

中图分类号: TP391

单位代号: 10280

密 级: 公开

学 号: 23721460

上海大学



硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	基于自蒸馏与混合专家机制 的农业果实目标检测方法研究
--------	-------------------------------

作 者 黄志怡

学科专业 计算机应用技术

导 师 韩越兴

完成日期 二〇二六年四月

姓 名：黄志怡

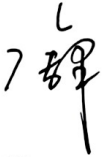
学号：23721460

论文题目：基于自蒸馏与混合专家机制的农业果实目标检测方法研究

上海大学

本论文经答辩委员会全体委员审查，确认
符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主 席： 
委 员：  陈 伟

导 师： 

答辩日期： 2026 年 6 月 9 日

姓 名: 黄志怡

学号: 23721460

论文题目: 基于自蒸馏与混合专家机制的农业果实目标检测方法研究

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下, 独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外, 论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 黄志怡

日期: 2026 年 6 月 9 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定, 即: 学校有权保留论文及送交论文复印件, 允许论文被查阅和借阅; 学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名: 黄志怡

导师签名: 韩继兴

日期: 2026 年 6 月 9 日

日期: 2026 年 6 月 9 日

上海大学工学硕士学位论文

基于自蒸馏与混合专家机制
的农业果实目标检测方法研究

作者: 黄志怡

导师: 韩越兴

学科专业: 计算机应用技术

计算机工程与科学学院

上海大学

2026 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in Engineering

**Research on Agricultural Fruit Object
Detection Method Based on
Self-Distillation and Mixture of
Experts**

Candidate: Zhiyi Huang

Supervisor: Yuexing Han

Major: Computer Application Technology

**School of Computer Engineering and Science
Shanghai University
April, 2026**

摘 要

随着智慧农业的发展,利用深度学习和计算机视觉技术实现果实成熟度自动检测,已成为现代农业研究的重要方向。然而,现有成熟度检测模型在复杂种植环境下仍面临诸多挑战:一方面,果园环境复杂,光照条件多变,果实尺寸差异显著,果实颜色与背景颜色相似,且枝叶遮挡严重,这些因素影响了现有目标检测方法在真实农业场景中的表现,导致对果实的定位以及成熟度检测不准确。另一方面,复杂农业环境造成特征多样性,同时成熟度类别之间存在相似性,现有的成熟度检测模型难以针对不同特征进行自适应处理,无法充分利用特征的语义信息,导致模型在复杂场景中泛化能力不足。此外,农业应用对实时性具有较高要求,高复杂度模型在实际场景中的应用受到限制。现有成熟度检测模型在复杂农业场景中难以兼顾准确性、高效性与鲁棒性。为了解决上述问题,本文以农业果实成熟度检测为研究对象,开展了系统性的研究工作。研究目标是提升复杂环境下果实目标检测的精度和稳定性,并探索端到端目标检测方法在农业场景中的应用潜力。主要开展了以下研究工作:

(1) 针对复杂环境下果实成熟度检测困难以及现有方法对多尺度信息的处理能力不足等挑战,提出了一种基于协同注意力与自蒸馏的目标检测模型 **FMD-DETR** (**Fruit Maturity Detection-DEtection TRansformer**)。在该模型中提出多尺度特征融合重加权模块,通过加强多尺度特征融合能力,以增强模型对不同尺度果实的检测能力;设计协同注意力解码器,结合交叉注意力和自注意力的优势,优化目标查询的交互方式,提升对遮挡目标的检测精度;同时在不额外增加推理计算量的前提下,提出层级自蒸馏策略,实现解码器内部不同层之间的知识传递,增强目标定位能力,进一步提高检测的准确性。实验结果表明, **FMD-DETR** 在番茄数据集上的平均精度为 75.4%,在真实场景番茄数据集上的平均精度为 52.7%,在草莓数据集上的平均精度为 41.5%,在 **FruitRipeness** 数据集上的平均精度为 87.1%。实验结果表明该模型在复杂环境下的检测精度和稳定性均优于现有方法,为果实自动化生产与采收提供了有效解决方案。

(2) 针对农业果实成熟度检测中特征多样性与模型自适应能力不足的问题,提出了一种基于动态稀疏混合专家机制的目标检测模型 **AgriMoE-DETR** (**Agricultural**

Mixture-of-Experts-DEtection TRansformer)。该模型通过高频细节增强模块，将浅层细节与深层语义相结合，提高对小目标特征的关注能力；在解码阶段构建了动态稀疏混合专家解码器，能够根据输入特征动态激活不同专家参与推理，增强模型的自适应能力；同时，提出负载均衡与专家多样性损失策略，促使不同专家学习到多样化的特征表示，提升了模型的泛化能力。实验结果表明，所提模型在番茄数据集、真实场景番茄数据集、草莓数据集以及 FruitRipeness 数据集上的平均精度分别为 76.1%、54.1%、42.1% 和 87.7%。相较于 FMD-DETR，该方法进一步提升了检测精度，增强了模型在复杂农业场景下的性能。

关键词：成熟度检测；目标检测；多尺度特征融合；层级自蒸馏；混合专家

ABSTRACT

With the development of smart agriculture, the use of deep learning and computer vision technologies to achieve automatic detection of fruit ripeness has become an important research direction in modern agricultural research. However, existing maturity detection models still face significant challenges in complex planting environments. On the one hand, orchard environments are complex, with diverse lighting conditions, variations in fruit size, fruit colors similar to the background, and severe occlusion by branches and leaves. These factors impact the performance of existing object detection methods in real-world agricultural environments, leading to inaccurate fruit localization and maturity detection. On the other hand, complex agricultural environments lead to feature diversity while inter-class similarities exist, existing maturity detection models cannot adaptively process different features and fail to fully leverage semantic information, resulting in insufficient generalization capabilities in complex environments. Furthermore, agricultural applications require high real-time performance, limiting the practical deployment of highly complex models. Existing maturity detection models face challenges in balancing accuracy, efficiency, and robustness in complex agricultural environments. To address these issues, this thesis focuses on agricultural fruit maturity detection and conducts a systematic study. The research aims to improve the accuracy and stability of fruit object detection in complex environments and to explore the potential applications of end-to-end object detection methods in agriculture. The main contributions are as follows:

(1) To address the challenges of detecting fruit maturity in complex environments and the limited ability of existing methods to handle multi-scale information, an object detection model FMD-DETR (Fruit Maturity Detection-DEtection TRansformer) is proposed, which is based on co-attention and self-distillation. The model proposes a multi-scale feature fusion re-weighting module, which is designed to enhance multi-scale feature integration, improving the detection accuracy of fruits with varying sizes. A co-attention decoder is designed which leverages the strengths of both cross-attention and self-attention to op-

timize the interaction between object queries, thereby improving the detection accuracy of occluded objects. In addition, without increasing the computational load of inference, a hierarchical self-distillation strategy is proposed for the decoder, which enables knowledge transfer across layers, enhancing object localization capability and further improving detection accuracy. Experiments show that the model achieves an average precision of 75.4% on the tomato dataset, 52.7% on the real-world tomato dataset, 41.5% on the strawberry dataset, and 87.1% on the FruitRipeness dataset. Experimental results demonstrate that the model surpasses existing methods in detection accuracy and robustness in complex environments, providing an effective solution for automated production and harvesting.

(2) To address the issue of diverse features in agricultural fruit maturity detection and insufficient adaptive capability of models, an object detection model AgriMoE-DETR (Agricultural Mixture-of-Experts-DEtection TRansformer) is proposed, which is based on a dynamic sparse mixture-of-experts mechanism. The method proposes a high-frequency detail enhancement module, which combines shallow-level details with deep-level semantics, designed to improve the ability to focus on small-object features. A dynamic sparse mixture-of-experts decoder is constructed in the decoding stage, which dynamically activates different experts for inference based on the features, thereby enhancing the model's adaptive capability. In addition, a load balancing and expert diversity loss strategy is proposed, which encourages different experts to learn diverse feature representations, improving the generalization ability of the model. Experiments show that the method achieves an average precision of 76.1%, 54.1%, 42.1%, and 87.7% on the tomato dataset, real-world tomato dataset, strawberry dataset, and FruitRipeness dataset, respectively. Experimental results demonstrate that compared with FMD-DETR, the method further improves detection accuracy and enhances model performance in complex agricultural environments.

Keywords: Maturity Detection; Object Detection; Multi-scale Feature Fusion; Hierarchical Self-Distillation; Mixture of Experts

目 录

摘 要	I
ABSTRACT	III
第一章 绪论	1
1.1 课题背景概述	1
1.2 课题研究目的与意义	2
1.3 国内外研究概况.....	3
1.3.1 农业果实成熟度检测研究现状	3
1.3.2 目标检测算法研究现状	4
1.4 论文主要工作	6
1.5 论文组织结构	8
第二章 相关理论和技术概述	9
2.1 目标检测算法	9
2.1.1 概述.....	9
2.1.2 Transformer	9
2.1.3 Detection Transformer	11
2.1.4 Real-Time Detection Transformer	12
2.1.5 评价指标	14
2.2 注意力机制	14
2.2.1 自注意力机制	15
2.2.2 多头自注意力机制	16
2.2.3 交叉注意力机制	17
2.3 多尺度特征融合.....	18
2.4 知识蒸馏	21
2.5 混合专家模型	23
2.6 本章小结	25

第三章 基于协同注意力和自蒸馏的农业果实目标检测方法	26
3.1 方法概述	26
3.1.1 模型框架	26
3.1.2 多尺度特征融合重加权模块	27
3.1.3 协同注意力解码器	29
3.1.4 层级自蒸馏策略	31
3.2 实验设计	33
3.2.1 实验环境与参数设置	33
3.2.2 定量实验	34
3.2.3 定性实验	39
3.2.4 消融实验	42
3.3 本章小结	45
第四章 基于动态稀疏混合专家机制的农业果实目标检测方法	47
4.1 方法概述	47
4.1.1 模型框架	47
4.1.2 高频细节增强模块	48
4.1.3 动态稀疏混合专家解码器	50
4.1.4 负载均衡与专家多样性损失策略	53
4.2 实验设计	54
4.2.1 实验环境与参数设置	54
4.2.2 定量实验	55
4.2.3 定性实验	60
4.2.4 消融实验	63
4.3 本章小结	68
第五章 总结与展望	69
5.1 总结	69
5.2 工作展望	70
参考文献	71

攻读硕士学位期间取得的研究成果	80
致 谢	81

第一章 绪论

1.1 课题背景概述

随着智慧农业与农业自动化技术的快速发展，对作物生长状态的精准监测已成为现代农业生产的重要基础，其中，准确判断果实成熟度是实现自动采摘^[1]、分级分选^[2]与产量评估^[3]的关键环节^[4]。果实在生长过程中，其颜色、质地与形态等外部特征随成熟阶段持续变化，这些变化直接反映了其生理状态与采收价值。然而，不同作物的成熟周期差异显著，同一果实不同光照条件、遮挡状态与生长环境下呈现出多样化的外观特征，使得机械设备难以在复杂自然环境中稳定、精确地完成成熟度判断。传统农业生产中，成熟度判断主要依赖人工经验，不仅效率低、成本高，而且具有较强的主观性，难以适应规模化与标准化生产需求。因此，构建一种稳定、可靠且可部署的果实成熟度自动检测技术，已成为智能农业领域亟需解决的重要问题^[5-6]。

随着深度学习与计算机视觉技术的快速发展，基于图像的果实成熟度检测逐渐成为农业信息化的重要技术路径。该任务本质上属于农业目标检测问题，其目标不仅是定位果实目标，还需进一步区分其成熟阶段，属于兼具目标检测与精细阶段识别需求的复杂视觉任务。与通用视觉场景相比，农业环境具有显著的复杂性与不确定性。自然光照随时间与天气变化，果实通常分布在背景复杂且外观相似的自然环境中，果实与叶片在颜色和纹理上具有较强相似性，并且常受到枝叶遮挡的影响；同一图像中果实尺度差异显著，需要同时兼顾大目标与小目标的准确检测；不同成熟阶段之间的外观差异往往连续而细微，类间边界模糊。这些因素使果实成熟度检测成为一项具有挑战性的任务，对视觉模型的特征表达能力、跨尺度特征学习能力以及在复杂环境中的鲁棒性提出了更高的要求。

早期研究主要依赖颜色、纹理和形态等人工设计特征对果实成熟状态进行分析^[7-8]，但此类方法对光照变化和背景干扰高度敏感，难以在真实果园环境中稳定工作^[9]。随着深度学习的发展，基于卷积神经网络的端到端检测框架逐渐成为主流，能够自动学习与检测任务相关的特征，从而显著提升性能^[10-11]。近年来，Transformer架构被应用于目标检测领域^[12-13]，其基于注意力机制的全局建模能力克服了传统局

部感受野的限制，为复杂场景下目标关系建模与多尺度表示提供了新的技术路径。

然而，现有模型在真实农业场景中检测果实成熟度时仍面临多重挑战：首先，现有模型捕捉多尺度特征的能力不足，难以准确检测小目标，且容易受到遮挡与背景干扰的影响；其次，果实成熟阶段的视觉差异细微，模型难以识别关键的区分特征；最后，农业应用对实时性具有较高要求，高复杂度模型在实际场景中的应用受到限制。

因此，当前果实成熟度检测研究亟需从模型结构层面探索新的技术路径：一方面，应强化多尺度特征之间的交互与融合机制，提升模型的表达能力，使模型能够在复杂场景中同时关注不同尺度的果实目标；另一方面，需要加强模型学习不同成熟阶段细微特征的能力，使其能够更加有效地捕捉颜色、纹理与形态变化等语义信息；同时，还需在保证检测准确性的前提下控制模型的复杂度，以提升其在真实农业环境中的运行效率。开发一种能在复杂农业场景中平衡准确性、效率与鲁棒性的成熟度检测框架已成为该领域的一个关键研究方向。

1.2 课题研究目的与意义

复杂农业环境下的果实成熟度检测面临多重挑战，包括光照变化、果实尺寸差异、遮挡以及背景干扰等问题，这些问题直接影响了现有目标检测方法在真实农业场景中的表现，容易导致小果实被遗漏，遮挡或重叠果实的定位不稳定，同时多尺度信息未能充分利用也会降低成熟度判断的准确性。为了克服上述困难，本研究的目的在于构建一种兼具准确性、鲁棒性与高效性的果实成熟度检测方法，以提升农业目标检测在复杂场景下的适应能力和应用价值。具体而言，提出一种面向复杂农业场景的果实成熟度检测框架，通过优化多尺度特征的融合与目标之间的交互，并在网络内部实现层间知识传递，增强模型对果实位置和成熟度的判定能力，有效提升检测精度和稳定性。同时，针对果实在视觉特征上的多样性及环境变化带来的挑战，构建动态稀疏混合专家网络，使模型能够根据输入特征动态选择不同专家进行处理，并结合负载均衡与专家多样性损失策略，提升模型在复杂场景检测中的表达能力和稳定性。

在理论层面，本研究聚焦于复杂农业环境下果实视觉特征的多样性与多尺度性，为目标检测技术在光照变化、遮挡和密集果实等挑战下的适应性提供理论参考，丰

富了农业目标检测的研究体系，并为未来相关技术发展提供基础性思路。在应用层面，所提出的方法面向果实成熟度自动识别这一农业生产中的关键环节，有助于降低对人工经验的依赖，为智能采摘、精准农业管理与自动化分级提供可靠的技术基础。该研究对于推动目标检测技术在农业场景中的落地应用，以及促进农业生产向智能化、精细化方向发展具有重要的现实意义。

1.3 国内外研究概况

1.3.1 农业果实成熟度检测研究现状

随着计算机视觉技术的快速发展，越来越多的研究将其应用于果实成熟度的检测与分析^[6,14-16]。在早期研究中，部分研究采用传统机器视觉技术检测果实成熟度，这类方法主要通过颜色、纹理和形状特征提取，结合传统分类器实现果实成熟度识别。Anindita 等人^[17]基于颜色和纹理特征提取特征，并通过人工神经网络（Artificial Neural Network, ANN）对油棕果进行分类以获得预测类别。Tu 等人^[18]提出了一种百香果成熟度检测分级模型，该模型通过提取颜色特征并使用 SVM 分类器，达到了 92.71% 的成熟度检测准确率。Guo 等人^[19]通过结合两个光谱范围的超光谱成像与支持向量机（Support Vector Machine, SVM）评估草莓成熟度，该方法通过提取果实的光谱特征来捕捉成熟度相关的细微变化，能够在实验条件下实现较高的精度。Suban 等人^[20]运用 K 近邻算法（K-Nearest Neighbor, KNN）判定木瓜成熟度，通过计算样本间的特征相似性来完成分级，展示了基于距离度量的分类方法在果实成熟识别中的可行性。Guo 等人^[21]则采用随机森林算法检测香蕉成熟度，并将其与 SVM、KNN 等其他机器学习算法进行对比。然而，大多数基于传统机器视觉技术的果实成熟度检测方法主要依赖于人工提取的低维视觉特征（颜色、纹理、形状等），这种方法的特征表达能力有限，难以描述果实表面细微的颜色渐变、局部纹理变化等成熟度关键信息，并且特征设计依赖先验知识，由于果实种植环境的复杂性，果实检测常受限于光照变化、枝叶遮挡、果实重叠等因素，这些挑战使成熟度检测任务尤为困难^[22]。

为解决传统方法在复杂环境下的局限性，近年来基于深度学习的多种解决方案应运而生。研究人员提出改进版 Mask R-CNN 方法，用于温室环境中不同成熟度番茄的分割与检测^[23]。Ko 等人^[24]提出基于多个卷积神经网络与随机决策融合（Stochastic Decision Fusion, SDF）技术的番茄成熟度检测方法。Li 等人^[25]通过在 YOLOv8 模型

中集成混合注意力变换器模块与四头自适应空间特征融合检测头，有效提升了柑橘成熟度检测精度。为适应实际农业应用对实时性和轻量化的要求，Zhang 等人^[26]采用轻量级 MobileNetV2 模型实现葡萄成熟度判定。最新研究进一步将端到端目标检测网络 DETR (Detection Transformer)^[27]应用到农业目标检测领域，以提升复杂场景下果实识别的精度与效率。Tang 等人^[28]提出的 ML-DETR 在结合 YOLOv8 与 RT-DETR 的基础上，通过多尺度自注意力与跨阶段特征增强机制，实现了成熟樱桃番茄的高精度实时检测，为农业果实自动识别提供了有效方法。相比于传统机器视觉方法，深度学习方法能够通过数据驱动的方式自动学习图像中的特征表示，无需人工手动设计颜色、纹理或形状等特征描述，因此在复杂环境下具有更强的适应能力。这类方法在提升识别精度与结果一致性的同时，降低了对人工干预的依赖，使其在多样化和复杂场景中的应用更加稳定且高效。

尽管现有成熟度检测方法已取得一定进展，但直接将当前成熟度检测模型或通用目标检测框架应用于农业成熟度检测仍面临挑战。首先，在农业成熟度检测中，同一类别目标在不同生长阶段、光照条件、天气环境、背景以及遮挡情况下会呈现出极大的外观差异，例如青涩与成熟果实在颜色和形态上差别明显，而叶片遮挡又进一步增加了检测难度。其次，农业目标具有显著的尺度变化，不同拍摄角度下的图像往往同时包含从大目标到小目标的多层级检测需求。复杂而相似的背景也是一大挑战，果实在颜色与纹理上与叶片接近，边界模糊，易造成误检与漏检。现有成熟度检测模型结构相对简单^[23]，对细微成熟度差异的特征捕捉能力有限，难以处理农业目标尺度变化大、遮挡严重等复杂场景问题。通用目标检测模型虽然在自然场景下表现优异^[27,29]，但是由于农业场景中存在高度的类间相似性与类内多样性，加之光照条件、遮挡及背景干扰的影响，这些模型在农业环境中的表现往往欠佳。因此，亟需设计一种在复杂农业环境下具有较高准确性、效率和鲁棒性的成熟度检测框架。

1.3.2 目标检测算法研究现状

目标检测 (Object Detection) 作为计算机视觉领域的核心任务，旨在通过自动化算法识别图像或视频序列中感兴趣的目标，并精确给出目标的类别标签及其所在的边界框坐标。随着深度学习技术的飞速发展，目标检测算法已从基于手工特征的传统方法全面转向基于深度神经网络的范式。目前，主流的目标检测算法根据其架构逻辑和演进路径，主要可分为基于卷积神经网络 (CNN) 的检测方法和基于 Transformer

的检测方法两大类^[30]。

基于卷积神经网络（CNN）的目标检测算法在过去十余年中占据主导地位，根据检测流程的阶段性的，通常可分为两阶段（Two-stage）和单阶段（One-stage）两类^[31]。两阶段目标检测算法以 R-CNN 系列为代表，其核心思想是“先生成候选区域，再对候选区域进行分类与边界框回归”。R-CNN^[32] 首次将卷积神经网络应用于目标检测，通过选择性搜索生成候选区域，并利用 CNN 提取特征、采用 SVM 完成分类，突破了传统手工特征方法的性能瓶颈。Fast R-CNN^[33] 针对 R-CNN 中重复卷积计算带来的效率问题，提出感兴趣区域池化（RoI Pooling）机制，实现整幅图像的特征共享，改善了训练与推理效率。Faster R-CNN^[34] 进一步提出区域提议网络（Region Proposal Network, RPN），利用神经网络自主生成候选框以替代低效的传统算法，使候选区域生成与目标检测过程统一于同一网络框架之中，提高了整体效率，并确立了以锚框为基础的 Anchor-based 检测范式。

为满足实时应用需求，单阶段检测算法省略了候选区域生成步骤，直接在特征图上预测目标类别与边界框位置。YOLO（You Only Look Once）系列^[29] 将目标检测建模为一个统一的回归问题，通过将图像划分为网格并直接预测边界框，实现了极高的推理帧率，在实时目标检测场景中得到了广泛应用。YOLOv8^[35]、YOLOv10^[36] 和 YOLO11^[37] 是近年来 YOLO 系列演进的代表性版本。YOLOv8 结合了无锚框分离式检测头，达到了准确性和速度平衡，具有较好的实际应用潜力。YOLOv10 解决了先前 YOLO 版本中存在的后处理和模型架构缺陷，通过消除非极大值抑制（NMS）并优化模型结构，有效降低了计算开销。YOLO11 采用了改进的骨干网络和颈部架构，增强了特征提取能力，从而实现更精确的目标检测。YOLO12^[38] 提出了以注意力机制为中心的架构，优化整体网络架构，实现了高目标检测准确性，但由于注意力模块较为复杂，可能带来训练不稳定以及内存消耗增加等问题。因此，目前 YOLO11 更适用于实际应用场景。SSD（Single Shot MultiBox Detector）^[39] 通过在不同尺度的特征图上设置多种尺度与长宽比的默认框，实现多尺度目标的并行检测，有效增强了对不同尺寸目标，尤其是小目标的检测能力。RetinaNet^[40] 针对单阶段检测器中正负样本极度不平衡的问题，提出了焦点损失函数（Focal Loss），从而在保持较高检测效率的同时提升了检测精度。FCOS（Fully Convolutional One-Stage Object Detection）^[41] 提出了一种全新的无锚框检测方法，摒弃了传统的预设锚框设计，通过将检测建模为逐像素的预测任务，并结合特征金字塔网络提高多尺度目标预测。

近年来, 基于 Transformer 的端到端目标检测方法逐渐成为研究热点。Carion 等人^[27] 首次提出 DETR (Detection Transformer), 利用 Transformer 编码器-解码器结构直接预测目标类别和边界框集合, 无需候选区域或锚框, 并通过二分匹配将预测结果与真实目标一一对应, 从而简化检测流程并消除非极大值抑制 (NMS) 带来的性能瓶颈。尽管 DETR 具有显著优势, 仍面临两大挑战: 训练收敛缓慢与查询优化困难。为解决这些问题, 已提出多种 DETR 变体方案。Deformable DETR^[42] 通过可变形注意力增强小目标和局部特征捕捉能力, 提高收敛速度; Conditional DETR^[43] 提出条件查询, 显式建立查询向量与空间位置之间的关联, 缓解了查询向量在高维特征空间中定位目标的学习难度, 降低查询优化难度; DAB-DETR^[44] 提出使用动态的锚框查询, 将查询初始化为可学习的锚框输出到 Transformer 解码器并动态更新, 使得目标定位更加精确, 加速训练收敛; DN-DETR^[45] 通过采用去噪训练, 将带噪声的真实框作为辅助监督信号, 使模型更稳定地学习匹配关系, 加速模型收敛; Group-DETR^[46] 采用分组一对多匹配机制, 在每组内进行一对一分配, 加快了模型训练收敛速度, 并提升了各种基于 DETR 的模型的性能; DINO^[47] 通过对比去噪训练策略, 并采用更加高效的查询设计机制, 有效缓解了 DETR 类模型收敛缓慢与查询向量优化困难的问题, 从而提升了模型的训练效率与检测性能。RT-DETR^[12] 通过高效混合编码器和优化查询选择方法, 提高了检测精度, 在速度和精度方面均优于以往的先进 YOLO 模型。RT-DETRv2^[48] 在此基础上, 在可变形注意力模块内为不同尺度的特征设置不同数量的采样点, 使得解码器能够实现选择性多尺度特征提取, 同时优化了训练策略, 提高模型性能而不损失速度。

总而言之, 目标检测算法经历了从传统候选区域或锚框方法到端到端 Transformer 方法的系统演进。但基于卷积神经网络的方法通常依赖候选区域或锚框进行局部特征建模, 而基于 Transformer 的方法通过全局信息建模和端到端训练, 在复杂、多尺度和密集目标识别中表现较好, 但在小目标检测上表现不佳, 在实时目标检测中达到精度与效率之间的平衡仍然是一项具有挑战性的任务。

1.4 论文主要工作

为了解决果实生长环境复杂、果实与背景颜色相近、光照条件变化剧烈以及枝叶遮挡严重等问题, 本文以农业果实成熟度检测为研究对象, 开展了系统性的研究

工作。研究目标是提升复杂环境下果实目标检测的准确性和稳定性，并探索端到端目标检测方法在农业场景中的应用潜力。技术路线如图1.1所示。主要研究工作和创新点概括如下：

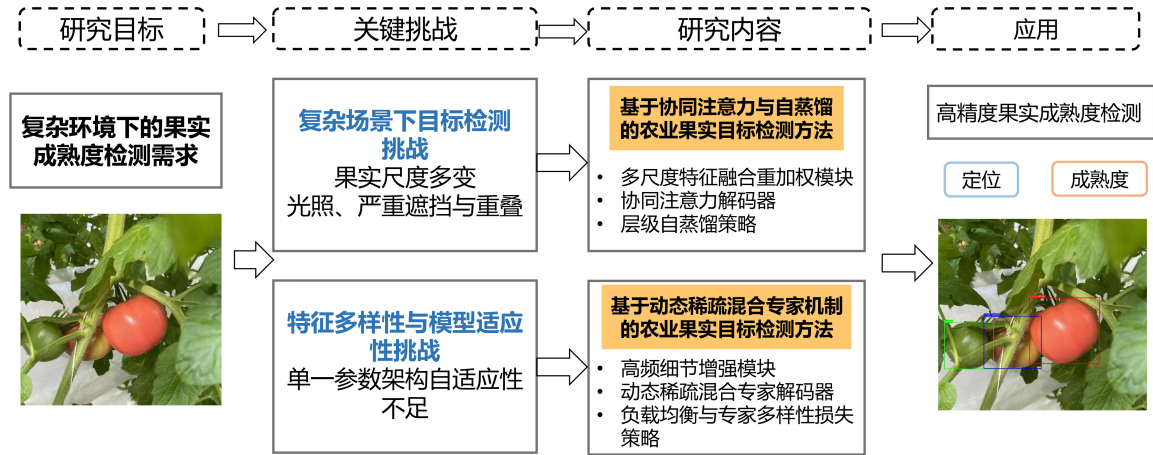


图 1.1 本研究的技术路线示意图

(1) 针对复杂农业场景下现有模型对多尺度、遮挡和密集果实定位与检测困难问题，提出一种基于协同注意力与自蒸馏的农业果实目标检测方法。具体而言，针对复杂场景下果实尺度变化大、小目标易漏检以及多尺度特征利用不足的问题，提出多尺度特征融合重加权模块，通过对不同尺度特征进行自适应重加权，提升模型对各类尺寸果实的检测能力。为进一步优化目标查询交互，构建协同注意力解码器，将交叉注意力与自注意力机制结合，增强解码器对遮挡和重叠目标的处理能力。此外，设计层级自蒸馏策略，实现解码器不同层之间的知识传递，提高空间感知与定位鲁棒性，从而在复杂果园场景中提升检测精度和稳定性。最终，在多个果实成熟度数据集上的实验结果验证了该方法在复杂农业环境下的有效性。

(2) 针对农业果实成熟度检测中特征多样性与模型自适应能力不足的问题，提出一种基于动态稀疏混合专家机制的农业果实目标检测方法。具体而言，针对农业目标检测中小目标定位难的问题，设计高频细节增强模块，将浅层细节与深层语义相结合，以弥补深层特征中细节的不足，从而提高对小目标的检测能力。针对复杂农业环境下光照、遮挡及果实尺度变化导致的特征差异问题，构建动态稀疏混合专家解码器，结合多尺度先验信息和稀疏激活策略，自适应地激活不同专家参与推理，增强模型的自适应能力。为缓解专家激活不均衡和表示同质化问题，设计负载均衡损失和专家多样性损失联合约束，使不同专家能够学习差异化的特征表示，从而进

一步提升模型对复杂农业场景的适应性。实验结果表明，该方法能够有效提升对不同成熟度状态果实的精准识别能力。

1.5 论文组织结构

本研究聚焦于复杂农业环境中果实成熟度检测任务面临的关键挑战，如尺度变化、遮挡严重、光照条件多变、小目标定位难以及模型自适应能力不足等问题，致力于提升复杂场景下果实成熟度检测的精度与稳定性，推动目标检测技术在农业领域的应用，促进智慧农业发展。全文共分为五章，论文结构安排如下：

第一章绪论主要介绍果实成熟度检测的研究背景和研究意义，阐述目标检测算法和果实成熟度检测的国内外研究现状，以及本文的研究内容和组织结构。

第二章阐述了本文所涉及的相关理论与技术基础，主要包括目标检测相关模型的基本概念与原理、本文采用的主要基线模型、注意力机制的基本原理、多尺度特征融合方法、知识蒸馏的概念与原理，以及混合专家模型的概念与原理。

第三章介绍了基于协同注意力与自蒸馏的农业果实目标检测方法。首先概述本文提出的果实成熟度检测框架，详细阐述提出的多尺度特征融合重加权模块、协同解码器以及层级自蒸馏策略，进而通过定量实验、定性实验以及消融实验，验证本文方法在农业目标检测任务中的可行性。

第四章介绍了基于动态稀疏混合专家机制的农业果实目标检测方法。首先概述面向农业果实成熟度检测的动态稀疏混合专家检测模型，随后介绍提出的高频细节增强模块、动态稀疏混合专家解码器以及负载均衡与专家多样性损失策略的实现，最后在农业果实数据集上通过实验验证所提方法的有效性。

第五章总结了全文工作，并对未来研究方向进行了展望。重点总结了提出的果实成熟度检测方法的创新点以及主要成果。此外，进一步分析了本文方法在农业目标检测领域存在的不足以及未来的研究方向。

第二章 相关理论和技术概述

本章主要阐述本文研究所涉及的相关理论与技术基础，包括目标检测模型的基本概念与原理、本文采用的基线模型、注意力机制、多尺度特征融合方法、知识蒸馏，以及混合专家模型的基本思想与实现原理。上述内容为后续模型构建与实验分析奠定了理论基础。

2.1 目标检测算法

2.1.1 概述

目标检测（Object Detection）是计算机视觉中的基础任务，其目标是在输入图像中同时完成目标类别识别与空间位置定位。具体而言，目标检测不仅需要判断图像中是否存在某一类别的目标，还需要给出该目标在图像中的位置，通常以边界框的形式表示。目标检测的核心任务可概括为两个方面：一是分类，即判断图像中是否包含目标物体并赋予正确的类别标签；二是定位，即通过边界框精确框出每个目标的位置。与传统分类任务不同，目标检测需同时处理图像中可能出现的多个物体、不同尺度以及重叠遮挡等复杂情况，因此具有更高的挑战性。目标检测通常包含三个基本组成部分：特征提取、定位与分类。其中，特征提取用于获取图像的特征表示，目标定位用于预测目标边界框位置，目标分类用于判断目标类别，三者共同完成目标检测任务。

2.1.2 Transformer

Transformer^[49]最初被提出用于自然语言处理任务，其核心思想是利用注意力机制对序列中的全局依赖关系进行建模。与传统循环神经网络（RNN）或卷积神经网络（CNN）不同，Transformer 不依赖固定的局部感受野或顺序传播结构，而是通过注意力机制在特征序列内部建立任意位置之间的关联，从而提升长距离依赖建模能力和特征表达能力。

Transformer 的基本结构包括编码器（Encoder）和解码器（Decoder）两部分，其结构如图2.1 所示。编码器由多头自注意力（Multi-Head Attention）模块与前馈神经

网络组成，用于对输入特征进行高维表示学习。解码器首先通过带掩码的多头自注意力机制对输入特征进行内部关系建模，然后通过多头注意力机制实现解码器特征与编码器输出特征之间的交互，增强模型对全局特征关系的建模能力。解码器输出特征经过前馈神经网络和线性层，得到最终的预测结果。

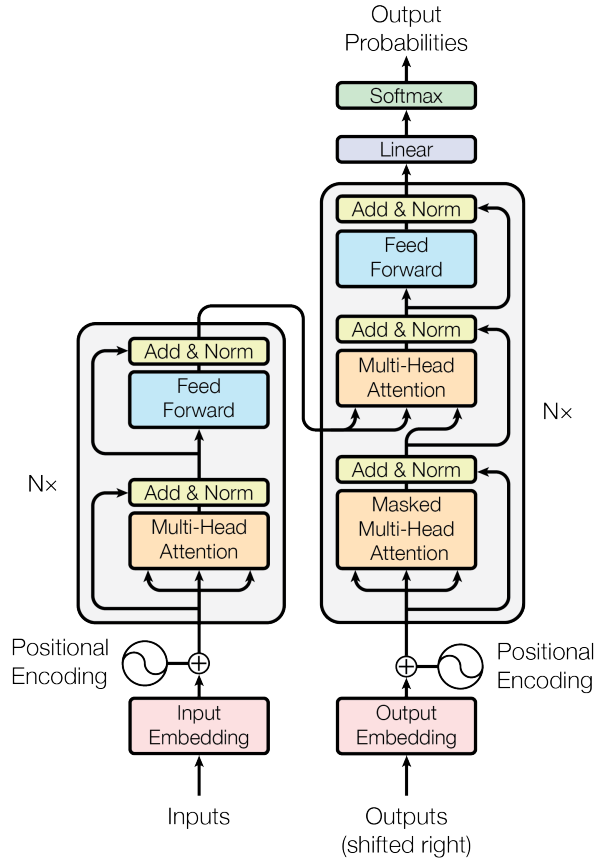


图 2.1 Transformer 模型结构图^[49]

在计算机视觉领域中，Transformer 逐渐被用于图像表示学习与视觉任务建模。Vision Transformer (ViT)^[50]将图像划分为固定大小的 Patch 序列，并将其视作类比文本词向量的输入序列，通过自注意力机制进行特征建模，如图2.2所示。相比卷积网络依赖局部卷积核逐层堆叠学习上下文信息的方式，Transformer 具备更强的全局建模能力，能够直接捕获远距离区域之间的依赖关系。后续研究提出了一系列经典的视觉 Transformer 变体，如 Swin Transformer^[51]和 Pyramid Vision Transformer^[52]等。

在目标检测任务中，Transformer 被成功应用到端到端检测框架中，例如 DETR 采用查询向量与图像特征之间的注意力交互机制完成目标预测，从而避免了锚框设计与候选区域生成等额外步骤。随着研究的不断推进，基于 Transformer 的检测模型

在特征表示、场景建模和结构一致性方面展现出优势。

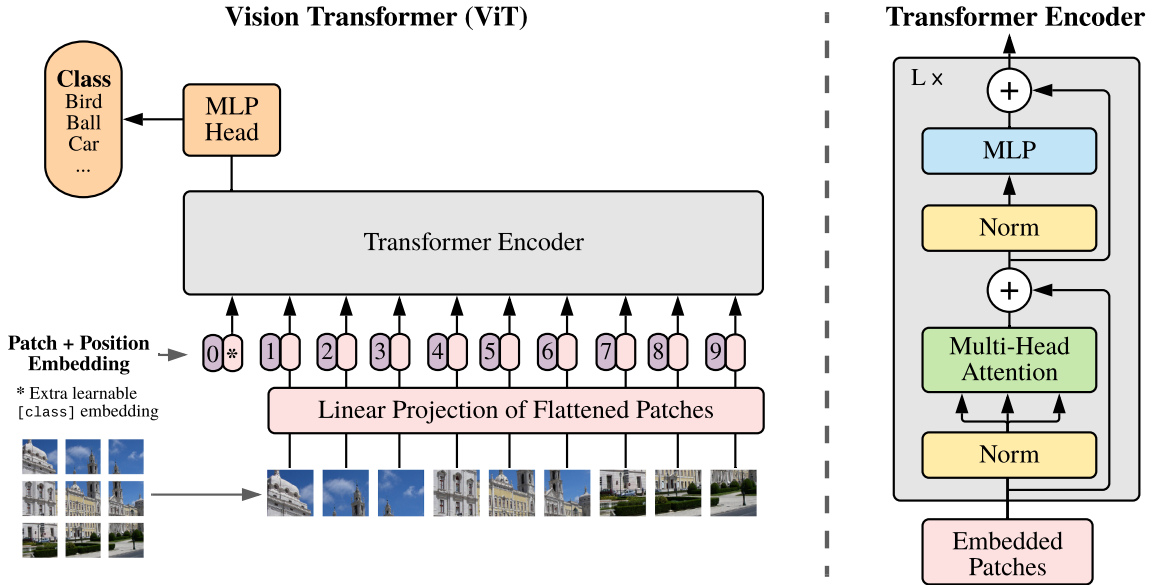


图 2.2 Vision Transformer 模型架构示意图^[50]

2.1.3 Detection Transformer

DETR (DEtection TRansformer)^[27]是首个基于 Transformer 架构的端到端目标检测器，DETR 消除了传统检测方法的锚框和非极大值抑制 (NMS) 组件，而是采用二分匹配将每个预测框与真实框进行匹配。通过这种策略，DETR 简化了检测流程并缓解了由 NMS 导致的性能瓶颈。DETR 的整体架构如图2.3所示。DETR 由一个 CNN 主干网络、一个 Transformer 编码器和一个 Transformer 解码器组成。

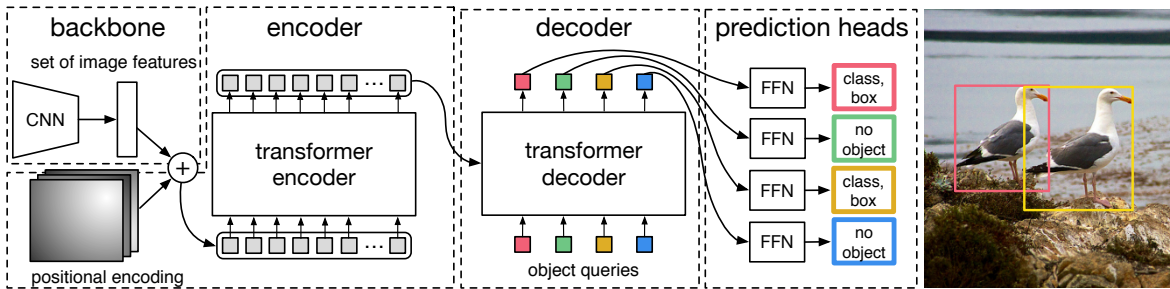


图 2.3 DETR 模型架构图^[27]

首先，DETR 使用 CNN 主干网络提取输入图像的特征图，并为其添加空间位置编码，然后将带有位置信息的特征输入编码器。编码器层的多头注意力机制通过捕捉特征图中不同位置之间的全局依赖关系来增强特征表示能力。解码器接收编码器

输出的特征序列和可学习的查询向量，通过解码器生成目标类别和边界框的预测结果。DETR 的训练过程中采用了匈牙利算法进行二分匹配，将每个预测框与真实框进行一一匹配，从而实现端到端的目标检测。DETR 的提出标志着基于 Transformer 的目标检测方法的兴起，尽管其在小目标检测和训练收敛速度方面存在挑战，但其简洁的架构和全局建模能力为后续研究提供了重要的方向。

2.1.4 Real-Time Detection Transformer

RT-DETR (Real-Time Detection Transformer)^[12] 是一种基于 Transformer 的端到端的实时目标检测算法。RT-DETR 的整体架构如图2.4所示。RT-DETR 由一个主干网络、一个高效的混合编码器和一个带有辅助预测头的 Transformer 解码器组成。首先将主干网络的最后三个阶段的特征输入到编码器中，高效的混合编码器通过基于注意力的尺度内特征交互 (AIFI) 和基于 CNN 的跨尺度特征融合 (CCFF) 将多尺度特征转化为一组图像特征。然后，不确定性最小查询选择模块选择固定数量的编码器特征作为解码器的初始对象查询。最后，带有辅助预测头的解码器迭代优化对象查询以生成类别和框。

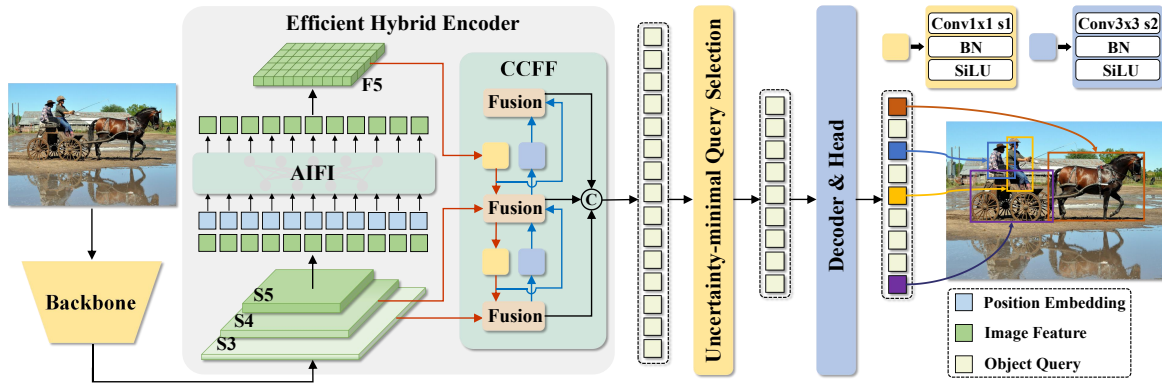


图 2.4 RT-DETR 模型架构图^[12]

RT-DETR 使用了主干网络输出的 S3、S4、S5 特征图，覆盖大、中、小不同尺度目标的特征信息。基于注意力的尺度内特征交互 (AIFI) 仅在最深层的 S5 特征图上应用 Transformer 编码器，过程如下：

$$Q = K = V = \text{Flatten}(F_5), \quad (2.1)$$

$$F'_5 = \text{Reshape}(\text{Attention}(Q, K, V)), \quad (2.2)$$

由于 S5 特征具有更深、更高级和更丰富的语义特征，因此仅计算 S5 特征图，而不是所有尺度的特征图，有助于显著减少计算量和提高计算速度，同时不会对模型的性能造成很明显的损害。

跨尺度融合模块（Cross-scale Feature Fusion, CCFF）是一种轻量级的基于卷积神经网络的结构，用于高效融合主干网络输出的不同尺度特征，它采用自上而下与自下而上相结合的双向融合方式完成跨尺度信息交互，先通过自上而下路径将高层特征逐级上采样并与相邻尺度特征对齐后进行融合以增强低层语义表达，再通过自下而上路径将融合后的低层特征逐级向高层回流并再次融合以补充细节与定位信息。在每次融合前后，CCFF 使用可重参数化卷积块（RepBlock）的轻量级卷积模块进行特征变换与增强。RepBlock 在推理时可通过结构重参数化技术合并为一个简单的卷积层，从而实现即插即用的零延迟提速。CCFF 中的融合模块如图2.5所示。经过 CCFF 处理后，得到一组语义增强且空间信息对齐的多尺度特征图，同时包含了深层的语义信息和浅层的细节信息。

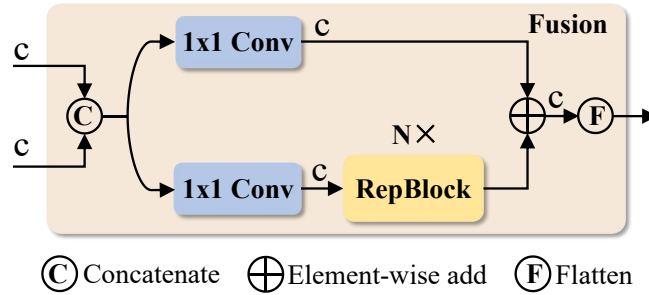


图 2.5 CCFF 中的融合模块示意图^[12]

在基于查询的检测框架中，查询的初始化质量会直接影响后续解码的难度。若初始查询来自定位置信度较低或噪声较大的特征位置，解码器需要更多层迭代才能将其修正为稳定的目标框，且更容易出现分类正确但定位不准确导致匹配困难等问题。RT-DETR 提出 IoU 感知查询选择，考虑分类置信度和 IoU 未必一致的问题，保留同时具备较高分类置信度与较高定位质量的查询，为后续解码器提供更高质量的输入特征表示。

RT-DETR 展示了在完全去除 NMS 的端到端检测范式下，通过针对实时推理的结构设计与优化，Transformer 检测器同样可以达到具有竞争力的实时速度，并取得优于主流实时检测器的速度和精度折中，从而为实时目标检测提供了一条新的端到

端技术路径。

2.1.5 评价指标

对于目标检测模型，通常使用两个主要指标来评估模型，即交并比（Intersection over Union, IoU）和平均精度（Average Precision, AP）^[53]。其中，交并比用于衡量预测边界框与真实边界框之间的重叠程度，其定义为预测框与真实框交集面积和并集面积之比：

$$\text{IoU} = \frac{\text{预测框} \cap \text{真实框}}{\text{预测框} \cup \text{真实框}} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}, \quad (2.3)$$

平均精度用于综合衡量模型在目标检测上的整体性能。对于每个类别，AP 分数主要通过绘制精确率-召回率（Precision-Recall, P-R）曲线下的面积来计算。精确率（Precision）为模型预测为“正样本”的所有结果中，有多少是真正的正样本。定义为：

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (2.4)$$

召回率（Recall）为所有真正的正样本中，有多少被模型成功检测出来了。定义为：

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (2.5)$$

其中， TP 表示真正例（True Positive），即被正确检测出的目标数量； FP 表示假正例（False Positive），即被错误检测为目标的数量； FN 表示假负例（False Negative），即未被检测出的真实目标数量。

$$\text{AP} = \sum_{k=1}^N (R_k - R_{k-1}) \cdot P(R_k), \quad (2.6)$$

其中， N 为检测结果总数， R_k 为前 k 个检测结果的召回率， $P(R_k)$ 为在召回率 R_k 处的插值精度。计算所有类别 AP 分数得到平均值 (mAP)，mAP 值越高，表明目标检测模型在所有目标类别上的整体性能越好。

2.2 注意力机制

在目标检测任务中，注意力机制旨在更有效地捕捉图像特征的全局上下文与关键局部信息，从而提升对目标的定位与识别性能。自注意力能够通过捕捉特征序列

内部的依赖关系，有效增强特征表示能力，而交叉注意力用于建模不同特征序列之间的交互，有效提取关键目标信息，因此自注意力机制和交叉注意力机制被广泛应用于 Transformer 及其变体中。

2.2.1 自注意力机制

自注意力机制（Self-Attention）是一种特殊形式的注意力机制，其核心思想是在输入序列内部通过元素间的相关性计算，为每个元素生成动态的上下文表示。自注意力直接对输入序列本身进行注意力权重分配，从而能够捕捉序列内部的全局依赖关系。该机制在 Transformer 模型中首次系统化提出，并已成为自然语言处理和计算机视觉任务的基础组件^[49]。

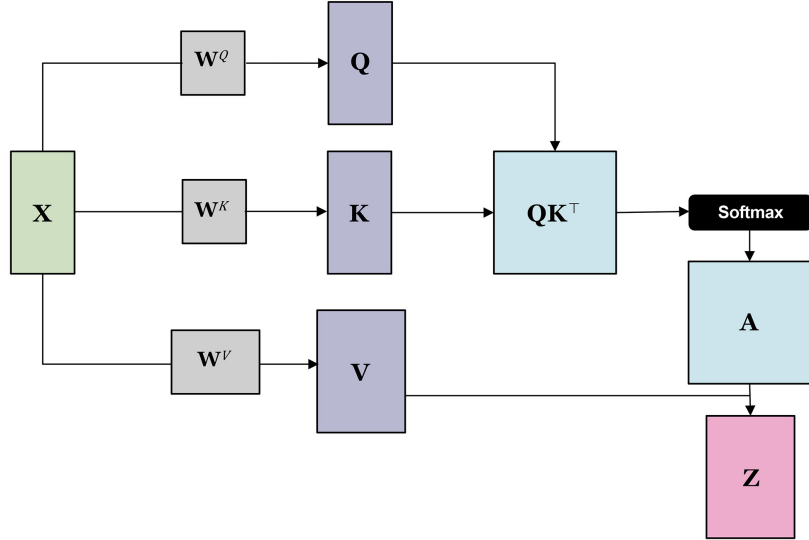


图 2.6 自注意力机制的计算流程示意图

在自注意力机制中，输入特征通过三个不同的线性投影分别映射为查询（Query）、键（Key）和值（Value）矩阵。如图2.6所示，设输入序列的特征表示为 $X \in \mathbb{R}^{n \times d}$ ，其中 n 为序列长度， d 为特征维度，则对应的线性变换可表示为：

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \quad (2.7)$$

其中， W^Q ， W^K ， W^V 分别为查询、键和值的线性映射权重矩阵。随后通过缩放点积计算注意力权重，并对值向量进行加权求和，得到注意力输出：

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (2.8)$$

其中, d_k 是键向量的维度, 缩放因子 $\sqrt{d_k}$ 用于稳定梯度和数值计算, softmax 操作用于将点积相似度归一化为注意力权重分布。该输出包含了输入序列中各位置之间的全局相关性信息, 从而实现了全局上下文的建模。

2.2.2 多头自注意力机制

多头自注意力 (Multi-Head Self-Attention) 机制是在自注意力的基础上进行扩展的一种机制, 其目的在于从多个不同的子空间并行学习和捕捉不同类型的特征关系, 从而增强模型的表示能力。与单一注意力头不同, 多头机制将输入特征分别投影到多个独立注意力空间, 每个注意力头都执行自注意力计算, 随后将各头的输出拼接并线性映射为最终输出。多头注意力的计算流程如图2.7所示。

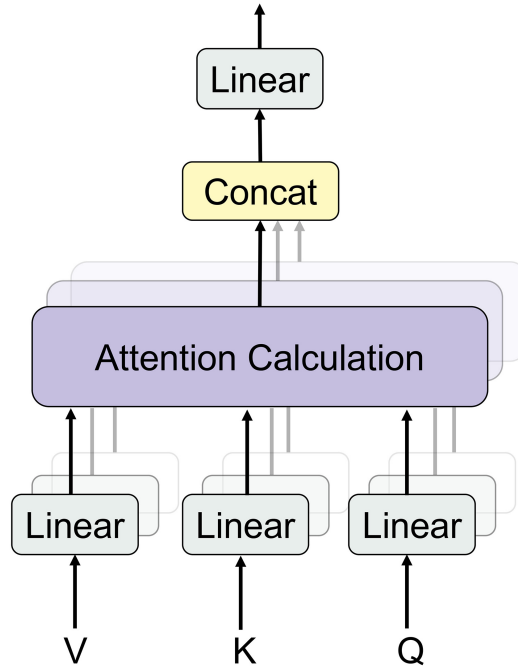


图 2.7 多头注意力的计算流程示意图

对于输入序列的特征表示 $X \in \mathbb{R}^{n \times d}$, 多头自注意力首先为每个头 $i = 1, \dots, h$ 分别计算查询、键和值:

$$Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V, \quad (2.9)$$

然后对每个头应用自注意力:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax} \left(\frac{Q_i K_i^\top}{\sqrt{d_k}} \right) V_i, \quad (2.10)$$

定义所有头的注意力输出拼接后再进行线性映射得到最终输出：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (2.11)$$

其中， W^O 为输出投影矩阵， h 表示注意力头的数量。多头机制允许模型在不同子空间中学习输入的不同依赖关系，从而提升对复杂结构的建模能力。

2.2.3 交叉注意力机制

交叉注意力机制（Cross-Attention）是在注意力框架下实现不同输入序列之间信息交互的一种机制，通常用于 Transformer 解码器中，使模型能够将一个序列的信息映射到另一个序列上。在自注意力中，查询（Query）、键（Key）和值（Value）均来自同一输入序列，而交叉注意力则将查询来自一个序列，而键和值来自另一个序列，从而实现序列间的关注与对齐，支持源序列与目标序列之间的上下文融合与信息检索。这一机制是 Transformer 解码器中连接编码器与解码器、实现序列间信息交互的核心模块。

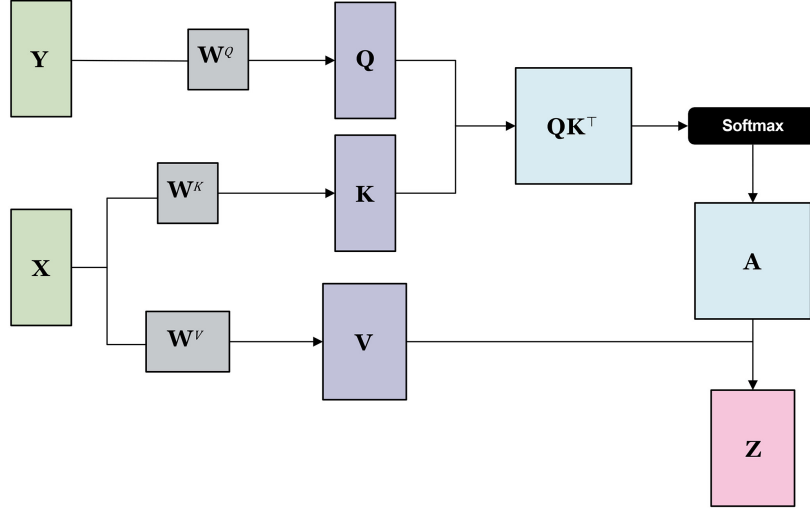


图 2.8 交叉注意力机制的计算流程示意图

如图2.8所示，设有两个不同输入序列 $X \in \mathbb{R}^{n \times d}$ 和 $Y \in \mathbb{R}^{m \times d}$ ，交叉注意力将 Y 映射为查询矩阵，将 X 映射为键矩阵和值矩阵：

$$Q = YW^Q, \quad K = XW^K, \quad V = XW^V, \quad (2.12)$$

然后进行缩放点积注意力计算：

$$\text{CrossAttention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (2.13)$$

其中， W^Q ， W^K ， W^V 分别为线性投影权重矩阵， $\sqrt{d_k}$ 为键向量维度的缩放因子。该机制能够动态分配权重，使目标序列在生成或预测时聚焦最相关的源序列信息。

2.3 多尺度特征融合

多尺度特征融合（Multi-scale Feature Fusion）是目标检测领域中的关键技术之一，多尺度特征融合旨在同时利用不同层级的互补信息，以克服单一尺度难以兼顾定位与语义识别的局限。浅层特征具有更高的空间分辨率，保留了丰富的边缘与细节信息，适合进行精确定位；而深层特征则拥有更多的语义信息，适合确定目标的类别与属性。通过构建跨尺度信息交互机制，实现语义信息与空间信息的融合，可以显著提升模型对多尺度目标的检测性能。

在此背景下，特征金字塔网络（Feature Pyramid Network, FPN）^[54] 成为多尺度特征融合的经典范式，通过自顶向下的路径与横向连接，将深层特征蕴含的强语义信息逐层传递至浅层高分辨率特征，提升目标检测能力。如图2.9所示，FPN 通过构建自顶向下的路径，将深层特征逐级上采样，并与对应尺度的浅层特征进行横向连接，使得不同尺度的特征都包含丰富的语义信息，显著增强了模型对多尺度目标的检测能力。

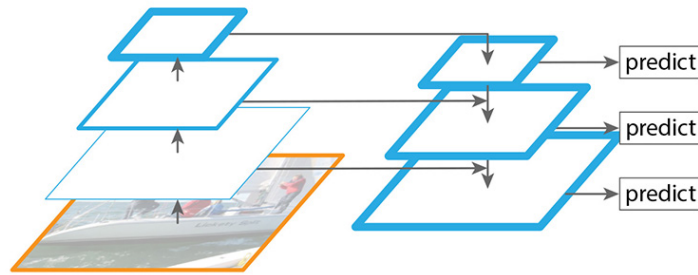


图 2.9 特征金字塔网络（FPN）结构示意图^[54]

路径聚合特征金字塔网络（Path Aggregation Feature Pyramid Network, PAFPN）^[55] 在 FPN 的基础上增加了自底向上的路径，使低层的精细定位信息能够更充分地传递至高层语义特征，缩短了低层到高层的信息路径，形成了双向的信

息流。该结构最初在实例分割网络 PAFNet^[55] 中作为其增强的特征金字塔（Augmented FPN）模块，这种双向融合机制不仅增强了不同尺度特征之间的信息交互，还提高了模型在检测物体位置和边缘时的准确性，在复杂场景和尺度变化显著的目标检测任务中表现出更优的鲁棒性。

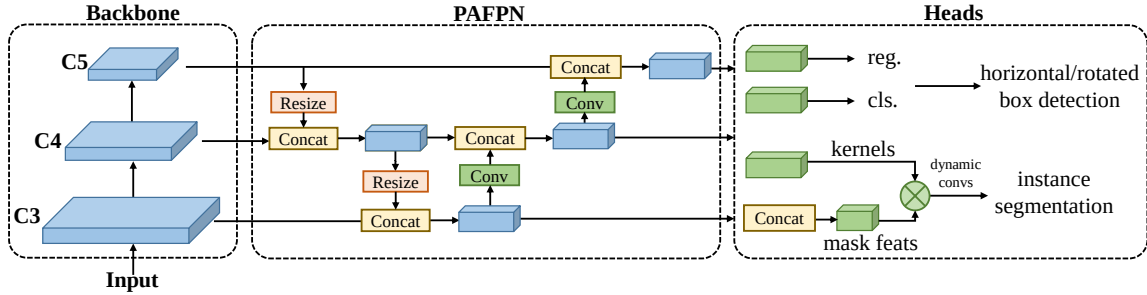
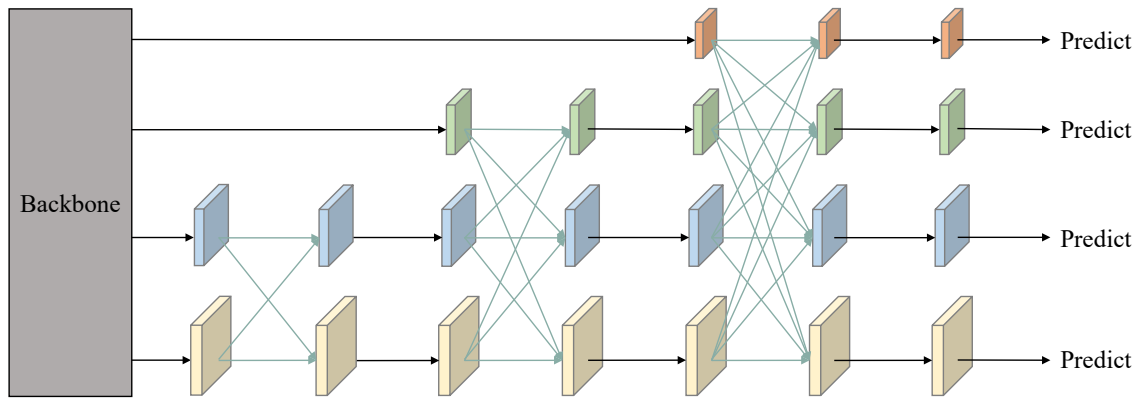


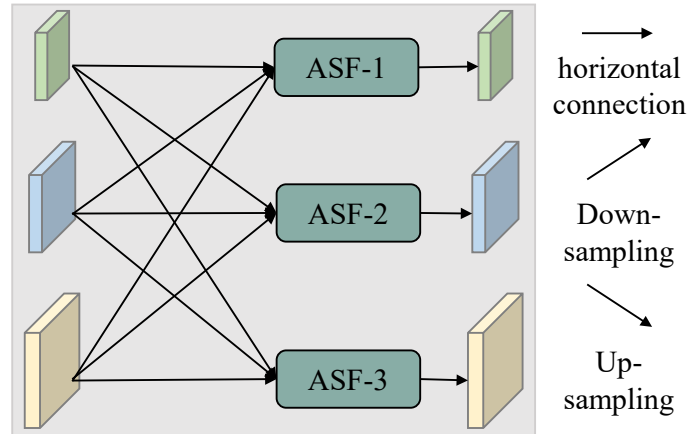
图 2.10 路径聚合特征金字塔网络（PAFPN）结构示意图^[55]

如图2.10所示，对主干网络输出的多尺度特征图（记为 C3、C4、C5），将高层语义特征 C5 上采样至与 C4 相同的空间分辨率，并与 C4 进行拼接，随后通过卷积操作得到融合特征。然后，再将该融合后的中间特征继续上采样至与 C3 相同的空间分辨率，并与 C3 进行拼接和卷积处理，得到融合后的低层特征。随后，PAFPN 的关键创新在于增加了第二条自底向上的融合路径，低层特征与上一层的中间特征再次进行拼接，并通过卷积操作生成新的高层融合特征。该过程实现浅层细节信息向高层语义特征的反向补充。通过该双向路径融合结构，不同层级特征之间能够实现更充分的信息交互，从而有效提升多尺度特征表达能力。

然而，FPN 的融合权重是固定的，无法根据图像内容动态分配不同层级特征的重要性，且缺乏对不同层次特征之间的信息矛盾有效抑制^[56-58]。为解决这一问题，渐进特征金字塔网络（Asymptotic Feature Pyramid Network, AFPN）^[59] 应运而生。如图2.11所示，在主干网络自下而上的特征提取过程中，AFPN 在初始阶段融合两个低级特征，后续阶段融合更高层次的特征，而最终阶段则将顶级特征加入到特征融合过程中。由于 AFPN 的架构是渐进的，这将使不同层级特征的语义信息在渐进融合过程中更加接近。

图 2.11 渐进特征金字塔网络（AFPN）结构示意图^[59]

在多级特征融合过程中，AFPN 利用自适应空间融合（Adaptively Spatial Fusion, ASF）机制，通过可学习参数为每个空间位置分配不同尺度特征的融合权重，使得网络能够根据局部内容特征选择性地融合来自不同感受野的信息。如图2.12所示，输入图像经过主干网络提取出三个尺度的特征层，自适应空间融合是针对每一个尺度层，将其他尺度层的特征通过上采样或下采样被调整到相同的空间尺寸，然后根据学习得到的权重图进行空间上的融合，从而增强多尺度特征的表达能力。该方法可以根据实际需求灵活调整，以适用于更多或更少层级的情况。

图 2.12 自适应空间融合（ASF）操作示意图^[59]

Transformer 在目标检测中的应用进一步推动了多尺度特征融合的发展，相关研究如 Deformable DETR^[42] 和 MViT^[60] 通过注意力机制实现了不同尺度特征之间的全局交互与融合。此外，结合卷积神经网络与 Transformer 的混合建模方式，通过协同利用局部细节与全局上下文信息，实现了多尺度特征的动态融合与互补表达^[61-62]。

2.4 知识蒸馏

知识蒸馏 (Knowledge Distillation, KD)^[63] 是一种广泛应用的高效模型压缩方法, 它通过将大型、复杂的教师模型中的知识传递给较小的学生模型, 从而实现性能提升的同时减小模型的规模。传统知识蒸馏方法通过对齐教师模型与学生模型输出层的预测分布实现知识传递。这种方式侧重于让学生模型学习教师模型在输出层的概率分布, 从而提高其预测能力。知识蒸馏方法的基本框架如图2.13所示。

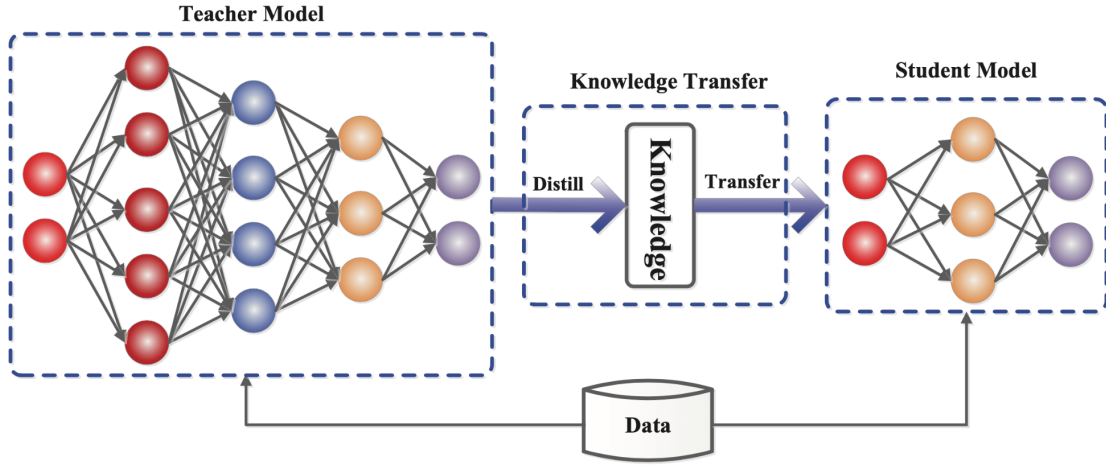


图 2.13 知识蒸馏框架示意图^[64]

基于这一思想, 经典知识蒸馏将“知识”定义为教师模型输出的类别概率分布, 即软标签 (soft targets)。设教师模型与学生模型在输出层的 logits 分别为 z^t 与 z^s , 其对应的类别概率分别为:

$$p_i^t = \frac{\exp(z_i^t/T)}{\sum_{j=1}^C \exp(z_j^t/T)}, \quad p_i^s = \frac{\exp(z_i^s/T)}{\sum_{j=1}^C \exp(z_j^s/T)}, \quad (2.14)$$

其中, p_i^t 与 p_i^s 分别表示教师模型与学生模型在第 i 类上的预测概率值, $j \in \{1, 2, \dots, C\}$, C 为类别总数。温度参数 $T > 0$ 用于调整 logits 经 softmax 后所得概率分布的平滑程度, 从而有助于学生模型学习教师模型所包含的类别间关系信息。

学生模型通过最小化其预测分布与教师分布之间的 Kullback-Leibler (KL) 散度

来学习该响应知识，其蒸馏损失定义为：

$$\begin{aligned}\mathcal{L}_{KD}^{\text{logit}} &= T^2 \cdot \text{KL}(p^t \| p^s) \\ &= T^2 \sum_{i=1}^C p_i^t \log \left(\frac{p_i^t}{p_i^s} \right),\end{aligned}\tag{2.15}$$

其中，系数 T^2 用于补偿温度缩放对梯度幅值的影响，确保损失在不同 T 下具有一致的优化尺度。

自蒸馏（Self-distillation）^[65]通过让模型的早期层从模型自身的精炼输出中进行学习，从而减少了对外部教师模型的依赖。通常情况下，网络中最深层或最后几层的特征表示或预测结果被认为具有更强的语义信息和泛化能力，因此将其作为教师，对网络的浅层或中间层进行监督训练。这种机制允许模型在训练过程中进行内部知识转移，从而提升模型整体的表征能力和收敛速度。假设一个神经网络被分为 K 个阶段，第 k 个阶段输出的逻辑向量（Logits）为 z^k 。其中， z^K 表示网络最终层（教师）的输出，而 $z^k (k < K)$ 表示中间浅层（学生）的输出，其概率分布分别定义为：

$$p_i^K = \frac{\exp(z_i^K/T)}{\sum_{j=1}^C \exp(z_j^K/T)}, \quad p_i^k = \frac{\exp(z_i^k/T)}{\sum_{j=1}^C \exp(z_j^k/T)},\tag{2.16}$$

其中， p_i^K 表示最终阶段输出对类别 i 的预测概率， p_i^k 表示第 k 个阶段预测为类别 i 的概率。自蒸馏损失定义为：

$$\begin{aligned}\mathcal{L}_{SD}^k &= T^2 \cdot \text{KL}(p^K \| p^k) \\ &= T^2 \sum_{i=1}^C p_i^K \log \left(\frac{p_i^K}{p_i^k} \right),\end{aligned}\tag{2.17}$$

在目标检测任务中，提出了多种针对性的知识蒸馏方法以提升检测性能^[66-70]。定位蒸馏（Localized Distillation, LD）^[66]表明，定位知识蒸馏比语义知识蒸馏更重要、更高效。如图2.14所示，定位蒸馏将边界框信息转换为概率分布来实现知识传递，得益于概率分布表示，可以更高效地将教师模型的定位知识传递给学生模型，从而提升目标检测的性能。教师模型和学生模型分别记为 z_T 与 z_S ，通过带有温度系数 τ 的 softmax 函数转换为对应的概率分布 p_T 和 p_S 。定位蒸馏通过 KL 散度量两个概率

分布之间的差异来实现，其计算公式为：

$$\begin{aligned}\mathcal{L}_{LD} &= \mathcal{L}_{KL}(\mathbf{p}_T, \mathbf{p}_S) \\ &= \mathcal{L}_{KL}(\mathcal{S}(\mathbf{z}_T, \tau), \mathcal{S}(\mathbf{z}_S, \tau)).\end{aligned}\quad (2.18)$$

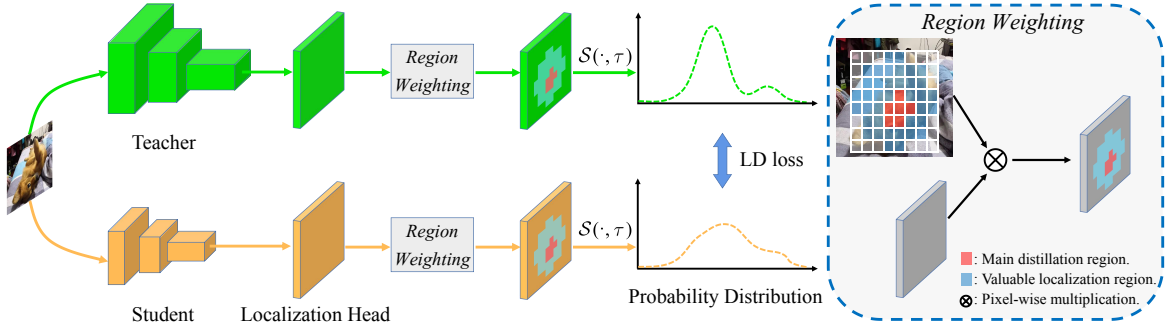


图 2.14 定位蒸馏 (LD) 方法示意图^[66]

全局最优定位自蒸馏 (GO-LSD)^[67] 结合了自蒸馏与定位蒸馏方法，如图2.15所示，通过将最终解码器层的精细分布的定位知识蒸馏到浅层，促使浅层预测通过与精确软标签的对齐来提升定位准确性。

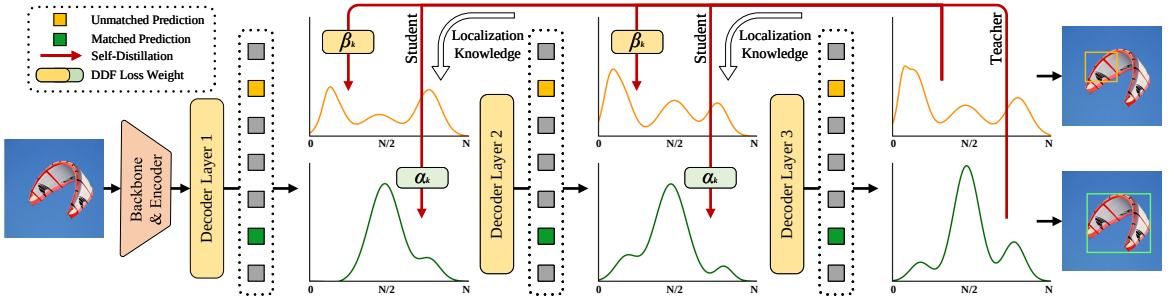


图 2.15 全局最优定位自蒸馏 (GO-LSD) 方法示意图^[67]

2.5 混合专家模型

混合专家模型 (Mixture of Experts, MoE)^[71] 是一类基于分治思想 (divide-and-conquer) 的模型结构，其核心思想是将复杂的学习任务划分为若干子任务，并由多个功能互补的子模型 (称为“专家”) 分别进行建模，再通过一个门控网络 (Gating Network) 对各专家的输出进行自适应加权融合，从而得到最终预测结果。MoE 模拟了人类专家团队的协作模式，每个专家专注于特定任务或领域，而门控网络则负责根据输入特征动态分配计算资源。这一设计使得模型能够在保持高性能的同时，显著

降低计算复杂度。稀疏门控混合专家层^[72]将 MoE 模型与 LSTM（长短期记忆网络）相结合，在稀疏性机制下实现了大规模模型的快速推理，主要应用于翻译领域，但也面临训练成本高和不稳定等挑战。随着深度学习的发展，研究将 MoE 模型应用到 Transformer 架构中^[73]，并提供了高效的分布式并行计算架构，使得 MoE 模型能够在分布式环境中进行大规模的训练和推理，进一步推动了 MoE 模型在自然语言处理领域的应用。Switch Transformer^[74]和 GLaM^[75]模型进一步挖掘了 MoE 技术在自然语言处理中的应用潜力，通过优化门控机制和专家设计，在多个 NLP 任务上取得优异表现。MoE 模型的应用范围不断扩大，除了在自然语言处理领域继续取得突破外，还在计算机视觉、多模态学习等领域得到了广泛的研究和应用。在视觉领域，相关研究如 V-MoE^[76]、M³vit^[77]和 ViMoE^[78]等，将 MoE 集成到 Vision Transformer (ViT) 模型中，用以替代标准 MLP 模块，实现了在图像分类任务、多视觉任务学习以及语义分割等任务上的性能提升。在多模态领域，MoE 在处理融合文本、图像、音频等多种数据类型多模态任务中展现出巨大潜力^[79-82]。通过 MoE，模型可以为不同模态或不同任务动态分配专家。尽管 MoE 在扩展性上优势显著，但依然存在一些挑战。首先是负载不均衡问题，门控网络可能存在选择偏向，导致部分专家被频繁激活，而其他专家则很少参与计算。这种不均衡的专家利用会造成部分参数训练不足，降低模型的训练效率，并可能对整体性能产生不利影响。其次是传统的 MoE 模型参数量高，限制了其在边缘计算设备上的部署与应用，这促使研究者转向稀疏化混合专家模型的研究。

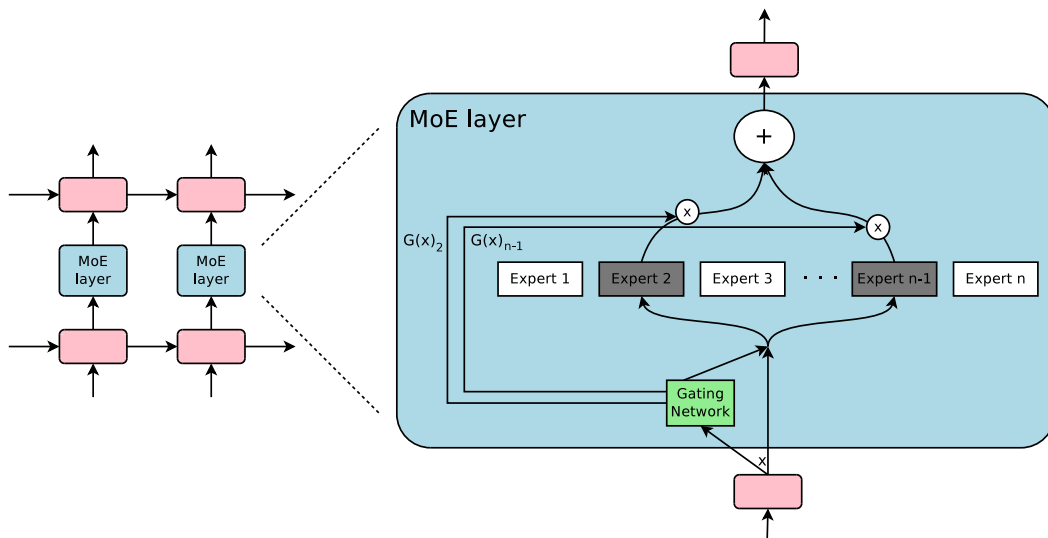


图 2.16 混合专家（MoE）模型的结构示意图^[72]

MoE 主要由两个关键部分组成：**MoE** 层和门控网络。如图2.16所示，**MoE** 层包含若干专家（**Experts**），每个专家本身是一个独立的神经网络。在实际应用中，这些专家通常是前馈网络（**FFN**），但它们也可以是更复杂的网络结构，甚至可以是 **MoE** 层本身，从而形成层级式的 **MoE** 结构。

门控网络用于决定每个输入特征应当路由到哪个专家进行处理。对于每个输入特征，门控网络根据其自身的表示分配权重，将其送入一个或多个专家网络。若同一输入特征被分配给多个专家进行计算，则各专家的输出结果会经过聚合操作，得到该输入特征的最终输出。这种机制实现了动态路由与专家分工，使模型能够针对不同输入自适应选择合适的计算路径，同时通过聚合多个专家的输出提升模型的表达能力和特征建模的丰富性。假设有 N 个专家网络 $\{E_1, E_2, \dots, E_N\}$ 和一个输入 $x \in \mathbb{R}^d$ ，专家网络的输出 y 定义为：

$$y = \sum_{i=1}^N G(x)_i E_i(x), \quad (2.19)$$

其中， x 是输入向量， $E_i(x)$ 是第 i 个专家网络对输入 x 的处理结果。在 **Transformer** 架构中，每个专家 E_i 通常由一个独立的前馈神经网络（**FFN**）构成。 $G(x)$ 是门控网络的输出， $G(x)_i$ 代表第 i 个专家的权重值。门控网络 $G(x)$ 的作用是根据输入 x 动态计算每个专家的激活权重，其定义如公式2.20所示：

$$G(x) = \text{softmax}(x \cdot W_g), \quad (2.20)$$

其中， $W_g \in \mathbb{R}^{d \times N}$ 是门控网络中可学习的权重矩阵。

2.6 本章小结

本章围绕本文研究所涉及的理论与技术基础进行了系统阐述。首先介绍了目标检测模型的基本概念与原理，为理解模型设计提供了基础认知。随后详细说明了本文所采用的基线模型 **RT-DETR** 及其结构特性。然后，对注意力机制、多尺度特征融合方法、知识蒸馏以及混合专家（**MoE**）模型的基本思想与实现原理进行了分析与讲解，为本研究工作的开展奠定了坚实的理论基础。

第三章 基于协同注意力和自蒸馏的农业果实目标检测方法

在实际农业生产中,果实的成熟度具有变化性,并且受环境光照、遮挡、果实大小差异以及背景复杂度等因素影响,使得成熟度检测任务具有较高的复杂性和挑战性。针对现有目标检测方法在复杂环境下存在的检测困难,以及在多尺度信息处理能力上的不足,提出一种面向农业果实成熟度检测的框架,以实现果实目标的精准定位与成熟度的准确判断。该模型提出多尺度特征融合重加权模块,通过加强多尺度特征融合能力,以增强模型对不同尺度果实的检测能力;设计协同注意力解码器,结合交叉注意力和自注意力的优势,优化目标查询的交互方式,有效提升检测精度;同时在不额外增加推理计算量的前提下,提出层级自蒸馏策略,实现解码器内部不同层之间的知识传递,增强目标定位能力,进一步提高检测的准确性。本方法为农业场景下实现更高效的自动化生产与果实采摘提供了可行的技术支撑。

3.1 方法概述

3.1.1 模型框架

本章提出了一种面向复杂环境下果实成熟度检测的多尺度 Transformer 模型 FMD-DETR (Fruit Maturity Detection-DEtection TRansformer)。该模型由三个核心模块构成:多尺度特征融合重加权模块 (Multi-scale Feature Fusion Re-weighting Module, MS-FFRM)、协同注意力解码器 (Co-Attention Decoder) 以及层级自蒸馏策略 (Hierarchical Self-Distillation strategy)。多尺度特征融合重加权模块对不同层级的特征进行融合,提升处理多尺度信息的能力,从而增强对不同尺寸果实的检测效果。协同注意力解码器结构进一步优化目标查询,提升对遮挡目标的检测能力。层级自蒸馏策略提取高层次的知识帮助浅层学习到更多复杂的空间和特征表示,增强其对物体位置的预测能力。

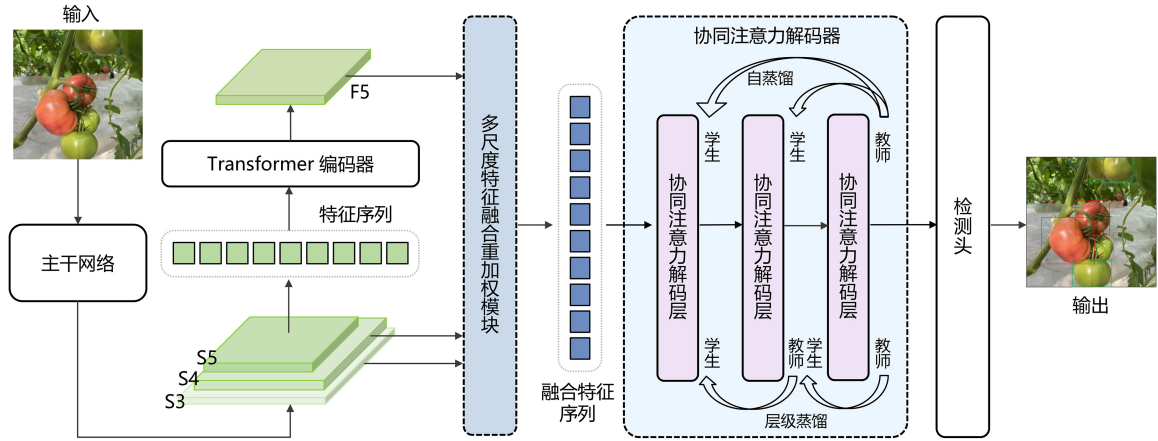


图 3.1 FMD-DETR 模型框架图

FMD-DETR 的架构如图3.1所示。首先通过主干网络 ResNet 提取到 S3、S4、S5 三个不同尺度特征图^[12]，分别对应浅层高分辨率细节信息、中层语义信息与深层强语义信息，并且能够覆盖不同尺度目标的表示需求。其中将具有更丰富语义的 S5 特征展平为特征序列后输入编码器，以进行全局上下文建模。仅对 S5 特征图进行该操作，目的是保留关键语义信息的同时降低计算复杂度。编码器输出的特征 F5 与 S3、S4 一起送入多尺度特征融合重加权模块，通过跨尺度特征融合与自适应权重分配，进一步提升了在不同尺度上检测关键特征的能力。随后，融合后的特征被输入到协同注意力解码器。在该解码过程中，目标查询与融合特征进行充分的特征匹配，通过协同注意力机制有效消除边界框预测的冗余与混淆。同时，通过层级自蒸馏策略，解码器深层网络的知识被传递至浅层网络，帮助浅层网络更好地学习复杂的空间与特征表示，进而提升目标定位的精度。最后，解码器输出的目标查询经过预测头输出目标边界框以及成熟度类别，完成端到端的高效检测任务。

3.1.2 多尺度特征融合重加权模块

图3.1所示的多尺度特征融合重加权模块（MS-FFRM）旨在整合不同尺度的特征，从而增强网络的特征表征能力。该模块包含两个核心组件：自适应特征融合模块（Adaptive Feature Fusion Module, AFFM）与特征通道重加权模块（Feature Channel Reweighting Module, FCRM），如图3.2所示。其中，AFFM 负责融合多尺度特征，而 FCRM 则通过加权重重要的特征来增强网络的特征表示能力。

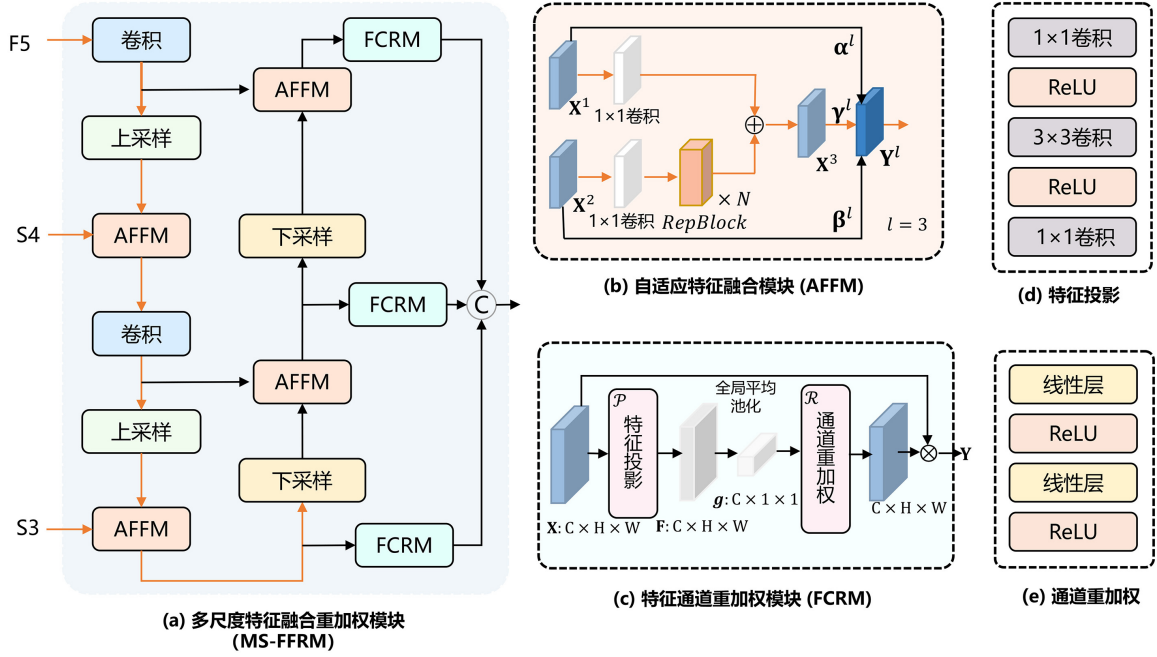


图 3.2 多尺度特征融合重加权模块结构示意图

自适应特征融合模块 (AFFM) 如图3.2(b) 所示。该模块有两个输入特征，分别设为 $\mathbf{X}^1, \mathbf{X}^2 \in \mathbb{R}^{C \times H \times W}$ ，其中 C 表示通道数， H 和 W 分别表示特征图的高度和宽度。 $\mathbf{X}^1, \mathbf{X}^2$ 分别输入到两条路径，包含两个 1×1 卷积层和 N 个 RepBlocks^[83]，两条路径的输出通过逐元素相加融合，得到融合特征 $\mathbf{X}^3 \in \mathbb{R}^{C \times H \times W}$ 。其中，RepBlocks 是可重新参数化的卷积模块，通过在训练阶段提升模型性能，并在部署时简化结构来加速推理。输入特征与融合特征通过自适应特征融合实现 l 级特征融合，该过程采用可学习的空间融合权重，能够根据不同层级特征的重要性动态调整融合权重，自适应特征融合过程如公式3.1所示：

$$\mathbf{Y}^l = \alpha^l \odot \mathbf{X}^1 + \beta^l \odot \mathbf{X}^2 + \gamma^l \odot \mathbf{X}^3, \quad (3.1)$$

其中， $\mathbf{Y}^l \in \mathbb{R}^{C \times H \times W}$ 表示输出特征， $\odot$ 表示逐元素相乘， α^l, β^l 和 γ^l 表示对应的空间重要性权重图。空间权重图的每个空间位置 (i, j) 的权重为 α_{ij}^l ，通过 softmax 函数归一化得到，其定义如公式3.2所示：

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}}, \quad \beta_{ij}^l = \frac{e^{\lambda_{\beta_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}}, \quad \gamma_{ij}^l = \frac{e^{\lambda_{\gamma_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}}, \quad (3.2)$$

其中， $\lambda_{\alpha_{ij}}^l, \lambda_{\beta_{ij}}^l$ 和 $\lambda_{\gamma_{ij}}^l$ 是通过 1×1 卷积层从特征图计算得到的。

特征通道重加权模块（FCRM）如图3.2(c)所示。该模块首先对输入特征图 $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ 进行特征投影处理，经过特征投影模块 $\mathcal{P}$ 将输入特征映射到一个新的特征空间，得到投影特征图 $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ ，该过程能够有效提取输入特征 $\mathbf{X}$ 中的关键空间信息。 $\mathbf{F}$ 的定义如公式3.3所示：

$$\mathbf{F} = \mathcal{P}(\mathbf{X}), \quad (3.3)$$

随后，投影特征图 $\mathbf{F}$ 经通道重新加权模块 $\mathcal{R}$ 处理，首先对 $\mathbf{F}$ 进行全局平均池化，即对每个通道在空间维度上求平均，再通过模块 $\mathcal{R}$ 获得通道权重 $\mathbf{g} \in \mathbb{R}^{C \times 1 \times 1}$ ，该权重反映了特征图 $\mathbf{F}$ 中不同通道的重要性， $\mathbf{g}$ 定义如公式3.4所示：

$$\mathbf{g} = \mathcal{R}(\text{Pool}_{avg}(\mathbf{F})), \quad (3.4)$$

其中， Pool_{avg} 表示全局平均池化。最后，将通道权重 $\mathbf{g}$ 应用于投影特征图 $\mathbf{F}$ 以调整不同通道的权重，从而增强网络的特征表示能力。输出特征图 $\mathbf{Y} \in \mathbb{R}^{C \times H \times W}$ 定义如公式3.5所示：

$$\mathbf{Y} = \mathbf{F} \odot \mathbf{g}, \quad (3.5)$$

其中， $\odot$ 表示逐元素相乘。

多尺度特征融合重加权模块（MS-FFRM）旨在利用跨尺度特征融合与通道重新加权的协同作用，整合不同尺度的特征，从而自适应地增强关键通道和关键区域，达到提升多尺度目标检测与定位准确性的目的。

3.1.3 协同注意力解码器

在 DETR 解码器中，交叉注意力机制和自注意力机制对目标查询产生相互制约的影响。具体来说，交叉注意力机制的优点在于能够有效地将多个查询集中在同一目标上，从而确保目标查询能够充分关注到与目标相关的特征，从而提升检测精度。然而，这也导致了在解码过程中，多个查询可能会关注到相同的物体，导致检测到的对象重叠。而自注意力机制则能够使得查询相互分散，优先保留与目标最匹配的查询，从而减少冗余关注^[84]。

为优化 DETR 解码器中目标查询之间的交互，本章提出了一种协同注意力解码器（Co-Attention Decoder），该解码器由一个去除自注意力层的解码器（C-Decoder）和

一个标准的 DETR 解码器 (SC-Decoder) 组成。协同注意力解码器层的结构如图3.3所示。

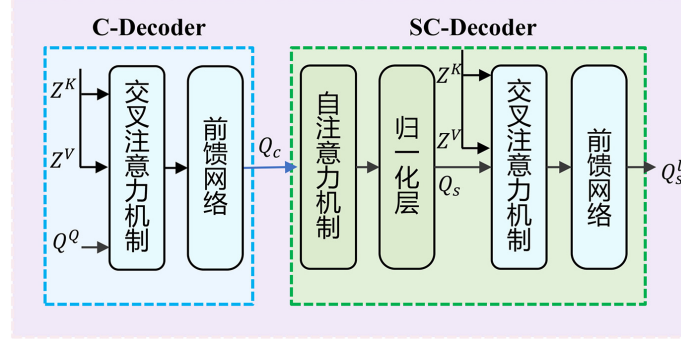


图 3.3 协同注意力解码器层结构示意图

C-Decoder 是一个专注交叉注意力 (Cross-Attention) 层的解码器，它不包含自注意力 (Self-Attention) 层，并且与 SC-Decoder 共享参数，包括交叉注意力层和前馈神经网络 (FFN)。C-Decoder 的设计目的是为每个真实目标分配多个正样本查询，从而提高对小目标或密集目标的检测能力。C-Decoder 中查询向量的更新公式3.6如下：

$$Q_c = Att_{cross}(Q^Q, Z^K, Z^V), \quad (3.6)$$

$$Q^Q = QW^Q, \quad Z^K = ZW^K, \quad Z^V = ZW^V, \quad (3.7)$$

其中， Z 表示编码器的输出特征， Q^Q 表示输入到解码器的查询向量， Z^K 表示键， Z^V 表示值。 W^Q , W^K , W^V 为可学习的线性投影矩阵。

SC-Decoder 与标准 DETR 解码器结构一致，每个解码器块由自注意力 (Self-Attention) 层、交叉注意力 (Cross-Attention) 层以及前馈神经网络 (FFN) 组成。SC-Decoder 的设计目标是进一步消除重复检测，实现一对一的精准匹配。SC-Decoder 巧妙地融合了 C-Decoder 的输出先验，其查询向量的更新过程如公式3.8和3.9所示：

$$Q_s = Att_{self}(Q_c^Q, Q_c^K, Q_c^V), \quad (3.8)$$

$$Q_s^l = Att_{cross}(Q_s^Q, Z^K, Z^V), \quad (3.9)$$

$$Q_c^Q = Q_c W^Q, \quad Q_c^K = Q_c W^K, \quad Q_c^V = Q_c W^V, \quad Q_s^Q = Q_s W^Q, \quad (3.10)$$

其中, Z 表示编码器输出特征, Q_c^Q 表示由 C-Decoder 输出 Q_c 生成的查询向量, Q_c^K 表示由 C-Decoder 输出 Q_c 生成的键, Q_c^V 表示由 C-Decoder 的输出 Q_c 生成的值, 而 Q_s^Q 表示经过自注意力后的查询向量。最终, 输出嵌入 Q_s^l 被输入到预测头进行类别标签和位置的预测。

协同注意力解码器旨在通过结合交叉注意力和自注意力的优势, 优化目标查询的交互方式, 从而提高对遮挡目标的检测精度, 减少冗余预测, 以提升目标检测的性能。

3.1.4 层级自蒸馏策略

自蒸馏是一种将模型深层信息传递至浅层的学习策略。如图3.1所示, 自蒸馏采用解码器最终层的输出作为教师模型, 将其提供的知识传递给学生模型的浅层结构。解码器最后一层生成的分布信息, 包括边界框概率分布与边界框回归参数, 为浅层学生模型提供了更精确的位置定位知识^[67]。通过最小化教师模型与学生模型输出之间的差异, 模型逐步从教师模型中提取高层语义知识, 帮助浅层学习更复杂的空间与特征表示, 从而提升目标位置预测精度。自蒸馏的损失函数定义如公式3.11-3.13所示:

$$\mathcal{L}_{SD} = T^2 \sum_{l=1}^{L-1} \left(\sum_{k=1}^{K_m} \alpha_k \cdot \text{KL} (P_k^L \| P_k^l) + \sum_{k=1}^{K_u} \beta_k \cdot \text{KL} (Q_k^L \| Q_k^l) \right), \quad (3.11)$$

以及

$$\alpha_k = IoU_k \cdot \frac{\sqrt{K_m}}{\sqrt{K_m} + \sqrt{K_u}}, \quad (3.12)$$

以及

$$\beta_k = Conf_k \cdot \frac{\sqrt{K_u}}{\sqrt{K_m} + \sqrt{K_u}}, \quad (3.13)$$

其中, KL 表示 Kullback-Leibler 散度, T 是用于平滑 logits 的温度参数, L 表示解码器的最后一层, $l \in \{1, 2, \dots, L-1\}$ 表示从第一层到倒数第二层的各个解码层。 P_k 表示与第 k 个预测结果对应的概率分布, Q_k 表示与第 k 个未匹配预测结果对应的概率分布。第 k 个匹配预测的蒸馏损失由权重 α_k 进行加权, 其中 IoU_k 通过第 k 个预

测的交并比（Intersection over Union, IoU）计算得到。 K_m 和 K_u 分别表示匹配预测与未匹配预测的数量。对于第 k 个未匹配预测，其对应的权重为 β_k ，其中 $Conf_k$ 表示分类置信度，其计算方式为教师模型 logits 中的最大类别概率。 IoU_k 和 $Conf_k$ 会在每次前向传播过程中根据当前的预测结果重新计算。

自蒸馏方法以解码器的最后一层作为教师模型、以浅层作为学生模型进行知识传递，但这种方式缺乏相邻层之间的知识交互。为进一步增强蒸馏效果，本章提出了一种层级知识蒸馏策略，如图3.1所示，该策略在解码器的多层之间进行蒸馏学习。在该策略中，后一层作为前一层的教师模型，通过层间知识传递对模型进行逐层优化。根据层级知识蒸馏的定义，蒸馏过程在解码器的各个层之间进行，并可为每一对相邻层定义对应的层级蒸馏损失函数。当蒸馏发生在第 l 层与第 $l+1$ 层之间时，其损失函数定义如公式3.14所示：

$$\mathcal{L}_{HD} = T^2 \sum_{l=1}^{L-1} \left(\sum_{k=1}^{K_m} \alpha_k \cdot \text{KL} (P_k^{l+1} \| P_k^l) + \sum_{k=1}^{K_u} \beta_k \cdot \text{KL} (Q_k^{l+1} \| Q_k^l) \right), \quad (3.14)$$

其中， α_k 与 β_k 的计算方式与自蒸馏相同，分别基于预测框的交并比（IoU）与分类置信度确定。层级蒸馏能够在模型的浅层与深层之间实现更加充分的知识传递，从而提升模型的整体性能。

自蒸馏使模型能够基于自身输出优化预测结果，从而摆脱对外部教师模型的依赖。由此降低了训练成本，同时增强模型的自我学习能力。然而，对于复杂的模型，单层知识传递可能无法充分利用模型不同层级的特征信息。而层级蒸馏通过在解码器的多层间传递知识，强化了不同层特征之间的关联性，使模型能够有效地利用各层的学习信息。本章提出的层级自蒸馏策略将自蒸馏与层级蒸馏相结合，旨在提升模型的整体学习能力，使其能够更好地学习定位知识，以提高目标定位的准确性。

3.2 实验设计

3.2.1 实验环境与参数设置

(1) 数据集

本章选取了四个数据集进行模型训练与验证：**Laboro** 番茄数据集^[85]、草莓数据集^[86]、真实世界番茄数据集和 **FruitRipeness** 数据集^[87]。

Laboro 番茄数据集包含 804 张图像,涵盖普通番茄 **b_** 和樱桃番茄 **l_**,数据集中果实在大小、颜色和遮挡程度上具有较大变化。根据番茄大小和成熟阶段分为六类:普通番茄-未成熟 (**b_green**)、樱桃番茄-未成熟 (**l_green**)、普通番茄-半熟 (**b_half_ripened**)、樱桃番茄-半熟 (**l_half_ripened**)、普通番茄-完全成熟 (**b_fully_ripened**) 以及樱桃番茄-完全成熟 (**l_fully_ripened**)。该数据集的训练集经过数据增强处理,包含随机色彩抖动、图像拉伸、裁剪、翻转及尺寸调整,最终获得 3215 张训练图像。草莓目标检测数据集由 813 张图像组成,拍摄于室外农田环境,数据集背景复杂,存在部分遮挡和光照变化,包括三类目标:成熟草莓 (**ripe**)、草莓花梗 (**peduncle**) 和未成熟草莓 (**unripe**)。真实场景番茄数据集来源于真实环境和温室环境的番茄,包含 1110 张番茄图像,分为三类:未成熟 (**unripe**)、半成熟 (**semi-ripe**) 和完全成熟 (**fully-ripe**)。该数据集突出了果实在真实环境下的复杂背景、遮挡及光照变化情况,适合用于评估模型在实际农业场景下的性能。**FruitRipeness** 数据集包含 5000 张室内外光照条件下采集的 RGB 图像,涵盖香蕉与芒果在成熟与未成熟状态下的外观特征,分为四个类别:生香蕉 (**Raw_Banana**)、生芒果 (**Raw_Mango**)、熟香蕉 (**Ripe_Banana**) 和熟芒果 (**Ripe_Mango**)。该数据集可用于训练模型并验证其在不同果实类别和环境下的泛化能力。图3.4展示了本章所用数据集图像示例。



图 3.4 本章使用的数据集图像示例

（2）实验环境与参数设置

本章实验使用 RT-DETR 作为基础模型实现，采用 ResNet-50 作为主干网络，初始权重是 ImageNet 上预训练的结果。所有模型在 PyTorch 框架下实现，并在 NVIDIA RTX 4090 GPU（24GB 显存）上训练和测试。模型训练过程中使用 AdamW 优化器进行参数更新，初始学习率设置为 0.0001，批处理大小为 4，权重衰减系数设为 1×10^{-4} 。训练总周期为 100 个 epoch，以充分保证模型的收敛性与稳定性。

（3）评价指标

在模型评估阶段，采用平均精度（Average Precision, AP）作为主要评价指标，用于综合衡量模型在目标检测任务中的整体精度。为了更全面地反映模型的检测能力，同时报告在不同交并比（IoU）阈值下的性能指标，包括 AP_{50} （IoU=0.50）和 AP_{75} （IoU=0.75）。此外，为进一步评估模型对不同尺度目标的检测能力，还统计了小目标（ AP_S ）、中等目标（ AP_M ）和大目标（ AP_L ）的 AP 值，从而全面衡量模型在不同场景和目标尺度下的检测性能。

除了检测精度指标外，本研究还关注模型的计算复杂度与实际运行效率，具体包括模型参数量（Params）、浮点运算量（GFLOPs）、以及推理速度（Frames Per Second, FPS）。参数量反映了模型的存储需求和网络规模，GFLOPs 衡量模型在单次前向推理中所需的计算量，而 FPS 用于评估模型在实际场景下的实时推理性能。通过这些指标能够更全面地评估所提出模型在精度、计算成本和实际应用可行性方面的综合性能，为农业果实成熟度检测任务的高效部署提供参考依据。

3.2.2 定量实验

表3.1展示了在 Laboro 番茄数据集上，所提算法与其他目标检测算法的对比结果，包括 Deformable-DETR^[42]、Conditional-DETR^[43]、DAB-DETR^[44]、DINO^[47]、Co-DETR^[88]、Rank-DETR^[89]、RT-DETR^[12]、Mobile-DETR^[12]、RT-DETRv2^[48]、Lite-DETR^[90]、YOLOv8s^[35]、YOLOv8x^[35]、YOLOv9s^[91]、YOLOv10s^[36]、YOLO11s^[37]以及 EdgeYOLO^[92]。其中 Mobile-DETR、Lite-DETR 和 EdgeYOLO 属于轻量级目标检测模型，主要面向资源受限的环境和实时推理应用。通过与这些轻量级及标准目标检测模型的对比，可以全面评估所提出的算法的综合性能。实验结果最优的指标以粗体标出，次优的指标以下划线标出。从表3.1的实验结果可见，所提出的算法在 Laboro 番茄数据集上表现出了显著的优越性。FMD-DETR 以 75.4% 的平均精度（AP）超越

现有目标检测算法。这表明，FMD-DETR 在保持 DETR 架构端到端检测优势的同时，有效克服了传统 Transformer 模型在特定农业场景下的性能瓶颈。

表 3.1 Laboro 番茄数据集上的模型定量比较

Model	Backbone	Params(M)	GFLOPs	FPS	AP	AP ₅₀	AP ₇₅	AP _S
Deformable-DETR ^[42]	R50	40	165	18	46.7	71.3	50.8	64.4
Conditional-DETR ^[43]	R50	44	86	14	68.7	<u>87.7</u>	77.8	77.4
DAB-DETR ^[44]	R50	48	87	18	68.2	86.3	76.5	79.9
DINO ^[47]	R50	48	235	18	74.8	87.3	<u>81.7</u>	86.9
Co-DETR ^[88]	R50	64	-	18	70.9	85.2	77.9	85.5
Rank-DETR ^[89]	R50	49	226	15	74.5	88.2	81.5	85.5
RT-DETR ^[12]	R50	42	66	20	71.1	82.7	77.8	86.4
Mobile-DETR ^[12]	MobileNet	15	18	20	60.4	76.0	66.6	77.5
RT-DETRv2 ^[48]	R50	42	66	20	70.5	81.7	76.8	<u>87.0</u>
Lite-DETR ^[90]	R50	47	157	10	68.5	83.5	75.0	85.0
YOLOv8s ^[35]	-	11.2	29	10	63.4	76.1	68.8	64.6
YOLOv8x ^[35]	-	68	257	10	72.9	82.1	78.7	73.9
YOLOv10s ^[36]	-	7.2	22	8	73.9	85.8	80.1	75.1
YOLO11s ^[37]	-	9.4	22	8	<u>75.4</u>	87.2	81.2	76.5
EdgeYOLO ^[92]	-	5.8	27	32	50.1	75.0	57.9	65.3
Ours	R50	44	66	22	75.4	86.7	81.9	88.3

在小目标检测性能方面，FMD-DETR 的小目标平均精度（AP_S）达 88.3%，明显优于所有对比算法，充分验证了该方法在小目标检测方面的有效性。Laboro 番茄数据集中存在大量尺度较小、分布密集且易受遮挡的果实目标，传统检测模型往往难以在保证高层语义表达的同时保留足够的空间细节信息。FMD-DETR 提出多尺度特征重加权模块，增强模型对多尺度信息的敏感性，使模型能够根据不同尺度目标的重要性动态调整特征响应强度，从而增强了对小尺度目标的检测能力。同时，层级自蒸馏策略在训练阶段引导浅层特征学习来自深层语义的定位知识，提升了模型对复杂背景下目标边界的定位精度，使得模型能够更精确地聚焦于细节特征，从而提升了小物体的检测性能。此外，相较于 RT-DETR，本章模型在计算量未显著增加的前提下，推理速度（FPS）达到了 22 FPS，提升了约 10%，同时显著提升精度，体现了在效率与性能之间的平衡。相比 DINO，FMD-DETR 减少了 8.3% 参数量、降低约 72% 计算量，在计算量大幅降低的同时保持了更高的检测精度。与轻量级模型 EdgeYolo 相比，尽管其推理速度较快，但检测精度明显不足，而 FMD-DETR 在保持

较高 FPS 的同时，实现了精度与速度的更优折中，兼顾了实时性与可靠性。

表3.2展示了所提算法与其他目标检测算法在真实番茄数据集上的对比结果。根据表3.2中的实验结果，提出的算法在真实场景下的番茄数据集上表现出了优越的检测能力。与其他算法相比，本章的 FMD-DETR 算法的平均精度（AP）达到 52.7%，AP₇₅ 达到 63.3%，表明了其在高精度目标检测中的优势。这说明所提出的协同注意力解码器通过优化目标查询，使其更有效地聚焦于目标特征，有效抑制了目标查询冗余，改善模型在复杂场景下的表现。此外，模型整体在不同尺度果实上的检测能力表现均衡，说明多尺度特征融合重加权模块能够有效地增强模型对不同尺度特征的敏感性和关注能力，有效整合来自不同尺度的特征信息，在复杂环境下保持较高的精度。在推理效率方面，FMD-DETR 算法在保证较高检测精度的同时，达到 16 FPS，相比大多数 DETR 系列方法具有更快的推理速度，显示出在效率与性能之间的良好平衡。

表 3.2 真实场景番茄数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Deformable-DETR ^[42]	R50	11	37.9	66.2	37.9	54.2	33.8	44.3
Conditional-DETR ^[43]	R50	11	<u>50.9</u>	83.6	57.5	59.5	46.9	55.2
DAB-DETR ^[44]	R50	11	50.5	<u>82.2</u>	56.9	41.0	47.6	54.9
DINO ^[47]	R50	11	46.7	73.5	54.5	63.8	45.3	50.3
Co-DETR ^[88]	R50	11	49.5	78.5	57.8	71.0	45.8	54.7
Rank-DETR ^[89]	R50	15	49.0	79.2	54.6	63.8	45.9	53.8
RT-DETR ^[12]	R50	14	49.8	80.9	57.7	55.1	44.0	54.0
Mobile-DETR ^[12]	MobileNet	12	48.3	79.5	51.9	58.5	45.4	52.0
RT-DETRv2 ^[48]	R50	14	50.0	78.0	57.5	<u>67.7</u>	44.7	<u>55.4</u>
Lite-DETR ^[90]	R50	11	48.5	75.3	58.5	64.3	46.9	51.8
YOLOv10s ^[36]	-	7	49.0	75.3	<u>59.3</u>	64.4	46.5	53.2
EdgeYOLO ^[92]	-	21	49.6	81.0	57.1	47.7	<u>47.2</u>	51.9
Ours	R50	16	52.7	81.0	63.3	64.7	46.1	57.3

真实场景中的番茄检测任务通常面临光照变化、背景干扰、部分遮挡以及目标大小不一等挑战，这对检测模型在精度和鲁棒性方面提出了更高要求。本章提出的 FMD-DETR 算法能够有效应对这些挑战，即使在部分果实被遮挡或光照不均的情况下，也能准确识别果实并判断成熟度，展示出其在实际应用中的可靠性和适应性。

草莓数据集与番茄数据集在成熟度判定标准上存在差异。为评估模型的泛化能力，表3.3展示了所提算法与其他目标检测算法在草莓数据集上的对比结果。根据表3.3所示，所提算法在草莓数据集上表现出具有竞争力的性能，超越了多数对比方法。该算法在不同尺寸目标的检测上表现出较强的鲁棒性。本章算法在整体性能上保持明显优势，取得了最高的 41.5% 的 AP。FMD-DETR 在 AP_{50} 达到 73.9%， AP_{75} 达到 44.2%，超过其他对比模型，表明模型对于遮挡的果实目标仍然具有较高的定位能力。尽管 FMD-DETR 在 AP_S 略低于某些对比方法，但较基线方法提升了 11.7 个百分点。并且 AP_M 达到 43.6%， AP_L 达到 42.4%，均为最高，总体上在多尺度目标检测中表现出色。以上结果表明，通过协同注意力机制使模型能够更有效地捕捉多尺度上下文信息，从而提升了不同尺度目标果实的检测精度。同时，层级自蒸馏策略将深层语义知识传递至浅层，进一步增强了浅层特征表达，使模型在复杂背景和部分遮挡情况下也能保持稳健的检测能力。此外，FMD-DETR 算法具有 20 FPS 的高推理速度，展现出精度与效率的良好平衡。

草莓数据集的类别间差异较为细微，果实尺寸大小不一，且容易被遮挡，这增加了模型在目标区分与精确定位方面的难度。本章的算法能够检测不同尺度的果实特征，即使在复杂的背景和被遮挡的情况下，也能准确定位目标，从而有效应对不同果实检测任务中的挑战，彰显了其强大的泛化能力。

表3.4展示了在 FruitRipeness 数据集上的对比结果。根据表3.4的实验结果，所提算法在 FruitRipeness 数据集上表现优异。模型平均精度 (AP) 达到 87.1%，超过了多种现有目标检测方法。 AP_{75} 达到 99.1%，体现出模型在边界框精确预测方面表现优异。同时，本章算法的 AP_M 和 AP_L 分别达到 84.7% 和 87.6%，在对比算法中表现突出，表明其对不同尺寸果实的定位能力有显著提升。同时推理速度为 30 FPS，高于基线方法，体现了性能与效率之间的良好平衡。此外，相比于轻量化模型 Lite-DETR 和 Mobile-DETR，本方法在保证快速推理的同时明显提升了检测精度，实现了更优的整体性能。实验结果进一步验证了多尺度特征融合重加权模块、协同注意力机制以及层级自蒸馏策略的有效性，使模型能够更有效地捕捉多尺度特征信息，同时强化目标位置表示，从而提升检测与定位精度。本章的算法能够对不同果实类别的成熟度进行准确检测，充分验证了所提方法在果实成熟度检测任务中的实际应用价值和优越性能。

表 3.3 草莓数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Deformable-DETR ^[42]	R50	11	26.1	49.9	26.5	0.8	20.4	29.6
Conditional-DETR ^[43]	R50	11	36.1	71.2	33.5	10.2	35.9	38.2
DAB-DETR ^[44]	R50	12	38.3	70.9	39.1	5.7	39.4	39.9
DINO ^[47]	R50	12	35.1	62.0	38.8	17.2	34.2	37.5
Co-DETR ^[88]	R50	12	35.4	61.0	40.9	24.9	38.2	36.2
Rank-DETR ^[89]	R50	15	36.7	68.2	39.4	<u>24.3</u>	39.6	37.3
RT-DETR ^[12]	R50	17	40.4	70.9	41.6	10.5	<u>42.7</u>	41.2
Mobile-DETR ^[12]	MobileNet	19	39.2	63.3	38.5	8.6	35.4	38.0
RT-DETRv2 ^[48]	R50	18	<u>40.5</u>	<u>73.3</u>	41.0	16.7	42.4	<u>41.4</u>
Lite-DETR ^[90]	R50	9	35.6	69.2	36.8	20.9	35.9	37.2
YOLOv8s ^[35]	-	12	35.9	68.0	33.9	11.4	35.6	37.5
YOLOv9s ^[91]	-	11	37.7	71.2	39.1	16.9	40.7	38.1
YOLOv10s ^[36]	-	13	35.5	67.3	35.1	11.1	38.1	36.2
YOLO11s ^[37]	-	12	38.2	69.1	40.6	7.6	36.8	39.9
EdgeYOLO ^[92]	-	30	38.7	68.7	<u>42.0</u>	19.5	40.5	39.1
Ours	R50	20	41.5	73.9	44.2	22.2	43.6	42.4

表 3.4 FruitRipeness 数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _M	AP _L
Deformable-DETR ^[42]	R50	27	76.1	95.7	93.0	69.6	77.4
Conditional-DETR ^[43]	R50	29	84.6	99.2	<u>98.8</u>	80.7	85.3
DAB-DETR ^[44]	R50	25	84.3	99.4	98.6	79.4	85.0
DINO ^[47]	R50	20	85.8	99.4	98.4	83.4	86.4
Co-DETR ^[88]	R50	10	86.4	99.4	99.1	84.0	86.8
Rank-DETR ^[89]	R50	15	84.3	98.9	97.9	78.1	84.9
RT-DETR ^[12]	R50	29	<u>86.7</u>	99.3	98.7	84.1	<u>87.0</u>
Mobile-DETR ^[12]	MobileNet	26	80.7	94.2	93.2	<u>84.2</u>	81.1
RT-DETRv2 ^[48]	R50	29	86.3	99.2	98.7	83.7	86.8
Lite-DETR ^[90]	R50	10	85.6	99.4	98.6	81.9	86.2
YOLOv10s ^[36]	-	16	85.0	99.6	98.7	82.9	85.5
YOLO11s ^[37]	-	25	85.4	99.6	98.8	82.8	85.8
EdgeYOLO ^[92]	-	34	83.1	99.1	98.2	83.5	83.6
Ours	R50	30	87.1	<u>99.4</u>	99.1	84.7	87.6

3.2.3 定性实验

图3.5展示了在 Laboro 番茄数据集上的可视化结果，图中实线框表示模型预测得到的目标检测框，框旁标签表示对应的目标类别。红色虚线框表示漏检，蓝色虚线框表示误检，黄色虚线框则表示冗余框。番茄数据集中的图像反映了真实农业环境中存在的挑战，包括复杂的背景、遮挡以及果实大小的差异。在同一个图像区域内，存在不同成熟度的番茄对象，例如未成熟的绿色果实和半成熟、正在变色的果实。受繁茂的枝叶遮挡以及果实间密集重叠的影响，果实的边界模糊且特征细节不明显，这增加了检测的难度。

如图3.5所示，Deformable-DETR、Conditional-DETR、DAB-DETR、DINO、Co-DETR、Rank-DETR 和 RT-DETR 算法在进行番茄成熟度检测时，普遍存在漏检现象，表明这些方法在检测时存在一定的局限性。RT-DETRv2 算法的检测结果中，可以看到冗余框的情况，说明该算法虽然能够检测到更多的目标，但也存在产生多余框的问题。尽管 YOLOv8x、YOLOv10s 和 YOLO11s 算法对比其他算法能够较好地进行检测，但仍然存在漏检问题。相比之下，本章算法在番茄成熟度检测中表现突出，不仅在准确性上高于其他方法，而且避免了冗余框的出现，展现了出色的检测能力和更高的效率。这表明，本章算法在精确定位目标和消除冗余框的生成方面具有明显的优势，能够为番茄成熟度检测提供更加可靠的解决方案。

图3.6展示了在草莓数据集上的可视化结果，红色虚线框代表漏检，蓝色虚线框代表误检，黄色虚线框代表冗余框。草莓数据集图像中需要同时识别草莓果实和花梗。然而，由于花梗常常遮挡果实，且其颜色、纹理与背景中叶片高度相似，因此极易产生视觉混淆，这增加了检测草莓果实的难度。

从图3.6中可以看出，Deformable-DETR、Conditional-DETR、DAB-DETR 和 RT-DETR 算法在进行草莓成熟度检测时，存在较多的冗余框；Co-DETR 算法存在漏检的情况；而 DINO、Rank-DETR、RT-DETR、RT-DETRv2 以及 YOLOv8s、YOLOv9s 和 YOLOv10s 均存在误检的情况。轻量化模型 Lite-DETR、Mobile-DETR 和 EdgeYOLO 同样出现了误检的情况。相比之下，本章算法在草莓成熟度检测中表现更好，既能有效定位目标，又能减少误检和冗余框的情况，展现了更高的检测精度和鲁棒性。这表明，本章算法在多种果实成熟度检测中的表现具有较强的适应性和泛化能力，能够在复杂环境下实现更准确的检测结果，为农业生产中的果实成熟度检测提供了更

可靠的技术支持。

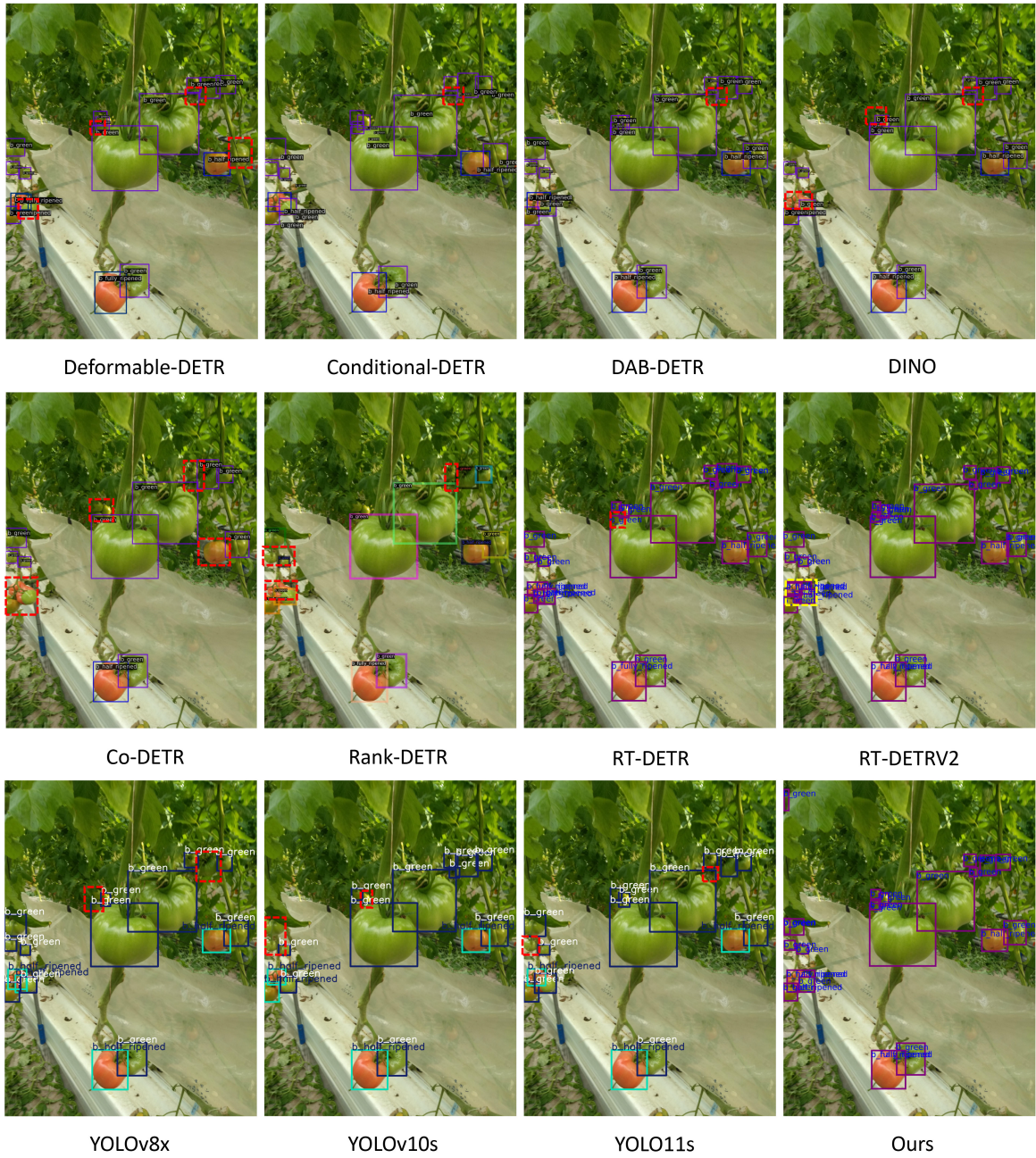


图 3.5 番茄成熟度检测结果定性比较



图 3.6 草莓成熟度检测结果定性比较

3.2.4 消融实验

模块有效性验证。为了验证本章所提出的多尺度特征融合重加权模块（MS-FFRM）、协同注意力解码器（CA-D）以及层级自蒸馏策略（HSD）在果实成熟度检测中的有效性，在 Laboro 番茄数据集上进行了消融实验。为了保证实验的一致性，所有模型均使用相同的超参数训练。在消融实验中，使用标准 RT-DETR 模型作为基线模型，逐步向基线模型添加不同模块，结果如表 3.5 所示。可以看到，基线模型平均精度（AP）为 71.1%。在仅加入多尺度特征融合重加权模块（MS-FFRM）后，AP 从 71.1% 提升至 75.0%， AP_{50} （+4.0%）、 AP_{75} （+3.2%）以及 AP_S （+2.1%）均有提升。仅使用协同注意力解码器（CA-D）时，AP 提升至 75.2%，当单独应用层级自蒸馏策略（HSD）时，AP 达到了 75.2%。进一步地，当将三者结合使用时，模型取得最高性能，AP 达到 75.4%。实验结果表明，每种策略都对模型性能产生积极贡献，而将三种策略组合使用能够实现最佳效果，提升了模型的整体检测能力，验证了所提出策略的有效性。

表 3.5 Laboro 番茄数据集各模块的消融实验结果

MS-FFRM	CA-D	HSD	AP	AP_{50}	AP_{75}	AP_S
			71.1	82.7	77.8	86.4
✓			75.0	86.7	81.0	88.5
	✓		75.2	86.4	81.2	88.1
		✓	75.2	85.9	81.2	88.7
✓	✓		75.0	85.9	80.8	88.4
	✓	✓	75.3	86.0	81.3	88.4
✓		✓	74.6	86.0	80.9	88.5
✓	✓	✓	75.4	86.7	81.9	88.3

协同注意力解码器的有效性。为了验证协同注意力解码器（Co-Attention Decoder）的有效性，本章进行了消融实验，结果如表 3.6 所示。

采用标准的 RT-DETR 解码器作为基线进行对比。实验结果表明，单独移除 C-Decoder 或 SC-Decoder 都会导致模型性能下降，降低了模型对特征的关注能力以及特征提取的精度，而完整的协同注意力解码器（结合 SC-Decoder 与 C-Decoder）则能够实现最佳的检测性能，并有效提升模型的检测精度。从消融实验的结果可以看出，C-Decoder 和 SC-Decoder 在提升检测性能方面均发挥了重要作用，其中 C-Decoder

表 3.6 协同注意力解码器消融实验结果

Method	AP	AP ₅₀	AP ₇₅	AP _S
Baseline Decoder	71.1	82.7	77.8	86.4
Co-Attention Decoder w/o C-Decoder	74.7	86.0	80.6	87.9
Co-Attention Decoder w/o SC-Decoder	73.0	84.0	78.9	87.8
Co-Attention Decoder (full)	75.2	86.4	81.2	88.1

强化目标相关特征表示，增强每个查询对目标特征的关注能力；SC-Decoder 则促进特征之间的信息交互，提高重要特征的选择能力。

层级自蒸馏策略的有效性。为了评估层级自蒸馏策略的作用，本章进行了消融实验，结果如表 3.7 所示。实验结果表明，单独使用自蒸馏（Self-Distillation）和层级蒸馏（Hierarchical Distillation）分别可将 AP 提升至 74.7%（+3.6%）和 75.0%（+3.9%），两者均对检测性能提升具有积极作用。当两种策略结合使用时，模型取得了最佳检测性能，AP 达到 75.2%，展现了优异的精度表现。同时，模型参数保持在 43M，未因引入层级蒸馏而增加，且推理速度为 22 FPS，表明该策略仅在训练阶段增加计算开销，而不会影响推理效率。

表 3.7 层级自蒸馏策略消融实验结果

Self-Distillation	Hierarchical Distillation	AP	AP ₅₀	AP ₇₅	AP _S	Params (M)	FPS
		71.1	82.7	77.8	86.4	42	20
✓		74.7	85.6	80.9	88.5	43	21
	✓	75.0	85.8	80.8	88.7	43	22
✓	✓	75.2	85.9	81.2	88.7	43	22

为了进一步验证层级自蒸馏策略的有效性，本章对特征图进行了可视化分析，如图 3.7 所示。图 (a) 代表输入图像，图 (b) 代表未进行蒸馏的特征图，图 (c) 代表使用自蒸馏的特征图，图 (d) 代表使用层级自蒸馏的特征图。相较于单独采用自蒸馏，自蒸馏与层级蒸馏的组合能生成更清晰、更精准的特征表示。通过采用层级自蒸馏，模型特征图能更准确地定位果实本身而非背景噪声，该策略显著增强了模型特征学习的鲁棒性，使模型在复杂场景下对目标的识别与定位更加可靠。

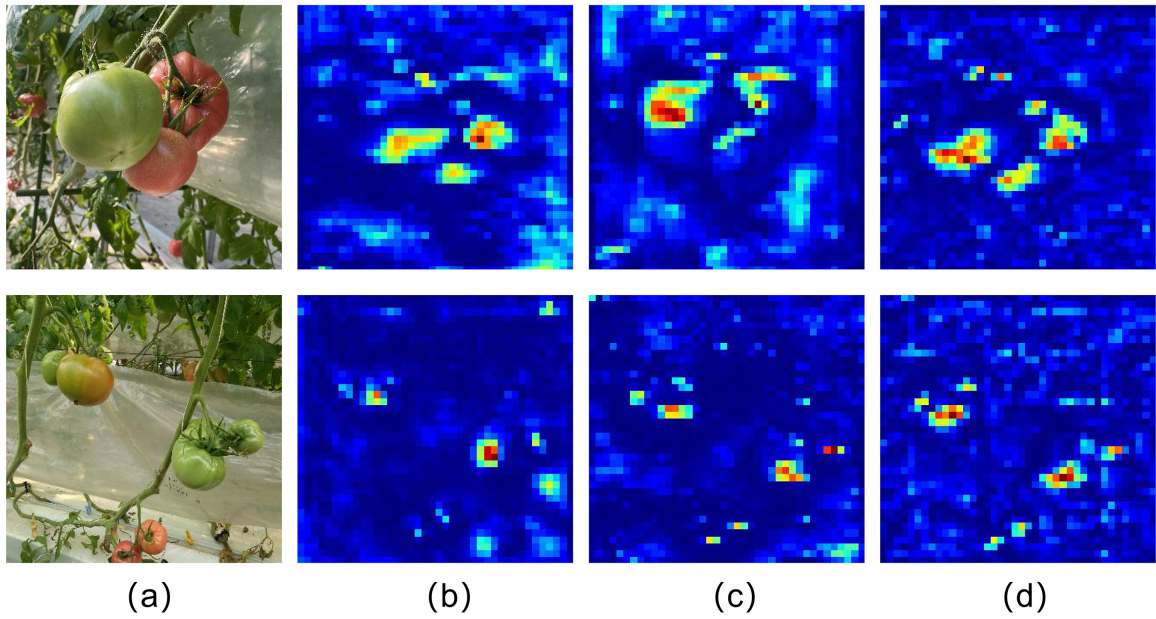


图 3.7 层级自蒸馏特征图可视化对比

温度参数 T 的影响。为了验证解码器中层级自蒸馏策略中温度参数 T 的作用，我们进行了消融实验，结果如表 3.8 所示。

表 3.8 层级自蒸馏中温度参数 T 对模型性能的影响

T	AP	AP₅₀	AP₇₅	AP_S
2	74.3	85.4	80.3	88.3
4	74.0	84.4	80.2	88.8
5	75.2	85.9	81.2	88.7
6	73.5	84.8	79.2	87.7
10	73.4	84.4	79.0	88.4

实验结果表明，将 T 提高到适当水平可以增强学生模型从教师模型中提取知识的能力，使学生模型能够更好地捕捉类别间的相似性。当 $T = 5.0$ 时，评估指标达到最佳性能。然而，温度过高可能导致输出过度平滑，使学生模型对细节的敏感性下降，从而对性能产生负面影响。因此，设置 $T = 5.0$ 更有利于提升模型性能。

多尺度特征融合重加权模块的有效性。为验证多尺度特征融合重加权模块（MS-FFRM）在特征融合中的有效性，对特征图进行了可视化分析，如图 3.8 所示。图 (a) 代表输入图像，图 (b) 代表未使用 MS-FFRM 模块的特征图，图 (c) 代表使用 MS-FFRM 模块的特征图。

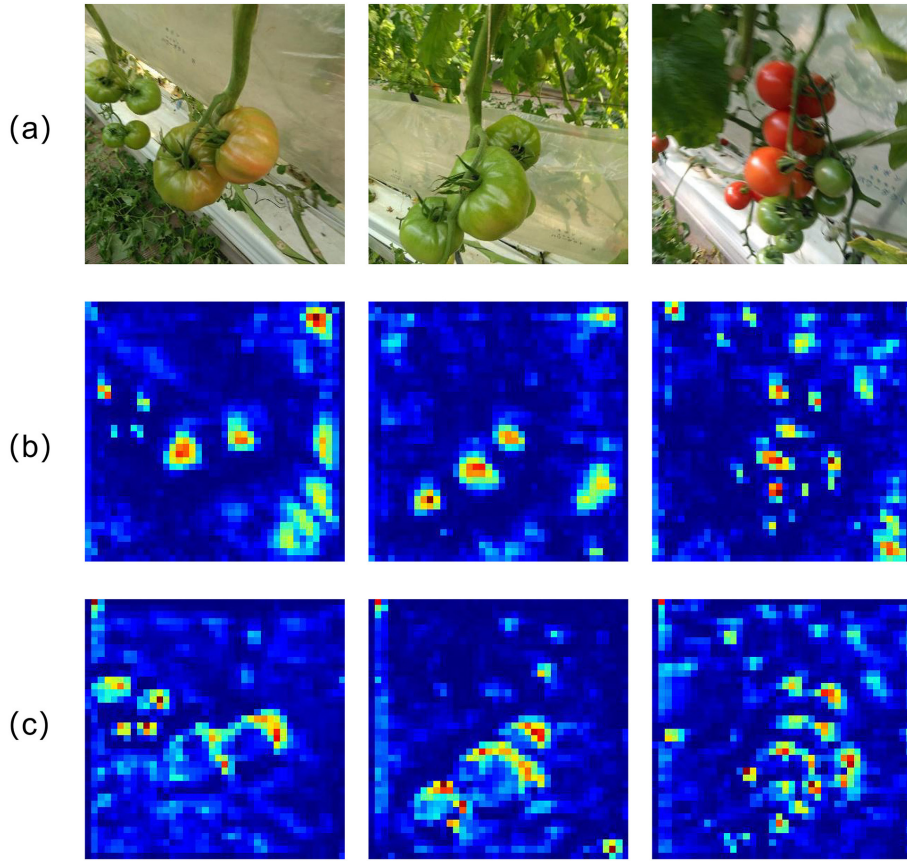


图 3.8 MS-FFRM 特征图可视化对比

如图 3.8 (b) 和 (c) 所示, 未使用 MS-FFRM 模块的特征融合方法在目标区域的关注不够准确, 容易受到背景干扰, 从而导致特征定位偏差。相比之下, 加入 MS-FFRM 模块后, 对目标果实的关注度显著增强, 同时有效抑制了背景区域的无关特征。这些结果表明, MS-FFRM 模块提升了对关键特征的检测能力, 从而提高了果实成熟度检测的性能。

3.3 本章小结

针对复杂环境下果实成熟度检测困难以及现有方法对多尺度信息的处理能力不足等挑战, 本章提出了目标检测模型 FMD-DETR。针对果实在尺度、遮挡和重叠方面的复杂性, 提出多尺度特征融合重加权模块, 通过对不同尺度特征进行自适应重加权, 提升模型对各类尺寸果实的检测能力。为进一步优化目标查询交互, 构建协同注意力解码器, 将交叉注意力与自注意力机制结合, 增强解码器对遮挡和重叠目标的处理能力。此外, 设计层级自蒸馏策略, 实现解码器不同层之间的知识传递, 提

高空间感知与定位鲁棒性。

在实验验证方面，FMD-DETR 在 Laboro 番茄数据集、真实场景下的番茄数据集、草莓数据集以及 FruitRipeness 数据集上均取得了显著的性能提升，平均精度（AP）分别达到 75.4%、52.7%、41.5% 和 87.1%，并且在精度与效率之间实现了良好的平衡。定性分析表明，FMD-DETR 在复杂环境下能够更准确地定位和识别不同成熟度的果实，展现出较强的鲁棒性和适应性。消融实验和可视化结果进一步验证了所提模块的有效性，展示了所提模块对整体性能提升的贡献。

本章模型的提出不仅为果实成熟度检测提供了一种新的解决方案，还为其他农产品智能采摘机器人的视觉识别、目标定位以及分级等工作提供了有效的支持。

第四章 基于动态稀疏混合专家机制的农业果实目标检测方法

在上一章中,针对复杂环境下果实成熟度检测过程中光照变化、背景干扰、遮挡与重叠目标等问题,提出了 FMD-DETR 模型,但其在小目标检测上还存在进一步提升空间。此外,在农业果实成熟度检测任务中,复杂场景下的果实图像特征存在显著的视觉差异,现有模型通常仅采用单一的参数共享结构,难以适应不同视觉模式下的特征变化,模型在针对性建模、自适应处理方面仍存在进一步提升空间。本章以提升小目标检测性能以及增强模型自适应能力为目标,提出一种面向农业果实成熟度检测的动态稀疏混合专家检测框架。该模型通过设计的高频细节增强模块,将浅层细节与深层语义相结合,使其更好地关注目标特征,以提升小目标的检测能力;构建动态稀疏混合专家解码器,赋予模型动态建模与专家分工能力,使其根据输入特征特点,自动选择最合适的专家组合参与推理;提出负载均衡与专家多样性损失策略,鼓励不同专家学习到不同的特征表示。本方法验证了混合专家机制在农业目标检测任务中的可行性,为应对复杂农业场景中的多样性挑战提供了有效方案。

4.1 方法概述

4.1.1 模型框架

本章提出一种面向农业果实成熟度检测的动态稀疏混合专家检测模型 AgriMoE-DETR (Agricultural Mixture-of-Experts-DEtection TRansformer)。该模型由三个核心模块构成:高频细节增强模块 (High-frequency Detail Enhancement Module, HDEM)、基于尺度先验的动态稀疏混合专家解码器 (Dynamic Sparse Mixture-of-Experts Decoder, DS-MoE Decoder) 和负载均衡与专家多样性损失策略 (Load Balancing & Expert Diversity Loss)。高频细节增强模块利用高频细节信息来增强深层特征,提升对关键细节特征的关注,从而提升对小目标的检测能力。动态稀疏混合专家解码器通过混合专家机制实现对特征的自适应处理,以提升对多尺度目标的适应能力。负载均衡与专家多样性损失策略则进一步缓解混合专家模型负载不均衡的问题,并提升对不同特征的学习能力。

AgriMoE-DETR 的框架图如4.1所示。首先,输入的果实图像经过主干网络 ResNet 提取多尺度特征 S3、S4、S5。其中, S5 特征作为高层语义特征,送入 Transformer 编码器进行编码处理,以提升深层特征的语义表达能力。随后,编码器输出的特征 F5 与 S3、S4 一同输入到高频细节增强模块 (HDEM),利用浅层特征 S3 中的高频结构信息作为细节补充,对深层特征 S4 和 F5 进行增强,分别得到增强后的特征 E4 和 E5。其中, S3 保持原始浅层细节表示, E4 和 E5 获得细节信息更加丰富的特征表示,用于后续多尺度特征融合。在此基础上,通过跨层特征融合过程,使浅层细节信息、中层语义信息和深层强语义信息得到充分融合,从而增强多尺度特征的表达能力。然后,融合后的多尺度特征序列输入到动态稀疏混合专家解码器 (DS-MoE Decoder) 中。解码器在每个解码层采用的尺度感知的专家选择机制,能够使不同尺度的目标特征自适应地选择更合适的专家进行处理。最后,解码器输出结果送入检测头完成分类与边界框回归,输出果实目标位置及其成熟度类别,得到检测结果。

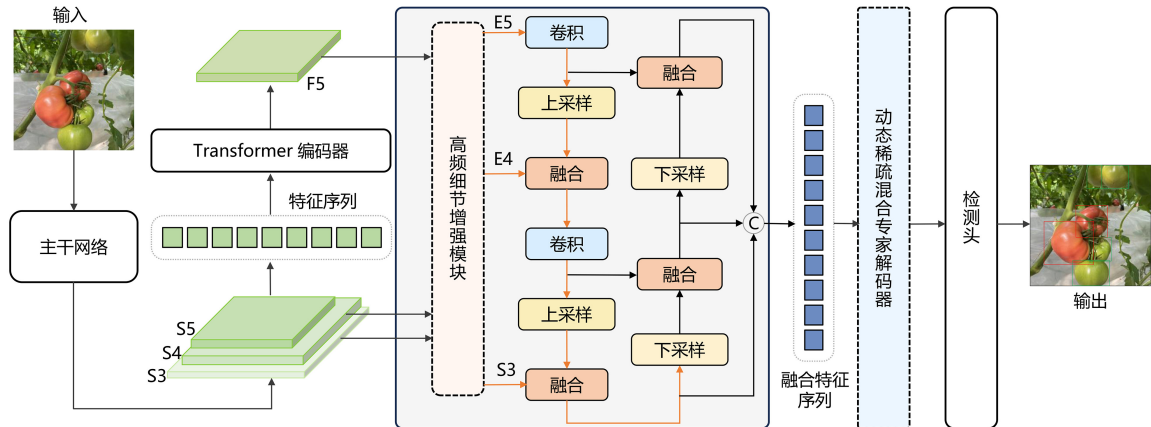


图 4.1 AgriMoE-DETR 模型框架图

4.1.2 高频细节增强模块

在目标检测任务中,浅层特征通常保留了较高的空间分辨率,但其响应往往受到大量低频背景纹理和冗余结构的干扰,而真正对目标定位起关键作用的是目标边缘、轮廓等高频结构信息。另一方面,深层特征虽然具备更强的语义区分能力,但其空间细节因多次下采样步骤而严重退化,这直接限制了目标的定位精度。因此,如何有效地从浅层特征中提取高频细节,并将其整合到深层语义特征中,已成为提升复杂背景下目标检测性能的关键问题。为解决这一问题,本章提出了高频细节增强

模块（High-frequency Detail Enhancement Module, HDEM），该模块从相对浅层特征中提取高频结构信息，并将其作为细节增强信息融入到更深层的特征中，以弥补深层特征中细节的不足。

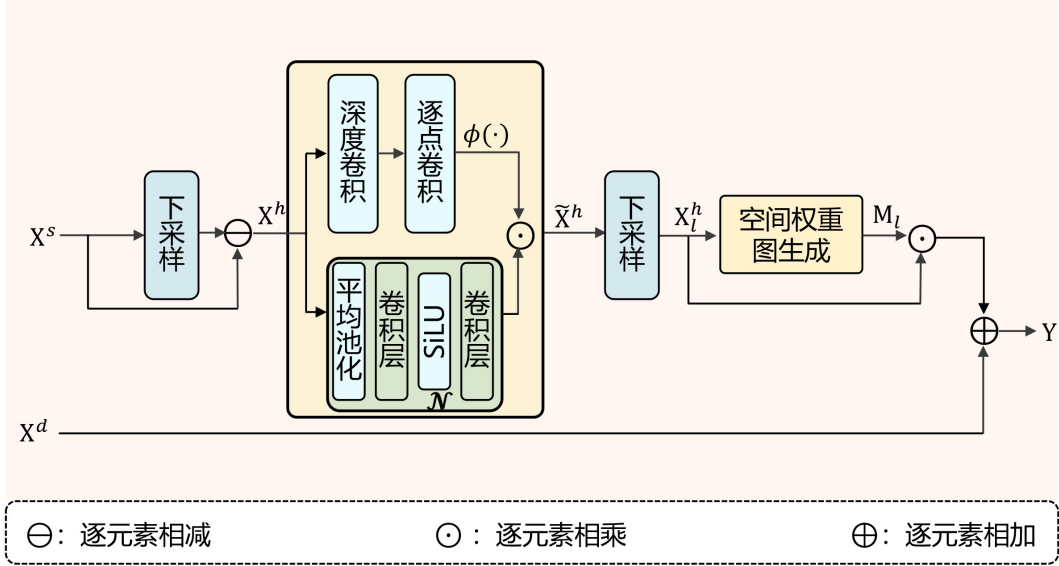


图 4.2 高频细节增强模块（HDEM）示意图

高频细节增强模块（HDEM）如图4.2所示。经过主干网络提取的特征 S3、S4 和 F5 输入到 HDEM 中，将浅层特征 S3 表示为 $\mathbf{X}^s \in \mathbb{R}^{B \times C \times H \times W}$ ，深层特征 S4 或 F5 表示为 $\mathbf{X}^d \in \mathbb{R}^{B \times C \times H' \times W'}$ ，其中 $H' < H, W' < W$ 。该模块首先将浅层特征 $\mathbf{X}^s$ 通过高频分解获得高频细节信息^[93]，得到对应的高频特征 $\mathbf{X}^h$ ，该过程能够突出目标边缘和纹理等关键信息，同时抑制低频背景干扰， $\mathbf{X}^h$ 的计算如公式4.1所示：

$$\mathbf{X}^h = \mathbf{X}^s - \text{AvgPool}_{3 \times 3}(\mathbf{X}^s), \quad (4.1)$$

随后，通过噪声抑制模块和轻量卷积模块对高频特征 $\mathbf{X}^h$ 进行增强，从而突出对目标定位更为关键的结构信息，并提高特征表达能力。增强后的高频特征 $\tilde{\mathbf{X}}^h$ 定义如公式4.2所示：

$$\tilde{\mathbf{X}}^h = \phi\left(\text{PW}(\text{DW}(\mathbf{X}^h))\right) \odot \mathcal{N}(\mathbf{X}^h), \quad (4.2)$$

其中， $\text{DW}(\cdot)$ 和 $\text{PW}(\cdot)$ 分别表示深度卷积和逐点卷积， $\mathcal{N}(\cdot)$ 代表噪声抑制模块， $\phi(\cdot)$ 为非线性激活函数。接下来，对增强后的高频特征 $\tilde{\mathbf{X}}^h$ 进行尺度对齐，以便将其有效

地整合到不同尺度的语义特征中。该过程通过平均池化操作对高频特征进行下采样，从而获得与深层特征相匹配的尺度表示 $\mathbf{X}_l^h$ ：

$$\mathbf{X}_l^h = \text{AvgPool}_{2 \times 2}(\tilde{\mathbf{X}}^h), \quad (4.3)$$

为了突出响应强度较高的区域，构建空间权重图对高频特征进行加权，空间权重图 $\mathbf{M}_l$ 的计算如公式4.4所示：

$$\mathbf{M}_l = \sigma \left(\frac{1}{C} \sum_{c=1}^C |\mathbf{X}_{l,c}^h| \right), \quad (4.4)$$

其中， C 表示特征图的通道数， $\mathbf{X}_{l,c}^h$ 表示 $\mathbf{X}_l^h$ 在第 c 个通道上的特征图， $\sigma(\cdot)$ 为 sigmoid 激活函数。最终，将加权后的高频特征与深层特征进行融合，得到增强后的输出特征 $\mathbf{Y} \in \mathbb{R}^{B \times C \times H' \times W'}$ ：

$$\mathbf{Y} = \mathbf{X}^d + \alpha \cdot (\mathbf{M}_l \odot \mathbf{X}_l^h), \quad (4.5)$$

其中， α 为可学习的系数， $\odot$ 表示逐元素相乘。

本章提出的高频细节增强模块（HDEM）打破了浅层细节与深层语义之间的融合壁垒，旨在增强模型捕捉目标边界和纹理细节的能力，以增强模型在复杂背景下对小目标的检测和定位能力。

4.1.3 动态稀疏混合专家解码器

为增强模型对农业场景中显著尺度变化目标的检测能力，在专家路由阶段利用多尺度先验信息进行条件约束，提出尺度感知的稀疏混合专家解码器（DS-MoE Decoder）模块。该模块利用不同特征层级的响应强度来推断每个特征对应的主导尺度，将其映射为可学习的尺度嵌入向量，作为门控网络的输入，从而实现了一种具有尺度感知能力的专家选择机制。该设计使得不同尺度的目标特征能够自适应地路由至更合适的专家进行处理，促进专家在不同空间尺度和语义层次上的分工协作。具体而言，为不同尺度的特征选择不同的专家，对于小尺度特征，选择局部纹理专家；而对于大尺度特征，优先选择高层语义专家。

除了尺度感知的专家选择，本模块还通过稀疏激活推理机制，以进一步提升计算效率与模型表达能力。具体来说，对于每个特征向量，仅激活 Top-K 个最相关的专家，从而实现动态稀疏计算。该机制将计算资源集中于最相关的专家上，避免冗

余计算，同时通过多个专家的加权组合实现隐式模型集成，提升整体特征表达能力和模型泛化性能。结合多尺度先验信息和稀疏激活策略，解码器能够在复杂农业场景中有效应对目标尺度差异，兼顾检测精度与计算效率，实现专家之间的协同工作与动态分工。

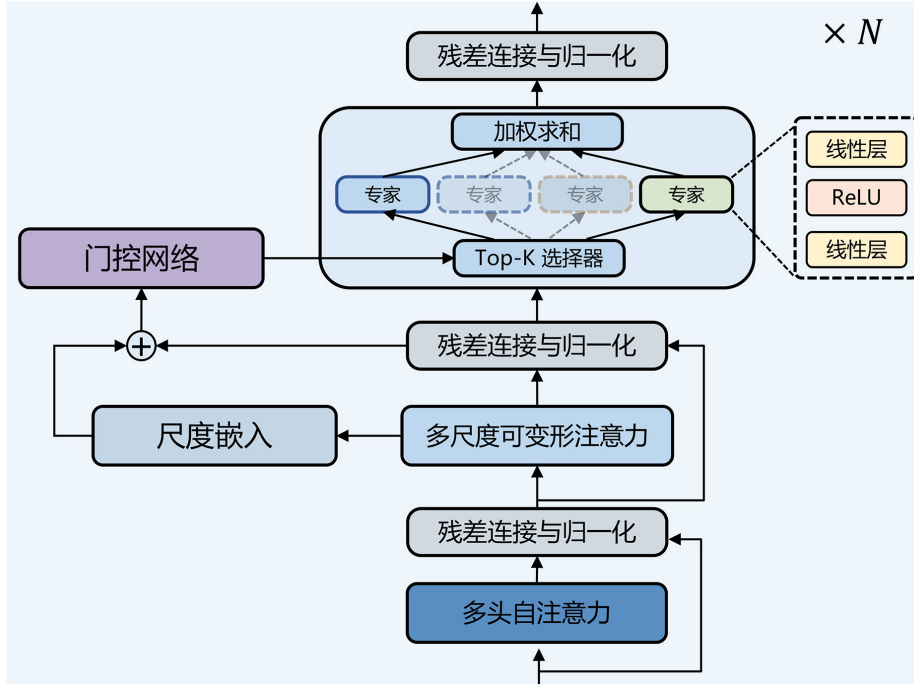


图 4.3 动态稀疏混合专家解码器（DS-MoE Decoder）模块示意图

动态稀疏混合专家解码器（DS-MoE Decoder）的结构如图4.3所示。传统的门控网络通常仅依赖查询向量的局部语义特征来进行专家路由决策，缺乏对输入特征在不同空间尺度上的响应强度的考虑，导致专家选择缺乏针对性和适应性。而动态稀疏混合专家解码器利用多尺度先验信息作为条件信息，以此指导门控网络进行尺度感知的专家路由决策。在解码器层中，设目标查询的序列长度为 N ，提取序列中的第 i 个查询向量为 $x_i \in \mathbb{R}^D$ ，其中 $i = 1, \dots, N$ ， D 为特征维度。多尺度先验信息的获取，需要得到不同尺度层的特征响应强度。具体而言，首先，利用解码器层中的多尺度可变形注意力（Multi-Scale Deformable Attention）机制，通过线性映射预测采样点注意力权重 $A_i \in \mathbb{R}^{H \times L \times P}$ ^[42]，如公式4.6所示：

$$A_i = \text{softmax}((W_{\text{att}}x_i + b_{\text{att}})_{h,l,p}), \quad (4.6)$$

其中， $W_{\text{att}} \in \mathbb{R}^{(HLP) \times D}$ 为可学习的权重矩阵， $b_{\text{att}} \in \mathbb{R}^{HLP}$ 为可学习的偏置向量。

$h = 1, \dots, H$ 为注意力头索引, $l = 1, \dots, L$ 为尺度层索引, $p = 1, \dots, P$ 为采样点索引。softmax 操作对每个注意力头 h , 在所有尺度层 l 和在采样点维度 p 上进行归一化。

每个尺度上有多个采样点的响应。接下来, 为了获得该尺度的整体响应强度, 将同一尺度层内的权重进行聚合, 并在多头之间取平均。将第 l 个尺度层的响应强度定义为:

$$\alpha_i^{(l)} = \frac{1}{H} \sum_{h=1}^H \sum_{p=1}^P A_{i,h,p}, \quad l = 1, \dots, L. \quad (4.7)$$

然后将每个尺度的响应组成向量, 由此得到尺度响应向量 $\alpha_i = [\alpha_i^{(1)}, \dots, \alpha_i^{(L)}] \in \mathbb{R}^L$, 并且满足 $\sum_{l=1}^L \alpha_i^{(l)} = 1$ 。此时, 从尺度响应向量中选出响应强度最大的尺度索引, 作为主导尺度索引, 如公式4.8所示:

$$\ell_i = \arg \max_l \alpha_i^{(l)}, \quad (4.8)$$

接下来, 定义一个可学习的尺度嵌入矩阵 $S \in \mathbb{R}^{L \times D}$, 每一行是一个尺度的嵌入向量。定义 $s^{(l)} \in \mathbb{R}^D$ 为第 l 个尺度层的嵌入向量, 则主导尺度索引对应的嵌入向量为 $s^{(\ell_i)}$ 。最终, 将尺度嵌入向量与原始查询向量相结合, 构造出输入门控网络的尺度感知查询特征 $\tilde{x}_i \in \mathbb{R}^D$, 如公式4.9所示:

$$\tilde{x}_i = x_i + s^{(\ell_i)}, \quad (4.9)$$

在获得尺度感知的查询特征 $\tilde{x}_i$ 后, 将其输入到门控网络中。对于每个查询向量 $\tilde{x}_i$, 门控网络通过 softmax 函数计算其对应的专家激活概率分布 g_i , g_i 定义如公式4.10所示:

$$g_i = \text{softmax}(W_g \tilde{x}_i), \quad (4.10)$$

其中, $W_g \in \mathbb{R}^{M \times D}$ 为可学习参数矩阵。上面的步骤使得门控网络能够根据输入特征在不同尺度上的响应强度, 动态选择更合适的专家, 从而实现了具有尺度感知能力的专家选择机制。

设混合专家模块包含 M 个专家 $\{E_k(\cdot)\}_{k=1}^M$ 。为降低计算开销并促进专家分工, 仅选择概率最大的 K 个专家参与计算, 并对这 K 个激活权重进行重归一化。第 k 个

专家的激活权重 $\hat{g}_{i,k}$ 定义如公式4.11所示：

$$\hat{g}_{i,k} = \begin{cases} \frac{g_{i,k}}{\sum_{j \in \text{TopK}(g_i, K)} g_{i,j}}, & k \in \text{TopK}(g_i, K), \\ 0, & \text{otherwise.} \end{cases} \quad (4.11)$$

最终，第 i 个查询向量的输出由被激活专家的加权组合得到，如公式4.12所示：

$$y_i = \sum_{k=1}^M \hat{g}_{i,k} E_k(x_i), \quad (4.12)$$

其中，每个专家的输出满足 $E_k(x_i) \in \mathbb{R}^D$ 。专家网络的输入仍采用原始查询特征 x_i ，将专家路由决策与特征表示学习进行解耦，使尺度先验信息仅影响专家分配过程，而不直接干扰特征表达。

本章提出的动态稀疏混合专家解码器，通过多尺度特征响应推导尺度先验作为条件信息约束专家路由，使不同空间尺度的特征能够被分配至更匹配的专家进行建模。同时，根据任务的需要动态选择激活的专家模型，实现对输入数据的灵活处理。此外，由于每次前向计算仅激活少数专家，大部分专家处于未激活状态，从而降低计算开销以提升模型的运行效率。

4.1.4 负载均衡与专家多样性损失策略

针对专家模型中可能出现的专家激活不均衡与表示同质化问题，本章设计了由负载均衡损失函数（Load Balancing Loss）与专家多样性损失函数（Expert Diversity Loss）组成的损失结构：其中，负载均衡损失函数用于鼓励不同专家被均衡利用；专家多样性损失函数则鼓励不同专家学习到不同的特征表示。

设共有 E 个专家，对于输入特征 $\mathbf{X}$ ，门控网络会将特征单元分配给不同专家。令 $\text{load}_i(\mathbf{X})$ 表示第 i 个专家在 $\mathbf{X}$ 的负载，则 $\text{load}(\mathbf{X})$ 定义为：

$$\text{load}(\mathbf{X}) = \{\text{load}_i(\mathbf{X})\}_{i=1}^E, \quad (4.13)$$

使用标准差来衡量各专家使用率的平衡程度，并使用均值对其进行归一化，通过平方增强对不均衡的惩罚，定义负载均衡损失为：

$$\mathcal{L}_{\text{load}}(\mathbf{X}) = \left(\frac{\text{std}(\text{load}(\mathbf{X}))}{\text{mean}(\text{load}(\mathbf{X}))} \right)^2, \quad (4.14)$$

由于不同专家可能学习到高度相关的表示，模型难以充分发挥多专家结构的优势。因此，通过多样性约束以鼓励专家学习具有区分性的特征表示。设 $\mathbf{o}_i$ 表示第 i 个专家的输出特征表示。计算所有专家对 (i, j) （其中 $i < j$ ）之间的相似度，并对其求和取平均，定义专家多样性损失为：

$$\mathcal{L}_{\text{div}} = \left(\binom{E}{2} \right)^{-1} \sum_{1 \leq i < j \leq E} \text{sim}(\mathbf{o}_i, \mathbf{o}_j), \quad (4.15)$$

其中 $\text{sim}(\cdot, \cdot)$ 表示余弦相似度函数， $\binom{E}{2}$ 为专家对的总数。

综合考虑上述两项约束，为了在确保检测性能的同时平衡专家利用率与特征多样性，最终的总损失函数定义为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}_{\text{load}} + \lambda_2 \mathcal{L}_{\text{div}}. \quad (4.16)$$

其中， $\mathcal{L}_{\text{det}}$ 表示目标检测任务的损失， λ_1, λ_2 表示权重系数，分别用于对负载均衡损失 $\mathcal{L}_{\text{load}}$ 与专家多样性损失 $\mathcal{L}_{\text{div}}$ 加权。

本章所提出的负载均衡与专家多样性损失策略，是为了促进专家之间的均衡利用，同时让专家学习到差异化的特征表示，以提升模型在复杂农业场景中的鲁棒性与泛化能力。

4.2 实验设计

4.2.1 实验环境与参数设置

（1）数据集

本章选取了四个数据集进行模型训练与验证：**Laboro** 番茄数据集^[85]、草莓数据集^[86]、真实世界番茄数据集以及 **FruitRipeness** 数据集^[87]。这些数据集涵盖了不同种类果实及其不同成熟阶段，能够较好地反映果实成熟度检测任务中的类别差异和外观变化。同时，数据集中包含遮挡、背景相似、目标尺度不一致、光照变化等多种复杂场景，有助于验证所提模型在真实农业环境下的检测性能和泛化能力。

（2）实验环境与参数设置

为保证与第三章实验具有可比性，本章在通用训练配置上与 **FMD-DETR** 保持一致。本章实验使用 **RT-DETR** 作为基础模型实现，采用 **ResNet-50** 作为主干网络，初始权重是 **ImageNet** 上预训练的结果。所有模型在 **PyTorch** 框架下实现，并在 **NVIDIA**

RTX 4090 GPU (24GB 显存) 上训练和测试。模型训练过程中使用 AdamW 优化器进行参数更新, 初始学习率设置为 0.0001, 批处理大小为 4, 权重衰减系数设为 1×10^{-4} 。训练总周期为 100 个 epoch, 以充分保证模型的收敛性与稳定性。此外, 本章的 MoE 模型专家数量设置为 4^[78], 即每一层混合专家模块包含 4 个专家分支。对于每个查询, 仅激活概率最高的 Top-2 个专家参与前向计算, 以在检测性能与计算效率之间取得平衡。

(3) 评价指标

在模型评估阶段, 采用平均精度 (Average Precision, AP) 作为主要评价指标, 用于综合衡量模型在目标检测任务中的整体精度。为了更全面地反映模型的检测能力, 同时报告在不同交并比 (IoU) 阈值下的性能指标, 包括 AP₅₀ (IoU=0.50) 和 AP₇₅ (IoU=0.75)。此外, 为进一步评估模型对不同尺度目标的检测能力, 还统计了小目标 (AP_S)、中等目标 (AP_M) 和大目标 (AP_L) 的 AP 值, 从而全面衡量模型在不同场景和目标尺度下的检测性能。

除了检测精度指标外, 本研究还关注模型的计算复杂度与实际运行效率, 具体包括模型参数量 (Params)、浮点运算量 (GFLOPs)、以及推理速度 (Frames Per Second, FPS)。参数量反映了模型的存储需求和网络规模, GFLOPs 衡量模型在单次前向推理中所需的计算量, 而 FPS 用于评估模型在实际场景下的实时推理性能。通过这些指标能够更全面地评估所提出模型在精度、计算成本和实际应用可行性方面的综合性能, 为农业果实成熟度检测任务的高效部署提供参考依据。

4.2.2 定量实验

表4.1展示了在 Laboro 番茄数据集上, 所提算法与其他目标检测算法以及上一章 FMD-DETR 算法的定量对比结果。从表4.1的结果可见, 本章算法在应对番茄成熟度检测任务时, 该算法平均精度 (AP) 达到 76.1%, 超越了其他目标检测算法, 比基线方法 RT-DETR 提升约 5.0 个百分点, 同时优于 YOLO11s 等当前主流方法, 表明所提方法在复杂农业场景中具有更强的检测能力。较 FMD-DETR 提升了 0.7%, 在 AP₅₀、AP₇₅、AP_S 分别提升了 0.4%、0.5%、0.8%。本章方法进一步提升了检测精度, 特别是在小目标 (AP_S) 上的提升最为显著, AP_S 平均精度达到 89.1%, 充分验证了该方法在复杂农业场景下对小目标检测的有效性。以上结果表明, 高频细节增强模块通过将浅层细节与深层语义相结合, 缓解了深层特征细节不足的问题, 提高了对小目

表 4.1 Laboro 番茄数据集上的模型定量比较

Model	Backbone	Params(M)	GFLOPs	FPS	AP	AP ₅₀	AP ₇₅	AP _S
Deformable-DETR ^[42]	R50	40	165	18	46.7	71.3	50.8	64.4
Conditional-DETR ^[43]	R50	44	86	14	68.7	<u>87.7</u>	77.8	77.4
DAB-DETR ^[44]	R50	48	87	18	68.2	86.3	76.5	79.9
DINO ^[47]	R50	48	235	18	74.8	87.3	81.7	86.9
Co-DETR ^[88]	R50	64	-	18	70.9	85.2	77.9	85.5
Rank-DETR ^[89]	R50	49	226	15	74.5	88.2	81.5	85.5
RT-DETR ^[12]	R50	42	66	20	71.1	82.7	77.8	86.4
Mobile-DETR ^[12]	MobileNet	15	18	20	60.4	76.0	66.6	77.5
RT-DETRv2 ^[48]	R50	42	66	20	70.5	81.7	76.8	87.0
Lite-DETR ^[90]	R50	47	157	10	68.5	83.5	75.0	85.0
YOLOv8s ^[35]	-	11.2	29	10	63.4	76.1	68.8	64.6
YOLOv8x ^[35]	-	68	257	10	72.9	82.1	78.7	73.9
YOLOv10s ^[36]	-	7.2	22	8	73.9	85.8	80.1	75.1
YOLO11s ^[37]	-	9.4	22	8	75.4	87.2	81.2	76.5
EdgeYOLO ^[92]	-	5.8	27	32	50.1	75.0	57.9	65.3
FMD-DETR	R50	44	66	22	<u>75.4</u>	86.7	<u>81.9</u>	<u>88.3</u>
Ours	R50	43	66	22	76.1	87.1	82.4	89.1

标的定位和检测能力。同时，动态稀疏混合专家解码器通过尺度感知的专家选择机制，提升了模型对不同尺度目标的适应能力，从而进一步增强了检测性能。此外，本章算法在提升精度的同时，并没有增加额外的负担，且模型参数量（Params）从 FMD-DETR 的 44M 降低至 43M；本章提出的动态稀疏专家解码器，有效解决了混合专家增加模型参数量的问题。同时，计算量（GFLOPs）保持不变，推理速度（FPS）达到了 22 FPS，体现了效率与性能之间的平衡。与其他目标检测算法及 FMD-DETR 相比，本章模型在参数量进一步压缩的前提下，全面刷新了 Laboro 番茄数据集上的各项检测精度指标，特别是增强了小目标番茄的检测能力，实现了在复杂农业场景下的高效检测。

表4.2展示了所提算法与其他目标检测算法以及 FMD-DETR 算法在真实番茄数据集上的定量对比结果。由表中结果可知，本章方法的平均精度（AP）达到了 54.1%，在真实场景下的番茄数据集上表现出了优越的检测能力，超越了其他对比算法。与 RT-DETR 相比，AP 提升了 4.3%，与 FMD-DETR 相比，AP 提升了 1.4%。同时，AP₅₀（+1.5%）、AP₇₅（+0.8%）、AP_S（+1.3%）、AP_M（+3.5%）及 AP_L（+0.7%）均有所提升，以上结果表明，所提出的基于尺度先验的动态稀疏混合专家解码器，利用多尺度信

表 4.2 真实场景番茄数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Deformable-DETR ^[42]	R50	11	37.9	66.2	37.9	54.2	33.8	44.3
Conditional-DETR ^[43]	R50	11	50.9	83.6	57.5	59.5	46.9	55.2
DAB-DETR ^[44]	R50	11	50.5	82.2	56.9	41.0	<u>47.6</u>	54.9
DINO ^[47]	R50	11	46.7	73.5	54.5	63.8	45.3	50.3
Co-DETR ^[88]	R50	11	49.5	78.5	57.8	71.0	45.8	54.7
Rank-DETR ^[89]	R50	15	49.0	79.2	54.6	63.8	45.9	53.8
RT-DETR ^[12]	R50	14	49.8	80.9	57.7	55.1	44.0	54.0
Mobile-DETR ^[12]	MobileNet	12	48.3	79.5	51.9	58.5	45.4	52.0
RT-DETRv2 ^[48]	R50	14	50.0	78.0	57.5	<u>67.7</u>	44.7	55.4
Lite-DETR ^[90]	R50	11	48.5	75.3	58.5	64.3	46.9	51.8
YOLOv10s ^[36]	-	7	49.0	75.3	59.3	64.4	46.5	53.2
EdgeYOLO ^[92]	-	21	49.6	81.0	57.1	47.7	47.2	51.9
FMD-DETR	R50	16	<u>52.7</u>	81.0	<u>63.3</u>	64.7	46.1	<u>57.3</u>
Ours	R50	20	54.1	<u>82.5</u>	64.1	66.0	49.6	58.0

息和动态稀疏计算，使得不同尺度的目标特征能够自适应选择合适的专家进行处理，增强了对多尺度目标的检测能力。小目标的提升得益于高频细节增强模块，该模块使模型能够更有效地捕捉小尺寸目标的边缘与纹理特征，从而提高检测精度。在推理效率方面，本章方法实现了 20 FPS 的推理速度，优于大多数 DETR 系列方法，体现了较好的实时性能，表明其在提升检测性能的同时保持了较高的推理效率。真实场景中的番茄检测任务通常面临光照变化、背景复杂、目标遮挡及尺度变化等问题。实验结果表明，所提出的高频细节增强模块、动态稀疏混合专家解码器以及负载均衡与专家多样性损失策略增强了模型对复杂环境的适应能力，提高检测的鲁棒性与泛化性能。

为了进一步验证本章所提算法的泛化能力以及在不同果实场景下的检测鲁棒性，表4.3展示了所提算法与其他目标检测算法以及 FMD-DETR 算法在草莓数据集上的定量对比结果。如表4.3所示，所提算法在草莓数据集上具有良好的检测性能。本章算法的平均精度（AP）达到了 42.1%，依然高于其他对比算法。相较于 FMD-DETR 提升了 0.6%，同时，在 AP₅₀ 和 AP₇₅ 两个关键指标上，本章算法分别达到了 75.9% 和 44.8%，分别提升了 2.0% 和 0.6%，表明模型改善了对于因重叠或遮挡而容易漏检的目标的检测。衡量小目标精度的 AP_S 指标提升了 1.3%，进一步验证了本章算法对

表 4.3 草莓数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Deformable-DETR ^[42]	R50	11	26.1	49.9	26.5	0.8	20.4	29.6
Conditional-DETR ^[43]	R50	11	36.1	71.2	33.5	10.2	35.9	38.2
DAB-DETR ^[44]	R50	12	38.3	70.9	39.1	5.7	39.4	39.9
DINO ^[47]	R50	12	35.1	62.0	38.8	17.2	34.2	37.5
Co-DETR ^[88]	R50	12	35.4	61.0	40.9	24.9	38.2	36.2
Rank-DETR ^[89]	R50	15	36.7	68.2	39.4	<u>24.3</u>	39.6	37.3
RT-DETR ^[12]	R50	17	40.4	70.9	41.6	10.5	42.7	41.2
Mobile-DETR ^[12]	MobileNet	19	39.2	63.3	38.5	8.6	35.4	38.0
RT-DETRv2 ^[48]	R50	18	40.5	73.3	41.0	16.7	42.4	41.4
Lite-DETR ^[90]	R50	9	35.6	69.2	36.8	20.9	35.9	37.2
YOLOv8s ^[35]	-	12	35.9	68.0	33.9	11.4	35.6	37.5
YOLOv9s ^[91]	-	11	37.7	71.2	39.1	16.9	40.7	38.1
YOLOv10s ^[36]	-	13	35.5	67.3	35.1	11.1	38.1	36.2
YOLO11s ^[37]	-	12	38.2	69.1	40.6	7.6	36.8	39.9
EdgeYOLO ^[92]	-	30	38.7	68.7	42.0	19.5	40.5	39.1
FMD-DETR	R50	20	<u>41.5</u>	<u>73.9</u>	<u>44.2</u>	22.2	<u>43.6</u>	<u>42.4</u>
Ours	R50	20	42.1	75.9	44.8	23.5	44.8	43.0

于小目标检测的提升。以上结果表明，本章提出的算法在其他果实数据集的检测中保持了良好的泛化性能。提出的高频细节增强模块能够关注重要特征，抑制不相关特征或背景的干扰，进而实现了更精准的目标定位；同时，所设计的动态稀疏混合专家解码器及专家损失策略，能够让模型自适应地学习到不同的特征，有效缓解了多类别特征之间的相互干扰，进一步提高了模型的特征区分能力。此外，本章算法在保持较高检测性能的同时，推理速度达到了 20 FPS，满足了实际应用中对实时性的需求。本章算法在草莓数据集上的优异表现，验证了其在不同果实类别和复杂环境下的检测能力与泛化能力。

为检验模型在不同果实类别和环境下的检测性能与泛化能力，表4.4展示了所提算法与其他目标检测算法以及 FMD-DETR 算法在 FruitRipeness 数据集上的定量对比结果。从表4.4中实验结果可以看出，所提算法在 FruitRipeness 数据集取得了最优异的检测性能，超过了现有多目标检测方法。具体而言，所提算法平均精度（AP）达到 87.7%，相较于 RT-DETR 提升了 1.0 个百分点，AP₅₀ 和 AP₇₅ 分别提升了 0.2 个百分点和 0.6 个百分点，AP_M 和 AP_L 分别提升了 1.1% 和 1.1%，说明模型对不同尺

表 4.4 FruitRipeness 数据集上的模型定量比较

Model	Backbone	FPS	AP	AP ₅₀	AP ₇₅	AP _M	AP _L
Deformable-DETR ^[42]	R50	27	76.1	95.7	93.0	69.6	77.4
Conditional-DETR ^[43]	R50	29	84.6	99.2	98.8	80.7	85.3
DAB-DETR ^[44]	R50	25	84.3	99.4	98.6	79.4	85.0
DINO ^[47]	R50	20	85.8	99.4	98.4	83.4	86.4
Co-DETR ^[88]	R50	10	86.4	99.4	99.1	84.0	86.8
Rank-DETR ^[89]	R50	15	84.3	98.9	97.9	78.1	84.9
RT-DETR ^[12]	R50	29	86.7	99.3	98.7	84.1	87.0
Mobile-DETR ^[12]	MobileNet	26	80.7	94.2	93.2	84.2	81.1
RT-DETRv2 ^[48]	R50	29	86.3	99.2	98.7	83.7	86.8
Lite-DETR ^[90]	R50	10	85.6	99.4	98.6	81.9	86.2
YOLOv10s ^[36]	-	16	85.0	99.6	98.7	82.9	85.5
YOLO11s ^[37]	-	25	85.4	99.6	98.8	82.8	85.8
EdgeYOLO ^[92]	-	34	83.1	99.1	98.2	83.5	83.6
FMD-DETR	R50	30	<u>87.1</u>	99.4	<u>99.1</u>	<u>84.7</u>	<u>87.6</u>
Ours	R50	29	87.7	<u>99.5</u>	99.3	85.2	88.1

寸的目标的检测能力得到了显著增强，而且增强了边缘和纹理信息的有效提取，使得果实区域的特征表达更加充分。同时，相较于 FMD-DETR，AP 提升了 0.6 个百分点。在 FMD-DETR 已实现较高精度的基础上，本章算法仍能取得稳定增益，充分验证了所提策略的有效性。FruitRipeness 数据集包含多种果实类别，且果实大小和成熟度存在差异，这要求模型能够在不同果实类别和多样环境下保持稳定表现。以上结果表明，本章方法通过增强对多尺度特征的感知和自适应能力，提升了对不同尺寸目标检测精度和定位精度，在果实成熟度检测任务中展现出优异的综合性能与泛化能力。

4.2.3 定性实验

图4.4展示了在 Laboro 番茄数据集上的可视化结果，其中红色虚线框表示漏检，蓝色虚线框表示误检，黄色虚线框则表示冗余框。番茄数据集图像中存在处于不同成熟阶段的番茄果实，以及大小各异的番茄果实和大量小目标和密集目标。番茄果实之间存在相互遮挡和重叠的情况，部分番茄果实颜色与叶片及背景相似，这些因素都增加了检测的难度。

检测结果如图4.4所示，从结果可以看出，Deformable-DETR、Conditional-DETR、DAB-DETR、DINO、Co-DETR、Rank-DETR 和 RT-DETR 算法在进行番茄成熟度检测时普遍存在漏检现象，尤其是对小目标和密集目标的检测上，说明这些算法在多尺度特征处理尤其是小目标的特征处理方面存在局限性。YOLOv8x、YOLOv10s 和 YOLO11s 算法在果实相互遮挡与重叠的情况下，存在漏检的问题。轻量化算法 Lite-DETR、Mobile-DETR 和 EdgeYOLO 存在漏检现象，且 EdgeYOLO 存在误检和冗余框情况，说明这些轻量化算法对于果实重叠的复杂场景的检测能力有限。相比之下，本章算法在番茄成熟度检测中展现出了优秀的检测能力，不仅避免了漏检、误检和冗余框的情况，而且检测框更贴合目标边界，对小目标同样实现了精准的检测。

图4.5展示了在草莓数据集上的可视化结果，其中红色虚线框代表漏检，蓝色虚线框代表误检，黄色虚线框代表冗余框。在草莓数据集的图像中，由于叶片造成的遮挡和重叠，部分草莓果实的边界并不清晰，而草莓和花梗的茂密生长进一步增加了检测难度。

从图4.5中可以看出，Deformable-DETR、Conditional-DETR、DAB-DETR 和 RT-DETR 算法在草莓成熟度检测时存在冗余框现象，并且 DINO、Rank-DETR、RT-DETR 和 RT-DETRv2 算法存在误检情况，说明这些算法在密集目标场景或目标近似重叠等情况存在局限性。YOLO 系列的 YOLOv9s 和 YOLOv10s 算法同样存在误检的情况。此外，轻量化模型 Lite-DETR、Mobile-DETR 和 EdgeYOLO 同样存在误检情况。相比之下，本文算法有效减少了与背景相似目标的误检，显示出更强的特征区分能力，能够更精准地捕捉目标的关键特征。本章算法在草莓成熟度检测中展现出了更好的检测能力与特征区分能力，为复杂场景下的农业目标检测提供了更精准的识别方案。



图 4.4 番茄成熟度检测结果定性比较



图 4.5 草莓成熟度检测结果定性比较

4.2.4 消融实验

模块有效性验证。为了验证本章所提出的高频细节增强模块（HDEM）、动态稀疏混合专家解码器（DS-MoE Decoder）以及负载均衡与专家多样性损失（LB&ED Loss）策略在果实成熟度检测中的有效性，在 Laboro 番茄数据集上进行了消融实验。为了保证实验的一致性，所有模型均使用相同的超参数训练。在消融实验中，使用标准 RT-DETR 模型作为基线模型。表4.5 显示，基线模型未加入任何模块时，AP 为

表 4.5 Laboro 番茄数据集各模块的消融实验结果

HDEM	DS-MoE Decoder	LB&ED Loss	AP	AP ₅₀	AP ₇₅	AP _S
			71.1	82.7	77.8	86.4
✓			75.2	86.8	81.1	88.1
	✓		75.5	86.5	81.7	88.2
	✓	✓	75.8	87.0	82.4	88.7
✓	✓		75.7	86.7	81.5	88.2
✓	✓	✓	76.1	87.1	82.4	89.1

71.1%。当单独加入高频细节增强模块（HDEM）时，AP 提升至 75.2%（+4.1%）。单独加入动态稀疏混合专家解码器（DS-MoE Decoder）时，AP 提升至 75.5%（+4.4%）。结合 DS-MoE Decoder 与 LB&ED Loss 时，AP 达 75.8%（+4.7%）。当三者同时使用时，AP 达到 76.1%，同时 AP₅₀（+4.4%）、AP₇₅（+4.6%）、AP_S（+2.7%）均有提升，在所有指标上均达到最佳表现，说明本章提出的模块一起使用能够协同提升模型在小目标定位、多尺度果实检测以及成熟度检测上的表现，有效增强了模型的整体性能和可靠性。

高频细节增强模块有效性。为了验证高频细节增强模块（HDEM）在特征关注和目标识别中的有效性，对特征图进行了可视化分析，如图4.6 所示，图 (a) 代表输入图像，图 (b) 代表未使用 HDEM 模块的特征图，图 (c) 代表使用 HDEM 模块的特征图。如图4.6 (b) 和 (c) 所示，未使用 HDEM 模块的特征图目标区域的定位受到背景的干扰，特征热力图在背景区域存在明显的响应。而使用 HDEM 模块的特征图对特征的边缘和纹理表示更清晰、更精准，模型的特征图更能定位在果实区域。结果表明，高频细节增强模块提升了模型对目标的定位准确性，从而提高了果实成熟度检测的性能。

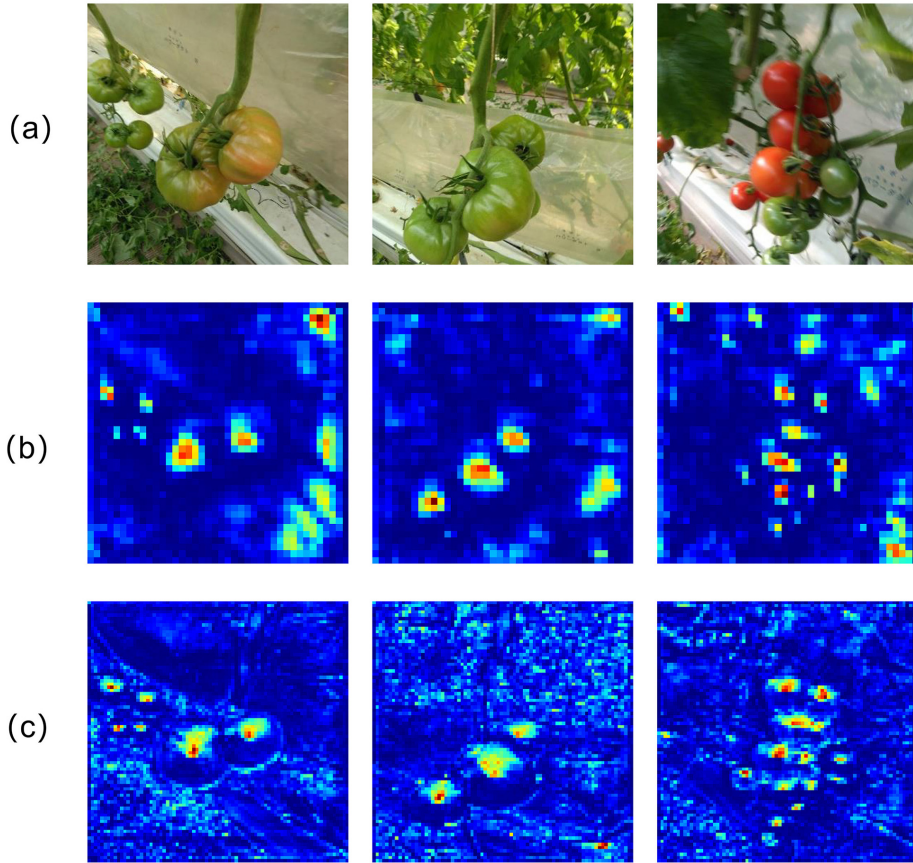


图 4.6 HDEM 特征图可视化对比

尺度感知路由与稀疏设计的有效性。为验证提出的动态稀疏混合专家解码器（DS-MoE Decoder）中尺度感知路由（Scale-Aware Routing）与稀疏设计（Sparse Design）的有效性，在 Laboro 番茄数据集上进行了消融实验。实验结果如表4.6。首先，将基线解码器与同时去除尺度感知路由和稀疏设计的模型进行对比，AP 从 71.1% 提升至 74.8%，表明基于 MoE 的解码器能够显著提升模型的检测性能。在去除稀疏设计后模型参数量由 42M 增加至 55M，而检测性能基本保持一致。这表明稀疏设计在有效降低模型复杂度的同时，能够保持模型性能，实现了良好的效率和性能平衡。结合尺度先验信息后，AP 从 74.9% 提升至 75.5%，说明尺度先验信息能够引导专家分配更加合理，从而提升特征表达能力。最后，提出的 DS-MoE Decoder 结合了尺度感知与动态稀疏设计，在综合性能上达到了最佳。实验结果验证了两种设计的有效性。

表 4.6 尺度感知路由与稀疏设计在 DS-MoE 解码器中的消融实验

Model	AP	AP ₅₀	AP ₇₅	AP _S	Params(M)
Baseline Decoder	71.1	82.7	77.8	86.4	42
w/o Scale-Aware Routing	74.9	86.4	81.6	87.7	42
w/o Sparse Design	75.3	86.9	81.2	87.6	55
w/o Sparse Design & Scale-Aware Routing	74.8	86.1	81.1	87.4	55
DS-MoE Decoder	75.5	86.5	81.7	88.2	42

负载均衡与专家多样性损失策略的有效性。为了验证针对混合专家（MoE）模块提出的辅助损失函数策略的有效性，针对负载均衡损失（Load Balancing Loss）和专家多样性损失（Expert Diversity Loss）进行了消融实验。结果如表4.7所示，未添加任何辅助损失的 MoE 模型（w/o MoE Loss）已经具备了较好的性能，其整体平均精度（AP）为 75.5%。单独使用负载均衡损失后，模型的 AP 达到 75.7%，单独使用专家多样性损失函数，模型的 AP 达到 75.7%，而同时使用两者共同构成混合专家的损失函数时，模型的精度达到 75.8%，相比于基础的混合专家模型，提升了 0.3 个百分点。说明负载均衡与专家多样性损失共同作用时能有效优化专家分配机制，使得模型更好地发挥作用。

表 4.7 负载均衡与专家多样性损失策略消融实验结果

Setting	AP	AP ₅₀	AP ₇₅	AP _S
w/o LB&ED Loss	75.5	86.5	81.7	88.2
+ Load Balancing Loss	75.7	86.9	81.9	88.6
+ Expert Diversity Loss	75.7	87.1	81.8	88.4
+ Both	75.8	87.0	82.4	88.7

负载均衡与专家多样性损失策略系数。负载均衡损失和专家多样性损失对于稀疏 MoE 模型的稳定训练至关重要。本章对权重系数 λ_1, λ_2 进行超参数消融实验，验证其对性能的影响。

从表4.8结果可知， λ_1 在所测试的取值范围内，在 $\lambda_1=1$ 时性能达到最佳。 λ_1 进一步增大时，模型性能会下降，这表明过大的负载均衡权重可能会过度约束专家分布，导致性能不再提升。因此，最终设置 $\lambda_1=1$ 作为负载均衡损失的系数。

从表4.9结果可知， λ_2 在所测试的取值范围内，最好的结果出现在 $\lambda_2 = 3$ 时，说明此时的权重值有助于充分激活更多专家，从而提升整体性能。当 λ_2 进一步增大时，

表 4.8 负载均衡损失系数 λ_1 对 MoE 模型性能的影响

λ_1	AP	AP ₅₀	AP ₇₅	AP _S
0	75.5	86.5	81.7	88.2
0.1	75.5	86.3	81.9	88.3
0.5	74.7	86.1	80.7	87.9
1	75.7	86.9	81.9	88.6
3	75.2	86.4	81.9	88.2

表 4.9 专家多样性损失系数 λ_2 对 MoE 模型性能的影响

λ_2	AP	AP ₅₀	AP ₇₅	AP _S
0	75.5	86.5	81.7	88.2
0.1	75.4	87.1	81.7	88.1
0.5	74.9	86.8	81.2	87.6
1	75.5	86.8	81.8	88.4
3	75.7	87.1	81.8	88.4
4	74.6	85.6	80.9	88.2

模型性能出现下降，表明过强的多样性约束可能会干扰门控网络的最优决策。因此，最终设置专家多样性损失系数 $\lambda_2 = 3$ 。

专家激活数量选择。为了研究所激活专家数量 k 对模型性能的影响，实验评估了 $k \in \{1, 2, 3, 4\}$ 的模型的性能指标。结果如表4.10所示，适当增加 k 能够带来一定的性能提升，例如从 $k = 1$ 到 $k = 2$ 时，AP 有所提高，在 Top-2 时达到了较高的精度。当 k 进一步增大时，性能呈现下降趋势，说明选择专家的数量越多并不能一直带来性能的提升。因此，Top-2 在当前实验设置下表现最好，作为最终设置。

表 4.10 专家激活数量 (k) 对模型性能的影响

Top- k	AP	AP ₅₀	AP ₇₅	AP _S	Params(M)
1	75.3	86.8	81.4	88.0	42
2	75.5	86.5	81.7	88.2	42
3	75.1	86.7	81.2	87.9	42
4	74.9	86.4	80.6	88.2	42

专家负载分布均衡性分析。本章对解码器各层中四个专家的负载分配比例进行了可视化与统计分析。如图4.7所示，从专家负载分布可以观察到，不同层的专家使用模式呈现出明显差异，表明模型能够根据输入特征动态选择合适的专家，从而实

现有效的专家分化。同时，各专家的负载保持在相对均衡的范围内，未出现明显的专家塌缩现象，说明所提出的方法在促进专家专门化的同时，能够维持合理的负载分布。该结果验证了所设计的路由机制在专家分工与负载均衡之间取得了良好的平衡。

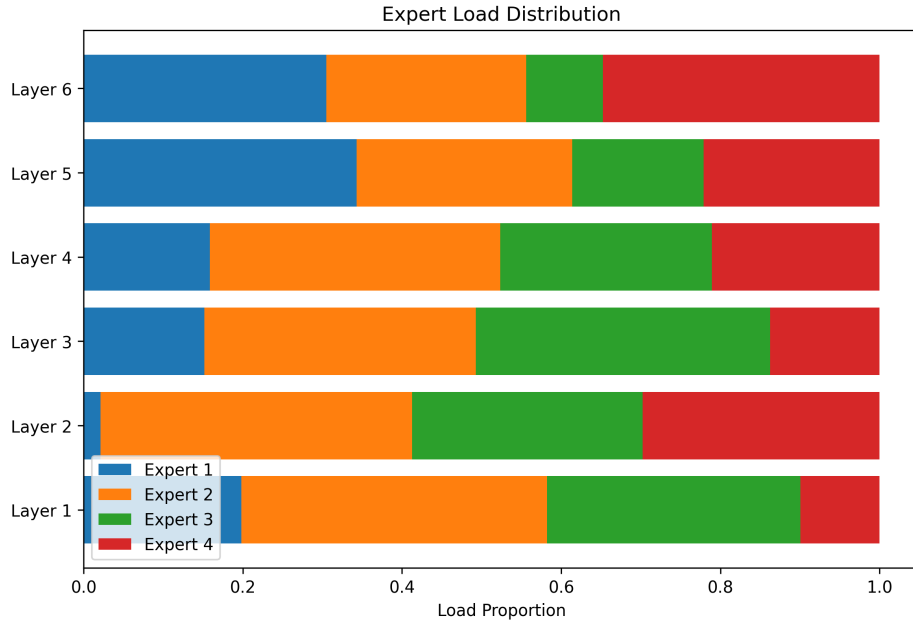


图 4.7 专家负载分布示意图

专家分配可视化分析。在图4.8中，展示了番茄数据集上的专家分配可视化结果，依次为原始图、预测结果图和专家分配图。通过将最终预测出的预测框与对应的专家分配结果进行可视化，直观呈现了目标检测中专家分配的情况，从而解释了专家真正学到了什么。在图4.8(a)中，一位专家主要负责处理尺度较大且特征清晰的目标，另一位专家专注于处理尺度较小且相对密集的目标，说明专家不仅根据目标的尺度进行分工，还根据目标的特征复杂度进行分工。在图4.8(b)中，对于尺度相近的目标，专家根据不同成熟度的颜色差异进行分工，红色的成熟番茄被精准分配给了其中一位专家，而绿色的未成熟番茄则被分配给了另一位专家。这表明所设计的动态稀疏混合专家机制具备自适应能力，能够准确抓住当前场景中最具区分度的特征来进行专家路由，从而提升模型的检测能力。



图 4.8 专家分配可视化结果

4.3 本章小结

针对农业果实成熟度检测中特征多样性、模型缺乏针对性建模与自适应能力不足的问题，本章提出了一种面向农业果实成熟度检测的基于动态稀疏混合专家机制的检测模型 AgriMoE-DETR。该模型通过高频细节增强模块，将浅层细节与深层语义相结合，提高对小目标特征的关注能力；在解码阶段构建了动态稀疏混合专家解码器，能够根据输入特征动态激活不同专家参与推理，增强模型的自适应能力；同时，提出负载均衡与专家多样性损失策略，促使不同专家学习到多样化的特征表示，提升了模型的泛化能力。在实验验证方面，所提模型在番茄数据集、真实场景番茄数据集、草莓数据集以及 FruitRipeness 数据集上的平均精度分别为 76.1%、54.1%、42.1% 和 87.7%。相较于 FMD-DETR，该方法进一步提升了检测精度，增强了模型在复杂农业场景下的性能。消融实验和可视化结果进一步说明，高频细节增强模块、动态稀疏混合专家解码器以及负载均衡与专家多样性损失策略对性能提升具有积极作用，验证了本章方法的有效性。

相较于上一章的 FMD-DETR 方法，本章提出的算法进一步提升了果实成熟度检测的精度，同时验证了动态稀疏混合专家机制在农业视觉任务中的有效性，为果实成熟度检测在农业场景中的应用提供了有力的技术支撑。

第五章 总结与展望

5.1 总结

随着深度学习与计算机视觉技术的快速发展，基于图像的果实成熟度检测逐渐成为农业信息化的重要技术路径。然而，在真实农业场景中，果实成熟度检测仍存在多重挑战：一方面，果实生长环境复杂，模型对果实的识别容易受到遮挡、重叠和背景干扰的影响，现有模型在多尺度目标尤其是小果实的检测能力有限。另一方面，复杂农业环境下不同光照、遮挡及果实尺度变化导致特征多样性，现有模型缺乏对不同视觉特征的自适应处理能力。同时，农业生产对检测的实时性要求高，而高复杂度模型在实际部署中往往难以满足速度与效率的需求。针对上述问题，本研究提出基于协同注意力与自蒸馏的农业果实目标检测方法和基于动态稀疏混合专家机制的农业果实目标检测方法，这两种方法在复杂农业场景下的目标检测任务中展现出了优异的性能，提升了定位精度与成熟度识别精度，为智能采摘、精准农业管理及自动化分级提供了可靠的技术支撑。

本文的主要研究工作如下：（1）提出一种基于协同注意力与自蒸馏的农业果实目标检测方法。具体而言，针对果实在尺度、遮挡和重叠方面的复杂性，提出多尺度特征融合重加权模块，通过对不同尺度特征进行自适应重加权，提升模型对各类尺寸果实的检测能力。为进一步优化目标查询交互，构建协同注意力解码器，将交叉注意力与自注意力机制结合，增强解码器对遮挡和重叠目标的处理能力。此外，设计层级自蒸馏策略，实现解码器不同层之间的知识传递，提高空间感知与定位鲁棒性，从而提升在复杂果园场景下的检测精度。最终，在多个果实成熟度数据集上的实验结果验证了该方法在复杂农业环境下的有效性。

（2）提出一种基于动态稀疏混合专家机制的农业果实目标检测方法。具体而言，针对农业目标检测中小目标定位难的问题，设计高频细节增强模块，利用高频细节信息增强深层特征，提升对目标边缘和细节特征的识别能力，从而改善对小目标的定位精度。针对复杂农业环境下光照、遮挡及果实尺度变化导致的特征差异问题，构建动态稀疏混合专家解码器，结合多尺度先验信息和稀疏激活策略，自适应地激活不同专家参与推理，增强模型的自适应能力。为缓解专家激活不均衡和表示同质化

问题，设计负载均衡损失和专家多样性损失联合约束，使不同专家能够学习互补的特征表示，从而进一步提升模型对复杂农业场景的适应性。实验结果表明，该方法能够有效提升对不同成熟度状态果实的精准识别能力。

本研究不仅为果实成熟度检测提供了一种新的解决方案，还为其他农产品智能采摘机器人的视觉识别、目标定位以及分级等工作提供了有效的支持。

5.2 工作展望

尽管本研究提出的农业果实成熟度检测方法在多个数据集上取得了较为优异的检测结果，并为农业智能化应用提供了一定的技术基础。然而，该领域仍存在一些有待进一步探索的研究方向。未来的研究工作可主要从以下两个方面展开：

（1）数据的多样性与模型泛化能力。尽管本研究使用了公开数据集并结合了多种数据增强方法，但训练数据在数量和多样性方面仍存在一定限制。因此，未来工作将进一步扩展数据集，覆盖更加多样的环境条件和不同生长阶段，同时，结合多模态数据展开研究，从而提升模型在真实农业场景中的鲁棒性与泛化能力。

（2）模型的参数规模在轻量化部署方面仍面临挑战。未来研究应重点关注模型压缩与结构优化，以降低计算复杂度，从而提升其在资源受限设备上的应用效率。

参考文献

- [1] LIU J, LIU Z. The vision-based target recognition, localization, and control for harvesting robots: A review[J]. International Journal of Precision Engineering and Manufacturing, 2024, 25(2): 409-428.
- [2] DAS P, YADAV J P S. Automated tomato maturity grading system using cnn[C]// 2020 International Conference on Smart Electronics and Communication. IEEE, 2020: 136-142.
- [3] AHMAD J, GUEAIEB W, EL SADDIK A, et al. Yield estimation and health assessment of temperate fruits: A modular framework[J]. Engineering Applications of Artificial Intelligence, 2024, 136: 108871.
- [4] ISLAM M, BIJJAHALLI S, FAHEY T, et al. Destructive and non-destructive measurement approaches and the application of ai models in precision agriculture: a review[J]. Precision Agriculture, 2024, 25(3): 1127-1180.
- [5] 孙梦梦, 鞠皓, 姜洪喆, 等. 水果成熟度无损检测技术研究进展[J]. 食品与发酵工业, 2023, 49(17): 354-362.
- [6] ZHOU X, HU X, SUN J. A review of fruit ripeness recognition methods based on deep learning[J]. Cyber-Physical Systems, 2025, 11(4): 508-542.
- [7] DEB J, BHATTACHARJEE D, CHAKRABARTY J. Colour, texture and geometric feature based classification of banana varieties using interpretable machine learning [J]. Model Assisted Statistics and Applications, 2026, 21(1): 50-64.
- [8] ISLAM M S, ISLAM M, et al. Improving mango ripeness grading accuracy: A comprehensive analysis of deep learning, traditional machine learning, and transfer learning techniques[J]. Machine Learning with Applications, 2025, 19: 100619.
- [9] MA J, LI M, FAN W, et al. State-of-the-art techniques for fruit maturity detection[J]. Agronomy, 2024, 14(12): 2783.

- [10] LI Z, DONG Y, SHEN L, et al. Development and challenges of object detection: A survey[J]. *Neurocomputing*, 2024, 598: 128102.
- [11] WEI W, CHENG Y, HE J, et al. A review of small object detection based on deep learning[J]. *Neural Computing and Applications*, 2024, 36(12): 6283-6303.
- [12] ZHAO Y, LV W, XU S, et al. Detsrs beat yolos on real-time object detection[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 16965-16974.
- [13] SONG B, ZHAO S, WANG Z, et al. Lw-detr: a lightweight transformer-based object detection algorithm for efficient railway crossing surveillance[J]. *Complex & Intelligent Systems*, 2025, 11(12): 1-13.
- [14] XIAO B, NGUYEN M, YAN W Q. Fruit ripeness identification using yolov8 model [J]. *Multimedia Tools and Applications*, 2024, 83(9): 28039-28056.
- [15] OLISAH C C, TREWHELLA B, LI B, et al. Convolutional neural network ensemble learning for hyperspectral imaging-based blackberry fruit ripeness detection in uncontrolled farm environment[J]. *Engineering Applications of Artificial Intelligence*, 2024, 132: 107945.
- [16] WU M, QIU Y, WANG W, et al. Improved rt-detr and its application to fruit ripeness detection[J]. *Frontiers in Plant Science*, 2025, 16: 1423682.
- [17] SEPTIARINI A, SUNYOTO A, HAMDANI H, et al. Machine vision for the maturity classification of oil palm fresh fruit bunches based on color and texture features[J]. *Scientia Horticulturae*, 2021, 286: 110245.
- [18] TU S, XUE Y, ZHENG C, et al. Detection of passion fruits and maturity classification using red-green-blue depth images[J]. *Biosystems Engineering*, 2018, 175: 156-167.
- [19] GUO C, LIU F, KONG W, et al. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine[J]. *Journal of Food Engineering*, 2016, 179: 11-18.
- [20] SUBAN I B, PARAMARTHA A, FORTWONATUS M, et al. Identification the maturity level of carica papaya using the k-nearest neighbor[C]// *Journal of Physics: Conference Series*: Vol. 1577. IOP Publishing, 2020: 012028.

- [21] MA L, LIANG C, CUI Y, et al. Prediction of banana maturity based on the sweetness and color values of different segments during ripening[J]. *Current Research in Food Science*, 2022, 5: 1808-1817.
- [22] ZHU X, CHEN F, ZHENG Y, et al. Detection of camellia oleifera fruit maturity in orchards based on modified lightweight yolo[J]. *Computers and Electronics in Agriculture*, 2024, 226: 109471.
- [23] LONG J, ZHAO C, LIN S, et al. Improved mask r-cnn-based segmentation method for tomato fruits at different maturity stages in greenhouse environment[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2021, 37(18): 100-108.
- [24] KO K, JANG I, CHOI J H, et al. Stochastic decision fusion of convolutional neural networks for tomato ripeness detection in agricultural sorting systems[J]. *Sensors*, 2021, 21(3): 917.
- [25] LI Y, LIU M, HE Z, et al. Detecting citrus fruit maturity using improved yolov8s [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2024, 40(24): 157-164.
- [26] ZHANG L, ZHOU W. Grape ripeness discrimination based on mobilenet v2[J]. *Xinjiang Agricultural Mechanization*, 2022(06): 29-31+42.
- [27] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//*European Conference on Computer Vision*. Springer, 2020: 213-229.
- [28] TANG Z, FANG L, SUN S, et al. Ml-detr: Multiscale-lite detection transformer for identification of mature cherry tomatoes[J]. *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 1-18.
- [29] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016: 779-788.
- [30] CAO J, PENG B, GAO M, et al. Object detection based on cnn and vision-transformer: A survey[J]. *IET Computer Vision*, 2025, 19(1): e70028.

- [31] PAGIRE V, CHAVALI M, KALE A. A comprehensive review of object detection with traditional and deep learning methods[J]. Signal Processing, 2025, 237: 110075.
- [32] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.
- [33] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.
- [34] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
- [35] JOCHER G, CHAURASIA A, QIU J. Ultralytics yolov8[CP/OL]. 2023. <https://github.com/ultralytics/ultralytics>.
- [36] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.
- [37] JOCHER G, QIU J. Ultralytics yolo11[CP/OL]. 2024. <https://github.com/ultralytics/ultralytics>.
- [38] TIAN Y, YE Q, DOERMANN D. Yolov12: Attention-centric real-time object detectors[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025.
- [39] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, 2016: 21-37.
- [40] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.
- [41] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9627-9636.

- [42] ZHU X, SU W, LU L, et al. Deformable detr: Deformable transformers for end-to-end object detection[C]//International Conference on Learning Representations. 2021.
- [43] MENG D, CHEN X, FAN Z, et al. Conditional detr for fast training convergence [C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 3651-3660.
- [44] LIU S, LI F, ZHANG H, et al. Dab-detr: Dynamic anchor boxes are better queries for detr[C]//International Conference on Learning Representations. 2022.
- [45] LI F, ZHANG H, LIU S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13619-13627.
- [46] CHEN Q, CHEN X, WANG J, et al. Group detr: Fast detr training with group-wise one-to-many assignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6633-6642.
- [47] ZHANG H, LI F, LIU S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[C]//International Conference on Learning Representations. 2023.
- [48] LV W, ZHAO Y, CHANG Q, et al. Rt-detr2: Improved baseline with bag-of-freebies for real-time detection transformer[EB/OL]. 2024. <https://arxiv.org/abs/2407.17140>.
- [49] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems: Vol. 30. 2017.
- [50] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. 2021.
- [51] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
- [52] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

- [53] PADILLA R, NETTO S L, DA SILVA E A. A survey on performance metrics for object-detection algorithms[C]//2020 International Conference on Systems, Signals and Image Processing. IEEE, 2020: 237-242.
- [54] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
- [55] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.
- [56] GUO C, FAN B, ZHANG Q, et al. Augfpn: Improving multi-scale feature learning for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12595-12604.
- [57] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10781-10790.
- [58] ZHOU J, ZHOU Z, ZHANG P. Dynaood-net: Dynamic progressive feature fusion and energy balancing for robust out-of-distribution detection[J]. IEEE Access, 2025, 13: 52758-52774.
- [59] YANG G, LEI J, ZHU Z, et al. Afpn: Asymptotic feature pyramid network for object detection[C]//2023 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2023: 2184-2189.
- [60] FAN H, XIONG B, MANGALAM K, et al. Multiscale vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6824-6835.
- [61] XIANG X, GONG W, LI S, et al. Tcnet: Multiscale fusion of transformer and cnn for semantic segmentation of remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 3123-3136.

- [62] HUANG Y, JIAO D, HUANG X, et al. A hybrid cnn-transformer network for object detection in optical remote sensing images: Integrating local and global feature fusion[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 18: 241-254.
- [63] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [C]//NIPS Deep Learning and Representation Learning Workshop. 2015.
- [64] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
- [65] ZHANG L, BAO C, MA K. Self-distillation: Towards efficient and compact neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4388-4403.
- [66] ZHENG Z, YE R, WANG P, et al. Localization distillation for dense object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9407-9416.
- [67] PENG Y, LI H, WU P, et al. D-fine: Redefine regression task of detr as fine-grained distribution refinement[C]//The Thirteenth International Conference on Learning Representations. 2025.
- [68] WANG J, CHEN Y, ZHENG Z, et al. Crosskd: Cross-head knowledge distillation for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16520-16530.
- [69] MI P, LIN J, ZHOU Y, et al. Active teacher for semi-supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14482-14491.
- [70] SU H, JIAN Z, WEI Y, et al. Classification and regression task integration in distillation for object detectors[J]. Neurocomputing, 2025, 624: 129386.
- [71] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive mixtures of local experts [J]. Neural Computation, 1991, 3(1): 79-87.

- [72] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[C]//Proceedings of the International Conference on Learning Representations. 2017.
- [73] LEPIKHIN D, LEE H, XU Y, et al. Gshard: Scaling giant models with conditional computation and automatic sharding[C]//Proceedings of the International Conference on Learning Representations. 2021.
- [74] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120): 1-39.
- [75] DU N, HUANG Y, DAI A M, et al. Glam: Efficient scaling of language models with mixture-of-experts[C]//International Conference on Machine Learning. PMLR, 2022: 5547-5569.
- [76] RIQUELME C, PUIGCERVER J, MUSTAFA B, et al. Scaling vision with sparse mixture of experts[J]. Advances in Neural Information Processing Systems, 2021, 34: 8583-8595.
- [77] FAN Z, SARKAR R, JIANG Z, et al. M³vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design[J]. Advances in Neural Information Processing Systems, 2022, 35: 28441-28457.
- [78] HAN X, WEI L, DOU Z, et al. Vimoe: An empirical study of designing vision mixture-of-experts[J]. IEEE Transactions on Image Processing, 2025, 34: 7209-7221.
- [79] XU X, KONG L, SHUAI H, et al. Limoe: Mixture of lidar representation learners from automotive scenes[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 27368-27379.
- [80] SHEN L, CHEN G, SHAO R, et al. Mome: Mixture of multimodal experts for generalist multimodal large language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 42048-42070.
- [81] SHU F, LIAO Y, ZHANG L, et al. Llava-mod: Making llava tiny via moe-knowledge distillation[C]//The Thirteenth International Conference on Learning Representations. 2025.

- [82] LIN B, TANG Z, YE Y, et al. Moe-llava: Mixture of experts for large vision-language models[J]. IEEE Transactions on Multimedia, 2026.
- [83] LI C, LI L, JIANG H, et al. Yolov6: A single-stage object detection framework for industrial applications[EB/OL]. 2022. <https://github.com/meituan/YOLOv6>.
- [84] HU Z, SUN Y, WANG J, et al. Dac-detr: Divide the attention layers and conquer[C]//Thirty-seventh Conference on Neural Information Processing Systems. 2023.
- [85] TRIGUBENKO R, HFUJIHARA. Laborotomato: Instance segmentation dataset [DS/OL]. 2020. <https://github.com/laboroai/LaboroTomato>.
- [86] PASTELL MATTIL M, VICTOR B. Strawberry dataset for object detection[DS/OL]. Zenodo, 2022. DOI: 10.5281/zenodo.6126677.
- [87] SUTAR A, NAIKARE A, JADHAV P, et al. Mango and banana dataset (ripe unripe): Indian rgb image datasets for yolo object detection[DS/OL]. Mendeley Data, 2023. DOI: 10.17632/y3649cmgg6.3.
- [88] ZONG Z, SONG G, LIU Y. Detrs with collaborative hybrid assignments training[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6748-6758.
- [89] PU Y, LIANG W, HAO Y, et al. Rank-detr for high quality object detection[J]. Advances in Neural Information Processing Systems, 2023, 36: 16100-16113.
- [90] LI F, ZENG A, LIU S, et al. Lite detr: An interleaved multi-scale encoder for efficient detr[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18558-18567.
- [91] WANG C Y, YEH I H, MARK LIAO H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European Conference on Computer Vision. Springer, 2024: 1-21.
- [92] LIU S, ZHA J, SUN J, et al. Edgeyolo: An edge-real-time object detector[C]//2023 42nd Chinese Control Conference. IEEE, 2023: 7507-7512.
- [93] POLESEL A, RAMPONI G, MATHEWS V J. Image enhancement via adaptive unsharp masking[J]. IEEE Transactions on Image Processing, 2000, 9(3): 505-510.

攻读硕士学位期间取得的研究成果

一、主要作者研究成果

[1] Han Y, Huang Z, Sun Y, et al. Agricultural object detection in complex environments via co-attention and self-knowledge distillation[J]. Information Sciences, 2025: 122711. (中科院二区, 导师一作, 本人二作)

二、软件著作权

[1] 基于自蒸馏模型的果实成熟度检测平台 V1.0, 开发人: 韩越兴、黄志怡、陈侨川、孙妍, 登记号: 2025SR2154696, 登记日期: 2025 年 11 月 5 日, 申请人: 上海大学。

[2] 基于双编码器和核预测网络的古代壁画图像修复平台 V1.0, 开发人: 韩越兴、黄志怡、陈侨川, 登记号: 2024SR1573964, 登记日期: 2024 年 10 月 21 日, 申请人: 上海大学。

致 谢

在上海大学的三年学习时光中，感谢每一位给予我帮助与温暖的师长、同窗和家人，是你们让我不曾孤单。

我要感谢我的导师韩越兴老师，在研究生期间，他一直给予我无私的指导和支持，在学术上，他耐心地指导我进行研究工作，提供了很多宝贵的建议和意见，帮助我克服了许多困难和挑战。在生活中，韩老师也是一个非常温暖和关心学生的人，他总是关心我的学习和生活情况，给予我很多帮助和支持。同时，我也要感谢陈侨川老师，在科研上给予了我很多指导和帮助，陈老师的专业知识和经验对我的研究工作起到了重要的推动作用。感谢孙妍老师，她一丝不苟的工作态度和对细节的关注让我受益匪浅，在论文写作过程中，她的建议和指导帮助我提升了论文的质量。祝愿他们在未来的工作和生活中一切顺利，取得更大的成就。

其次，我要感谢我的家人和朋友们，在我攻读硕士学位期间，他们一直是我坚强的后盾和支持者。感谢他们的理解、鼓励和支持，让我能够专注于学业和研究工作，克服了许多困难和挑战。感谢师兄师姐们的帮助和指导，他们在学术上给予了我很多启发和支持。感谢同学们的陪伴和支持，我们一起度过了许多难忘的时光，互相鼓励和帮助，共同成长。

最后，我要感谢始终陪伴我的自己，读研给你带来了迷茫，也带来了成就感，每一次成功的喜悦都离不开你的坚持和努力。每个在实验室奋斗的夜晚，每个孤单的瞬间，你都以一颗强大的心脏和乐观的心态坚持了下来。祝愿你在未来的道路上继续保持赤诚和乐观，不要磨灭对学习的热情和对生活的热爱，健康快乐。

论文有页码，但求知没有终点；校园有围墙，但人生天地广阔。我将带着这段岁月赋予我的理性与坚韧，奔赴下一场山海。愿我们此去前程似锦，他日相逢，依旧赤诚如初。

黄志怡

上海大学

2026 年 4 月