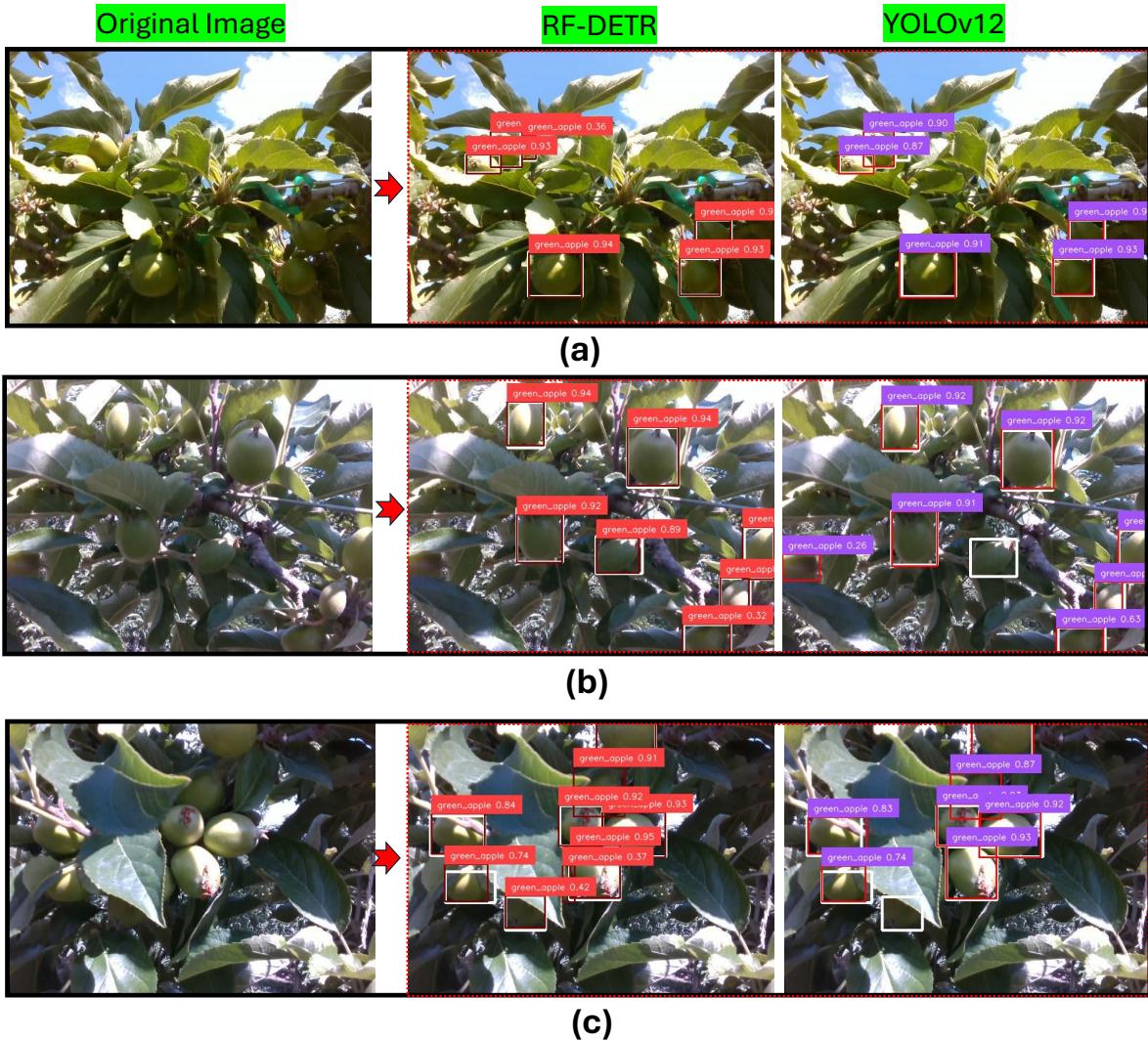


图5：复杂果园场景下RF-DETR与YOLOv12单类青果检测的视觉对比。a) 三枚簇拥的青果被茂密树冠部分遮挡；RF-DETR全部检出，YOLOv12漏检一枚。b) 一枚与树冠颜色融为一体伪装青果；RF-DETR正确识别，YOLOv12检测失败。c) 低光照条件下仅花萼可见的严重遮挡青果；RF-DETR成功定位，YOLOv12未能检出。

表1：基于RF-DETR（Transformer架构）与YOLOv12（CNN架构）目标检测算法在单类别与多类别青果检测中的精确率、召回率及F1分数对比分析。该表展示了不同YOLOv12配置（X、L、N）下的模型性能，并突出其在果园环境中应对不同复杂度与类别条件时检测青果的有效性。

Models	Single-Class			Multi-Class		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
RF-DETR	0.8663	0.8828	0.8744	0.7652	0.8109	0.7874
YOLOv12X	0.8797	0.8595	0.8694	0.6986	0.8261	0.7570
YOLOv12L	0.8892	0.8631	0.8759	0.7692	0.7827	0.7759
YOLOv12N	0.8671	0.8901	0.8784	0.7569	0.7406	0.7487

### 3.2. Analysis of Mean Average Precision (mAP)

在单类青果检测任务中，RF-DETR以0.9464的最高mAP@50值超越了所有其他模型，这表明其在实现充分重叠条件下精准检测与定位青果方面具有卓越能力。此外，RF-DETR还取得了0.7433的mAP@50:95值，位列所有模型中的第二高位。

测试的模型中。尽管YOLOv12N取得了稍高的mAP@50:95值0.7620，但RF-DETR持续更高的mAP@50表明在实际果园检测场景中性能更为可靠，尤其是在50%阈值下的边界框精度至关重要时。在多类别检测场景下，绿色果实被标记为被遮挡或

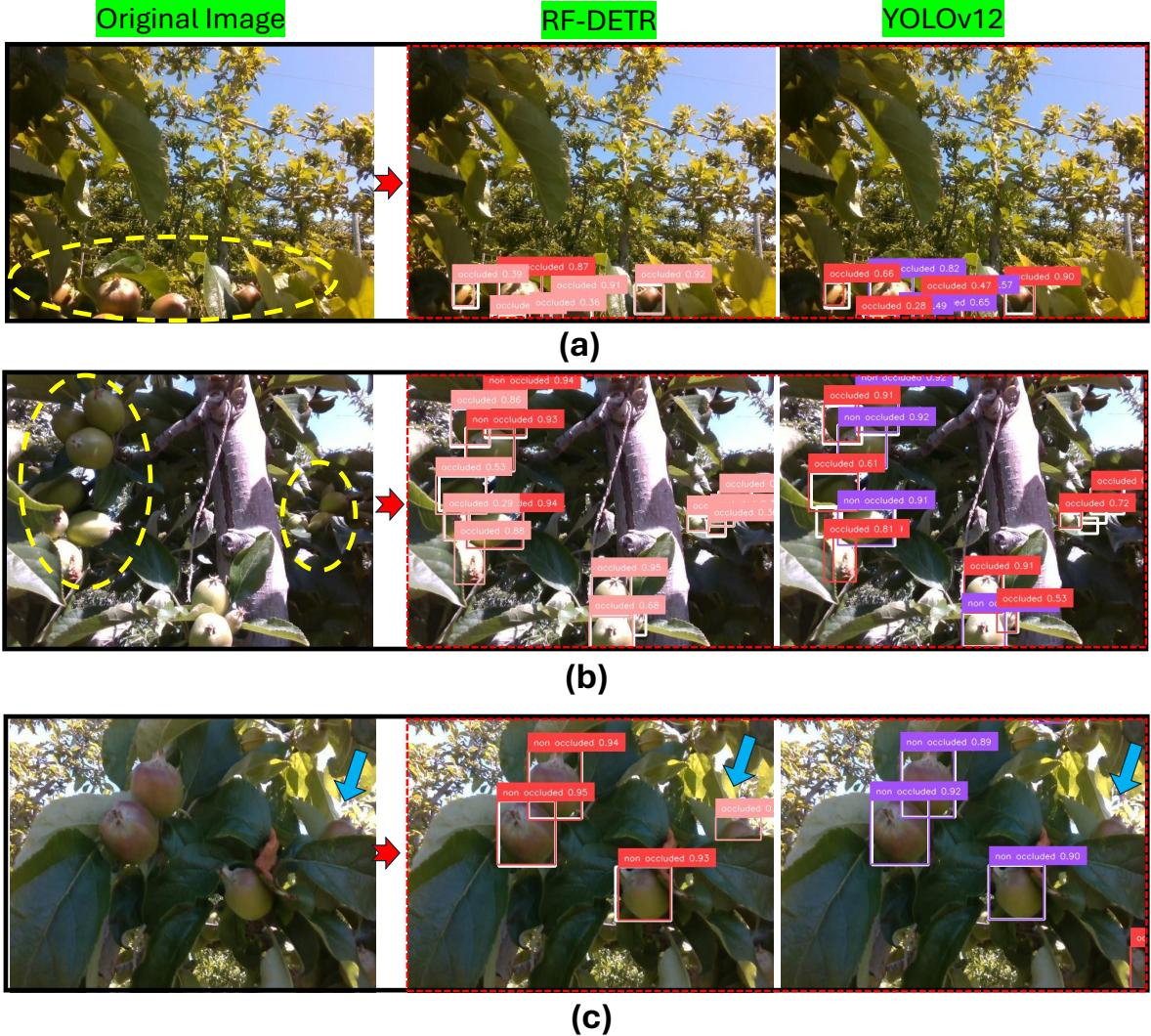


Figure 6: **Visual comparison of multi-class greenfruit detection by RF-DETR and YOLOv12 under label ambiguity.** (a) A dense fruit cluster at the image edge; YOLOv12 over-detected with false positives, while RF-DETR correctly detected 5 true greenfruits. (b) An occluded apple at the bottom; RF-DETR correctly labeled it as occluded, while YOLOv12 misclassified it as non-occluded. (c) A highly occluded fruit with only 10% (approx) visibility; RF-DETR detected it as occluded, whereas YOLOv12 missed the detection entirely.

non-occluded, YOLOv12L achieved the highest mAP@50:95 of 0.6622, slightly outperforming YOLOv12X and RF-DETR object detection, which achieved 0.6609 and 0.6530 respectively. This suggests that YOLOv12L was marginally better in maintaining detection consistency across varying degrees of overlap under label ambiguity. However, RF-DETR object detection model achieved the highest mAP@50 of 0.8298 in the multi-class setting, confirming its strength in confidently detecting objects with at least 50% spatial alignment. These findings indicate that RF-DETR excels in spatially accurate detections, particularly for clearly visible fruits, while YOLOv12L performs slightly better under complex classification scenarios involving occlusion. The complete visualization of these metrics is presented in Figure 7, where Figure 7a shows the mAP@50 for single-class detection, and Figure 7b illustrates mAP@50 and mAP@50:95 for multi-class detection.

### 3.3. Training Dynamics and Model Convergence Analysis

Figure 8 provides a detailed visualization of the mean Average Precision (mAP@50) against the number of training epochs for both RF-DETR and YOLOv12X models, shedding light on their learning efficiency and stability during the training phase. In Figure 8a, which tracks performance for single-class greenfruit detection, RF-DETR, a transformer-based object detection model, demonstrates an impressive early convergence, plateauing before 10 epochs. This rapid stabilization underscores RF-DETR's swift adaptability to complex orchard scenes, a significant advance over traditional CNN-based models like YOLOv12X.

Similarly, Figure 8b illustrates the training progression for multi-class detection scenarios. Here, RF-DETR also shows superior convergence, reaching stability at around 20 epochs, far sooner than its CNN counterpart, which continues to seek equilibrium. This faster convergence of RF-DETR in both single

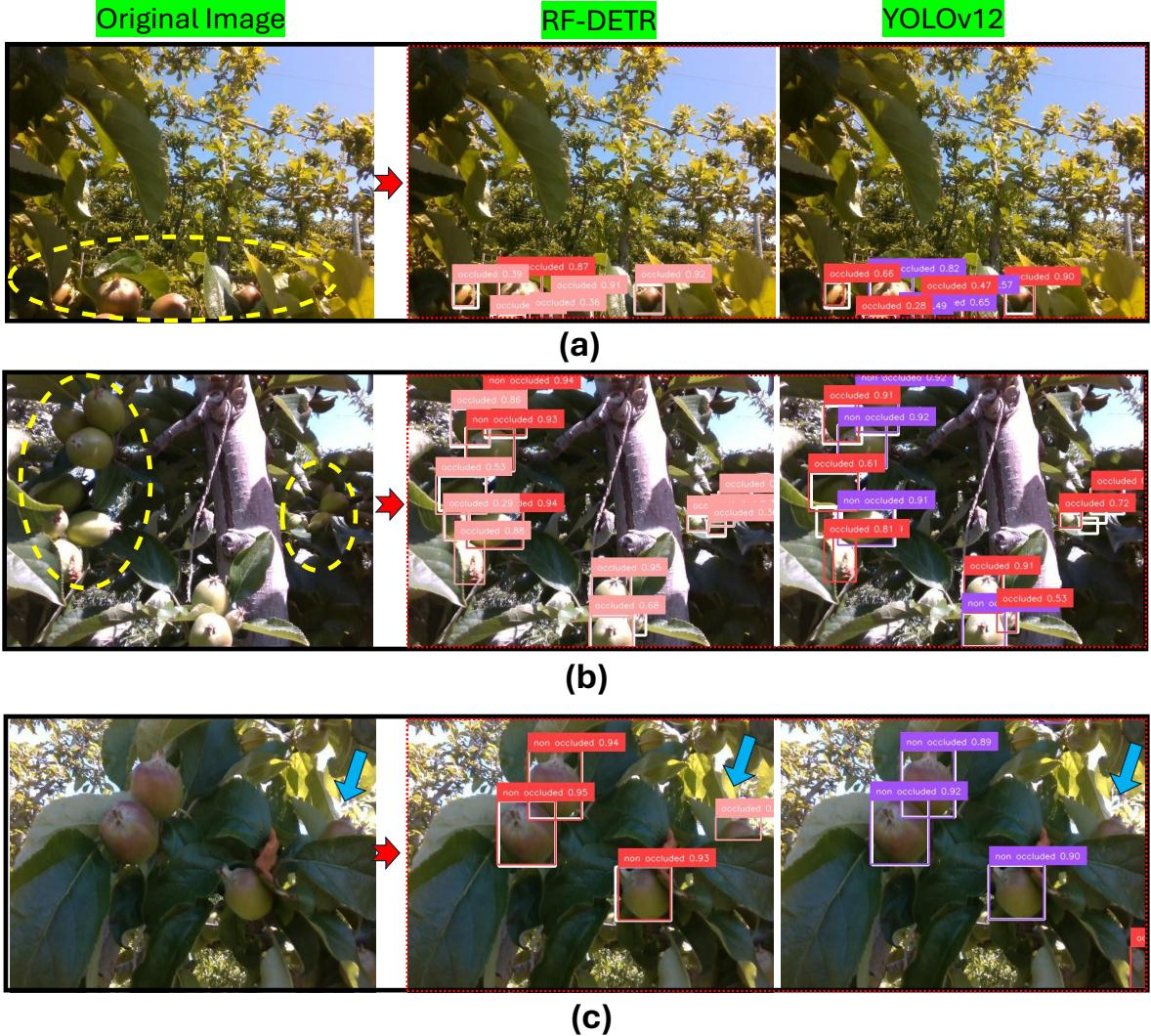


图6：RF-DETR与YOLOv12在标签模糊场景下的多类青果检测视觉对比。(a) 图像边缘处的密集果实簇；YOLOv12出现误检产生假阳性，而RF-DETR正确检测出5个真实青果。(b) 底部被遮挡的苹果；RF-DETR正确标记为遮挡状态，YOLOv12则错误分类为未遮挡。(c) 可见度仅约10%的高度遮挡果实；RF-DETR成功检测为遮挡目标，YOLOv12则完全漏检。

在非遮挡情况下，YOLOv12L以0.6622的mAP@50:95最高值略胜一筹，小幅超越YOLOv12X（0.6609）和RF-DETR目标检测模型（0.6530）。这表明在标签模糊条件下，YOLOv12L能更稳定地保持不同重叠程度下的检测一致性。然而在多类别场景中，RF-DETR模型以0.8298的mAP@50值占据优势，印证了其在空间对齐度达50%以上的目标检测任务中的强项。这些发现说明：RF-DETR擅长空间精度要求高的检测任务（尤其针对清晰可见的水果目标），而YOLOv12L在涉及遮挡的复杂分类场景中表现更优。完整指标可视化结果如图7所示，其中图7a展示单类别检测的mAP@50，图7b则呈现多类别检测的mAP@50与mAP@50:95指标。

### 3.3. Training Dynamics and Model Convergence Analysis

图8详细展示了RF-DETR与YOLOv12X模型在训练周期数下的平均精度均值(mAP@50)变化，揭示了二者在训练阶段的学习效率与稳定性差异。在图8a的单类青果检测性能曲线中，基于Transformer架构的目标检测模型RF-DETR表现出惊人的早期收敛性，在10个周期前即达到稳定平台。这种快速稳定态势凸显了RF-DETR对复杂果园场景的敏捷适应能力，相较YOLOv12X等传统CNN模型实现了显著突破。

同样地，图8b展示了多类别检测场景下的训练进程。在此，RF-DETR同样展现出更优的收敛性，约20个周期即达到稳定状态，远早于其CNN对照模型——后者仍在持续寻求平衡。RF-DETR在单类别检测中更快的收敛速度

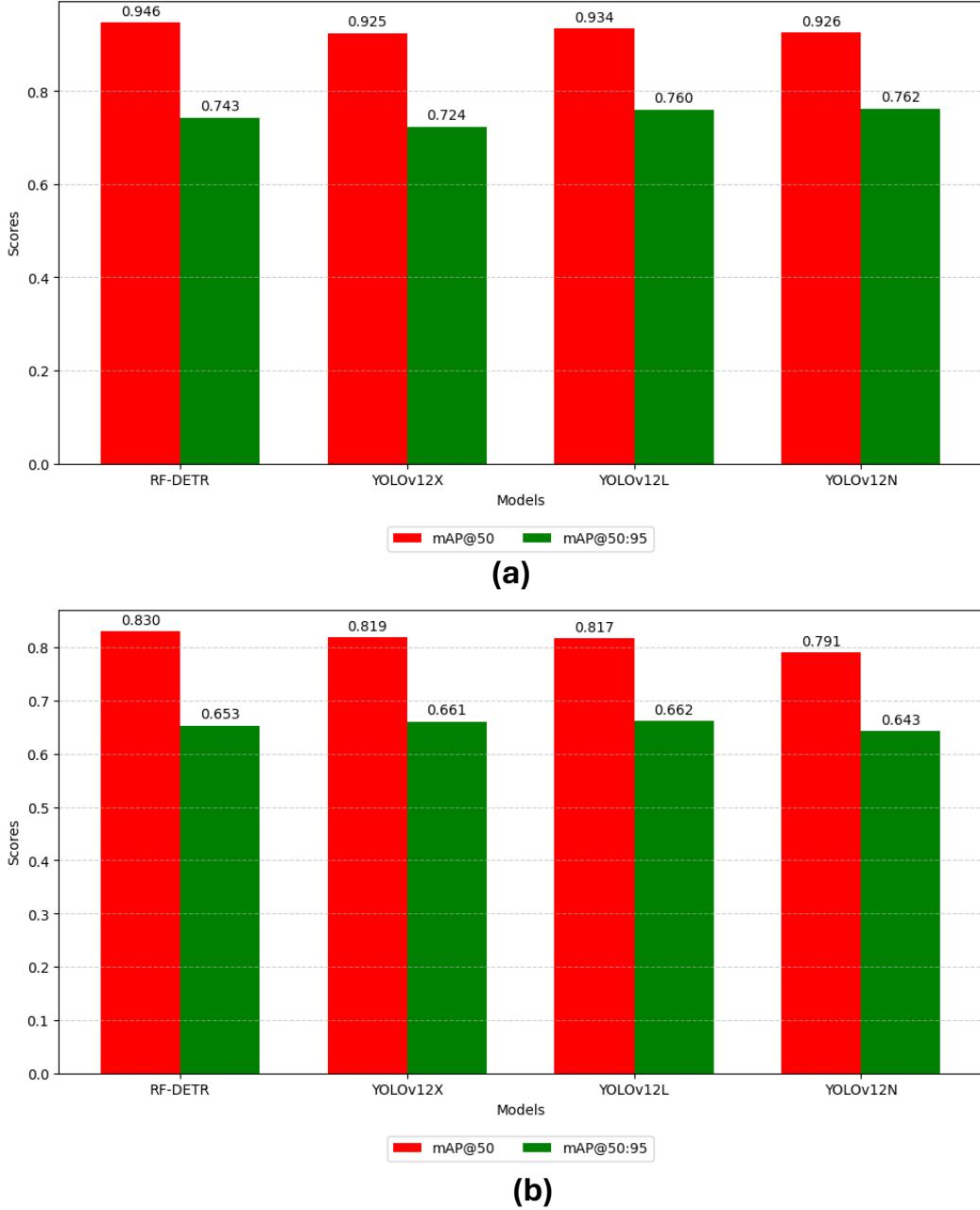


Figure 7: **Mean Average Precision (mAP) comparison for greenfruit detection using RF-DETR and YOLOv12 object detection models:** a) mAP@50 for single-class detection. b) mAP@50 and mAP@50:95 for multi-class detection

and multi-class settings is emblematic of the inherent strengths of transformer technology in handling dynamic and visually cluttered environments efficiently.

These following five observations highlight several key advantages of employing transformer-based models like RF-DETR in object detection tasks:

1. **Accelerated Learning Curve:** RF-DETR’s ability to reach peak performance quickly reduces the computational resources and time required for training, enhancing productivity and reducing operational costs.
2. **Stable Performance:** The model maintains consistent ac-

curacy over time, indicating robustness against overfitting and the ability to generalize well from limited epoch training.

3. **Adaptability:** RF-DETR’s architecture is evidently well-suited for complex detection environments, such as those in precision agriculture, where occlusions and varying object appearances prevail.
4. **Efficient Resource Utilization:** By converging quickly, RF-DETR maximizes the utility of computational resources, allowing for more tasks to be performed in the same computational budget.
5. **Edge Deployment:** The model’s quick adaptation and sta-

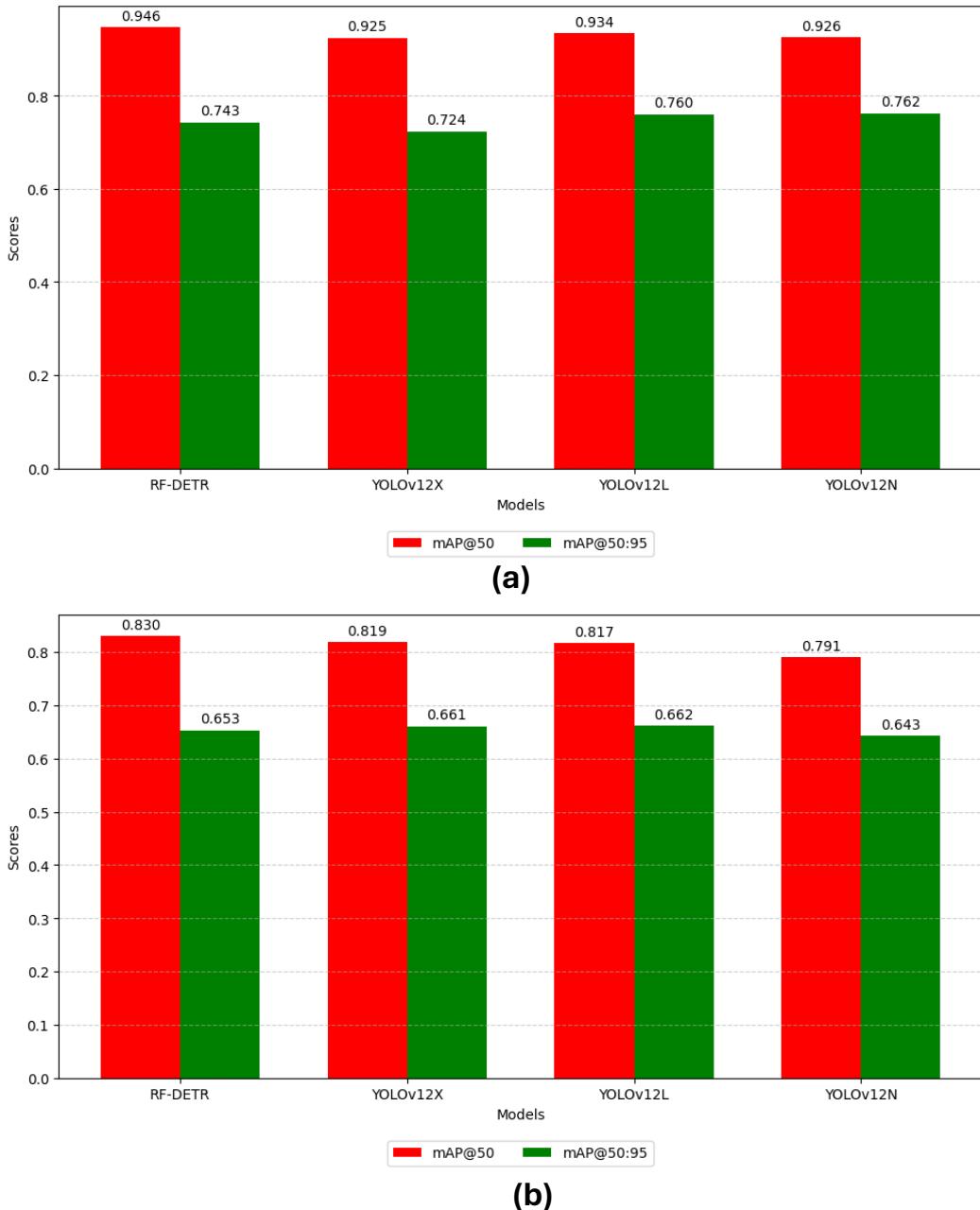


图7 使用RF-DETR与YOLOv12目标检测模型进行青果检测的平均精度均值(mAP)对比: a) mAP@50 {v\*} 或  
单c) 多类别检测。b) 多类别检测的mAP@50和mAP@50:95

多类别设置体现了Transformer技术在高效处理动态和视觉杂乱环境中的固有优势。

以下五点观察凸显了在目标检测任务中采用基于Transformer的模型（如RF-DETR）的若干关键优势：

1. 加速学习曲线：RF-DETR能够快速达到峰值性能，减少了训练所需的计算资源和时间，从而提升生产力并降低运营成本。
2. 稳定性能：该模型始终保持一致的准

随时间推移的准确性，表明其对抗过拟合的稳健性以及从有限周期训练中良好泛化的能力。

3. 适应性：RF-DETR的架构显然非常适合复杂的检测环境，例如精准农业中普遍存在的遮挡和物体外观多变的情况。
4. 高效资源利用：通过快速收敛，RF-DETR最大限度地提升了计算资源的利用率，使得在相同的计算预算下能够执行更多任务。
5. 边缘部署：模型的快速适应与稳-

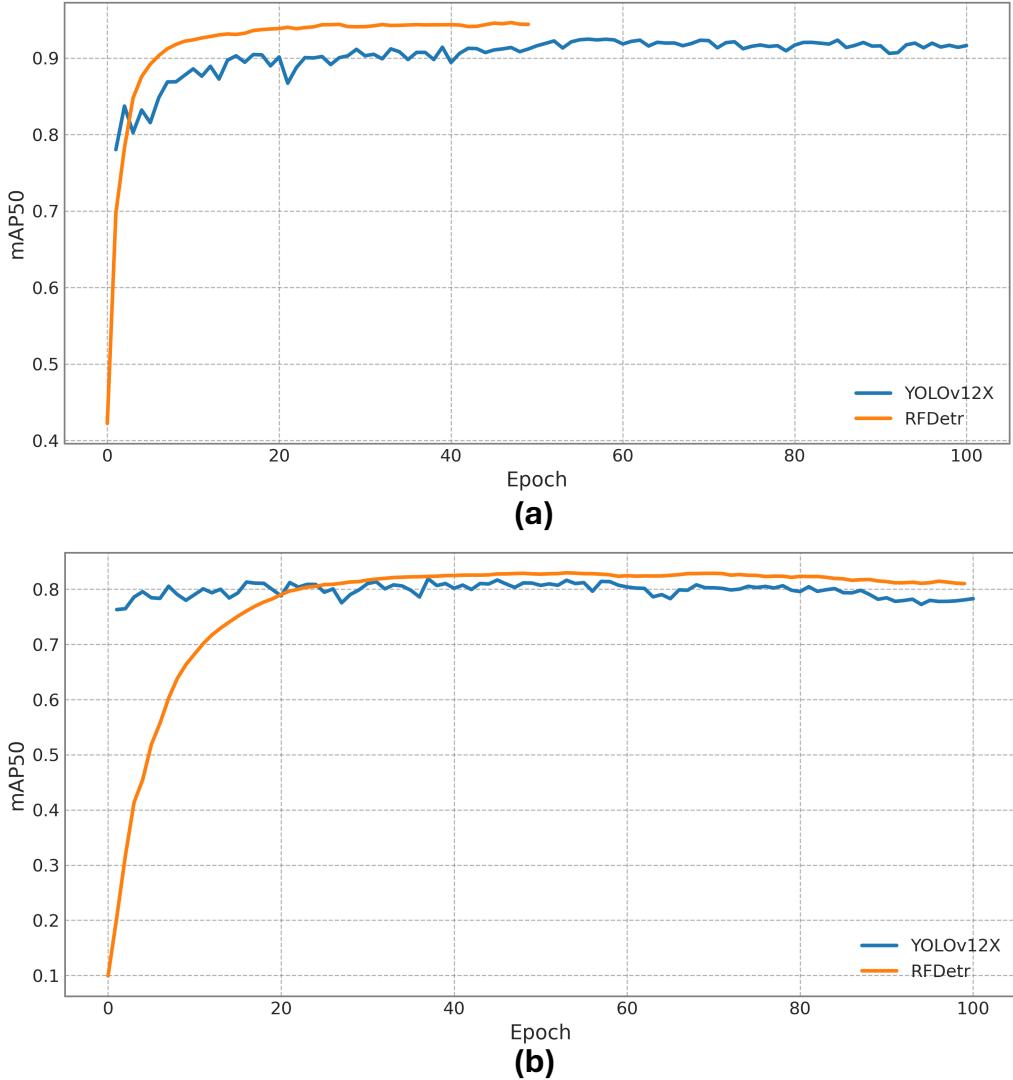


Figure 8: Training Dynamics and Model Convergence Analysis: mAP@50 vs. Epoch Curves for Object Detection Models. (a) Single-class greenfruit detection showing the performance trajectory of RF-DETR and YOLOv12X models over training epochs. (b) Multi-class greenfruit detection comparing the convergence patterns of both models throughout the training period

ble performance make it ideal for deployment in edge devices where computational resources and power are limited.

#### 4. Discussion

The progression of greenfruit detection technologies is closely aligned with recent advancements in computer vision, where each new model iteration introduces more nuanced capabilities, especially in complex agricultural settings. Notable contributions include those by [31, 49], which provided a comparative analysis of YOLOv11 and YOLOv8, focusing on their efficacy in segmenting occluded and non-occluded immature green fruits. Similarly, [50] explored size estimation techniques using YOLOv8 combined with geometric shape fitting on 3D point cloud data, aiming to enhance yield predictions and crop management decisions. These studies underscore the ongoing

efforts to refine the accuracy and efficiency of detection systems in variable orchard environments.

In this context, our study leverages the RF-DETR model, which has set new benchmarks in detection performance. RF-DETR's transformer-based architecture has achieved a mAP@50 of 0.9464, surpassing YOLOv12 and demonstrating superior spatial detection accuracy, particularly under conditions of partial visibility and camouflage [51]. This model's efficiency is further highlighted by its rapid convergence during training, indicating a significant advancement over traditional CNN-based models.

The integration of Vision-Language Models (VLMs) and open-vocabulary detection also represents a pivotal shift towards more dynamic and adaptable detection systems. These technologies, as reviewed in [52, 53], allow for the identification of a broader range of fruit types and characteristics without retraining. This adaptability is crucial for managing

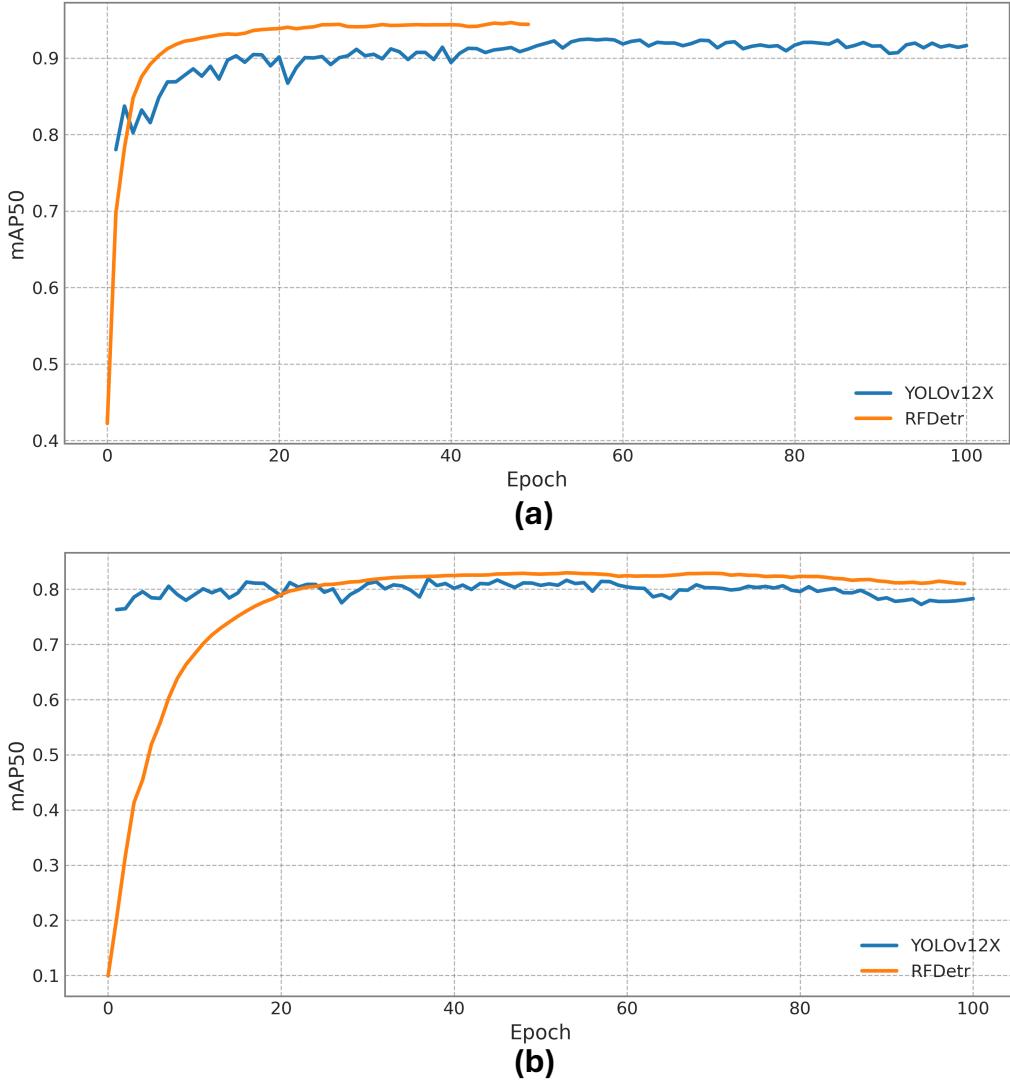


图8：训练动态与模型收敛性分析：目标检测模型的mAP@50随训练轮次变化曲线。(a) 单类青果检测展示了RF-DETR与YOLOv12X模型在训练周期内的性能轨迹。(b) 多类青果检测对比了两模型在整个训练期间的收敛模式

出色的性能使其非常适合部署在计算资源和电力有限的边缘设备中。

#### 4. 讨论

绿色果实检测技术的进步与计算机视觉领域的最新发展紧密相连，其中每个新模型迭代都引入了更精细的能力，尤其是在复杂的农业环境中。值得注意的贡献包括[31,49]的研究，它们对YOLOv11和YOLOv8进行了比较分析，重点关注这些模型在分割被遮挡与未遮挡未成熟绿色果实方面的效能。同样地，[50]探索了利用YOLOv8结合三维点云数据的几何形状拟合进行尺寸估算的技术，旨在提升产量预测和作物管理决策的准确性。这些研究凸显了持续不断的

在多变果园环境中提升检测系统准确性和效率的努力。

在此背景下，我们的研究采用了RF-DETR模型，该模型在检测性能上树立了新的标杆。基于Transformer架构的RF-DETR实现了0.9464的mAP@50，超越了YOLOv12，展现出卓越的空间检测精度，尤其在部分可见及伪装条件下表现突出[51]。该模型在训练过程中快速收敛的特性，进一步凸显了其高效性，标志着相对于传统CNN模型的重大进步。

视觉语言模型（VLMs）与开放词汇检测的融合，标志着检测系统向更动态、更适应性的方向迈出了关键一步。如[52,53]所述，这些技术无需重新训练即可识别更广泛的水果类型及特征，这种适应性对于管理至关重要。

the diverse conditions typical of agricultural settings, where fruit appearances and environmental factors such as lighting and occlusion vary considerably. the application of multimodal learning approaches that incorporate various sensory data types promises to resolve longstanding challenges such as camouflage and label ambiguity. The exploration of semi-supervised and few-shot learning paradigms could reduce reliance on extensive labeled datasets, facilitating quicker adaptation to new orchard environments [54]. Furthermore, the deployment of lightweight transformer variants and efficient VLMs for real-time field applications will be instrumental. These advancements will enable the development of mobile or edge-based systems that offer real-time analytics, crucial for immediate agricultural decision-making [55]. Continued advancements in these areas will undoubtedly forge detection systems that are not only highly accurate but also capable of semantic and contextual understanding. Such systems will drive the next wave of innovations in precision agriculture, ensuring that detection technologies are not only effective but also robust and adaptable to the complex dynamics of natural orchard environments.

## 5. Conclusion

This study provided an in-depth evaluation of RF-DETR (Transformer-based) and YOLOv12 (CNN-based) object detection models for detecting greenfruits in commercial orchards with complex visual environments. The research process included gathering real-world images, preparing datasets with occlusion-based labels for both single-class and multi-class detection, and assessing the models under standardized conditions. The comparison was based on precision, recall, F1-score, and mean average precision (mAP@50 and mAP@50:95). The analysis extended to training dynamics, revealing that RF-DETR demonstrated a quicker convergence, achieving stable performance in fewer epochs compared to YOLOv12. This insight highlights RF-DETR's effectiveness in adapting to the variable conditions of orchard environments while maintaining accuracy across extensive training phases.

**Key Findings:**

- **Single-Class Detection:** RF-DETR object detection model showcased superior performance with the highest mAP@50 of 0.9464, effectively localizing and detecting greenfruits amidst complex backgrounds. While YOLOv12N achieved the highest mAP@50:95 of 0.7620, RF-DETR remained consistently more accurate in cluttered and occluded scenarios.
- **Multi-Class Detection:** RF-DETR excelled in distinguishing between occluded and non-occluded fruits, registering the highest mAP@50 of 0.8298. YOLOv12L performed marginally better in mAP@50:95 with 0.6622, showcasing enhanced classification accuracy in detailed occlusion conditions.
- **Model Training Dynamics and Convergence:** The RF-DETR object detection model was notable for its rapid

training convergence, particularly in single-class scenarios where it plateaued at under 10 epochs, demonstrating both the efficiency and robustness of transformer-based architectures in handling dynamic visual data.

## Acknowledgment

This research is funded by the National Science Foundation and the United States Department of Agriculture, National Institute of Food and Agriculture through the “AI Institute for Agriculture” Program (Award No.AWD003473). We extend our heartfelt gratitude to Zhichao Meng, Astrid Wimmer, Randall Cason, Diego Lopez, Giulio Diracca for the data preparation; Martin Churuvija and Priyanka Upadhyaya for their invaluable efforts in logistical support during the data collection throughout this project. Special thanks to Dave Allan for granting commercial orchard access for this research experiment.

## Declarations

The authors declare no conflicts of interest.

## References

- [1] M. Masmoudi, H. Ghazzai, M. Frika, Y. Massoud, Object detection learning techniques for autonomous vehicle applications, in: 2019 IEEE international conference on vehicular electronics and safety (ICVES), IEEE, 2019, pp. 1–5.
- [2] M. Hnewa, H. Radha, Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques, IEEE Signal Processing Magazine 38 (2020) 53–67.
- [3] R. Elakkia, V. Subramaniyaswamy, V. Vijayakumar, A. Mahanti, Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks, IEEE Journal of Biomedical and Health Informatics 26 (2021) 1464–1471.
- [4] P. K. Mishra, G. Saroha, A study on video surveillance system for object detection and tracking, in: 2016 3rd international conference on computing for sustainable global development (INDIACom), IEEE, 2016, pp. 221–226.
- [5] C. M. Badgugar, A. Poulose, H. Gan, Agricultural object detection with you only look once (yolo) algorithm: A bibliometric and systematic literature review, Computers and Electronics in Agriculture 223 (2024) 109090.
- [6] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, C. McCool, Deepfruits: A fruit detection system using deep neural networks, sensors 16 (2016) 1222.
- [7] P. Singh, R. Krishnamurthi, Iot-based real-time object detection system for crop protection and agriculture field security, Journal of Real-Time Image Processing 21 (2024) 106.
- [8] L. Yang, T. Noguchi, Y. Hoshino, Development of a pumpkin fruits pick-and-place robot using an rgb-d camera and a yolo based object detection ai model, Computers and Electronics in Agriculture 227 (2024) 109625.
- [9] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroud, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern recognition 77 (2018) 354–377.
- [10] D. Nimma, Z. Zhou, Intelpvt: intelligent patch-based pyramid vision transformers for object detection and classification, International Journal of Machine Learning and Cybernetics 15 (2024) 1767–1778.
- [11] H. Liu, Y. Zhan, J. Sun, Q. Mao, T. Wu, A transformer-based model with feature compensation and local information enhancement for end-to-end pest detection, Computers and Electronics in Agriculture 231 (2025) 109920.
- [12] Y. Zang, W. Li, J. Han, K. Zhou, C. C. Loy, Contextual object detection with multimodal large language models, International Journal of Computer Vision 133 (2025) 825–843.

农业环境的典型多样性条件，其中果实外观及光照、遮挡等环境因素差异显著。采用融合多种感官数据类型的多模态学习方法，有望解决诸如伪装和标签模糊性等长期存在的难题。探索半监督和小样本学习范式可减少对大量标注数据集的依赖，从而更快适应新的果园环境[54]。此外，部署轻量级Transformer变体和高效视觉语言模型(VLMs)以实现实时田间应用将至关重要。这些进步将推动开发能提供实时分析的移动或边缘系统，这对即时农业决策至关重要[55]。这些领域的持续发展必将打造出不仅精度极高、还能进行语义和上下文理解的检测系统。此类系统将引领精准农业的新一轮创新浪潮，确保检测技术不仅高效，而且足够稳健，能适应自然果园环境的复杂动态变化。

## 5. 结论

本研究对基于Transformer的RF-DETR和基于CNN的YOLOv12两种目标检测模型在复杂视觉环境下的商业果园绿果检测进行了深入评估。研究过程包括采集真实场景图像、为单类别及多类别检测任务准备带有遮挡标注的数据集，并在标准化条件下评估模型性能。对比指标涵盖精确率、召回率、F1分数以及平均精度均值（mAP@50和mAP@50:95）。研究进一步分析了训练动态特征，发现RF-DETR展现出更快的收敛速度，在较少的训练周期内即可达到稳定性能，这一优势突显了RF-DETR在适应果园环境多变条件的同时，仍能在长期训练阶段保持准确性的卓越效能。

主要发现：

- 单类别检测：RF-DETR目标检测模型展现了卓越性能，其mAP@50最高达到0.9464，在复杂背景中有效定位并检测出绿色果实。尽管YOLOv12N以0.7620的mAP@50:95位居榜首，但在杂乱和遮挡场景下，RF-DETR始终保持更精准的检测能力。
- 多类别检测：RF-DETR在区分遮挡与非遮挡果实方面表现卓越，其mAP@50最高达到0.8298。而YOLOv12L在mAP@50:95指标上略胜一筹，取得了0.6622的成绩，展现了在复杂遮挡条件下更优的分类精度。
- 模型训练动态与收敛性：RF-DETR目标检测模型因其快速

训练收敛性，特别是在单类别场景下，仅需不到10个周期即达到稳定状态，这充分展示了基于Transformer架构在处理动态视觉数据时的高效性与鲁棒性。

## 致谢

本研究由美国国家科学基金会和美国农业部国家食品与农业研究所通过“农业人工智能研究所”计划（资助编号AWD003473）资助。我们衷心感谢孟志超、Astrid Wimmer、Randall Cason、Diego Lopez和Giulio Diracca在数据准备方面的工作；感谢Martin Churuvija和Priyanka Upadhyay在本项目数据收集过程中提供的宝贵后勤支持。特别感谢Dave Allan为本研究实验开放商业果园的准入权限。

## 声明

作者声明无利益冲突。

## 参考文献

- [1] M. Masmoudi, H. Ghazzai, M. Frikha, Y. Massoud, 面向自动驾驶应用的物体检测学习技术, 见: 2019年IEEE国际车辆电子与安全会议(ICVES), IEEE, 2019年, 第1–5页.
- [2] M. Hnewa, H. Radha, 自动驾驶车辆在雨天条件下的物体检测: 最新技术与新兴方法综述, IEEE信号处理杂志 38 (2020年) 53–67.
- [3] R. Elakkia, V. Subramaniyaswamy, V. Vijayakumar, A. Mabantu, 基于混合物体检测对抗网络的宫颈癌诊断医疗系统, IEEE生物医学与健康信息学杂志 26 (2021年) 1464–1471.
- [4] P. K. Mishra, G. Saroha, 视频监控系统中物体检测与跟踪的研究, 见: 2016年第三届可持续发展全球计算国际会议(INDIACOM), IEEE, 2016年, 第221–226页.
- [5] C. M. Badgugar, A. Poulose, H. Gan, 基于YOLO算法的农业物体检测: 文献计量与系统综述, 农业计算机与电子 223 (2024年) 109090.
- [6] I. Sa, Z. Ge, F. Da youb, B. Upcroft, T. Perez, C. McCool, Deepfruits: 基于深度神经网络的水果检测系统, 传感器 16 (2016年) 1222.
- [7] P. Singh, R. Krishnamurthi, 基于物联网的实时物体检测系统用于作物保护与农田安全, 实时图像处理杂志 21 (2024年) 106.
- [8] L. Yang, T. Noguchi, Y. Hoshino, 基于RGB-D相机与YOLO物体检测AI模型的南瓜采摘机器人开发, 农业计算机与电子 227 (2024年) 109625.
- [9] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, 等, 卷积神经网络的最新进展, 模式识别 77 (2018年) 354–377.
- [10] D. Nimma, Z. Zhou, Intelpvt: 基于智能补丁的金字塔视觉变换器用于目标检测与分类, 《国际机器学习与控制论期刊》15卷 (2024年) 1767–1778页。
- [11] H. Liu, Y. Zhan, J. Sun, Q. Mao, T. Wu, 一种基于Transformer的特征补偿与局部信息增强端到端害虫检测模型, 《农业中的计算机与电子》231卷 (2025年) 109920.
- [12] Y. Zang, W. Li, J. Han, K. Zhou, C. C. Loy, 基于多模态大语言模型的上下文目标检测, 国际计算机视觉杂志 133 (2025) 825–843.

- [13] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, W.-S. Zheng, Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models, arXiv preprint arXiv:2501.18954 (2025).
- [14] C.-Y. Fu, M. Shvets, A. C. Berg, Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free, arXiv preprint arXiv:1901.03353 (2019).
- [15] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
- [16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
- [17] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, M. Karkee, Comprehensive performance evaluation of yolo11, yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments, arXiv preprint arXiv:2407.12040 (2024).
- [18] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [19] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang, Dynamic detr: End-to-end object detection with dynamic attention, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2988–2997.
- [20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020).
- [21] J. Hosang, R. Benenson, B. Schiele, Learning non-maximum suppression, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4507–4515.
- [22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.
- [23] X. Ren, D. Ramanan, Histograms of sparse codes for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3246–3253.
- [24] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, Y. Zhang, Discriminative object tracking via sparse representation and online dictionary learning, IEEE transactions on cybernetics 44 (2013) 539–553.
- [25] K. O’shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).
- [26] D. Soydナー, Attention mechanism in neural networks: where it comes and where it goes, Neural Computing and Applications 34 (2022) 13371–13385.
- [27] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, U. Farooq, A survey of the vision transformers and their cnn-transformer based variants, Artificial Intelligence Review 56 (2023) 2917–2970.
- [28] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, cnn architectures, challenges, applications, future directions, Journal of big Data 8 (2021) 1–74.
- [29] S. Chen, Y. Liu, X. Gao, Z. Han, Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, in: Chinese conference on biometric recognition, Springer, 2018, pp. 428–438.
- [30] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
- [31] R. Sapkota, M. Karkee, Comparing yolo11 and yolov8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment, arXiv preprint arXiv:2410.19869 (2024).
- [32] Y. Tian, Q. Ye, D. Doermann, Yolov12: Attention-centric real-time object detectors, arXiv preprint arXiv:2502.12524 (2025).
- [33] R. Sapkota, R. Qureshi, M. F. Calero, C. Badjugar, U. Nepal, A. Poulose, P. Zeno, U. B. P. Vaddevolu, S. Khan, M. Shoman, et al., Yolov10 to its genesis: a decadal and comprehensive review of the you only look once (yolo) series, arXiv preprint arXiv:2406.19407 (2024).
- [34] R. Sapkota, M. Karkee, Improved yolov12 with llm-generated synthetic data for enhanced apple detection and benchmarking against yolo11 and yolov10, arXiv preprint arXiv:2503.00057 (2025).
- [35] Z. Meng, X. Du, R. Sapkota, Z. Ma, H. Cheng, Yolov10-pose and yolov9-pose: Real-time strawberry stalk pose detection models, Computers in Industry 165 (2025) 104231.
- [36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.
- [37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
- [38] D. Wang, Z. Li, X. Du, Z. Ma, X. Liu, Farmland obstacle detection from the perspective of uavs based on non-local deformable detr, Agriculture 12 (2022) 1983.
- [39] H. Lin, J. Liu, X. Li, L. Wei, Y. Liu, B. Han, Z. Wu, Dcea: Detr with concentrated deformable attention for end-to-end ship detection in sar images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).
- [40] Z. Zong, G. Song, Y. Liu, Detrs with collaborative hybrid assignments training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 6748–6758.
- [41] Y. Zhang, Y. Wu, H. Xu, Y. Xie, Y. Zhang, Improved co-detr with dropkey and its application to hot work detection, Concurrency and Computation: Practice and Experience 37 (2025) e70020.
- [42] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, W. Liu, You only look at one sequence: Rethinking transformer in vision through object detection, Advances in Neural Information Processing Systems 34 (2021) 26183–26197.
- [43] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detra beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 16965–16974.
- [44] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, et al., Simple open-vocabulary object detection, in: European conference on computer vision, Springer, 2022, pp. 728–755.
- [45] G. Heigold, M. Minderer, A. Gritsenko, A. Bewley, D. Keysers, M. Lučić, F. Yu, T. Kipf, Video owl-vit: Temporally-consistent open-world localization in video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13802–13811.
- [46] B. Wang, K. Huang, B. Li, Y. Yan, L. Zhang, H. Lu, Y. He, Effowl: Transfer visual language models to open-world tracking efficiently and effectively, arXiv preprint arXiv:2504.05141 (2025).
- [47] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, arXiv preprint arXiv:2203.03605 (2022).
- [48] P. Robicheaux, M. Popov, A. Madan, I. Robinson, J. Nelson, D. Ramanan, N. Peri, Roboflow100-vl: A multi-domain object detection benchmark for vision-language models, Roboflow (2025).
- [49] R. Sapkota, A. Paudel, M. Karkee, Zero-shot automatic annotation and instance segmentation using llm-generated datasets: Eliminating field imaging and manual annotation for deep learning model development, arXiv preprint arXiv:2411.11285 (2024).
- [50] R. Sapkota, D. Ahmed, M. Churuvija, M. Karkee, Immature green apple detection and sizing in commercial orchards using yolov8 and shape fitting techniques, IEEE Access 12 (2024) 43436–43452.
- [51] R. Sapkota, M. Karkee, Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning, arXiv preprint arXiv:2410.19846 (2024).
- [52] Q. Liu, H. Meng, R. Zhao, X. Ma, T. Zhang, W. Jia, Green apple detector based on optimized deformable detection transformer, Agriculture 15 (2024) 75.
- [53] R. Sapkota, S. Raza, M. Shoman, A. Paudel, M. Karkee, Multimodal large language models for image, text, and speech data augmentation: A survey, arXiv preprint arXiv:2501.18648 (2025).
- [54] Q. Liu, J. Lv, C. Zhang, Mae-yolov8-based small object detection of green crisp plum in real complex orchard environments, Computers and Electronics in Agriculture 226 (2024) 109458.
- [55] J. Lv, Z. Wu, P. Zhou, J. Huang, G. Liu, Y. Gu, H. Rong, L. Zou, Fcae-yolov8n: a target detection method for immature grape clusters, New Zealand Journal of Crop and Horticultural Science (2024) 1–19.

[13] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, W.-S. Zheng, LMDet：在大语言模型监督下学习强大的开放词汇目标检测器，arXiv预印本 arXiv:2501.18954 (2025)。 [14] C.-Y. Fu, M. Shvets, A. C. Berg, Retin aMask：通过学习预测掩码免费提升最先进单阶段检测性能，arXiv预印本 arXiv:1901.03353 (2019)。 [15] M. Tan, R. Pang, Q. V. Le, EfficientDet：可扩展高效的目标检测，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第10781–10790页。 [16] J. Redmon, S. Divvala, R. G irshick, A. Farhadi, YOLO：统一实时目标检测，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第779–788页。 [17] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, M. Karkee, YOLOv11、YOLOv10、YO LOv9与YOLOv8在复杂果园环境中幼果检测与计数的综合性能评估，arXiv预印本 arXiv:2407.12040 (2024)。 [18] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN，载于《IEEE国际计算机视觉会议论文集》，2017年，第2961–2969页。 [19] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L . Zhang, Dynamic DETR：基于动态注意力的端到端目标检测，载于《IE EE/CVF国际计算机视觉会议论文集》，2021年，第2988–2997页。 [20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR：端到端目标检测的可变形Transformer，arXiv预印本 arXiv:2010.04159 (2020)。 [21] J. Hosang, R. Benenson, B. Schiele, 学习非极大值抑制，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第4507–4515页。 [22] N. Ca rion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, 基于Tran sformer的端到端目标检测，载于《欧洲计算机视觉会议》，Springer, 2020年，第213–229页。 [23] X. Ren, D. Ramanan, 基于稀疏编码直方图的目标检测，载于《IEEE计算机视觉与模式识别会议论文集》，2013年，第3246–3253页。 [24] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, Y. Zhang, 通过稀疏表示与在线字典学习的判别性目标跟踪，《IEEE控制论汇刊》44 (2013) 539–553。 [25] K. O’ shea, R. Nash, 卷积神经网络导论，arXiv预印本 arXiv:1511.08458 (2015)。 [26] D. Soydancer, 神经网络中的注意力机制：起源与发展，《神经计算与应用》34 (2022) 13371–13385。 [27] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, U. Farooq, Vision Tra nsformer及其CNN-Transformer混合变体综述，《人工智能评论》56 (202 3) 2917–2970。 [28] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamar'a, M. A. Fadhel, M. Al-Amidie, L. Farhan, 深度学习综述：概念、CNN架构、挑战、应用与未来方向，《大数据期刊》8 (2021) 1–74。 [29] S. Chen, Y. Liu, X. Gao, Z. Han, MobileFaceNets：移动设备上高效精准的实时人脸验证CNN，载于《中国生物特征识别会议》，Springer, 2018年，第428–438页。 [30] R. Girshick, J. Donahue, T. Darrell, J. Malik, 用于精确目标检测与语义分割的丰富特征层次结构，载于《IEEE计算机视觉与模式识别会议论文集》，2014年，第580–587页。 [31] R. Sapkota, M. Karkee, 对比YOLOv11与YOLOv8在复杂果园环境中遮挡与非遮挡未成熟绿色果实的实例分割表现，arXiv预印本 arXiv:2 410.19869 (2024)。 [32] Y. Tian, Q. Ye, D. Doermann, YOLOv12：以注意力为核心的实时目标检测器，arXiv预印本 arXiv:2502.12524 (2025)。 [33] R. Sapkota, R. Qureshi, M. F. Calero, C. Badjugar, U. Nepal, A. Poulose, P. Zeno, U. B. P. Vaddevolu, S. Khan, M. Shoman等，YOLOv10溯源：YOLO系列十年全面回顾，arXiv预印本 arXiv:2406.19407 (2024)。 [34] R. Sapkota, M. Karkee, 基于LLM生成合成数据改进的YOLOv12增强苹果检测及与YOLOv11、YOLOv10的基准测试，arXiv预印本 arXiv:2503.00057 (202 5)。 [35] Z. Meng, X. Du, R. Sapkota, Z. Ma, H. Cheng, YOLOv10-Pose与YOLOv9-

姿态：实时草莓茎秆姿态检测模型，《工业计算机》165卷 (2025年) 104231页。 [36] 刘伟、D. Anguelov、D. Erhan、C. Szegedy、S. Reed、傅佳伟、A. C. Berg, SSD：单次多框检测器，载于《计算机视觉-EC CV 2016：第14届欧洲会议论文集，第一部分》，荷兰阿姆斯特丹，2016年10月11-14日，Springer出版社，2016年，第21–37页。 [37] 林泰宇、P. Goyal、R. Girshick、K. He、P. Dollár, 密集目标检测的焦点损失，《IE EE国际计算机视觉会议论文集》，2017年，第2980–2988页。 [38] 王栋、李政、杜晓、马志、刘旭，基于非局部可变形DETR的无人机视角农田障碍物检测，《农业》12卷 (2022年) 1983页。 [39] 林浩、刘杰、李翔、魏磊、刘洋、韩冰、吴钊，DCEA：基于集中可变形注意力的端到端S AR图像舰船检测DETR，《IEEE应用地球观测与遥感精选期刊》 (2024年)。 [40] 宗政、宋歌、刘洋，协作混合分配训练的DETRs，《IEEE/C VF国际计算机视觉会议论文集》，2023年，第6748–6758页。 [41] 张毅、吴越、徐浩、谢宇、张阳，采用Dropkey改进的Co-DETR及其在热作业检测中的应用，《并行计算：实践与经验》37卷 (2025年) e70020。 [42] 方毅、廖博、王旭、方杰、齐健、吴锐、牛杰、刘伟，你只需看一个序列：通过目标检测重新思考视觉中的Transformer，《神经信息处理系统进展》34卷 (2021年) 第26183–26197页。 [43] 赵毅、吕伟、徐松、魏杰、王刚、党强、刘洋、陈杰，实时目标检测中DETRs超越YOLOs，《IE EE/CVF计算机视觉与模式识别会议论文集》，2024年，第16965–16974页。 [44] M. Minderer、A. Gritsenko、A. Stone等，简单开放词汇目标检测，《欧洲计算机视觉会议》，Springer出版社，2022年，第728–755页。 [45] G. Heigold、M. Minderer等，Video OWL-ViT：视频中时间一致性的开放世界定位，《IEEE/CVF国际计算机视觉会议论文集》，2023年，第13802–13811页。 [46] 王博、黄凯、李斌等，EffOWT：高效迁移视觉语言模型至开放世界跟踪，arXiv预印本 arXiv:2504.05141 (2025年)。 [47] 张浩、李飞、刘松等，DINO：改进去噪锚框的端到端目标检测DETR，arXiv预印本 arXiv:2203.03605 (2022年)。 [48] P. Robicheaux等，Robo Flow100-VL：视觉语言模型多领域目标检测基准，RoboFlow (2025年)。 [49] R. Sapkota等，基于LLM生成数据集的零样本自动标注与实例分割：消除深度学习模型开发中的实地成像与人工标注，arXiv预印本 arXiv:2 411.11285 (2024年)。 [50] R. Sapkota等，商用果园中基于YOLOv8与形状拟合技术的未成熟青苹果检测与尺寸测量，《IEEE Access》12卷 (20 24年) 第43436–43452页。 [51] R. Sapkota等，基于YOLOv11与视觉Transfo rmer的商用苹果园未成熟青果三维姿态估计（机器人蔬果应用），arXiv 预印本 arXiv:2410.19846 (2024年)。 [52] 刘强、孟浩等，基于优化可变形检测Transformer的青苹果检测器，《农业》15卷 (2024年) 75页。 [53] R. Sapkota等，多模态大语言模型在图像、文本及语音数据增强中的应用综述，arXiv预印本 arXiv:2501.18648 (2025年)。 [54] 刘强、吕杰等，基于MAE-YOLOv8的复杂果园环境下青脆李小目标检测，《农业计算机与电子技术》226卷 (2024年) 109458页。 [55] 吕杰、吴钊等，FCAE-Y OLOv8n：未成熟葡萄串目标检测方法，《新西兰作物与园艺科学杂志》 (2024年) 1–19页。