

中图分类号: TP391

单位代号: 10280

密 级: 公开

学 号: 23721525

上海大学



硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	基于 Transformer 的复杂农田 杂草检测方法研究
--------	----------------------------------

作 者 沈新宇

学科专业 计算机应用技术

导 师 孙妍

完成日期 二〇二六年四月

姓 名：沈新宇

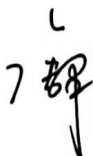
学号：23721525

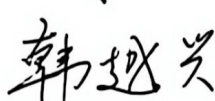
论文题目：基于 Transformer 的复杂农田杂草检测方法研究

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：  

导 师： 

答辩日期： 2026 年 6 月 9 日

姓 名：沈新宇

学号：23721525

论文题目：基于 Transformer 的复杂农田杂草检测方法研究

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：沈新宇

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：沈新宇

导师签名：孙妍

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 Transformer 的复杂农田
杂草检测方法研究

作者: 沈新宇

导师: 孙妍

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 Transformer-Based Weed Detection Methods in Complex Agricultural Environments

Candidate: Xinyu Shen

Supervisor: Yan Sun

Major: Technology of Computer Application

School of Computer Engineering and Science

Shanghai University

April, 2026

摘 要

随着智慧农业的发展，复杂农田场景下的杂草目标检测在智能除草与精准施药等应用中具有重要意义。然而，在实际田间场景中，杂草与农作物外观相似，目标密集且分布不均，并伴随大范围遮挡，导致现有方法易出现漏检、误检及预测不稳定等问题。尽管卷积神经网络在视觉任务中得到了广泛应用，但其受限于局部感受野的建模机制，难以有效捕获全局上下文信息。而基于 Transformer 的方法虽具备全局建模优势，但在复杂场景下仍存在特征判别能力不足及解码过程不稳定的挑战。为此，本文提出一种基于 Transformer 的杂草目标检测方法，以提升复杂农田场景下的检测性能。具体而言，本文围绕特征表征增强与解码建模优化两个方面开展研究：

1. 针对复杂背景下特征表征不足的问题，提出双域融合 Transformer 检测模型 FS-DETR (Frequency-Spatial Domain Integrated Detection Transformer)。该方法通过混合特征融合模块与双域注意力机制，实现空间域与频域特征的联合建模，提升细粒度特征表达能力；并引入基于高斯分布和约束引导的标签分配策略，优化正负样本匹配过程，提高训练稳定性与检测精度。

2. 针对解码过程不稳定的问题，本文在 FS-DETR 基础上提出其解码增强版本 FS-DETR-D (FS-DETR with Decoder Enhancement)。该方法通过基于置信度熵与空间感知的动态查询分配机制，实现查询数量对场景复杂度的自适应调节；同时结合超边增强的高阶查询交互机制，以建模密集区域目标间的结构关系；并设计匹配质量自适应感知损失函数，使监督强度随匹配质量动态变化，从而提升解码过程的稳定性与预测一致性。

在 WeedCrop、LincolnBeet 和 MH-Weed16 等多个公开数据集上的实验结果表明，本文提出的 FS-DETR 在检测精度与复杂场景适应性方面优于多种主流方法，其解码增强版本 FS-DETR-D 在三个数据集上相较于基线模型 FS-DETR 分别实现了 1.1%、0.8% 和 0.9% 的检测精度提升，进一步验证了模型的有效性与稳定性。

关键词：杂草检测；Transformer；双域特征融合；动态查询分配；高阶关系建模

ABSTRACT

With the development of smart agriculture, weed object detection in complex farmland scenarios plays an important role in applications such as intelligent weeding and precision spraying. However, in real field environments, weeds and crops often exhibit similar appearances, with densely packed and unevenly distributed targets accompanied by large-scale occlusions, which lead existing methods to suffer from missed detections, false positives, and unstable predictions. Although convolutional neural networks have been widely used in vision tasks, they are limited by local receptive field modeling and struggle to effectively capture global contextual information. Transformer-based methods, while capable of global modeling, still face challenges such as insufficient feature discriminability and instability in the decoding process in complex scenarios. To address these issues, this paper proposes a Transformer-based weed detection method to improve detection performance in complex farmland environments. Specifically, the proposed work focuses on two aspects, namely feature representation enhancement and decoding modeling optimization.

1. To address the insufficient feature representation in complex backgrounds, a dual-domain Transformer-based detection model, FS-DETR (Frequency-Spatial Domain Integrated Detection Transformer), is proposed. The method achieves joint modeling of spatial structural information and frequency-domain discriminative features through a Hybrid Feature Fusion module and a Dual-Domain Attention Mechanism, thereby enhancing fine-grained feature representation capability. In addition, a Gaussian Distribution-based and Constraint-guided Label Assignment strategy is introduced to improve positive - negative sample matching, thereby enhancing training stability and detection accuracy.

2. To address the instability in the decoding process, a decoder-enhanced variant of FS-DETR, named FS-DETR-D (FS-DETR with Decoder Enhancement), is proposed. This method introduces Entropy-Guided and Spatially-Aware Dynamic Query Allocation to adaptively adjust the number of queries according to scene complexity. Furthermore, Hyperedge Query Interaction is employed to model structural relationships among densely distributed

objects. Finally, Adaptive Matchability-Aware Loss is designed to dynamically adjust supervision strength according to matching quality, thereby improving decoding stability and prediction consistency.

Experimental results on multiple public datasets, including WeedCrop, LincolnBeet, and MH-Weed16, demonstrate that the proposed FS-DETR outperforms various mainstream methods in terms of detection accuracy and robustness in complex scenarios. Furthermore, its decoder-enhanced variant, FS-DETR-D, achieves improvements of 1.1%, 0.8%, and 0.9% in detection accuracy over the baseline model FS-DETR on the three datasets, respectively, further validating the effectiveness and stability of the proposed model.

Keywords: Weed Detection; Transformer; Dual-Domain Feature Fusion; Dynamic Query Allocation; High-Order Relationship Modeling

目 录

摘 要	I
ABSTRACT	II
第一章 绪论	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	3
1.2.1 通用目标检测研究现状与分析	3
1.2.2 杂草目标检测研究现状与分析	7
1.3 研究内容	8
1.4 论文组织结构	9
第二章 相关理论和技术概述	11
2.1 目标检测技术	11
2.1.1 基于卷积神经网络的目标检测方法.....	11
2.1.2 基于 Transformer 的目标检测方法	14
2.2 多尺度特征融合技术	19
2.3 频域特征建模技术	20
2.4 基于超边的高阶关系建模技术	21
2.5 数据集和指标	23
2.5.1 数据集	23
2.5.2 目标检测评价指标	25
第三章 面向复杂场景的双域 Transformer 杂草检测模型	27
3.1 引言	27
3.2 模型整体框架	28
3.3 混合特征融合模块	29
3.3.1 频域特征选择	29
3.3.2 多尺度特征融合	30
3.4 双域注意力机制模块	32

3.4.1	频域全局注意力	32
3.4.2	空间域可变形注意力	34
3.4.3	双域特征自适应融合	35
3.5	基于高斯分布与约束引导的标签分配方法	36
3.5.1	中心与位置约束	36
3.5.2	高斯分布相似性	37
3.5.3	代价函数与标签分配	38
3.6	实验结果与分析	39
3.6.1	实验数据及实验设置	39
3.6.2	实验评估和对比分析	40
3.6.3	消融实验	42
3.6.4	可视化结果分析	48
3.7	本章小结	52
第四章	面向复杂场景的解码自适应机制	54
4.1	引言	54
4.2	模型整体框架	55
4.3	基于置信度熵与空间感知的动态查询分配机制	55
4.3.1	置信度分布建模	56
4.3.2	空间密度感知建模	57
4.3.3	动态查询分配机制	58
4.4	超边增强的高阶查询交互机制	60
4.4.1	超边自适应构建	61
4.4.2	超边引导的高阶查询交互	62
4.4.3	超边与自注意力特征融合	63
4.5	匹配质量自适应感知的分类损失函数	64
4.5.1	匹配质量建模与监督冲突分析	64
4.5.2	自适应分类损失函数设计	65
4.6	实验结果与分析	66
4.6.1	实验数据及实验设置	66
4.6.2	实验评估和对比分析	66

4.6.3	消融实验	69
4.6.4	可视化结果分析	75
4.6.5	跨数据集泛化性能验证	76
4.7	本章小结	77
第五章	总结与展望	78
5.1	总结	78
5.2	未来研究展望	79
插图索引		81
表格索引		83
参考文献		84
攻读硕士学位期间取得的科研成果		93
致 谢		94

第一章 绪论

1.1 研究背景及意义

随着智慧农业的快速发展，杂草防治已成为保障农作物产量和品质的关键环节之一。传统的全面覆盖式化学除草不仅会导致农药的大量浪费，还会引发严重的土壤与水体污染，甚至导致杂草产生抗药性。因此，研发能够实现“精准靶向喷洒”或“激光除草”的智能除草机器人，成为了现代农业工程的重要发展趋势。在这一过程中，目标检测作为智能除草系统视觉感知的核心基础任务，其在复杂农业场景下的检测精度与计算效率，直接决定了智能农机作业的质量与可行性。近年来，随着深度学习的快速发展，基于深度神经网络的目标检测方法能够从大规模数据中自动学习判别性特征，从而突破传统方法依赖人工特征设计且泛化能力受限的瓶颈，并已广泛应用于农业视觉领域。

然而，如图1.1所示，绿色边界框对应农作物目标，红色边界框表示杂草，在真实自然农业环境中，杂草目标检测面临复杂多变的场景条件。苗期阶段的杂草通常以小尺度形式存在，且分布密集且不均匀；同时，其与作物在颜色与纹理上具有较高相似性，易受到复杂背景的干扰。此外，作物与杂草之间普遍存在相互遮挡，并伴随光照变化产生的明暗交错，进一步加剧了特征表达的不稳定性。这些因素共同导致杂草目标的判别信息易被掩盖或混淆，从而对检测模型的特征建模能力与判别鲁棒性提出了更高要求。

在此背景下，模型不仅需要具备精细的作物与杂草区分能力，还需能够有效建模目标之间的空间结构关系，并在训练与推理过程中保持稳定一致的预测表现。因此，杂草目标检测的核心目标旨在从复杂农田图像中准确识别并定位杂草目标，实现作物与杂草之间的有效区分，为后续精准喷洒或机械除草提供可靠的视觉感知支持。该任务要求模型在复杂背景与多变环境条件下，具备精确的目标定位能力与良好的鲁棒性，从而实现稳定可靠的视觉理解与决策基础。

基于上述任务需求，杂草目标检测研究对于推动智能农业发展具有重要意义。一方面，该任务有助于提升智能除草系统的自动化与精细化水平，减少化学农药的使用，从而降低农业生产对生态环境的影响；另一方面，其研究成果可为计算机视觉

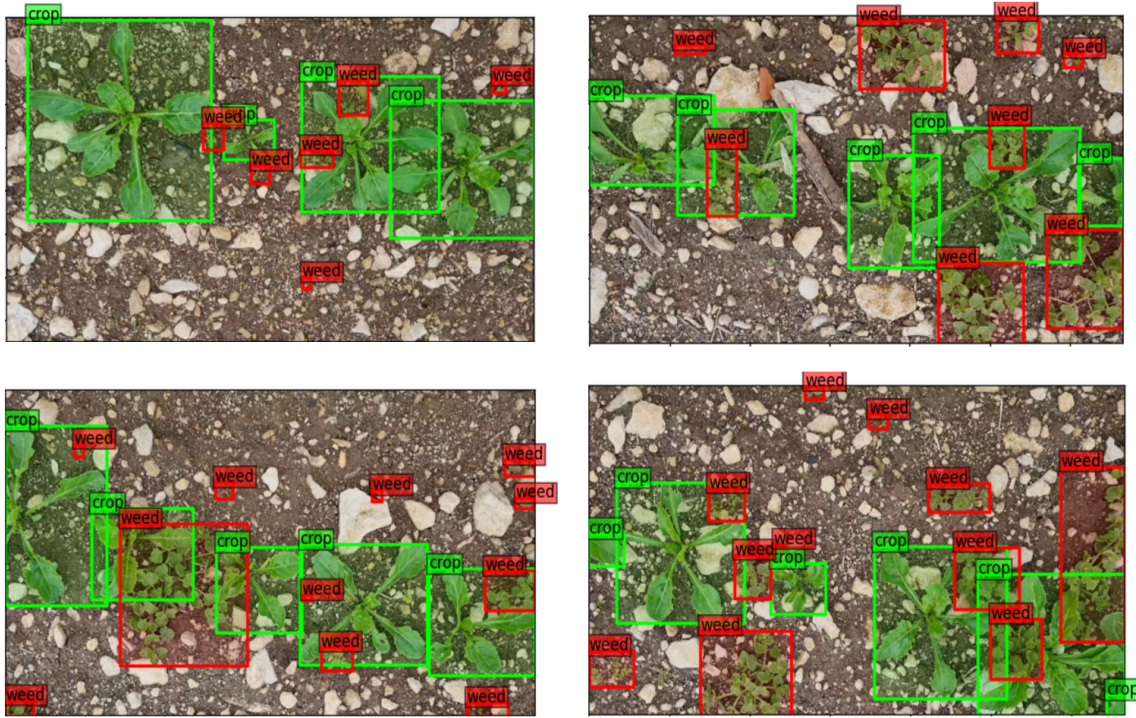


图 1.1 复杂农田场景杂草目标检测面临的典型挑战示意图

技术在复杂农田场景中的应用提供支撑，推动智能感知算法向更高鲁棒性与实用性方向发展，并促进智能农机装备的实际应用。

现有通用目标检测方法在复杂农田场景中仍普遍存在漏检与误检的问题，例如将作物误判为杂草，或在密集区域中产生不稳定的检测结果。以 Faster R-CNN^[1]、YOLO^[2]为代表的卷积神经网络方法虽然显著提升了特征表达能力，但由于主要依赖局部卷积运算，其感受野受限，难以在保持高分辨率特征的同时有效建模全局上下文信息。而近年来兴起的 Transformer 检测框架虽在全局建模能力上具有明显优势，但在复杂杂草场景中的适用性仍面临多方面挑战。现有方法多集中于空间域特征建模，对频域中潜在的判别信息挖掘不足。由于杂草与作物在叶脉边缘及表面纹理上的差异主要体现在高频成分中，因此仅依赖空间域特征在外观相似的情况下难以实现有效区分。同时，Transformer 检测器通常采用固定数量的查询参与解码，难以适应农田中杂草分布密度的剧烈变化，容易在密集区域产生漏检，在稀疏区域又引入不必要的计算开销。此外，杂草形态不规则且伴随遮挡，其边界框中往往混入大量背景或作物像素，使得特征表达本身具有较强的不确定性。在这种情况下，查询与真实目标之间的匹配关系变得更加模糊，传统依赖离散标签分配的训练策略也更容易导致解码过程不稳定。

针对上述现有模型在复杂农业场景中的局限性,本文基于 Transformer 框架提出一种改进的目标检测模型,从特征建模与解码优化两个层面展开研究。在特征建模方面,引入空间域结构信息与频域判别信息的双域融合机制,增强模型对复杂背景与细粒度目标的特征表达能力;在解码优化方面,设计动态查询分配、高阶查询交互及匹配质量自适应监督机制,提升模型在密集场景下的自适应能力与训练稳定性。通过上述两方面的协同优化,本文构建了一种高效的杂草检测方法,有效提升了模型在复杂环境下的检测精度与鲁棒性。

1.2 国内外研究现状

1.2.1 通用目标检测研究现状与分析

目标检测是计算机视觉领域的重要研究方向,其任务是在给定图像或视频中同时完成目标类别识别与空间位置定位。与图像分类等任务相比,目标检测不仅需要区分目标与背景,还需要精确刻画目标在图像中的空间分布,因此对特征表达能力、空间建模能力以及判别决策稳定性提出了更高要求。目标检测方法大体经历了传统手工特征方法、基于卷积神经网络的深度学习方法以及基于 Transformer 的端到端检测方法三个主要发展阶段。

(1) 传统的目标检测算法

传统目标检测算法的特征提取主要依赖人工设计,通常基于模板匹配与滑动窗口策略。其整体检测流程相对固定,主要包括区域选择、特征提取和分类判别三个阶段。在区域选择阶段,通过多尺度滑动窗口或候选区域生成方法对图像进行遍历;随后在特征提取阶段,从候选区域中提取具有判别性的人工特征;最后在分类判别阶段,利用分类器对候选区域进行判断并输出检测结果。

在早期研究中,Viola 等人提出的 VJ 人脸检测算法^[3]通过将 Haar-like 特征与 AdaBoost 分类器相结合,实现了对人脸目标的快速检测,在当时兼顾了较高的检测精度与实时性能。Dalal 等人提出的 HOG 行人检测算法^[4]通过构建方向梯度直方图描述目标的局部结构特征,并结合支持向量机(SVM)分类器,有效提升了行人检测在复杂背景下的鲁棒性。随后,Felzenszwalb 等人提出的可变形部件模型^[5]在 HOG 和 SVM 的基础上引入多尺度部件建模思想,将目标表示为多个具有几何约束关系的可变形部件,从而在一定程度上增强了模型对目标形变和局部遮挡的适应能力。

尽管上述传统目标检测方法在特定任务和相对简单的场景中取得了一定效果，但其整体性能仍存在局限。这类方法对特征设计高度依赖人工经验，当应用场景或目标类别发生变化时，模型的通用性和鲁棒性较差。同时，基于滑动窗口的穷举式搜索策略计算量大，且各检测环节相对独立，缺乏对图像全局信息的统一建模。因此，在目标尺度差异明显、目标密集或背景复杂的情况下，传统检测方法难以在精度与效率之间取得平衡，逐渐难以满足实际应用需求。

（2）基于卷积神经网络的目标检测算法

随着深度学习技术的快速发展，卷积神经网络（Convolutional Neural Networks, CNN）^[6] 依托其层级化特征表示能力和端到端可训练特性，在视觉感知任务中展现出显著优势，推动目标检测技术由依赖人工设计特征与规则约束的传统方法，逐步转向以大规模数据驱动为核心的发展范式。CNN 通过多层卷积与非线性映射结构，能够从原始像素中自动学习由低级纹理特征到高级语义特征的层次化表示，有效提升了目标类别判别与空间定位的鲁棒性。

2012 年，AlexNet^[7] 在 ImageNet^[8] 大规模图像分类任务中取得突破性成果，显著超越了传统方法的性能水平，充分验证了深层卷积网络在复杂视觉特征建模方面的潜力。这一成果不仅推动了图像分类领域的发展，也为后续将深度特征引入目标检测、语义分割等高层视觉任务奠定了重要基础^[9]。在此背景下，研究者开始探索基于 CNN 的目标检测方法，并根据检测流程与预测策略的不同，将其主要划分为两阶段检测方法和单阶段检测方法两大类。

两阶段检测方法以 R-CNN 系列为代表，其基本思想是将目标检测任务分解为候选区域生成与候选区域精细识别两个相对独立的阶段。Girshick 等人提出的 R-CNN^[10] 首次将深度卷积特征引入目标检测框架，通过选择性搜索算法生成候选区域，并对每个候选区域单独进行特征提取与分类，在当时显著提升了检测精度。然而，该方法计算开销较大，推理效率受限。为提升检测效率，SPP-Net^[11] 通过引入空间金字塔池化结构，实现了对不同尺寸候选区域的统一特征表示，避免了重复卷积计算。Fast R-CNN^[12] 进一步采用感兴趣区域池化机制，使候选区域能够共享特征图表示，从而显著提升训练与推理效率。Faster R-CNN 在此基础上引入区域建议网络，实现候选框生成与目标检测的端到端联合训练，成为两阶段检测框架的重要里程碑。随后，Mask R-CNN^[13] 进一步引入像素级掩码预测分支，实现了目标检测与实例分割任务的统一建模，拓展了两阶段检测框架的应用范围。

单阶段检测方法以 YOLO 和 SSD^[14] 为代表, 其核心思想是在特征图上进行密集预测, 直接回归目标类别概率与边界框位置, 从而省略候选区域生成过程, 在推理速度和实时性能方面具有明显优势。YOLO 系列方法通过将目标检测视为一个端到端的回归问题, 实现了较高的检测效率; SSD 通过在多层特征图上进行预测, 引入多尺度检测机制, 在一定程度上缓解了单尺度预测对小目标不敏感的问题。随后, YOLOv2^[15]、YOLOv3^[16] 等改进版本通过引入锚框机制、多尺度特征预测以及更深的网络结构, 在检测精度与实时性能之间取得了更为合理的平衡。针对复杂场景下背景干扰强、目标分布密集等问题, 一些改进方法如 YOLO-ACE^[17] 和 YOLO-EV^[18] 通过引入上下文信息建模、注意力机制以及多层特征融合策略, 增强了模型对关键区域的关注能力, 在复杂背景条件下取得了较好的检测效果。

总体而言, 基于卷积神经网络的目标检测方法在特征自动学习、多尺度建模以及端到端训练方面取得了显著进展。然而, 由于卷积运算本质上具有局部感受野和空间平移不变性, 其对长程依赖关系和全局结构信息的建模能力仍然受限。因此, 在目标高度重叠、外观相似或对比度较低的复杂场景中, 模型难以有效捕捉目标之间的全局关系, 仍容易出现误检或漏检现象。

(3) 基于 Transformer 的目标检测算法

为克服卷积神经网络在全局建模方面的固有局限, Transformer 架构逐渐被引入目标检测领域。Carion 等人提出的 DETR^[19] 首次将 Transformer 引入端到端目标检测框架, 通过引入查询机制与全局自注意力, 实现了无需锚框设计和非极大值抑制的统一检测流程。该方法通过一对一匹配机制建立预测结果与真实目标之间的对应关系, 在结构设计上具有较强的简洁性和理论一致性, 并显著增强了模型对全局上下文关系的建模能力。然而, 由于全局注意力计算复杂度较高, 且查询缺乏有效的空间先验, DETR 在训练初期收敛缓慢, 对小目标和密集目标的检测效果仍然有限。针对这些问题, 后续研究从特征建模、查询机制、注意力结构以及训练策略四个方面进行了改进, 以提升收敛速度、局部特征建模能力和检测性能。

在特征建模方面, Deformable DETR^[20] 通过引入可变形注意力机制, 使每个查询仅关注少量与其位置相关的关键采样点, 在显著降低计算复杂度的同时增强了模型对局部细粒度特征的建模能力, 改善了小目标检测效果。然而, 该方法仍主要基于空间域特征进行建模, 对复杂场景中背景干扰严重、目标外观高度相似等问题的判别能力提升有限。

在查询机制建模方面, DAB-DETR^[21]将查询显式建模为四维锚框形式, 引入更强的位置先验, 使查询在解码初期即可聚焦潜在目标区域, 从而加速模型收敛。Conditional DETR^[22]通过为每个查询生成独立的空间嵌入, 并采用拼接方式与内容嵌入融合, 缓解了原始 DETR 查询难以学习有效空间语义的问题; Conditional DETR v2^[23]进一步通过行列解耦注意力机制提升了查询的表达能力。这类方法在一定程度上增强了查询的空间感知能力, 但其查询数量通常仍保持固定, 难以根据目标数量和分布复杂度的变化进行自适应调整, 在目标密集或稀疏场景中仍可能引入计算冗余或漏检风险。为进一步提高召回率, 一些研究尝试通过查询筛选策略优化查询的使用效率。DDQ^[24]通过引入密集查询并筛选区分性较高的查询参与一对一匹配, 提高了目标召回能力; Saliency DETR^[25]则通过显著性过滤机制, 仅对判别性较高的查询执行 Transformer 编码, 从而减少冗余计算并提升检测效率。

在注意力结构优化方面, 不少方法着重降低全局自注意力的计算复杂度, 以提升推理效率。Anchor DETR^[26]和 Dynamic DETR^[27]通过简化注意力计算或以 CNN 替换编码器结构, 在一定程度上加快了模型推理速度。Efficient DETR^[28]则通过减少编码器和解码器层数并结合关键参考点提取, 间接降低了注意力计算开销, 实现了更高效的检测流程。尽管如此, 这类方法的优化重点主要集中在计算效率层面, 对于复杂场景中目标间结构关系及高阶依赖的建模能力仍显不足, 因此在密集分布或严重遮挡情况下的判别性能提升有限。

在训练策略和匹配机制方面, DN-DETR^[29]通过引入去噪训练, 将带噪声的真实标签和边界框作为辅助输入, 引导模型学习稳定的查询表示; DINO^[30]在此基础上引入对比去噪机制, 进一步增强了模型对无目标查询的区分能力。在匹配策略上, DETR 系列方法逐步由严格的一对一匹配向一对多匹配演进, 如 Group DETR^[31]、DETA^[32]、H-DETR^[33]、Rank DETR^[34]以及 Co-DETR^[35]等方法通过引入多组查询、辅助分支或并行检测头, 有效提升了检测性能和训练稳定性。尽管匹配机制不断得到改进, 但在目标密集且遮挡严重的复杂场景中, 查询与真实目标之间的匹配关系仍具有较强不确定性, 固定的监督强度难以适应匹配质量的动态变化, 解码阶段的预测稳定性问题尚未得到根本解决。

由此可见, 尽管基于 Transformer 的目标检测方法在全局建模和端到端学习方面取得了显著进展, 但在复杂场景下仍面临若干关键挑战。一方面, 现有方法多侧重于空间域特征建模, 对频域等潜在判别信息的挖掘不足, 且在复杂背景和外观相似

条件下难以形成稳定且具有区分性的特征表达；另一方面，在解码阶段，查询的数量与形式通常固定，难以随目标数量及分布复杂度的变化进行自适应调整，同时匹配与监督机制对匹配质量变化的适应能力有限，导致在目标密集或严重遮挡场景中预测稳定性不足。这些局限在一定程度上制约了 Transformer 检测框架在复杂真实场景中的进一步应用。

1.2.2 杂草目标检测研究现状与分析

杂草检测是实现精准农业和变量施药的关键前提。尽管传统目标检测器在通用视觉任务中取得了显著进展，但在农业领域仍面临一系列独特挑战。与通用场景不同，田间环境具有作物与杂草外观高度相似、生长密度极不均匀以及背景干扰严重等特殊特性，这使得杂草检测成为一项极具挑战的任务。

针对上述挑战，早期的杂草检测研究主要基于传统图像处理技术，利用颜色指数、形状因子及纹理特征来区分杂草与作物，并结合支持向量机、随机森林等浅层分类器进行识别^[36]。在具体实现上，一类方法基于可见光图像，利用超绿指数结合 Otsu 阈值分割提取植被区域，并通过形态学开闭运算去除噪声，进一步依据几何特征完成作物与杂草的区分^[37]。另一类方法则利用近红外与可见光之间的反射差异，增强植被与土壤之间的对比度，从而在一定程度上减弱光照变化的影响^[38]。然而，这类基于人工特征的方法对环境变化较为敏感。当光照条件发生变化或土壤背景复杂时，其检测性能容易下降。此外，在作物与杂草处于苗期阶段时，由于二者在颜色、纹理及形态上的差异较小，特征区分能力进一步降低。因此，其在复杂田间环境中的鲁棒性与泛化能力仍然有限^[39]。

随着深度学习技术的普及，基于卷积神经网络的检测模型凭借其强大的特征自动提取能力，被广泛应用于农业领域。早期模型如 WeedNet^[40]，通过结合多光谱成像与语义分割方法实现杂草检测，相较于传统方法在检测性能上取得了显著提升。随后，两阶段和单阶段目标检测框架，包括 Faster R-CNN 和 YOLO 等方法，被逐渐引入农业数据集。研究者一方面利用这些模型解决部分遮挡问题，另一方面利用 YOLO 的高效性满足农机作业的实时需求。尽管基础 CNN 模型在检测具有明显特征的目标时表现良好，但在面对田间小尺度且密集分布的目标时仍存在困难。为缓解这一问题，一系列基于 YOLO 的改进模型被提出。例如，YOLO-ACE^[17]在 YOLOv5 的基础上引入上下文增强和选择性卷积核注意力机制，从而提升模型对小尺度及遮挡杂草

的检测能力。YOLO-EV^[18]则在 YOLOv8 框架上结合先进的注意力模块与优化的特征融合策略，以提高模型在复杂、杂乱田间环境中的检测性能。

然而，上述方法大多仍基于卷积结构，其局部感受野限制了模型对全局上下文信息的建模能力，难以有效刻画田间场景中“簇状分布”和“多尺度共存”的结构特性。这导致模型在杂草密集区域易出现漏检，而在稀疏区域则易产生误检。为弥补卷积模型在长程依赖建模方面的不足，近年来基于 Transformer 的目标检测方法逐渐受到关注。注意力机制能够在全局范围内建立特征关联，从而提升复杂场景下的建模能力。例如，Reedha 等人^[41]将 Vision Transformer 应用于作物与杂草的识别任务，Zhou 等人^[42]利用 Swin Transformer 进行小麦穗检测，并取得了良好效果。尽管 Transformer 在全局建模方面具有天然优势，但现有方法在农业场景中仍存在一定局限。一方面，在低对比度与复杂背景条件下，模型的特征判别能力受限，难以对目标进行充分表征；另一方面，由于对作物与杂草之间细粒度差异的刻画不足，在目标密集且外观高度相似的场景中，解码过程易出现不稳定，从而引发类别混淆。

1.3 研究内容

复杂场景下的杂草目标检测是一项旨在从农业环境中准确识别杂草目标的关键任务。然而，在实际田间场景中，杂草与作物外观高度相似，目标呈现密集且不均匀分布，并伴随严重遮挡。这些复杂特性对检测模型的特征判别能力、查询建模灵活性以及监督机制稳定性提出了更高要求。针对上述问题，本文从两个层面展开研究：一是提升复杂背景下的特征判别能力；二是优化解码阶段的查询建模与监督机制，以应对密集遮挡与动态分布带来的不稳定问题。

本文构建了双域融合的 Transformer 杂草检测模型（Frequency-Spatial Domain Integrated Detection Transformer, FS-DETR），用于增强复杂场景下的特征表征能力；并进一步提出解码增强框架 FS-DETR-D（FS-DETR with Decoder Enhancement），用于缓解复杂场景下解码过程不稳定问题，从而提升整体检测性能与预测稳定性。如图1.2所示，具体研究内容如下：

（1）在复杂场景特征判别方面，提出混合特征融合模块与双域注意力机制，通过协同建模空间域结构信息与频域纹理信息，增强模型对小尺度及外观相似杂草目标的细粒度判别能力。

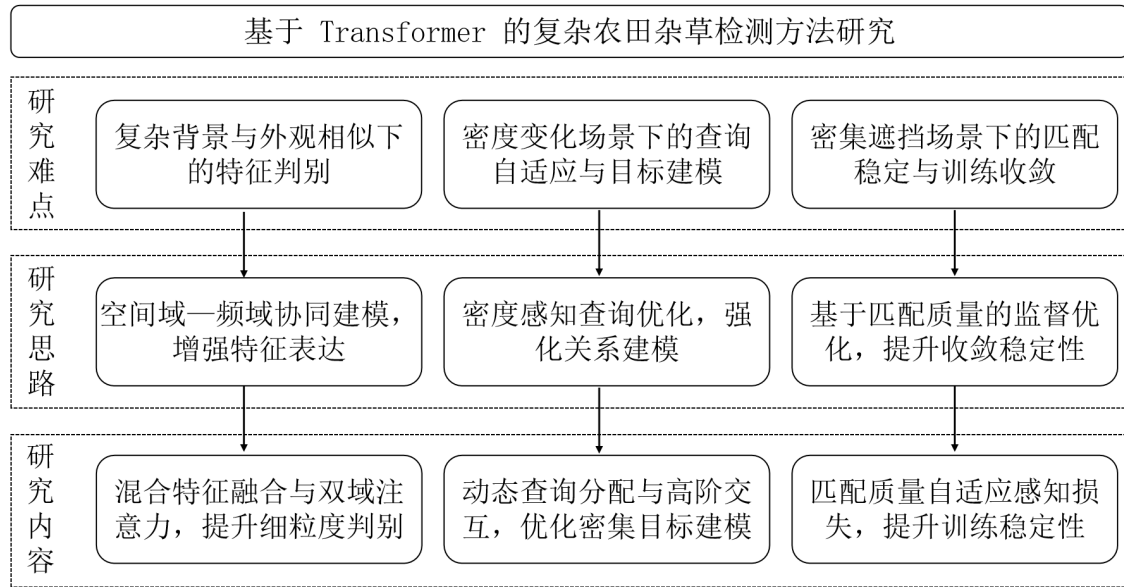


图 1.2 本研究的主要研究内容框架

(2) 在查询自适应与密集目标建模方面，通过引入基于置信度熵与空间感知的动态查询分配策略，实现解码阶段查询资源对场景复杂度的自适应调节；同时引入超边增强的高阶查询交互机制，通过建模查询之间的高阶关系以强化密集区域目标表达，从而降低漏检并提升检测完整性。

(3) 在匹配与监督机制方面，设计匹配质量自适应感知损失函数，从监督角度进一步提升模型的训练稳定性与检测精度；同时提出基于高斯分布与约束引导的标签分配方法，通过优化预测与真实目标之间的匹配过程提升分配合理性。

1.4 论文组织结构

本文共分为五章，整体组织结构如下：

第一章介绍复杂农田场景下杂草目标检测的研究背景与意义，综述国内外相关研究现状，分析现有方法在复杂场景建模中的不足，并阐述本文的研究内容与组织结构。

第二章对本文涉及的关键理论与技术进行系统综述，主要包括基于卷积神经网络与 Transformer 的目标检测方法、多尺度特征融合技术、频域特征建模方法以及基于超边的高阶关系建模技术，并对实验所采用的数据集与评价指标进行介绍，为后续方法设计与实验分析提供理论基础。

第三章介绍了基于双域融合的 Transformer 杂草检测方法。首先概述本文提出的

复杂农业场景杂草检测整体框架，随后介绍所设计的混合特征融合模块、双域注意力机制以及基于高斯分布和约束引导的标签分配策略，最后通过对比实验与消融实验验证所提方法在复杂农业场景中的有效性。

第四章在第三章方法的基础上，进一步围绕解码过程不稳定的问题，提出解码自适应优化的杂草检测方法。首先概述面向复杂场景的解码优化检测框架，随后介绍提出的基于置信度熵与空间感知的动态查询分配策略、超边增强的高阶查询交互机制以及匹配质量自适应感知损失函数的实现，最后在相关数据集上通过实验验证所提方法的有效性。

第五章对全文的研究工作进行总结，并对未来研究方向进行展望。

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，包括基于卷积神经网络与 Transformer 的目标检测方法、多尺度特征融合技术、频域特征建模方法以及基于超边的高阶关系建模技术等内容，并对实验所采用的数据集与评价指标进行介绍。上述内容为后续模型构建与实验设计奠定了坚实的理论基础。

2.1 目标检测技术

2.1.1 基于卷积神经网络的目标检测方法

卷积神经网络在图像分类任务中取得了显著成果，但仅依靠卷积层、池化层和全连接层无法直接实现目标检测。目标检测不仅需要识别图像中目标的类别信息，还需要精确地确定目标在图像中的位置，因此在 CNN 的基础上发展出多种目标检测网络结构。根据是否采用候选区域生成和锚框机制，基于卷积神经网络的目标检测方法可分为两阶段目标检测网络、单阶段目标检测网络以及无锚框目标检测网络。

（1）两阶段目标检测网络

两阶段目标检测网络遵循“由粗到细”的思想，首先生成可能包含目标的候选区域，再对候选区域进行分类和边界框回归，从而实现高精度的检测。早期的 R-CNN 算法首次将 CNN 引入目标检测，通过选择性搜索生成候选区域，并对每个候选区域提取特征，再通过 SVM 分类器进行分类，同时使用回归器优化边界框位置。虽然 R-CNN 开创了深度学习在目标检测领域的应用，但训练过程繁琐，特征提取重复，检测速度较慢。为解决这些问题，Fast R-CNN 提出对整幅图像仅进行一次卷积特征提取，并使用感兴趣区域池化层将候选区域映射为固定尺寸的特征向量，从而实现端到端训练，提高了训练速度和精度。随后，Faster R-CNN 在 Fast R-CNN 的基础上，新增了区域建议网络（Region Proposal Network, RPN）模块。

如图2.1所示，输入图片先经卷积层提取特征图，RPN 会基于该特征图通过滑动窗口生成锚框（anchor），并同步预测候选框的偏移量，以此实现候选区域的端到端生成；生成的候选区域经感兴趣区域池化统一特征尺度后，再通过分类与回归模块完成目标类别的判定与边界框的精准调整。Mask R-CNN 在 Faster R-CNN 的基础上

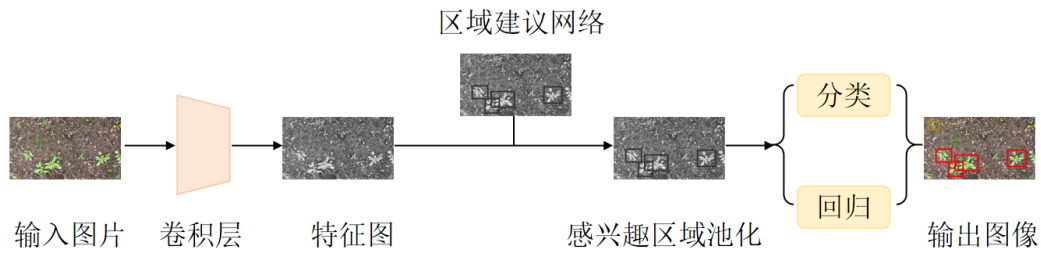


图 2.1 Faster R-CNN 整体框架结构图

进一步加入全卷积掩码分支，实现了实例分割功能，同时通过感兴趣区域对齐解决特征下采样量化导致的信息损失问题。此外，Cascade R-CNN 采用多阶段级联结构逐级增加交并比阈值，提高了正负样本区分能力和候选框质量，从而进一步提升了检测精度。

（2）单阶段目标检测网络

单阶段目标检测网络则通过在输入图像上划分网格，直接预测每个网格中的目标类别和边界框信息，无需候选区域生成，因此检测速度显著提升。YOLO 系列算法是最具代表性的单阶段检测方法，其核心思想是将目标检测任务转化为回归问题，直接在特征图上预测目标类别和边界框位置。

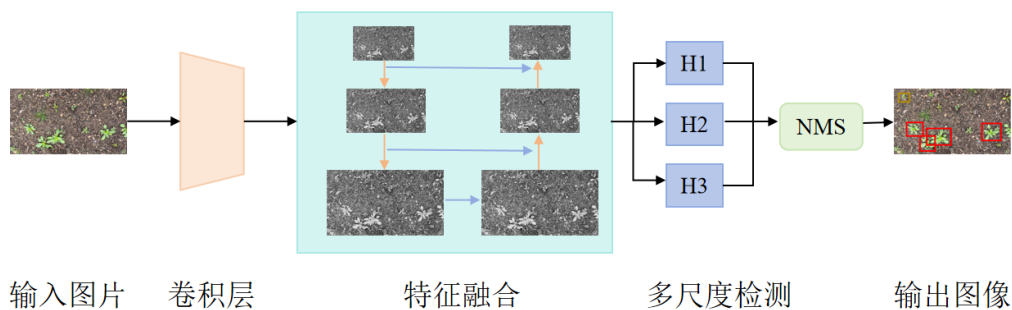


图 2.2 YOLOv5 整体框架结构图

YOLOv1 通过卷积层和全连接层输出预测张量，完成端到端检测，但每个网格预测能力有限，存在小目标漏检问题。YOLOv2、YOLOv3 在网络结构上引入批量标准化、残差连接和特征金字塔网络结构，并采用 anchor 机制提升小目标检测精度。在 YOLOv5 及其后续版本的迭代过程中，模型进一步引入了 Focus 层与路径聚合网络 PAN 结构，对特征处理与融合流程进行了优化。其中，Focus 层通过对输入特征图进行切片与拼接操作，在完成下采样的同时最大程度地保留了原始图像中的细粒度信息，从而提升了浅层特征的表达能力；PAN 结构则在特征金字塔的基础上构建了

自上而下与自下而上相结合的双向特征融合路径，增强了不同层级特征图之间的信息传递与语义互补，有效提升了模型对多尺度目标的特征表示能力。通过上述改进，YOLO 系列模型在目标定位精度和检测性能方面均取得了进一步提升，其整体工作流程如图2.2所示。与 YOLO 系列类似，SSD 通过多层卷积特征图进行多尺度预测，每层负责不同大小目标的检测，并在卷积层上生成多比例 anchor，实现单次前向传播完成目标检测，同时自适应选择多层次 anchor 提升检测精度。

(3) 无锚框目标检测网络

无锚框目标检测网络的核心目标在于解决传统基于锚框方法所存在的固有问题，主要包括对锚框尺寸和长宽比例等先验信息的强依赖、密集锚框采样所带来的冗余计算与存储开销，以及训练过程中正负样本数量严重不平衡等问题。该类方法通过直接预测目标的关键点位置或边界框相关参数，摆脱了锚框机制的约束，从而实现更加灵活和高效的目标定位方式。

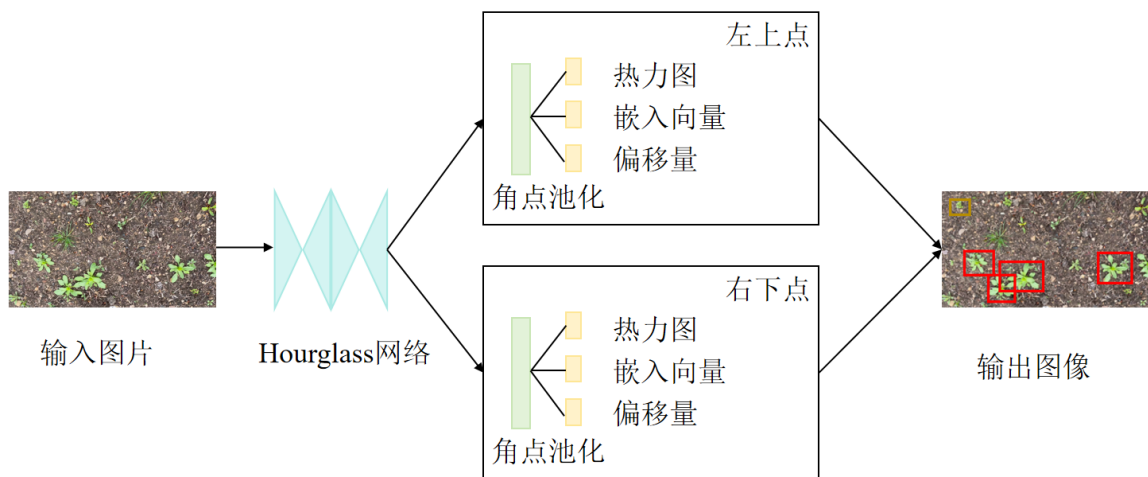


图 2.3 CornerNet 整体框架结构图

以 CornerNet 为代表的无锚框检测方法，通过预测目标左上角和右下角两个关键点的位置偏移来构建目标边界框。具体而言，输入图像首先经 Hourglass 网络提取特征，随后分别生成左上角和右下角关键点对应的热力图、嵌入向量以及偏移量。其中，引入的角点池化操作能够沿目标边界方向聚合全局上下文信息，从而增强关键点的特征表达能力并提升边界框的定位精度，其整体工作流程如图2.3所示。此外，YOLOX 在 YOLO 系列基础上引入无锚框检测器，并采用解耦的分类与回归模块，将分类任务和边界框回归任务分别处理，不仅降低了网络复杂度，还显著提升了检测性能。

2.1.2 基于 Transformer 的目标检测方法

传统基于卷积神经网络的目标检测方法在局部特征提取方面表现优异，但在复杂场景下对全局依赖的建模以及小目标的检测仍存在一定局限。近年来，Transformer 在自然语言处理领域取得显著成功，其自注意力机制能够有效捕捉序列中的全局信息。这一特性逐渐被引入计算机视觉任务，尤其是目标检测，使模型能够在特征图上获取全局上下文，从而弥补卷积操作在全局建模方面的不足，并提升对复杂场景和小目标的理解能力。

(1) Transformer 的基本结构与机制

Transformer 模型是一种基于自注意力机制的深度学习架构，由 Vaswani^[43]等人于 2017 年提出。与传统循环神经网络和长短期记忆网络依赖序列化计算不同，Transformer 完全基于自注意力机制对输入序列进行建模，从而能够高效捕捉长距离依赖关系，并显著提升并行计算能力。该模型通过多层注意力结构与前馈神经网络的堆叠，实现对输入特征中全局上下文信息的建模，为复杂关系建模提供了有效手段。

其中，自注意力机制是 Transformer 的核心组成部分，其通过计算序列中各元素之间的相关性，自适应地分配不同位置的关注权重，从而在处理当前元素时综合考虑全局信息。设输入序列为 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ ，经嵌入层映射后得到向量表示 $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ 。通过线性变换将输入特征映射为查询 (Query)、键 (Key) 和值 (Value) 向量，分别记为 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ ：

$$\mathbf{Q} = \mathbf{E} \cdot \mathbf{W}_Q \quad (2.1)$$

$$\mathbf{K} = \mathbf{E} \cdot \mathbf{W}_K \quad (2.2)$$

$$\mathbf{V} = \mathbf{E} \cdot \mathbf{W}_V \quad (2.3)$$

其中， $\mathbf{W}_Q$ 、 $\mathbf{W}_K$ 和 $\mathbf{W}_V$ 为可学习的线性投影矩阵，用于映射特征并建模其相关性与信息交互。

随后通过点积计算查询向量与键向量之间的相似度，并进行缩放处理：

$$\mathbf{S} = \frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}} \quad (2.4)$$

其中， d_k 表示键向量的维度，用于抑制点积值过大带来的梯度不稳定问题。

对相似度矩阵进行 Softmax 归一化可得到注意力权重：

$$\mathbf{W} = \text{Softmax}(\mathbf{S}) \quad (2.5)$$

最终，利用注意力权重对值向量进行加权求和，得到自注意力输出：

$$\mathbf{Output} = \mathbf{W} \cdot \mathbf{V} \quad (2.6)$$

上述过程可统一表示为：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} \quad (2.7)$$

自注意力机制不受元素间距离限制，能够同时建模局部与全局依赖关系，为后续将 Transformer 引入视觉任务奠定了理论基础，其计算流程示意如图2.4所示。

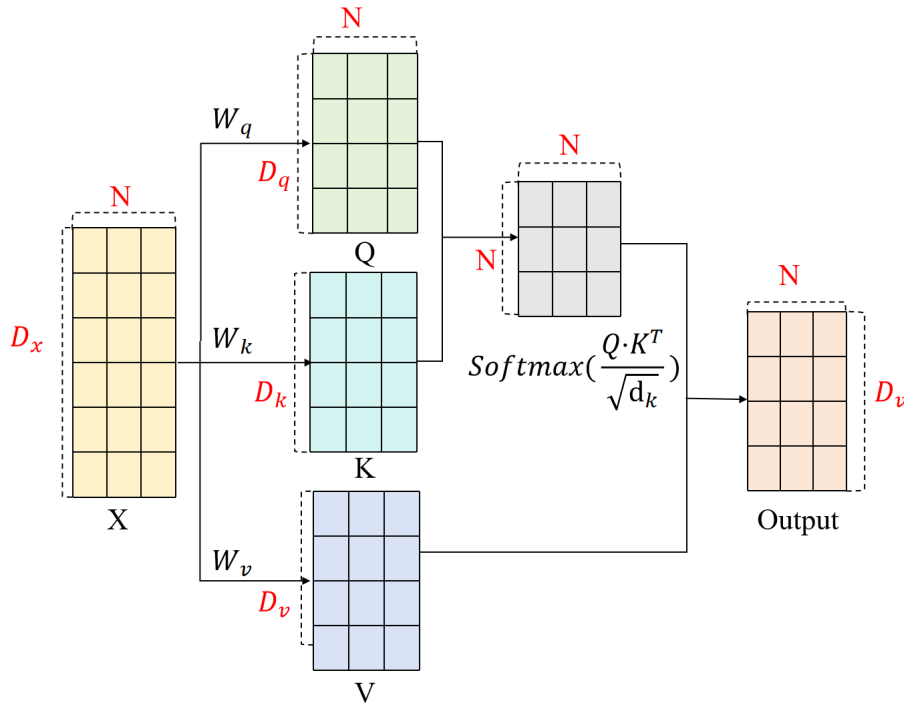


图 2.4 自注意力机制计算流程图

尽管自注意力机制能够有效建模全局信息，但其仅在单一表示空间中进行特征交互，表达能力仍存在一定局限。为了增强模型在不同特征子空间中的表达能力，Transformer 引入多头自注意力机制。该机制通过并行计算多个自注意力头，使模型能够从不同角度关注输入特征中的多种关系。设注意力头数为 h ，则多头注意力的计算形式为：

$$\text{MH}(\mathbf{E}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot \mathbf{W}_o \quad (2.8)$$

其中, $\text{head}_i = \mathbf{A}_i(\mathbf{E})$ 表示第 i 个注意力头对输入特征 $\mathbf{E}$ 的映射结果, $\mathbf{A}_i(\cdot)$ 表示对应的注意力计算过程, $\mathbf{W}_o$ 为多头特征拼接后的线性映射矩阵。

多头自注意力能够在不同表示子空间中建模多样化的依赖关系, 有助于提升模型对复杂结构和上下文信息的表达能力, 其结构示意图如图2.5所示。

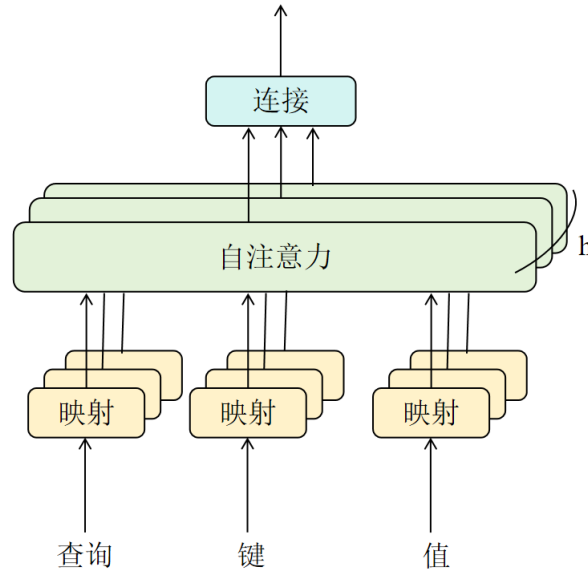


图 2.5 多头自注意力机制结构图

在此基础上, Transformer 将多头自注意力机制与前馈神经网络等模块进行有机组合, 构建了完整的编码器解码器整体架构。模型由编码器和解码器两部分组成, 均由若干相同结构的层堆叠而成, 图中以 “ $n \times$ ” 表示重复次数, 其总体结构如图2.6所示。

编码器用于对输入序列进行特征建模与表示学习。具体而言, 输入序列首先通过嵌入层映射为高维特征, 并与位置编码相结合以引入位置信息。随后, 特征进入编码器层, 每层由多头自注意力机制、前馈神经网络、残差连接和层归一化依次组成, 并重复堆叠 n 次。多头自注意力机制能够建模序列中不同位置的全局依赖关系, 残差连接和层归一化有助于缓解梯度消失与梯度爆炸, 提高训练稳定性和收敛速度。最终, 编码器输出融合全局上下文信息的高层特征表示。

解码器在编码器输出的基础上生成目标序列的预测结果。目标序列同样经过嵌

入层与位置编码处理，然后依次通过掩蔽多头自注意力模块、跨注意力模块、前馈神经网络、残差连接与层归一化，每个解码器层重复 n 次。掩蔽自注意力机制屏蔽未来时刻信息，保证解码过程中仅依赖已生成的历史序列，避免信息泄露；跨注意力模块以编码器输出作为键和值，使解码器能够利用全局上下文信息。最终，解码器输出通过全连接层映射，得到目标序列预测结果。

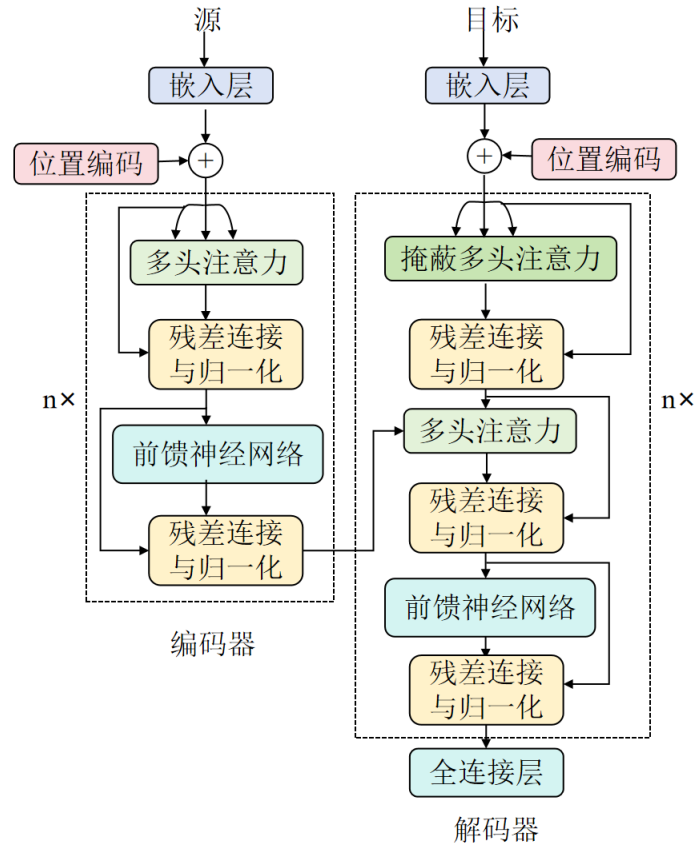


图 2.6 编码器和解码器结构图

(2) Transformer 在目标检测中的引入

Transformer 的核心优势在于自注意力机制，它能够对输入特征序列的任意位置进行信息交互。在目标检测中，将图像特征展平为序列后，Transformer 可以有效建模目标之间的全局关系，弥补卷积操作感受野有限的问题。此外，Transformer 具有天然的可扩展性和端到端训练能力，使其在复杂场景下的小目标检测表现出更强的适应性。例如，DETR 首次将全连接注意力引入目标检测，将目标检测问题转化为集合预测问题，直接回归目标类别和边界框，实现端到端检测，无需传统的候选区域生成和非极大值抑制操作。

(3) DETR 的端到端目标检测框架

DETR 的网络框架主要由特征提取 Backbone、Transformer 编解码模块、预测头以及匹配模块四个部分构成。首先，输入图像通过基于 CNN 的 Backbone 网络提取多尺度特征表示，并将其展平为特征序列；随后引入位置编码以补偿序列化过程中丢失的空间位置信息，并共同送入 Transformer 编码器。编码器利用多头自注意力机制对特征序列进行全局上下文建模，生成包含丰富语义信息的全局特征表示。

随后，Transformer 解码器以一组可学习的目标查询向量作为输入，通过与编码器输出特征之间的跨注意力机制，逐步聚焦于潜在目标区域。解码器的输出被送入由前馈网络等组成的预测头，同时完成目标类别判别以及边界框位置回归。最后，采用匈牙利匹配算法在预测结果与真实标注之间建立一一对应关系，并在训练阶段联合优化分类损失与边界框回归损失，从而实现无需候选框和非极大值抑制的端到端目标检测。DETR 的整体工作流程如图2.7所示。

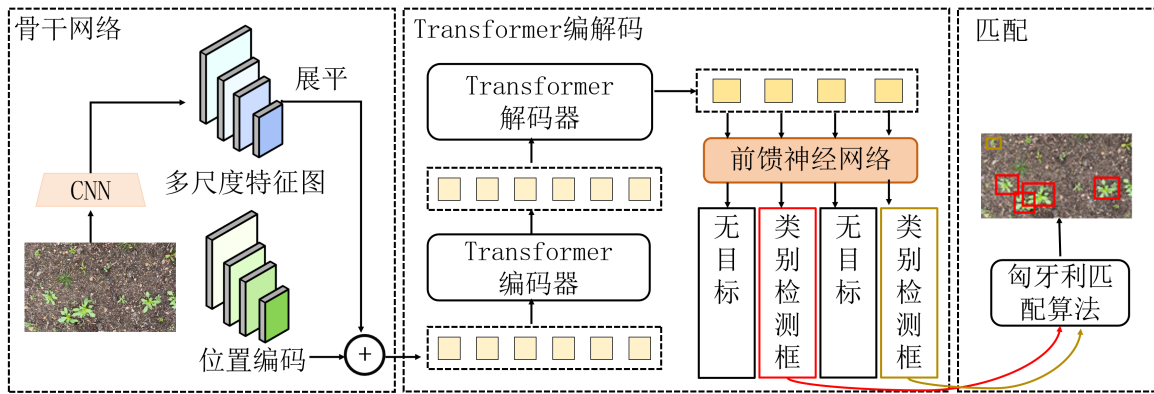


图 2.7 DETR 网络结构图

与传统基于 CNN 的目标检测方法相比，DETR 不依赖预定义锚框或候选区域生成机制，将目标检测统一建模为集合预测问题，从而显著简化了检测流程。同时，借助 Transformer 自注意力机制的全局建模能力，DETR 能够有效刻画目标之间的上下文关系，在复杂背景和多尺度目标场景中具有一定优势。然而，DETR 也存在明显局限性。由于其直接基于全局注意力和随机初始化的目标查询向量进行预测，模型在训练初期难以快速建立稳定的匹配关系，导致收敛速度较慢，通常需要较长的训练周期；此外，在小目标密集或尺度变化较大的场景中，其对局部细节特征的建模能力相对不足，检测性能往往不及成熟的卷积式检测方法。针对上述问题，后续研究提出了 Deformable DETR、DAB-DETR 等改进方案，通过引入多尺度特征表示和可变形注意力机制，有效提升了模型的训练效率和小目标检测精度。

2.2 多尺度特征融合技术

多尺度特征融合是目标检测中用于提升模型对尺度变化目标感知能力的重要技术，其核心在于整合不同层级特征中的空间细节信息与高层语义信息，从而构建具有更强表达能力的特征表示。在复杂场景中，目标通常呈现尺度分布不均、小目标占比高以及多尺度目标共存等特点，单一尺度特征难以完整描述目标属性，因此多尺度特征融合成为目标检测中的关键组成部分。

在多尺度特征建模中，特征金字塔结构（Feature Pyramid Network, FPN）^[44]提供了一种经典的融合范式。FPN 通过自顶向下的路径与横向连接机制，将高层语义信息逐层传递至低层特征，使高语义、低分辨率特征与低语义、高分辨率特征实现有效融合，从而在不同尺度上同时保留语义信息与空间细节，其结构示意图如图2.8所示。具体而言，FPN 利用主干网络提取多层级特征，其中高层特征语义丰富但分辨率较低，低层特征细节充足但语义较弱。在融合过程中，当前层通过横向连接引入低层特征并进行通道对齐，同时结合上层特征的上采样结果进行逐元素相加，实现语义与细节的协同表达。基于自顶向下路径与横向连接，FPN 构建了统一的多尺度特征表示，从而提升不同尺度目标的检测能力。

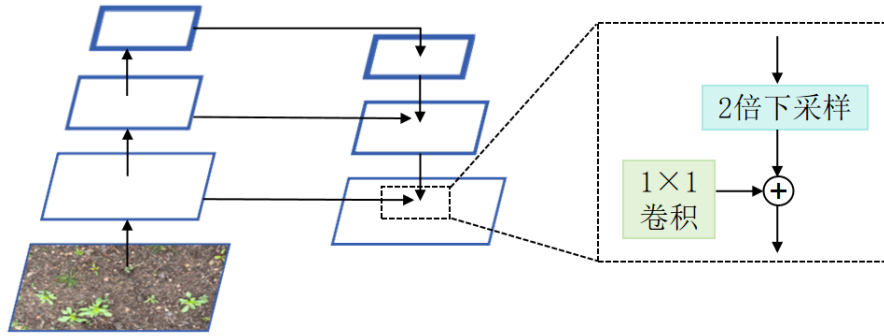


图 2.8 FPN 多尺度特征融合结构图

尽管 FPN 在多尺度特征融合方面取得了良好效果，但其融合路径仍相对固定，难以充分适应复杂场景中多尺度目标的多样性。基于此，研究者在 FPN 框架基础上围绕特征融合路径与信息流动方式进行了进一步优化。

在信息流动建模方面，相关方法通过扩展特征传递路径以增强跨层信息交互能力。例如，PANet^[45]在自顶向下路径之外引入自底向上的路径增强结构，使底层特征中的定位信息能够更直接地传递至高层，从而缩短信息传播路径并缓解特征衰减问

题。在此基础上，部分方法进一步构建双向信息流动结构，使不同层级特征能够在多个方向上进行充分交互。

在融合机制建模方面，研究重点由固定的逐元素融合逐步转向自适应特征融合。**BiFPN**^[46]通过引入可学习的加权机制，对不同尺度特征的重要性进行建模，使融合过程能够根据特征贡献度进行动态调整，从而在保证性能的同时提升计算效率。**NAS-FPN**^[47]则将特征融合结构设计转化为神经架构搜索问题，通过自动探索多尺度特征之间的连接关系，获得性能更优的融合拓扑结构。与此同时，**TFPN**^[48]与**EFPN**^[49]分别从跨尺度信息交互增强与特征精炼角度出发，进一步提升融合特征的一致性与表达能力。随着研究的深入，多尺度特征融合逐渐由结构设计优化转向更加灵活的融合策略建模。相关方法通过引入动态权重学习或结构自适应机制，使模型能够根据输入特征分布对不同尺度信息进行灵活组合。其中，**Dynamic Feature Fusion**^[50]通过动态预测融合权重，实现对多尺度特征的自适应调节，从而在目标密集或尺度变化显著的复杂场景中获得更具判别性的特征表示。

总体而言，多尺度特征融合通过在不同层级特征之间建立有效的信息交互机制，使模型能够同时利用全局语义信息与局部细节信息，为目标检测任务提供了重要的特征表示基础。

2.3 频域特征建模技术

频域特征建模是近年来视觉任务中逐渐受到关注的重要研究方向，其核心思想在于将图像从空间域映射至频域，在全局频率空间中对图像的结构信息与纹理分布进行建模。相比仅在空间域进行特征学习的方法，频域表示能够显式地对不同频率成分进行解耦，其中低频信息主要反映图像的整体结构与轮廓，而高频信息则主要对应边缘与细节纹理特征。因此，频域特征在细节增强、噪声抑制以及特征判别能力提升等方面具有重要优势。

如图2.9所示，输入图像首先通过傅里叶变换被映射至频域空间，在频谱域中通过低通与高通滤波操作实现频率分解，从而得到低频分量与高频分量。随后，两类频率分量分别通过逆傅里叶变换重建回空间域，形成对应的低频特征与高频特征表示，用于后续特征建模与融合。在复杂农业场景中，高频特征对于刻画杂草目标的细粒度纹理与局部结构具有更强的适应性与表达能力。

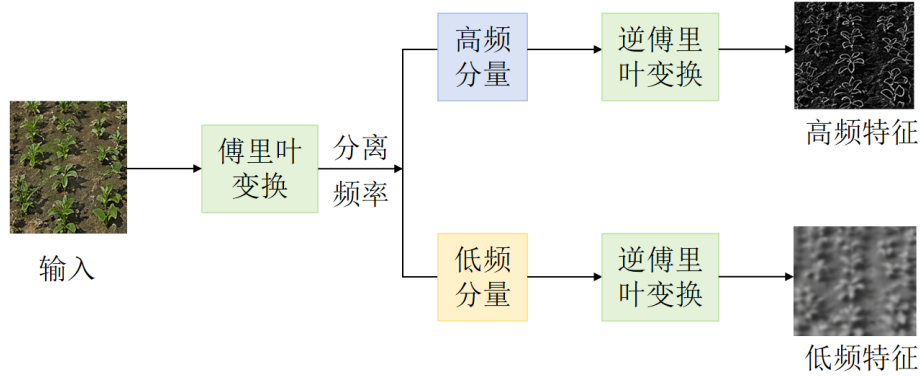


图 2.9 傅里叶变换频域分解与重建示意图

在理论建模层面，频域方法通常基于傅里叶变换等数学工具，将图像表示为不同频率分量叠加的形式，为多尺度信息分析与全局结构建模提供了统一的表达框架。基于该表示方式，早期研究主要集中于频域特征的显式建模与调控。例如，SPAM^[51]通过引入频域掩模对不同频带进行自适应加权，实现高低频信息的平衡表达；FECNet^[52]通过频率分解机制对图像的全局结构与局部细节关系进行建模，从而增强特征表达能力。近年来，随着深度学习的发展，频域建模逐渐由独立特征设计转向与空间域特征的联合建模与协同融合。在该方向上，研究重点不再局限于频域信息的单独利用，而是更加关注跨域信息交互机制的设计。例如，SFDG^[53]通过动态图建模方式显式刻画空间域与频域之间的依赖关系，实现多域特征的自适应融合；在 Transformer 框架下，HAT^[54]进一步强化高频信息建模能力，通过引入频域增强机制提升模型对局部纹理细节的表达能力。

总体而言，频域特征建模通过引入频率域信息补充了传统空间域特征表达的不足，在结构建模与细节表征方面具有重要作用，并逐渐发展为与空间域特征学习相互补充的重要技术路径。

2.4 基于超边的高阶关系建模技术

在复杂视觉场景中，目标通常呈现密集分布与频繁遮挡，不同目标在空间位置与语义属性上存在显著关联。传统方法多依赖局部特征或成对关系建模，难以刻画多目标之间的群体级依赖关系，从而限制了模型在复杂场景下的表达能力。为此，引入超图结构对高阶关系进行建模。与普通图仅描述两两关系不同，超图通过超边连接多个节点，可直接刻画多目标之间的联合关系，实现由“成对建模”向“群体建模”的扩展^[55-56]。相关研究表明，超图结构在复杂场景的群组关系建模中具有显著

优势，并在结构表示与特征传播等方面得到广泛验证^[57-62]。

在结构表示上，超图通过超边将多个相关节点组织为统一的关系单元。如图2.10所示，在普通图结构中，节点 n_1 至 n_5 之间的关系通过多条成对边分别建模，例如边 e_1 、 e_2 、 e_3 和 e_4 仅连接两个节点，每条边只能刻画局部关联，多节点关系需要依赖多条边间接表达。相比之下，在超图结构中，超边能够同时连接多个节点，例如超边 e_1 连接节点 n_1 、 n_3 和 n_6 ，超边 e_2 连接节点 n_2 、 n_4 和 n_5 ，从而直接刻画多个节点之间的整体关系。该表示方式将分散的成对关系转化为统一的群组关系表达，使结构更加紧凑且更适用于描述复杂场景中的多目标关联。

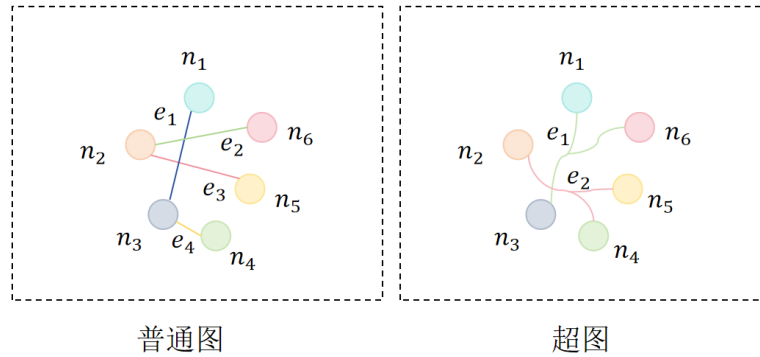


图 2.10 普通图与超图结构对比图

在数学建模层面，为对上述结构进行形式化描述，超图通常采用关联矩阵 H 描述节点与超边之间的关系，其中 $H_{i,m}$ 表示节点 i 与超边 m 的关联强度。基于该表示，节点特征可通过“节点—超边—节点”的信息传递过程实现高阶关系建模，即通过对同一超边内多个节点特征进行联合聚合，并将聚合结果反馈至相关节点，从而增强节点对群体结构信息的感知能力。

在基于 Transformer 的目标检测框架中，目标通常以查询形式进行建模。通过在查询层面引入高阶关系建模机制，有助于增强模型对多目标结构依赖关系的刻画能力，从而提升其在复杂场景中的表达能力。因此，基于超边的高阶关系建模为复杂场景下的结构感知提供了有效支撑，并为后续查询交互机制的设计奠定了理论基础。

2.5 数据集和指标

2.5.1 数据集

近年来,随着深度学习在农业视觉领域中的广泛应用,多个面向作物与杂草识别任务的公开数据集相继发布,为相关算法的训练与性能评估提供了重要数据基础。与早期在受控环境下采集的小规模农业图像数据不同,近年来发布的数据集在目标类别、图像规模以及场景复杂度方面均有显著提升,能够更真实地反映田间环境中作物与杂草共存、目标密集分布以及背景复杂等特点。

为全面评估本文提出方法在复杂农业场景中的检测性能,本文选取 Weed-Crop^[63]、Lincolnbeet^[64]和 MH-Weed16^[65]三个典型农业数据集进行实验验证。这些数据集均来源于真实田间环境,包含作物与杂草混合分布、多目标密集以及尺度变化显著等复杂检测场景,能够有效评估模型在农业视觉任务中的鲁棒性与检测精度。

(1) WeedCrop 数据集

WeedCrop 数据集包含 2,822 张图像,涵盖受控与田间环境及多个生长阶段,涉及甜菜、胡萝卜、西葫芦、南瓜、萝卜和黑萝卜六类作物。数据采用方向校正与增强策略,包括对图像以相同概率进行 0°、顺时针 90° 以及逆时针 90° 的随机旋转,在 -15° 至 $+15^{\circ}$ 范围内施加随机水平和垂直剪切变换,以及在 -25° 至 $+25^{\circ}$ 范围内进行随机亮度调整。该数据集共划分为 2,469 张训练图像、235 张验证图像和 118 张测试图像,并提供了详细的实例级标注信息,共包含 8,894 个作物目标边界框和 14,025 个背景干扰目标边界框。图2.11给出了 WeedCrop 数据集的部分典型样本。

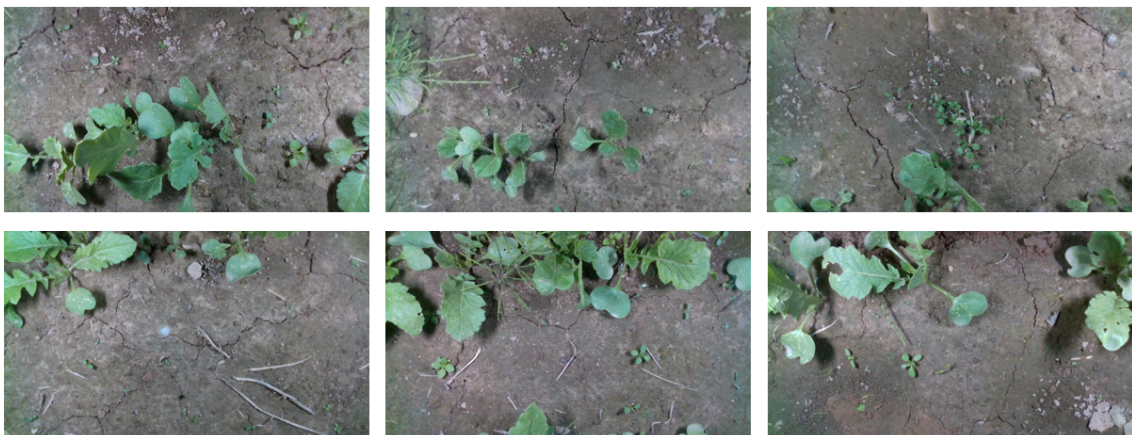


图 2.11 WeedCrop 数据集示例图

(2) Lincolnbeet 数据集

Lincolnbeet 数据集由 4,405 张高分辨率图像构成，图像分辨率为 1920×1080 像素，均从英国林肯地区三个不同甜菜田块在不同时期采集的视频序列中截取获得。不同田块在土壤条件、植株分布及目标类别组成等方面存在明显差异，进一步增强了数据集的场景多样性和复杂性。该数据集划分为 3,521 张训练图像、440 张验证图像和 441 张测试图像，包含 16,399 个作物目标边界框和 22,847 个背景干扰目标边界框。图 2.12 展示了 Lincolnbeet 数据集中的代表性复杂场景样本。



图 2.12 Lincolnbeet 数据集示例图

(3) MH-Weed16 数据集



图 2.13 MH-Weed16 数据集示例图

MH-Weed16 是一个大规模真实场景数据集，完整数据集共包含 25,972 张田间图像，其中包括 18,395 张单目标样本图像以及 7,577 张多目标混合场景图像，数据由 Intel RealSense、Canon 以及 iPhone 等多种成像设备采集得到。在本文实验中，选取

了 MH-Weed16 数据集中标注最为完整的子集，即 6,656 张基于 IntelRealSense 设备采集的多目标场景图像。该子集被划分为 5,324 张训练图像、666 张验证图像和 666 张测试图像，涵盖 15 类细粒度目标。图2.13展示了 MH-Weed16 数据集的部分样本示例。

2.5.2 目标检测评价指标

在目标检测任务中，评价指标用于从分类性能、定位精度及运行效率等多个维度对模型进行定量分析，是衡量检测算法性能优劣的重要依据。本文选取目标检测任务中具有代表性的评价指标，对模型性能进行系统分析。

精确率（Precision）与召回率（Recall）是衡量模型分类性能的基础指标，分别用于刻画误检与漏检情况。其定义如下：

$$Precision = \frac{TP}{TP + FP} \quad (2.9)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.10)$$

其中， TP 表示被正确预测为正样本的数量， FP 表示被错误预测为正样本的负样本数量， FN 表示被错误预测为负样本的正样本数量。

在定位精度评估方面，引入交并比（Intersection over Union, IoU）用于衡量预测边界框与真实边界框之间的重叠程度，其定义为两者交集面积与并集面积之比：

$$IoU = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \quad (2.11)$$

其中， B_p 表示预测边界框， B_{gt} 表示真实边界框。IoU 的取值范围为 $[0, 1]$ ，数值越大表示定位越准确，当其超过预设阈值时，通常认为该检测结果为正确匹配。

在综合评价方面，平均精度（Average Precision, AP）通过精确率-召回率（PR）曲线对模型性能进行刻画。在给定交并比阈值 τ 的条件下，其定义为 PR 曲线下的面积：

$$AP_\tau = \int_0^1 P_\tau(R) dR \quad (2.12)$$

其中， R 表示召回率， $P_\tau(R)$ 表示在 IoU 阈值为 τ 时，通过对检测结果按置信度排序并逐步累积得到的对应精确率。

为提高评估结果的稳定性与全面性，本文采用与 COCO 数据集一致的评测标准，该指标能够综合反映模型在不同定位精度要求下的检测性能。具体而言，在多个 IoU 阈值下对 AP 进行平均计算，其中 IoU 阈值范围设定为 0.50 至 0.95，步长为 0.05，其定义为：

$$AP = \frac{1}{|\gamma|} \sum_{\tau \in \gamma} AP_{\tau}, \quad \gamma = \{0.50, 0.55, \dots, 0.95\} \quad (2.13)$$

其中， $|\gamma|$ 表示 IoU 阈值集合 γ 中的元素个数， AP_{τ} 表示在 IoU 阈值为 τ 时计算得到的平均精度。

在多类别目标检测任务中，进一步采用平均精度均值（mean Average Precision, mAP）对模型整体性能进行评估，其定义为所有类别 AP 的算术平均：

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (2.14)$$

其中， N 表示目标类别的总数， AP_i 表示第 i 类目标的平均精度。mAP 的取值范围为 $[0, 1]$ ，数值越大表明模型整体检测性能越优。

此外，为进一步分析模型在不同检测条件下的性能表现，本文还引入 AP_{50} 、 AP_{75} 以及针对不同尺度目标的评价指标，包括 AP_S 、 AP_M 和 AP_L ，分别用于评估宽松定位、严格定位以及小、中、大目标的检测性能。

除检测精度外，模型运行效率同样是衡量其实用性的重要因素。通常采用每秒处理帧数（Frames Per Second, FPS）来衡量检测速度，其数值越高表示模型实时性越强。总体而言，检测精度与推理速度之间存在一定的权衡关系，因此在实际应用中需要结合具体任务需求进行综合考虑。

第三章 面向复杂场景的双域 Transformer 杂草检测模型

3.1 引言

复杂农业环境中的杂草检测是实现精准农业与智能除草的重要基础，对于提高作物产量和降低环境污染具有重要意义。传统依赖化学除草剂的方法虽具有较高效率，但其非选择性喷洒易引发生态退化与环境污染问题。因此，面向可持续农业发展的精准检测与机械除草技术逐渐成为研究热点。然而，在实际田间场景中，受外观相似、分布密集不均及遮挡等因素影响，杂草检测易出现漏检、误检及预测不稳定问题，对模型的精细判别能力与鲁棒性提出了更高要求。

针对上述问题，现有方法主要基于卷积神经网络与 Transformer 检测框架进行改进。基于卷积神经网络的方法通常依赖锚框或无锚框机制实现目标检测，代表性方法包括 Faster R-CNN、YOLO 等。这类方法具备较高的推理效率，但其特征建模主要依赖局部卷积操作，难以有效捕获长距离依赖关系，在目标密集或外观高度相似的复杂农业场景中容易受到背景噪声干扰，从而产生误检与漏检。此外，无锚框方法在小尺度目标定位方面仍存在一定不足。为克服卷积网络在全局建模上的局限，基于 Transformer 的目标检测方法逐渐受到关注，典型代表为 DETR 及其系列模型。该类方法通过自注意力机制引入全局上下文信息，并将目标检测建模为集合预测问题，在目标关系建模方面具有优势。然而，由于缺乏空间归纳偏置，其在小目标定位、细粒度边界建模以及复杂场景稳定性方面仍存在不足，尤其在外观高度相似与高遮挡场景下容易出现定位不准确及前景背景混淆的问题。近年来，研究者针对上述问题进行了大量改进，在一定程度上提高了目标定位精度与检测召回能力。然而，多数方法仍主要依赖空间域特征表达，在外观高度相似场景下对细粒度差异信息的建模能力有限。此外，密集查询或锚框先验更多侧重于提升目标召回率，而对精细边界区分及前景与背景特征解耦的改善相对有限。

为此，本文在 Transformer 检测框架基础上提出一种融合频域与空间域特征的杂草检测方法（Frequency-Spatial Domain Integrated Detection Transformer, FS-DETR）。该方法构建混合特征融合模块，实现空间结构信息与频域判别特征的协同建模；在编码阶段引入双域注意力机制，将频域注意力与空间域可变形注意力自适应融合，以

增强模型对复杂背景、小尺度目标及外观相似目标的表征能力。此外，针对目标密集及样本分配不均问题，设计基于高斯分布与约束引导的标签分配策略，以提升训练稳定性并改善整体检测性能。

3.2 模型整体框架

本章提出的双域融合 Transformer 杂草检测模型 FS-DETR 如图3.1所示，该模型主要由骨干网络、混合特征融合模块（Hybrid Feature Fusion, HFF）、双域注意力（Dual Domain Attention Mechanism, DDAM）编码器、多层 Transformer 解码器以及基于高斯分布与约束引导的标签分配模块（Gaussian Distribution-based and Constraint-guided Label Assignment, GCLA）组成。

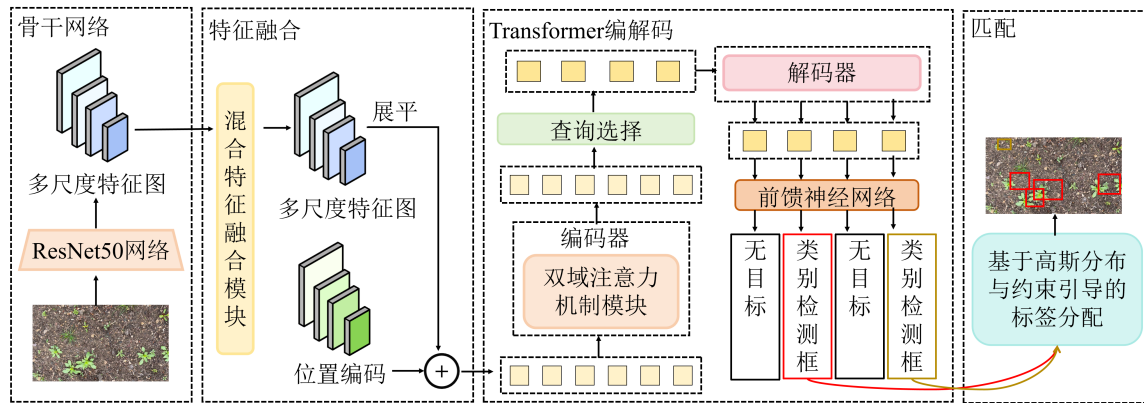


图 3.1 FS-DETR 整体架构图

在特征提取阶段，模型采用 ResNet-50 作为骨干网络，对输入图像进行多尺度空间特征提取，为后续特征建模提供基础表示。随后，引入 HFF 模块，将频域高频信息与多尺度空间特征进行协同融合，以增强细粒度纹理与多尺度目标的表达能力，从而提升复杂背景下的特征判别性。在进入 Transformer 编码器之前，引入位置编码以补充特征的空间位置信息。在编码阶段，设计 DDAM 模块，同时建模空间域结构信息与频域判别信息，使特征在全局上下文建模的同时保留对局部纹理差异的敏感性，从而提升模型对小尺度杂草及外观高度相似目标的区分能力。编码器输出融合全局语义上下文的特征序列，并作为解码器输入。在解码阶段，通过查询选择策略生成初始目标查询，并输入多层 Transformer 解码器进行逐层交互更新，逐步细化类别预测与边界框回归结果，实现由粗到细的目标定位优化。在匹配阶段，引入 GCLA 模

块，对预测框与真实标注之间的匹配关系进行建模，以实现更稳定且合理的正负样本分配，从而提升模型收敛稳定性与检测精度。

3.3 混合特征融合模块

为有效应对复杂场景下细粒度目标检测的挑战，本章节提出混合特征融合模块，其结构如图3.2所示。现有多尺度特征融合方法^[44-45]主要在空间域内进行跨层语义交互，其核心目标是增强不同尺度特征之间的信息传递。然而，此类方法通常缺乏对频率结构信息的显式建模，难以有效保留复杂农业场景下杂草目标的高频纹理与边缘细节。针对这一局限，本文提出的 **HFF** 模块在浅层特征阶段引入频域选择分支，通过提取高频分量增强对边缘轮廓与叶片纹理的表达能力，同时结合分层多尺度特征融合，从低层到高层整合不同尺度的空间特征。通过频域信息与多尺度空间特征的协同建模，**HFF** 模块能够从频域和空间域两个维度同时捕捉目标的细粒度细节，显著提升模型在复杂场景下的目标定位与识别精度。

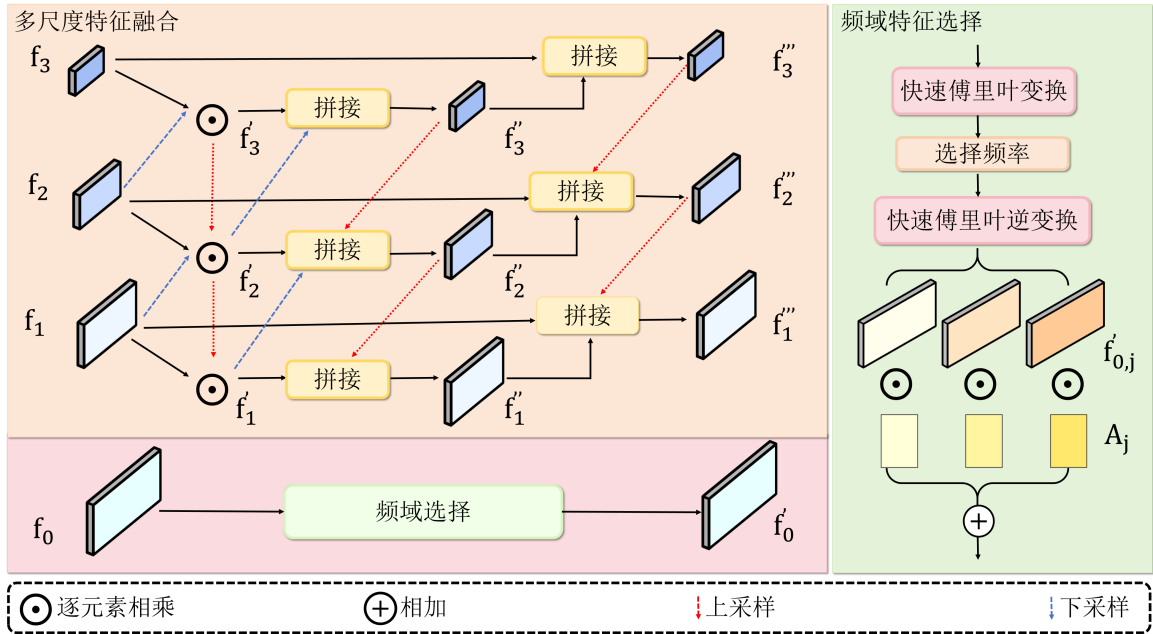


图 3.2 混合特征融合模块结构图

3.3.1 频域特征选择

模型最浅层的特征图 f_0 主要包含丰富的局部细节和边缘信息，这些信息通常集中在频域的高频分量中。若直接融合最大特征图，会无差别地聚合所有特征，可能导

致关键目标信息被背景噪声掩盖，同时增加计算开销。因此，本文采用频域选择特征融合方法，有针对性地提取高频细节，以提升特征表示质量并增强检测性能。具体而言，首先对最浅层特征图 $\mathbf{f}_0 \in \mathbb{R}^{C \times H \times W}$ （其中 C 表示通道数， H 和 W 表示空间高度和宽度）每个通道独立应用二维快速傅里叶变换转换到频域，生成频域表示 $\hat{\mathbf{f}}_0$ ，公式如下：

$$\hat{\mathbf{f}}_0[u, v] = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{f}_0[h, w] e^{-j2\pi(\frac{uh}{H} + \frac{vw}{W})}, \quad u = 0, \dots, H-1, v = 0, \dots, W-1 \quad (3.1)$$

其中， h, w 表示空间域中的像素位置索引， u, v 表示频域中的离散频率索引， j 为虚数单位（ $j^2 = -1$ ）。

为了提取不同频段的特征，对 $\hat{\mathbf{f}}_0$ 应用一组预定义的二进制掩码 $\{\mathbf{M}_j\}_{j=1}^n$ ，得到第 j 个频段的表示 $\hat{\mathbf{f}}_{0,j}$ ，公式如下：

$$\hat{\mathbf{f}}_{0,j} = \hat{\mathbf{f}}_0 \odot \mathbf{M}_j, \quad j = 1, \dots, n \quad (3.2)$$

其中， $\odot$ 表示逐元素乘法， n 为频段数量。

随后，通过逆快速傅里叶变换将每个掩码后的频率分量映射回空间域，得到捕捉不同频段互补信息的空间特征图 $\mathbf{f}_{0,j}$ ，公式如下：

$$\mathbf{f}_{0,j}[h, w] = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \hat{\mathbf{f}}_{0,j}[u, v] e^{j2\pi(\frac{uh}{H} + \frac{vw}{W})}, \quad h = 0, \dots, H-1, w = 0, \dots, W-1 \quad (3.3)$$

为自适应平衡这些频率分量的贡献，通过可学习的注意力图 $\mathbf{A}_j \in \mathbb{R}^{H \times W}$ 施加空间动态加权。通过聚合所有加权分量，得到最终的频率增强浅层特征 $\mathbf{F}_0$ ，公式如下：

$$\mathbf{F}_0 = \sum_{j=1}^n \mathbf{A}_j \odot \mathbf{f}_{0,j} \quad (3.4)$$

3.3.2 多尺度特征融合

在复杂农业场景中，杂草目标通常具有显著的尺度变化与结构差异。浅层特征图 $\mathbf{f}_1$ 具有较高的空间分辨率，能够保留丰富的边缘与纹理细节信息，但语义表达能力较弱；深层特征图 $\mathbf{f}_3$ 具有较强的语义表达能力，但空间分辨率较低，细节信息相对不足；中层特征图 $\mathbf{f}_2$ 在语义信息与空间细节之间起到过渡作用。因此，有效融合不同层级特征对于提升复杂场景下的检测性能具有重要意义。

针对上述问题, 本文提出一种跨尺度逐级融合策略。与现有 FPN、PANet 等方法采用逐元素相加或通道拼接的融合方式不同, 本文通过在不同特征层之间建立显式的逐元素乘积交互, 实现语义信息的逐级传递与细节信息的逐层增强, 从而提升整体特征表示能力。该过程以高层语义特征为引导, 在不同尺度之间逐步完成信息补充与特征增强。

在深层语义建模阶段, 引入中层特征 $\mathbf{f}_2$ 的结构信息对 $\mathbf{f}_3$ 进行补充, 从而增强其对中尺度空间结构的感知能力, 其计算过程如下:

$$\mathbf{f}'_3 = \mathbf{f}_3 \odot \text{Downsample}(\mathbf{f}_2) \quad (3.5)$$

其中, $\text{Downsample}(\cdot)$ 表示下采样操作, 用于将高分辨率特征映射至较低分辨率尺度, 从而引入更高层语义信息。

在中层特征建模过程中, $\mathbf{f}_2$ 同时接收来自深层语义信息与浅层细节信息的约束。其中, 深层特征提供语义引导, 浅层特征提供局部结构补充, 使中层特征成为跨尺度信息融合的中介表示:

$$\mathbf{f}'_2 = \mathbf{f}_2 \odot \text{Downsample}(\mathbf{f}_1) \odot \text{Upsample}(\mathbf{f}'_3) \quad (3.6)$$

其中, $\text{Upsample}(\cdot)$ 表示上采样操作, 用于将低分辨率特征恢复至较高分辨率尺度, 以补充空间细节信息。

在浅层特征建模阶段, 通过引入中层融合后的语义增强信息, 对 $\mathbf{f}_1$ 进行补充, 使其在保留高分辨率细节信息的同时获得更强的语义表达能力:

$$\mathbf{f}'_1 = \mathbf{f}_1 \odot \text{Upsample}(\mathbf{f}'_2) \quad (3.7)$$

在完成逐层特征交互后, 为进一步增强不同尺度特征之间的信息耦合能力, 本文采用多尺度特征聚合与跳跃连接策略, 对不同层级的特征进行统一整合, 从而同时保留语义信息与空间细节信息。具体而言, 在交互融合特征 $\mathbf{f}'_1$ 、 $\mathbf{f}'_2$ 和 $\mathbf{f}'_3$ 的基础上, 通过通道维度拼接操作对相邻尺度特征进行聚合, 以实现跨尺度信息的充分交互。

首先, 对深层特征 $\mathbf{f}_3$ 进行多尺度信息聚合, 将其与中层特征的下采样结果以及高层补充信息进行融合, 得到增强后的特征表示:

$$\mathbf{f}''_3 = \text{Concat}(\mathbf{f}'_3, \text{Downsample}(\mathbf{f}'_2)) \quad (3.8)$$

其中， $\text{Concat}(\cdot)$ 表示沿通道维度的拼接操作。

在此基础上，对中层特征 $\mathbf{f}_2$ 进行多尺度聚合操作，同时融合来自浅层与深层的补充信息，以增强其跨尺度表达能力：

$$\mathbf{f}_2' = \text{Concat}(\mathbf{f}_2, \text{Downsample}(\mathbf{f}_1'), \text{Upsample}(\mathbf{f}_3')) \quad (3.9)$$

对于浅层特征 $\mathbf{f}_1$ ，将其与来自中层的增强特征进行融合，从而在保留高分辨率细节信息的同时增强语义表达能力：

$$\mathbf{f}_1'' = \text{Concat}(\mathbf{f}_1', \text{Upsample}(\mathbf{f}_2')) \quad (3.10)$$

在完成多尺度特征聚合后，为进一步融合不同阶段的特征信息，引入跳跃连接机制，将原始特征与聚合后的特征进行拼接，得到最终融合特征表示：

$$\mathbf{f}_i''' = \text{Concat}(\mathbf{f}_i, \mathbf{f}_i''), \quad i \in \{1, 2, 3\} \quad (3.11)$$

其中， $i \in \{1, 2, 3\}$ 表示不同尺度层级的特征索引。

3.4 双域注意力机制模块

在复杂环境下，杂草常呈现密集重叠分布，且背景中存在土壤和作物等多种干扰。现有空域注意力机制如可变形注意力虽能适应杂草形态的变化，但在复杂背景中难以充分捕捉全局结构信息；而现有频域注意力方法虽能建模全局依赖，但局部几何细节刻画能力不足，且通常独立于空域注意力使用。针对这一问题，本文提出双域注意力机制，通过融合频域注意力与可变形注意力增强编码器对杂草的特征表示能力，如图3.3所示。该机制在频域提取叶片纹理、边缘和局部细节信息，并结合可变形注意力建模杂草与作物之间的空间关系，有效提升了杂草在低对比度、遮挡和高相似度条件下的检测精度。

3.4.1 频域全局注意力

频域注意力机制通过在频域空间中建模全局依赖关系，从频率角度刻画特征之间的关联性，能够有效捕捉图像中的周期性结构与多尺度变化模式。相比仅在空间域进行局部建模的方法，频域表示能够显式区分不同频率成分，其中低频信息主要

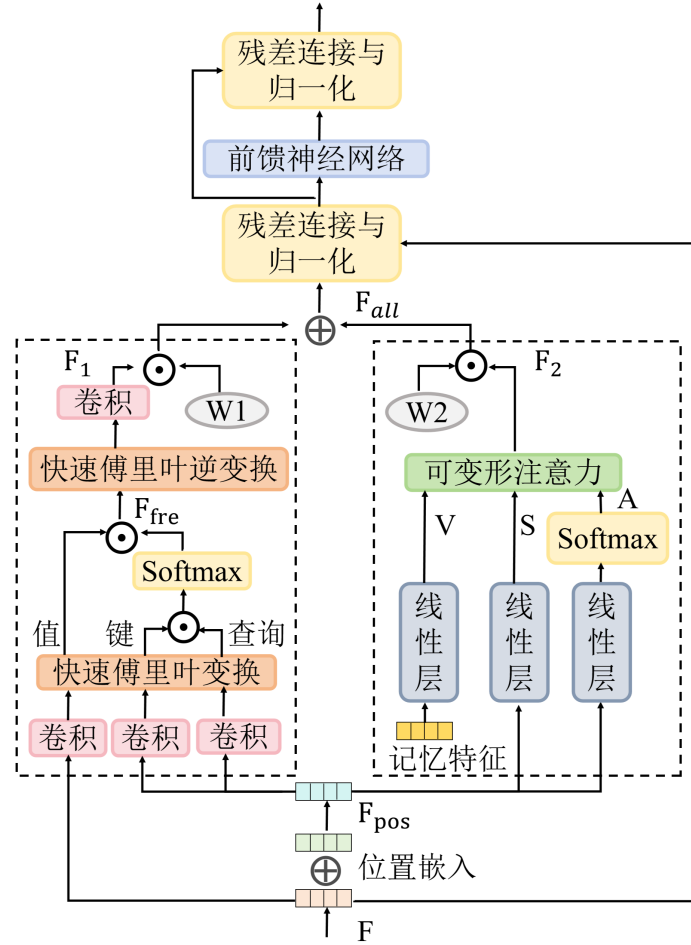


图 3.3 双域注意力机制模块结构图

对应图像的整体结构，高频信息则反映细节纹理与边缘变化。因此，该机制对于杂草检测任务中叶片细粒度纹理、复杂背景干扰以及小目标的表达具有重要作用。

设输入特征图为 $\mathbf{F} \in \mathbb{R}^{C \times L}$ ，其中 C 表示特征通道数， L 表示展平后的空间维度序列长度。为引入位置信息，将加入位置编码后的特征表示为 $\mathbf{F}_{\text{pos}}$ 。在特征建模过程中，首先对输入特征施加 1×1 卷积操作，以增强通道间的表达能力。随后，对特征沿空间序列维度 L 执行快速傅里叶变换（FFT），将其从空间域映射至频域空间。FFT 的离散形式中，第 k 个频率分量定义为：

$$\text{FFT}(\mathbf{X})_k = \sum_{n=0}^{L-1} \mathbf{X}_n e^{-j2\pi \frac{kn}{L}}, \quad k = 0, \dots, L-1 \quad (3.12)$$

其中， n 表示空间域中的离散位置索引， k 表示频率分量索引， j 为虚数单位（ $j^2 = -1$ ）。

在频域表示下，我们构建查询（ $\mathbf{Q}$ ）、键（ $\mathbf{K}$ ）与值（ $\mathbf{V}$ ）特征。具体而言， $\mathbf{Q}$ 与 $\mathbf{K}$ 基于含位置编码的特征 $\mathbf{F}_{\text{pos}}$ 生成，而 $\mathbf{V}$ 基于不含位置编码的原始特征 $\mathbf{F}$ 生成，计

算方式如下：

$$\mathbf{Q} = \text{FFT}(\text{Conv}(\mathbf{F}_{\text{pos}})) \quad (3.13)$$

$$\mathbf{K} = \text{FFT}(\text{Conv}(\mathbf{F}_{\text{pos}})) \quad (3.14)$$

$$\mathbf{V} = \text{FFT}(\text{Conv}(\mathbf{F})) \quad (3.15)$$

在获得频域表示后，通过计算 $\mathbf{Q}$ 与 $\mathbf{K}$ 之间的相关性来建模全局依赖关系。该过程通过 Softmax 函数进行归一化处理，从而得到频域注意力权重，并用于对 $\mathbf{V}$ 特征进行加权融合：

$$\mathbf{F}_{\text{fre}} = \text{Softmax} \left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d}} + \mathbf{B} \right) \cdot \mathbf{V} \quad (3.16)$$

其中， $\mathbf{B}$ 表示可学习的相对位置偏置， d 表示特征维度。

最后，将频域增强后的特征映射回空间域。该过程通过逆快速傅里叶变换实现频域到空间域的转换，并结合卷积操作进行进一步特征整合，从而得到最终输出特征 $\mathbf{F}_1$ ：

$$\mathbf{F}_1 = \text{Conv}(\text{IFFT}(\mathbf{F}_{\text{fre}})), \quad \mathbf{F}_{1,k} = \text{Conv} \left(\frac{1}{L} \sum_{n=0}^{L-1} \mathbf{F}_{\text{fre},n} e^{j2\pi \frac{kn}{L}} \right), \quad k = 0, \dots, L-1 \quad (3.17)$$

其中， $\text{IFFT}(\cdot)$ 表示逆快速傅里叶变换操作，用于将频域特征恢复至空间域表示， $\mathbf{F}_{1,k}$ 表示输出特征 $\mathbf{F}_1$ 在第 k 个空间位置上的响应值， L 表示特征序列的长度，即将空间特征图展平后的离散采样点总数。

3.4.2 空间域可变形注意力

为弥补频域建模在局部几何结构刻画方面的不足，本文在空间域引入可变形注意力机制，以实现对局部区域的自适应建模。在复杂田间场景中，杂草与作物目标通常具有形态不规则、尺度变化显著及空间分布稀疏等特点，传统基于固定采样位置的注意力机制难以适应此类几何变化。通过引入可学习的采样偏移，该机制能够动态调整关注区域，从而增强局部细节特征的表达能力。

具体而言，给定含位置编码的特征图 $\mathbf{F}_{\text{pos}}$ ，首先通过线性映射生成采样偏移量 $\mathbf{S}$ 与注意力权重 $\mathbf{A}$ ，用于刻画空间采样位置及其重要性分布。同时，对编码器输出的记忆特征 **Memory** 进行线性变换，得到基础特征表示 $\mathbf{M}$ ，用于后续特征聚合。上述过程可表示为：

$$\mathbf{S} = \text{Linear}(\mathbf{F}_{\text{pos}}) \quad (3.18)$$

$$\mathbf{A} = \text{SoftMax}(\text{Linear}(\mathbf{F}_{\text{pos}})) \quad (3.19)$$

$$\mathbf{M} = \text{Linear}(\mathbf{Memory}) \quad (3.20)$$

其中， $\text{Linear}(\cdot)$ 表示线性映射层，用于特征投影与参数生成， $\text{SoftMax}(\cdot)$ 用于对注意力权重进行归一化处理。

在获得采样偏移与注意力权重后，采用可变形注意力机制对空间特征进行自适应聚合。该过程首先根据采样偏移 $\mathbf{S}$ 在记忆特征 $\mathbf{M}$ 上进行双线性插值，实现对非规则位置特征的连续采样，从而突破固定网格采样的限制。在此基础上，通过注意力权重对各采样位置的特征响应进行自适应加权，实现不同空间位置贡献的动态调制。最终，将所有采样结果进行加权融合，得到空间域增强特征 $\mathbf{F}_2$ ，其计算过程如下：

$$\mathbf{F}_2 = \text{DeformAttention}(\mathbf{S}, \mathbf{A}, \mathbf{M}) = \sum_{n=1}^N \mathbf{W}_n \left[\sum_{k=1}^K A_{nk} \cdot \mathbf{W}'_n \text{BilinearInterpolation}(\mathbf{M}, \mathbf{S}) \right] \quad (3.21)$$

其中， $\text{BilinearInterpolation}(\cdot)$ 表示双线性插值操作，用于在非整数采样位置获取连续特征值； A_{nk} 表示第 n 个注意力头在第 k 个采样点的注意力权重； $\mathbf{W}'_n$ 用于对采样得到的特征进行线性变换； $\mathbf{W}_n$ 用于对各注意力头的输出进行加权融合； K 表示每个注意力头的采样点数量； N 表示注意力头个数。

3.4.3 双域特征自适应融合

为了充分发挥频域特征与空间域特征在复杂农业场景中的互补优势，本文在特征融合阶段设计了一种自适应加权策略。频域特征侧重于建模图像的全局结构信息与频率分布特征，在刻画整体轮廓与周期性纹理方面具有优势；空间域特征则更关注局部几何结构与细粒度纹理信息，在目标边缘与局部形态表达方面更为有效。因此，两类特征在表达层面具有明显互补性，如何实现有效融合成为提升模型性能的关键问题。

针对该问题，本文引入可学习权重对双域特征进行自适应融合，通过数据驱动方式动态调整不同特征分支的贡献，从而实现对不同场景特征分布的自适应建模。具体而言，将频域注意力输出 $\mathbf{F}_1$ 与空间域注意力输出 $\mathbf{F}_2$ 进行加权融合，得到最终特

征表示：

$$\mathbf{F}_{\text{all}} = w_1 \cdot \mathbf{F}_1 + w_2 \cdot \mathbf{F}_2 \quad (3.22)$$

其中， w_1 和 w_2 为可学习融合权重，在训练过程中通过反向传播进行端到端优化，使模型能够自适应调整频域与空间域特征的贡献比例。

与固定融合策略相比，该方法无需人工设定融合系数，能够根据输入特征的分布自动调节不同模态信息的权重，从而提升模型在复杂农业场景下的适应性与表达稳定性。

3.5 基于高斯分布与约束引导的标签分配方法

尽管双域注意力机制能够增强复杂场景下的特征表征能力，但模型性能的提升不仅依赖于特征表达，还受到训练阶段标签分配策略的关键影响，其直接决定预测查询与真实目标之间的匹配关系，并影响分类与定位分支的优化效果。现有标签分配方法多依赖 IoU 阈值、几何约束或质量度量对样本进行划分，并结合动态分配或质量感知机制以提升复杂场景下的适应性与训练稳定性^[66-74]。然而，在 DETR 系列方法中，标签分配通过匈牙利算法基于固定 IoU 或几何距离构建匹配代价，在杂草遮挡严重、尺度差异大或目标密集分布的情况下，易导致正负样本分配不稳定，从而影响模型收敛与检测精度。为此，本章节对匈牙利算法中的匹配代价函数进行针对性改进，提出一种基于高斯分布与约束引导的标签分配方法，通过引入中心约束、位置约束以及分布相似性度量，实现更稳健且符合杂草目标空间结构的匹配策略。

3.5.1 中心与位置约束

为进一步提升预测框与真实目标之间的匹配精度，本文在匹配代价设计中引入中心约束与位置约束机制，从中心偏移与整体重叠关系两个层面对预测框进行联合约束，以增强模型对目标几何位置的敏感性。

在中心约束方面，通过衡量预测边界框与真实边界框在水平与垂直方向中心坐标的偏移程度，从而构建中心约束代价。具体而言，分别计算 x 与 y 方向上的中心距离，并根据其与真实框宽高的比例关系设置约束阈值，其表达如下：

$$\text{Cost}_{\text{cenx}} = \begin{cases} 1 & \text{L1}(\text{Cen}_x^p, \text{Cen}_x^{gt}) \geq \alpha W_{gt} \\ 0 & \text{L1}(\text{Cen}_x^p, \text{Cen}_x^{gt}) < \alpha W_{gt} \end{cases} \quad (3.23)$$

$$\text{Cost}_{\text{ceny}} = \begin{cases} 1 & \text{L1}(\text{Cen}_y^p, \text{Cen}_y^{gt}) \geq \alpha H_{gt} \\ 0 & \text{L1}(\text{Cen}_y^p, \text{Cen}_y^{gt}) < \alpha H_{gt} \end{cases} \quad (3.24)$$

其中, $\text{Cost}_{\text{cenx}}$ 与 $\text{Cost}_{\text{ceny}}$ 分别表示预测边界框与真实边界框在 x 、 y 方向上的中心约束代价; Cen_x^p , Cen_y^p 表示预测边界框的中心坐标, Cen_x^{gt} , Cen_y^{gt} 表示真实边界框的中心坐标; $\text{L1}(\cdot)$ 表示 L_1 距离; W_{gt} 与 H_{gt} 分别表示真实边界框的宽度与高度; α 为取值范围在 $(0, 1)$ 内的可学习参数, 用于控制中心偏移约束的敏感程度。

在位置约束方面, 通过引入预测框与真实框之间的 IoU 关系对整体空间重叠程度进行建模。当预测框与真实框的重叠程度低于设定阈值时, 认为其匹配质量较差, 从而施加惩罚约束, 其计算方式如下:

$$\text{Cost}_{\text{pos}} = \begin{cases} 1 & \text{IoU}(B^p, B^{gt}) < \beta \\ 0 & \text{IoU}(B^p, B^{gt}) \geq \beta \end{cases} \quad (3.25)$$

其中, Cost_{pos} 表示预测框与真实框的位置约束代价; B^p 表示预测边界框, B^{gt} 表示真实边界框; $\text{IoU}(\cdot)$ 表示交并比; β 为取值范围在 $(0, 1)$ 内的可学习参数。

最终, 将中心约束与位置约束的代价进行累加, 构建整体约束引导项:

$$\text{Cost}_{\text{constraint}} = \text{Cost}_{\text{cenx}} + \text{Cost}_{\text{ceny}} + \text{Cost}_{\text{pos}} \quad (3.26)$$

3.5.2 高斯分布相似性

为更准确刻画目标区域的空间分布特性, 本文将边界框从传统的几何表示扩展为概率分布建模形式。考虑到真实目标的前景像素通常集中于目标中心区域, 而边缘区域逐渐衰减, 这种空间不均匀性难以通过矩形框的硬边界进行有效表达。因此, 本文采用二维高斯分布对目标进行建模, 使其能够以连续概率形式描述目标的空间响应强度分布, 从而更细致地刻画目标的结构信息。

对于两个检测框 $\mathbf{b}_a = (c_{x_a}, c_{y_a}, w_a, h_a)$ 和 $\mathbf{b}_b = (c_{x_b}, c_{y_b}, w_b, h_b)$, 其中 (c_x, c_y) 表示边界框中心坐标, w 和 h 分别表示宽度与高度, 分别构建其对应的高斯分布 $\mathcal{N}_a$ 和 $\mathcal{N}_b$ 。相较于传统 IoU 仅关注区域重叠关系, 该表示方式能够同时编码空间位置与尺度结构信息, 从而为后续匹配提供更丰富的几何约束。

为了衡量两个高斯分布之间的差异，本文采用二阶 Wasserstein 距离进行度量。该距离能够同时考虑分布均值之间的偏移以及结构尺度差异，从而更全面地反映预测框与真实框之间的几何一致性，其定义如下：

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2 \quad (3.27)$$

其中， $\boldsymbol{\mu}_a = [c_{x_a}, c_{y_a}, \frac{w_a}{2}, \frac{h_a}{2}]^T$ ， $\boldsymbol{\mu}_b = [c_{x_b}, c_{y_b}, \frac{w_b}{2}, \frac{h_b}{2}]^T$ 。

与 IoU 等基于重叠面积的度量方式不同，该距离在目标不重叠或小目标场景下仍然具有连续梯度变化特性，因此在复杂农业场景中具有更好的稳定性与鲁棒性。

在此基础上，为将距离度量转化为统一尺度下的相似性表示，本文引入归一化高斯 Wasserstein 距离，通过指数映射将距离压缩至 $(0, 1)$ 范围内，从而用于匹配代价构建，其表达如下：

$$\text{Cost}_{\text{NWD}} = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C}\right) \quad (3.28)$$

其中， $\exp(\cdot)$ 表示以自然常数 e 为底的指数函数， C 为尺度归一化系数，用于控制距离到相似性映射的衰减程度。

3.5.3 代价函数与标签分配

为实现预测框与真实目标之间的精确匹配，本文构建由分类代价、约束引导代价以及高斯分布相似性代价组成的联合匹配代价函数。与仅依赖 IoU 或单一几何度量的匹配方式不同，该方法同时引入类别信息、中心几何约束以及分布相似性约束，从而从语义层面与几何层面对匹配过程进行联合建模。

为了避免各类代价在数值尺度与作用强度上的不均衡问题，本文进一步引入可学习权重参数 A_1 、 A_2 和 A_3 ，通过数据驱动方式自适应调整不同代价分量的贡献比例，使模型能够根据训练阶段的优化需求动态平衡分类信息与几何结构信息之间的影响。具体而言，联合代价函数定义如下：

$$\text{Cost} = A_1 \cdot \text{Cost}_{\text{class}} + A_2 \cdot \text{Cost}_{\text{constraint}} + A_3 \cdot \text{Cost}_{\text{NWD}} \quad (3.29)$$

其中， $\text{Cost}_{\text{class}}$ 采用 Focal Loss 对预测类别与真实类别之间的差异进行度量，以缓解正负样本不均衡问题。

在标签分配阶段，基于新的联合代价函数，本文采用匈牙利算法在预测集合与真实标注之间进行全局最优一对一匹配，通过最小化整体匹配代价获得最优分配结果，从而确保每个真实目标都能找到唯一对应的预测样本，保证训练过程的监督质量。

3.6 实验结果与分析

3.6.1 实验数据及实验设置

（1）实验环境

本章节提出的各项算法均基于 Python 编程语言实现，并依托于 PyTorch 深度学习框架构建。为确保实验环境的一致性，所有训练与测试过程均在相同版本镜像配置的 Docker 容器中完成。软件环境包括 PyTorch 1.13 和 CUDA 11.6。硬件平台为配备 Intel Xeon E5-2620（4 核，主频 2.10GHz）处理器的工作站，内存 32GB，并搭载 NVIDIA RTX 3090（24GB）显卡，用于满足大规模深度学习计算需求。

（2）实验数据

本章节实验数据使用 WeedCrop、Lincolnbeet 以及 MH-Weed16 数据集，这些数据集均包含复杂环境下的多目标图像，能够真实反映农业场景中杂草密集分布、遮挡频繁以及植株形态多样的情况。

（3）训练参数

模型训练在单块 NVIDIA RTX 3090 GPU 上进行，批量大小设置为 2。优化器选用 AdamW^[75]，Transformer 模块初始学习率为 1×10^{-4} ，ResNet-50 骨干网络初始学习率为 1×10^{-5} ，权重衰减设置为 1×10^{-4} ，训练总轮数为 36 个 epoch。为加速模型收敛，本研究在训练前加载了 ResNet-50 骨干网络的最新预训练权重，并在训练过程中对其进行更新。

（4）网络模型参数

FS-DETR 模型包含 6 层 Transformer 编码器和 6 层 Transformer 解码器，隐藏特征维度为 256。骨干网络 ResNet-50 在训练前已进行 ImageNet 预训练，以增强特征提取能力。编码器与解码器的各层结构均遵循标准多头自注意力机制设计，用于在复杂场景中捕获目标间的全局上下文关系与细粒度特征。

3.6.2 实验评估和对比分析

为验证所提出的 FS-DETR 方法在复杂场景杂草目标检测中的有效性与优越性，本章节将其与当前领域内的多种先进检测算法进行了多维度对比。实验不仅评估了 FS-DETR 在整体检测性能上的提升，还特别分析了其在不同目标尺度下的表现，以全面展示频域-空间域协同建模策略的优势。具体而言，本章节选取了以 YOLO 系列为代表的轻量化实时检测方法、以 DETR 系列为代表的高精度 Transformer 检测器，以及采用 InternImage 等特殊骨干网络的代表性模型，从而保证对比的全面性与公平性。

表 3.1 WeedCrop 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	44.9	75.0	48.9	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	45.1	77.1	46.7	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	<u>46.3</u>	78.8	47.7	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	45.4	77.3	47.4	42.9	48.5	68.7
RT-DETRv2 ^[80]	2024CVPR	ResNet50	45.9	77.4	<u>50.0</u>	44.5	<u>50.3</u>	57.1
InternImage ^[81]	2023CVPR	InternImage	42.0	74.6	43.8	45.0	37.1	69.7
DETR ^[19]	2020ECCV	ResNet50	34.5	60.2	36.3	32.4	32.8	44.6
Deformable-DETR ^[20]	2021ICLR	ResNet50	41.2	77.4	40.6	41.3	41.9	57.7
Co-DETR ^[35]	2023ICCV	ResNet50	42.9	72.4	48.5	45.3	41.2	64.1
DINO ^[30]	2023ICLR	ResNet50	45.8	78.2	48.5	43.9	49.9	59.5
DDQ ^[24]	2023CVPR	ResNet50	45.6	80.1	46.4	<u>46.0</u>	47.8	66.4
Align-DETR ^[82]	2024BMVC	ResNet50	41.3	69.2	44.6	40.8	40.7	67.8
H-DETR ^[33]	2023CVPR	ResNet50	36.7	64.4	39.2	37.1	37.5	48.4
DEIM ^[83]	2025CVPR	ResNet50	45.7	75.4	49.4	44.9	47.5	63.2
FS-DETR	—	ResNet50	47.2	<u>79.3</u>	53.0	47.2	51.1	<u>68.8</u>

表中加粗与下划线分别表示最优与次优结果。如表3.1、表3.2和表3.3所示，FS-DETR 在 WeedCrop、Lincolnbeet 和 MH-Weed16 三个数据集上均取得了最高的平均精度 (AP)，在不同尺度目标上的检测性能均优于现有方法。其中，FS-DETR 在 WeedCrop 数据集的整体 AP、小目标 AP 及中等目标 AP 上均显示出明显优势。与 CNN 方法 InternImage 相比，整体 AP 提升 5.2%；与最新 YOLO 系列相比，FS-DETR 在 YOLOv9-C 和 YOLOv10-L 上分别取得了 2.3% 和 2.1% 的性能提升，同时较 YOLOv13-L 提升 0.9%，有效增强了小目标和中等目标的检测性能。相比 DETR 系列，FS-DETR 在

DINO 和 RT-DETRv2 上分别提升 1.4% 和 1.3%，进一步验证了其在复杂农业场景下的稳健性。

表 3.2 Lincolnbee 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	57.2	78.6	60.9	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	57.8	79.4	61.6	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	58.7	79.7	62.6	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	59.3	81.2	63.2	14.2	41.5	<u>72.8</u>
RT-DETRv2 ^[80]	2024CVPR	ResNet50	59.0	80.1	62.8	13.5	41.2	71.7
InternImage ^[81]	2023CVPR	InternImage	58.3	80.6	61.7	14.6	41.2	70.5
DETR ^[19]	2020ECCV	ResNet50	44.6	60.9	42.0	4.3	28.7	58.6
Deformable-DETR ^[20]	2021ICLR	ResNet50	50.4	76.0	53.9	4.9	30.9	62.8
Co-DETR ^[35]	2023ICCV	ResNet50	58.8	80.9	62.1	17.0	41.1	71.1
DINO ^[30]	2023ICLR	ResNet50	59.4	81.6	63.2	13.0	41.8	71.7
DDQ ^[24]	2023CVPR	ResNet50	<u>59.5</u>	<u>81.7</u>	<u>64.2</u>	15.1	<u>42.5</u>	72.3
Align-DETR ^[82]	2024BMVC	ResNet50	58.5	80.0	62.1	12.9	41.0	70.6
H-DETR ^[33]	2023CVPR	ResNet50	58.3	80.6	61.7	14.6	41.2	70.5
DEIM ^[83]	2025CVPR	ResNet50	58.5	80.2	62.3	14.4	40.8	70.6
FS-DETR	—	ResNet50	60.4	82.2	64.3	<u>16.5</u>	43.3	73.0

在 Lincolnbee 数据集上, FS-DETR 保持了优异的性能, 尤其在中、大目标检测方面表现突出。与 InternImage 相比, 整体 AP 提升 2.1%; 较 YOLOv9-C 和 YOLOv10-L 分别提升 3.2% 和 2.6%, 同时在 YOLOv13-L 上取得了 1.7% 的性能提升, 展示了其对密集目标分布的良好适应能力。此外, 相较于 DDQ 和 RT-DETRv2, FS-DETR 分别提升 0.9% 和 1.4%, 进一步验证了其在复杂环境下对多样化杂草目标的稳健检测能力。

在 MH-Weed16 数据集上, FS-DETR 在中、大目标上表现尤为突出。实验结果显示, FS-DETR 相比 InternImage 整体 AP 提升 19.1%, 体现出其在复杂农业场景下更强的特征表达能力。同时分别超过 YOLOv9-C、YOLOv10-L 和 YOLOv13-L, 提升 0.7%、0.6% 和 0.5%, 说明在轻量化一阶段检测器中仍具有稳定优势。此外, 相较于 DDQ 和 RT-DETRv2, 分别提升 1.6% 和 6.3%, 表明其在查询建模与目标匹配机制上的改进有效提升了密集目标场景下的检测稳定性。

实验结果表明, FS-DETR 在不同数据集上均实现了 AP 的持续提升, 验证了其

表 3.3 MH-Weed16 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	31.8	63.7	26.3	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	31.9	64.6	26.8	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	<u>32.0</u>	63.4	27.4	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	25.6	53.1	21.2	4.6	16.8	27.9
RT-DETRv2 ^[80]	2024CVPR	ResNet50	26.2	54.1	21.7	<u>4.7</u>	16.9	28.4
InternImage ^[81]	2023CVPR	InternImage	13.4	29.6	11.2	3.5	11.8	14.8
DETR ^[19]	2020ECCV	ResNet50	15.4	40.3	9.8	0.1	4.6	14.8
Deformable-DETR ^[20]	2021ICLR	ResNet50	19.7	46.4	14.0	0.5	9.3	20.9
Co-DETR ^[35]	2023ICCV	ResNet50	31.2	<u>66.1</u>	24.7	4.2	20.7	33.6
DINO ^[30]	2023ICLR	ResNet50	30.9	64.5	25.3	4.3	16.7	34.0
DDQ ^[24]	2023CVPR	ResNet50	30.9	63.8	27.2	4.6	16.6	33.9
Align-DETR ^[82]	2024BMVC	ResNet50	31.1	63.2	27.6	4.0	17.5	<u>34.1</u>
H-DETR ^[33]	2023CVPR	ResNet50	30.4	62.6	25.6	4.2	20.1	33.0
DEIM ^[83]	2025CVPR	ResNet50	30.6	60.1	27.4	4.4	19.3	33.5
FS-DETR	—	ResNet50	32.5	66.7	<u>27.5</u>	4.9	<u>20.3</u>	35.3

在复杂农业场景中的有效性，其性能提升主要得益于多模块的联合优化设计。HFF 在特征构建阶段融合空间域多尺度结构信息与频域高频纹理特征，使模型能够同时捕获目标的结构轮廓与细粒度纹理，从而增强对细粒度杂草目标的表达能力。在此基础上，DDAM 在编码阶段进一步联合建模空间结构依赖与频域判别信息，其中频域分支用于抑制背景噪声干扰，可变形注意力则增强对关键区域的自适应聚焦能力，从而提升模型在低对比度、高相似度及密集分布场景下的目标定位与表征能力。此外，GCLA 通过引入高斯分布与约束引导机制对预测框与真实标注进行匹配建模，使正负样本分配更合理，从而减少误检并提升检测稳定性。

3.6.3 消融实验

本章节旨在验证 FS-DETR 中各个模块在复杂场景下的有效性及其协同作用，所有实验均在 Lincolnbeet 数据集上进行逐步集成实验。实验均基于 DINO^[30] 的 ResNet-50 骨干网络 and 自注意力机制，使用与 3.6.2 节相同的 6 层 Transformer 编码器和 6 层解码器配置及训练设置。基线模型在 Lincolnbeet 数据集上的 AP 为 59.0%，作为后续各项实验的参考标准。实验结果如表3.4所示。

(1) 模块的有效性

表 3.4 模块有效性的消融实验

HFF	DDAM	GCLA	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
			59.0	81.6	62.2	12.7	41.2	71.5
✓			60.1	81.8	64.2	16.0	43.2	72.7
	✓		60.2	82.0	64.3	16.7	42.0	72.9
		✓	60.0	81.8	64.1	15.3	43.0	72.5
✓	✓		60.2	81.9	64.5	15.1	42.8	72.8
✓		✓	60.2	82.0	64.3	15.6	42.9	72.7
	✓	✓	60.2	82.2	64.3	14.7	42.8	72.9
✓	✓	✓	60.4	82.2	64.3	16.5	43.3	73.0

单模块实验显示，无论是 HFF、DDAM 还是 GCLA，均能在复杂背景下显著改善模型性能。HFF 模块通过引入频域选择分支并结合多尺度空间特征，有效增强了对杂草边缘轮廓、叶片纹理及细微结构的捕捉能力，从而在小目标检测上 AP 从 12.7% 提升至 16.0%。DDAM 模块通过自适应融合频域与可变形空间注意力，使模型在密集或高相似目标场景下更容易区分杂草与作物，其小目标 AP 达到 16.7%，在保持整体 AP 提升的同时，改善了低对比度及遮挡情况下的检测效果。GCLA 模块引入中心、位置约束和归一化高斯沃瑟斯坦距离作为样本匹配代价，使标签分配更贴合杂草在田间的真实空间分布，从而在训练稳定性和正负样本平衡上产生明显改善。

当两模块联合使用时，例如 HFF 与 DDAM、HFF 与 GCLA、DDAM 与 GCLA，AP 均提升至 60.2%，这一增幅不仅体现了各模块的互补性，也反映了它们在应对复杂场景中多尺度目标、密集分布以及背景干扰时的协同作用。最显著的提升出现在三个模块同时启用的情况下，整体 AP 达到 60.4%，小目标 AP 上升至 16.5%，中型和大型目标 AP 分别为 43.3% 和 73.0%。这一结果充分体现了 HFF、DDAM 与 GCLA 的联合优化能力：HFF 提供了丰富的细节信息以增强定位精度，DDAM 利用频域与空间域特征的互补性强化判别能力，而 GCLA 则保证训练阶段样本分配的合理性和稳定性。因此，FS-DETR 在复杂田间环境下的优异性能正是基于三个创新模块在捕捉细节特征、整合全局上下文以及优化标签分配方面的协同作用，使模型在处理小型、密集及视觉相似杂草目标时表现出更高的鲁棒性与精度。

（2）混合特征融合模块的作用

为了验证 HFF 的有效性，本章节对比了不同特征融合策略在复杂场景中的表现，如图3.4所示。传统方法如 FPN^[44]通过自顶向下路径整合多尺度特征，可以在

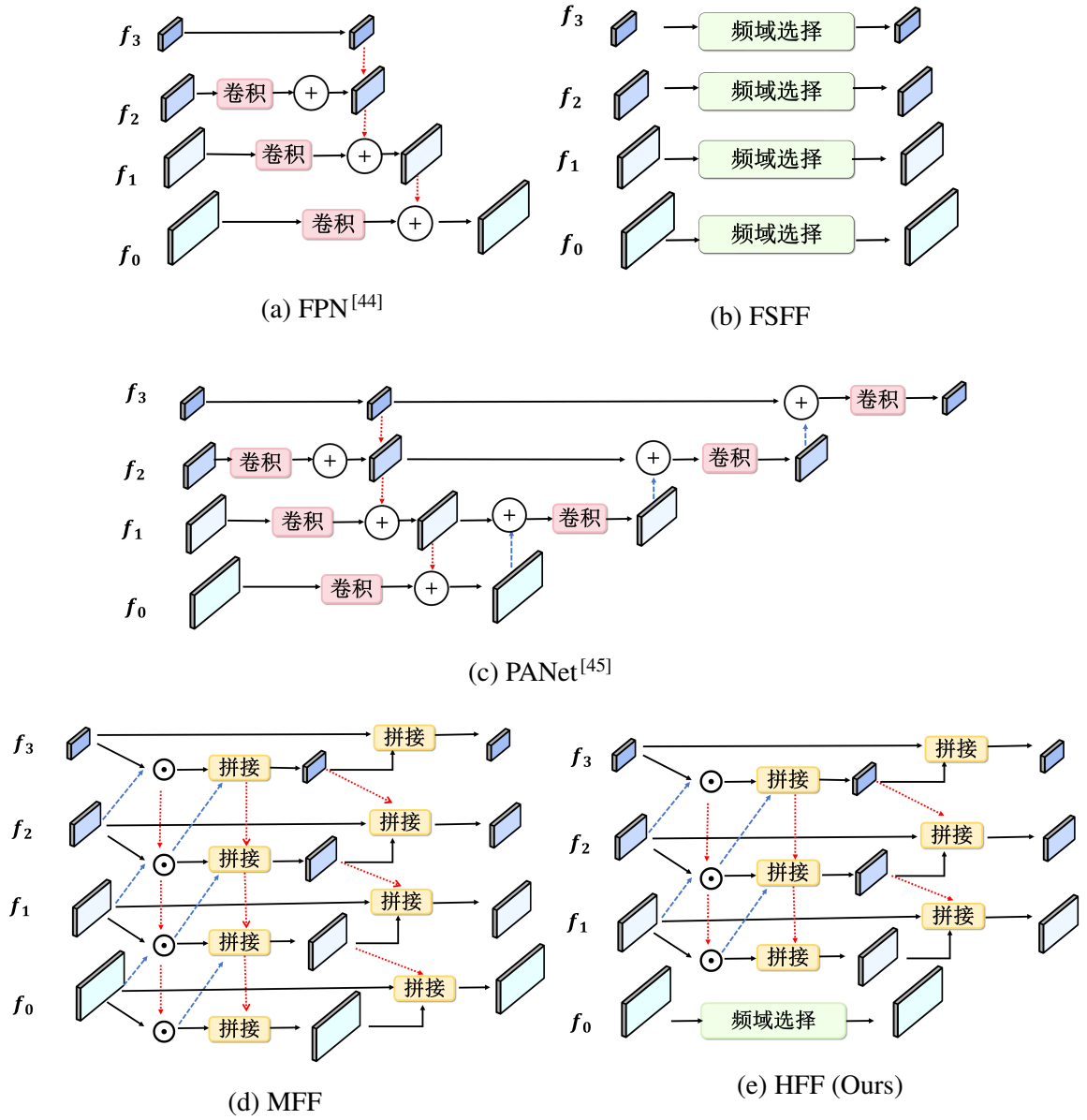


图 3.4 不同特征融合方法对比图

一定程度上增强模型对大中型目标的感知能力，但在小目标检测中仍存在明显短板；PANet^[45]则引入双向路径优化特征传播，使浅层语义能够向高层传递，但其在处理背景复杂、目标密集或遮挡严重的场景时仍容易受到干扰。在此基础上，我们设计了三种融合策略以探索最优特征表示方式：多尺度特征融合（MFF）、频域选择特征融合（FSFF）以及同时结合频域选择与空间多尺度融合的 HFF。

MFF 通过逐层特征整合增强了模型的多尺度感知能力，但在杂草小目标或叶片纹理细节的捕捉上仍显不足，导致在高遮挡或低对比度区域出现部分漏检。FSFF 强调从频域提取高频信息，有效提升了细粒度特征的判别力，但当背景中存在土壤纹

理、枯叶杂乱或作物叶片遮挡时，单一频域信息仍难以完全抑制噪声，AP 的提升较为有限。相比之下，HFF 模块融合频域选择与多尺度空间信息，通过高频分量增强小目标与边缘表征，并借助跨层特征交互提升多尺度建模能力，从而在复杂背景下兼顾细粒度表达与空间鲁棒性。

实验结果如表3.5所示，HFF 在整体 AP 上达到 60.2%，优于 FPN、PANet、MFF 和 FSFF 等方法。在小目标检测场景下，其 AP 由 12.7% 提升至 16.7%，表明该融合策略在细粒度目标表征方面具有更强能力。在实际复杂场景中，杂草常常呈现小型、密集分布，并且容易被作物叶片部分遮挡或与背景土壤和植被纹理融为一体。HFF 模块通过空间多尺度特征融合与频域高频信息增强的协同建模，有效强化细粒度边缘与纹理信息，使模型在低对比度和遮挡条件下仍能准确区分杂草与背景，从而减少漏检和误检。

表 3.5 混合特征融合模块的消融实验

特征融合策略	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
基线模型	59.0	81.6	62.2	12.7	41.2	71.5
+ 特征金字塔模块 (FPN) ^[44]	59.9	81.6	63.7	14.6	41.8	72.4
+ 路径聚合模块 (PANet) ^[45]	59.7	81.5	63.6	15.4	42.2	72.2
+ 多尺度特征融合模块 (MFF)	59.8	81.8	63.7	15.2	41.9	72.2
+ 频域选择特征融合模块 (FSFF)	59.9	81.6	63.6	16.0	42.4	72.7
+ 混合特征融合 (HFF)	60.2	82.0	64.3	16.7	42.0	72.9

(3) 双域注意力机制的作用

表3.6展示了 DDAM 在复杂场景下的检测效果。在仅采用自注意力机制时，模型的 AP 为 59.0%。当将传统自注意力替换为可变形注意力时，AP 小幅提升至 59.4%，这说明可变形注意力在捕捉杂草形态变化和局部结构方面具有一定优势。然而，单纯依靠空间域注意力仍难以充分刻画细微纹理和边缘信息。当仅使用频域注意力，AP 提升至 59.9%，表明频域高频分量对小型杂草边缘轮廓和幼苗叶片纹理的识别至关重要，尤其在背景复杂、土壤和作物叶片纹理相似的区域，频域信息能够有效增强微小目标的可区分性。

通过将空间可变形注意力与频域注意力进行有机融合，双域注意力机制模块将 AP 提升至 60.2%，小目标检测精度达到 16.7%。这一结果表明，双域信息的协同建模显著增强了模型在复杂环境下对细粒度目标的识别能力。该融合机制不仅充分利

表 3.6 双域注意力机制的消融实验

注意力机制	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
自注意力机制	59.0	81.6	62.2	12.7	41.2	71.5
可变形注意力机制 ^[20]	59.4	81.6	63.2	13	41.8	71.7
频域注意力机制	59.9	81.7	63.6	15.3	42.3	72.3
双域注意力机制	60.2	82.0	64.3	16.7	42.0	72.9

用了空间域的局部几何信息和频域的纹理细节，还提升了模型对多尺度及遮挡目标的适应性，从而进一步提高了细粒度目标的检测精度。

此外，表3.7对双域注意力机制中可学习融合权重 w_1 和 w_2 的作用进行了分析。固定权重设置虽然能够在一定程度上整合空间和频域特征，但不同组合对小目标检测效果存在明显差异。例如，当 $w_1 = 0.7$ ， $w_2 = 0.3$ 时，小目标 AP 为 16.0%，而固定均衡权重 $w_1 = w_2 = 0.5$ 时仅为 13.2%。相比之下，允许模型在训练过程中自适应学习 w_1 和 w_2 后，小目标 AP 提升至 16.7%，整体 AP 达到 60.2%。这表明模型能够根据输入内容动态调整空间与频域特征的贡献，从而在微小、遮挡及背景相似目标的检测中表现出更高的准确性与稳定性，体现了自适应融合机制在特征平衡方面的有效性。

表 3.7 双域注意力机制中融合权重的消融实验

w_1	w_2	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
0.5	0.5	59.5	81.8	63.0	13.2	41.7	71.8
0.3	0.7	59.7	81.9	63.5	14.8	41.9	72.2
0.7	0.3	59.9	81.8	64.0	16.0	41.7	72.5
可学习	可学习	60.2	82.0	64.3	16.7	42.0	72.9

(4) 基于高斯分布与约束引导的标签分配的作用

为了深入分析 GCLA 在复杂环境下的杂草检测效果，本文在统一的匈牙利匹配框架下，对不同标签分配策略进行了系统评估，结果如表3.8所示。传统 DETR 采用基于匈牙利匹配的标签分配策略，其代价函数由分类代价与边界框回归代价构成，其中回归部分结合 L1 距离与 IoU 进行度量。在该设置下，模型整体 AP 为 59.0%。然而，在作物与杂草外观高度相似的复杂场景中，这种基于几何重叠与位置偏差的度量方式仍存在一定局限，导致分类准确率和 small 目标识别能力不足。引入基于高斯分

布的标签分配方法后，AP 提升至 59.7%，这表明通过将目标特征建模为高斯分布并优化匹配，可以有效强化小尺度目标的识别，尤其是在杂草被遮挡或分布密集的情况下，其检测性能明显优于传统方法。相比之下，基于约束引导的标签分配方法通过引入中心位置、边界约束以及空间分布规则，实现了 59.8% 的 AP，进一步减少了误匹配现象，提高了整体检测精度。将高斯分布与约束引导相结合的 GCLA 方法，AP 提升至 60.0%，小目标检测精度达到 15.3%，显示了两种策略的互补性能够显著增强模型在复杂背景下对微小目标的判别能力和稳健性。

表 3.8 基于高斯分布与约束引导标签分配的消融实验

标签分配策略	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
基于原始代价函数的标签分配	59.0	81.6	62.2	12.7	41.2	71.5
基于高斯分布的标签分配	59.7	81.7	63.4	15.2	42.2	72.0
基于约束引导的标签分配	59.8	81.7	63.4	14.3	42.3	72.1
基于高斯分布与约束引导的标签分配	60.0	81.8	64.1	15.3	43.0	72.5

此外，为进一步验证 GCLA 对 DETR 类模型收敛性能的影响，本文绘制了基线模型 DINO 与 FS-DETR 的收敛曲线对比，如图3.5所示。相比 DINO，FS-DETR 在训练早期即可获得更高的检测精度，并表现出更快的收敛速度与更稳定的优化过程。这表明 GCLA 能够通过优化正负样本匹配关系，为模型提供更加稳定的监督信号，从而缓解 DETR 类模型训练初期匹配不稳定的问题，并提升最终检测性能。

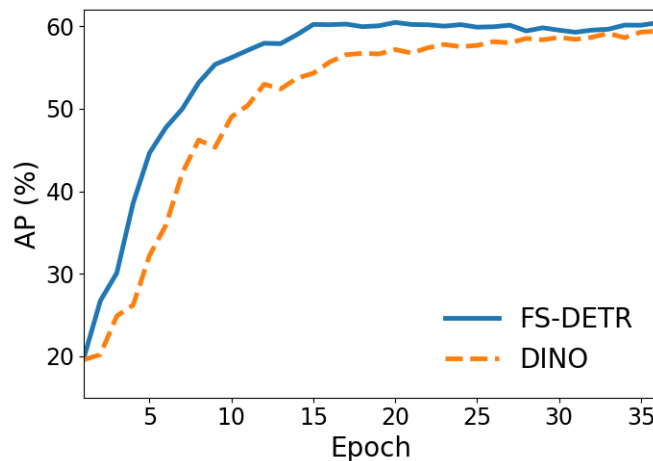


图 3.5 DINO 与 FS-DETR 的收敛曲线对比

为了进一步理解基于高斯分布与约束引导的标签分配中各成本组件的贡献，我

们对可学习权重 A_1 , A_2 , A_3 的作用进行了分析, 如表3.9所示。当将这些权重固定为均衡值或预设组合时, 各项 AP 指标均低于训练中自适应优化的结果, 表明不同任务和场景下, 单一固定比例难以兼顾分类准确性、位置约束以及分布相似性。然而, 允许模型在训练过程中自适应学习 A_1 , A_2 , A_3 后, 不仅整体 AP 达到 60.0%, 小目标检测也获得 15.3% 的提升, 这说明模型能够根据图像内容灵活平衡分类成本、约束成本与高斯分布相似性成本, 从而更准确地分配正负样本, 提高细粒度杂草的识别能力。

表 3.9 基于高斯分布与约束引导的标签分配中代价权重的消融实验

A_1	A_2	A_3	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
1/3	1/3	1/3	59.6	81.7	63.0	14.3	41.8	72.0
0.5	0.3	0.2	59.5	81.8	63.3	14.0	41.9	72.1
0.2	0.5	0.3	59.7	81.8	63.4	14.5	42.2	72.2
0.2	0.3	0.5	59.8	81.7	63.6	14.8	42.3	72.3
可学习	可学习	可学习	60.0	81.8	64.1	15.3	43.0	72.5

3.6.4 可视化结果分析

(1) 整体检测性能可视化

图3.6、图3.7和图3.8分别展示了 FS-DETR 在 WeedCrop、Lincolnbeet 和 MH-Weed16 杂草数据集上的检测可视化结果。为便于直观分析, 图中采用统一标注方式展示检测结果, 其中 GT 表示真实标注框, 白色边界框用于标识未被模型检测到的目标。在图3.6和图3.7中, 绿色边界框对应农作物目标, 红色边界框表示杂草; 而在图3.8中, 不同颜色的边界框用于区分不同类别的杂草目标。从图中可以直观地看出, FS-DETR 在各类场景下均能有效识别杂草目标。与其他主流方法相比, 本模型在处理小目标时表现出明显优势, 显著降低了漏检率, 这对于早期杂草防治至关重要。

(2) 复杂场景下的对比分析

为深入评估模型在挑战性环境下的鲁棒性, 图3.9将 FS-DETR 与强基线模型 DINO 在 Lincolnbeet 数据集上的检测结果进行了对比, 其中第一行为真实标注 (GT), 第二行为 DINO 的预测结果, 第三行为本文方法的检测结果。从图中可以看到, 所选场景涵盖了多种具有代表性的挑战情况, 例如目标密集分布以及低对比度目标。在这些情况下, 杂草往往紧密分布且与背景或作物在颜色与纹理上较为接近, 同时部

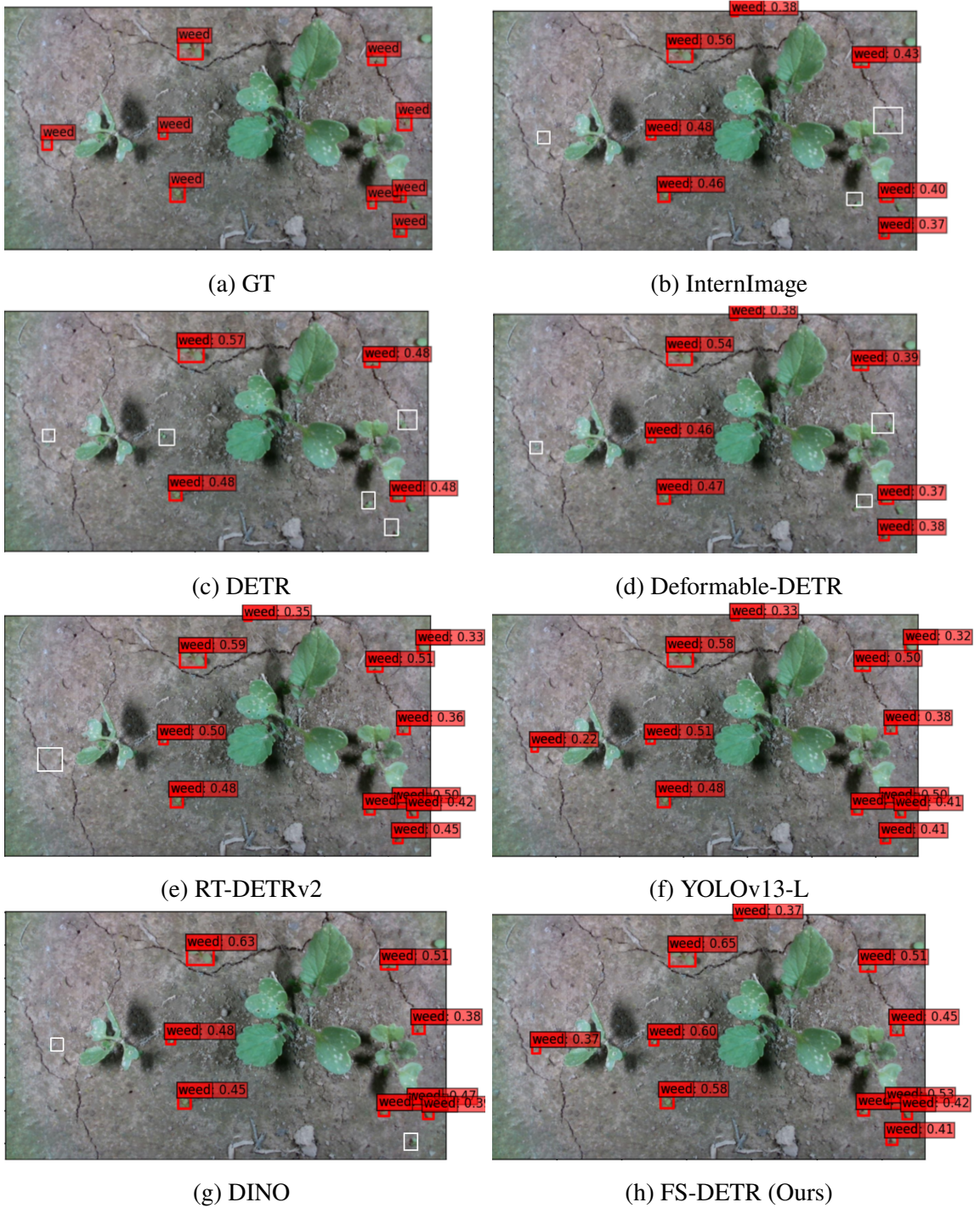


图 3.6 不同模型在 WeedCrop 数据集上的可视化对比

分图像还存在一定程度的模糊现象，使得目标边界不清晰、判别难度进一步增加。

在此类复杂条件下，FS-DETR 依然能够稳定检测，其预测框数量接近真实目标，边界与 Ground Truth 重合度高，定位精确。模型不仅能有效区分相邻或部分重叠的个体，还能保持较高且稳定的预测置信度，有效避免 DINO 中的误检与漏检，显示出更强的泛化能力。

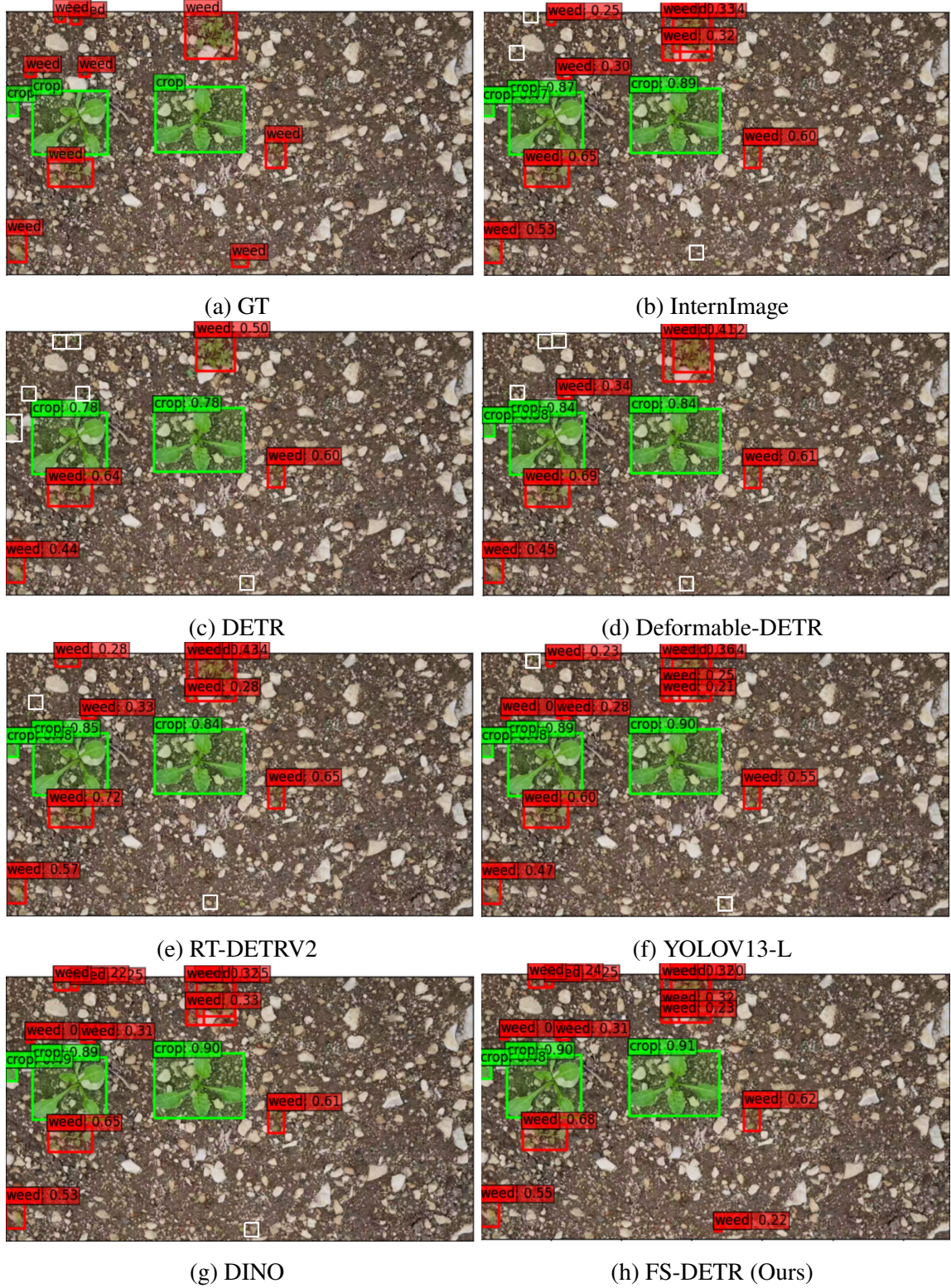


图 3.7 不同模型在 Lincolnbeet 数据集上的可视化对比

(3) 核心模块贡献分析

FS-DETR 的性能提升主要源于几个关键设计的有机协同。在特征构建阶段，模型通过混合特征融合机制，将空间域的局部结构信息与频域的全局纹理特征有机整

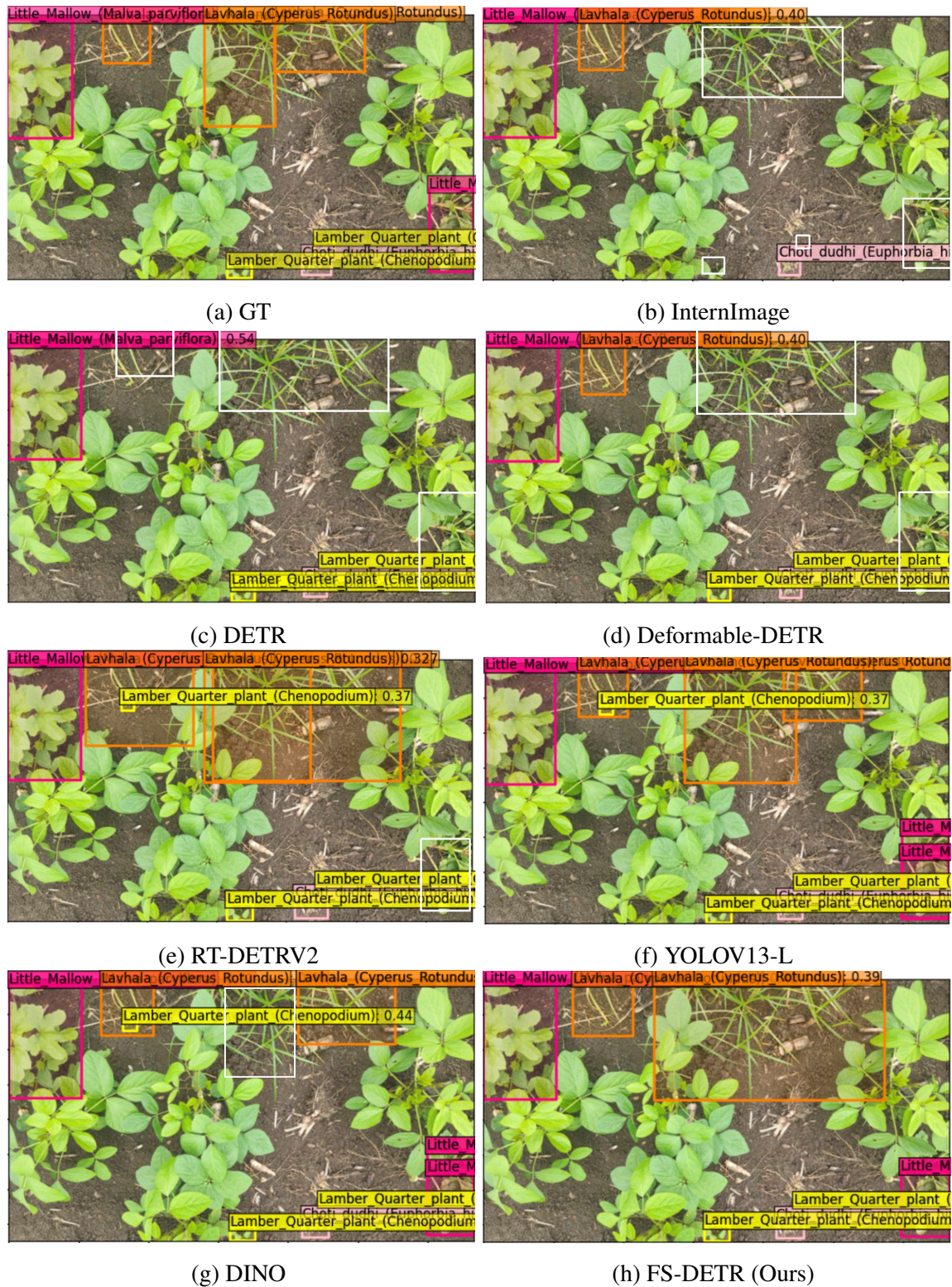


图 3.8 不同模型在 MH-Weed16 数据集上的可视化对比

合，使得对多尺度目标，尤其是小尺寸杂草的表征能力显著增强。在此基础上，编码器通过双域注意力机制对特征进行深度提炼，该机制将频域注意力的全局上下文感知优势与可变形注意力的局部几何建模能力巧妙结合，使模型能够适应杂草形态多

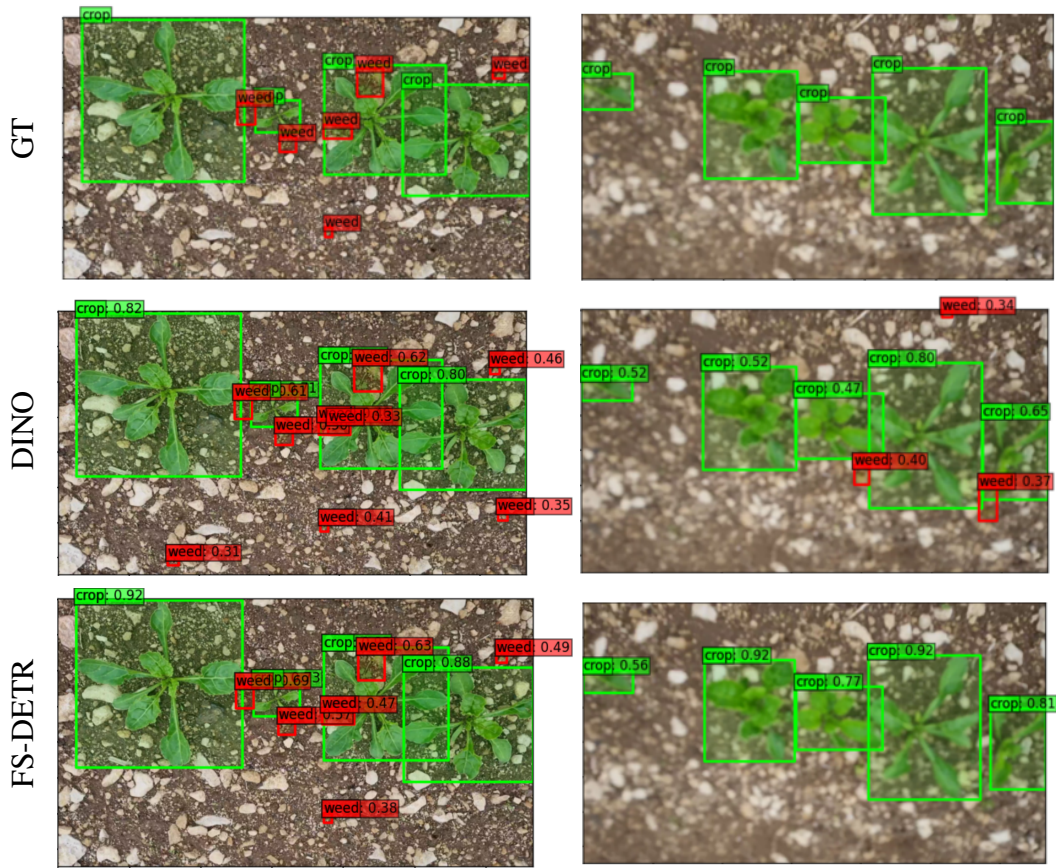


图 3.9 DINO 与 FS-DETR 在复杂场景下可视化对比

变、部分遮挡以及背景复杂等情况，从而提高定位和识别的准确性。同时，模型引入基于高斯分布和约束引导的标签分配策略，通过综合考虑中心位置、边界框重叠以及样本与真实目标的特征分布相似性，实现正样本的动态匹配。这一策略有效减少了因传统匹配不准确而产生的漏检，提高了模型在复杂田间环境下的召回率和检测稳健性。这三者的协同作用，使 FS-DETR 在小目标检测、复杂背景干扰抑制及整体精度上都表现出明显优势。

3.7 本章小结

本章节提出了一种面向复杂农田场景的双域融合 Transformer 杂草检测模型 FS-DETR。该模型通过构建混合特征融合模块与双域注意力机制，实现空间域与频域信息的协同建模，从而增强复杂场景下杂草目标的细粒度特征表达能力；同时，引入基于高斯分布与约束引导的标签分配策略，以提升正负样本匹配的准确性与训练过程的稳定性。通过上述模块的协同作用，模型在目标密集、遮挡严重、背景复杂及外观高度相似的场景中表现出更强的鲁棒性与判别能力。实验结果及可视化分析表

明, FS-DETR 在 Lincolnbeet、WeedCrop 及 MH-Weed16 数据集上的表现均优于现有方法, 在复杂场景下的杂草检测任务中展现出明显优势。

第四章 面向复杂场景的解码自适应机制

4.1 引言

在第三章中，本文针对复杂农田场景中杂草目标密集分布、小尺度目标占比高、背景干扰严重以及外观高度相似等问题，提出了双域融合 Transformer 杂草检测模型 FS-DETR。然而，在 DETR 框架中，编码器主要负责对输入特征进行全局建模，而目标类别判别与位置回归主要依赖解码阶段对目标查询的逐层更新。因此，当场景复杂度进一步增加时，模型性能往往不仅取决于特征表达能力，还受到解码阶段优化稳定性的影响。

从解码机制来看，复杂场景下的不稳定性主要来源于若干在标准 DETR 中默认采用的建模假设。大多数 DETR 方法采用固定数量的目标查询参与解码，其隐含假设是不同图像之间目标数量差异较小。然而在真实农业场景中，杂草目标数量与分布具有明显的不均衡特征：在目标密集区域，固定查询可能不足以覆盖所有潜在目标；而在目标稀疏区域，则会产生大量无效查询，从而降低查询与真实目标之间的匹配效率。同时，解码器中查询关系通常通过自注意力机制进行成对建模。然而，在杂草密集或外观相似区域，多个查询往往对应相邻目标，仅依赖两两关系难以刻画其整体关联结构，从而限制了对复杂目标布局的建模能力。此外，在训练过程中，查询与真实目标之间的匹配质量会随解码层逐步变化。然而，现有方法通常对所有正匹配查询施加统一强度的分类监督，忽略匹配质量差异，导致训练初期易产生噪声梯度，影响解码稳定性。

针对上述问题，本文在第三章 FS-DETR 模型基础上提出解码增强框架 FS-DETR-D (FS-DETR with Decoder Enhancement)，从查询分配、查询交互和分类监督三个方面对解码阶段进行改进。具体而言，提出基于置信度熵与空间感知的动态查询分配机制，以实现查询规模的自适应调整；引入基于超边增强的高阶查询交互机制，以刻画多查询之间的群组关系；并设计匹配质量自适应的分类损失函数，使监督强度与匹配可靠性相一致，从而提升模型在复杂农业场景下的检测稳定性与性能。

4.2 模型整体框架

本章在 FS-DETR 编码器输出特征的基础上，提出一种面向复杂农业场景杂草检测的解码增强框架 FS-DETR-D，以应对杂草目标密集、空间分布不均以及匹配关系动态变化等问题，从而提升模型解码过程的稳定性与检测性能。该框架在不改变第三章网络结构的前提下，围绕解码阶段的查询生成、特征交互及监督建模过程进行针对性优化，其整体结构如图4.1 所示。

在解码输入构建阶段，编码器输出候选查询特征及其对应的类别预测结果。基于多层预测信息，引入基于置信度熵与空间感知的动态查询分配机制（Entropy-Guided and Spatially-Aware Dynamic Query Allocation, ESQA），通过对场景复杂度的联合建模，自适应确定参与解码的查询数量，从而实现查询规模与场景复杂度之间的动态匹配。在查询特征建模阶段，通过引入超边增强的高阶查询交互机制（Hyperedge Query Interaction, HQI），解码器将查询表示为图结构中的节点，以刻画查询之间的高阶结构关系。在解码监督阶段，引入匹配质量自适应感知的分类损失函数（Adaptive Matchability-Aware Loss, AMAL），将查询与真实目标之间的匹配质量作为调节因子融入分类损失建模，使监督强度随匹配质量变化自适应调整，从而引导解码过程逐步生成更加稳定的目标预测结果。

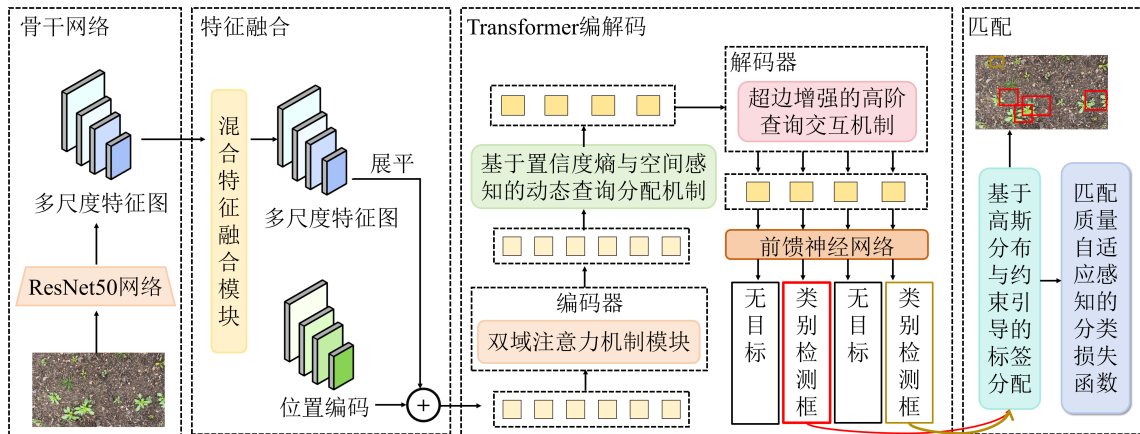


图 4.1 FS-DETR-D 整体架构图

4.3 基于置信度熵与空间感知的动态查询分配机制

在复杂农业场景中，杂草目标不仅在数量上具有不确定性，其空间分布也呈现出聚集与离散并存的特点，不同区域之间往往存在显著的密度差异。传统 DETR 系

列方法通常采用固定数量的目标查询，难以适应这种动态变化。在杂草密集区域，有限的查询数量难以覆盖全部目标，容易导致漏检；而在杂草稀疏区域，过多的查询则会引入冗余计算，并对模型训练产生干扰。此外，即使目标数量相近，目标在空间上的聚集性或离散性所对应的检测难度仍存在显著差异，这进一步表明仅依赖固定数量的查询难以有效应对复杂场景下的目标检测问题。

针对上述问题，本文提出基于置信度熵与空间感知的动态查询分配机制。该方法从置信度分布的不确定性与候选目标的空间紧密度两个维度对场景复杂度进行联合建模，并通过轻量级查询分配网络自适应预测最优查询数量，从而实现查询规模与场景复杂度之间的动态匹配。

4.3.1 置信度分布建模

不同编码器层在特征抽象层级和感受野范围上存在显著差异。浅层编码器更关注局部纹理、边缘及细粒度结构信息，对小尺度杂草目标和密集区域更为敏感；而深层编码器则侧重于全局语义建模，能够捕获目标之间的上下文关系及整体空间布局。在复杂农业场景中，杂草目标的密度、尺度变化及遮挡情况往往同时存在，仅依赖单一编码层的预测结果难以全面反映场景复杂度。

因此，首先对编码器多层输出的候选目标预测结果进行联合置信度分析，以获得对场景复杂程度更加稳健的估计。设 Transformer 编码器由 L 个层级堆叠而成，第 l 层编码器输出的类别分数矩阵表示为 $\mathbf{C}_{\text{enc}}^l \in \mathbb{R}^{B \times N \times D}$ ，其中 B 为批次大小， N 为候选框总数， D 为目标类别的数量。针对每一编码层，提取每个候选目标在所有类别上的最大预测分数，并通过 Sigmoid 函数进行归一化处理：

$$\mathbf{S}_{\text{conf}}^l = \sigma \left(\max_d (\mathbf{C}_{\text{enc}}^l) \right) \quad (4.1)$$

其中， $\sigma(\cdot)$ 为 Sigmoid 函数，用于将预测分数归一化到 $(0, 1)$ 区间； $\max_d(\cdot)$ 表示在类别维度 D 上进行最大值池化，用于获取各候选目标的最大类别响应作为置信度。

随后，对不同编码层的置信度分布进行融合，得到统一的候选目标置信度表示：

$$\mathbf{S}_{\text{conf}} = \sum_{l=1}^L \alpha_l \cdot \mathbf{S}_{\text{conf}}^l, \quad \sum_{l=1}^L \alpha_l = 1 \quad (4.2)$$

其中， α_l 是可学习参数，使模型能够根据训练数据自适应调整不同层级特征在复杂度评估中的贡献比例。

在获得融合后的置信度分布后，将其视为候选目标的概率分布，用于刻画目标在图像中的整体分布特性。第 i 个候选目标对应的概率定义为：

$$P_{\text{conf}}^i = \frac{S_{\text{conf}}^i}{\sum_{i=1}^N S_{\text{conf}}^i + \epsilon} \quad (4.3)$$

其中， ϵ 为避免数值不稳定的小量常数， N 表示所有候选目标的总数。

基于上述概率分布，定义场景复杂度的置信度熵为：

$$H_{\text{conf}} = - \sum_{i=1}^N P_{\text{conf}}^i \cdot \log(P_{\text{conf}}^i + \epsilon) \quad (4.4)$$

通过该指标，高熵值对应目标密集且分布复杂的图像，低熵值对应目标稀疏且背景简单的图像，可用于表征图像的整体复杂程度。

4.3.2 空间密度感知建模

仅依赖置信度分布难以充分刻画目标在空间上的结构特性，因此本文引入空间密度感知机制，从候选目标的几何分布角度对场景复杂度进行补充建模。

首先，通过编码器输出的候选框回归结果提取目标中心点坐标，并将其归一化至 $[0, 1]$ 空间，得到候选点集合 $\mathbf{X} \in \mathbb{R}^{B \times N \times 2}$ 。为降低计算复杂度并提升建模稳定性，仅保留置信度最高的 K_s 个候选点，形成候选点集合 $\{\mathbf{x}_i\}_{i=1}^{K_s} \subset \mathbb{R}^2$ 。为了刻画候选点之间的空间相对关系，本文计算任意两个候选点 $\mathbf{x}_i = (x_i, y_i)$ 与 $\mathbf{x}_j = (x_j, y_j)$ 之间的欧氏距离，其定义为：

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad i, j = 1, 2, \dots, K_s \quad (4.5)$$

为了避免传统最近邻方法中基于硬阈值的邻域划分带来的不连续性，本研究引入基于距离的软权重机制，将候选点的邻域关系连续化建模。具体地，对于第 i 个候选点，其与其他候选点之间的归一化权重定义为：

$$w_{ij} = \frac{\exp(-D_{ij}/\tau)}{\sum_{k=1}^{K_s} \exp(-D_{ik}/\tau)} \quad (4.6)$$

其中， $\exp(\cdot)$ 表示以自然常数 e 为底的指数函数， w_{ij} 表示候选点 $\mathbf{x}_j$ 对 $\mathbf{x}_i$ 的邻域权重， τ 为温度参数，用于控制权重分布范围。当 τ 较小时，权重更加集中于近邻点；当 τ 较大时，分布更加平滑。

基于上述权重，定义第 i 个候选点的加权邻域距离为：

$$d_i = \sum_{j=1}^{K_s} w_{ij} D_{ij} \quad (4.7)$$

该量刻画第 i 个候选点的局部空间尺度。在此基础上，将每个候选点的局部尺度与其置信度概率 P_{conf}^i 相结合，对所有候选点进行加权汇聚，得到全局空间结构量：

$$S_{\text{nn}} = \sum_{i=1}^{K_s} P_{\text{conf}}^i \cdot d_i \quad (4.8)$$

最终，将空间密度定义为：

$$\rho = \frac{1}{S_{\text{nn}} + \epsilon} \quad (4.9)$$

其中， ϵ 为防止数值不稳定的微小常数。

通过上述建模方式，候选目标的空间聚集程度能够得到有效量化。当目标在局部区域较为密集时，点间距离减小， S_{nn} 随之降低，空间密度 ρ 增大；反之，当目标分布较为稀疏时， S_{nn} 增大， ρ 相应减小。该空间密度指标可为动态查询分配提供空间复杂度参考。

4.3.3 动态查询分配机制

为了实现查询数量在语义复杂度与空间结构复杂度共同作用下的连续自适应调节，本文引入查询分配网络（Query Allocation Network, QAN）。首先对置信度熵与空间密度进行标准化处理：

$$\hat{H} = \frac{H_{\text{conf}}}{\max(H_{\text{conf}}) + \delta}, \quad \hat{\rho} = \frac{\rho}{\max(\rho) + \delta} \quad (4.10)$$

其中， δ 为防止数值不稳定的微小常数。

随后，将标准化后的置信度熵 $\hat{H}$ 与空间密度 $\hat{\rho}$ 进行联合建模，构建二维复杂度特征向量：

$$\mathbf{z} = [\hat{H}, \hat{\rho}] \quad (4.11)$$

联合特征 $\mathbf{z}$ 被输入轻量级查询分配网络 QAN，该网络能够将多维复杂度信号映射为连续比例系数 $r \in (0, 1)$ ，用于动态调节查询数量。QAN 首先通过全连接层对输入特征进行线性变换，并利用 ReLU 激活函数引入非线性表达能力，以建模语义不

确定性与空间分布结构之间的耦合关系。随后，通过第二层线性映射并结合 Sigmoid 激活函数，将输出约束至 $(0, 1)$ 区间，得到查询分配比例。整个过程表示为：

$$r = \text{QAN}(\mathbf{z}) = \text{Sigmoid}(W_2 \cdot \text{ReLU}(W_1 \mathbf{z} + b_1) + b_2) \quad (4.12)$$

其中， W_1 和 W_2 为可学习权重参数，分别用于特征映射与比例系数回归， b_1 和 b_2 为偏置项。为抑制训练初期的熵值波动，QAN 的输出 r 在前 3 次迭代中仅用于前向计算，不参与梯度回传，待模型初步收敛后再更新 QAN 参数。

基于该比例系数，动态计算查询数量：

$$K_{\text{dyn}} = \lceil K_{\min} + (K_{\max} - K_{\min}) \cdot r \rceil \quad (4.13)$$

其中， $K_{\min} = 100$ 与 $K_{\max} = 500$ 分别为查询数量的下界与上界，用于约束动态分配范围。该设定使模型能够应对极端场景，当图像中目标极度密集时动态比例 r 趋近于 1，查询数量自动上调至上限 $K_{\max}$ 以保证召回率并防止计算开销失控，当图像中无目标时置信度熵趋近于 0，查询数量自动下调至下限 $K_{\min}$ ，此时查询置信度均处于低值区间，可有效判别为背景从而避免误检或漏检。

根据计算得到的动态查询数量 K_{dyn} ，首先从候选框置信度 $\mathbf{S}_{\text{conf}}$ 中选取置信度最高的前 K_{dyn} 个索引，记为 I ：

$$I = \text{TopK}(\mathbf{S}_{\text{conf}}, K_{\text{dyn}}) \quad (4.14)$$

随后，根据索引 I 从查询候选集合 QueryProposals 中提取对应查询，得到动态查询集合：

$$\mathbf{Q} = \text{Gather}(\text{QueryProposals}, I) \quad (4.15)$$

其中， $\text{Gather}(\cdot)$ 表示按照索引 I 从查询候选集合中提取对应元素的操作。

在批量训练中，为解决不同图像查询数量不一致的问题，引入零填充与掩码机制。具体而言，将查询数不足 $K_{\max}$ 的样本通过零填充扩展至统一尺寸 $B \times K_{\max} \times C$ 。同时，构建二值掩码矩阵 $\mathbf{M} \in \{0, 1\}^{B \times K_{\max}}$ 标识查询有效性，其中有效位置为 0，填充位置为 1。该掩码作为 Transformer 解码器的键填充掩码参数，在自注意力计算中将填充位置注意力分数置为 $-\infty$ ，保证其在 Softmax 后权重接近零，从而在保持批量训练效率的同时显著减少冗余计算。

上述各模块的整体流程可统一表示为算法 4.1。

算法 4.1 基于置信度熵与空间感知的动态查询分配机制 (ESQA)

输入： 编码器各层类别分数矩阵 $\{\mathbf{C}_{enc}^l\}_{l=1}^L$ ；候选查询特征 **QueryProposals**；最小查询数量 $K_{\min}$ ；最大查询数量 $K_{\max}$ ；稳定常数 ϵ, δ ；温度参数 τ ；空间建模规模 K_s ；候选目标总数 N ；查询分配网络 QAN。

输出： 动态查询特征 **Q**；掩码向量 **M**。

- 1: **for** $l = 1$ **to** L **do**
 - 2: $\mathbf{S}_{\text{conf}}^l = \sigma(\max(\mathbf{C}_{enc}^l))$ // 第 l 层第 i 个候选置信度
 - 3: $\mathbf{S}_{\text{conf}} = \mathbf{S}_{\text{conf}} + \alpha_l \cdot \mathbf{S}_{\text{conf}}^l$ // 多层置信度融合
 - 4: **end for**
 - 5: $P_{\text{conf}}^i = \frac{\mathbf{S}_{\text{conf}}^i}{\sum_{j=1}^N \mathbf{S}_{\text{conf}}^j + \epsilon}$ // 构建置信度概率分布
 - 6: $H_{\text{conf}} = -\sum_{i=1}^N P_{\text{conf}}^i \log(P_{\text{conf}}^i + \epsilon)$ // 计算置信度熵
 - 7: $I_s = \text{TopK}(\mathbf{S}_{\text{conf}}, K_s)$ // 选择置信度最高的 K_s 个候选点
 - 8: $D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, $i, j \in I_s$ // 计算欧氏距离
 - 9: $w_{ij} = \frac{\exp(-D_{ij}/\tau)}{\sum_{k \in I_s} \exp(-D_{ik}/\tau)}$ // 计算软邻域权重
 - 10: $d_i = \sum_{j \in I_s} w_{ij} D_{ij}$ // 加权邻域距离
 - 11: $S_{\text{nn}} = \sum_{i \in I_s} P_{\text{conf}}^i \cdot d_i$ // 全局空间结构量
 - 12: $\rho = \frac{1}{S_{\text{nn}} + \epsilon}$ // 空间密度
 - 13: $\hat{H}_{\text{conf}} = \frac{H_{\text{conf}}}{\max(H_{\text{conf}}) + \delta}$, $\hat{\rho} = \frac{\rho}{\max(\rho) + \delta}$ // 归一化
 - 14: $r = \text{QAN}([\hat{H}_{\text{conf}}, \hat{\rho}])$ // 查询比例预测
 - 15: $K_{\text{dyn}} = \lceil K_{\min} + (K_{\max} - K_{\min}) \cdot r \rceil$ // 动态查询数量
 - 16: $I = \text{TopK}(\mathbf{S}_{\text{conf}}, K_{\text{dyn}})$ // 选择 Top- K_{dyn} 候选
 - 17: **Q** = **Gather(QueryProposals, I)** // 收集对应查询
 - 18: **Q** = **Pad(Q, $K_{\max}$)** // 查询填充
 - 19: **M** $\leftarrow$ **Ones($K_{\max}$)**; **M** $_{1:K_{\text{dyn}}} \leftarrow 0$ // 构建掩码
 - 20: **return Q, M**
-

4.4 超边增强的高阶查询交互机制

在复杂农业场景中，杂草目标通常呈现出密集分布、小尺度混杂及相互遮挡等特征，使得解码阶段的目标查询之间不仅存在显著的相关性，还往往表现为局部聚集的群组结构。然而，现有 DETR 系列方法在解码器中主要依赖基于成对关系的自

注意力机制，其注意力计算建立在两两相似度之上。尽管通过多层堆叠可以在一定程度上聚合上下文信息，但该过程缺乏对多查询群组结构的显式建模，使得高阶语义依赖难以被有效捕获。在杂草密集或形态相似区域，这种建模不足会限制查询之间的协同信息传递，降低特征表达的判别性，进而影响解码器对细粒度杂草目标的区分与定位能力。

针对上述问题，本文提出一种超边增强的高阶查询交互机制。在该机制中，将解码器中的目标查询建模为图结构中的节点，并引入可学习的超边对多个相关查询进行联合建模，从而将查询间的关系由传统的成对连接扩展为多节点协同交互。通过显式刻画查询之间的群组级关联结构，HQI 能够有效捕获杂草密集区域中潜在的高阶语义一致性与空间结构约束，进而增强查询特征的判别能力与稳定性，提高模型在复杂农业场景中的杂草目标检测性能。

4.4.1 超边自适应构建

在复杂农业场景中，杂草目标往往呈现出密集分布与结构性聚集特征，不同目标之间不仅存在显著的语义相关性，还伴随一定的空间共现关系。在基于 Transformer 的检测框架中，解码器输出的查询通常需要在全局范围内建模目标间的关联信息。然而，现有基于两两交互的建模方式难以充分刻画密集场景下的群组结构关系。如图4.2所示，设解码器输出的查询集合为 $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \dots, \mathbf{q}_N\}$ 。为建模查询之间的全局高阶关联关系，本文引入超边结构，通过构建可学习的超边原型对查询集合进行关系建模，从而增强模型对整体上下文信息的表达能力。

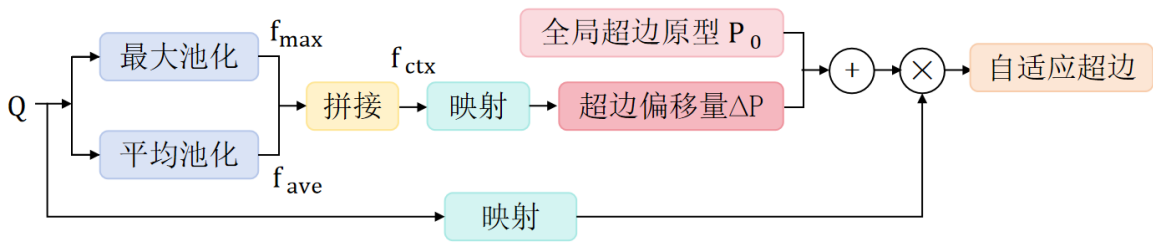


图 4.2 超边自适应构建结构图

具体而言，首先从查询特征中提取全局上下文信息。通过全局平均池化与最大池化分别获得整体分布特征 $\mathbf{f}_{ave}$ 与显著响应特征 $\mathbf{f}_{max}$ ，并在通道维度进行融合，得到上下文表示：

$$\mathbf{f}_{ctx} = \text{Concat}(\mathbf{f}_{ave}, \mathbf{f}_{max}) \in \mathbb{R}^{2C} \quad (4.16)$$

其中, C 表示特征维度。

在此基础上, 通过轻量级线性映射生成超边偏移量 $\Delta \mathbf{P} \in \mathbb{R}^{M \times C}$, 其中 M 表示超边数量, 并与可学习的初始超边原型 $\mathbf{P}_0 \in \mathbb{R}^{M \times C}$ 进行融合, 使超边结构能够自适应地响应不同输入场景中的查询分布变化, 从而得到自适应超边表示:

$$\mathbf{P} = \mathbf{P}_0 + \Delta \mathbf{P} \quad (4.17)$$

4.4.2 超边引导的高阶查询交互

在获得动态超边原型后执行基于超边的高阶查询特征交互, 其整体流程如图4.3所示。首先, 将节点特征通过线性投影得到查询向量 $\mathbf{z}_i \in \mathbb{R}^C$, 并使用多头机制将向量划分为多个子空间。在每个子空间中, 计算查询节点与超边原型的相似度 $S_{i,m}$, 并通过 Softmax 归一化得到参与矩阵:

$$A_{i,m} = \frac{e^{S_{i,m}}}{\sum_{j=1}^N e^{S_{j,m}}}, \quad i = 1, \dots, N, \quad m = 1, \dots, M \quad (4.18)$$

其中, $S_{i,m}$ 表示第 i 个查询与第 m 个超边原型之间的相似度得分, $A_{i,m}$ 表示第 i 个查询对第 m 个超边的共享权重, 这不仅保留了节点与超边的匹配信息, 也显式捕捉了查询之间的高阶关系。

随后, 基于参与矩阵完成节点到超边的特征聚合过程, 每个超边根据参与矩阵聚合所有查询特征, 并通过线性映射生成超边特征:

$$\mathbf{f}_m = \sigma \left(\mathbf{W}_e \sum_{i=1}^N A_{i,m} \mathbf{q}_i \right) \quad (4.19)$$

其中, $\mathbf{W}_e$ 为超边映射权重。

在此基础上, 进一步执行超边到节点的传播操作, 将超边特征反向分发至查询节点, 更新查询表示:

$$\mathbf{q}'_i = \sigma \left(\mathbf{W}_v \sum_{m=1}^M A_{i,m} \mathbf{f}_m \right) \quad (4.20)$$

其中, $\mathbf{W}_v$ 为节点映射权重。

最终通过残差连接与原始查询特征融合, 得到高阶增强后的查询特征:

$$\hat{\mathbf{Q}} = \mathbf{Q} + \mathbf{Q}' \quad (4.21)$$

通过这一机制，每个查询不仅保留自身的局部特征，还整合了全局的高阶语义信息，使解码器在处理密集、小尺度或遮挡目标时更加准确。

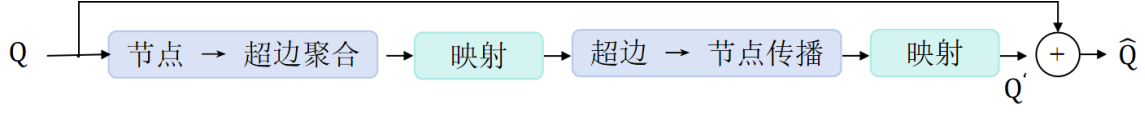


图 4.3 超边引导的高阶查询交互结构图

4.4.3 超边与自注意力特征融合

在解码器中，自注意力机制通过建模查询之间的全局依赖关系，为目标级特征提供了丰富的上下文信息，是 Transformer 解码过程中的关键组成部分。然而，自注意力主要基于成对关系进行特征交互，对于复杂农业场景中由多个杂草目标形成的高阶结构关系，其建模能力仍然有限。相比之下，超边增强查询交互模块能够显式刻画查询节点之间的群组关联，但若单独使用，可能削弱自注意力在上下文建模方面的优势。因此，本文将超边增强特征与解码器自注意力特征进行有机融合，以充分发挥二者在不同层面的互补优势。

如图4.4所示，基于超边的查询高阶特征交互模块以并行分支的形式嵌入解码器层结构。在解码器第 L 层中，上一层输出的查询特征 $\mathbf{Q}_{L-1}$ 既经过自注意力分支生成全局上下文特征 $\mathbf{Q}_{SA}$ ，又经过超边增强分支得到高阶语义增强特征 $\hat{\mathbf{Q}}$ 。为了在保持自注意力上下文建模能力的同时引入超边结构提供的高阶语义信息，本章节采用残差融合策略，将两类特征整合，形成第 L 层解码器的最终输出查询特征：

$$\mathbf{Q}_{out} = \mathbf{Q}_{SA} + \lambda \cdot \hat{\mathbf{Q}} \quad (4.22)$$

其中， $\mathbf{Q}_{SA}$ 表示自注意力输出的查询特征， $\hat{\mathbf{Q}}$ 为超边增强后的高阶特征表示， λ 为可学习的融合权重参数，在训练过程中通过反向传播自适应更新，用于动态调节高阶语义信息在融合特征中的贡献比例。

通过上述融合策略，解码器查询在各层能够同时建模目标间的全局上下文关系与高阶结构信息，从而提升特征的语义判别能力与结构表达能力。在不显著增加模型复杂度的前提下，该机制有效增强了模型在杂草密集、小尺度及遮挡场景下的特征表示能力，并为后续分类与位置回归提供更加稳定的特征基础。

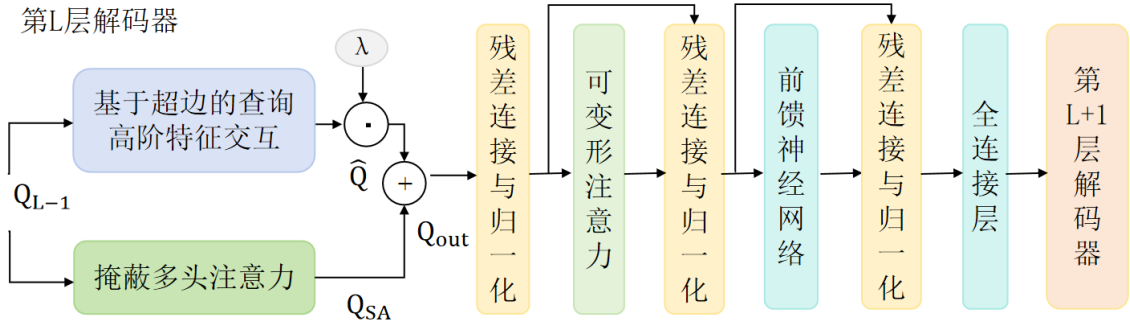


图 4.4 超边与自注意力特征融合结构图

4.5 匹配质量自适应感知的分类损失函数

在查询驱动的 Transformer 目标检测框架中，解码阶段通过多层迭代不断修正目标查询的类别预测与边界框位置，其优化过程高度依赖查询预测结果与真实目标之间的匹配关系。然而，在复杂农业场景中，杂草目标往往呈现出密集分布、尺度差异显著以及相互遮挡等特征，导致解码初期大量查询仅能形成低一致性甚至不可靠的匹配关系。现有 DETR 系列方法在解码阶段通常采用与匹配质量弱相关或离散建模的分类损失，难以刻画匹配质量随解码逐步演化的连续变化特性，从而削弱分类监督与定位优化之间的协同作用。针对这一问题，本章节提出一种匹配质量自适应感知的分类损失函数，通过将查询与真实杂草目标之间的匹配质量显式引入分类损失建模，使分类监督强度能够根据匹配可靠性进行自适应调整，从而提高模型在复杂农业场景中的杂草目标检测稳定性与检测性能。

4.5.1 匹配质量建模与监督冲突分析

在 DETR 解码阶段，查询与真实目标之间的匹配结果由匈牙利算法确定，其匹配质量通常通过预测框与真实框之间的交并比（Intersection over Union, IoU）进行度量。记第 i 个查询的匹配质量为 $q \in [0, 1]$ ，对应其 IoU 值。

该指标反映了当前解码阶段查询预测结果在空间定位上的可靠程度。然而，现有方法普遍将匹配关系视为离散事件，仅在分类标签层面区分正样本（ $y = 1$ ）与背景样本（ $y = 0$ ），而未在分类损失中显式建模匹配质量的连续差异。这种设定在解码初期尤为明显，大量查询虽然已被分配为正样本，但其定位质量仍然较低，分类置信度与空间定位之间容易出现不一致现象。为缓解上述矛盾，VariFocalLoss（VFL）通过引入 IoU 权重，在一定程度上增强了对高质量预测的分类监督，其定义如下：

$$\text{Loss}_{\text{VFL}}(p, q) = \begin{cases} -q(q \log p + (1 - q) \log(1 - p)) & q > 0 \\ -\alpha p^\gamma \log(1 - p) & q = 0 \end{cases} \quad (4.23)$$

其中, p 表示预测框对应类别的预测概率, q 表示预测框与真实框之间的交并比; α 与 γ 为平衡因子与聚焦因子, 用于调节正负样本之间的损失比例。

尽管 VFL 在一定程度上强化了对高质量预测的分类监督, 但其监督强度与 IoU 数值高度耦合, 当匹配质量较低时, 分类损失显著减弱, 使得低质量查询在解码初期难以对模型参数更新产生有效影响。

4.5.2 自适应分类损失函数设计

为使分类监督能够适应匹配质量在解码过程中的动态变化, 本文将匹配质量敏感系数建模为匹配质量的单调函数, 使其随匹配可靠性连续变化, 定义为:

$$\gamma(q) = \gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot q \quad (4.24)$$

其中, q 表示预测框与真实框之间的 IoU 值, $\gamma_{\min}$ 和 $\gamma_{\max}$ 分别表示低质量与高质量匹配对应的损失调节强度。

在此基础上, 定义匹配质量自适应感知损失函数 (AMAL) 如下:

$$\text{Loss}_{\text{AMAL}}(p, q, y) = \begin{cases} -q^{\gamma(q)} \log p - (1 - q^{\gamma(q)}) \log(1 - p) & y = 1 \\ -p^{\gamma(q)} \log(1 - p) & y = 0 \end{cases} \quad (4.25)$$

其中, p 表示类别预测概率, $y \in \{0, 1\}$ 为分类标签, $y = 1$ 表示该查询与某一真实目标匹配, $y = 0$ 表示该查询被判定为背景; $\gamma(q)$ 为匹配质量敏感系数, 用于根据 IoU 动态调整损失权重。

从优化机理上看, AMAL 能够根据查询与真实目标之间匹配质量的变化, 自适应调节分类监督强度。对于匹配质量较低的查询, 匹配质量敏感系数 $\gamma(q)$ 接近其下界 $\gamma_{\min}$, 即使预测置信度较高但定位尚不可靠, 仍会对分类结果施加有效惩罚, 从而避免由于匹配不确定性引入的梯度噪声, 保证解码早期训练的稳定性。随着解码过程推进, 预测框位置逐步收敛, 查询与真实目标之间的匹配质量不断提升, $\gamma(q)$ 随之增大。在此过程中, 高匹配质量查询在分类损失中的权重相应提高, 从而对其施

加更强的约束，促使模型重点优化定位更为准确的预测结果，使分类置信度与空间定位精度保持一致收敛。

需要说明的是，本文以 IoU 作为匹配质量的唯一度量指标，而未引入分类置信度。这是因为在解码阶段，IoU 能够直接反映预测框的空间定位精度与真实目标之间的几何一致性，而分类置信度本身尚处于优化过程中，将其纳入匹配质量度量可能引入额外的不确定性与耦合噪声。此外，在训练后期当匹配质量普遍较高时，AMAL 并不会退化为标准的二值离散监督模式，因为 $\gamma(q)$ 随 IoU 连续变化，即使 q 接近 1， $\gamma(q)$ 仍处于较大值，损失函数始终保持对高质量样本的差异化强化监督。对于匹配质量极差的困难样本，AMAL 适当降低其分类监督强度，但这并非放弃学习，而是将优化重心转移至定位分支；随着解码层数加深，这些样本的匹配质量逐步提升， $\gamma(q)$ 随之增大，分类监督强度自动恢复，从而实现对难样本的有效学习。

通过上述基于匹配质量的自适应调节机制，AMAL 在整个解码过程中实现了从低匹配质量查询的稳定修正到高匹配质量预测的强化优化，在统一损失函数框架内协调了分类学习与定位精度的协同提升。

4.6 实验结果与分析

4.6.1 实验数据及实验设置

本章实验设置在整体上与第三章保持一致，实验环境与训练策略均采用相同配置。在数据集方面，在第三章所用数据集基础上进一步引入 PASCAL VOC 2007+2012^[84]数据集，以验证模型的泛化能力，具体细节参见第三章相关内容。为验证所提出解码机制改进的有效性，在实验过程中除相关模块外，其余训练参数均保持不变，以确保实验结果的公平性与可比性。

4.6.2 实验评估和对比分析

为验证第四章提出的 FS-DETR-D 模型在复杂农业场景杂草目标检测任务中的有效性，本文在 WeedCrop、Lincolnbeet 和 MH-Weed16 三个杂草数据集上进行了对比实验，并与主流 CNN 检测方法、典型 Transformer 系列检测器以及第三章提出的 FS-DETR 基线模型进行了系统比较。整体结果如表4.1、表4.2和表4.3所示。可以看出，本文方法在三个数据集上均取得了最优的检测性能，在整体 AP 以及不同 IoU 阈

表 4.1 WeedCrop 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	44.9	75.0	48.9	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	45.1	77.1	46.7	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	46.3	78.8	47.7	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	45.4	77.3	47.4	42.9	48.5	68.7
RT-DETRv2 ^[80]	2024CVPR	ResNet50	45.9	77.4	50.0	44.5	50.3	57.1
InternImage ^[81]	2023CVPR	InternImage	42.0	74.6	43.8	45.0	37.1	<u>69.7</u>
DETR ^[19]	2020ECCV	ResNet50	34.5	60.2	36.3	32.4	32.8	44.6
Deformable-DETR ^[20]	2021ICLR	ResNet50	41.2	77.4	40.6	41.3	41.9	57.7
Co-DETR ^[35]	2023ICCV	ResNet50	42.9	72.4	48.5	45.3	41.2	64.1
DINO ^[30]	2023ICLR	ResNet50	45.8	78.2	48.5	43.9	49.9	59.5
DDQ ^[24]	2023CVPR	ResNet50	45.6	<u>80.1</u>	46.4	46.0	47.8	66.4
Align-DETR ^[82]	2024BMVC	ResNet50	41.3	69.2	44.6	40.8	40.7	67.8
H-DETR ^[33]	2023CVPR	ResNet50	36.7	64.4	39.2	37.1	37.5	48.4
DEIM ^[83]	2025CVPR	ResNet50	45.7	75.4	49.4	44.9	47.5	63.2
FS-DETR	—	ResNet50	<u>47.2</u>	79.3	<u>53.0</u>	<u>47.2</u>	<u>51.1</u>	68.8
FS-DETR-D	—	ResNet50	48.3	80.2	53.9	48.4	52.2	69.8

值指标下均表现出稳定优势。

在 WeedCrop 数据集上, 本文方法 FS-DETR-D 取得了 48.3% 的整体 AP, 相比 FS-DETR 基线模型提升了 1.1%。与 CNN 方法相比, 本文方法相较 InternImage 提升了 6.3%, 相较 YOLOv9-C、YOLOv10-L 和 YOLOv13-L 分别提升 3.4%、3.2% 和 2.0%。此外, 在小目标检测方面, AP_S 达到 48.4%, 相比基线模型进一步提升, 表明模型在杂草密集分布和背景干扰较强的区域中能够更有效地识别细粒度目标。与 Transformer 方法相比, 相较于 DINO 和 DDQ 分别提升 2.5% 和 2.7%, 进一步验证了解码阶段优化策略的有效性。

在 Lincolnbeet 数据集上, FS-DETR-D 同样保持了稳定优势。整体 AP 达到 61.2%, 相比 FS-DETR 提升 0.8%, 在 AP₅₀ 和 AP₇₅ 指标上分别达到 82.9% 和 65.0%。与 YOLO 系列方法相比, 本文方法在 YOLOv9-C、YOLOv10-L 和 YOLOv13-L 上分别取得 4.0%、3.4% 和 2.5% 的性能提升, 说明在目标分布密集且尺度差异较大的场景下, 改进模型能够更好地建模复杂目标关系。同时, 小目标 AP 提升至 17.2%, 中等目标 AP 达到 43.6%, 均优于基线模型, 表明模型在密集杂草场景中具有更稳定的目标区分能力。

在更具挑战性的 MH-Weed16 数据集上, FS-DETR-D 依然取得了最优结果, 整体

表 4.2 Lincolnbeet 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	57.2	78.6	60.9	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	57.8	79.4	61.6	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	58.7	79.7	62.6	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	59.3	81.2	63.2	14.2	41.5	72.8
RT-DETRv2 ^[80]	2024CVPR	ResNet50	59.0	80.1	62.8	13.5	41.2	71.7
InternImage ^[81]	2023CVPR	InternImage	58.3	80.6	61.7	14.6	41.2	70.5
DETR ^[19]	2020ECCV	ResNet50	44.6	60.9	42.0	4.3	28.7	58.6
Deformable-DETR ^[20]	2021ICLR	ResNet50	50.4	76.0	53.9	4.9	30.9	62.8
Co-DETR ^[35]	2023ICCV	ResNet50	58.8	80.9	62.1	<u>17.0</u>	41.1	71.1
DINO ^[30]	2023ICLR	ResNet50	59.4	81.6	63.2	13.0	41.8	71.7
DDQ ^[24]	2023CVPR	ResNet50	59.5	81.7	64.2	15.1	42.5	72.3
Align-DETR ^[82]	2024BMVC	ResNet50	58.5	80.0	62.1	12.9	41.0	70.6
H-DETR ^[33]	2023CVPR	ResNet50	58.3	80.6	61.7	14.6	41.2	70.5
DEIM ^[83]	2025CVPR	ResNet50	58.5	80.2	62.3	14.4	40.8	70.6
FS-DETR	—	ResNet50	<u>60.4</u>	<u>82.2</u>	<u>64.3</u>	16.5	<u>43.3</u>	<u>73.0</u>
FS-DETR-D	—	ResNet50	61.2	82.9	65.0	17.2	43.6	73.2

AP 达到 33.4%，相比 FS-DETR 提升 0.9%。与 InternImage 相比，整体性能提升 20.0%，同时也优于 YOLOv9-C、YOLOv10-L 和 YOLOv13-L 等方法。值得注意的是，该数据集中小目标比例较高且背景复杂，本文方法在 AP_S 指标上由 4.9% 提升至 5.5%，说明模型对微小杂草目标具有更强的感知能力。此外，在 AP₇₅ 指标上提升至 28.4%，进一步表明模型在复杂背景条件下仍能保持较稳定的定位精度。

本文方法在三个数据集上均表现出稳定的性能提升，这与解码阶段建模方式的改进密切相关。动态查询分配能够随场景复杂度变化调整有效查询数量，使检测过程在目标密集区域获得更充分的表达，从而减少漏检。同时，高阶查询交互进一步强化了相邻目标之间的区分能力，使密集或重叠目标不易混淆。此外，匹配质量自适应监督机制根据预测与真实目标之间的匹配质量动态调节监督强度，使模型更加关注高质量匹配样本，同时抑制低质量匹配带来的干扰，从而提升训练过程的稳定性与有效性。

表 4.3 MH-Weed16 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv9-C ^[76]	2024arXiv	CSPDarknet	31.8	63.7	26.3	—	—	—
YOLOv10-L ^[77]	2024arXiv	CSPDarknet	31.9	64.6	26.8	—	—	—
YOLOv13-L ^[78]	2025arXiv	—	32.0	63.4	27.4	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	25.6	53.1	21.2	4.6	16.8	27.9
RT-DETRv2 ^[80]	2024CVPR	ResNet50	26.2	54.1	21.7	4.7	16.9	28.4
InternImage ^[81]	2023CVPR	InternImage	13.4	29.6	11.2	3.5	11.8	14.8
DETR ^[19]	2020ECCV	ResNet50	15.4	40.3	9.8	0.1	4.6	14.8
Deformable-DETR ^[20]	2021ICLR	ResNet50	19.7	46.4	14.0	0.5	9.3	20.9
Co-DETR ^[35]	2023ICCV	ResNet50	31.2	66.1	24.7	4.2	20.7	33.6
DINO ^[30]	2023ICLR	ResNet50	30.9	64.5	25.3	4.3	16.7	34.0
DDQ ^[24]	2023CVPR	ResNet50	30.9	63.8	27.2	4.6	16.6	33.9
Align-DETR ^[82]	2024BMVC	ResNet50	31.1	63.2	<u>27.6</u>	4.0	17.5	34.1
H-DETR ^[33]	2023CVPR	ResNet50	30.4	62.6	25.6	4.2	20.1	33.0
DEIM ^[83]	2025CVPR	ResNet50	30.6	60.1	27.4	4.4	19.3	33.5
FS-DETR	—	ResNet50	<u>32.5</u>	<u>66.7</u>	27.5	<u>4.9</u>	20.3	<u>35.3</u>
FS-DETR-D	—	ResNet50	33.4	67.3	28.4	5.5	<u>20.6</u>	35.4

4.6.3 消融实验

本章节在第三章提出的 FS-DETR 模型基础上, 进一步构建解码增强框架 FS-DETR-D, 用于研究复杂农业场景下解码阶段的自适应优化机制。针对杂草目标密集分布、尺度差异显著以及匹配关系动态变化等问题, 本文从解码过程入手, 引入动态查询分配、高阶查询交互以及解码监督自适应建模等关键策略, 以提升模型在复杂农业场景中的检测稳定性与识别精度。

所有消融实验均以第三章提出的 FS-DETR 作为基线模型, 在保持网络主体结构及训练策略不变的前提下, 采用逐步集成的方式进行评估。在 WeedCrop 数据集上, 基线 FS-DETR 的检测性能 AP 为 47.2%, 该结果作为本章各项消融实验的性能参考基准。

(1) 模块的有效性

为系统评估第四章各解码增强模块的有效性, 本文对 ESQA、HQI 以及 AMAL 进行逐一及组合消融实验, 实验结果如表4.4所示。

当仅引入 ESQA 模块时, 模型 AP 由 47.2% 提升至 47.7%, 小目标检测性能提升更为明显, 表明基于置信度熵与空间感知的动态查询分配机制能够有效缓解固定

表 4.4 模块有效性的消融实验

ESQA	HQI	AMAL	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
			47.2	79.3	53.0	47.2	51.1	68.8
✓			47.7	79.5	53.4	48.0	51.6	69.1
	✓		47.9	79.6	53.4	47.9	51.6	69.3
		✓	47.5	79.5	53.5	47.6	51.5	69.2
✓	✓		48.1	79.8	53.8	48.1	51.8	69.5
✓		✓	47.9	79.6	53.6	48.1	52.0	69.5
	✓	✓	48.0	79.8	53.7	48.2	52.0	69.5
✓	✓	✓	48.3	80.2	53.9	48.4	52.2	69.8

查询数量在复杂场景下带来的冗余与漏检问题。单独引入 HQI 模块后, AP 提升至 47.9%, 验证了通过高阶结构关系建模能够增强查询间的信息交互, 从而提升杂草密集分布及遮挡情况下的检测一致性。仅使用 AMAL 时, AP 提升至 47.5%, 验证了基于匹配质量的自适应监督机制在稳定解码阶段分类优化中的有效性。

在两两组合实验中, ESQA+HQI、ESQA+AMAL 以及 HQI+AMAL 均取得进一步性能提升, 其中 AP 均稳定接近或达到 48.0%, 且在小目标检测指标上表现更为突出, 表明动态查询建模、高阶查询交互与自适应监督机制之间具有良好的互补性。当同时引入 ESQA、HQI 和 AMAL 三个模块时, 模型 AP 达到 48.3%, 相较基线提升 1.1%, AP_S 提升至 48.4%, 充分验证了所提出解码自适应机制在复杂农业场景杂草检测任务中的协同优化效果。

(2) 基于置信度熵与空间感知的动态查询分配机制的作用

为验证 ESQA 在复杂农业场景下的有效性, 本文将其与不同固定查询数量策略进行了对比实验。考虑到杂草密集区域的目标数量可能远超常规设置, 实验设置了固定查询数量 $K = 100$ 、 $K = 300$ 以及更大容量的 $K = 500$ 作为基准。同时, 为了验证不同感知维度的贡献, 分别设置了仅基于置信度熵、仅基于空间感知以及二者融合的动态分配策略, 动态查询范围均设定为 $[100, 500]$ 。实验重点评估模型在检测精度 (AP) 以及推理速度 (FPS) 之间的权衡, 结果如表 4.5 所示。

从表中可以看出, 随着固定查询数量从 100 增加至 300, 模型整体 AP 由 46.7% 提升至 47.2%, 但推理速度 FPS 下降至 12.0。进一步增至 $K = 500$ 时, 虽然计算负载增加导致 FPS 降至 10.5, AP 却仅微幅提升至 47.3%, 说明单纯增加查询数量已无法显著改善检测精度, 且对推理效率造成明显拖累。相比之下, 引入动态查询分配

表 4.5 不同查询策略下的性能与效率对比

查询策略	FPS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
固定查询 $K = 100$	13.5	46.7	78.8	52.5	46.5	50.5	68.4
固定查询 $K = 300$	12.0	47.2	79.3	53.0	47.2	51.1	68.8
固定查询 $K = 500$	10.5	47.3	79.1	53.1	47.3	51.2	68.9
基于置信度熵的动态查询分配机制	11.8	47.6	79.3	53.3	47.8	51.7	69.1
基于空间感知的动态查询分配机制	11.9	47.5	79.1	53.2	47.6	51.5	69.0
基于置信度熵与空间感知的动态查询分配机制	11.6	47.7	79.5	53.4	48.0	51.6	69.1

机制能够显著提升模型性能。单独使用置信度熵或空间感知策略，分别将 AP 提升至 47.6% 和 47.5%，且推理速度 FPS 保持在 11.8 和 11.9 左右，这表明语义不确定性与几何分布特征在衡量场景复杂度时各自具有独立作用。基于二者的互补性，将置信度熵与空间感知信息融合，构建基于置信度熵与空间感知的动态查询分配机制，以实现更加精细的查询资源分配。该机制能够根据图像场景自适应调整查询数量，使整体 AP 达到 47.7%。尤其在小目标检测上，ESQA 相比固定 $K = 300$ 的策略提升了 0.8%，达到 48.0%，有效缓解了密集小目标的漏检问题。尽管 ESQA 增加了查询数量，使 FPS 微降至 11.6，但仍明显优于固定 $K = 500$ 的效率，并接近 $K = 300$ 的基准水平。这一结果充分证明，ESQA 能够根据场景的语义与空间复杂度自适应地分配计算资源，在保证推理实时性的前提下实现了检测精度的最大化。

（3）超边增强的高阶查询交互机制的作用

为进一步分析 HQI 模块在查询特征交互中的作用，本文对不同查询交互机制进行消融实验。具体而言，在解码器结构中分别采用三种查询交互方式：仅使用自注意力机制、仅使用基于超边增强的高阶查询交互机制，以及自注意力与基于超边增强的高阶查询交互机制联合使用，实验结果如表 4.6 所示。

表 4.6 不同查询交互机制的消融实验

查询交互机制	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
自注意力机制	47.2	79.3	53.0	47.2	51.1	68.8
超边增强的高阶查询交互机制	47.7	79.5	53.3	47.7	51.4	69.0
自注意力 + 超边增强的高阶查询交互机制	47.9	79.6	53.4	47.9	51.6	69.3

从表中可以看出，仅使用自注意力机制通过计算查询间的成对相似度来捕获全局依赖关系获得了 47.2% 的 AP。当仅采用基于超边增强的高阶查询交互机制替代自

注意力时，利用超图结构显式建模多查询间的高阶复杂关联，有效增强了特征的结构化表达，使模型性能提升至 47.7%。将自注意力机制与 HQI 模块联合使用，通过融合全局上下文信息与局部高阶结构约束实现特征互补，进一步将 AP 提升至 47.9%，并在 AP_S 等指标上取得更好的表现。实验结果表明，自注意力机制与超边交互机制在特征建模方面具有互补性，二者结合能够进一步提升解码阶段查询特征的表达能力，从而提高模型在复杂场景中的检测性能。

在此基础上，为进一步分析高阶查询交互信息在融合过程中的作用强度，本文对融合权重 λ 进行消融实验，通过设置不同取值，探究其在特征融合中的贡献程度，实验结果如表 4.7 所示。

表 4.7 融合权重 λ 的消融实验

λ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.25	47.4	79.3	53.0	47.3	51.2	68.9
0.5	47.7	79.5	53.2	47.6	51.4	69.2
0.75	47.8	79.6	53.3	47.8	51.5	69.2
1.0	47.6	79.4	53.2	47.7	51.3	69.0
可学习	47.9	79.6	53.4	47.9	51.6	69.3

从表中可以看出，融合权重 λ 对模型性能具有显著影响。随着 λ 由 0.25 增大至 0.75，模型性能呈现稳步提升趋势，表明引入高阶查询交互信息有助于增强特征表达能力。当 $\lambda = 0.75$ 时模型性能达到最优，说明此时自注意力机制与超边交互机制之间实现了较为合理的权衡。当 λ 进一步增大至 1.0 时，模型性能出现小幅下降，表明过度引入高阶结构信息可能对全局上下文建模产生一定干扰。相比之下，采用可学习的 λ 能够在训练过程中自适应调节不同信息源的融合比例，从而取得最佳检测性能。此外，在小目标检测指标 AP_S 上的提升更为明显，进一步验证了该机制在密集及小尺度目标建模中的有效性。

(4) 匹配质量自适应感知的分类损失函数的作用

为验证所提出 AMAL 的有效性，本文在相同实验设置下，对不同分类损失函数进行了消融对比实验。实验选取当前具有代表性的 VariFocal Loss (VFL) 作为对比方法，并与本文提出的 AMAL 进行性能评估。VFL 通过引入 IoU 权重对高质量匹配样本施加更强监督，在一定程度上缓解了分类置信度与定位质量不一致的问题。然而，其监督强度与匹配质量呈固定耦合关系，在匹配质量较低阶段对样本的约束较

弱，限制了模型在解码早期的稳定优化能力。相比之下，AMAL 将匹配质量显式建模为动态调节因子，通过构建随匹配质量连续变化的自适应监督机制，使分类损失能够根据预测可靠性自动调整优化强度，从而在不同解码阶段实现更合理的梯度分配。实验结果如表4.8所示。相较于 VFL，AMAL 在整体检测性能上进一步提升，将 AP 提升至 47.5%。同时，小目标检测指标 AP_S 由 47.4% 提升至 47.6%，表明所提出方法在复杂场景下对小尺度目标具有更强的建模能力与鲁棒性。

表 4.8 不同分类损失函数的消融实验

分类损失函数	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
VFL	47.3	79.3	53.2	47.4	51.2	69.0
AMAL	47.5	79.5	53.5	47.6	51.5	69.2

为更直观地分析损失函数在不同匹配质量下的行为差异，本文分别绘制其关于预测置信度 p 的变化曲线。

在 IoU 为 0.1 的低匹配质量条件下，如图4.5所示，AMAL 对高置信度预测表现出更强的惩罚能力。当模型对定位不准确的候选框给出较高置信度时，AMAL 能够显著提高损失值，从而有效抑制由错误高置信度预测带来的梯度噪声，提升解码初期的训练稳定性。相比之下，VFL 在该区域的约束相对较弱。

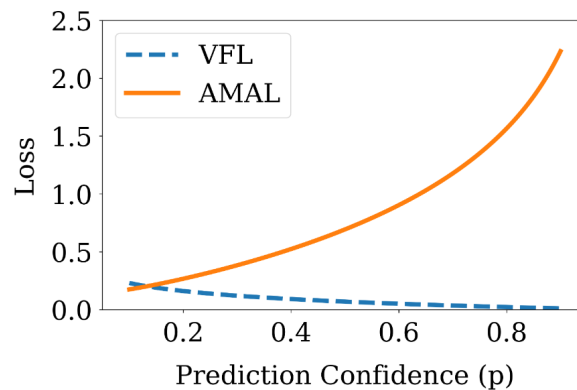


图 4.5 IoU=0.1 条件下 AMAL 与 VFL 损失-置信度曲线对比

在 IoU 为 0.9 的高匹配质量条件下，如图4.6所示，AMAL 在高置信度区间仍保持较高的损失值，对预测结果施加更严格的约束，从而促使模型进一步提升预测质量。而 VFL 在高置信度区域的惩罚逐渐减弱，约束能力相对有限。

因此，AMAL 在不同匹配质量条件下均表现出更强的自适应约束能力，能够在

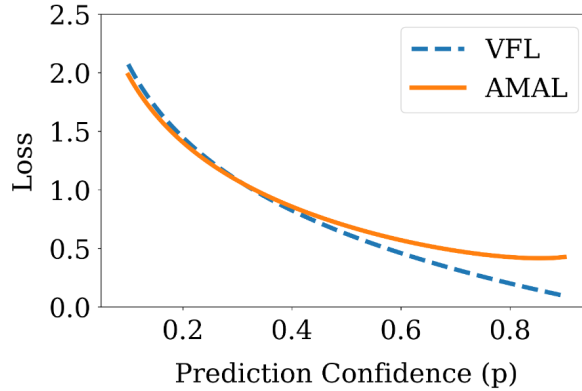


图 4.6 IoU=0.9 条件下 AMAL 与 VFL 损失-置信度曲线对比

低质量区域抑制噪声，在高质量区域强化监督，从而整体提升模型的训练稳定性与检测性能。

在验证 AMAL 有效性后，为进一步探究匹配质量敏感系数 $\gamma(q)$ 的动态调节机制对模型性能的具体影响，并确定最佳的超参数区间，本文对 $\gamma(q)$ 的不同取值策略进行了详细的敏感性分析实验。实验设置包括退化为固定值的静态策略以及不同范围的动态调节策略，具体结果如表4.9所示。

表 4.9 匹配质量敏感系数 $\gamma(q)$ 参数区间的消融实验

$\gamma_{\min}$	$\gamma_{\max}$	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
2.0	2.0	47.1	79.0	52.8	46.8	50.8	68.7
0.0	2.0	47.3	79.1	53.2	47.0	51.2	68.9
1.0	3.0	47.5	79.5	53.5	47.6	51.5	69.2
2.0	4.0	47.2	79.3	53.1	47.3	51.0	69.0

从表4.9可以看出，匹配质量敏感系数 $\gamma(q)$ 的调节方式对模型性能具有显著影响。当采用静态设置 $\gamma = 2.0$ 时，模型无法根据匹配质量变化调整监督强度，其性能相对较低。相比之下，引入动态调节机制后，模型性能均有所提升，验证了匹配质量自适应调节策略的有效性。当 $\gamma(q)$ 在 $[1.0, 3.0]$ 范围内变化时，模型取得最优性能。这表明较小的 γ 有助于低匹配质量阶段的稳定优化，而较大的 γ 能够强化对高质量预测的约束，从而实现分类与定位的协同提升。若区间设置过低或过高，则会破坏这种平衡关系，进而限制模型性能的进一步提升。

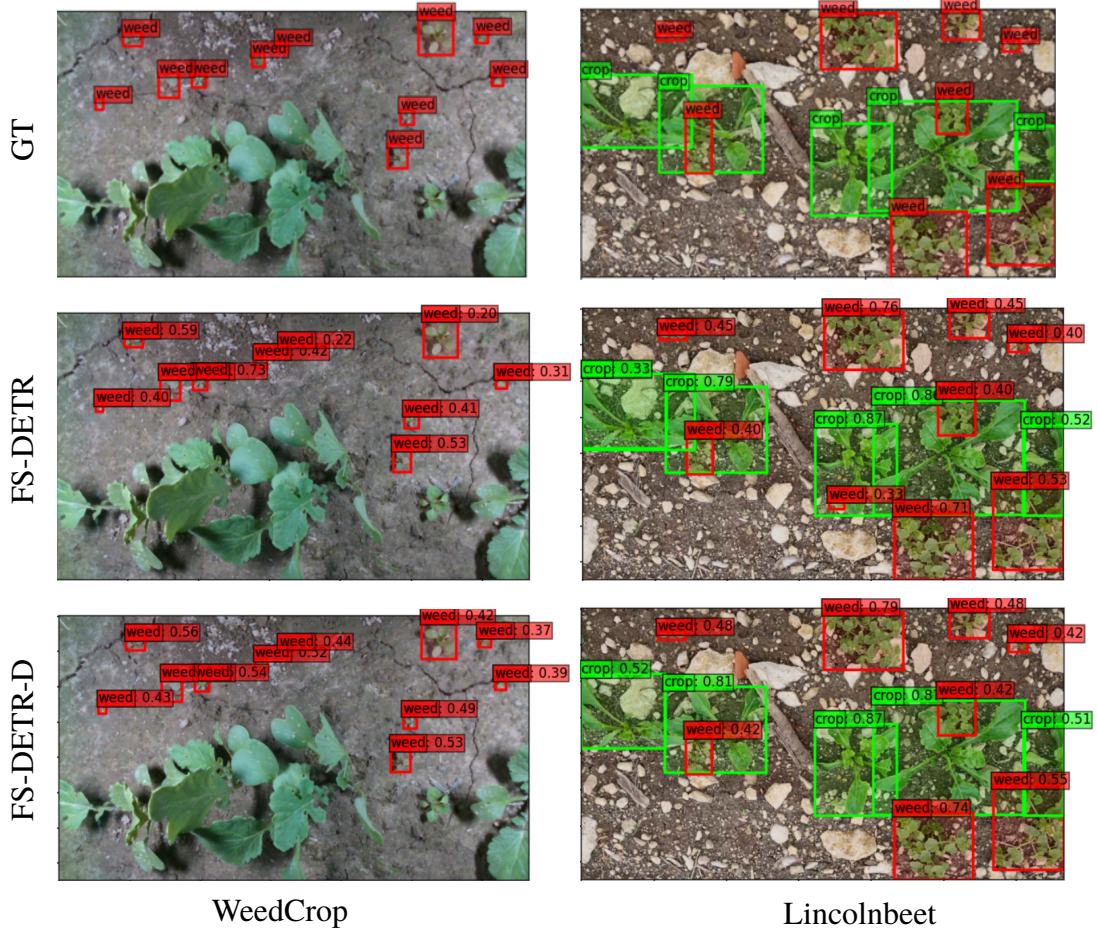


图 4.7 FS-DETR-D 与 FS-DETR 在不同数据集上的可视化对比

4.6.4 可视化结果分析

图4.7展示了第四章改进模型与基线 FS-DETR 在 WeedCrop 和 Lincolnbeet 数据集上的检测可视化结果。图中，GT 表示真实标注，不同颜色的边界框对应不同类别的目标，其中绿色表示作物，红色表示杂草，白色表示漏检目标。整体来看，两种模型在尺度较大的显著目标上均能取得较好的检测效果。然而，在杂草密集分布及小尺度目标占比较高的场景中，改进模型在目标识别与实例区分方面表现更为稳定，能够有效减少漏检及相邻目标混淆现象。此外，在作物与杂草高度重叠或背景纹理复杂的情况下，基线模型在局部密集区域仍存在漏检或误检问题，而改进模型在边界完整性与定位精度方面表现更优，与真实标注具有更高的一致性。

上述性能提升主要得益于第四章提出的解码阶段三项关键改进的协同作用。基于置信度熵与空间感知的动态查询分配机制能够根据场景复杂度自适应调整有效查询数量，使解码资源更加集中于目标密集区域，从而缓解固定查询数量带来的漏检问题。基于超边增强的高阶查询交互模块在保留自注意力建模能力的基础上，引入

多查询协同建模，有效强化了遮挡目标与相邻目标之间的结构一致性表达。匹配质量感知的自适应损失建模机制则通过动态调节监督强度，降低匹配噪声对解码分类学习的干扰，提高了复杂背景和低对比度条件下的检测稳定性。

4.6.5 跨数据集泛化性能验证

为进一步验证所提出模型在不同场景下的泛化能力，我们在通用目标检测数据集 PASCAL VOC 2007+2012^[84] 上进行了补充实验，比较了本文方法与多种先进 DETR 系列和 YOLO 系列方法的性能，结果如表4.10所示。实验中，本文方法在 AP 指标上取得了 68.8%，较 FS-DETR 基线模型的 67.7% 提升了 1.1%，同时在 AP₅₀ 和 AP₇₅ 上也分别达到了 88.7% 和 75.6%，显示出在大、中、小尺度目标上的稳健性。此外，对于小目标的检测精度达到 29.3%，相比 FS-DETR 提升 1.5%，进一步体现了模型在复杂场景下对细粒度目标的捕获能力，而中尺度和大尺度目标的 AP 分别为 54.7% 和 78.8%，均优于基线和其他对比方法。这些结果不仅证明了本文所提出的解码自适应优化机制在农业杂草检测中表现出的优势能够迁移至通用目标检测任务，而且也显示出该方法对不同数据分布、目标密度和尺度变化具有良好的适应性。

表 4.10 PASCAL VOC 2007+2012 数据集上本章方法与先进方法的性能对比

Method	Reference	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
YOLOv13-L ^[78]	2025arXiv	—	67.3	86.5	73.6	—	—	—
RT-DETR ^[79]	2024CVPR	ResNet50	64.4	84.3	70.1	26.2	50.1	74.1
RT-DETRv2 ^[80]	2024CVPR	ResNet50	65.4	85.0	70.8	26.5	51.6	75.0
DETR ^[19]	2020ECCV	ResNet50	42.6	65.7	47.2	9.6	29.9	52.7
Deformable-DETR ^[20]	2021ICLR	ResNet50	46.5	70.5	51.3	11.2	33.2	56.2
Co-DETR ^[35]	2023ICCV	ResNet50	64.3	83.9	70.5	24.7	50.5	74.2
DINO ^[30]	2023ICLR	ResNet50	66.6	86.0	72.8	25.4	52.5	76.6
DDQ ^[24]	2023CVPR	ResNet50	66.7	86.1	74.0	26.8	<u>54.0</u>	76.8
Align-DETR ^[82]	2024BMVC	ResNet50	62.1	82.0	67.3	21.8	47.4	72.2
H-DETR ^[33]	2023CVPR	ResNet50	56.9	75.6	62.2	19.2	42.4	67.0
DEIM ^[83]	2025CVPR	ResNet50	66.4	84.9	72.0	24.4	51.9	75.7
FS-DETR	—	ResNet50	<u>67.7</u>	<u>87.3</u>	<u>74.1</u>	<u>27.8</u>	53.4	<u>77.8</u>
FS-DETR-D	—	ResNet50	68.8	88.7	75.6	29.3	54.7	78.8

尽管本文方法在 PASCAL VOC 2007+2012 数据集上展现出了良好的泛化性能，但智慧农业场景中仍存在复杂的跨域分布差异。例如，不同采集环境下的光照条件、

土壤背景以及作物生长期变化,均会对模型特征表征与检测稳定性产生显著影响。为进一步验证模型在农业场景中的跨域泛化能力,本文补充进行了跨数据集验证实验,即在 WeedCrop 数据集上训练模型,并直接迁移至 LincolnBeet 数据集进行测试而不进行微调。如表4.11所示,在明显域偏移条件下,各方法检测性能均出现不同程度下降,说明复杂农业环境中的跨域适应仍具有较大挑战性。然而,相较于 DINO 与 FS-DETR,本文提出的 FS-DETR-D 在跨数据集测试中仍取得了更优的检测性能,其 AP、AP₅₀ 与 AP₇₅ 分别达到 49.1%、75.2% 和 53.6%。这表明所提出的双域特征建模与解码自适应优化机制能够有效缓解光照变化、背景差异及目标尺度变化带来的特征分布偏移问题,从而提升模型在复杂农业场景中的跨域泛化能力与鲁棒性。

表 4.11 跨数据集泛化性能比较 (WeedCrop→LincolnBeet, 无微调)

方法	训练集	测试集	AP	AP ₅₀	AP ₇₅
DINO	WeedCrop	LincolnBeet	46.5	72.1	50.2
FS-DETR	WeedCrop	LincolnBeet	47.8	73.6	51.4
FS-DETR-D	WeedCrop	LincolnBeet	49.1	75.2	53.6

4.7 本章小结

本章在第三章 FS-DETR 模型基础上构建解码增强框架 FS-DETR-D,从查询分配、查询交互和监督建模三个方面对解码过程进行优化。在查询分配方面,引入基于置信度熵与空间感知的动态查询分配机制,联合建模置信度分布与空间结构信息,自适应调整有效查询数量,使查询资源更聚焦于密集区域;在查询交互方面,提出超边增强的高阶查询交互机制,通过联合建模多个相关查询,增强密集目标场景下的特征表达能力;在监督策略方面,设计匹配质量自适应感知分类损失函数,根据匹配质量动态调节监督强度,提高解码阶段的训练稳定性。实验结果表明,改进后的模型 FS-DETR-D 在 WeedCrop、Lincolnbeet 和 MH-Weed16 数据集上均优于 FS-DETR,尤其在目标密集且分布不均匀、遮挡严重、背景复杂的场景中表现出显著提升。消融实验进一步验证了各模块的有效性及其互补性。

第五章 总结与展望

5.1 总结

复杂农业场景下的杂草目标检测是精准农业感知的重要基础。然而，在实际田间环境中，由于杂草与作物外观相似、目标密集且分布不均，并伴随严重遮挡，检测任务面临较大挑战，现有方法易出现漏检、误检及预测不稳定等问题。传统的以 CNN 为代表的检测方法受限于局部卷积运算，难以在保持高分辨率特征的同时有效捕捉作物行距、叶片形态等全局上下文信息。而近年来兴起的 Transformer 架构虽具备全局建模优势，但在复杂农田场景中的适用性仍受到多方面因素制约，具体表现为对频域判别信息利用不足、固定查询机制难以适应目标分布密度的动态变化，以及由目标遮挡与形态不规则所带来的匹配不确定性问题，从而影响解码过程的稳定性与最终检测性能。针对上述问题，本文基于 Transformer 提出改进的目标检测模型，从特征建模与解码优化两个方面进行设计，以提升复杂场景下的检测精度与稳定性。

在特征建模方面，本文构建了双域融合的 Transformer 杂草检测模型（Frequency-Spatial Domain Integrated Detection Transformer, FS-DETR），通过同时利用空间域结构信息与频域特征，提高模型在复杂背景下的特征表达能力。具体而言，提出混合特征融合模块，实现多尺度空间特征与频域信息的协同融合，从而增强模型对细粒度纹理和边缘信息的表达能力。同时设计双域注意力机制，结合频域注意力与可变形注意力，在突出目标响应的同时加强对局部关键区域的建模能力，使模型在杂草密集或背景干扰较强的场景中仍能够保持稳定的检测性能。此外，为提升训练阶段正负样本匹配的稳定性，本文提出基于高斯分布与约束引导的标签分配策略，通过引入中心约束与分布相似性度量，使标签分配更加合理，从而提高模型的整体检测精度。

在解码阶段，针对固定查询建模与静态监督机制难以适应复杂场景动态变化的问题，本文在 FS-DETR 基础上提出解码增强框架 FS-DETR-D（FS-DETR with Decoder Enhancement）。该框架从查询分配、特征交互与监督学习三个方面对解码过程进行协同优化。在解码输入构建方面，提出基于置信度熵与空间感知的动态查询分配机制，通过联合建模置信度分布与空间结构信息，自适应评估场景复杂度，并动态调整

参与解码的查询数量，从而在目标密集与稀疏场景中实现查询资源的合理分配。在查询特征交互方面，提出基于超边增强的高阶查询交互机制，通过超图建模查询之间的高阶关系，增强了解码阶段对目标间结构依赖的表达能力。在解码监督建模方面，设计了匹配质量自适应感知的分类损失函数，将匹配质量作为连续调节因子引入分类损失建模，使监督强度随匹配质量动态变化，有效提升了解码过程的稳定性与预测一致性。

在 WeedCrop、Lincolnbeet 和 MH-Weed16 等杂草数据集上的实验结果表明，本文提出的 FS-DETR 及其解码增强框架 FS-DETR-D 在检测精度和复杂场景适应能力方面均优于多种主流目标检测方法，尤其在小尺度杂草检测和密集分布场景下表现更加稳定。实验结果充分验证了双域特征建模与解码自适应机制在复杂环境下杂草检测任务中的有效性。

5.2 未来研究展望

基于本文在复杂场景杂草目标检测领域的研究成果，围绕双域特征建模与解码自适应优化等关键问题，未来研究仍具有广阔的拓展空间，可从以下几个方面进一步深入与完善。

（1）双域特征建模机制的轻量化与自适应融合

本文通过联合建模空间域结构信息与频域判别特征，有效提升了模型在目标密集和背景复杂场景下的特征判别能力。然而，双域特征建模在提升性能的同时也引入了一定的计算开销。未来研究可探索更加轻量化的双域建模方式，例如通过自适应频域选择、动态通道激活或特征共享机制，在保证检测精度的前提下降低计算复杂度，从而提升其在实际农业设备上的部署能力。

（2）动态查询分配机制的细粒度建模扩展

本文结合置信度熵与空间感知信息，实现了基于场景复杂度的查询数量动态分配，缓解了固定查询机制在复杂场景下的局限性。但当前方法仅基于整幅图像的整体复杂度进行建模，难以反映不同区域之间的差异。未来可考虑对图像不同区域分别建模复杂度，从而实现更加细致的查询分配。

（3）高阶查询结构建模的效率与泛化能力研究

本文通过超边增强的高阶查询交互机制刻画查询间的结构关系，有效提升了解

码阶段的特征表达一致性。未来研究可进一步探索更加高效的结构建模方法，例如层次化关系建模或稀疏结构建模，以减少计算冗余并提升模型在不同农业场景中的泛化能力。

（4）方法在实际应用场景中的高效部署与拓展验证

本文主要针对静态图像中的杂草检测任务展开研究。未来可将该方法拓展至农业机器人、无人机遥感监测以及视频连续检测等应用场景，并结合模型压缩、知识蒸馏和推理加速等技术，提高模型在实际农业生产环境中的实时性与应用价值。

综上所述，复杂田间场景下的杂草检测仍面临诸多挑战。未来通过在特征建模、动态查询机制以及模型效率优化等方面的持续研究，有望构建更加高效、鲁棒且具有良好的泛化能力的杂草检测模型，为智能农业和精准农业管理提供有力的技术支撑。

插图索引

图 1.1	复杂农田场景杂草目标检测面临的典型挑战示意图	2
图 1.2	本研究的主要研究内容框架	9
图 2.1	Faster R-CNN 整体框架结构图.....	12
图 2.2	YOLOv5 整体框架结构图	12
图 2.3	CornerNet 整体框架结构图.....	13
图 2.4	自注意力机制计算流程图	15
图 2.5	多头自注意力机制结构图	16
图 2.6	编码器和解码器结构图	17
图 2.7	DETR 网络结构图.....	18
图 2.8	FPN 多尺度特征融合结构图	19
图 2.9	傅里叶变换频域分解与重建示意图	21
图 2.10	普通图与超图结构对比图	22
图 2.11	WeedCrop 数据集示例图	23
图 2.12	Lincolnbeet 数据集示例图	24
图 2.13	MH-Weed16 数据集示例图	24
图 3.1	FS-DETR 整体架构图.....	28
图 3.2	混合特征融合模块结构图	29
图 3.3	双域注意力机制模块结构图	33
图 3.4	不同特征融合方法对比图	44
图 3.5	DINO 与 FS-DETR 的收敛曲线对比	47

图 3.6	不同模型在 WeedCrop 数据集上的可视化对比	49
图 3.7	不同模型在 Lincolnbeet 数据集上的可视化对比	50
图 3.8	不同模型在 MH-Weed16 数据集上的可视化对比	51
图 3.9	DINO 与 FS-DETR 在复杂场景下可视化对比	52
图 4.1	FS-DETR-D 整体架构图	55
图 4.2	超边自适应构建结构图	61
图 4.3	超边引导的高阶查询交互结构图	63
图 4.4	超边与自注意力特征融合结构图	64
图 4.5	IoU=0.1 条件下 AMAL 与 VFL 损失-置信度曲线对比	73
图 4.6	IoU=0.9 条件下 AMAL 与 VFL 损失-置信度曲线对比	74
图 4.7	FS-DETR-D 与 FS-DETR 在不同数据集上的可视化对比	75

表格索引

表 3.1	WeedCrop 数据集上本章方法与先进方法的性能对比	40
表 3.2	Lincolnbee 数据集上本章方法与先进方法的性能对比	41
表 3.3	MH-Weed16 数据集上本章方法与先进方法的性能对比	42
表 3.4	模块有效性的消融实验	43
表 3.5	混合特征融合模块的消融实验	45
表 3.6	双域注意力机制的消融实验	46
表 3.7	双域注意力机制中融合权重的消融实验	46
表 3.8	基于高斯分布与约束引导标签分配的消融实验	47
表 3.9	基于高斯分布与约束引导的标签分配中代价权重的消融实验	48
表 4.1	WeedCrop 数据集上本章方法与先进方法的性能对比	67
表 4.2	Lincolnbeet 数据集上本章方法与先进方法的性能对比	68
表 4.3	MH-Weed16 数据集上本章方法与先进方法的性能对比	69
表 4.4	模块有效性的消融实验	70
表 4.5	不同查询策略下的性能与效率对比	71
表 4.6	不同查询交互机制的消融实验	71
表 4.7	融合权重 λ 的消融实验	72
表 4.8	不同分类损失函数的消融实验	73
表 4.9	匹配质量敏感系数 $\gamma(q)$ 参数区间的消融实验	74
表 4.10	PASCAL VOC 2007+2012 数据集上本章方法与先进方法的性能对比	76
表 4.11	跨数据集泛化性能比较 (WeedCrop→LincolnBeet, 无微调)	77

参考文献

- [1] 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.
- [2] 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.
- [3] VIOLA P, JONES M J. Robust real-time face detection[J]. International journal of computer vision, 2004, 57(2): 137-154.
- [4] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05): Vol. 1. Ieee, 2005: 886-893.
- [5] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.
- [6] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014: 806-813.
- [7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
- [8] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.
- [9] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2021, 32(4): 1201-1227.

- [10] 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.
- [11] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
- [12] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [13] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
- [14] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016: 21-37.
- [15] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
- [16] REDMON J, FARHADI A. Yolov3: an incremental improvement[A/OL]. arXiv (2018). <https://arxiv.org/abs/1804.02767>.
- [17] ZHOU Q, LI H, CAI Z, et al. YOLO-ACE: enhancing YOLO with augmented contextual efficiency for precision cotton weed detection[J]. Sensors (Basel, Switzerland), 2025, 25(5): 1635.
- [18] ZHOU Q, WANG Z, ZHONG Y, et al. Efficient optimized yolov8 model with extended vision[J]. Sensors, 2024, 24(20): 6506.
- [19] 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.
- [20] ZHU X, SU W, LU L, et al. Deformable detr: deformable transformers for end-to-end object detection[A/OL]. arXiv (2020). <https://arxiv.org/abs/2010.04159>.
- [21] LIU S, LI F, ZHANG H, et al. Dab-detr: dynamic anchor boxes are better queries for detr[A/OL]. arXiv (2022). <https://arxiv.org/abs/2201.12329>.

- [22] MENG D, CHEN X, FAN Z, et al. Conditional detr for fast training convergence[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3651-3660.
- [23] CHEN X, WEI F, ZENG G, et al. Conditional detr v2: efficient detection transformer with box queries[A/OL]. arXiv (2022). <https://arxiv.org/abs/2207.08914>.
- [24] ZHANG S, WANG X, WANG J, et al. Dense distinct query for end-to-end object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7329-7338.
- [25] HOU X, LIU M, ZHANG S, et al. Saliencetr: enhancing detection transformer with hierarchical salience filtering refinement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 17574-17583.
- [26] WANG Y, ZHANG X, YANG T, et al. Anchor DETR: query design for transformer-based object detection[A/OL]. arXiv (2021). <https://arxiv.org/abs/2109.07107>.
- [27] DAI X, CHEN Y, YANG J, et al. Dynamic detr: end-to-end object detection with dynamic attention[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2988-2997.
- [28] YAO Z, AI J, LI B, et al. Efficient detr: improving end-to-end object detector with dense prior[A/OL]. arXiv (2021). <https://arxiv.org/abs/2104.01318>.
- [29] 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.
- [30] ZHANG H, LI F, LIU S, et al. Dino: detr with improved denoising anchor boxes for end-to-end object detection[A/OL]. arXiv (2022). <https://arxiv.org/abs/2203.03605>.
- [31] 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.
- [32] OUYANG-ZHANG J, CHO J H, ZHOU X, et al. Nms strikes back[A/OL]. arXiv (2022). <https://arxiv.org/abs/2212.06137>.

- [33] JIA D, YUAN Y, HE H, et al. Detrs with hybrid matching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 19702-19712.
- [34] 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.
- [35] 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.
- [36] LE V N T, APOPEI B, ALAMEH K. Effective plant discrimination based on the combination of local binary pattern operators and multiclass support vector machine methods[J]. Information processing in agriculture, 2019, 6(1): 116-131.
- [37] HAMUDA E, GLAVIN M, JONES E. A survey of image processing techniques for plant extraction and segmentation in the field[J]. Computers and electronics in agriculture, 2016, 125: 184-199.
- [38] SLAUGHTER D C, GILES D, DOWNEY D. Autonomous robotic weed control systems: a review[J]. Computers and electronics in agriculture, 2008, 61(1): 63-78.
- [39] WANG A, ZHANG W, WEI X. A review on weed detection using ground-based machine vision and image processing techniques[J]. Computers and electronics in agriculture, 2019, 158: 226-240.
- [40] SA I, CHEN Z, POPOVIĆ M, et al. weednet: dense semantic weed classification using multispectral images and mav for smart farming[J]. IEEE robotics and automation letters, 2017, 3(1): 588-595.
- [41] REEDHA R, DERICQUEBOURG E, CANALS R, et al. Transformer neural network for weed and crop classification of high resolution UAV images[J]. Remote sensing, 2022, 14(3): 592.
- [42] ZHOU Q, HUANG Z, ZHENG S, et al. A wheat spike detection method based on transformer[J]. Frontiers in Plant Science, 2022, 13: 1023924.

- [43] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] GHIASI G, LIN T Y, LE Q V. Nas-fpn: learning scalable feature pyramid architecture for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7036-7045.
- [48] LIANG Y, CHANGJIAN W, FANGZHAO L, et al. TFPN: twin feature pyramid networks for object detection[C]//2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019: 1702-1707.
- [49] DENG C, WANG M, LIU L, et al. Extended feature pyramid network for small object detection[J]. IEEE Transactions on Multimedia, 2021, 24: 1968-1979.
- [50] HU Y, CHEN Y, LI X, et al. Dynamic feature fusion for semantic edge detection [A/OL]. arXiv (2019). <https://arxiv.org/abs/1902.09104>.
- [51] YUN G, YOO J, KIM K, et al. Spanet: frequency-balancing token mixer using spectral pooling aggregation modulation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6113-6124.
- [52] HUANG J, LIU Y, ZHAO F, et al. Deep fourier-based exposure correction network with spatial-frequency interaction[C]//European conference on computer vision. Springer, 2022: 163-180.

- [53] WANG Y, YU K, CHEN C, et al. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7278-7287.
- [54] BAI J, YUAN L, XIA S T, et al. Improving vision transformers by revisiting high-frequency components[C]//European conference on computer vision. Springer, 2022: 1-18.
- [55] ANTELM I A, CORDASCO G, POLATO M, et al. A survey on hypergraph representation learning[J]. ACM Computing Surveys, 2023, 56(1): 1-38.
- [56] LEE G, BU F, ELIASSI-RAD T, et al. A survey on hypergraph mining: patterns, tools, and generators[J]. ACM Computing Surveys, 2025, 57(8): 1-36.
- [57] FU X, XIAO J, ZHU Y, et al. Continual image deraining with hypergraph convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9534-9551.
- [58] DI D, ZOU C, FENG Y, et al. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 5800-5815.
- [59] FAN H, ZHANG F, WEI Y, et al. Heterogeneous hypergraph variational autoencoder for link prediction[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(8): 4125-4138.
- [60] FENG Y, HUANG J, DU S, et al. Hyper-yolo: when visual object detection meets hypergraph computation[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 47(4): 2388-2401.
- [61] FENG Y, YOU H, ZHANG Z, et al. Hypergraph neural networks[C]//Proceedings of the AAAI conference on artificial intelligence: Vol. 33. 2019: 3558-3565.
- [62] GAO Y, FENG Y, JI S, et al. HGNN+: general hypergraph neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3181-3199.
- [63] SUDARS K, JASKO J, NAMATEVS I, et al. Dataset of annotated food crops and weed images for robotic computer vision control. data brief 31, 105833[Z]. 2020.

- [64] DARBYSHIRE M, SALAZAR-GOMEZ A, GAO J, et al. Towards practical object detection for weed spraying in precision agriculture[J]. *Frontiers in Plant Science*, 2023, 14: 1183277.
- [65] SHINDE S, ATTAR V. An indian annotated weed dataset for computer vision tasks in precision farming[J]. *Data in Brief*, 2025: 111691.
- [66] 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.
- [67] XU C, WANG J, YANG W, et al. RFLA: gaussian receptive field based label assignment for tiny object detection[C]//*European conference on computer vision*. Springer, 2022: 526-543.
- [68] ZHANG S, CHI C, YAO Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020: 9759-9768.
- [69] KIM K, LEE H S. Probabilistic anchor assignment with iou prediction for object detection[C]//*European conference on computer vision*. Springer, 2020: 355-371.
- [70] GE Z, LIU S, LI Z, et al. Ota: optimal transport assignment for object detection[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 303-312.
- [71] GE Z, LIU S, WANG F, et al. YoloX: exceeding yolo series in 2021[A/OL]. *arXiv* (2021). <https://arxiv.org/abs/2107.08430>.
- [72] LI X, WANG W, WU L, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[J]. *Advances in neural information processing systems*, 2020, 33: 21002-21012.
- [73] ZHANG H, WANG Y, DAYOUB F, et al. Varifocalnet: an iou-aware dense object detector[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 8514-8523.

- [74] FENG C, ZHONG Y, GAO Y, et al. Toood: task-aligned one-stage object detection [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2021: 3490-3499.
- [75] KINGMA D P, BA J. Adam: a method for stochastic optimization[A/OL]. arXiv (2014). <https://arxiv.org/abs/1412.6980>.
- [76] 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.
- [77] 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.
- [78] LEI M, LI S, WU Y, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception[A/OL]. arXiv (2025). <https://arxiv.org/abs/2506.17733>.
- [79] ZHAO Y, LV W, XU S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.
- [80] LV W, ZHAO Y, CHANG Q, et al. Rt-detr2: improved baseline with bag-of-freebies for real-time detection transformer[A/OL]. arXiv (2024). <https://arxiv.org/abs/2407.17140>.
- [81] WANG W, DAI J, CHEN Z, et al. Internimage: exploring large-scale vision foundation models with deformable convolutions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14408-14419.
- [82] CAI Z, LIU S, WANG G, et al. Align-detr: improving detr with simple iou-aware bce loss[A/OL]. arXiv (2023). <https://arxiv.org/abs/2304.07527>.
- [83] HUANG S, LU Z, CUN X, et al. Deim: detr with improved matching for fast convergence[C]//Proceedings of the computer vision and pattern recognition conference. 2025: 15162-15171.

- [84] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

攻读硕士学位期间取得的研究成果

一、论文

[1] Yan Sun, Xinyu Shen, Qiaochuan Chen, Yuexing Han. A dual-domain detection transformer for fine-grained weed detection in complex agricultural scenes. Information Sciences, Vol. 741, doi: <https://doi.org/10.1016/j.ins.2026.123243>, 2026. (中科院二区, 本人第二作者)

二、软件著作权

[1] 软件名称: 基于深度学习的杂草智能检测软件 V1.0, 开发人: 孙妍, 沈新宇, 韩越兴, 陈侨川, 登记号: 2025SR2154502, 登记日期: 2025.11.05, 申请人: 上海大学。(导师一作, 本人二作)

致 谢

时光荏苒，转眼间硕士阶段的学习已接近尾声。回首这段充实而难忘的求学旅程，我不仅在专业知识与科研能力方面得到了系统提升，更在不断探索与实践实现了自我成长，也更加懂得了责任与感恩的意义。

首先，我要衷心感谢我的导师孙妍老师。在论文选题、研究开展以及撰写修改的全过程中，孙妍老师始终给予我耐心细致的指导与悉心帮助。无论是在科研上的严格要求，还是在生活中的关心与鼓励，都让我倍感温暖、受益匪浅。老师严谨求实的治学态度和精益求精的科研精神，不仅为本论文的顺利完成奠定了坚实基础，也将长久影响我未来的学习与工作。同时，我要诚挚感谢课题组的导师韩越兴老师和陈侨川老师。感谢两位老师在科研思路、学术视野以及问题分析方面给予我的指导与启发，使我在研究过程中不断拓展思考深度、提升研究能力。

此外，感谢课题组各位同学在学习与科研中的帮助与支持。特别感谢马唯一同学，不仅是并肩前行的伙伴，也是我十分珍贵的朋友。在这段求学时光中，我们朝夕相处、相互陪伴，使科研之路多了一份温暖与坚定；也特别感谢黄志怡同学，作为课题组的大管家，在科研与日常事务中给予我极大帮助与支持，使各项工作得以更加顺利、高效地推进。同时，感谢同门刘斐然和王志梁同学。在共同学习与科研探索过程中，我们彼此鼓励、携手前行。刘斐然同学不仅是并肩奋斗的伙伴，也是我十分要好的朋友，在学习与生活中给予我诸多支持与陪伴；王志梁同学性格内敛沉稳，但在科研过程中始终给予我踏实而关键的帮助，使我受益匪浅。

我还要特别感谢我的室友于新宇、刘姿和付羽道同学。我们在研一时共同获得了“最美寝室”的称号，这不仅是一份荣誉，更是我们温馨相处的见证。感谢你们一路以来的陪伴与关心，让平凡的日子变得有声有色。那些一起聚餐时的欢声笑语、结伴出游时的轻松自在，构成了这段时光中最鲜活的记忆。在一次次相聚与分享中，我们彼此治愈、相互支持，也让我在紧张的科研生活之余得以放松心情、重新蓄力出发。这些闪闪发光的日常，将成为我心中最温暖而难忘的片段。

同时，衷心感谢我的家人和朋友们。感谢你们一路以来的理解、包容与陪伴，在我迷茫时给予指引，在我疲惫时给予安慰，在我前行时默默守望。你们的支持如细

水长流，温柔而坚定，始终给予我前进的勇气与力量，让我能够心怀从容，坚定地走到今天。

最后，我也要感谢一路坚持努力的自己。在面对科研压力与挑战时始终未曾放弃，让这段旅程得以完整而有意义地完成。

谨以此文，向所有在我硕士阶段给予关心、帮助与支持的老师、同学、家人和朋友们致以最诚挚的感谢！

沈新宇

上海

2026 年 4 月 16 日