

中图分类号: TP391

单位代号: 10280

密 级: 公开

学 号: 23721520

上海大学



硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	面向微小目标的 实时目标检测方法研究
--------	-----------------------

作 者 徐天洋

学科专业 计算机应用技术

导 师 陈侨川

完成日期 二〇二六年四月

姓 名：徐天洋

学号：23721520

论文题目：面向微小目标的实时目标检测方法研究

上海大学

本论文经答辩委员会全体委员审查，确认
符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主 席：

7 薛

委 员：

张瑞

王华

导 师：

陈伟

答辩日期：

2026 年 6 月 9 日

姓名：徐天洋

学号：23721520

论文题目：面向微小目标的实时目标检测方法研究

上海大学学位论文原创性声明

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

学位论文作者签名：徐天洋

日期：2026年6月10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：徐天洋

导师签名：陈钢

日期：2026年6月10日

日期：2026年6月10日

上海大学工学硕士学位论文

面向微小目标的
实时目标检测方法研究

作者: 徐天洋

导师: 陈侨川

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 Real-time Object Detection Methods for Tiny Objects

Candidate: Tianyang Xu

Supervisor: Qiaochuan Chen

Major: Technology of Computer Application

School of Computer Engineering and Science

Shanghai University

April, 2026

摘 要

在当今智能化时代，微小目标检测已成为智能农业、自动驾驶、无人机侦察及医疗影像分析等领域的关键技术之一。微小目标泛指图像中尺寸小于 16×16 像素以下的目标，这些目标在图像中像素占比极低且缺乏辨识性特征，导致传统深度学习模型面临三大严峻挑战：一是深度卷积网络下采样过程中微小目标特征信息流失严重；二是现有标签分配策略难以在极高信噪比环境下为小尺度样本提供充足且高质量的正样本；三是处理高分辨率特征图以捕捉细节时，巨大的计算开销与显存占用限制了算法在边缘设备上的实时部署。针对上述痛点，本文深入探究了深度学习模型的视觉表征机制，分别从提升检测精度与优化推理效率两个维度出发，提出了两种互补的微小目标检测方案：

(1) 提出了一种基于隐式特征融合与混合自适应标签分配的高精度目标检测方法。针对特征融合中的细节丢失问题，隐式特征融合利用隐式神经表示将多尺度特征映射至统一分辨率进行融合，有效保留了微小目标的精细特征。针对微小目标正样本质量与数量难以兼顾的缺陷，混合自适应标签分配策略结合 IoU 与感受野距离构建混合度量得分，并利用自适应机制精准挖掘高质量正样本。

(2) 提出了一种基于选择稀疏编码器与尺度感知 Query 选择的实时目标检测方法。针对基于 Transformer 的检测架构处理高分辨率特征图计算开销过大的问题，选择稀疏编码器利用特征选择机制与线性自注意力，大幅降低了显存占用与计算复杂度。为了解决微小目标特征在 Query 筛选阶段的置信度劣势，尺度感知 Query 选择机制通过引入基于特征下采样步长的折损系数，动态提升了模型对微小目标的关注度。

实验结果表明，本文提出的方法在 AI-TOD、SODA-D、VisDrone 及 AgriPest 等多个挑战性数据集上均表现优越。两种方法分别在微小目标检测精度和模型推理速度上取得了显著突破，适用于多种微小目标检测场景。本研究不仅为微小目标检测方法的优化提供了新思路，也为微小目标检测技术在资源受限的边缘终端上的实时部署提供了坚实的理论支撑与实践方案。

关键词：微小目标检测；特征融合；标签分配；实时目标检测

ABSTRACT

In the current era of artificial intelligence, tiny object detection has become a key technology in fields such as smart agriculture, autonomous driving, unmanned aerial vehicles (UAVs), and medical image analysis. Tiny objects occupy an extremely low proportion of pixels in images and lack discriminative features, causing traditional deep learning models to face three severe challenges: first, the severe loss of feature information for tiny objects during the downsampling process of deep convolutional networks; second, the difficulty for existing label assignment strategies to provide sufficient, high-quality positive samples for small-scale instances in environments with an extremely high signal-to-noise ratio; third, the massive computational overhead and memory consumption required to process high-resolution feature maps for detail capture, which limits the real-time deployment of algorithms on edge devices. To address these pain points, this paper deeply explores the visual representation mechanisms of deep learning models. From the dual perspectives of enhancing detection accuracy and optimizing inference efficiency, two complementary tiny object detection schemes are proposed:

(1) A high-accuracy object detection method based on implicit feature fusion and hybrid adaptive label assignment. To tackle the problem of detail loss during feature fusion, the implicit feature fusion employs implicit neural representations to map multi-scale features to a unified resolution for fusion, effectively preserving the fine-grained features of tiny objects. To overcome the difficulty in balancing the quality and quantity of positive samples for tiny objects, the hybrid adaptive label assignment strategy integrates Intersection over Union (IoU) and receptive field distance to construct a hybrid metric score, leveraging an adaptive mechanism to precisely mine high-quality positive samples.

(2) A real-time object detection method based on a selective sparse encoder and scale-aware Query selection. To address the excessive computational overhead of Transformer-based detection architectures when processing high-resolution feature maps, the selective sparse encoder utilizes a feature selection mechanism and linear self-attention, substan-

tially reducing memory consumption and computational complexity. To resolve the confidence disadvantage of tiny object features during the Query selection phase, a scale-aware Query selection mechanism introduces a decay coefficient based on the feature downsampling stride, dynamically elevating the model's attention to tiny objects. Furthermore, a one-to-many matching auxiliary detection head is introduced to strengthen the supervision signals for tiny objects during training to accelerate convergence. This auxiliary head is seamlessly detached during inference to maintain an end-to-end architecture.

Experimental results demonstrate that the methods proposed in this paper exhibit superior performance across multiple challenging datasets, including AI-TOD, SODA-D, Vis-Drone, and AgriPest. The two methods achieve significant breakthroughs in tiny object detection accuracy and model inference speed, respectively, making them suitable for diverse tiny object detection scenarios. This research not only offers novel perspectives for the optimization of tiny object detection methods but also provides solid theoretical support and practical solutions for the real-time deployment of tiny object detection technologies on resource-constrained edge terminals.

Keywords: Tiny Object Detection; Feature Fusion; Label Assignment; Real-time Object Detection

目 录

摘 要	I
ABSTRACT	II
第一章 绪论	1
1.1 课题来源	1
1.2 课题背景概述	1
1.3 课题研究目的与意义	2
1.4 国内外研究现状	3
1.4.1 微小目标检测研究现状	3
1.4.2 多尺度特征融合研究现状	4
1.4.3 目标检测标签分配研究现状	5
1.4.4 实时目标检测研究现状	6
1.5 论文主要工作	7
1.6 论文组织结构	8
第二章 相关理论和技术概述	10
2.1 视觉特征提取与融合	10
2.1.1 主干网络特征提取与残差学习	10
2.1.2 多尺度特征融合与特征金字塔	11
2.2 目标检测的标签分配	13
2.2.1 基于锚框的标签分配	13
2.2.2 无锚框与动态标签分配	14
2.2.3 DETR 的一对一二分图匹配	15
2.3 经典目标检测方法	16
2.3.1 双阶段目标检测模型 Faster R-CNN	16
2.3.2 单阶段目标检测模型 YOLOv8	18
2.3.3 端到端目标检测模型 RT-DETR	19
2.4 评价指标	21

2.4.1	精度评价指标	21
2.4.2	速度与计算复杂度评价指标	22
2.5	本章小结	23
第三章 基于隐式特征融合与混合自适应标签分配的		
	目标检测	24
3.1	方法概述	24
3.1.1	隐式特征融合 IFF	25
3.1.2	混合自适应标签分配策略 HALA	27
3.2	实验分析	31
3.2.1	实验设置	31
3.2.2	对比实验	32
3.2.3	消融实验	37
3.3	本章小结	45
第四章 基于稀疏编码器与尺度感知 Query 选择的实时微小目标检测		
4.1	方法概述	46
4.1.1	选择稀疏编码器 SSE	47
4.1.2	尺度感知 Query 选择策略 SQS	49
4.1.3	一对多匹配辅助检测头 O2MAH	50
4.2	实验分析	53
4.2.1	实验设置	53
4.2.2	对比实验	53
4.2.3	消融实验	55
4.2.4	实时检测实验	62
4.3	本章小结	63
第五章 总结与展望		
5.1	结论	64
5.2	工作展望	65
插图索引		67
表格索引		69

参考文献	70
攻读硕士学位期间取得的研究成果	81
致 谢	82

第一章 绪论

1.1 课题来源

本课题得到国家自然科学基金青年基金（编号：62401352），上海大学硅酸盐文物保护教育部重点实验室开放课题项目（编号：SCRC2024ZZ07TS）的资助。

1.2 课题背景概述

目标检测是计算机视觉中的基础任务之一，旨在从图像或视频序列中识别出感兴趣对象的位置与类别。随着成像设备分辨率与视场范围的提升，图像中可被采集到的目标尺度范围显著扩展，其中尺度极小、像素占比极低的目标对现有检测方法提出了新的挑战，由此形成了微小目标检测这一具有代表性的研究问题。

目前，关于“小目标”与“微小目标”，目前学术界尚未形成统一的尺度阈值定义，常见做法是依据数据集或具体任务场景给出相对标准。MS COCO^[1]数据集将像素面积小于或等于 32×32 像素的对象界定为小目标；而在航空遥感、低空无人机成像等更高分辨率的任务中，定义则更为严格，例如 AI-TOD^[2]数据集将尺寸小于 16×16 像素的目标界定为微小目标。在该尺度下，目标在图像中通常仅表现为弱点状信号，缺乏可辨识的纹理、颜色与拓扑结构。为表述统一，本文若无特别说明，统一采用“微小目标”一词，泛指上述以 AI-TOD 为代表、尺寸通常在 16×16 像素以下的目标；在涉及具体数据集（如 VisDrone^[3]）时，则保留其原始“小目标”定义。

微小目标检测在多个应用场景中具有现实需求。在植保预警系统中，针对尺度极小且特征隐蔽的早期病斑及迁飞性害虫，高精度的检测算法能够突破环境噪声干扰，实现灾情的早期发现与精准施药；在反无人机预警系统中，需要在复杂背景下对低空慢速小型无人机进行快速定位与跟踪；在自动驾驶感知中，远距离的行人、交通信号灯和路面障碍物在车载传感器中常以微小像素簇的形式出现，对其稳定检测有助于提升感知系统的可靠性；在医学影像分析中，早期肿瘤病灶或钙化点通常仅占图像的极小区域，对其进行准确检测对提升计算机辅助诊断性能具有积极意义。

然而，微小目标检测面临着严峻的技术瓶颈。首先是信息丢失问题，目标检测方法常用的卷积神经网络通过层层下采样获取高层语义特征，但这一过程伴随着特

征图分辨率的指数级下降，导致微小目标的像素信号在深层特征中几乎被抹除。其次是背景干扰与信噪比低，微小目标极易与背景中的噪声、复杂纹理混淆，甚至在成像过程中受到光照、遮挡和运动模糊的影响而发生畸变。最后，实时性要求与算力限制的矛盾日益凸显。在无人机机载设备或嵌入式监控终端上，微小目标检测算法必须在毫秒级内完成推理，而处理高清视频流所带来的计算负担往往超出了边缘设备的算力上限。

综合分析上述技术瓶颈，目前的研究亟需从以下三个维度寻求突破。在特征表征层面，应探索能够有效保留微小目标信息的特征融合方法，以在不同尺度间实现更精细的特征对齐与细节恢复；在模型训练层面，应结合微小目标特性，为微小目标提供更高质量的正样本监督信号；在计算效率层面，应针对高分辨率图像带来的计算负担，设计轻量化方法优化微小目标特征，以实现精度与实时性的平衡。本文正是基于上述思路，分别从优化微小目标检测精度与速度两个方向出发，为解决微小目标检测难题提供新的技术路径。

1.3 课题研究目的与意义

本文以微小目标检测为研究对象，针对其在低分辨率成像、低信噪比背景与有限算力条件下的检测难题，研究面向微小目标的特征表征、标签分配与高效推理方法，旨在缓解现有检测器在该类目标上精度与效率难以兼顾的问题。本文的研究目的与意义主要体现在以下三个方面：

第一，优化多尺度特征融合机制。传统的特征金字塔在融合深浅层特征时，往往无法兼顾浅层的细节与深层的语义。本文通过研究如何在高分辨率特征图上保持微小目标的判别力，能够推动视觉表征学习中精细特征建模的发展。

第二，优化目标检测标签分配策略。微小目标在定位过程中由于其边界框对微小位移的极端敏感，导致训练过程中的正样本数量极少，引发梯度震荡与收敛困难。通过引入混合指标评估、动态分配等策略，有助于改善微小目标场景下检测器的正样本质量与训练稳定性。

第三，突破微小目标检测的实时性计算瓶颈。现有的高精度检测模型通常依赖复杂的网络结构，伴随着庞大的计算开销与较高的推理延迟，难以满足无人机侦察、自动驾驶等应用场景对高实时性的严苛要求。本文针对高分辨率特征图带来的显存

与计算开销问题, 研究面向微小目标的稀疏化计算与轻量化结构设计, 有助于提升该类方法在资源受限的边缘设备上的可行部署。

1.4 国内外研究现状

围绕微小目标检测, 本节按照四个维度梳理代表性研究工作。第一, 面向微小目标的检测方法整体研究现状, 包括训练样本扩充、特征增强与聚焦检测等思路; 第二, 多尺度特征融合方法研究现状; 第三, 目标检测标签分配方法研究现状; 第四, 实时目标检测方法研究现状。其中后三个维度对应本文后续章节中分别关注的特征表征、标签分配与高效推理三个核心问题。

1.4.1 微小目标检测研究现状

微小目标检测的相关工作主要集中在面向样本优化、特征表示优化和聚焦检测等方面。在面向样本优化方面, 为了应对微小目标训练样本不足的问题, 一些研究人员开发了数据增强技术。例如, **DS-GAN**^[4] 是一种数据增强流程, 旨在通过添加高质量的微小目标合成样本来提升检测性能。**Poseidon**^[5] 则利用训练集中已有的实例, 并结合图像元数据生成一致性的样本, 以解决类别不平衡问题。另一种面向样本方法是直接优化标签分配策略, 为微小目标分配更多正样本以用于训练。**DotD**^[6] 是一种新的评估指标, 可以作为 **IoU** 的替代方案, 为微小目标检测带来优势。

在特征表示优化方面, 超分辨率思路被广泛用于提升微小目标的可辨识度。**EFPN**^[7] 利用转置卷积获得高分辨率特征图来提升检测精度。**Bashir** 等人^[8] 应用生成对抗网络对低分辨率遥感图像进行超分辨率重建, 并结合残差特征聚合来捕捉复杂特征。在 **ALPD-Net**^[9] 中, 通过两阶段解码过程逐步融合多模态特征以生成高质量的显著性图, 旨在准确检测显著目标。在检测微小目标时, 上下文信息有时比目标特征本身能提供更多的决策支持, 因此上下文信息建模也被广泛应用于特征表示优化。**VSS-Net**^[10] 结合了基于图的视觉上下文表示和语义编码器, 以获取更优的特征表示。对于上下文建模方法来说, 考虑到原始的 **RoI** 池化操作会破坏小物体的结构特征, **SINet**^[11] 引入了一个上下文感知的 **RoI** 池化来维持上下文信息。**Cui** 等人^[12] 设计了一种上下文感知块, 用于将多尺度上下文信息与金字塔膨胀卷积组合在一起, 赋予高分辨率特征更多语义信息。**Kiobya** 等人^[13] 使用多尺度语义信息增强模块来丰富特征金字塔网络最低预测层的特征, 旨在提高小目标的检测精度。**SemDiff**^[14] 提出

了一种语义差异化方法,该方法通过显式融合高级语义信息来增强特征判别力,从而更有效地将微小目标从复杂的背景中剥离出来。**CFANet**^[15] 通过有效聚合特征取代了传统的逐层特征融合,这种方法将空间细节与语义信息相结合,实现了准确且高效的目标检测。**FFAGRNet**^[16] 引入了一种分组特征重建机制,通过对特征图进行分组和重新加权,以较低的计算成本增强了关键特征,同时也提高了检测精度。

在聚焦检测方面,由于高分辨率图像中微小目标分布往往不均匀,常规检测方法会在大量无目标区域上消耗冗余计算。聚焦检测思路通过先粗定位、再精细化检测的方式提升整体效率^[17]。**ClusDet**^[18] 利用物体之间的语义和空间信息生成聚类区域,然后将其裁剪出来以进行后续的精细检测。**QueryDet**^[19] 设计了级联查询策略,在低分辨率特征上预测小物体的粗略位置,然后使其引导高分辨率特征计算准确的检测结果。**F&D**^[20] 引入了 **Focus & Detect** 框架,其中 **Focus** 网络检测候选区域,然后将其裁剪并调整到更高分辨率,从而实现对小物体的精确检测。**Zhang** 等人引入了 **SS-Net**^[21] 以消除多尺度融合中的冗余,从而实现了高效且准确的预测。

总体来看,上述方法在微小目标检测精度上取得了一定进展,但仍存在局限性。超分辨率类方法引入了额外的生成计算开销,且生成细节可能产生伪影;上下文增强类方法依赖较深的特征融合结构,进一步加剧了计算负担;聚焦检测类方法虽然提升了局部精度,但裁剪与多次推理在一定程度上破坏了图像的全局空间一致性,易造成微小目标的漏检。

1.4.2 多尺度特征融合研究现状

多尺度特征融合是缓解微小目标信息丢失的常用手段。**Zhang** 等人^[22] 提出了一种从粗到细的检测框架,并设计细粒度特征对齐模块以自适应学习细粒度像素变换,缓解特征融合时的空间错位问题。**Quan** 等人^[23] 提出全局显式集中调节特征金字塔,在捕获全局长程依赖关系的同时获得更具区分性的特征表示。**MENet**^[24] 利用门控单元和高效注意力机制选择性地聚合微小目标的上下文信息,并抑制无关背景噪声。**Ge** 等人^[25] 提出 **CAFR** 模块,利用交叉注意力机制建立多尺度特征之间的全局相关性,对上采样后的深层特征进行微调,缓解直接上采样所导致的特征错位。**DN-FPN**^[26], 通过对比学习抑制通道压缩带来的特征噪声,并保留来自横向特征的几何信息和高层特征的语义信息。**MSFF-Net**^[27] 在多尺度上分别提取空间特征与时间特征并加以融合;**ADSFNet**^[28] 利用自适应差异化模块对特征进行融合与区分;**Zhang** 等人^[29] 提出

显著性上下文增强模块，自适应整合多尺度上下文信息，强调信息量丰富的特征同时抑制背景干扰。

从上述工作可见，现有多尺度特征融合方法主要从缓解空间错位和增强上下文信息两个角度展开。然而，这些方法在面向尺寸更小、像素更稀疏的微小目标时仍存在一定不足。单纯基于线性插值的上采样难以恢复因下采样而损失的高频细节；基于注意力的聚合通常假设各尺度特征已对齐到相同的离散网格，缺乏对任意尺度之间连续映射关系的建模能力。如何在融合阶段更充分地保留微小目标的细节信号，仍有进一步研究的空间。

1.4.3 目标检测标签分配研究现状

正负训练样本的分配方式对检测器性能具有重要影响。在通用目标检测任务中，已有多种改进的标签分配策略。**FreeAnchor**^[30] 基于面向目标检测设计的似然函数确定正样本锚框；**ATSS**^[31] 根据样本的统计特性自适应调整正负样本分配；**PAA**^[32] 使用高斯混合模型对锚框得分分布建模并据此区分正负样本；**VFNet**^[33] 为训练样本分配软标签，以减少二值标签带来的噪声。

针对微小目标，研究者进一步设计了专用的标签分配策略。**NWD**^[34] 基于高斯 Wasserstein 距离使用固定阈值分配训练样本；**RFLA**^[35] 直接度量高斯感受野与真实值的相似性，并使用 **Top-K** 策略实现微小目标的样本数量平衡；**Cof-Net** 中的 **CoF-SA** 模块^[22] 通过在训练过程中逐步引入几何和分类约束，从粗到细动态选择样本；**Zhang** 等人^[24] 提出了基于中心区域的标签分配方法，将位于中心区域内的锚框视为正样本，从而为微小目标提供了更多正样本；**KLDet**^[36] 结合定位和分类信息选择高质量的正样本，而不仅仅依赖于中心点距离或 **IoU** 指标。**Wang** 等人^[37] 提出了一种自动标签分配框架，该框架集成了实例属性模块和目标性预测模块，以实现外观感知和质量引导的标签分配。

从上述工作可以看出，针对微小目标的标签分配方法在度量准则与正样本数量上已有显著改进。其中 **NWD**、**RFLA**、**CoF-SA** 等方法在度量维度或阈值机制上各有侧重，**KLDet** 和 **Wang** 等人的工作进一步引入了多维度信息或学习式分配。然而，多数方法仍以预设阈值为主或仅利用单一类型的度量进行匹配，对训练过程中模型自身学习状态的动态变化考虑相对有限；同时，部分以扩充正样本为主要目标的策略在弱信号区域容易引入低质量样本，从而对最终的定位精度与分类置信度产生负面

影响。

1.4.4 实时目标检测研究现状

在实时目标检测的发展历程中，基于卷积神经网络的单阶段检测框架始终占据着绝对主导地位。但在感受野固定和网格分辨率粗糙的限制下，早期框架对小尺度和密集目标的检测能力较弱。为了解决这一缺陷，研究界开始探索利用不同层级的特征图来分别负责不同尺度目标的检测任务。Liu 等人提出的 SSD^[38] 利用不同尺度的特征图预测固定默认边界框的类别得分和坐标偏移量，以极简单的端到端训练实现了高精度的单次前向推断。Redmon 和 Farhadi 提出的 YOLOv3^[39] 利用多尺度预测架构并引入了强大的 Darknet-53 主干网络，通过上采样与特征拼接机制在速度与精度之间实现了当时的最佳平衡，确立了单阶段检测器在未来数年内的基本设计规范和评测基准。Wang 等人提出的 YOLOv7^[40] 通过创新的动态标签分配策略和重参数化模块设计，在不增加推理计算成本的前提下显著增强了网络的特征表达能力。Wang 等人提出的 YOLOv10^[41] 引入了一致性双重分配策略，成功在 YOLO 框架中实现了完全无非极大值抑制的端到端训练。为了在实时推理的极致速度下榨取注意力机制的全部理论潜力，Tian 等人提出的 YOLOv12^[42] 是一种以注意力机制为核心的颠覆性框架，其引入的区域注意力和残差高效层聚合网络有效化解了纯注意力模型计算复杂度极高的问题。Liu 等人提出的 ESOD^[43] 利用了一个基于自适应缩放和特征增强的轻量化框架，通过在保持高分辨率空间信息的同时，减少冗余计算来平衡小目标检测的精度与实时性需求。

在传统 CNN 架构不断优化的同时，基于 Transformer 体系的检测模型（如 DETR 及其变体）也逐渐展现出其精度优势。然而，标准 Transformer 对高分辨率多尺度特征图的处理会产生巨大的计算开销，因为自注意力的计算复杂度随输入序列长度呈二次方剧烈增长，导致早期 DETR 模型完全无法满足实时性需求。Lv 等人提出的 RT-DETR^[44] 是首个真正意义上的实时端到端目标检测 Transformer，其通过设计创新的混合编码器，将多尺度特征的尺度内交互与跨尺度融合进行物理层面的彻底解耦，大幅降低了计算冗余。Roboflow 团队基于强大的 DINOv2^[45] 视觉 Transformer 主干网络开发了 RF-DETR^[46]，证明了参数量庞大的基础大模型完全可以通过合理的架构裁剪与蒸馏，无缝应用至实时推理任务中。Zhang 等人提出的 SO-DETR^[47] 通过双域（频域与时域）特征融合增强小目标表征，并结合知识蒸馏技术，在保持实时检

测速度的同时显著提升了对极小目标的捕捉能力。

总体来看, 现有实时检测方法在常规尺度目标上已较为成熟, 但在面向微小目标时精度与速度之间的平衡较为困难。基于 CNN 的 YOLO 系列为追求速度通常减少对高分辨率特征的使用, 导致微小目标的细节信号被进一步压缩; 基于 Transformer 的实时检测器虽然引入了较强的全局建模能力, 但在高分辨率特征图上的计算开销仍然较大。如何在显著降低高分辨率特征带来冗余计算的同时, 保留对微小目标的有效监督信号, 仍是当前实时微小目标检测研究中亟待解决的问题。

1.5 论文主要工作

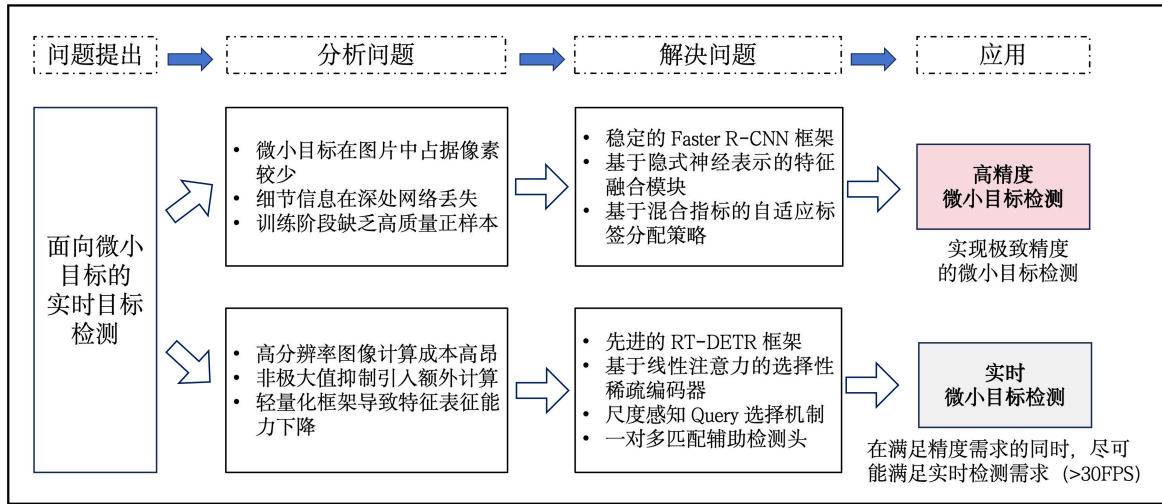


图 1.1 本文的研究路线示意图

针对上述微小目标检测在特征表征、标签分配与实时推理方面的不足, 本文围绕高精度与实时性两条主线开展研究, 提出两种检测方法, 技术路线如图1.1所示。本文的主要工作和创新点如下:

(1) 提出一种基于隐式特征融合和混合自适应标签分配的微小目标检测方法, 旨在应对微小目标信息丢失与训练阶段缺乏高质量正样本的问题。隐式特征融合通过利用隐式神经表示, 在特征融合前将来自不同金字塔层级的特征图映射至统一尺度, 从而缓解了多尺度融合中的特征对齐偏差。通过将特征图建模为连续表示, 隐式特征融合能够在任何分辨率下实现有效融合, 在保留微小目标细节的同时减少了信息损失。混合自适应标签分配策略则通过将 IoU 与感受野距离相结合, 改进了标签分

配策略,使其在处理微小目标时表现更佳。随后,利用一种自适应选择策略来识别高质量训练样本,增强了模型的训练效果和检测性能。在 AI-TOD、SODA-D、VisDrone 和 AgriPest 数据集上的广泛实验表明,该方法在遥感、自动驾驶、低空无人机成像和农业影像等多种微小目标检测场景中均表现优异。

(2) 提出了一种高效的实时微小目标检测方法,旨在应对微小目标检测中高分辨率特征图导致的计算开销过大问题,平衡检测精度与推理效率。首先,选择性稀疏编码器利用特征选择机制与线性自注意力,剔除冗余信息并筛选显著特征,大幅降低了显存占用与计算量,使模型能在资源受限设备上处理更高分辨率的特征;其次,尺度感知 Query 选择在不确定性筛选的基础上引入尺度折损系数,针对性地提升了模型对微小目标的关注度与定位精度;最后,一对多匹配辅助检测头通过引入基于锚框的监督机制,增强了训练阶段的信号强度与稳定性。实验证明,该方法在保持微小目标检测精度的同时,极大地降低了显存占用并实现了较快的推理速度。

1.6 论文组织结构

针对复杂背景下微小目标特征提取困难、定位精度低以及实时性要求高等挑战,本文深入研究了基于深度学习的微小目标检测方法,提出了隐式特征融合、混合自适应标签分配及选择性稀疏编码器等改进策略,旨在提升检测性能并兼顾推理效率。论文具体结构安排如下:

第二章介绍了本文的相关理论与技术基础。首先,对目标检测的相关理论进行了概述,包括视觉特征提取与融合、目标检测的标签分配方法以及本文方法涉及到的经典目标检测模型;其次,详细阐述了用于衡量检测性能的精度评价指标、速度与计算复杂度评价指标,为后续改进方案的性能评估提供依据。

第三章提出了基于隐式特征融合与混合自适应标签分配的目标检测方法。该章首先分析了微小目标特征丢失的问题,阐述了隐式特征融合模块的设计原理,以增强对微小目标的特征表达能力;接着,介绍了混合自适应标签分配策略,以优化微小目标的正样本质量。最后,通过在 AI-TOD、SODA-D、VisDrone 和 AgriPest 等多个典型微小目标数据集上的对比实验与详尽的消融实验,验证了所提方法在提升微小目标检测精度方面的有效性。

第四章提出了基于选择性稀疏编码器与尺度感知 Query 选择的实时微小目标检

测方法。该章聚焦于提升模型的推理效率与实时性能，首先分析了微小目标图片包含大量冗余信息的问题，构建了高效的选择性稀疏编码器；同时，阐述了尺度感知 Query 选择策略与一对多匹配辅助检测头的设计，以进一步优化检测效果。通过在多个数据集上的对比试验和消融实验数据集上的实验，证明了该方法在保持高性能的同时，具有优异的实时检测能力。

第五章对全文工作进行总结，归纳本文在微小目标检测精度提升与推理效率优化两方面的主要研究成果，分析现有方法存在的局限，并对后续可能的研究方向进行展望。

第二章 相关理论和技术概述

本章旨在为后续章节中提出的实时微小目标检测方法方法创新与实验验证提供必要的背景知识。首先剖析主干网络与特征金字塔在微小目标场景下的有效性边界；其次给出标签分配的形式化定义并梳理三种主流方法，揭示其对小目标监督信号的影响；进而以 Faster R-CNN、YOLOv8 与 RT-DETR 为代表，分别讨论双阶段、单阶段与端到端范式在精度—速度权衡下对微小目标的适配性与固有局限；最后给出本文实验所采用的评价指标体系。

2.1 视觉特征提取与融合

目标检测任务的核心在于从复杂的视觉背景中高效、准确地提取出目标的特征表示。现代目标检测模型通常将这一基础过程解耦为用于提取多层级特征的主干网络，以及用于整合多尺度特征的颈部网络两部分。本节将以 ResNet 和特征金字塔网络 FPN 为代表，探讨视觉特征的提取与融合机制，这也是后续理解并目标检测算法的重要理论基石。

2.1.1 主干网络特征提取与残差学习

主干网络负责从原始输入图像中逐层提取特征，其特征的表达能力直接决定了检测器的性能上限。在早期的卷积神经网络中，研究人员通常认为增加网络的深度可以获取更丰富的全局上下文语义信息。然而，随着网络层数的不断加深，模型在训练过程中容易遭遇梯度消失或梯度爆炸现象，导致网络出现退化，即深层网络的训练误差反而高于浅层网络。

为了解决深度网络的退化问题，ResNet^[48]提出了一种基于跨层恒等映射的残差网络。ResNet 摒弃了让网络层直接拟合底层映射的传统做法，转而让网络去拟合残差映射，巧妙地解决了梯度消失的问题，残差模块是残差网络的重要组成部分，残差模块如图2.1所示。一般的深度神经网络学习的是输入到输出的映射：

$$f : X \rightarrow Y \quad (2.1)$$

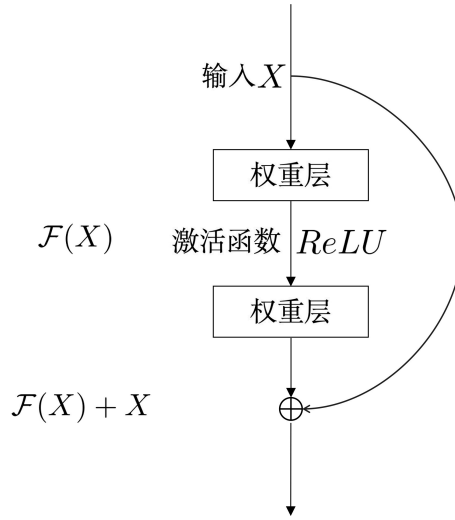


图 2.1 残差模块示意图

其中 X 为网络输入， Y 为输出。在残差模块中，输入不仅通过权重层往前传播，还通过跳跃连接作为恒等部分加入到输出中，因此，残差模块得到的输出表示为：

$$y = \sigma(\mathcal{F}(X) + X) \quad (2.2)$$

其中， σ 是激活函数， $\mathcal{F}(X)$ 是残差模块学习的映射。通过跳跃连接，网络层和顶层之间有了更短的传播路径，从而可以接收到更多来自顶层的梯度信息，梯度消失的问题有效地得到了解决。在目标检测任务中，ResNet 的引入使得主干网络能够堆叠至上百层，在有效保留底层边缘纹理特征的同时，提取出极具判别力的高层语义特征。

2.1.2 多尺度特征融合与特征金字塔

尽管深层网络能够提取极强的语义特征，但经过多次下采样后，其特征图的空间分辨率已大幅降低。这种低分辨率的深层特征对大尺度的目标较为友好，但对于微小目标而言，其位置信息与空间细节在深层网络中往往已大量丢失。相反，浅层特征图虽然具有较高的分辨率，保留了丰富的细节信息，但其缺乏足够的宏观语义支撑。

为了解决这一空间分辨率与语义信息的矛盾，特征金字塔网络 FPN^[49] 应运而生。如图 2.2 所示，FPN 提出了一种自顶向下结合横向连接的架构，旨在构建一个在所有尺度上都具备强语义信息的高质量特征金字塔。具体而言，FPN 首先对深层低分辨率、强语义的特征图进行空间上采样，使其分辨率与上一级浅层特征图保持一致；随

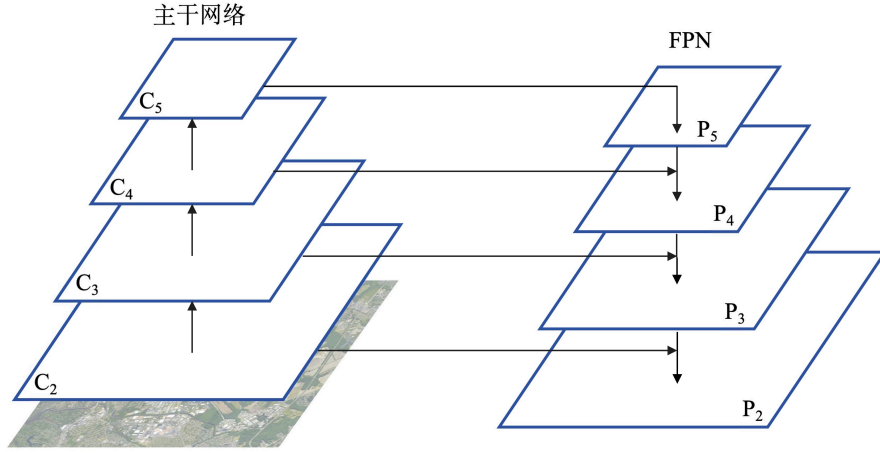


图 2.2 特征金字塔网络 FPN 示意图

后，通过 1×1 卷积调整浅层特征的通道数，并将两者进行逐元素相加。其特征融合的计算过程可简化表示为：

$$P_i = \text{Conv}_{1 \times 1}(C_i) + \text{Upsample}(P_{i+1}) \quad (2.3)$$

其中， C_i 代表主干网络自底向上路径中第 i 阶段的输出特征， P_{i+1} 为更高层级的融合特征， P_i 则为当前层生成的金字塔特征。

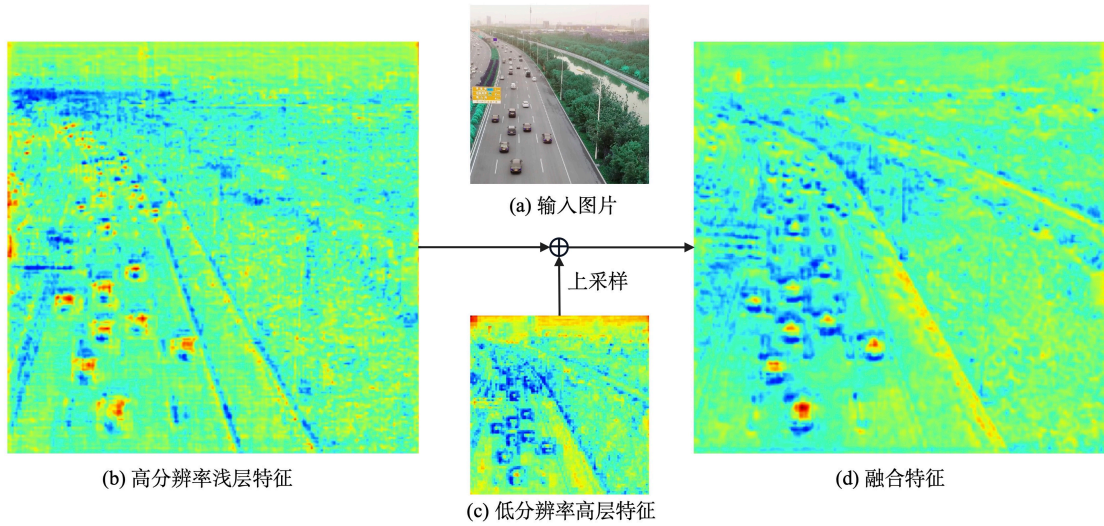


图 2.3 FPN 对微小目标特征的影响示意图

通过这种多尺度特征的深度交融，FPN 能够为不同尺度的目标分配最适合的特征层进行预测，例如利用底层的 P_2 高分辨率特征来检测微小目标，利用顶层的 P_5

特征来检测大型目标。这种架构现已成为绝大多数目标检测模型的标准配置，极大提升了模型在多尺度场景下的鲁棒性。然而，FPN 基于双线性上采样和逐元素相加的融合过程在多尺度特征之间引入了空间不对齐问题，从而对微小目标产生了负面影响。如图 2.3 所示，与原始的浅层特征图相比，融合后特征图中的微小目标细节信息变得模糊。与能够容忍一定信息丢失的正常尺寸目标不同，微小目标由于像素极少，任何程度的细节丢失都可能导致它们无法被有效检测。

2.2 目标检测的标签分配

在目标检测模型的训练过程中，模型通常会生成大量的预测结果。如何将这海量的预测结果与图像中数量有限的真实目标进行匹配，从而划分出用于计算损失函数的正样本和负样本，这一过程被称为标签分配。标签分配策略的优劣直接决定了模型能否获得高质量的监督信号，对于极易漏检或产生定位偏差的微小目标而言，合理的标签分配机制尤为关键。本节将依次介绍基于锚框、无锚框以及 DETR 架构下的一对一标签匹配机制。

2.2.1 基于锚框的标签分配

在以 Faster R-CNN 和早期 YOLO（如 YOLOv3/v4/v5）为代表的模型中，基于锚框的标签分配是最经典的策略。该策略的核心依赖于预测锚框与真实目标框之间的空间重叠度，通常采用交并比 IoU 作为核心度量指标：

$$IoU = \frac{\text{Area}(B_{\text{anchor}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{anchor}} \cup B_{\text{gt}})} \quad (2.4)$$

其中 B_{anchor} 表示锚框， B_{gt} 表示 GT 框。基于 IoU 的匹配通常遵循最大 IoU 分配与阈值划分原则。首先，模型会计算每个真实目标框与所有锚框的 IoU，并将与该真实框 IoU 最大的锚框无条件设为正样本，以保证每个目标至少有一个正样本负责预测。其次，设定两个超参数阈值 IoU_{high} 和 IoU_{low} ，如果某个锚框与任意 GT 的 IoU 大于 IoU_{high} ，则将其划分为正样本；如果小于 IoU_{low} ，则划分为背景负样本；介于两者之间的锚框则通常被忽略，不参与梯度反向传播。

这种静态的分配机制对微小目标极不友好。如图 2.4 所示，由于微小目标的绝对面积很小，预测框哪怕只有几个像素的微小偏移，其 IoU 值也会急剧下降。这导致

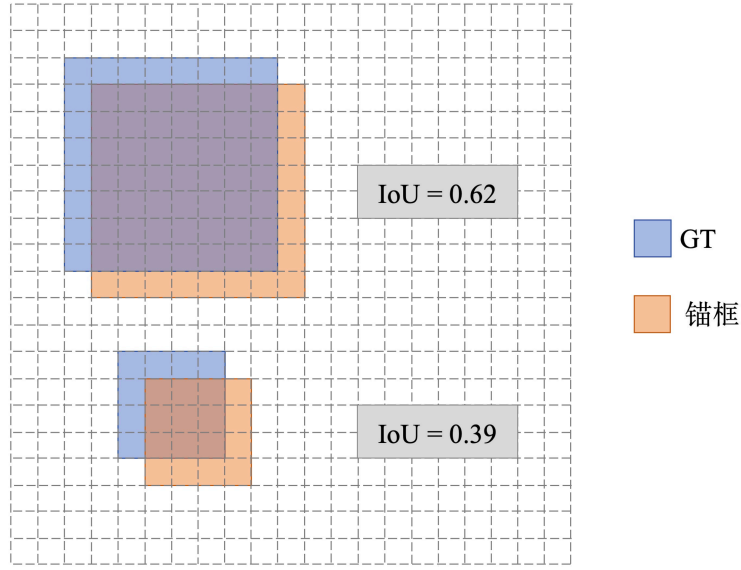


图 2.4 不同尺寸目标预测框偏移对 IoU 的影响示意图

在常规 IoU_{high} 阈值（如 0.5）下，大量微小目标无法匹配到足够的高质量正样本，陷入缺乏正样本的困境，从而严重限制了模型对小目标的召回率。

2.2.2 无锚框与动态标签分配

为了彻底摆脱预设锚框带来的超参数敏感、正负样本极度不平衡以及 IoU 计算复杂等缺陷，无锚框检测器应运而生。这类模型借鉴了语义分割的全卷积逐像素预测思想，将目标检测重构为直接在特征图网格点上进行的逐点预测任务。例如 FCOS^[50] 将落入真实目标框中心区域的特征网格点定义为正样本，并利用中心度分支抑制低质量的边缘预测。

在 FCOS 中，特征图上的每一个空间位置（即像素点）都被视为一个潜在的预测样本。假设特征图上某一点的坐标为 (x, y) ，传统的基于锚框的方法会将该点作为中心生成多个锚框再计算 IoU；而 FCOS 的空间静态分配策略则简单直接，如果该点 (x, y) 落入了任何一个真实目标框的内部，该点即被分配为正样本。

对于分配为正样本的点，FCOS 不再回归边界框的中心点与宽高偏移量，而是直接回归当前点 (x, y) 到真实目标框左、上、右、下四个边界的绝对距离。设真实框的左上角和右下角坐标分别为 (x_0, y_0) 和 (x_1, y_1) ，其回归目标 (l^*, t^*, r^*, b^*) 的数学定义为：

$$l^* = x - x_0, \quad t^* = y - y_0 \quad (2.5)$$

$$r^* = x_1 - x, \quad b^* = y_1 - y \quad (2.6)$$

其中, l^* 表示当前点到左边界的水平距离; t^* 表示当前点到上边界的垂直距离; r^* 表示当前点到右边界的水平距离; b^* 表示当前点到下边界的垂直距离。这种极简的分配方式会面临两个严峻问题, 一是当两个不同尺寸的真实目标发生重叠时, 重叠区域内的点将产生分配歧义; 二是目标框边缘位置的正样本点虽然被包含在框内, 但由于远离目标中心, 往往只能提取到背景特征, 从而产生大量低质量的预测框。

针对第一个问题, **FCOS** 深度利用了 **FPN** 的多尺度特性, 通过硬性限制不同特征层负责回归的距离范围 (例如底层负责回归距离较小的小目标, 顶层负责大目标), 巧妙地缓解了重叠歧义。针对第二个问题, **FCOS** 创新性地引入了一个并行的中心度预测分支。中心度用于衡量一个正样本点距离目标真实中心的远近程度, 其真实标签 $Center-ness^*$ 定义为:

$$Center-ness^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}} \quad (2.7)$$

中心度的值域为 $(0, 1]$, 越靠近目标中心的值越大。在推理阶段, 模型将预测的中心度与分类得分相乘作为最终的置信度, 从而利用非极大值抑制 **NMS** 自然地过滤掉那些远离目标中心的低质量预测框。

2.2.3 DETR 的一对一二分图匹配

无论是基于锚框还是无锚框, 传统的卷积神经网络检测器通常采用一对多的分配策略, 即一个真实目标会被分配给多个正样本进行预测。这必然导致推理阶段产生大量冗余的重叠边界框, 从而必须依赖非极大值抑制 **NMS** 这一人工设计的启发式算法进行过滤。

基于 **Transformer** 的 **DETR** 及其衍生模型通过引入二分图匹配, 确立了一种新的范式。**DETR** 将预测输出视为一个固定长度为 N 的无序集合 $\hat{y}$, 并将真实标签补充空集 (背景类) 扩展为长度为 N 的集合 y 。标签分配的过程被转化为寻找两个集合之间代价最小的一对一排列映射 σ 。该过程通常利用匈牙利算法求解:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^N \mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) \quad (2.8)$$

其中, $\mathfrak{S}_N$ 为所有可能的排列组合, 匹配代价函数 $\mathcal{L}_{match}$ 综合了分类损失 $\mathcal{L}_{cls}$ 、L1

距离损失 $\mathcal{L}_{L1}$ 以及广义交并比 (GIoU) 损失 $\mathcal{L}_{giou}$:

$$\mathcal{L}_{match} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou} \quad (2.9)$$

通过这种全局最优的一对一匹配机制, DETR 强迫每个目标只由一个 Query 负责预测, 从根本上消除了冗余框的产生, 实现了真正的无 NMS 端到端检测。然而, 这种严苛的一对一分配机制在训练早期阶段缺乏一对多分配带来的丰富监督信号, 导致模型收敛缓慢, 这也是目前基于 DETR 架构进一步优化微小目标检测亟需克服的理论难点之一。

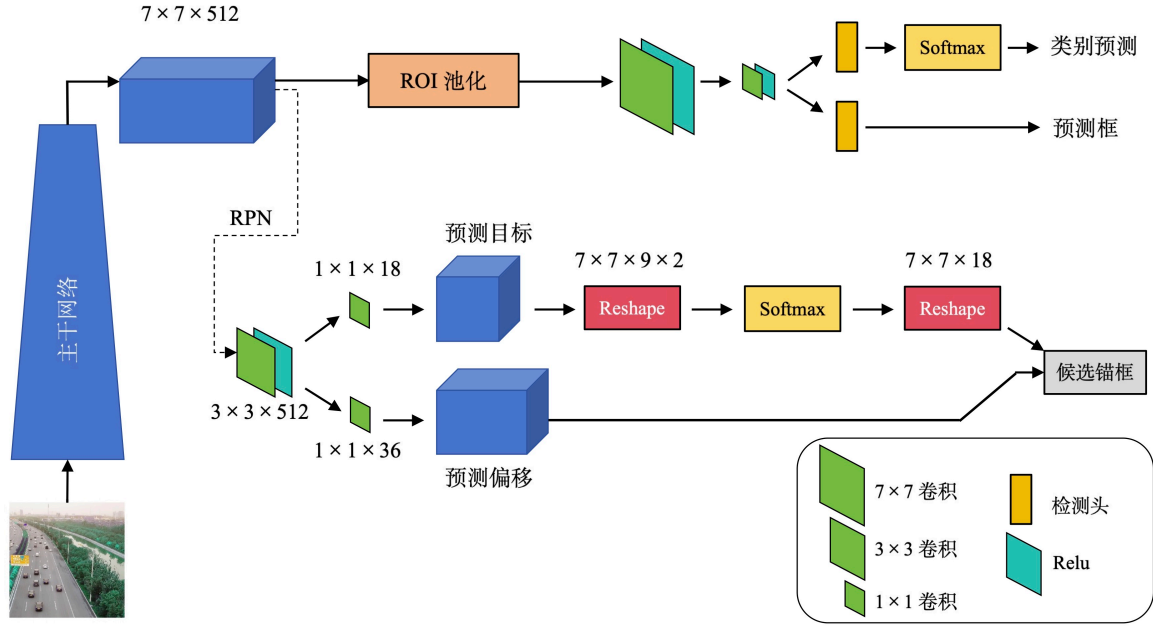
2.3 经典目标检测方法

在明确了特征提取与融合、标签分配等基础机制后, 本节将聚焦于目标检测模型的核心架构演进。随着深度学习技术的不断发展, 目标检测领域逐渐衍生出三大主流范式, 分别为追求极致精度的双阶段检测模型、兼顾速度与精度的单阶段检测模型, 以及近年来兴起的基于 Transformer 的端到端检测模型。本节将分别以 Faster R-CNN、YOLOv8 和 RT-DETR 作为这三大范式的代表进行详细阐述, 为后续微小目标检测算法的改进提供模型层面的理论支撑, 本文第三章方法基于 Faster R-CNN 改进, 第四章方法基于 RT-DETR 改进。

2.3.1 双阶段目标检测模型 Faster R-CNN

双阶段目标检测模型的核心思想是将目标检测任务解耦为两个连续的子阶段。首先, 通过特定的启发式算法或神经网络在图像中提取出可能包含目标的候选区域; 其次, 利用卷积神经网络对这些候选区域进行特征提取, 并输入到后续的分类器和回归器中, 完成目标类别的判定和边界框坐标的精细调整。这种先粗筛、后精调的级联架构, 使得双阶段模型在检测精度和召回率上具有显著优势, 尤其是在复杂场景和多尺度目标检测中表现优异。

Faster R-CNN^[51] 作为双阶段目标检测模型的里程碑式工作, 其最大贡献在于提出了区域候选网络 RPN, 将候选框的生成过程完全融入深度神经网络中, 首次实现了目标检测的端到端训练。本文第三章的改进工作主要基于 Faster R-CNN 框架, 此处对其核心组件及数学原理进行重点概述。

图 2.5 Faster R-CNN^[51] 模型结构示意图

如图2.5所示，在 Faster R-CNN 的架构中，输入图像首先通过共享的主干网络提取出特征图。随后，RPN 在特征图的每个空间位置上预设多种尺度和长宽比的锚框。RPN 作为一个全卷积网络，通过二分类判断锚框内是否包含目标，并初步调整锚框的坐标。RPN 的多任务损失函数定义为：

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (2.10)$$

其中， i 表示一个批次中锚框的索引； p_i 是锚框被预测为目标概率； p_i^* 为真实标签（正样本为 1，负样本为 0）； t_i 表示预测的边界框四个参数化坐标向量， t_i^* 则是与正样本锚框对应的真实边界框坐标向量。 L_{cls} 为交叉熵分类损失， L_{reg} 为边界框回归损失， λ 为平衡权重的超参数。经过 RPN 筛选出的高质量候选区域将通过 RoI 池化层映射到固定维度的特征向量，最后送入全连接网络完成最终的类别判定和边界框精细回归。

尽管 Faster R-CNN 具备极高的检测精度，但其复杂的级联结构导致推理速度较慢；同时，对于微小目标而言，传统 RPN 预设的锚框尺寸往往偏大，难以与微小目标的真实感受野实现 IoU 匹配，导致正样本匮乏。因此，如何增强高分辨率特征的代表能力、优化锚框机制，是利用该框架提升微小目标检测精度的关键切入点。

2.3.2 单阶段目标检测模型 YOLOv8

为了突破双阶段模型在推理速度上的瓶颈，单阶段模型摒弃了耗时的区域候选提取步骤，将目标检测重构为直接在特征图上进行的密集预测任务。YOLOv8^[52] 作为 YOLO 系列的先进模型，在模型架构、正负样本分配以及损失函数设计上进行了全面的现代化重构，是目前工业界应用最广泛的实时检测基准模型之一。

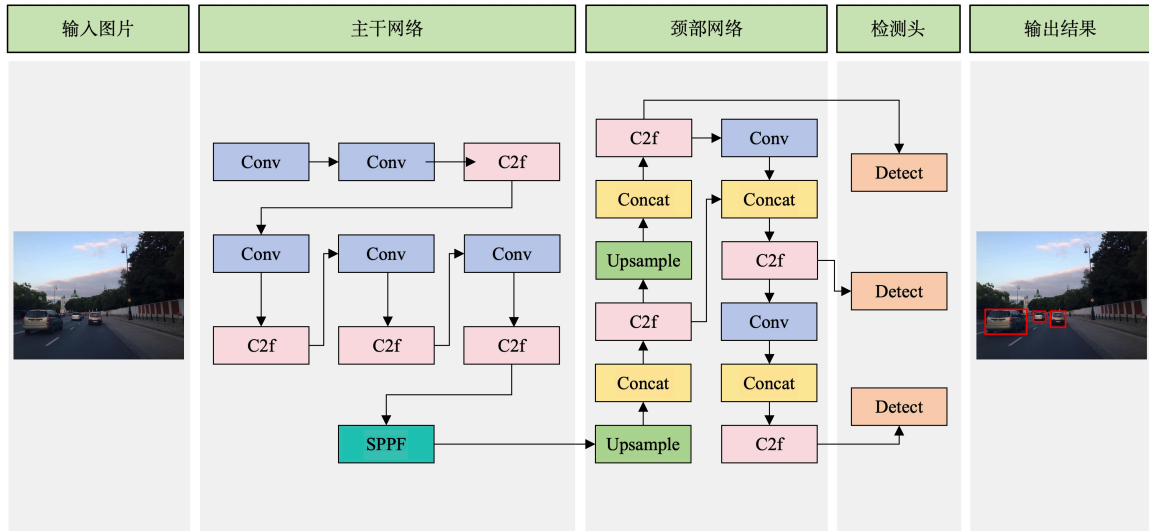


图 2.6 YOLOv8^[52] 模型结构示意图

如图2.6所示，YOLOv8 采用了跨阶段局部网络的设计思想，引入了轻量且高效的 C2f 模块，与传统的 Bottleneck 结构相比，C2f 模块通过增加更多的跳跃连接和梯度流分支，不仅能在保证轻量化的同时提取更丰富的特征表示，还能有效缓解深层网络中的梯度消失问题，从而更好地保留图像中微小目标的浅层纹理信息。

在特征聚合阶段，YOLOv8 沿用了 PANet 的双向融合架构。该结构同时包含自顶向下的语义信息传递与自底向上的定位特征馈送。这种深浅层特征的充分交融，使得模型能够构建出高质量的多尺度特征金字塔，显著提升了对不同尺度目标的感知能力。

解耦头与无锚框设计是 YOLOv8 区别于早期 YOLO 模型的最显著特征。传统检测头将分类和回归任务耦合在同一卷积层中，由于两项任务对特征的内在需求不同（分类倾向于平移不变性，回归倾向于平移敏感性），极易产生空间对齐冲突。YOLOv8 采用了解耦头结构，将类别预测和边界框回归分为两个独立的物理分支。此外，YOLOv8 全面转向了无锚框范式，这一设计彻底摆脱了对锚框尺寸、长宽比等

启发式超参数的依赖，极大地提升了模型的泛化能力。设特征图上某网格的中心点坐标为 (x, y) ，模型回归分支直接预测该中心点到真实边界框左、上、右、下四个边界的距离 (l, t, r, b) ，从而解码出预测框的坐标 $(x_{min}, y_{min}, x_{max}, y_{max})$ ：

$$x_{min} = x - l, \quad y_{min} = y - t \quad (2.11)$$

$$x_{max} = x + r, \quad y_{max} = y + b \quad (2.12)$$

在正负样本匹配策略上，YOLOv8 采用了任务对齐动态分配策略。该策略旨在解决分类得分与定位精度不对齐的问题，通过构建对齐度量指标 t 来动态计算样本权重：

$$t = s^\alpha \cdot u^\beta \quad (2.13)$$

其中， s 表示模型预测的分类得分， u 表示预测边界框与真实框之间的交并比 IoU， α 和 β 为调节分类与定位任务权重的超参数。该策略能够高效筛选出具备高分置信度且定位准确的高质量正样本。

通过单次前向传播与高效的特征融合，YOLOv8 实现了极佳的速度与精度平衡，但尽管融入了多尺度特征和优化的损失函数，基于 CNN 的 YOLOv8 在极致的微小目标检测任务中依然面临不可忽视的瓶颈。由于取消了区域候选阶段，模型缺乏对特定区域特征的二次对齐和精细化裁剪，导致微小目标在复杂背景下的定位极易出现偏差。此外，密集预测必然依赖非极大值抑制作为后处理步骤来消除冗余框，而 NMS 的超参数（如 IoU 阈值）对密集排列的小目标极其敏感，极易造成小目标的漏检。CNN 局部感受野和手工后处理带来的固有缺陷，促使学术界开始探索具备全局感受野且无需 NMS 的端到端检测范式。

2.3.3 端到端目标检测模型 RT-DETR

近年来，随着 Transformer 架构在自然语言处理领域的巨大成功，以自注意力机制为核心的全局上下文建模范式被引入到计算机视觉领域。2020 年，DETR 模型的提出，标志着目标检测技术迈入了一个全新的端到端时代。该模型彻底抛弃了传统 CNN 模型中依赖人工先验设计的锚框以及非极大值抑制后处理步骤，将目标检测任务重构为集合预测问题。

DETR 的核心思想是利用卷积神经网络提取图像的二维空间特征，将其展平并

叠加位置编码后输入 **Transformer** 编码器。在解码器端，模型引入了一组可学习的对象查询，即 **Query**，通过交叉注意力机制从编码器输出的全局上下文中聚合特定目标的特征。但尽管 **DETR** 实现了极其优雅的端到端架构，但其核心的全局自注意力机制在计算时需要将所有 **Query** 与所有的键进行点积交互。设输入特征序列长度为 N ，通道数为 C ，其计算复杂度为 $\mathcal{O}(N^2C)$ 。为了控制这种随序列长度呈平方级增长的巨大计算开销，**DETR** 只能依赖深层低分辨率特征图进行预测。这一妥协导致模型对图像局部细节的感知能力极弱，在微小目标检测任务上表现堪忧，同时模型面临着收敛速度极其缓慢的理论缺陷。

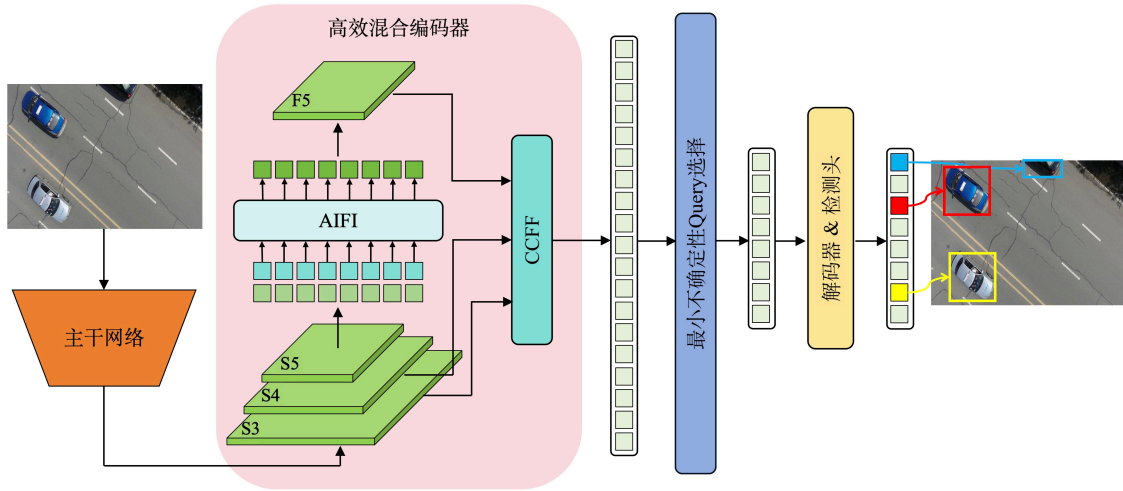


图 2.7 RT-DETR^[44] 模型结构示意图

在此背景下，百度提出的 **RT-DETR**^[44] 通过对模型架构进行深度的计算解耦与冗余消除，首次在基于 **Transformer** 的检测器上实现了超越先进 **YOLO** 模型的实时推理速度。由于本文第四章的改进工作主要围绕该模型展开，此处对其核心架构进行详细解析。如图2.7所示，**RT-DETR** 的高效架构主要由两个创新组件构成。首先是高效混合编码器，传统多尺度 **Transformer** 编码器中，跨尺度与同尺度特征的自注意力计算相互耦合，导致了严重的计算冗余。**RT-DETR** 创新性地将编码器解耦为基于注意力的尺度内特征交互模块 **AIFI** 和基于 **CNN** 的跨尺度特征融合模块 **CCFM**。鉴于深层特征图具备丰富的语义信息且分辨率较低，**AIFI** 仅在主干网络输出的最高层（如 S_5 特征层）执行单尺度的自注意力计算，在捕获全局概念的同时极大地压缩了计算量。对于包含丰富细粒度信息的浅层特征，**RT-DETR** 放弃了昂贵的注意力机制，转而采用经过优化的 **CNN** 模块进行自上而下与自下而上的多尺度特征融合。这种混

合设计兼顾了 **Transformer** 的全局语义感知能力与 **CNN** 局部特征提取的高效性。

其次是最小化不确定性 **Query** 选择。在 **Decoder** 的 **Query** 初始化阶段，以往模型多采用启发式选取或随机初始化，导致分类得分与定位精度往往存在不对齐现象。**RT-DETR** 提出利用编码器输出特征的联合不确定性指标来筛选最优的初始特征。其度量分数 S_i 综合了分类置信度 $\hat{p}_i$ 与预测边界框的 **IoU**：

$$S_i = \hat{p}_i(c_i) \cdot \text{IoU}(\hat{b}_i, b_i) \quad (2.14)$$

模型通过该分数对编码器特征进行排序，筛选出最具代表性的前 K 个特征作为初始 **Query**。这一机制为 **Decoder** 提供了更高质量的区域先验信息，极大加速了二分图匹配的收敛过程。

此外，**RT-DETR** 还支持在不重新训练网络的情况下，根据实际部署的算力平台灵活调整解码器的层数，从而在检测精度与推理速度之间实现平滑的动态权衡。**RT-DETR** 的出现，为实时微小目标检测提供了一种无需后处理且具备全局感知能力的新范式。

尽管 **RT-DETR** 在多数目标检测任务中取得了令人满意的效果，但其在微小目标检测任务中仍存在一定局限。受实时性要求限制，**RT-DETR** 难以对高分辨率浅层特征进行充分建模，使模型对微小目标依赖的细粒度空间信息利用不足。此外，**RT-DETR** 的 **Query** 选择机制更倾向于保留响应显著的候选区域，而微小目标特征响应弱，容易在筛选阶段被忽略，影响目标召回率。因此，如何高效利用高分辨率特征、优化 **Query** 选择机制，是利用该框架提升微小目标检测精度的关键优化方向。

2.4 评价指标

为了客观、定量地评估目标检测模型在特定任务上的综合性能，必须建立科学的评价体系。鉴于本文的后续研究分别侧重于提升微小目标的检测精度（第三章）与追求模型的实时推理速度（第四章），本节将评价指标划分为精度评价指标与速度、模型计算复杂度评价指标两大类进行概述。

2.4.1 精度评价指标

目标检测的精度评估依赖于预测边界框与真实边界框之间的重合度，该重合度通常由交并比 **IoU** 来衡量。基于预设的 **IoU** 阈值（通常设为 0.5），可以将模型的预

测结果划分为以下三种基础统计量：

(1) 真正例 (True Positive, TP)：预测框包含目标，且与 GT 框的 $IoU \geq$ 阈值。

(2) 假正例 (False Positive, FP)：预测框不包含目标，或与 GT 框的 $IoU <$ 阈值 (即误检)。

(3) 假负例 (False Negative, FN)：真实目标未被模型检测出 (即漏检)。

基于上述统计量，可以进一步推导出精确率 (Precision) 和召回率 (Recall)。精确率衡量的是模型预测出的正样本中有多少是真正的正样本，而召回率衡量的是所有真实的样本中有多少被模型成功找了出来：

$$Precision = \frac{TP}{TP + FP} \quad (2.15)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.16)$$

在实际评估中，精确率和召回率往往是一对相互制约的指标。为了综合衡量模型在不同置信度阈值下的表现，通常引入平均精度 (Average Precision, AP) 与平均精度均值 (mean Average Precision, mAP)。AP 定义为以召回率 R 为横轴、精确率 P 为纵轴所绘制的 P-R 曲线下的积分面积：

$$AP = \int_0^1 P(R) dR \quad (2.17)$$

MS COCO^[1] 的 AP 指标被用于评估各种检测器的性能，AP 值越高代表检测器性能越好。具体而言， AP_{50} 表示定义真正例 TP 的 IoU 阈值为 0.5； AP_{75} 表示定义 TP 的 IoU 阈值为 0.75；AP 则表示从 AP_{50} 到 AP_{95} (IoU 间隔为 0.05) 的平均值。 AP_S 、 AP_M 和 AP_L 分别对应基于目标尺寸计算的 AP 值。特别地，在本文后续第三章、第四章的实验中，对于 AI-TOD^[2] 数据集，这些指标被替换为 AP_{vt} 、 AP_t 、 AP_s 和 AP_m ；对于 SODA-D^[53] 数据集，则分别被替换为 AP_{eS} 、 AP_{rS} 、 AP_{gS} 和 AP_N 。评价指标与目标尺寸之间的关系如表 2.1 所示。

2.4.2 速度与计算复杂度评价指标

为了全面评估本文第四章所提轻量化及端到端模型的部署可行性与实时性，除了精度外，还需引入以下速度与复杂度评价指标：

(1) 每秒帧数 (FPS)：衡量模型推理速度的最直观指标，即模型在一秒钟内能够成功处理并输出检测结果的图像数量。在自动驾驶或视频监控等实际工业场景中，

表 2.1 不同数据集评价指标与目标尺寸之间的对应关系

数据集	评价指标	目标尺寸 (像素数)
AI-TOD	AP_{vt}	[4, 64]
	AP_t	(64, 256]
	AP_s	(256, 1024]
	AP_m	(1024, ∞]
SODA-D	AP_{es}	(0, 144]
	AP_{rs}	(144, 400]
	AP_{gs}	(400, 1024]
	AP_N	(1024, 2000]
VisDrone & AgriPest	AP_S	(0, 1024]
	AP_M	(1024, 4096]
	AP_L	(4096, ∞)

通常要求 $FPS \geq 30$ 以满足实时性要求。

(2) 参数量 (Params): 衡量模型空间复杂度的指标, 即网络中所有需要通过反向传播学习的权重参数的总数 (通常以兆, 即 M 为单位)。参数量直接决定了模型部署时所需的磁盘存储空间和内存占用。

(3) 浮点运算次数 (FLOPs): 衡量模型时间复杂度与计算开销的指标 (通常以 GFLOPs 为单位)。FLOPs 越低, 表明模型在进行单次前向推理时所需的硬件算力越小, 越有利于向算力受限的边缘设备迁移。

2.5 本章小结

本章对实时微小目标检测所涉及的相关理论和核心技术进行了系统性的梳理与概述, 旨在为后续章节的算法改进与网络设计提供详实的理论支撑。首先介绍了视觉特征提取与融合的基础框架; 其次重点探讨了目标检测任务中至关重要的标签分配机制; 然后深入剖析了经典目标检测模型的三大主流范式及其演进脉络; 最后为了客观地评价本论文的方法, 本章介绍了使用的评价指标。

第三章 基于隐式特征融合与混合自适应标签分配的目标检测

微小目标在图像中所占像素极为有限，这对特征提取、多尺度融合以及主流检测方法所采用的标签分配策略均提出了挑战。针对特征金字塔在跨尺度融合时存在的空间不对齐与微小目标信息丢失问题、以及基于 IoU 的标签分配无法为微小目标分配足够数量正样本的问题，本章提出了一种基于隐式特征融合和混合自适应标签分配的微小目标检测方法，通过隐式特征融合缓解多尺度特征融合中的细节丢失问题，并利用结合 IoU 与感受野距离的自适应标签分配策略提升训练样本质量，旨在实现高精度的微小目标检测。方法的有效性将在本章的对比实验与消融实验中加以验证。

3.1 方法概述

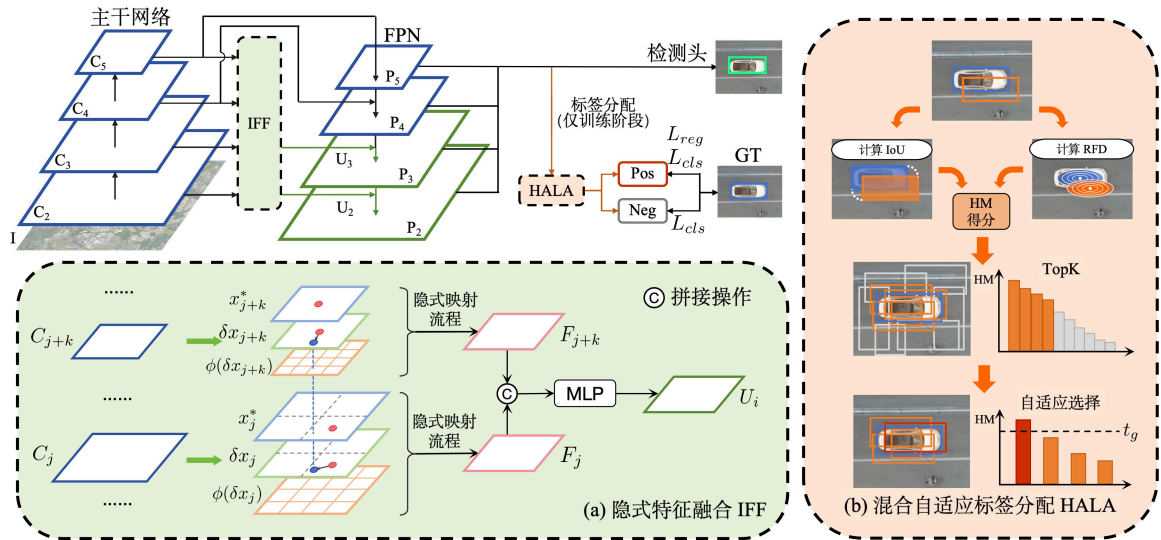


图 3.1 基于隐式特征融合与混合自适应标签分配的目标检测方法架构

图 3.1 展示了所提方法的整体框架。给定输入图像 I ，利用主干网络 ResNet^[48] 从不同阶段生成四个特征图 $\{C_i\}_{i=2}^5$ 。在该框架中，隐式特征融合 (Implicit Feature Fusion, IFF) 利用 $\{C_i\}_{i=2}^5$ 进行特征融合，生成不同金字塔层级的融合特征图 U_2, U_3 ，随后与 C_2, C_3 一起被输入到特征金字塔 FPN 中，并进一步融合得到最终的融合特征图

$\{P_i\}_{i=2}^5$ ，这些特征图被输入到检测头中。**IFF** 缓解了特征融合过程中的不对齐问题，并输出更加准确的融合特征。在训练阶段，混合自适应标签分配策略（**Hybrid Metric Adaptive Label Assignment, HALA**）负责选择合适的正负样本，用于计算与真实边框的相应损失，指导模型优化。

3.1.1 隐式特征融合 IFF

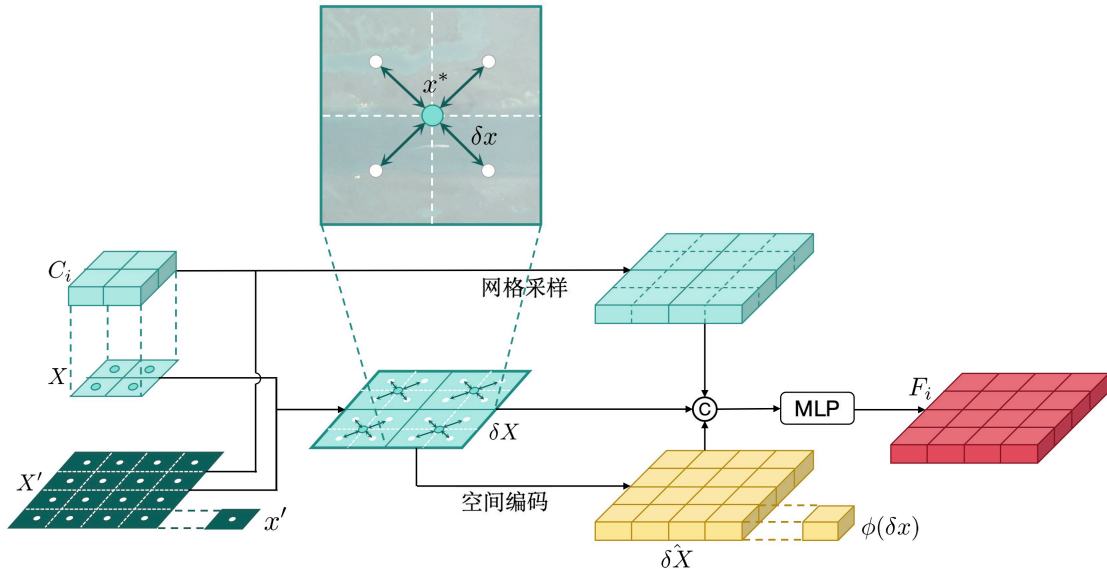


图 3.2 隐式映射流程示意图

现有的大多数多尺度特征融合方法往往在融合前将高层特征图上采样至与低层特征图相同的尺寸，这一过程容易导致特征不对齐以及细节特征的丢失。**IFF** 的目标是利用隐式神经表示将多尺度特征映射到统一的尺寸，同时，它采用空间编码来挖掘特征图之间的潜在联系。最后，将所有映射后的特征进行整合，生成融合后的特征图。

隐式神经表示的原理是利用神经网络来拟合连续信号或几何结构，并学习输入坐标与其对应信号之间的映射关系。为了将特征图表示为连续信号并将其映射到任意尺度，本章方法定义了隐式映射流程，如图 3.2 所示。隐式映射流程的输入为原始特征图 $C_i \in \mathbb{R}^{C \times H_i \times W_i}$ ，两个坐标集合 $X = \{(\frac{j-0.5}{H_i}, \frac{k-0.5}{W_i}) \mid 1 \leq j \leq H_i, 1 \leq k \leq W_i\}$ 与 $X' = \{(\frac{j-0.5}{H'}, \frac{k-0.5}{W'}) \mid 1 \leq j \leq H', 1 \leq k \leq W'\}$ ，输出为映射特征图 $F_i \in \mathbb{R}^{C \times H' \times W'}$ 。

其中 H', W' 表示隐式映射流程输出特征图 F_i 的尺度大小, X, X' 分别表示 C_i, F_i 中通道维度向量的坐标集合。

首先, 对于每个 $x' \in X'$, 隐式映射流程需要计算 x' 与 X 中距离最近坐标 x^* 的差值 δx , δx 的集合为 δX 。 δx 可以用下式表示:

$$\delta x = x' - x^* \quad (3.1)$$

$$x^* = \underset{x \in X}{\operatorname{argmin}} \|x' - x\| \quad (3.2)$$

其中, $\|\cdot\|$ 表示坐标间的欧式距离。相关研究^[54]指出, 空间编码能够在特征融合过程中帮助模型学习更精细的细节, 从而提升微小目标检测的性能。因此, 隐式映射流程采用如下的编码函数对 δx 进行空间编码:

$$\begin{aligned} \phi(\delta x) = & (\sin(\omega_1 \delta x^T), \cos(\omega_1 \delta x^T), \sin(\omega_2 \delta x^T), \cos(\omega_2 \delta x^T), \\ & \dots, \sin(\omega_{2N} \delta x^T), \cos(\omega_{2N} \delta x^T)) \end{aligned} \quad (3.3)$$

其中, $\omega_1, \omega_2, \dots, \omega_N$ 为空间编码的可学习频率参数, 其被初始化为 $\omega_n = (2^n, 0), \omega_{n+N} = (0, 2^n), n \in \{1, 2, \dots, N\}$, N 为空间编码的超参数, 默认设置下 $N = 24$ 。该空间编码函数将 δx 转换为一个 $4N$ 维向量, $\phi(\delta x)$ 的集合记作 $\delta \hat{X}$ 。

然后, 隐式映射流程以 X' 为坐标网格使用最近邻插值将 C_i 采样至目标尺寸, 便于 **MLP** 建模。最终, 隐式映射流程可以用下式表示:

$$F_i = \text{MLP}(\text{Concat}(\text{GridSample}(C_i, X'), \delta X, \delta \hat{X})) \quad (3.4)$$

其中, MLP 表示多层感知机建模, Concat 表示拼接操作, GridSample 表示最近邻插值网格采样操作。

在特征融合阶段, 如图 3.1(a) 所示, **IFF** 采用上述隐式映射流程, 将来自不同级别的多尺度特征映射至统一的分辨率进行拼接, 随后使用 **MLP** 挖掘特征层之间的关系。以融合特征 $U_i \in \mathbb{R}^{C \times H_i \times W_i}$ 为例, U_i 可以通过下式进行计算:

$$U_i = \text{MLP}(\text{Concat}(\{F_j\}_{j=i}^5)) \quad (3.5)$$

由于隐式映射流程每次只能将原始特征图映射至特定尺度, 因此同一特征图需经历多次映射才能获得不同尺度的融合特征图。然而, 计算高层融合特征对提升微小目标检测性能的贡献有限, 并且会引入冗余计算。因此, 本章方法可以根据下游

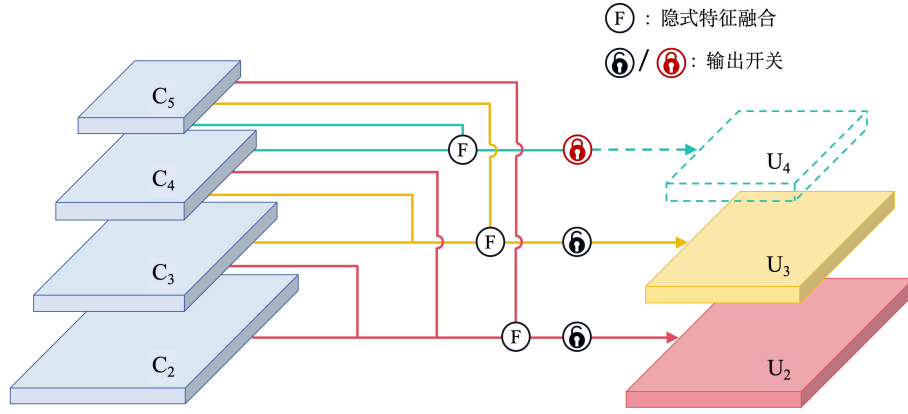


图 3.3 特征融合层级示意图

任务来选择特定的融合层，而非获取所有层级的融合特征，如图 3.3 所示。第 3.2.3.2 节中开展的消融实验验证了该设计的有效性。

为了将 IFF 无缝集成到大多数目标检测框架中，本章方法将其与 FPN 相结合。得益于所学习到的连续信号，隐式映射流程能够将特征图映射至任意分辨率。为了适配 FPN，本章方法将 IFF 融合特征图 U_i 的分辨率设定为与 C_i 一致，并使用 U_2, U_3 来替换 FPN 中用于融合 P_2, P_3 的横向连接；而对于没有输出融合特征的金字塔层级，则继续沿用原始的 FPN 设置，使用 C_4, C_5 特征图，在此框架下 P_i 可以表示为：

$$P_i = \begin{cases} P_{i+1} + U_i, & \text{if } i \in [2, 3], \\ P_{i+1} + C_i, & \text{if } i = 4, \\ C_5, & \text{if } i = 5. \end{cases} \quad (3.6)$$

3.1.2 混合自适应标签分配策略 HALA

在训练目标检测模型时，首先需要划分正负样本以计算分类损失，随后利用正样本来计算回归损失，引导模型收敛。因此，正负样本的选择过程至关重要。如 2.2.1 节所述，由于微小目标 GT 框与锚框的轻微像素偏移会 IoU 骤降，传统基于 IoU 的标签分配策略往往无法为微小目标分配足够数量的正样本。近年来，研究者们引入了一些专门针对微小目标设计的新度量指标，用于衡量先验框与 GT 框区域之间的匹配度（例如 RFLA^[35] 中的 RFD，以及 NWD^[34]）。基于这些新度量指标的标签分配策略在微小目标检测中取得了更优的性能。然而，仅依赖单一指标的标签分配策略往往无法达到最优效果。本章针对如何更有效地综合考虑这些新指标来提升微小

目标检测性能进行了深入研究，并基于精心设计的混合度量得分提出了 HALA。

HALA 的流程如图 3.1(b) 所示。本章方法将由锚框与 GT 框计算得出的感受野距离 (Receptive Field Distance, RFD) 与 IoU 相结合，从而获得混合度量 (Hybrid Metric, HM) 得分，该得分被用于评估锚框质量。随后，利用 Top-K 机制与自适应选择策略来确定候选正样本和最终正样本 (即图中的橙色和红色边界框)，有效地提升了微小目标检测中正样本的质量，进而改善了检测性能。本章下文将分别对 HALA 中的关键组成部分进行介绍：

感受野距离 RFD RFD 直接衡量特征图中通道维度特征向量的有效感受野 (Effective Receptive Field, ERF) 与 GT 区域之间的匹配程度。相关研究表明，ERF 可以建模为高斯分布^[55]。本章方法将每个特征向量的 ERF 和 GT 框分别建模为二维高斯分布，并将对应高斯分布之间的 KL 散度作为 RFD。

为了将 ERF 建模为二维高斯分布，首先需要获取 ERF 的半径。然而，ERF 的半径无法直接计算，需要通过理论感受野 (Theoretical Receptive Field, TRF) 的半径近似得到^[55]。本章方法遵循 RFLA^[35] 的设定，将 ERF 半径视为 TRF 半径的一半。每个特征向量的 ERF 半径可以通过以下公式推导得出：

$$tr_n = tr_{n-1} + (k_n - 1) \prod_{i=1}^{n-1} s_i \quad (3.7)$$

$$er_n = \frac{1}{2} tr_{n-1} \quad (3.8)$$

其中， er_n ， tr_n 分别表示经过第 n 个卷积层后每个特征向量的 ERF 和 TRF 半径， k_n 和 s_n 分别表示第 n 层卷积操作的卷积核大小和步长。随后，可以将 ERF 建模为相应的二维高斯分布 $N_e(\mu_e, \sigma_e)$ 。具体而言， $N_e(\mu_e, \sigma_e)$ 的均值向量为每个特征向量对应输入图片中的坐标 (x_n, y_n) ，而协方差矩阵则由 er_n 构建而成，具体表示如下：

$$\mu_e = \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \sigma_e = \begin{bmatrix} er_n^2 & 0 \\ 0 & er_n^2 \end{bmatrix} \quad (3.9)$$

在获得 ERF 的高斯分布后，需要将 GT 框建模为另一个二维高斯分布 $N_g(\mu_g, \sigma_g)$ ，以便评估两者之间的匹配度。该高斯分布的均值向量为每个 GT 框的中心点坐标

(x_g, y_g) ，而协方差矩阵则由 GT 框宽度 w_g 和高度 h_g 构建而成，具体表示如下：

$$\mu_g = \begin{bmatrix} x_g \\ y_g \end{bmatrix}, \sigma_g = \begin{bmatrix} \frac{w_g^2}{4} & 0 \\ 0 & \frac{h_g^2}{4} \end{bmatrix} \quad (3.10)$$

在获得 ERF 和 GT 的二维高斯分布后，需要选择合适的度量指标来衡量它们之间的匹配度。本章方法选择了基于 KL 散度的 RFD，其定义如下：

$$D_{\text{KL}}(N_e \| N_g) = \frac{2er_n^2}{w_g^2} + \frac{2er_n^2}{h_g^2} + \frac{2(x_n - x_g)^2}{w_g^2} + \frac{2(y_n - y_g)^2}{h_g^2} + \ln \frac{w_g h_g}{4er_n^2} - 1 \quad (3.11)$$

$$\text{RFD} = \frac{1}{1 + D_{\text{KL}}(n_e \| n_g)} \quad (3.12)$$

其中， $D_{\text{KL}}(N_e \| N_g)$ 表示 N_e 和 N_g 之间的 KL 散度。

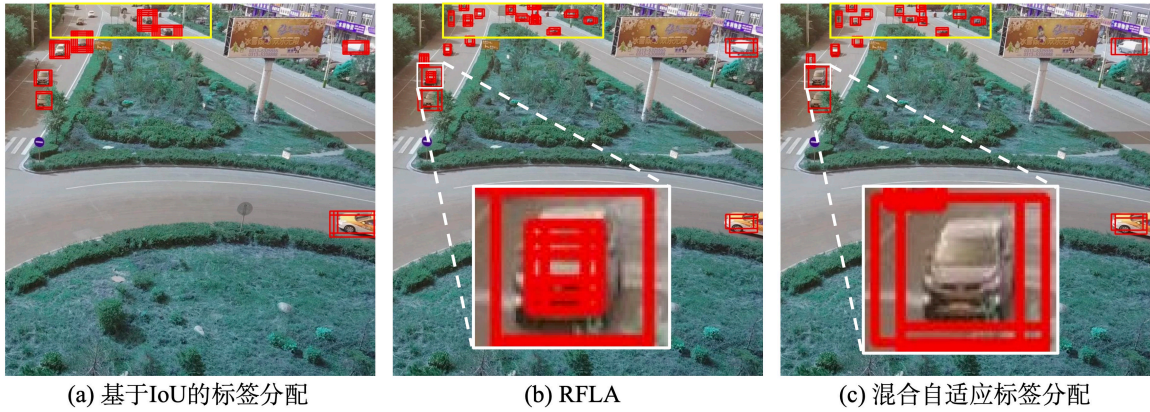


图 3.4 不同标签分配方法的正样本分配结果可视化

混合度量 HM 得分 如图3.4所示，完全基于 IoU 的策略无法为微小目标分配正样本，而完全基于 RFD 的标签分配结果中往往包含过小的正样本，这是由于 KL 散度过于侧重 GT 框与锚框之间的中心点距离。然而，较小的锚框无法完全包裹目标，从而遗漏关键信息，影响检测精度。为了缓解这一问题，本章方法在标签分配阶段同时综合考虑 RFD 和 IoU，采用 HM 得分来评估 GT 框与锚框之间的匹配度。HM 得分可以用下式计算：

$$HM = \frac{1}{2}(RFD + \alpha \times IoU) \quad (3.13)$$

其中, α 为平衡 RFD 和 IoU 的超参数。

算法 3.1: 混合自适应标签分配 HALA

Require:

- $\mathcal{G}$: 图像上的 GT 框集合
- L : 特征层级的数量
- $\mathcal{A}$: 所有锚框的集合
- $\mathcal{K}$: 每个层级的候选锚框数量, 默认设置为 12

Ensure:

- $\mathcal{P}$: 正样本集合
 - $\mathcal{N}$: 负样本集合
 - $\mathcal{I}$: 忽略样本集合
 - 1: **for** 每个 GT 框 $g \in \mathcal{G}$ **do**
 - 2: 为 g 的候选正样本及其 HM 分数构建两个空集: $\mathcal{C}_g \leftarrow \emptyset, \mathcal{S}_g \leftarrow \emptyset$
 - 3: 计算 $\mathcal{A}$ 和 g 之间的 HM 分数: $S \leftarrow \text{HM}(\mathcal{A}, g)$
 - 4: **for** $s \in S$ **do**
 - 5: **if** $0 \leq s \leq 0.8$ **then**
 - 6: $\mathcal{N} \leftarrow \mathcal{N} \cup s$
 - 7: **end if**
 - 8: **end for**
 - 9: 从 $\mathcal{A}$ 中选择 $\mathcal{K}$ 个具有最高 HM 分数的锚框: $\mathcal{A}_k, S_k \leftarrow \text{TopK}(\mathcal{A}, S, K)$
 - 10: $\mathcal{C}_g \leftarrow \mathcal{C}_g \cup \mathcal{A}_k, \mathcal{S}_g \leftarrow \mathcal{S}_g \cup S_k$
 - 11: 计算 $\mathcal{S}_g$ 的均值: $m_g \leftarrow \text{Mean}(\mathcal{S}_g)$
 - 12: 计算 $\mathcal{S}_g$ 的标准差: $v_g \leftarrow \text{Std}(\mathcal{S}_g)$
 - 13: 计算 GT 框 g 的 HM 阈值: $t_g \leftarrow m_g + v_g$
 - 14: **for** 每个候选样本 $c \in \mathcal{C}_g$ **do**
 - 15: **if** $\text{HM}(c, g) \geq t_g$ **then**
 - 16: $\mathcal{P} \leftarrow \mathcal{P} \cup c$
 - 17: **end if**
 - 18: **end for**
 - 19: **end for**
 - 20: $\mathcal{N} \leftarrow \mathcal{N} - \mathcal{P}$
 - 21: $\mathcal{I} \leftarrow \mathcal{A} - \mathcal{N} - \mathcal{P}$
 - 22: **return** $\mathcal{P}, \mathcal{N}, \mathcal{I}$
-

自适应标签分配 本章方法先选取一个候选正样本子集, 然后利用这些候选样本 HM 得分的均值和标准差计算出一个浮动阈值, 从而自适应地选择最终的正样本。目标检测网络通常会根据不同的金字塔层级来设置锚框, 并分别在单一的金字塔层级下进行标签分配。与上述传统策略不同, 本章方法将不同金字塔层级生成的所有锚框聚合到一起进行标签分配。采取这种策略是因为微小目标检测主要依赖特征金字塔中的浅层特征, 与微小目标相匹配的锚框也集中在这些较低的金字塔层级中。这一改动有助于防止引入较高金字塔层级的低质量正样本, 从而优化模型训练效果。HALA 的完整流程如算法 3.1 所示。

首先, HALA 针对图像中的每个 GT 框 g 初步划定负样本。正如第 3 至 8 行所述, 计算所有锚框与 GT 框之间的 HM 得分, 并将得分介于 0 到 0.8 之间的锚框初步加入负样本集合。接下来, 确定候选正样本。如第 9 行所示, HALA 根据每个 GT 的 HM 得分, 选取排名前 $\mathcal{K}$ 的锚框。随后, 计算这些候选正样本 HM 得分的均值和标准差, 这在第 11 和 12 行中分别记作 m_g 和 v_g 。如第 13 行所示, 该 GT 的 HM 得分阈值可通过 $t_g = m_g + v_g$ 计算得出。最后, 在第 14 至 21 行中, 挑选出 HM 得分大于阈值 t_g 的候选正样本作为最终的正样本。任何被判定为正样本的锚框都将从负样本集合中移除, 而剩余所有既非正样本也非负样本的锚框则在当前训练步被忽略。

对于单个 GT, 较大的 m_g 表明该 GT 周围存在较多匹配程度高的锚框, 而较大的 v_g 表明候选样本质量差异大, 只有得分较高的少数锚框与该 GT 较匹配。因此用 $m_g + v_g$ 作为阈值, 可以在候选样本质量普遍较高时提高筛选标准, 在质量普遍较低时避免阈值过高导致无正样本。此外, 本章方法中正样本的选择与目标尺寸无关。对于标准正态分布而言, 理论上约有 16% 的样本值会大于 $m_g + v_g$ 。尽管候选正样本的 HM 得分并不严格服从标准正态分布, 但如图 3.9 所示, 这种基于统计的方法仍能平衡不同尺度目标之间的正样本数量。传统方法往往难以给微小目标分配足够的正样本, 同时又会给常规尺寸的目标分配过多的正样本, 这种差异通常会导致模型倾向于优先学习检测常规尺寸的目标。本章方法平衡了不同尺度目标的正样本数量, 从而增强了对微小目标的检测性能。

3.2 实验分析

3.2.1 实验设置

本章方法基于 PyTorch^[56] 和 MMDetection^[57] 实现。所有实验均在单张 NVIDIA RTX 3090 GPU 上进行, Batch Size 设置为 2。选择在 ImageNet^[58] 上预训练的 ResNet-50^[48] 作为主干网络。对于 AI-TOD 数据集, 输入图像分辨率设定为 800×800。对于 SODA-D 数据集, 原始图像被裁剪为 800×800 的图像块, 裁剪步长为 650 像素, 并在训练和测试时将其缩放至 1200×1200。对于 VisDrone 数据集, 图片预处理在保持原始长宽比的同时缩放输入图像, 将短边设置为 800 像素, 并确保长边不超过 1333 像素。所有模型均使用随机梯度下降优化器训练 12 个周期, 动量设置为 0.9, 权重衰减为 0.0001, 初始学习率为 0.005, 并在第 8 和第 11 个周期时按 0.1 的比例进行衰

减。锚框生成器遵循 RFLA^[35] 的设置, RPN 候选框数量设定为 3000。在推理阶段, 非极大值抑制前的边界框数量同样设置为 3000, 以应对密集预测场景。其他参数设置均遵循 MMDetection 的默认值, 除非另有说明, 所有实验均采用上述配置。

3.2.2 对比实验

本小节分析了本章方法与其他最先进的方法在 AI-TOD^[2]、SODA-D^[53]、VisDrone-DET2019^[3] 以及 AgriPest^[59] 数据集上的对比实验。上述四个数据集对微小目标的尺寸定义并不一致, 例如, 在 AI-TOD 数据集中, 小于 256 像素的目标被视为微小目标; 而在 SODA-D 中, 这一阈值为 1024 像素。为了确保公平对比, 本章实验遵循各数据集原始论文中指定的评价指标来评估本章方法的性能。在这四个数据集中, AI-TOD 的目标平均尺寸最小, 挑战性也最大。因此, 本章实验主要在 AI-TOD 数据集上对所提方法进行评估。

3.2.2.1 AI-TOD 数据集上的对比实验

表 3.1 本章方法与其他先进方法在 AI-TOD 数据集上的对比实验

方法	主干网络	训练周期	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	参数量 (M)
RetinaNet ^[60]	ResNet-50	1×	8.7	22.3	4.8	2.4	8.9	12.2	16.0	36.25
Faster-RCNN ^[51]	ResNet-50	1×	11.1	26.3	7.6	0.0	7.2	23.3	33.6	41.16
AutoAssign ^[61]	ResNet-50	1×	12.2	30.4	8.1	2.3	12.2	17.2	25.0	35.98
ATSS ^[31]	ResNet-50	1×	12.8	30.6	8.5	1.9	11.6	19.5	29.2	32.30
Cascade R-CNN ^[62]	ResNet-50	1×	13.8	30.8	10.5	0.0	10.6	25.5	36.6	68.95
M-CenterNet ^[2]	DLA-34	1×	14.5	40.7	6.4	6.1	15.0	19.4	20.4	-
DotD ^[6]	ResNet-50	1×	16.1	39.2	10.6	8.3	17.6	18.1	22.1	-
FSANet ^[63]	ResNet-50	1×	16.3	41.4	9.8	4.4	14.6	23.4	33.3	31.9
Faster R-CNN w/ NWD ^[34]	ResNet-50	1×	17.8	43.8	11.0	2.5	17.0	26.1	34.3	41.16
KLDet ^[36]	ResNet-50	1×	19.6	-	-	-	-	-	-	32.00
ADAS-GPM ^[64]	ResNet-50	1×	20.1	49.7	12.4	7.4	19.8	24.9	32.1	-
MENet ^[24]	ResNet-50	1×	20.4	50.0	12.9	8.9	21.4	23.2	31.0	41.95
Faster R-CNN w/ RFLA ^[35]	ResNet-50	1×	21.1	51.6	13.1	9.5	21.2	26.1	31.5	41.16
本章方法	ResNet-50	1×	23.4	55.5	15.5	10.2	24.2	27.3	33.0	42.85
FSANet ^[63]	Swin-T	2×	22.6	52.8	15.6	7.4	21.6	29.1	38.5	34.63
MENet ^[24]	Swin-T	2×	23.2	56.2	15.0	9.7	23.9	25.3	34.4	44.68
KLDet ^[36]	ResNet-50	2×	25.2	-	-	-	-	-	-	32.00
DNTR ^[26]	ResNet-50	3×	26.2	56.7	20.2	12.8	26.4	31.0	37.0	42.6
CAF ² ENet-S ^[25]	DarkNet-53	300e	28.1	61.4	21.8	12.4	27.9	34.9	39.2	11.32
DFEM ^[65]	DFEM	-	28.4	59.1	21.3	11.0	26.2	32.1	40.6	8.2
本章方法 *	ResNet-50	2×	29.1	60.2	23.9	12.3	29.1	34.1	41.4	42.85

训练周期表示检测器训练的周期数, 1× 表示训练 12 个周期. 300e 表示训练 300 个周期.

* 表示将 IFF 和 HALA 集成到 DetectoRS^[66] 中.

表 3.1 展示了本章方法与 AI-TOD 测试集上其他先进方法的对比结果。AI-TOD 中的实例根据其绝对像素面积被划分为四个子尺度，分别为极微小 (vt)、微小 (t)、较小 (s) 和中等 (m)。通过将 IFF 和 HALA 集成到 Faster R-CNN^[51] 中，本章方法达到了 23.4% 的平均 AP，在相同的主干网络和训练策略下优于大多数现有方法。此外，当与 DetectoRS^[66] 集成并训练 24 个周期后，本章方法展现出了更先进的性能，平均 AP 达到 29.1%，比 CAF²ENet-S 高出 1%。本章方法在特定指标上表现出优势，尤其是在具有挑战性的 AP_{vt} 和 AP_t 指标上。此外，在同样基于 Faster R-CNN 改进的情况下，本章方法在 AP 指标上超过了专门针对微小目标检测的方法 RFLA^[35]，在 AP_{vt} 上高出 0.7%，在 AP_t 上高出 3.0%。大多数微小目标检测方法通常会牺牲中等尺寸目标的检测性能， AP_m 很少能超过 Cascade R-CNN 的 36.6%。而 HALA 中引入的 IoU 约束使得 AP_m 较其他方法有小幅提升，例如相比 RFLA 提升了 1.5%。AI-TOD 的一些典型检测结果可视化如图 3.5 所示。图 3.6 的第一行展示了本章方法与其他两种模型在 AI-TOD 数据集上的检测结果对比，图中所有目标的大小均小于 256 像素。Faster R-CNN 难以识别这些微小目标；RFLA 虽比 Faster R-CNN 具有更高的召回率，但往往会针对单个目标生成多个重复的边界框。相比之下，本章方法成功检测到了图中所有目标，且结果更加精确。

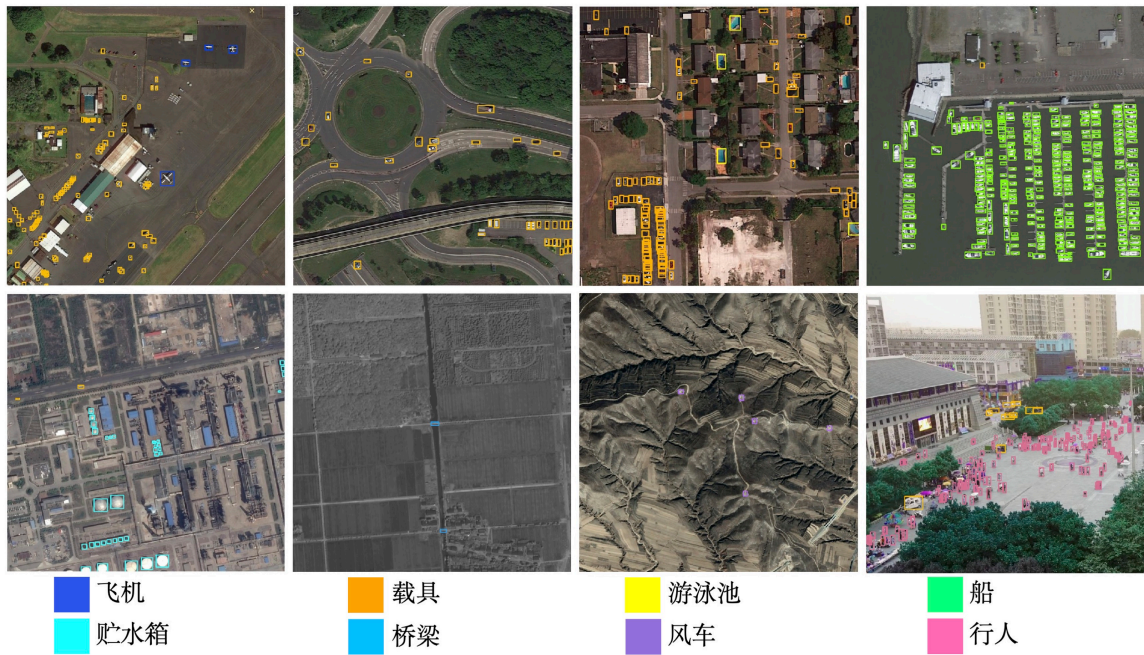


图 3.5 本章方法在 AI-TOD 数据集上的检测结果可视化

3.2.2.2 SODA-D 数据集上的对比实验

表 3.2 本章方法与其他先进方法在 SODA-D 数据集上的对比实验

方法	训练周期	AP	AP ₅₀	AP ₇₅	AP _{es}	AP _{rs}	AP _{gs}	AP _N
Deformable-DETR ^[67]	50e	19.2	44.8	13.7	6.3	15.4	24.9	34.2
CenterNet ^[68]	70e	21.5	48.8	15.6	5.1	16.2	29.6	42.4
FCOS ^[50]	1×	23.9	49.5	19.9	6.9	19.4	30.9	40.9
Sparse R-CNN ^[69]	1×	24.2	50.3	20.3	8.8	20.4	30.2	39.4
CornerNet ^[70]	2×	24.6	49.5	21.7	6.5	20.5	32.2	43.8
KLDet ^[36]	1×	25.9	53.8	21.4	10.7	22.2	31.9	41.6
YOLOX ^[50]	70e	26.7	53.4	23.0	13.6	25.1	30.9	30.4
ATSS ^[31]	1×	26.8	55.6	22.1	11.7	23.9	32.2	41.3
ESOD ^[43]	-	27.0	55.4	24.1	9.7	23.0	34.1	43.8
DyHead ^[71]	1×	27.5	56.1	23.2	12.4	24.4	33.0	41.9
RepPoints ^[72]	1×	28.0	55.6	24.7	10.1	23.8	35.1	45.3
RetinaNet ^[60]	1×	28.2	57.6	23.7	11.9	25.2	34.1	44.2
Faster R-CNN ^[51]	1×	28.9	59.4	24.1	13.8	25.7	34.5	43.0
Cascade R-CNN ^[62]	1×	29.1	56.5	25.9	12.5	25.5	35.4	44.7
Faster-RCNN w/ RFLA ^[35]	1×	29.7	60.2	25.2	13.2	26.9	35.4	44.6
CFINet ^[73]	1×	30.7	60.8	26.7	14.7	27.8	36.4	44.6
本章方法	1×	31.1	61.4	27.7	15.0	27.3	37.4	47.0

表 3.2 展示了本章方法与几种主流方法在 SODA-D 测试集上的对比结果。通过将 IFF 和 HALA 集成到 Faster R-CNN^[51] 中，本章方法在整体检测性能上表现优异，取得了 31.1% 的平均 AP 值，超越了包括 Deformable-DETR、Sparse R-CNN 以及专门针对微小目标优化的 CFINet 和 Faster-RCNN w/ RFLA 等先进方法。特别值得注意的是，在衡量极微小目标和普通微小目标检测能力的 AP_{es} 和 AP_{gs} 指标上，本章方法分别达到了 15.0% 和 37.4% 的最好结果，这充分验证了 IFF 在缓解特征不对齐问题以及 HALA 在挖掘高质量微小目标正样本方面的优势。此外，本章方法的 AP₅₀ 达到 61.4%，AP₇₅ 达到 27.7%，证明该方法不仅能有效召回微小目标，还能实现高精度的边界框回归。尽管在 AP_{rs} 指标上以 0.5% 的微小差距次于 CFINet，但综合来看，本章方法在微小目标检测的精度和鲁棒性上均展现出了先进的水平。

表 3.3 本章方法与其他先进方法在 VisDrone 数据集上的对比实验

方法	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster R-CNN ^[51]	21.4	40.7	19.9	11.7	33.9	54.7
Cascade R-CNN ^[62]	23.2	39.9	23.4	16.5	36.8	39.4
Yolov5 ^[74]	24.1	44.1	24.7	15.3	35.6	38.4
RetinaNet ^[60]	24.3	44.3	18.7	16.5	36.8	39.4
ClusDet ^[18]	26.7	50.6	24.7	17.6	38.9	51.4
DetectoRS ^[66]	26.8	43.2	28.0	17.5	38.2	41.7
CEASC ^[75]	28.7	50.7	28.4	-	-	-
CFPT ^[76]	29.7	50.0	30.4	19.7	41.9	48.0
SDP ^[77]	30.2	52.5	30.6	22.6	39.6	39.8
DCTDdet ^[78]	30.5	54.4	29.6	22.7	41.1	55.3
AMRNet ^[79]	31.7	52.7	33.1	23.0	43.4	58.1
DNTR ^[26]	33.1	53.8	34.8	-	-	-
CZDet ^[80]	33.2	58.3	33.2	26.1	42.6	43.4
本章方法 *	34.3	57.0	35.1	24.3	45.9	51.8

* 表示将 IFF 和 HALA 集成到 DetectoRS^[66]中.

3.2.2.3 VisDrone 数据集上的对比实验

为了进一步验证所提方法在复杂航拍和尺度变化剧烈场景下的泛化能力与鲁棒性,表 3.3 展示了本章方法与现有主流目标检测算法在 VisDrone 数据集上的对比结果,评估工具及数据集设置均遵循数据集官方指南^[81]。通过将 IFF 和 HALA 集成到 DetectoRS^[66] 中,本章方法实取得了 34.3% 的平均 AP,超越了 Faster R-CNN、Yolov5 等经典检测方法以及多种专门针对无人机或小目标场景设计的 AMRNet、DNTR 和 CZDet。与同样基于 DetectoRS^[66] 实现的 DNTR 相比,本章方法的平均 AP 高出 1.2%。尽管由于实验中未采用任何图像裁剪策略,导致本章方法的 AP₅₀ 略低于 CZDet,但其平均 AP 和 AP₇₅ 仍具优势,分别超过 CZDet 1.1% 和 1.9%。总体而言,得益于 HALA 对正样本的有效平衡与挖掘,本章方法在应对 VisDrone 数据集中目标尺度差异大、分布密集等挑战时,展现出了更优的尺度适应性和先进的综合检测能力。

3.2.2.4 AgriPest 数据集上的对比实验

为了评估本章方法在农业场景下检测微小害虫的有效性,本小节在 AgriPest^[59] 数据集上进行了实验。实验结果如表 3.4 所示。在与众多经典算法以及先进检测方法相比,本章方法展现出了优越的综合性能。具体而言,本章方法取得了全场最高的

表 3.4 本章方法与其他先进方法在 AgriPest 数据集上的对比实验

Method	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster RCNN ^[51]	37.1	75.0	32.2	19.4	36.4	41.0
Cascade RCNN ^[62]	37.8	74.8	33.7	19.4	37.2	40.5
RetinaNet ^[60]	36.2	73.1	31.0	17.6	34.6	38.0
FCOS ^[50]	35.1	73.3	27.7	17.8	34.6	39.5
YOLOv7 ^[40]	38.9	77.5	34.1	22.1	37.2	43.3
YOLOv8 ^[52]	39.1	77.9	34.5	22.8	37.6	43.8
DINO ^[82]	36.8	76.6	30.7	22.0	36.1	40.8
Align DETR ^[83]	39.4	78.2	35.8	22.1	39.0	42.3
本章方法 *	39.8	78.7	36.9	23.6	39.5	41.3

* 表示将 IFF 和 HALA 集成到 DetectoRS^[66]中.

39.8% 的平均 AP。在定位精度方面，无论是较为宽松的 AP₅₀ 还是更为严苛的 AP₇₅，本章方法均取得最好效果，分别达到 78.7% 和 36.9%，这表明该模型能够生成精准的边界框，满足农业害虫精准识别的需求。更为重要的是，针对该数据集中大量存在的微小目标，本章方法在 AP_S 指标上达到了 23.6% 的最佳水平。与强大的通用目标检测方法 YOLOv8 相比，本章方法在 AP_S 和 AP_M 指标上也展现出优势，分别提升了 0.8% 和 1.9%。此外，与基于 Transformer 的先进模型 DINO 和 Align DETR 相比，本章方法依然具有竞争力，平均 AP 超过了 Align DETR 0.4%。

3.2.2.5 对比实验小结

为了进一步验证所提方法的有效性，本小节在四个代表性数据集上进行了检测结果可视化。如图 3.6 所示，本章方法在各种场景下始终能产生更准确、更稳定的预测结果。与 Faster R-CNN 和 RFLA 相比，本章方法展现出更清晰的目标边界、更少的误检，以及在遮挡和复杂背景等挑战性情况下的更强鲁棒性。这些视觉结果与定量评价指标相一致，进一步证实了本章方法在不同数据集上的泛化能力和有效性。

对于包含小于 64 像素目标（每个目标占图像面积不足 0.01%）的 AI-TOD 数据集，现有方法往往难以分配充足的高质量正样本，导致结果指标较低。在这一极具挑战性的场景下，本章方法取得了极具竞争力的性能。对于 SODA-D、VisDrone 和 AgriPest 数据集，由于大多数目标所占图像面积仍小于 0.1%，构成了巨大的检测挑战，很少有方法能取得显著效果。在这些数据集上，本章方法在 AP_{es} 和 AP_S 等指标

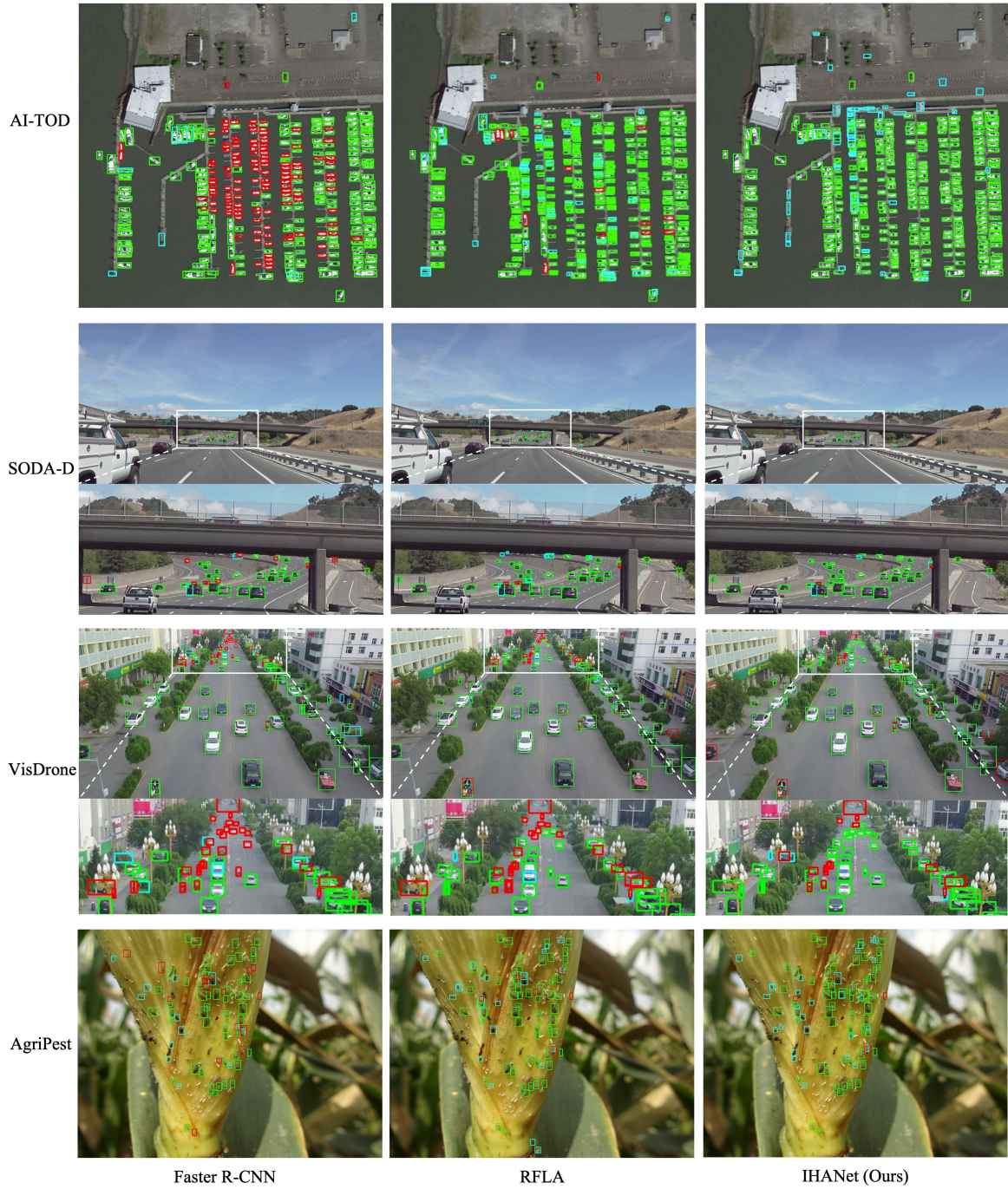


图 3.6 不同模型在 AI-TOD 数据集上的检测性能对比。绿色、红色和蓝色框分别代表正确检测 (TP)、漏检 (FN) 和误检 (FP)。

上表现出色。

3.2.3 消融实验

本小节在 AI-TOD^[2] 测试集上进行了消融实验，以验证各组件的有效性并确定本章方法的最佳配置。在接下来的实验中，基线方法由配置了 RFLA^[35] 先验生成器

的 Faster R-CNN 组成。为了确保公平对比，除所对比的参数外，所有配置下的其他参数均保持一致。

3.2.3.1 本章方法不同模块的消融实验

表 3.5 不同模块的消融实验结果

IFF	HALA	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	Params(M)	GFLOPs
		20.3	49.1	13.4	4.9	21.1	27.0	32.7	41.16	134.42
✓		22.1	53.5	14.8	5.8	22.9	28.5	33.5	42.85	179.94
	✓	21.2	51.1	13.6	9.6	21.5	25.8	32.7	41.16	134.42
✓	✓	23.4	55.5	15.5	10.2	24.2	27.3	33.0	42.85	179.94

为了深入验证本章所提出的 IFF 与 HALA 的有效性及其对网络整体性能的具体贡献，本小节在相同的实验设置下开展了详细的消融实验，结果如表 3.5 所示。当单独引入 IFF 时，模型的 AP 提升至 22.1%，且在 AP_s 和 AP_m 上均达到了最高值。这表明 IFF 能够有效解决多尺度融合中的特征不对齐问题，增强了特征表示能力，尽管这也带来了少量参数量和计算量的增加。另一方面，当单独引入 HALA 时，模型在 AP_{vt} 上的表现迎来了跃升，从 4.9% 提升至 9.6%，这归功于其在训练过程中能够分配更高质量的正样本，如图 3.4 所示。特别值得强调的是，由于 HALA 仅在训练阶段优化正负样本的匹配过程，它能够在不增加任何推理参数和计算开销的前提下实现性能的提升。最后，当同时集成 IFF 和 HALA 两个模块时，网络性能达到了最佳，平均 AP 达到了 23.4%，并在 AP₅₀、AP₇₅、AP_{vt} 以及 AP_t 等关键指标上均取得了最佳性能。这一结果充分证明了 IFF 与 HALA 之间存在高度的互补性，IFF 为微小目标提供了更具判别性的高质量特征，而 HALA 则通过挖掘高质量的正样本来增大这些特征的训练收益，两者协同工作，构建出了一个高效且鲁棒的微小目标检测架构。

3.2.3.2 IFF 中不同输出特征图的消融实验

为了探究 IFF 在不同金字塔层级输出融合特征图对检测性能的具体影响，并确定最优的网络配置，表 3.6 展示了针对不同融合层级组合的消融实验结果。实验数据表明，相较于基线方法，仅在最底层引入融合特征 U_2 就能带来显著的性能跃升，特别是 AP_{vt} 从 4.9% 大幅增长至 9.6%，这印证了底层高分辨率特征对于微小目标定位的核心价值。当同时输出 U_2 和 U_3 两个层级的融合特征时，网络整体达到了最佳性

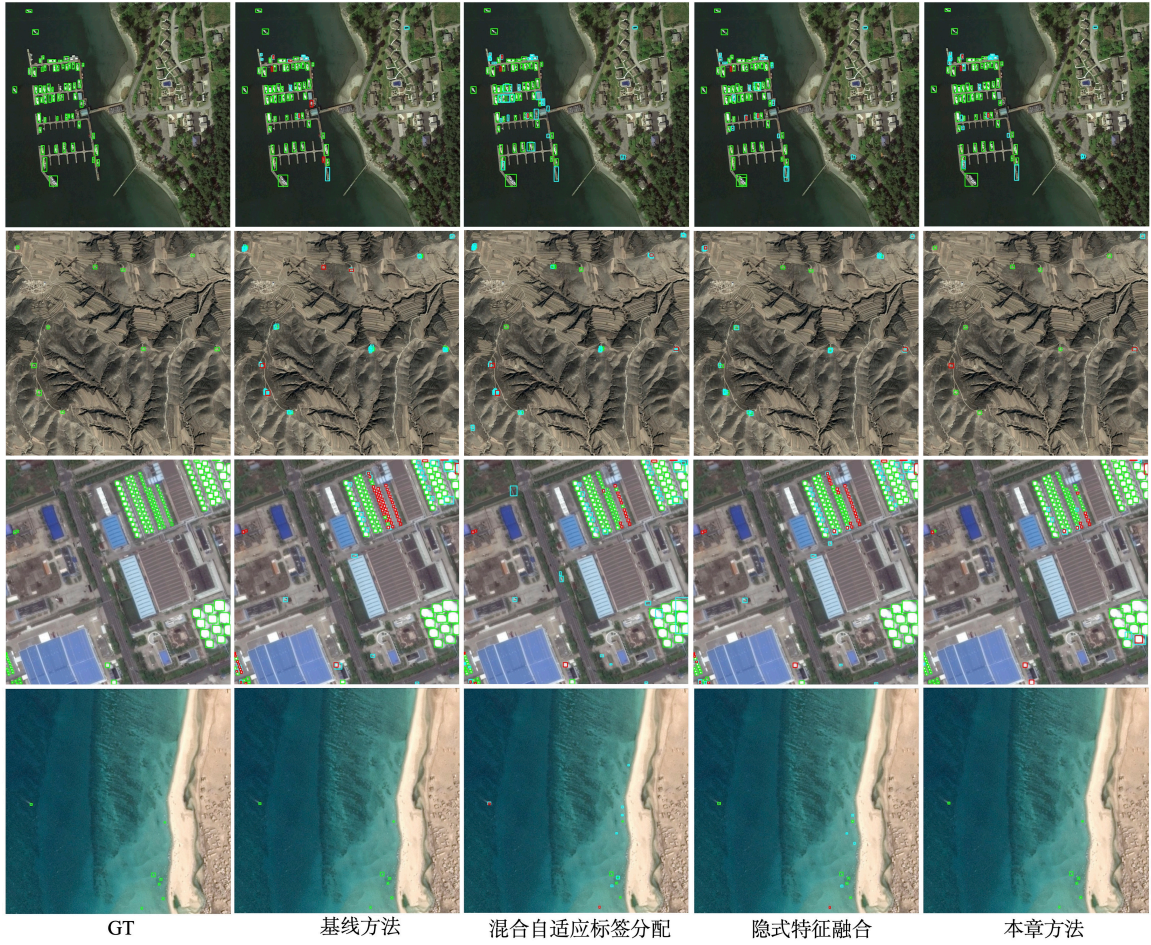


图 3.7 本章方法在 AI-TOD 数据集上的消融实验结果可视化。绿色、红色和蓝色框分别代表正确检测 (TP)、漏检 (FN) 和误检 (FP)。

表 3.6 IFF 输出不同融合特征图的消融实验结果

方法	输出融合特征	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	Params(M)	GFLOPs
基线方法	-	20.3	49.1	13.4	4.9	21.1	27.0	32.7	41.16	134.42
	U_2	23.0	55.3	15.1	9.6	24.0	26.8	33.4	42.11	172.47
IFF	U_2, U_3	23.4	55.5	15.5	10.2	24.2	27.3	33.0	42.85	179.94
	U_2, U_3, U_4	22.6	54.7	15.0	8.9	23.2	26.1	33.5	43.40	181.29

能, 不仅平均 AP 取得全场最高的 23.4%, 而且在 AP_{vt} 、 AP_t 和 AP_s 等代表小尺度目标的关键指标上均表现出最佳水平, 在精度与计算开销之间取得了极佳的平衡。然而, 当进一步计算并引入更高层级的融合特征 U_2, U_3, U_4 时, 尽管 AP_m 获得了微弱提升, 但综合性能以及微小目标的检测精度反而出现了明显回落, 例如 AP_{vt} 降至 8.9%, 同时参数量和计算量也进一步增加。这一现象有力地验证了本章前文提出的观点, 对于像素信息本就极其有限的微小目标而言, 过度计算和融合高层特征不仅对检测增

益极其有限，反而可能引入冗余的背景噪声或引发特征冲突。因此，实验结果充分证明了按需选择 U_2 和 U_3 作为融合层输出，是实现高效微小目标检测的最佳策略。

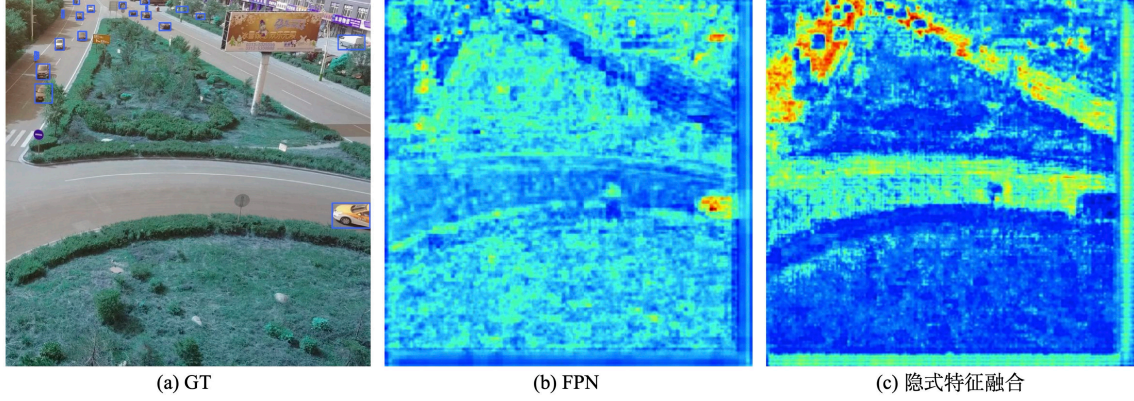


图 3.8 IFF 融合特征图的可视化结果

为了更好地展示 IFF 的效果，本小节对 FPN 输出的底层特征图 P_2 进行了可视化。可视化图是通过沿通道维度对特征取平均值生成的，其中明亮区域表示存在目标。如图 3.8 所示，在不使用 IFF 的情况下，模型的注意力局限于大目标区域；相比之下，引入 IFF 后，模型更侧重于微小目标所在的位置。

3.2.3.3 IFF 中不同融合层级的消融实验

表 3.7 IFF 中不同融合层级的消融实验结果

m	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	Params(M)	GFLOPs
2	23.0	55.3	14.8	10.8	23.4	27.1	33.4	42.24	161.55
3	23.3	56.0	15.2	9.9	24.1	27.3	34.0	42.65	171.77
4	23.4	55.5	15.5	10.2	24.2	27.3	33.0	42.85	179.94

为了探索不同融合层级的影响，本小节将 IFF 从全局融合调整为窗口化融合。与图 3.3 中利用特定金字塔层级以上的所有特征图进行融合的方法不同，本小节采用大小为 m 的滑动窗口来选择 FPN 特征金字塔，输出的融合特征图 U_i 通过融合 $\{C_i, C_{i+1}, \dots, C_{i+m-1}\}$ 获得。如表 3.7 所示，不同 m 取值在处理不同尺度的目标时展现出了各自的优势。当 $m = 2$ 时，即仅融合相邻两层特征，模型在 AP_{vt} 上取得了 10.8% 的最高精度，且具有最少的参数量和计算开销。这表明较小的融合跨度能够有效避免高层强语义信息对底层细粒度空间特征的负面影响，从而更有利于极端微

小特征的保留。然而，随着融合层级 m 的增加，模型能够捕获更广阔的感受野和更丰富的多尺度上下文信息。当 $m = 4$ 时，网络达到了 23.4% 的最佳平均 AP，并且在 AP_{75} 、 AP_t 和 AP_s 上均取得了最优或并列最优的表现。尽管 $m = 4$ 相比 $m = 2$ 略微牺牲了一点极小目标的精度，并引入了少量的额外计算开销，但它整体上实现了最为全面的性能提升。综合考量模型在多尺度小目标检测上的泛化能力与鲁棒性，本章最终采用 $m = 4$ 作为 IFF 模块的标准配置。

3.2.3.4 IFF 中空间编码的消融实验

表 3.8 IFF 中空间编码的消融实验结果。

空间编码	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	Params(M)	GFLOPs
	23.2	54.6	15.5	9.9	24.0	27.3	33.2	42.81	178.77
✓	23.4	55.5	15.5	10.2	24.2	27.3	33.0	42.85	179.94

为了验证 IFF 中引入空间编码对微小目标检测性能的具体贡献，表 3.8 展示了相关的消融实验结果。从实验数据可以清晰地看出，引入空间编码后，模型的平均 AP 从 23.2% 提升至 23.4%，且在 AP_{50} 这一指标上获得了 0.9% 的显著增长，从 54.6% 提升至 55.5%。更为关键的是，空间编码的加入促进了极小目标和微小目标的检测表现，使得 AP_{vt} 从 9.9% 提升至 10.2%， AP_t 从 24.0% 提升至 24.2%。这一结果充分印证了本章前文的理论分析，空间编码能够为隐式映射流程提供相对位置关系与高频先验信息，从而在多尺度特征对齐和重构过程中有效保留并引入更精细的局部细节。对于像素特征极其微弱的微小目标而言，这种细节的保留至关重要。尽管引入该机制导致 AP_m 出现了 0.2% 的微幅回落，并且仅增加了 0.04M 的参数和 1.17 GFLOPs 的计算开销，但综合考量其在微小目标核心指标上的稳定提升，将空间编码作为 IFF 的标准组件是必要且具有性价比的。

3.2.3.5 HALA 中不同阈值策略的消融实验

为了探究 HALA 中自适应阈值的计算策略对最终检测性能的具体影响，表 3.9 展示了采用不同统计特征组合作为筛选阈值的消融实验结果。HALA 的核心在于利用候选样本得分的统计分布来动态确定正样本。为了验证本章方法提出的“均值 + 标准差”策略的有效性，本小节分别将“中位数 + 标准差”和“仅均值”作为阈值的两种替

表 3.9 HALA 中不同阈值策略的消融实验结果

方法	阈值策略	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
基线方法	-	20.3	49.1	13.4	4.9	21.1	27.0	32.7
	中位数 + 标准差	22.7	54.1	14.8	7.7	23.2	29.0	34.8
HALA	仅均值	23.1	55.1	15.4	8.8	23.6	27.6	33.3
	均值 + 标准差	23.4	55.5	15.5	10.2	24.2	27.3	33.0

代方案。实验数据显示，当采用“中位数 + 标准差”的策略时，模型倾向于为小目标和中等尺寸目标分配高质量正样本，具体表现为 AP_s 达到峰值 29.0%， AP_m 达到 34.8%，但代价是 AP_{vt} 大幅滑落至 7.7%。这主要是因为中位数在统计上抗噪能力强，容易忽略候选样本得分分布中的长尾变化，从而导致在极小目标候选样本得分普遍偏低且分布不均时，无法自适应地给出合理的阈值下限。当采用“仅均值”作为阈值时，虽然平均 AP 达到 23.1% 有所回升，但由于缺乏标准差，忽略了各个 GT 候选样本得分的差异程度，导致阈值缺乏针对个体 GT 的弹性调节能力，8.8% 的 AP_{vt} 仍未达到最优。相比之下，本章最终采用的“均值 + 标准差”策略完美实现了质量与数量的平衡。通过将均值与标准差相结合，该策略能够针对每一个 GT 动态收放匹配阈值，既保证了极小目标能够分配到足够数量的正样本，又严格移除了背景噪声严重的低质量锚框。最终，该策略不仅获得了 23.4% 的最高平均 AP，且在 AP_{vt} 、 AP_t 以及 AP_{75} 上均取得最优，充分证明了该自适应选择机制的科学性与有效性。

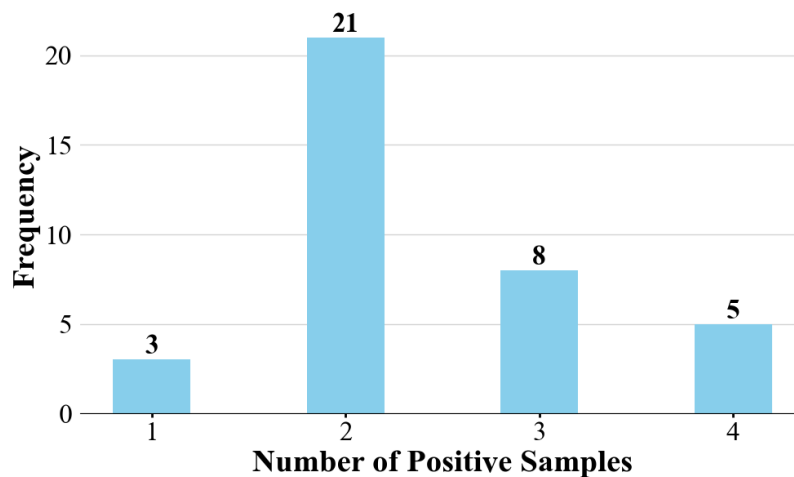


图 3.9 单张图像中不同实例最终正样本数量的可视化

在实验中，当 K 设为 12 时，如图 3.9 所示，无论候选正样本的质量如何，HALA 都能确保大多数实例被分配到 2 到 4 个正样本。与分配固定数量样本、可能引入低

质量正样本的传统 Top-K 策略不同, HALA 能够自适应地选择高质量样本, 从而提高训练的稳定性。

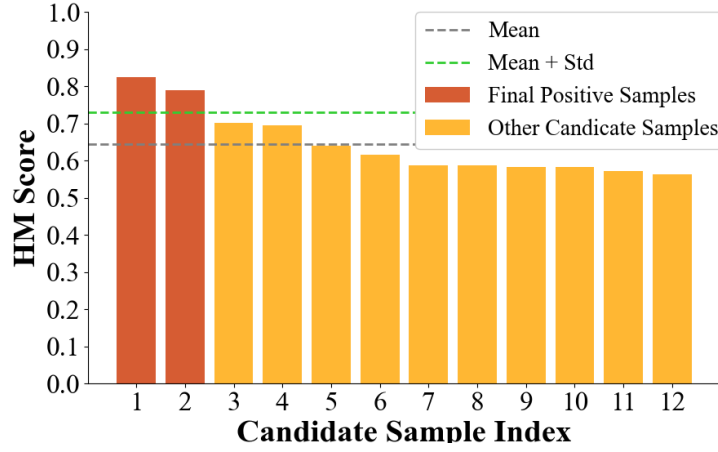
3.2.3.6 HALA 中不同 $\mathcal{K}$ 值的消融实验

表 3.10 不同 HALA 中不同 $\mathcal{K}$ 值的消融实验结果

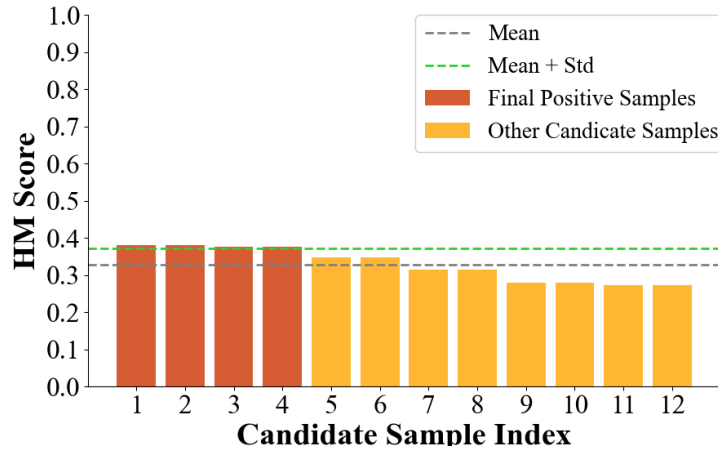
方法	$\mathcal{K}$	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
基线方法	-	20.3	49.1	13.4	4.9	21.1	27.0	32.7
	9	22.8	55.9	14.1	9.3	22.4	27.8	35.8
	12	23.4	55.5	15.5	10.2	24.2	27.3	33.0
	15	23.1	56.0	14.9	9.7	23.6	28.0	33.3
	18	22.8	55.0	14.3	10.5	23.3	28.3	33.7

为了探讨 HALA 策略中候选正样本数量 $\mathcal{K}$ 对检测性能的具体影响, 本小节进行了 HALA 中不同 $\mathcal{K}$ 值的消融实验, 表 3.10 中展示了相应的实验结果。在 HALA 的自适应阈值计算中, $\mathcal{K}$ 值直接决定了每个 GT 框初步筛选出的候选正样本规模。实验数据表明, 当 $\mathcal{K} = 12$ 时, 模型达到了最佳的性能平衡, 取得了 23.4% 的最高平均 AP, 并且在 AP₇₅ 以及 AP_t 上均表现出最优水平。当设定较小的候选数量如 $\mathcal{K} = 9$ 时, 虽然更为严苛的初步筛选保证了候选样本的高质量, 使得 AP_m 达到了峰值 35.8%, 但这种保守的策略容易导致微小目标分配不到足够的正样本, 从而限制了整体召回率。相反, 当放宽候选数量使如 $\mathcal{K} = 18$ 时, 模型被迫为 GT 匹配更多的锚框, 这使得微小目标获得了充足的正样本, AP_{vt} 和 AP_s 分别达到了 10.5% 和 28.3% 的峰值; 然而, 过大的 $\mathcal{K}$ 值也不可避免地将大量低质量的锚框纳入正样本之中, 直接导致了平均 AP 以及 AP₇₅ 的显著下滑。综合考量微小目标正样本数量的充足性与其质量的可靠性, 本章最终将 $\mathcal{K} = 12$ 设定为 HALA 的默认超参数。

图 3.10(a) 展示了一个候选正样本质量普遍较高的实例, 其中仅有两个候选样本超过了计算出的阈值 t_g 。因此, HALA 选择这两个高质量候选样本作为最终的正样本。相比之下, 图 3.10(b) 展示了一个候选正样本质量较低的实例, 在这种情况下, HALA 仍能识别出四个正样本用于训练, 从而确保了充足的训练信号并提升了召回率。



(a)



(b)

图 3.10 单张图像中两个实例候选正样本的 HM 得分分布可视化

3.2.3.7 HALA 中不同 α 值的消融实验

本小节探究了 HALA 中 HM 得分的平衡系数 α 对最终检测性能的影响，表 3.11 中展示了相关的消融实验结果。系数 α 的核心作用是调节 IoU 在 HM 得分中的权重，以此来弥补 RFD 在匹配时容易导致正样本锚框过小的问题。从实验结果可以清晰地看到，当 $\alpha = 0.9$ 时，模型达到了理想的性能平衡，不仅平均 AP 达到了最高的 23.4%，并且在 AP_{75} 、 AP_{vt} 和 AP_t 上均达到最高值。当 $\alpha = 0.8$ 时，模型更加依赖 RFD 进行匹配，这虽然使得 AP_s 达到了峰值 28.4%，但在极小目标上的表现和综合 AP 却受到了一定限制，这印证了前文所述的单纯依赖 RFD 会引入低质量正样本的缺陷。相反，当 $\alpha = 1.0$ 时，IoU 约束开始占据主导，这使得模型在 AP_m 、 AP_{50} 上表现更好，但却削弱了 RFD 对微小特征特异性匹配的优势，导致 AP_{vt} 回落至 9.4%。

表 3.11 不同 α 值的消融实验结果

方法	α	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
基线方法	-	20.3	49.1	13.4	4.9	21.1	27.0	32.7
	0.8	22.8	55.1	15.1	9.6	23.1	28.4	32.9
HALA	0.9	23.4	55.5	15.5	10.2	24.2	27.3	33.0
	1.0	23.2	55.7	15.1	9.4	23.6	27.8	33.4

综合以上分析，将 α 设置为 0.9 能够平衡 RFD 对微小目标的敏感性与 IoU 对目标的空间约束，从而最大化增益微小目标检测任务。

3.3 本章小结

本章针对卷积神经网络在微小目标检测中特征提取受限、融合不对齐与标签分配极度不平衡的痛点，提出一种基于隐式特征融合与混合自适应标签分配的高精度检测方法。IFF 引入隐式神经表示与空间编码，将离散特征建模为连续信号并映射至统一分辨率融合，有效保留微小目标的局部精细特征并减少背景噪声。混合自适应标签分配策略结合感受野距离与 IoU 构建混合指标，利用均值与标准差动态生成阈值进行筛选，在零额外推理成本下实现高质量正样本的精准挖掘与分配。在 AI-TOD 等四个复杂数据集上的对比与消融实验表明，该方法有效突破了微小目标的特征与样本瓶颈，在极端场景下展现出卓越的精度与泛化能力，整体性能显著优于主流先进算法。

第四章 基于稀疏编码器与尺度感知 Query 选择的实时微小目标检测

继第三章探索了卷积神经网络对微小目标检测任务的潜力后，本章将聚焦于基于 Transformer 的实时目标检测框架 RT-DETR^[44]。检测微小目标依赖浅层的高分辨率特征图，引入了大量计算开销。不同于上一章着重于提升为小目标检测的精度，本章工作旨在维持微小目标检测精度的同时，实现更高的模型推理速度并降低计算量和显存占用，提升微小目标检测模型部署在资源受限设备上的可行性。基于此，本章提出了一种基于稀疏编码器与尺度感知 Query 选择的实时微小目标检测方法，它结合了选择稀疏编码器、尺度感知 Query 选择机制与一对多匹配的辅助检测头，旨在实现高精度的实时微小目标检测。

4.1 方法概述

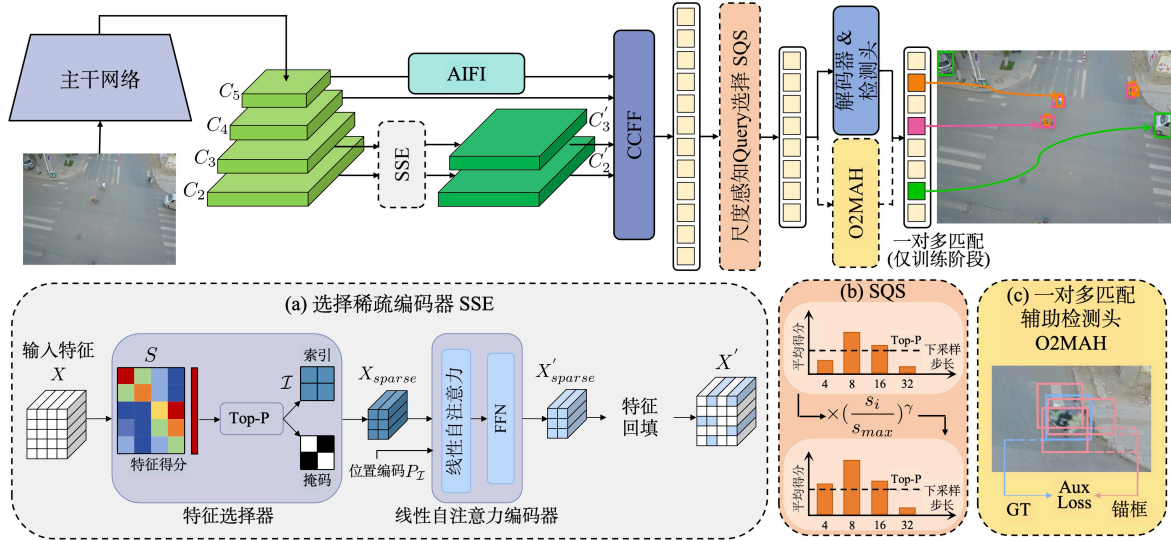


图 4.1 基于选择稀疏编码器与尺度感知 Query 选择的实时微小目标检测方法架构

图 4.1 展示了本章所提方法的整体框架。该方法以 RT-DETR 为基础框架，其整体流程始于主干网络对输入图像进行多尺度特征提取得到 $\{C_i\}_{i=2}^5$ ，随后将包含丰富空间细节但计算开销较大的浅层特征 C_2, C_3 输入选择稀疏编码器 (Selective Sparse Encoder, SSE)，该模块通过评估特征重要性并利用 Top-P 机制筛选出显著特征，配

合线性自注意力机制在保留微小目标细节信息的同时降低高分辨率特征建模的计算成本；紧接着，经过优化的特征图 C'_2, C'_3 进入 RT-DETR 的 CCFF 颈部网络进行特征融合，并由尺度感知 Query 选择机制（Scale Aware Query Selection, SQS）对候选 Query 的置信度得分进行动态调整，通过引入基于下采样步长的调节系数补偿微小目标特征在筛选中的得分劣势，确保高质量的针对微小目标的 Query 能顺利进入解码器进行精细化预测；最后，在训练阶段引入基于锚框的一对多匹配辅助检测头（One to Many Auxiliary Head, O2MAH），为每个微小目标 GT 分配多个正样本锚框来强化监督信号并加速模型收敛，而推理时则仅保留主检测头的一对一预测，从而在不增加部署负担的前提下实现了精准、实时的微小目标检测。本章后续将对提出的选择稀疏编码器、尺度感知 Query 选择机制以及一对多匹配辅助检测头分别进行介绍。

4.1.1 选择稀疏编码器 SSE

针对微小目标检测任务，高分辨率特征图（ C_2, C_3 ）包含的关键空间细节对于捕捉极小尺度目标的几何特征至关重要。然而，传统的 Transformer 编码器由于其自注意力机制的计算复杂度随序列长度 N 呈平方级增长（即 $O(N^2)$ ）^[84]，导致直接处理高分辨率特征图会消耗极高的计算资源与显存。为解决这一矛盾，本章提出了选择稀疏编码器 SSE。

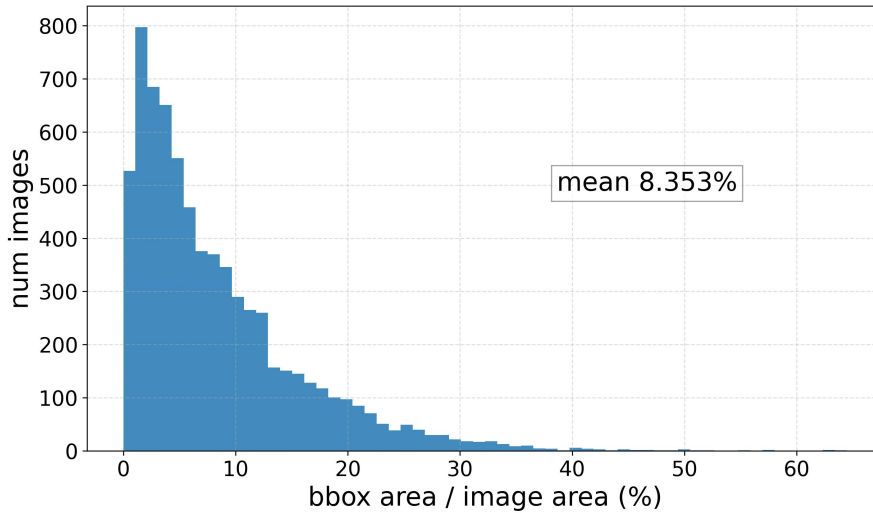


图 4.2 VisDrone 数据集中目标区域占比分布

显著特征选择 如图4.2所示, 对于 VisDrone^[3]数据集, 每张图片中目标区域平均占比为 8.353%, 因此对特征图进行全局优化会造成大量的计算资源浪费。基于上述发现, SSE 首先利用一个特征选择器, 通过简单的特征通道维度均值为输入特征图生成显著得分图:

$$S = \text{Mean}_{\text{Channel}}(X) \quad (4.1)$$

其中 $X \in \mathbb{R}^{H \times W \times C}$ 表示输入的原始高分辨率特征图, H, W, C 分别代表特征图高度、宽度和通道数, $S \in \mathbb{R}^{H \times W \times 1}$ 表示输出的显著得分图, $\text{Mean}_{\text{Channel}}$ 表示通道维度均值操作。随后, 根据得分图 S 提取稀疏显著特征向量集 X_{sparse} :

$$\mathcal{I} = \text{Top-P}(S, p) \quad (4.2)$$

$$X_{\text{sparse}} = \{X_i \mid i \in \mathcal{I}\} \quad (4.3)$$

其中 $\text{Top-P}(S, p)$ 表示选择 S 中得分前 $p\%$ 的索引, p 为选择显著特征的比例, X_i 表示特征图 X 中索引为 i , 通道数为 C 的特征向量。该特征选择方法通过 Top-P 机制近似的筛选潜在的显著区域, 本章在4.2.3.3对不同 p 进行了消融实验。

线性注意力编码器 在特征选择器选出原始特征图中的重要特征后, 本章方法对这些特征使用 Transformer 编码器进一步建模优化, 提升模型对微小目标特征的判别能力。由于 SSE 筛选出的显著特征在空间分布上具有稀疏性且打破了原始特征图的连续性, 为了使模型能够感知特征点间的空间几何关系, 在进入自注意力层前, 必须引入位置编码。将位置编码 $P \in \mathbb{R}^{k \times d}$ 与稀疏特征 X_{sparse} 进行融合, 生成带有位置信息的输入向量 Z :

$$Z = X_{\text{sparse}} + P_{\mathcal{I}} \quad (4.4)$$

其中 $P_{\mathcal{I}}$ 是根据索引集 $\mathcal{I}$ 从原始二维正弦位置编码中提取的对应位置编码向量。随后, 基于带有位置信息的 Z 生成 Q 、 K 和 V :

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \quad (4.5)$$

其中 W_Q, W_K, W_V 为自注意力层的可学习投影矩阵。为了尽可能减少 SSE 带来的额外计算, 本章方法使用了线性注意力机制, 它通过将计算复杂度从空间维度的平方

级降低至线性级，能够在保持全局建模能力的同时显著降低计算开销，其计算公式如下：

$$\text{LinearAttention}(Q, K, V) = \frac{\phi(Q)(\phi(K)^T V)}{\phi(Q)(\phi(K)^T \mathbf{1}_k)} \quad (4.6)$$

其中 $\phi(\cdot)$ 表示特征映射函数，通常采用 $\text{elu}(x) + 1$ ； $\mathbf{1}_k$ 表示长度为 k 的全 1 向量。

被选中的稀疏特征 $X_{\text{sparse}} \in \mathbb{R}^{k \times C}$ 加入位置编码后，送入线性自注意力编码器，利用线性自注意力层与前馈神经网络 FFN 完成进一步的精细建模特征：

$$Z' = \text{LayerNorm}(\text{Linear-Attention}(Q, K, V) + Z) \quad (4.7)$$

$$\hat{X}_{\text{sparse}} = \text{LayerNorm}(\text{FFN}(Z') + Z') \quad (4.8)$$

其中 $\hat{X}_{\text{sparse}}$ 表示为经过线性自注意力编码器后的增强稀疏特征； Z' 表示 Z 经过线性自注意力计算得到的中间结果。

特征回填 为了使增强后的特征能够进入后续的特征融合模块，必须将这些稀疏分布的特征向量映射回原始的二维空间 ($H \times W$)。本章将这一过程称为特征回填。

特征回填是一个基于索引的反向映射过程，如图4.1(a)所示。上文方法中已经获得了线性自注意力编码器输出的增强稀疏特征集合 $\hat{X}_{\text{sparse}} \in \mathbb{R}^{k \times C}$ ，其在原始特征图中对应的空间索引集合为 $\mathcal{I}$ 。回填操作生成的特征图 $X' \in \mathbb{R}^{H \times W \times C}$ 可表示为：

$$X' = \text{ScatterBack}(X, \hat{X}_{\text{sparse}}, \mathcal{I}) \quad (4.9)$$

其中 $\text{ScatterBack}(X, \hat{X}_{\text{sparse}}, \mathcal{I})$ 表示基于索引 $\mathcal{I}$ 用 $\hat{X}_{\text{sparse}}$ 替换 X 中对应特征。

SSE 通过特征选择与线性自注意力机制大幅降低了模型处理高分辨率特征图的计算开销，使得模型能够在相同的硬件约束下，支持对更高分辨率特征图的实时处理，有效地缓解了计算压力，并使模型注意力高度聚焦于微小目标可能存在的区域。

4.1.2 尺度感知 Query 选择策略 SQS

在 RT-DETR 等检测模型中，为了评估多尺度特征图上每个空间位置作为潜在目标候选 (Query) 的质量，模型需要利用编码器输出的高维特征来预测其包含前景目标的概率。Query 选择机制通常依据编码器输出的特征来预测分类置信度，并挑选 Top-K 个高分特征向量作为解码器的初始 Query。然而，对于微小目标而言，由于其

所占像素极少且特征表达较弱，其特征的置信度得分往往显著低于具有丰富纹理和显著结构的大中型目标。这种硬性筛选机制会导致大量用于检测微小目标的 Query 在进入解码器之前就被当作背景噪声过滤掉，造成漏检。

为了弥补这一缺陷，本章提出了尺度感知 Query 选择机制 SQS。该机制在 RT-DETR 原有的不确定性得分基础上，引入了一个基于特征层级下采样步长的调节系数，通过显式地干预不同尺度特征的得分分布，使得更多微小目标 Query 能够进入后续的解码器阶段。设第 l 层特征图的下采样步长为 s_l （例如 $s_l \in \{4, 8, 16, 32\}$ ）。第 l 层特征图对应的调节系数 λ_l 定义为：

$$\lambda_l = \left(\frac{s_{max}}{s_l} \right)^\gamma \quad (4.10)$$

其中 s_l 表示第 l 层特征图的下采样步长，它直接反映了该层特征的感受野大小与目标尺度的对应关系。步长越小，特征图分辨率越高，越倾向于包含微小目标信息。 s_{max} 表示当前特征金字塔中的最大下采样步长； γ 是控制调节幅度的超参数。

随后，将各层级的调节系数 λ_l 与对应层级的特征置信度得分进行逐元素相乘，得到调整后的尺度感知置信度得分 $\hat{C}_l$ ：

$$\hat{C}_l = C_l \cdot \lambda_l \quad (4.11)$$

其中 C_l 表示第 l 层特征图通道维度特征向量的置信度得分。

最后，SQS 依据调整后的得分 $\hat{C}_l$ 在全球范围内执行排序，选出前 k_Q 个最具潜力的 Query 作为输入解码器的初始 Query 集合：

$$\mathcal{Q}_{init} = \text{Top-K}(\{\hat{C}_1, \hat{C}_2, \dots, \hat{C}_N\}, k_Q) \quad (4.12)$$

如图4.3所示，通过这种基于特征层级下采样步长的重加权机制，SQS 有效且高效地平衡了不同尺度特征向量在 Query 选择中的得分差异。由于该操作仅依赖于固定的层级步长而无需计算预测框面积，其在提升微小目标精度的同时，几乎没有引入任何额外的计算开销。

4.1.3 一对多匹配辅助检测头 O2MAH

DETR 类模型通过引入匈牙利算法进行二分图匹配，成功实现了端到端的一对一目标检测，彻底摒弃了非极大值抑制等后处理步骤。然而，这种严格的一对一匹配

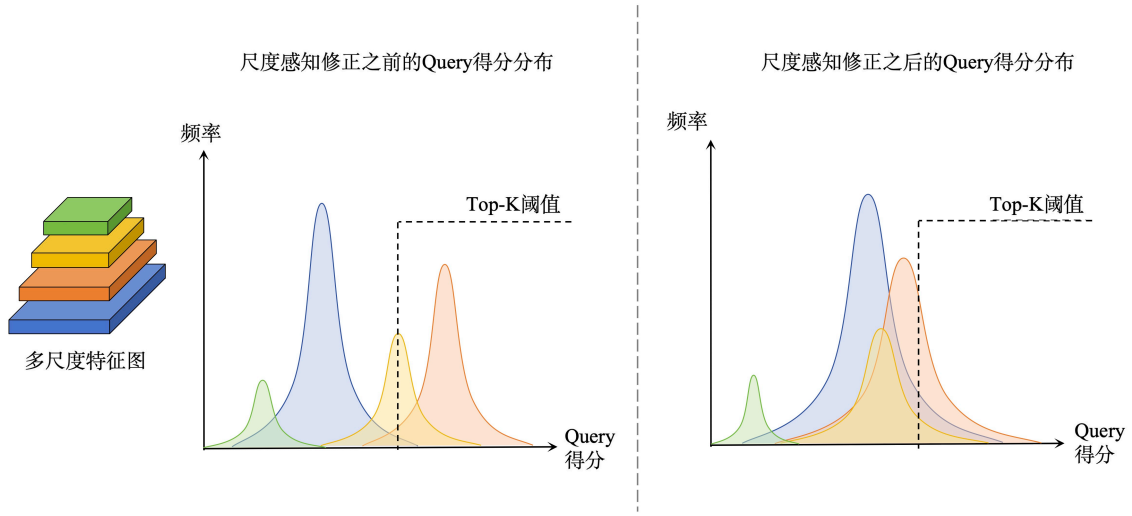


图 4.3 尺度感知 Query 选择效果示意图

机制在微小目标检测任务中暴露出显著的局限性，由于每个 GT 在每次迭代中只能被分配给一个预测 Query，导致微小目标能够获取的正样本监督信号极其稀疏，这种稀疏性不仅拖慢了模型的收敛速度，还容易导致网络在训练初期陷入局部最优，难以充分学习微小目标的特征表征。为了突破这一瓶颈，本章在模型中引入一对多匹配辅助检测头 O2MAH。该检测头在训练阶段与主检测头并行工作，通过提供密集的监督信号来加速模型收敛并提升微小目标的检测鲁棒性。

一对多分配策略 与主检测头采用的二分图匹配不同，O2MAH 放弃了严格的排他性分配，转而采用基于锚框的密集标签分配策略。对于每一个微小目标 GT 框 y_i ，O2MAH 依据锚框与 GT 框的 IoU，为其分配一个正样本集合 $\mathcal{P}_i$ ，集合中的锚框数量 $K_i > 1$ ，匹配过程可以用下式表示：

$$\mathcal{P}_i = \text{TopK}_{\hat{y}_j \in \hat{Y}_{aux}} \left(\text{IoU}(\hat{y}_j, y_i), K_i \right) \quad (4.13)$$

其中 $\hat{Y}_{aux}$ 表示 O2MAH 预定义的锚框集合，这些锚框仅参与辅助检测头的标签分配，该过程使得一个 GT 可以同时受到多个预测框的监督，显著增加了特征图上微小目标对应区域的激活频率。

辅助损失函数的构建 在训练过程中，总损失函数由主检测头的一对一损失与辅助检测头的一对多损失共同构成。对于主检测头，本章方法沿用了 RT-DETR 的标准损失函数设计。在通过匈牙利算法获得 GT 框与预测框的最优二分图匹配 σ 后，主损

失 $\mathcal{L}_{main}$ 由分类损失、边界框 L_1 损失和 GIoU^[85] 损失三部分加权求和而成：

$$\mathcal{L}_{cls}(c_i, \hat{p}_{\sigma(i)}) = -\log \hat{p}_{\sigma(i)}(c_i) \quad (4.14)$$

$$\mathcal{L}_{giou}(b_i, \hat{b}_{\sigma(i)}) = 1 - \text{GIoU}(b_i, \hat{b}_{\sigma(i)}) \quad (4.15)$$

$$\mathcal{L}_{main} = \sum_{i=1}^M [\lambda_{cls} \mathcal{L}_{cls}(c_i, \hat{p}_{\sigma(i)}) + \lambda_{L1} \|b_i - \hat{b}_{\sigma(i)}\|_1 + \lambda_{giou} \mathcal{L}_{giou}(b_i, \hat{b}_{\sigma(i)})] \quad (4.16)$$

其中 M 表示当前图像中 GT 的总数； $\sigma(i)$ 表示匈牙利匹配算法为第 i 个真实目标分配的最优预测 Query 索引； $\mathcal{L}_{cls}$ 表示分类损失，在 RT-DETR 架构中，为缓解正负样本极度不平衡的问题，通常采用 Focal Loss^[60] 或 Varifocal Loss^[86] (VFL)； c_i 为真实类别标签， $\hat{p}_{\sigma(i)}$ 为预测的分类置信度； $\|\cdot\|_1$ 为 L_1 表示范数损失，用于约束预测框 $\hat{b}_{\sigma(i)}$ 与 GT 框 b_i 在中心点坐标及宽高绝对数值上的差异； $\mathcal{L}_{giou}$ 表示广义交并比 GIoU 损失，用于衡量预测框与 GT 框在空间形状与重叠度上的尺度不变形误差。

对于 O2MAH，其损失计算采用与主检测头完全相同的损失组件，但计算范围扩展至所有被一对多分配策略选中的正样本集合 $\mathcal{P}_i$ ，一对多辅助损失 $\mathcal{L}_{aux}$ 定义为：

$$\mathcal{L}_{aux} = \sum_{i=1}^M \sum_{j \in \mathcal{P}_i} [\lambda_{cls} \mathcal{L}_{cls}(c_i, \hat{p}_j) + \lambda_{L1} \|b_i - \hat{b}_j\|_1 + \lambda_{giou} \mathcal{L}_{giou}(b_i, \hat{b}_j)] \quad (4.17)$$

最终，网络优化的总目标函数 $\mathcal{L}_{total}$ 定义为：

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \alpha \mathcal{L}_{aux} \quad (4.18)$$

其中 α 为平衡辅助监督信号强度的超参数（默认 $\alpha = 1.0$ ），以确保模型在密集强化微小目标特征学习的同时，不会破坏主检测头一对一匹配的收敛稳定性。

O2MAH 的另一个核心优势在于其仅在训练阶段启用，在模型完成训练进入推理部署阶段时，整个 O2MAH 分支将被完全剥离，网络仅保留执行一对一匹配的主检测头输出结果。这种设计使得本章方法既能在训练期间受到密集正样本的监督，又能在推理期保持端到端的无 NMS 极简架构，在不增加任何推理计算量与延迟的前提下，实现了微小目标检测精度的显著跃升。

4.2 实验分析

4.2.1 实验设置

本章方法基于 PyTorch^[56] 实现。所有实验均在单张 NVIDIA RTX 3090 GPU 上进行并开启混合精度训练，训练时 Batch Size 设置为 4，推理时 Batch Size 设置为 8。选择在 ImageNet^[58] 上预训练的 ResNet-18^[48] 作为主干网络。对于 VisDrone^[3] 数据集，输入图像分辨率默认设定为 640×640。对于 AI-TOD^[2] 数据集，输入图像分辨率设定为 800×800。本章方法相关模型均使用 AdamW 优化器训练 120 个周期，一阶动量、二阶动量权重分别设置为 0.9、0.999，权重衰减为 0.0001，初始学习率为 0.0001，在训练开始的前 2000 步进行线性热身训练，并在第 72 个周期时按 0.1 的比例进行衰减。除非另有说明，所有实验均采用上述配置。

4.2.2 对比实验

本小节分析了本章方法与其他先进检测方法在 VisDrone-DET2019^[3]、AI-TOD^[2] 数据集上的对比实验。

4.2.2.1 VisDrone 数据集上的对比实验

如表 4.1 所示，本章所提出的方法在 VisDrone 数据集上与当前主流的先进目标检测方法进行了量化对比。在低 GFLOPs 对比组中，以 ResNet18 为骨干网络的本章方法展现出了优越的精度与计算资源平衡能力，其凭借 21.7M 的参数量和 63.8 GFLOPs 的计算开销，不仅超越了最新的 YOLO 系列中型模型 YOLOv8/9/10/11-M，相较于基线模型 RT-DETR-R18 更是实现了 AP 和 AP₅₀ 分别高达 3.6% 和 6.6% 的提升。而在高 GFLOPs 对比组中，同样配备 ResNet18 主干网络并使用 C₂ 高分辨率特征图的本章方法实现了大幅的性能提升。其 33.4% 的 AP 与 54.4% 的 AP₅₀ 不仅超越了诸如 Deformable DETR、DetectoRS 等以 ResNet50 为主干的经典检测器以及 YOLOv11-X 这样的重型模型，也优于同类先进的 RT-DETR-R50 与 UAV-DETR-R50，且其 23.8M 的参数量仅为上述 R50 系列模型的一半左右。这一具有竞争力的实验结果证明，本章所引入的 SSE 与 SQS 通过高效过滤背景冗余并强化微小目标的检测精度，缓解了高分辨率特征处理与计算瓶颈之间的矛盾，配合 O2MAH 在训练期的密集监督，使得本章方法能够在较低的显存和计算负担下实现先进的微小目标检测精度。VisDrone

表 4.1 本章方法与其他先进方法在 VisDrone 数据集上的对比实验

模型	主干网络	出版	输出尺寸	Params (M)	GFLOPs	AP	AP ₅₀
YOLOv10-M ^[41]	CSPDarknet (SCDown)	arXiv2024	640 × 640	15.4	59.1	24.5	40.5
YOLOv11-M ^[87]	CSPDarknet (C3K2, C2PSA)	arXiv2024	640 × 640	20.0	67.7	25.9	43.1
YOLOv9-M ^[88]	CSPDarknet (GELAN, PGI)	ECCV2025	640 × 640	20.1	76.8	25.2	42.0
YOLOv8-M ^[52]	CSPDarknet (C2F)	-	640 × 640	25.9	78.9	24.6	40.7
RT-DETR-R18 ^[44]	ResNet18	CVPR2024	640 × 640	20.0	57.3	26.7	44.6
本章方法 *	ResNet18	-	640 × 640	21.7	63.8	28.6	48.6
YOLOv10-L ^[41]	CSPDarknet (SCDown)	arXiv2024	640 × 640	24.4	120.3	26.3	43.1
PP-YOLOE-P2-Alpha-1 ^[89]	CSPRepResNet	-	640 × 640	54.1	111.4	30.1	48.9
YOLOv8-L ^[52]	CSPDarknet (C2F)	-	640 × 640	43.7	165.2	26.1	42.7
DetectoRS w/ RFLA ^[35]	ResNet50	ECCV2022	800 × 800	123.2	160.0	27.4	45.3
DCFL ^[90]	ResNet50	CVPR2023	1024 × 1024	36.1	157.8	-	32.1
Deformable DETR ^[67]	ResNet50	ICLR2020	1333 × 800	40.0	173.0	27.1	42.2
DETR ^[91]	ResNet50	ECCV2020	1333 × 750	60.0	187.0	24.1	40.1
YOLOv11-X ^[87]	CSPDarknet (C3K2, C2PSA)	arXiv2024	640 × 640	56.8	194.5	28.1	45.6
QueryDet ^[19]	ResNet50	CVPR2022	2400 × 2400	33.9	212.0	28.3	48.1
ClusDet ^[18]	ResNet50	ICCV2019	1000 × 600	30.2	207.0	26.7	50.6
UAV-DETR-R50 ^[92]	ResNet50	arXiv2025	640 × 640	42.0	170.0	31.5	51.1
RT-DETR-R50 ^[44]	ResNet50	CVPR2024	640 × 640	42.0	129.9	28.4	47.0
SO-DETR ^[47]	ResNet50	-	640 × 640	44.4	161.4	31.5	51.5
本章方法	ResNet18	-	640 × 640	23.8	146.4	33.4	54.4

* 表示不使用 C_2 高分辨率特征图。

数据集上的一些典型检测结果如图 4.4 所示。

4.2.2.2 AI-TOD 数据集上的对比实验

如表 4.2 所示，为了进一步验证本章方法在极端微小尺度上的检测性能，本小节在专门针对微小目标检测的 AI-TOD 数据集上进行了量化对比。实验结果表明，尽管本章方法仅采用了轻量级的 ResNet-18 作为主干网络，其参数量仅为 23.8M，但其平均 AP 达到 27.1%，超越了表格中所有基于更深层 ResNet-50 骨干网络的先进模型。特别值得注意的是，在要求更为严格的高 IoU 阈值指标 AP₇₅ 上，本章方法达到了 21.3%，相较于次优模型实现了较大提升，这表明该模型不仅能找到微小目标，更具备较高的边界框定位精度。此外，在针对极微小目标的细分指标上，本章方法在 AP_{vt} 和 AP_t 上分别达到 14.7% 和 29.3%，这证明了 SQS 有效克服了极小目标特征在 Query 筛选中的得分劣势，同时配合一对多辅助检测头在训练时提供的密集监督。实验中本章方法在微小目标核心指标上的优势与极低的参数开销，展现出了该方法出色的部署潜力与实用价值。



图 4.4 本章方法在 VisDrone-2019 数据集上的检测结果可视化

4.2.3 消融实验

本小节在 VisDrone-2019^[3] 数据集上进行了消融实验，以验证各组件的有效性并确定本章方法的最佳配置。在接下来的实验中，基准模型为 RT-DETR。为了确保公平对比，除所对比的参数外，所有配置下的其他参数均保持一致。

表 4.2 本章方法与其他先进方法在 AI-TOD 数据集上的对比实验

方法	主干网络	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m	参数量 (M)
RetinaNet ^[60]	ResNet-50	8.7	22.3	4.8	2.4	8.9	12.2	16.0	36.3
Faster-RCNN ^[51]	ResNet-50	11.1	26.3	7.6	0.0	7.2	23.3	33.6	41.2
AutoAssign ^[61]	ResNet-50	12.2	30.4	8.1	2.3	12.2	17.2	25.0	36.0
ATSS ^[31]	ResNet-50	12.8	30.6	8.5	1.9	11.6	19.5	29.2	32.3
Cascade R-CNN ^[62]	ResNet-50	13.8	30.8	10.5	0.0	10.6	25.5	36.6	69.0
M-CenterNet ^[2]	DLA-34 ^[93]	14.5	40.7	6.4	6.1	15.0	19.4	20.4	-
DotD ^[6]	ResNet-50	16.1	39.2	10.6	8.3	17.6	18.1	22.1	-
FSANet ^[63]	Swin-T ^[94]	22.6	52.8	15.6	7.4	21.6	29.1	38.5	34.6
Faster R-CNN w/ NWD ^[34]	ResNet-50	17.8	43.8	11.0	2.5	17.0	26.1	34.3	41.2
ADAS-GPM ^[64]	ResNet-50	20.1	49.7	12.4	7.4	19.8	24.9	32.1	-
Faster R-CNN w/ RFLA ^[35]	ResNet-50	21.1	51.6	13.1	9.5	21.2	26.1	31.5	41.2
MENet ^[24]	Swin-T ^[94]	23.2	56.2	15.0	9.7	23.9	25.3	34.4	44.7
KLDet ^[36]	ResNet-50	25.2	-	-	-	-	-	-	32.0
DNTR ^[26]	ResNet-50	26.2	56.7	20.2	12.8	26.4	31.0	37.0	42.6
本章方法	ResNet-18	27.1	58.6	21.3	14.7	29.3	35.1	39.1	23.8

4.2.3.1 不同模块的消融实验

表 4.3 本章方法不同模块的消融实验结果

SSE	SQS	O2MAH	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	Params(M)	GFLOPs
			32.0	52.0	32.7	23.8	42.6	51.0	21.4	135.3
✓			32.7	53.2	33.3	24.1	43.4	54.0	23.8	146.5
	✓		32.7	53.2	33.4	24.4	43.5	56.1	21.4	135.3
		✓	32.1	52.3	33.1	23.5	42.6	55.6	21.4	135.3
✓	✓		33.2	54.0	34.5	24.7	44.1	56.1	23.8	146.5
✓		✓	33.1	54.0	33.9	24.6	44.1	52.1	23.8	146.5
	✓	✓	32.0	52.0	32.7	23.8	42.6	51.0	21.4	135.3
✓	✓	✓	33.4	54.4	33.9	25.0	44.0	55.3	23.8	146.5

为了深入评估本章方法中各个核心模块的有效性，本小节在相同实验设置下进行了模块消融实验。如表 4.3 所示，以未添加任何改进的 RT-DETR 作为基线模型，其初始的 AP 为 32.0%，针对小目标的 AP_S 为 23.8%。

首先对单一模块的贡献进行分析。单独引入 SSE 后，模型 AP 提升至 32.7%，AP_S 提升至 24.1%。虽然这引入了 2.4M 的少量额外参数与 11.2 GFLOPs 的计算量，但考虑到其对浅层高分辨率特征的高效建模，这一开销在可接受范围内。而单独引入 SQS 展现出了较高的性价比，在完全不增加任何参数量和计算开销的前提下，将 AP 提

升至 32.7%，尤其使得小目标精度 AP_S 提升了 0.6%，验证了调节系数在修正微小目标置信度得分上的有效性。单独引入 O2MAH 也带来了大部分指标的精度提升，但 AP_S 略有下降，可能是在引入 SSE 和 SQS 之前，O2MAH 带来的正样本更偏向中大目标。

进一步的模块组合实验揭示了它们之间的协同增强效应。当 SSE 与 SQS 结合使用时，模型 AP 进一步提升至 33.2%， AP_S 达到 24.7%，表明 SSE 与 SQS 能够形成互补。最终，在同时集成三个模块的完整架构下，模型的平均 AP 取得 33.4% 的最高分，较基线提升 1.4%，且微小目标核心指标 AP_S 达到 25.0%，较基线提升 1.2%。综合来看，本章提出的三个模块在几乎没有大幅改变原模型轻量化特性的基础上，成功实现了对微小目标检测精度的显著提升。

4.2.3.2 SSE 中不同显著特征选择方式的消融实验

表 4.4 SSE 不同显著特征选择方式的消融实验结果

方法	选择方式	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params(M)	GFLOPs
基线方法	-	32.0	52.0	32.7	23.8	42.6	51.0	21.4	135.3
	基于轻量卷积神经网络	32.8	53.7	33.4	24.3	43.6	54.2	24.4	163.3
SSE	基于自注意力层	32.8	53.7	33.5	24.1	44.0	56.5	23.9	150.7
	基于 Top-P 策略	33.4	54.4	33.9	25.0	44.0	55.3	23.8	146.5

为了验证 SSE 中显著特征选择方式的合理性与高效性，本小节对比了基于轻量级卷积神经网络、基于自注意力层以及本章方法的基于特征图通道均值 Top-P 策略三种不同的特征筛选机制。如表 4.4 所示，基于卷积神经网络和自注意力层的选择方式虽然引入了较为复杂的参数学习过程，但它们在微小目标核心指标 AP_S 上的表现受限于复杂网络带来的特征平滑效应，并未达到最优。此外，这两种方法还导致模型的计算开销大幅增加，GFLOPs 分别飙升至 163.3 和 150.7。相比之下，本章所采用的 Top-P 策略展现出了优异的能效比，它在引入零参数量和极小计算开销的情况下，使模型在平均 AP 和微小目标精度 AP_S 上实现最优，分别达到 33.4% 和 25.0%。这一实验结果充分证明，简单直接的 Top-P 策略能够在避免引入冗余计算负担的前提下，精准地锁定并保留微小目标的显著空间细节，契合本章旨在实现高精度、实时且轻量化微小目标检测的设计初衷。

4.2.3.3 SSE 中不同显著特征选择比例 p 的消融实验

表 4.5 SSE 不同显著特征选择比例 p 的消融实验结果

方法	p	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	Params(M)	GFLOPs
基线方法	-	32.0	52.0	32.7	23.8	42.6	51.0	21.4	135.3
全局注意力	1	-	-	-	-	-	-	21.4	135.3
SSE	0.1	32.7	53.0	33.8	24.2	43.8	53.6	23.1	141.3
	0.15	33.0	53.8	33.9	24.3	43.8	55.6	23.4	143.9
	0.2	33.4	54.4	33.9	25.0	44.0	55.3	23.8	146.5
	0.25	33.1	54.0	33.8	24.5	43.9	55.6	24.1	149.3

为了探究 SSE 中显著特征选择比例 p 对网络性能与计算效率的具体影响，本节对比了全局注意力以及不同采样比例下的模型表现。如表 4.5 所示，值得注意的是，当尝试放弃特征筛选，直接在浅层高分辨率特征图上应用全局注意力时，由于自注意力机制的计算复杂度随空间尺度呈平方级激增，直接导致了模型训练时的显存爆炸并出现 OOM 报错。这一现象也证明了在微小目标检测任务中，传统 Transformer 编码器难以直接处理包含丰富空间细节的高分辨率特征，而引入 SSE 进行特征稀疏化采样是有必要的。

在应用 SSE 后，随着 p 从 0.1 提升至 0.2，模型捕获的关键空间细节逐渐丰富，各项检测指标呈现稳步上升趋势。当 p 为 0.2 时，模型性能达到最优，平均 AP 和小目标精度 AP_S 分别达到了 33.4% 和 25.0%。然而，当 p 进一步扩大至 0.25 时，模型的检测精度并未随之提升，反而出现了轻微回落，AP 降至 33.1%，AP_S 降至 24.5%，且计算开销继续攀升至 149.3 GFLOPs。这种性能衰退表明，微小目标在图像中占据的像素极少，过高的采样比例跨越了显著特征的临界点，将冗余的背景噪声和无效纹理引入了编码器；这些噪声干扰了模型对关键特征的深度建模，进而分散了自注意力机制对微小目标的聚焦能力。综合来看，20% 的显著特征选择契合了目标特征在背景中的分布先验，在成功规避显存爆炸并保持轻量级计算开销的同时，实现了微小目标检测精度的提升。

为了更好的展示 SSE 的作用，本小节对 SSE 选择的显著特征区域进行的可视化。如图 4.5 所示，该可视化图片直观地展示了 SSE 在复杂场景下对微小目标的特征选择能力。VisDrone 数据集图像中包含大量极小尺度的目标，背景高度复杂且冗余。对这些高分辨率特征图进行全量自注意力编码将不可避免地浪费计算资源和显存。从图中可以看到，大多数 GT 都落在了选定的绿色区域内。这表明 SSE 能够精准地从

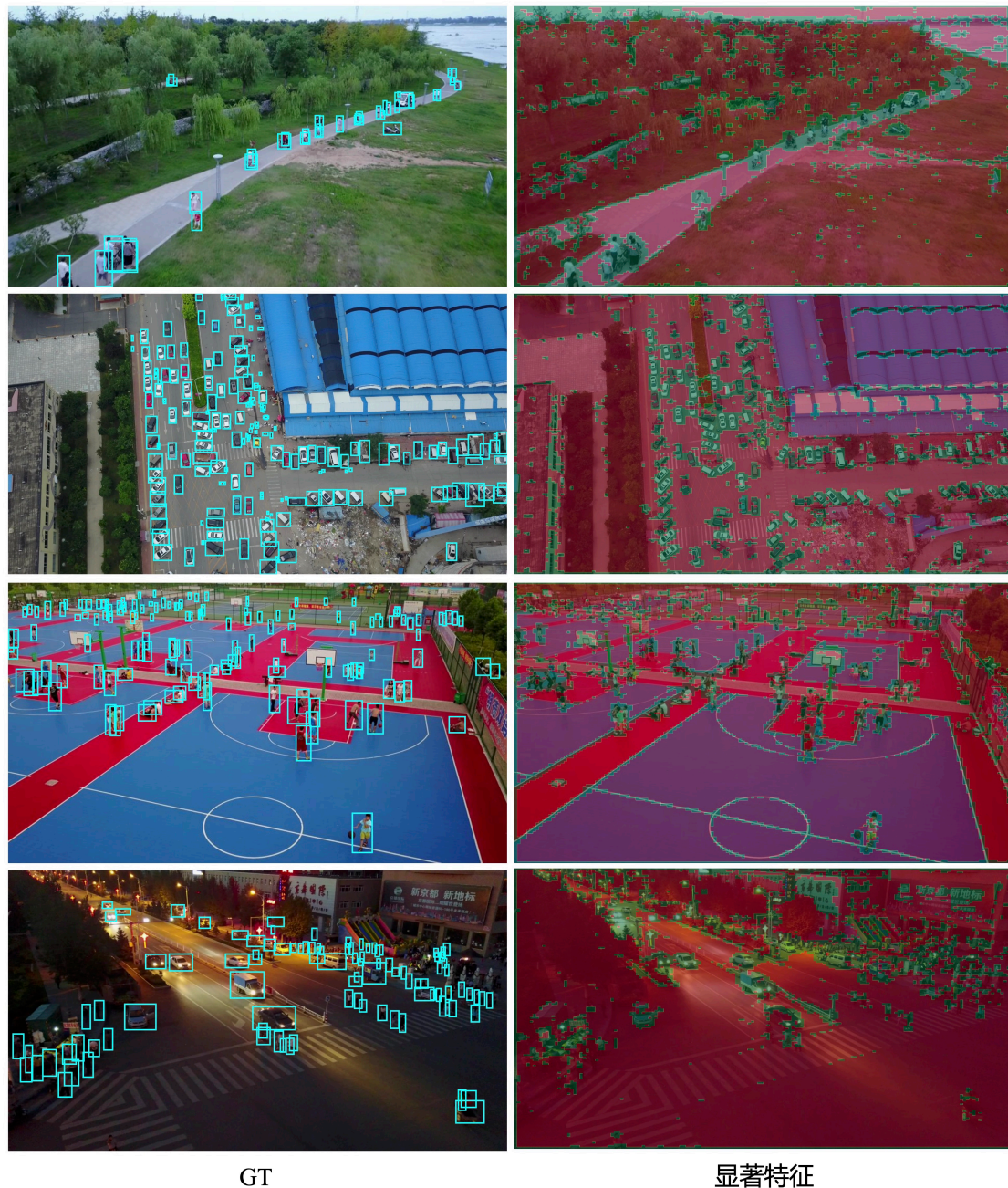


图 4.5 SSE 选择的显著特征区域可视化。

冗余的背景中筛选出微小目标的几何细节，缓解关键空间信息丢失。大片非目标区域特征被在线性自注意力编码器中被忽略，有效地剔除了冗余信息，验证了 SSE 能够以极小的计算开销优化显著特征。在第 4 行夜间场景中，目标在低光照条件下对比度极低。然而，SSE 仍能准确地选择出这些潜在的绿色目标区域，说明基于 Top-P 的显著特征选择具有良好的鲁棒性。综上所述，显著特征区域进行的可视化证明了 SSE 能够高效且精准地提取对任务最有贡献的区域特征，还验证了其能够有效减少

冗余计算，从而支持实时检测更高分辨率特征图。

4.2.3.4 SSE 中使用不同自注意力层的消融实验

表 4.6 SSE 中不同自注意力层的消融实验结果

方法	自注意力层	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	Params(M)	GFLOPs
基线方法	-	32.0	52.0	32.7	23.8	42.6	51.0	21.4	135.3
SSE	多头自注意力	33.0	54.2	33.8	24.5	43.9	54.1	23.8	175.0
	线性注意力	33.4	54.4	33.9	25.0	44.0	55.3	23.8	146.5

为了验证 SSE 中引入线性注意力机制的必要性与优越性，本节在保持其他网络结构与特征选择策略一致的前提下，对比了传统多头自注意力与线性注意力在处理稀疏显著特征时的性能表现。如表 4.6 所示，相较于基线方法，直接应用传统的多头自注意力机制虽然能够使平均 AP 提升至 33.0%，微小目标 AP_S 提升至 24.5%，但由于其自注意力矩阵的计算复杂度随输入序列长度呈平方级增长，导致模型的计算开销剧增至 175.0 GFLOPs，违背了在资源受限设备上实时检测的设计初衷。

相比之下，本章所采用的线性注意力机制展现出了较好的计算性能优势。在模型参数量保持一致的情况下，线性注意力将计算开销削减了约 16%，降至 146.5 GFLOPs。这种计算复杂度的降低并未带来精度折损，反而使其性能超越了传统的多头自注意力，模型的平均 AP 达到了 33.4%，核心的微小目标精度 AP_S 更是达到了 25.0% 的最优水平。这一实验结果不仅证明了线性注意力在降低计算开销上的作用，更表明其机制相较于传统的平方级注意力，能够更为高效且精准地拟合 SSE 所提取的稀疏特征分布，从而平衡了精度与速度。

4.2.3.5 SQS 中不同 γ 值的消融实验

表 4.7 SQS 中不同 γ 值的消融实验结果

方法	γ	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
基线方法	-	32.0	52.0	32.7	23.8	42.6	51.0
	0.7	33.0	54.2	33.7	24.9	43.5	59.8
SQS	0.8	33.4	54.4	33.9	25.0	44.0	55.3
	0.9	32.7	53.5	33.5	24.2	43.7	54.4

γ 是 SQS 中控制置信度调节幅度的超参数，为了确认所提出的 SQS 及其核心超参数 γ 的最佳配置，本小节进行了消融实验。如表4.7所示，当 γ 设定为 0.8 时，模型达到了最佳的综合表现，平均 AP 从 32.0% 提升至 33.4%， AP_{50} 与 AP_{75} 也分别获得了 2.4 % 和 1.2 % 的增长。更重要的是，得益于 SQS 通过调节系数有效平衡了不同尺度特征向量的置信度得分差异，微小目标在 Query 竞争阶段的劣势得到了弥补。在该最优配置下， AP_S 从 23.8% 明显提升至 25.0%， AP_M 也同步提升了 1.4 %，但引入了更多微小目标特征， AP_L 相较于 $\gamma = 0.7$ 有所下降。此外，实验数据也反映出模型性能随 γ 值变化呈现合理的先升后降趋势， $\gamma = 0.9$ 时各项指标开始轻微回落，表明适度的尺度干预才能达到最佳的多尺度特征平衡。实验结果充分证明了 SQS 在有效缓解微小目标漏检问题、提升模型多尺度检测能力方面的有效性。

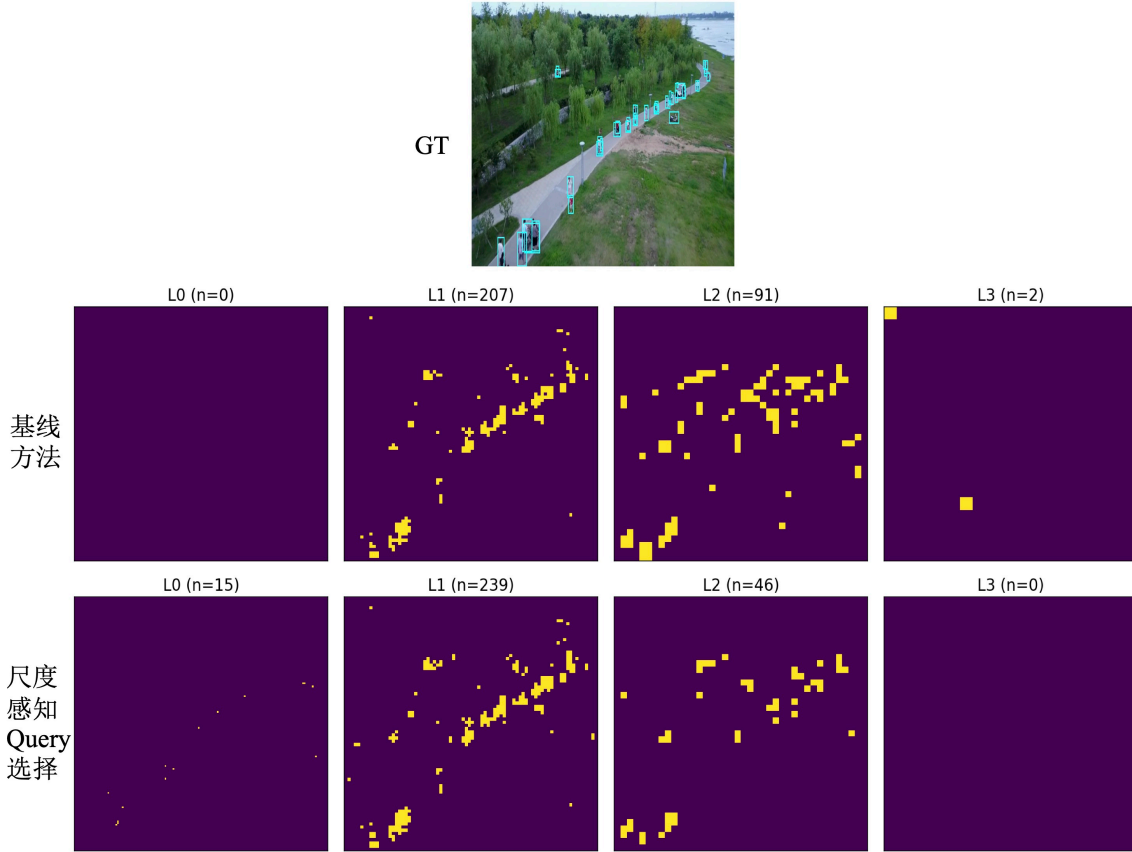


图 4.6 Top-300 Query 在不同特征金字塔层级（L0 至 L3）的分布可视化

如图 4.6 所示，为了直观验证 SQS 对微小目标检测任务 Query 选择的干预效果，本小节对模型在进入解码器前提取的 Top-300 初始化 Query 在不同特征金字塔层级 L0 至 L3 上的空间分布进行了可视化。其中，L0 代表具有最小下采样步长、最高分

分辨率的浅层特征图，包含了最丰富的微小目标细节；而 L3 代表最低分辨率的深层特征图，主要用于检测大尺度目标。在基线方法的筛选机制中，由于浅层特征的初始置信度得分普遍偏低，导致在 Top-K 竞争中处于劣势。这直接体现在 L0 层被选中的 Query 数量为 0。而在引入 SQS 后，L0 层的 Query 数量增加至 15。观察整体的 Query 分布变化，在总 Query 数量保持不变的前提下，代表高分辨率的 L0 和 L1 层分配到的 Query 总数从基线方法的 207 个增加至 254 个；与之对应，代表低分辨率的 L2 和 L3 层的 Query 数量则从 93 个减少至 46 个。这种资源的重分配说明 SQS 通过引入基于特征层级下采样步长的调节系数，有效抑制了深层特征图中的特征得分，从而为浅层微小目标特征腾出了更多的 Query 名额。综上所述，该可视化结果充分印证了 SQS 能够有效平衡不同尺度特征在 Query 竞争中的得分差异，使得网络在不增加任何额外 Query 数量的前提下，缓解了微小目标的漏检问题。

4.2.4 实时检测实验

为了全面验证本章提出方法的实时推理能力，本小节在与训练环境一致的硬件平台使用单张 NVIDIA RTX 3090 GPU 进行了推理速度评估。在测试过程中，所有模型均采用 FP32 全精度格式，并在未引入额外后处理优化的纯 PyTorch 环境下记录推理每秒推理帧数 FPS。为了模拟真实的流视频处理场景，实验时将推理时的 Batch Size 设置为 1。

表 4.8 本章方法与其他先进模型在 VisDrone 数据集上的推理速度与精度对比

模型	Params (M)	GFLOPs	AP	FPS
YOLOv8-M	25.9	78.9	24.6	65.24
YOLOv9-M	20.1	76.8	25.2	47.24
YOLOv10-M	15.4	59.1	24.5	63.72
YOLOv11-M	20.0	67.7	25.9	53.73
RT-DETR-R18	20.0	57.3	26.7	59.82
本章方法 *	21.7	63.8	28.6	51.21

* 表示不使用 C_2 高分辨率特征图。

如表 4.8 所示，尽管本章方法为了增强微小目标检测能力对浅层高分辨率特征

进行了深度建模，但其推理速度依然与以检测速度著称的 YOLO 系列保持在同一水平，满足微小目标检测场景的实时性需求。

本章方法在检测精度与推理速度之间取得良好权衡的关键，在于对网络架构的精细化设计。首先，SSE 通过 Top-P 策略排除了大部分背景冗余特征，并配合线性自注意力机制，成功将高分辨率特征图的处理耗时控制在较低范围内；其次，SQS 仅通过简单的动态标量乘法调整得分，几乎不引入任何额外耗时；最后，一对多辅助检测头仅在训练阶段激活，在推理阶段被完全剥离，在不增加额外负担的情况下实现了的精度提升。实验证明，本章方法能够在不牺牲实时推理性能的前提下，实现微小目标检测精度的提升。

4.3 本章小结

本章针对微小目标检测中高分辨率特征计算开销大、特征易受背景干扰及监督信号稀疏等问题，提出了一种兼顾高精度与实时性的目标检测方法。首先，选择稀疏编码器利用基于特征通道均值的特征选择机制与线性自注意力，极大地降低了高分辨率特征处理的计算与显存消耗。其次，尺度感知 Query 选择依据特征金字塔的下采样步长动态调整候选 Query 的置信度得分，在不引入额外计算量的前提下显著提升了微小目标的检测精度。最后，一对多匹配辅助检测头在训练阶段采用密集的分配策略为微小目标提供更强的监督信号以加速收敛，并在推理阶段无缝剥离，保持了端到端的极简架构。在 VisDrone 和 AI-TOD 数据集上的结果表明，该方法以极低的参数量与复杂度显著超越主流模型，为资源受限设备上的部署提供了高效方案。

第五章 总结与展望

5.1 结论

在当今数字化与智能化高速发展的背景下，微小目标检测在农业监测、自动驾驶、无人机侦察和医学影像分析等关键领域发挥着不可替代的作用。然而，由于微小目标占据像素极少，在深度卷积神经网络的前馈过程中极易出现特征丢失、信噪比低以及正负样本分配极度不均衡等问题。此外，为保留微小目标特征而依赖的高分辨率特征图，给模型带来了巨大的计算开销，使得检测精度与推理速度之间存在着难以调和的矛盾。

针对上述挑战，本文的核心思想是从基于深度学习的目标检测理论出发，深入研究微小目标检测的特征融合和标签分配理论，结合实时目标检测方向的最新进展，分别从提升检测精度和优化实时部署效率两个维度出发，提出了一种超高精度微小目标检测方法和一种偏向实时应用的微小目标检测方法。

本文的主要研究工作和贡献总结如下：

(1) 提出了一种基于隐式特征融合与混合自适应标签分配的高精度微小目标检测方法。针对卷积神经网络在特征提取容易造成微小目标信丢失的问题，以及传统标签分配策略对与微小目标的局限性，本文在双阶段检测框架的基础上提出了该方法。在特征融合层面，设计了隐式特征融合模块。该模块通过隐式神经表示与空间编码，将离散特征图建模为连续信号，有效缓解了多尺度融合中的特征对齐偏差，成功保留了微小目标的局部精细特征。在标签分配层面，提出了混合自适应标签分配策略。该策略创新性地将感受野距离与 IoU 结合构建混合度量得分，并利用样本分布的均值与标准差动态生成匹配阈值。实验表明，该方法的综合性能显著优于现有主流方法，特别是在极端微小目标的召回与精确定位上实现了大幅提升，证明了该方法在提升极小尺度目标检测能力上的有效性。

(2) 提出了一种基于选择性稀疏编码器与尺度感知 Query 选择的实时微小目标检测方法。为了突破 Transformer 架构在处理高分辨率特征图时的计算与显存瓶颈，本文基于 RT-DETR 框架提出了一种兼顾精度与速度的轻量化端到端检测方法。在特征编码层面，设计了选择性稀疏编码器。通过基于通道均值的 Top-K 特征选择机制

与线性自注意力，以极低的计算代价排除了大量背景冗余，实现了对高分辨率显著特征的深度编码。在 Query 筛选层面，提出了尺度感知 Query 选择策略。通过引入基于下采样步长的动态折损系数，有效补偿了微小目标特征在置信度评估中的固有劣势，帮助高质量的微小目标 Query 能够顺利进入解码阶段。在训练监督层面，引入了一对多匹配辅助检测头，通过密集的监督信号加速了模型对微小目标的收敛，并在推理阶段无缝剥离，保持了极简的端到端架构。实验表明，该方法在参数量和计算量大幅缩减的情况下，实现了与复杂重型模型相媲美甚至更高的精度，同时满足了与其他实时目标检测方法相当的实时推理性能，成功打破了微小目标检测中精度与计算开销的壁垒。

5.2 工作展望

尽管本文在微小目标的特征表征、标签分配以及轻量化端到端检测方面取得了突破性进展并达到了业内先进水平，但该研究方向仍有广阔的探索空间和诸多值得深入研究的方面。基于本文的研究成果，未来的研究工作可以从以下几个维度进一步深化和拓展：

(1) 跨模态融合与多源信息互补。本文的方法主要基于 RGB 图像进行特征提取。然而，在夜间、雨雪、雾霾等复杂恶劣气象条件下，微小目标原本微弱的光学特征会受到进一步破坏，甚至完全不可见。未来的研究可以考虑引入红外热成像、毫米波雷达或事件相机等多模态数据。通过探索高效的跨模态特征对齐与融合网络，利用不同传感器在空间分辨率和物理感知上的互补性，全天候、全天时地提升微小目标检测的鲁棒性。

(2) 面向极端边缘场景的底层硬件协同优化。虽然本文提出的实时微小目标检测方法已通过选择性稀疏编码和线性注意力机制大幅降低了模型的计算复杂度，但在搭载于微型无人机或低功耗物联网探头等极端受限的边缘芯片上时，依然存在算力约束。未来可结合网络结构搜索、结构化剪枝、低比特量化（如 INT8 或 INT4 量化）以及算子级别的重构技术，开展算法与底层硬件的协同优化，以实现微小目标检测模型的极致压缩与加速部署。

(3) 时序上下文建模与视频流微小目标检测。本文的研究目前主要集中在单帧静态图像的检测上。在实际的无人机航拍或监控视频中，微小目标往往伴随着复杂

的背景运动和自身的快速位移。未来可以引入时序上下文建模，通过光流法或时序 **Transformer** 模块提取帧间运动先验。利用多帧之间的时间一致性和运动轨迹规律，可以有效弥补单帧图像中因遮挡、运动模糊或极小尺度带来的信息缺失，进一步降低视频流中的漏检率和误检率。

（4）弱监督与小样本微小目标学习范式。当前的高精度微小目标检测模型高度依赖于类似 **AI-TOD**、**VisDrone** 这样规模庞大且标注精细的数据集。然而，在诸如深空探测、罕见医学病灶识别等特殊领域，获取海量且高质量的微小目标像素级标注成本极其高昂。未来研究可探索自监督学习、大语言模型驱动的开放词汇检测或基于主动学习的小样本检测范式。通过挖掘海量无标注数据中的通用视觉特征，减少模型对人工边界框标注的依赖，提升算法在低资源场景下的泛化迁移能力。

插图索引

图 1.1	本文的研究路线示意图	7
图 2.1	残差模块示意图	11
图 2.2	特征金字塔网络 FPN 示意图	12
图 2.3	FPN 对微小目标特征的影响示意图	12
图 2.4	不同尺寸目标预测框偏移对 IoU 的影响示意图	14
图 2.5	Faster R-CNN ^[51] 模型结构示意图	17
图 2.6	YOLOv8 ^[52] 模型结构示意图	18
图 2.7	RT-DETR ^[44] 模型结构示意图	20
图 3.1	基于隐式特征融合与混合自适应标签分配的目标检测方法架构	24
图 3.2	隐式映射流程示意图	25
图 3.3	特征融合层级示意图	27
图 3.4	不同标签分配方法的正样本分配结果可视化	29
图 3.5	本章方法在 AI-TOD 数据集上的检测结果可视化	33
图 3.6	不同模型在 AI-TOD 数据集上的检测性能对比。绿色、红色和蓝色框 分别代表正确检测 (TP)、漏检 (FN) 和误检 (FP)。	37
图 3.7	本章方法在 AI-TOD 数据集上的消融实验结果可视化。绿色、红色和 蓝色框分别代表正确检测 (TP)、漏检 (FN) 和误检 (FP)。	39
图 3.8	IFF 融合特征图的可视化结果	40
图 3.9	单张图像中不同实例最终正样本数量的可视化	42
图 3.10	单张图像中两个实例候选正样本的 HM 得分分布可视化	44

图 4.1	基于选择稀疏编码器与尺度感知 Query 选择的实时微小目标检测方法架构	46
图 4.2	VisDrone 数据集中目标区域占比分布	47
图 4.3	尺度感知 Query 选择效果示意图	51
图 4.4	本章方法在 VisDrone-2019 数据集上的检测结果可视化	55
图 4.5	SSE 选择的显著特征区域可视化。	59
图 4.6	Top-300 Query 在不同特征金字塔层级 (L0 至 L3) 的分布可视化	61

表格索引

表 2.1	不同数据集评价指标与目标尺寸之间的对应关系	23
表 3.1	本章方法与其他先进方法在 AI-TOD 数据集上的对比实验	32
表 3.2	本章方法与其他先进方法在 SODA-D 数据集上的对比实验	34
表 3.3	本章方法与其他先进方法在 VisDrone 数据集上的对比实验	35
表 3.4	本章方法与其他先进方法在 AgriPest 数据集上的对比实验	36
表 3.5	不同模块的消融实验结果	38
表 3.6	IFF 输出不同融合特征图的消融实验结果	39
表 3.7	IFF 中不同融合层级的消融实验结果	40
表 3.8	IFF 中空间编码的消融实验结果。	41
表 3.9	HALA 中不同阈值策略的消融实验结果	42
表 3.10	不同 HALA 中不同 K 值的消融实验结果	43
表 3.11	不同 α 值的消融实验结果	45
表 4.1	本章方法与其他先进方法在 VisDrone 数据集上的对比实验	54
表 4.2	本章方法与其他先进方法在 AI-TOD 数据集上的对比实验	56
表 4.3	本章方法不同模块的消融实验结果	56
表 4.4	SSE 不同显著特征选择方式的消融实验结果	57
表 4.5	SSE 不同显著特征选择比例 p 的消融实验结果	58
表 4.6	SSE 中不同自注意力层的消融实验结果	60
表 4.7	SQS 中不同 γ 值的消融实验结果	60
表 4.8	本章方法与其他先进模型在 VisDrone 数据集上的推理速度与精度对比 ..	62

参考文献

- [1] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M/OL]//Computer Vision –ECCV 2014. Springer International Publishing, 2014: 740-755. http://dx.doi.org/10.1007/978-3-319-10602-1_48.
- [2] WANG J, YANG W, GUO H, et al. Tiny object detection in aerial images[C/OL]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 3791-3798. <http://dx.doi.org/10.1109/icpr48806.2021.9413340>.
- [3] DU D, ZHU P, WEN L, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results[C/OL]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, 2019. <http://dx.doi.org/10.1109/iccvw.2019.00030>.
- [4] BOSQUET B, CORES D, SEIDENARI L, et al. A full data augmentation pipeline for small object detection based on generative adversarial networks[J/OL]. Pattern Recognition, 2023, 133: 108998. <http://dx.doi.org/10.1016/j.patcog.2022.108998>.
- [5] RUIZ-PONCE P, ORTIZ-PEREZ D, GARCIA-RODRIGUEZ J, et al. POSEIDON: a data augmentation tool for small object detection datasets in maritime environments [J/OL]. Sensors, 2023, 23(7): 3691. <http://dx.doi.org/10.3390/s23073691>.
- [6] XU C, WANG J, YANG W, et al. Dot distance for tiny object detection in aerial images [C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2021: 1192-1201. <http://dx.doi.org/10.1109/cvprw53098.2021.00130>.
- [7] DENG C, WANG M, LIU L, et al. Extended feature pyramid network for small object detection[J/OL]. IEEE Transactions on Multimedia, 2022, 24: 1968-1979. <http://dx.doi.org/10.1109/tmm.2021.3074273>.

- [8] BASHIR S M A, WANG Y. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network [J/OL]. Remote Sensing, 2021, 13(9): 1854. <http://dx.doi.org/10.3390/rs13091854>.
- [9] ZHANG Y, WANG S, ZHANG Y, et al. Asymmetric light-aware progressive decoding network for RGB-thermal salient object detection[J/OL]. Journal of Electronic Imaging, 2025, 34(01). <http://dx.doi.org/10.1117/1.jei.34.1.013005>.
- [10] ZHANG Y, LIU Y, KANG W, et al. VSS-net: visual semantic self-mining network for video summarization[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(4): 2775-2788. <http://dx.doi.org/10.1109/tcsvt.2023.3312325>.
- [11] HU X, XU X, XIAO Y, et al. SINet: a scale-insensitive convolutional neural network for fast vehicle detection[J/OL]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(3): 1010-1019. <http://dx.doi.org/10.1109/tits.2018.2838132>.
- [12] CUI L, LV P, JIANG X, et al. Context-aware block net for small object detection [J/OL]. IEEE Transactions on Cybernetics, 2022, 52(4): 2300-2313. <http://dx.doi.org/10.1109/tcyb.2020.3004636>.
- [13] KIOBYA T, ZHOU J, MAISELI B. A multi-scale semantically enriched feature pyramid network with enhanced focal loss for small-object detection[J/OL]. Knowledge-Based Systems, 2025, 310: 113003. <http://dx.doi.org/10.1016/j.knosys.2025.113003>.
- [14] YUAN X, CHENG G, YAO R, et al. Semantic differentiation aids oriented small object detection[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2025: 1-1. <http://dx.doi.org/10.1109/tcsvt.2025.3532243>.
- [15] ZHANG Y, WU C, GUO W, et al. CFANet: efficient detection of UAV image based on cross-layer feature aggregation[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-11. <http://dx.doi.org/10.1109/tgrs.2023.3273314>.
- [16] ZHANG Y, WU C, ZHANG T, et al. Full-scale feature aggregation and grouping feature reconstruction-based UAV image target detection[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-11. <http://dx.doi.org/10.1109/tgrs.2024.3392794>.

- [17] CHENG G, YUAN X, YAO X, et al. Towards large-scale small object detection: survey and benchmarks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023: 1-20. <http://dx.doi.org/10.1109/tpami.2023.3290594>.
- [18] YANG F, FAN H, CHU P, et al. Clustered object detection in aerial images[C/OL]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 8310-8319. <http://dx.doi.org/10.1109/iccv.2019.00840>.
- [19] YANG C, HUANG Z, WANG N. QueryDet: cascaded sparse query for accelerating high-resolution small object detection[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 13658-13667. <http://dx.doi.org/10.1109/cvpr52688.2022.01330>.
- [20] KOYUN O C, KESER R K, AKKAYA I B, et al. Focus-and-detect: a small object detection framework for aerial images[J/OL]. Signal Processing: Image Communication, 2022, 104: 116675. <http://dx.doi.org/10.1016/j.image.2022.116675>.
- [21] ZHANG Y, ZHANG T, WANG S, et al. An efficient perceptual video compression scheme based on deep learning-assisted video saliency and just noticeable distortion [J/OL]. Engineering Applications of Artificial Intelligence, 2025, 141: 109806. <http://dx.doi.org/10.1016/j.engappai.2024.109806>.
- [22] ZHANG C, LAM K M, WANG Q. CoF-net: a progressive coarse-to-fine framework for object detection in remote-sensing imagery[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-17. <http://dx.doi.org/10.1109/tgrs.2022.3233881>.
- [23] QUAN Y, ZHANG D, ZHANG L, et al. Centralized feature pyramid for object detection[J/OL]. IEEE Transactions on Image Processing, 2023, 32: 4341-4354. <http://dx.doi.org/10.1109/tip.2023.3297408>.
- [24] ZHANG T, ZHANG X, ZHU X, et al. Multistage enhancement network for tiny object detection in remote sensing images[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12. <http://dx.doi.org/10.1109/tgrs.2024.3363614>.

- [25] GE L, WANG G, ZHANG T, et al. Regression-guided refocusing learning with feature alignment for remote sensing tiny object detection[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14. <http://dx.doi.org/10.1109/tgrs.2024.3407122>.
- [26] LIU H I, TSENG Y W, CHANG K C, et al. A DeNoising FPN with transformer R-CNN for tiny object detection[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15. <http://dx.doi.org/10.1109/tgrs.2024.3396489>.
- [27] ZHANG Y, ZHANG T, WU C, et al. Multi-scale spatiotemporal feature fusion network for video saliency prediction[J/OL]. IEEE Transactions on Multimedia, 2024, 26: 4183-4193. <http://dx.doi.org/10.1109/tmm.2023.3321394>.
- [28] ZHANG Y, ZHEN J, LIU T, et al. Adaptive differentiation siamese fusion network for remote sensing change detection[J/OL]. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 1-5. <http://dx.doi.org/10.1109/lgrs.2024.3516775>.
- [29] ZHANG Y, XIAO Y, ZHANG Y, et al. Video saliency prediction via single feature enhancement and temporal recurrence[J/OL]. Engineering Applications of Artificial Intelligence, 2025, 160: 111840. <http://dx.doi.org/10.1016/j.engappai.2025.111840>.
- [30] ZHANG X, WAN F, LIU C, et al. Learning to match anchors for visual object detection[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3096-3109. <http://dx.doi.org/10.1109/tpami.2021.3050494>.
- [31] ZHANG S, CHI C, YAO Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020. <http://dx.doi.org/10.1109/cvpr42600.2020.00978>.
- [32] KIM K, LEE H S. Probabilistic anchor assignment with IoU prediction for object detection[M/OL]//Computer Vision –ECCV 2020. Springer International Publishing, 2020: 355-371. http://dx.doi.org/10.1007/978-3-030-58595-2_22.
- [33] ZHANG H, WANG Y, DAYOUB F, et al. VarifocalNet: an IoU-aware dense object detector[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. <http://dx.doi.org/10.1109/cvpr46437.2021.00841>.

- [34] WANG J, XU C, YANG W, et al. A normalized gaussian wasserstein distance for tiny object detection[A/OL]. arXiv (2021). <https://arxiv.org/abs/2110.13389>.
- [35] XU C, WANG J, YANG W, et al. RFLA: gaussian receptive field based label assignment for tiny object detection[M/OL]//Computer Vision –ECCV 2022. Springer Nature Switzerland, 2022: 526-543. http://dx.doi.org/10.1007/978-3-031-20077-9_31.
- [36] ZHOU Z, ZHU Y. KLDet: detecting tiny objects in remote sensing images via kullback–leibler divergence[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16. <http://dx.doi.org/10.1109/tgrs.2024.3382099>.
- [37] WANG H, JIA T, WANG Q, et al. Automatic label assignment for object detection [J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35 (11): 11448-11461. <http://dx.doi.org/10.1109/tcsvt.2025.3578021>.
- [38] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [M/OL]//Computer Vision –ECCV 2016. Springer International Publishing, 2016: 21-37. http://dx.doi.org/10.1007/978-3-319-46448-0_2.
- [39] REDMON J, FARHADI A. YOLOv3: an incremental improvement[A/OL]. arXiv (2018). <https://arxiv.org/abs/1804.02767>.
- [40] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C/OL]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 7464-7475. <http://dx.doi.org/10.1109/cvpr52729.2023.00721>.
- [41] WANG A, CHEN H, LIU L, et al. YOLOv10: real-time end-to-end object detection [A/OL]. arXiv (2024). <https://arxiv.org/abs/2405.14458>.
- [42] TIAN Y, YE Q, DOERMANN D. YOLOv12: attention-centric real-time object detectors[A/OL]. arXiv (2025). <https://arxiv.org/abs/2502.12524>.
- [43] LIU K, FU Z, JIN S, et al. ESOD: efficient small object detection on high-resolution images[J/OL]. IEEE Transactions on Image Processing, 2025, 34: 183-195. <http://dx.doi.org/10.1109/tip.2024.3501853>.

- [44] ZHAO Y, LV W, XU S, et al. DETRs beat YOLOs on real-time object detection [C/OL]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 16965-16974. <http://dx.doi.org/10.1109/cvpr52733.2024.01605>.
- [45] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision[A/OL]. arXiv (2024). <https://arxiv.org/abs/2304.07193>.
- [46] ROBINSON I, ROBICHEAUX P, POPOV M, et al. RF-DETR: neural architecture search for real-time detection transformers[A/OL]. arXiv (2026). <https://arxiv.org/abs/2511.09554>.
- [47] ZHANG H, ZHANG H, MEI A, et al. SO-DETR: leveraging dual-domain features and knowledge distillation for small object detection[C/OL]//2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025: 1-8. <http://dx.doi.org/10.1109/ijcnn64981.2025.11229240>.
- [48] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016. <http://dx.doi.org/10.1109/cvpr.2016.90>.
- [49] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. <http://dx.doi.org/10.1109/cvpr.2017.106>.
- [50] TIAN Z, SHEN C, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019. <http://dx.doi.org/10.1109/iccv.2019.00972>.
- [51] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149. <http://dx.doi.org/10.1109/tpami.2016.2577031>.

- [52] VARGHESE R, M. S. YOLOv8: a novel object detection algorithm with enhanced performance and robustness[C/OL]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, 2024: 1-6. <http://dx.doi.org/10.1109/adics58448.2024.10533619>.
- [53] CHENG G, YUAN X, YAO X, et al. Towards large-scale small object detection: survey and benchmarks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023: 1-20. <http://dx.doi.org/10.1109/tpami.2023.3290594>.
- [54] XU X, WANG Z, SHI H. Ultrasr: spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution[A/OL]. arXiv (2021). <https://arxiv.org/abs/2103.12716>.
- [55] LUO W, LI Y, URTASUN R, et al. Understanding the effective receptive field in deep convolutional neural networks[J]. Advances in neural information processing systems, 2016, 29.
- [56] PASZKE A, GROSS S, MASSA F, et al. Pytorch: an imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32.
- [57] CHEN K, WANG J, PANG J, et al. MMDetection: open mmlab detection toolbox and benchmark. arXiv 2019[A/OL]. arXiv (1906). <https://arxiv.org/abs/1906.07155>.
- [58] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C/OL]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. <http://dx.doi.org/10.1109/cvprw.2009.5206848>.
- [59] WANG R, LIU L, XIE C, et al. AgriPest: a large-scale domain-specific benchmark dataset for practical agricultural pest detection in the wild[J/OL]. Sensors, 2021, 21 (5): 1601. <http://dx.doi.org/10.3390/s21051601>.
- [60] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017. <http://dx.doi.org/10.1109/iccv.2017.324>.
- [61] ZHU B, WANG J, JIANG Z, et al. Autoassign: differentiable label assignment for dense object detection[A/OL]. arXiv (2020). <https://arxiv.org/abs/2007.03496>.

- [62] CAI Z, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C/OL]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 6154-6162. <http://dx.doi.org/10.1109/cvpr.2018.00644>.
- [63] WU J, PAN Z, LEI B, et al. FSANet: feature-and-spatial-aligned network for tiny object detection in remote sensing images[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-17. <http://dx.doi.org/10.1109/tgrs.2022.3205052>.
- [64] FU R, CHEN C, YAN S, et al. Gaussian similarity-based adaptive dynamic label assignment for tiny object detection[J/OL]. Neurocomputing, 2023, 543: 126285. <http://dx.doi.org/10.1016/j.neucom.2023.126285>.
- [65] ZHANG Y, LIU T, ZHEN J, et al. Adaptive downsampling and scale enhanced detection head for tiny object detection in remote sensing image[J/OL]. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 1-5. <http://dx.doi.org/10.1109/lgrs.2025.3532983>.
- [66] QIAO S, CHEN L C, YUILLE A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 10208-10219. <http://dx.doi.org/10.1109/cvpr46437.2021.01008>.
- [67] 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>.
- [68] DUAN K, BAI S, XIE L, et al. CenterNet: keypoint triplets for object detection[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019. <http://dx.doi.org/10.1109/iccv.2019.00667>.
- [69] SUN P, ZHANG R, JIANG Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. <http://dx.doi.org/10.1109/cvpr46437.2021.01422>.
- [70] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[J/OL]. International Journal of Computer Vision, 2019, 128(3): 642-656. <http://dx.doi.org/10.1007/s11263-019-01204-1>.

- [71] DAI X, CHEN Y, XIAO B, et al. Dynamic head: unifying object detection heads with attentions[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 7369-7378. <http://dx.doi.org/10.1109/cvpr46437.2021.00729>.
- [72] YANG Z, LIU S, HU H, et al. RepPoints: point set representation for object detection[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 9656-9665. <http://dx.doi.org/10.1109/iccv.2019.00975>.
- [73] YUAN X, CHENG G, YAN K, et al. Small object detection via coarse-to-fine proposal generation and imitation learning[C/OL]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 6294-6304. <http://dx.doi.org/10.1109/iccv51070.2023.00581>.
- [74] JOCHER G, STOLEN A, BORVEC J, et al. yolov5, GitHub repository[J]. 2020-06-09)[2021-07-09].[Article], 2020.
- [75] DU B, HUANG Y, CHEN J, et al. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images[C/OL]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 13435-13444. <http://dx.doi.org/10.1109/cvpr52729.2023.01291>.
- [76] DU Z, HU Z, ZHAO G, et al. Cross-layer feature pyramid transformer for small object detection in aerial images[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-14. <http://dx.doi.org/10.1109/tgrs.2025.3572706>.
- [77] MA Y, CHAI L, JIN L. Scale decoupled pyramid for object detection in aerial images [J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14. <http://dx.doi.org/10.1109/tgrs.2023.3298852>.
- [78] LI J, GUO S, YI S, et al. DMCTDet: a density map-guided composite transformer network for object detection of UAV images[J/OL]. Signal Processing: Image Communication, 2025, 136: 117284. <http://dx.doi.org/10.1016/j.image.2025.117284>.
- [79] WEI Z, DUAN C, SONG X, et al. Amrnet: chips augmentation in aerial images object detection[A/OL]. arXiv (2020). <https://arxiv.org/abs/2009.07168>.

- [80] MEETHAL A, GRANGER E, PEDERSOLI M. Cascaded zoom-in detector for high resolution aerial images[C/OL]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2023: 2046-2055. <http://dx.doi.org/10.1109/cvprw59228.2023.00198>.
- [81] ZHU P, WEN L, BIAN X, et al. Vision meets drones: a challenge[A/OL]. arXiv (2018). <https://arxiv.org/abs/1804.07437>.
- [82] 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>.
- [83] 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>.
- [84] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[A/OL]. arXiv (2023). <https://arxiv.org/abs/1706.03762>.
- [85] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and A loss for bounding box regression[A/OL]. arXiv (2019). <https://arxiv.org/abs/1902.09630>.
- [86] ZHANG H, WANG Y, DAYOUB F, et al. VarifocalNet: an IoU-aware dense object detector[A/OL]. arXiv (2021). <https://arxiv.org/abs/2008.13367>.
- [87] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements[A/OL]. arXiv (2024). <https://arxiv.org/abs/2410.17725>.
- [88] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: learning what you want to learn using programmable gradient information[A/OL]. arXiv (2024). <https://arxiv.org/abs/2402.13616>.
- [89] XU S, WANG X, LV W, et al. PP-YOLOE: an evolved version of YOLO[A/OL]. arXiv (2022). <https://arxiv.org/abs/2203.16250>.
- [90] XU C, DING J, WANG J, et al. Dynamic coarse-to-fine learning for oriented tiny object detection[C/OL]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 7318-7328. <http://dx.doi.org/10.1109/cvpr52729.2023.00707>.

- [91] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[A/OL]. arXiv (2020). <https://arxiv.org/abs/2005.12872>.
- [92] ZHANG H, ZHANG H, LIU K, et al. UAV-DETR: efficient end-to-end object detection for unmanned aerial vehicle imagery[C/OL]//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025: 15143-15149. <http://dx.doi.org/10.1109/iros60139.2025.11246176>.
- [93] YU F, WANG D, SHELHAMER E, et al. Deep layer aggregation[A/OL]. arXiv (2019). <https://arxiv.org/abs/1707.06484>.
- [94] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[A/OL]. arXiv (2021). <https://arxiv.org/abs/2103.14030>.

攻读硕士学位期间取得的研究成果

一、主要作者研究成果

[1] Qiaochuan Chen, Tianyang Xu, Yan Sun, Bing Wang, Yuexing Han*, Tiny object detection via implicit feature fusion and hybrid metric adaptive label assignment, Knowledge-Based Systems, Volume 347, doi: <https://doi.org/10.1016/j.knosys.2026.116218>. 2026. 中科院一区 Top, JCR Q1

二、软件著作权

[1] 软件名称：平流雾能见度观测预警系统 V1.0，开发人：陈侨川，徐天洋，韩越兴。登记号：2024SR0542281，申请人：上海大学，开发完成日期：2024 年 01 月 31 日，登记日期：2024 年 4 月 22 日。

[2] 软件名称：基于高斯建模标签分配的密集小目标害虫检测软件 V1.0，开发人：陈侨川，徐天洋，韩越兴。登记号：2024R11L1474448，申请人：上海大学，开发完成日期：2024 年 02 月 8 日，登记日期：2024 年 8 月 28 日。

[3] 软件名称：基于混合指标自适应标签分配的遥感图像检测软件 V1.0，开发人：陈侨川，徐天洋，韩越兴。登记号：2025R11L1697365，申请人：上海大学，开发完成日期：2025 年 03 月 10 日，登记日期：2025 年 9 月 25 日。

致 谢

还记得三年前，一位青春懵懂的学生踏入了上海大学的校园，惊诧于上海这座城市所带来的震撼。旅途很漫长，三年中发生了太多事情，以至于我需要像今天一样抛开所有杂念，来细细回忆并写下这段论文致谢。这趟旅程并不轻松，我不断的向着山顶攀爬，又不断的一落千丈，但很高兴幸运从未离我而去。我为这期间经历的无数美好瞬间和至暗时刻感到骄傲，这些经历就像是幽暗矿藏中熠熠生辉的宝石，成就了现在的我。我想要由衷的感谢无数鼓励支持我的老师、家人和朋友，并借此机会向所有帮助过我的人们致以最诚挚的谢意。

首先，我想诚挚地感谢我的老师。感谢我的导师陈侨川教授，以及课题组内的韩越兴教授、孙妍教授等老师对我科研工作的用心指导，帮助我顺利完成了大小论文的撰写，我论文中的每一句话都凝聚着他们的心血，老师们专业的学术见解和严谨的治学态度会让我受用终身。生活中陈老师和韩老师也对我悉心照顾，两位老师风趣的谈吐和宽和的为人也帮我缓解了学业上的压力，积极面对挑战。

其次，我要由衷地感谢我的家人。我深知自己衣食无忧的生活来之不易，这样的生活由家人对我的无私奉献和牺牲编织而成。平时忙碌的生活让我没能空出足够的时间来陪伴家人，我时常对此感到懊恼。我有时也没法轻松表达对家人的爱，但这份爱也是切切实实的保存在我的心底。家人也许不是最知心的朋友，但永远是最温暖的港湾。感谢家人对我的宽容、理解，帮助家人过上幸福生活是我目前最大的奋斗动力。

然后，我想真切的感谢我的同门伙伴。感谢师兄师姐、师弟师妹、同届朋友们为我的生活带来欢乐，并在学习、生活上给我带来的宝贵经验。是你们让一个孤僻不善言辞的小子感受到与人产生连接的重要性，并在校园生活中感受到了家的温暖。我们不仅在学业上互勉互助，更在生活中建立了深厚的友谊。我可以自豪的说，与你们相处的这三年是我生命目前为止最幸福的时光。

最后，感谢三年中的每一个点滴。我想我会永远记住这三年中耳机里播放的音乐、健身房里沉重的哑铃、校门口常去的“星级”餐馆以及和朋友们的欢声笑语。这个地方承载了我太多的回忆，将思绪跳转至当下时才反应过来快到离别的时候了。我

觉得自己是个很幸运的人，可以把最好的时光留在这里，在这里经历的一切将我塑造成了现在这个纯粹、乐观、充满正能量的人，对我来说没有什么比这更加宝贵。

真诚的希望上面提到的每一个人都能幸福安康，过上自己想要的生活，我对你们的感激之情难以言表，只能压缩在这一页小小的论文致谢上面。我也希望我在未来的道路上能够不忘初心，让心中赤诚的火焰继续燃烧。

徐天洋

上海大学

2026 年 5 月 6 日