

西安交通大学

博士学位论文

多光谱融合智能光电处理算法与系统设计

学位申请人：陈炜煌

指导教师：孙宏滨教授

学科名称：控制科学与工程

2025 年 12 月

**Intelligent Electro-Optical Processing Algorithm and Systems
Design based on Multispectral Fusion**

A dissertation submitted to
Xi'an Jiaotong University
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

By
Weihuang Chen
Supervisor: Prof. Hongbin Sun
Control Science and Technology
December 2025

博士学位论文答辩委员会

多光谱融合智能光电处理算法与系统设计

答辩人：陈炜煌

答辩委员会委员：

西安交通大学教授：辛景民 _____ (注：主席)

西北工业大学教授：王鹏 _____

西安电子科技大学教授：董伟生 _____

西安交通大学教授：魏平 _____

西安交通大学教授：杜少毅 _____

答辩时间：2023 年 05 月 14 日

答辩地点：西安交通大学科学馆 324

摘要

自动驾驶在节约驾驶成本、提高交通效率、减少环境污染等方面拥有巨大优势，成为了学术界和工业界的热门研究课题。为了实现安全可靠、稳定高效的驾驶行为，自动驾驶车辆需要精准地预测出周围环境中交通参与者的未来行为轨迹，并规划出自身无碰撞且运动学可行的短时运动轨迹。传统轨迹预测方法无法保证长期预测的精度，严重依赖启发式设计的传统运动规划方法也无法保证其泛化性能。近年来，基于数据驱动的深度学习方法得到了快速发展，为完成预测规划任务带来了新思路。从数据输入输出的角度考虑，预测和规划都是对交通参与者的历史特征进行建模后输出未来轨迹。因此，这两项具有共性的任务均可以采用具有强大特征拟合能力的深度学习方法来完成。然而，此类方法仍然存在交通参与者异质性处理能力差，缺少概率性预测结果以及无法保证轨迹平滑性等问题，使得自动驾驶的安全性受到威胁，阻碍了自动驾驶技术进一步的发展。

本文聚焦于利用 Transformer 网络解决上述核心难点问题：1) 如何构造更精准、更快的实用化轨迹预测网络模型；2) 如何在保证完成自动驾驶任务的前提下促使运动规划方法尽可能地减少交通违规行为。主要研究工作如下。

1. 提出了一种基于时空 Transformer 网络的单模态轨迹预测网络模型，弥补了之前方法只能有效预测同质交通参与者的缺陷，提高了密集交通环境下时空交互建模能力。针对之前方法对时间序列数据进行串行处理造成记忆能力弱以及空间邻域范围设置不合理等问题，该方法采用 Transformer 网络并构建了全感知域的时空图模型。整个网络包括时空 Transformer 编码器、时间 Transformer 编码器和时间 Transformer 解码器三个部分。时空 Transformer 编码器能够对时空图特征按照不同维度交替提取，从而充分融合时空信息。经过时间 Transformer 编码器对于时间信息的进一步处理后，时间 Transformer 解码器生成了关于异质交通参与者的单模态轨迹。在自动驾驶轨迹预测公开数据集上的实验结果表明，该方法比当时最好的方法在主要性能指标上提高了至少 7.2%。

2. 提出了一种基于概率性候选轨迹网络的多模态轨迹预测网络模型，在加快模型推理速度的同时，提高了多模态轨迹预测的精度。针对当前多模态轨迹预测方法无法提供概率性预测结果的问题，该方法设计了一种既能生成目标点引导信息，又能提供概率性结果的三阶段轨迹预测过程。首先，该方法利用无监督学习自动获取交通参与者的潜在意图集合，并应用分类网络筛选出符合当前交通参与者运动趋势的概率性目标点集合。然后，通过 Transformer 网络生成中间位置锚点。最后，使用连续曲线光滑连接当前位置、锚点和目标点，形成表达能力更强的概率性候选轨迹集。多个公开轨迹预测数据集的实验结果验证了该方法在提供高性能、高效率的概率性预测结果的同时，能够确保概率较高的预测结果更符合交通参与者的下一步行为。

3. 提出了一种基于安全轨迹树网络的运动规划网络模型，减少了之前基于学习的运动规划方法在完成自动驾驶任务时出现的大量交通违规行为。针对之前方法因不能满足相关运动约束而造成的违规问题，该方法提出了一种具有曲率连续性和运动学可行性的轨迹树。该轨迹树既能够用于运动规划主任务，也能够作用于共性的轨迹预测辅助任务，从而帮助模型通过学习预测规划间的交互提升性能。针对高维栅格化特征输入可解释性差、计算效率低的问题，该方法采用包含交通参与者和局部任务路线的离散化输入表达方式，增加了模型的可解释性。该方法还利用 Transformer 主干网络精准提取不同输入之间的空间交互信息。针对自动驾驶汽车在复杂场景中保持长期静止不动的问题，该方法在训练过程中引入了焦点损失函数，鼓励自动驾驶车辆安全高效地完成导航任务。多个自动驾驶闭环测试基准的实验结果表明，该方法不仅在自动驾驶任务完成度和违规得分方面比之前最好的方法分别提高了 39.2% 和 10.6%，而且推理速度加快了 1.5 倍。

综上所述，本文所提出的单模态轨迹预测、多模态轨迹预测和运动规划方法获得了高性能的表现，具有精度高、速度快和违规驾驶行为少的优势，为保证自动驾驶安全性发挥了重要作用。

关 键 词：自动驾驶；轨迹预测；运动规划；自注意力模型

论 文 类型：应用研究

ABSTRACT

Autonomous driving is an innovative and advanced research field in academia and industry, with potential to reduce road fatalities, improve traffic efficiency, and decrease environmental pollution. To achieve safe, reliable, stable, and efficient driving behavior, autonomous vehicles need to accurately predict the future trajectories of surrounding traffic participants and plan collision-free, kinematically feasible short-term motion trajectories. Traditional trajectory prediction methods often lack accuracy in long-term predictions, while motion planning methods based on heuristic design may lack generalization performance. In recent years, rapid advancements in data-driven deep learning methods have revolutionized prediction and planning tasks. Deep learning methods offer powerful feature fitting capabilities that enable accurate modeling of historical traffic patterns and output of future trajectories. Consequently, both prediction and planning tasks can be achieved using deep learning techniques. Despite these remarkable benefits, these methods still face various challenges, such as poor ability to deal with traffic participant heterogeneity, lack of probabilistic prediction results, and inability to guarantee trajectory smoothness. These issues pose significant safety concerns for autonomous driving and impede the further advancement of this technology.

This dissertation proposes to leverage Transformer network to address these core difficulties. The objectives of this study are twofold: 1) improving the accuracy and inference speed of practical trajectory prediction models, 2) enhancing motion planning methods to minimize traffic violations while guaranteeing the completion of autonomous driving tasks. The main contributions of this research are as follows.

1. This dissertation proposes a new Spatio-Temporal Transformer Network for unimodal trajectory prediction, which addresses the limitations of previous homogeneous prediction methods and improves spatio-temporal interactive modeling capabilities. To address the problems of weak memory ability and unreasonable setting of spatial neighborhood range caused by the serial processing of time series data in the previous method, we adopt Transformer network and constructs a spatio-temporal graph of the whole perceptual domain. The network consists of three parts, i.e. spatio-temporal Transformer encoder, temporal Transformer encoder, and temporal Transformer decoder. The first Transformer encoder extracts spatio-temporal features by alternating between different dimensions to fully integrate spatio-temporal information. The second Transformer encoder further processes temporal information, and the temporal Transformer decoder generates unimodal trajectories for heterogeneous traffic participants. Experimental results demonstrate that the proposed method enhances key performance metrics by at least 7.2% over state-of-the-art methods.

2. This dissertation proposes a new Probabilistic Proposal Network for multimodal trajectory prediction which not only enhances the prediction accuracy of multimodal trajectory prediction, but also accelerates the inference speed. To address the problem that previous multimodal trajectory prediction methods cannot provide probabilistic prediction results, we devise a three-stage trajectory prediction process that generates target point guidance information and provides probabilistic outcomes. Firstly, the proposed method employs unsupervised learning to automatically obtain the potential intention set of traffic participants and applies a classification network to filter out a set of probabilistic target points that comply with the current movement trend of traffic participants. Next, Transformer network generates intermediate position anchors. Finally, a continuous curve is used to smoothly link the current position, anchors, and target point, producing a more expressive set of probabilistic trajectory candidates. Experimental results demonstrate that the proposed method yields high-performance and high-efficiency probabilistic prediction results while ensuring that the prediction results with higher probability align more closely with the next behavior of traffic participants.

3. This dissertation proposes a new safe Trajectory Tree Network for motion planning, which can effectively reduce traffic violations while completing autonomous driving tasks. The key component of TTNet is a predefined trajectory tree that conforms to vehicle dynamics constraints and explicitly reflects different intentions. This tree is used for both the main planning task and an auxiliary trajectory prediction task. To enhance interpretability, we introduce input expressions typically used in traditional planning algorithms into our integrated framework. Additionally, to promote safe and efficient navigation, we incorporate a focal loss during training and employ a Transformer-based backbone network to accurately capture spatial interactions not only among the ego vehicle and its surroundings, but also among dynamic agents and the reference line. Experimental results demonstrate that the proposed method significantly improves task completion and violation scores by 39.2% and 10.6%, respectively, compared to SOTA methods while accelerating the inference speed by 1.5 times.

In summary, our proposed methods achieve outstanding performance for unimodal trajectory prediction, multimodal trajectory prediction and motion planning, with the advantages of high precision, high speed and less driving violations, thus playing crucial roles in ensuring the safety of autonomous driving.

KEY WORDS: Autonomous Driving; Trajectory Prediction; Motion Planning; Transformer

TYPE OF DISSERTATION: Application Research

目 录

摘要	I
ABSTRACT	III
1 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状	4
1.2.1 机载光电系统总体研究进展	4
1.2.2 机载光电系统小目标检测研究现状	5
1.2.3 机载光电系统目标跟踪研究现状	11
1.3 论文的主要贡献	15
1.4 论文的组织结构	17
2 智能光电处理算法与系统设计研究基础	19
2.1 智能光电系统基本组成	19
2.2 机载光电系统目标检测算法研究基础	22
2.2.1 基本方法	23
2.2.2 常用数据集	28
2.2.3 评价指标	32
2.3 机载光电系统目标跟踪算法研究基础	33
2.3.1 基本方法	34
2.3.2 常用数据集	37
2.3.3 评价指标	41
2.4 本章小结	42
3 基于双重注意力处理的高效可见光航拍图像小目标检测算法	43
3.1 引言	43
3.2 相关工作	45
3.2.1 基于 CNN 的小目标检测方法	45
3.2.2 基于 DETR 的小目标检测方法	46
3.3 基于双重注意力处理的目标检测网络	46
3.3.1 整体架构	46
3.3.2 双重注意力处理块	47
3.3.3 双融合特征编码器	49
3.3.4 损失函数	50
3.4 实验结果与分析	51
3.4.1 数据集与评价指标	51
3.4.2 实现细节	51
3.4.3 与 SOTA 的对比实验	53

3.4.4 消融实验	54
3.4.5 可视化结果	57
3.5 本章小结	58
4 基于多特征聚焦与跨阶段 Transformer 的红外小目标检测网络	59
4.1 引言	59
4.2 相关工作	61
4.2.1 面向嵌入式平台的红外小目标检测	61
4.2.2 多尺度特征学习方法	62
4.3 基于多特征聚焦与跨阶段 Transformer 的检测网络	62
4.3.1 模型架构	62
4.3.2 多特征聚焦模块	64
4.3.3 基于深度分离卷积的跨阶段 Transformer	64
4.4 实验结果与分析	67
4.4.1 数据集与评价指标	67
4.4.2 与 SOTA 的对比实验	68
4.4.3 消融实验	70
4.4.4 可视化结果	73
4.5 本章小结	75
致谢	76
参考文献	77
攻读学位期间取得的研究成果	87
答辩委员会会议决议	88
常规评阅人名单	89
声明	

CONTENTS

ABSTRACT (Chinese)	I
ABSTRACT (English)	III
1 Introductions.....	1
1.1 Research Background and Significance	1
1.2 Domestic and International Research Status	4
1.2.1 Overall Research Progress of Electro-Optical Systems	4
1.2.2 Research of Small Object Detection for Electro-Optical Systems	5
1.2.3 Research of Single Object Tracking for Electro-Optical Systems	11
1.3 Major Contributions	15
1.4 Thesis Organization	17
2 Research Foundation of Intelligent Electro-Optical Algorithms and System Design	19
2.1 Fundamental Architecture of Intelligent Electro-Optical Systems.....	19
2.2 Research Foundations of Object Detection for Airborne Electro-Optical Systems.	22
2.2.1 Fundamental Approaches.....	23
2.2.2 Commonly Used Datasets	28
2.2.3 Evaluation Metrics.....	32
2.3 Research Foundations of Object Tracking for Airborne Electro-Optical Systems..	33
2.3.1 Fundamental Approaches.....	34
2.3.2	37
2.3.3	41
2.4 Chapter Summary.....	42
3 Efficient Drone Object Detection Network Based on Bipartite Attentive Processing....	43
3.1 Introduction.....	43
3.2 Related Work	45
3.2.1 CNN-based Small Object Detection Methods	45
3.2.2 DETR-based Small Object Detection Methods.....	46
3.3 Bipartite Attentive Processing Detection Network.....	46
3.3.1 Overall Architecture	46
3.3.2 Bipartite Attentive Processing Block	47
3.3.3 Dual-Fusion Feature Encoder	49
3.3.4 Loss Function	50
3.4 Experimental Results and Analysis	51
3.4.1 Datasets and Evaluation Metrics.....	51
3.4.2 Implementation Details	51
3.4.3 Comparison With State-of-The-Art Methods	53

3.4.4 Ablation Study	54
3.4.5 Visualization.....	57
3.5 Chapter Summary.....	58
4 Multi-Feature Focus and Cross-Stage Transformer Network for Infrared Small Object Detection	59
4.1 Introduction.....	59
4.2 Related Work	61
4.2.1 Infrared Small Object Detection for Embedded Platforms.....	61
4.2.2 Multiscale Feature Learning Methods	62
4.3 Multi-Feature Focus and Cross-Stage Transformer Network.....	62
4.3.1 Model Architecture	62
4.3.2 Multi-Feature Focus Module	64
4.3.3 Depth-wise Cross-stage transFormer.....	64
4.4 Experimental Results and Analysis	67
4.4.1 Datasets and Evaluation Metrics.....	67
4.4.2 Comparison With State-of-The-Art Methods	68
4.4.3 Ablation Study	70
4.4.4 Visualization.....	73
4.5 Chapter Summary.....	75
Acknowledgements.....	76
References	77
Achievements	87
Decision of Defense Committee	88
General Reviewers List	89
Declarations	

1 绪论

1.1 研究背景与意义

在新型飞行平台技术成熟与低空应用生态加速扩张的双重推动下，低空经济作为融合空域资源开发与前沿技术应用的战略性新兴产业，正成为培育新质生产力、推动经济结构优化升级的新增长引擎。低空经济泛指在垂直高度 1000 米以下的空域范围内，以各类有人驾驶和无人驾驶航空器为载体，以多元化的低空飞行活动为牵引，辐射带动相关领域融合发展的综合性经济形态。它不仅旨在破解传统地面交通拥堵、提升偏远地区通达性，更是构建“空天地一体化”综合立体交通网、推动经济社会高质量发展的战略支点。作为低空经济的核心载体，以无人机（Unmanned Aerial Vehicle, UAV）和电动垂直起降航空器（electric Vertical Take-Off and Landing, eVTOL）为代表的智能航空器正发挥着前所未有的关键作用，其卓越的灵活性与适应性，催生了城市空中交通（Urban Air Mobility, UAM）等新兴交通运输范式，并推动应用场景从传统的通用航空运营，广泛渗透至物流配送、农业植保、电力巡检、环境监测、应急救援乃至公共安全等众多领域，产生了颠覆性的影响。这些智能航空器正从根本上提升作业效率并重塑行业生态，凭借其低碳环保、噪声低、运行成本低等技术优势，正推动基础设施巡检、广域态势感知、应急救援等应用场景从传统人工操作向无人化智能作业模式的深刻变革。低空经济的蓬勃发展与广泛应用，对其核心载体无人机的智能化水平提出了极高要求。然而，实现复杂动态环境下的自主感知、决策与行动，其关键在于飞行平台能否具备实时、精准、可靠的环境感知与信息处理能力。在此背景下，智能光电系统的重要性日益凸显，并已成为无人机的核心组件与实现其智能化的关键使能技术。

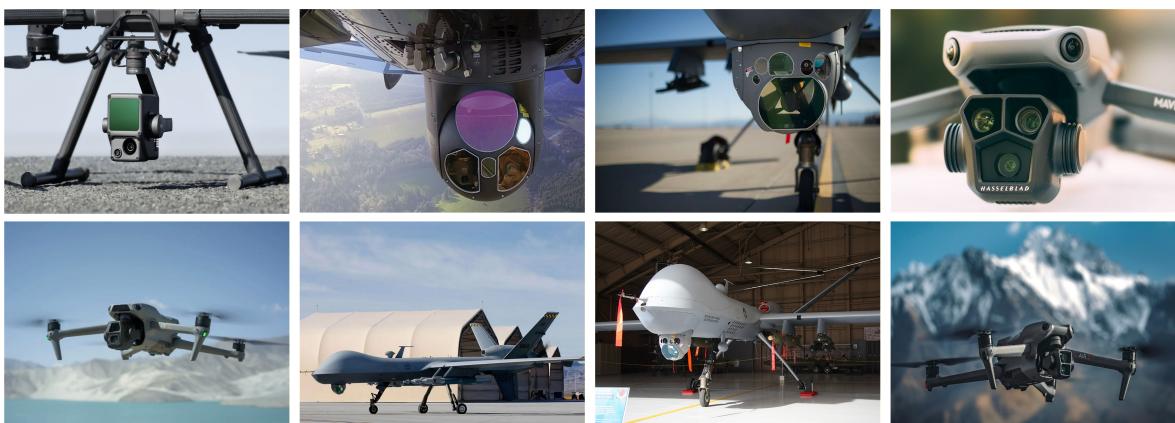


图 1-1 无人机与其搭载的光电系统

无人机光电系统是一种基于光电转换原理的综合性探测系统，其核心功能是通过接收目标反射或辐射的电磁波，将其转换为电信号并进行处理，从而实现对目标的探测、识别、跟踪与测距。该系统是现代无人机实现环境感知与自主飞行的核心，其基本

构成包含三大核心组件：光电传感器，稳定平台和信息处理系统。光电传感器负责捕获从紫外到远红外波段的电磁波信息。典型系统采用多谱段集成设计，主要包括：1) 可见光摄像机，基于 CCD 或 CMOS 技术，在昼间提供高分辨率图像；2) 红外热像仪，通过检测目标与背景的温差实现夜间观测，常用中波红外与长波红外波段以契合大气传输窗口；3) 激光测距仪，作为主动探测组件，通过测量激光飞行时间提供精确的目标距离信息。此外，先进系统正逐步集成短波红外、高光谱成像仪乃至激光雷达等传感器，极大丰富了信息的维度与层次。稳定平台用于隔离无人机飞行过程中产生的振动和姿态变化，确保光电传感器视轴的稳定指向。它采用多轴陀螺稳定技术，通过惯性测量单元实时感知无人机飞行带来的振动与姿态变化，并驱动伺服机构产生反向补偿运动，以维持传感器视轴的稳定指向。信息处理系统对传感器获取的原始数据进行处理、分析和利用，包括图像增强、降噪、和电子稳像等基础处理任务，运用计算机视觉与深度学习算法实现目标的自动检测、识别与持续跟踪的智能分析任务，以及对于多谱段图像、激光测距与惯性导航数据的整合压缩以及融合，生成统一的目标信息和态势数据。

无人机光电系统的发展经历了从功能单一的观测设备到多传感器融合的智能任务系统的演进。这一历程始于 20 世纪 60-80 年代，彼时的系统功能相对简单，主要以可见光电视摄像为核心，配备基础的光学望远镜和简单的机械稳定装置。该系统体积重量大、分辨率和稳定性有限，且基本无法进行夜间观测。直至 70 年代，随着红外技术起步，系统开始集成第一代需要光机扫描的红外热像仪，采用单元或线列探测器，灵敏度较低，且整个系统的智能化程度极低，完全依赖操作员的肉眼识别和手动控制。进入 20 世纪 90 年代至 21 世纪初，凝视型焦平面阵列技术和数字信号处理技术的成熟带来了性能的质的飞跃。基于 InSb 或 HgCdTe 材料的焦平面阵列红外探测器成为主流，实现了高灵敏度红外成像，彩色 CCD 相机也取代了早期的黑白电视相机。在稳定平台方面，基于光纤陀螺或激光陀螺的多轴稳定平台成为标准配置，显著提升了稳定精度。此阶段，可见光、红外与激光测距/指示器构成的典型“三光系统”成为高端无人机的标配，并开始引入自动目标识别和电子稳像等初步智能算法，实现了从侦察到打击的一体化能力，其代表性系统如美国 MQ-1B “捕食者” 无人机搭载的 AN/AAS-52 光电系统。从 2010 年至今，无人机光电系统向着多光谱、高分辨率与高度智能化方向迅猛发展。传感器范畴已远超传统可见光与红外，扩展至短波红外、中波红外、高光谱成像及偏振成像等新体制传感器，例如美国 Raytheon 公司的 RAVEN 系统还融合了激光雷达以创建环境 3D 图像。在信息处理层面，人工智能与深度学习算法的深度集成，赋予了系统更强大的自动目标识别、跟踪与智能分析能力。同时，系统架构走向开放化，支持功能的快速扩展与升级。现代系统，如以色列 RAFEL 公司的 Litening-5 多功能吊舱，已演变为集情报、监视、目标捕获与侦察于一身的完整任务系统，标志着其从独立的“传感器”向“传感-互联-情报生成”综合节点的彻底转型。

尽管无人机光电系统在技术上取得了显著进步，但就其智能化程度而言，整体仍处于初级发展阶段，尚未达到完全自主的高级智能水平。现有的智能化功能更多体现在辅助决策层面，而非完全的自主决策。在目标识别与跟踪方面，现有系统已能实现一

定程度的自动检测和识别，但多局限于预设类别的目标，且对图像质量、环境条件有较高要求。目前的目标识别算法在背景简单、目标清晰的情况下表现良好，但在复杂背景或伪装欺骗条件下，性能会显著下降。具体体现在以下三个方面，首先，现有识别算法对于航拍视角下的小目标检测效果普遍不佳。由于无人机通常在百米乃至千米级高度进行广域侦察，地面对目标在图像中所占的像素占比极小，往往不足图像的 0.1%。这种小目标所携带的图像特征非常微弱，在经历了模型的下采样操作后，其细节信息几乎损失殆尽，导致主流检测模型（如 YOLO、DETR 等）的特征提取网络难以有效捕获其差异性特征，从而造成大量的漏检与误检。其次，现有跟踪算法在面对目标遮挡及复杂背景干扰时，难以实现鲁棒且稳定的持续跟踪。无人机在执行任务时，目标易被树木、建筑物等静态物体短暂遮挡，或混入人群、车流等相似动态背景中。当前以相关滤波和孪生网络为代表的跟踪算法，通常不具备判丢和重捕获的能力，目标在被遮挡时模型无法及时感知，现象通常是跟踪框定在遮挡物上或跳变至其他目标，在目标重现后因模型更新错误或判别能力不足无法重新锁定正确的目标，导致跟踪失败。最后，红外图像中弱小目标识别面临着比可见光图像更为严峻的挑战。红外成像依赖于目标与背景的热辐射差异，但其固有的低分辨率、低对比度和缺乏丰富纹理细节的特性，使得小目标信噪比极低，在图像中几乎与背景噪声融为一体。这不仅使人眼观察困难，也使得依赖形状、纹理等高层语义特征的通用识别算法效能骤降。如何从信噪比低、背景杂波强的红外图像中稳定、准确地分离出弱小目标，是目前智能光电处理领域一个尚未完全解决的痛点。

由于功耗和尺寸的限制，无人机光电系统通常采用嵌入式边缘计算设备作为信息处理核心，这类设备的计算资源、内存带宽和功耗预算都极为有限，严重制约了复杂智能算法的实时性能。目前基于 Transformer 的主流目标识别和跟踪算法参数量大，计算复杂度高，难以在边缘设备实现实时处理。同时，算法复杂度与精度之间的平衡面临严峻挑战，为了提升对小目标的检测能力，现有研究往往通过增加网络深度、增大图像分辨率等增加计算负载的方法，使其更加难以在边缘设备上部署，无法在任务现场实现端到端的实时智能感知，严重削弱了无人机在动态环境中自主决策与快速响应的能力。

综上所述，面向无人机光电系统智能化的迫切需求，以及当前光电系统在算法层面与硬件层面上面临的双重挑战，本文旨在研究一套高效、鲁棒的多光谱融合智能光电处理算法，并据此设计一套与之适配的嵌入式系统架构，以显著提升无人机在真实复杂场景下的自主感知与决策能力。本文将重点研究可见光航拍图像中的小目标识别，通过设计专门的特征增强模块与多尺度融合机制来提升微小目标的检测精度，探索复杂场景下的抗遮挡长时目标跟踪方法，结合重识别技术与自适应模型更新策略，确保目标在遮挡干扰下的持续稳定跟踪，同时，针对红外航拍图像中的小目标识别难题，开展面向低对比度红外图像的弱小目标检测网络架构研究，结合轻量化网络与边缘推理优化，有效提升低信噪比环境下弱小目标的检测能力，最终通过智能光电系统嵌入式架构的软硬件协同设计，优化算法在边缘设备上的计算效率与资源分配，实现从算法创新到系统集成的完整技术闭环，为推进无人机光电系统的智能化升级提供理论与实

践支撑。

1.2 国内外研究现状

1.2.1 机载光电系统总体研究进展

机载光电系统指的是集成在有人或无人飞行平台，用于执行侦察、监视、目标识别与跟踪等多重任务的光电系统，是一个涉及了光学、电子、机械、计算机软硬件等领域的综合系统。其工作原理是通过集成多种光电传感器，将所处环境的光学特征或红外辐射等信息转换为电信号，并经后续处理形成可供人工判读或机器自动识别的图像与数据，从而为侦察、跟踪、测量等任务提供关键信息输入。机载光电系统的发展经历了功能单一到高度集成化的过程，早期的光电系统只配备单一传感器，而现代光电系统则集成了多波段、多类型的传感器，并朝着与飞行平台深度耦合的方向发展。

基于多光谱探测与稳定平台技术的持续突破，国际上已形成多条技术路线各异、功能侧重点不同的代表性系统。美国 FLIR 系统公司生产的 Star SAFIRE HD 光电系统是高度集成化的典范，其最多可同时装载八种传感器，包括高分辨率中波红外焦平面传感器、高清彩色可见光传感器、微光与短波红外传感器，以及激光测距仪、指示器和惯性测量组件。其可见光与红外相机均能实现高达 $\times 120$ 倍的光学连续变焦，视场分别覆盖 $29^\circ \sim 0.25^\circ$ 和 $40^\circ \sim 0.35^\circ$ 的广阔范围，从而无缝衔接大视场搜索与小视场精细识别的任务需求。在高端战略侦察领域，Northrop Grumman 公司为“全球鹰”高空长航时无人机设计的综合传感器套件（EISS）展现了系统级融合的先进理念。该套件集成了合成孔径雷达、移动目标指示器、高清可见光相机和第三代红外传感器，具备全天候、全时段侦察能力。其光电载荷采用全反射式双波段共孔径光学系统成像，并创新性地运用高精度两轴稳定平台与快速反射镜相结合的全数字复合控制技术，有效实现了广域扫描搜索、动目标检测、高精度稳像及像移补偿，从而在高速飞行条件下仍能获取稳定的高分辨率图像。

在当前技术发展背景下，前沿研究主要聚焦于系统轻量化、智能化和实时处理能力的协同突破。美国 Logos 公司开发的“轻型多模式”战术光谱和侦察成像”（SPRITE）吊舱是该领域的典型代表。它将超轻型广域运动成像（WAMI）传感器、高清侦察相机和短波红外高光谱传感器集于一身，并在吊舱内集成了一个手掌大小的“多模式边缘处理器”（MMEP）。该处理器能以每秒 10 亿像素和每秒 300 万光谱的极高速度处理不同类型数据，使得三种传感器能够相互指示、协同工作，极大地提升了无人机在复杂环境下发现、识别与跟踪目标的自主能力。高精度视轴稳定技术构成了机载光电系统成像质量保障的核心基础，其中 L3Harris WESCAM 公司开发的 MX 系列光电系统在此方面取得了显著突破。该系列产品采用创新的五轴稳定架构与独特的传感器-稳定机构隔离设计，使其集成的多种高性能传感器（涵盖 1080p 高清中波红外传感器、连续变焦高清可见光相机及激光测距仪等）能够获得一致的稳定性能。其采用的模块化现场可更换单元设计，使得传感器更换后无需重新校准，显著提升了系统的可维护性与适应性。

以色列 RAFAEL 公司的 RecceLite XR 光电侦察系统，集成了可见光、近红外、中波红外和短波红外四种波长传感器，探测范围扩展至 80 公里以上，并结合先进的图像处理算法，支持广域持续监视与防区外侦察。Elbit Systems 公司推出的新一代先进多传感器有效载荷系统 (AMPS NG) 采用了双前视红外传感器设计并增加了短波红外传感器，这种配置能有效克服高湿度、烟尘或灰尘等恶劣环境的干扰，实现了在数十公里外对地面及空中目标的高分辨率成像与稳定跟踪，展现了其在复杂电磁与环境条件下的强大生存与作战能力。

当前机载光电系统的发展主要集中体现在硬件层面的持续优化，包括但不限于传感器谱段的扩展、探测器分辨率的提升以及稳定平台精度的改进。这些技术进步提升了系统数据采集与前端处理的性能与可靠性，其智能化水平依然存在明显局限。具体而言，现有系统普遍缺乏对感知内容的深度理解能力，所搭载的算法多局限于基础图像增强与人工辅助识别，难以在复杂战场环境下实现真正意义上的自主探测、智能识别与决策支持。在行业应用深化与技术发展的内在驱动下，现代机载光电系统呈现出多维度的协同演进趋势。在感知层面，为适应复杂环境与多样化探测需求，系统正朝着多光谱融合、高分辨率与高帧频的方向发展，以全面提升环境感知精度。为实现精准观测与稳定跟踪，系统需具备更强的扰动抑制能力，并与导航定位系统深度融合，实现高精度的自主定位与目标运动分析。面对海量异构数据，信息处理系统需引入边缘计算与智能算法，实现高效的特征提取与目标识别。同时，小型化、轻量化与成本控制成为系统规模化应用的关键，这依赖于光机电一体化设计与新型材料的创新应用。最终，系统功能正从单一感知向多任务集成方向发展，通过智能化处理逐步完成从数据采集设备到智能感知节点的能力升级。

1.2.2 机载光电系统小目标检测研究现状

机载光电系统的小目标检测技术是实现远程精准感知的核心挑战之一，其核心任务是在远距离成像条件下，从复杂背景中识别出像素占比极低、特征微弱的兴趣目标。在军事侦察、灾害监测、安防巡逻等应用中，目标的远距离成像特性导致其在图像中仅占据极少数像素点，同时受到复杂背景杂波、低信噪比以及目标纹理信息匮乏等多重因素的严重制约。早期的研究多基于传统模型驱动方法，例如利用局部对比度测量或低秩稀疏分解理论来增强目标与背景的差异性。这类方法在简单场景下具有一定效果，但其性能高度依赖于人工设计的先验假设，在面临强背景干扰、目标尺度变化或低照度条件时，常出现虚警率升高、鲁棒性不足的问题。随着人工智能技术的发展，基于深度学习的方法已成为当前研究的主流。深度神经网络通过端到端的学习方式，能够自动从大规模数据中提取更具判别力的特征，显著提升了模型在复杂环境下的泛化能力。为应对上述挑战，现有小目标检测方法通常在通用目标检测的成熟框架中引入针对性设计，通过改进网络结构、优化特征提取机制和增强上下文建模能力等方式，解决小目标检测中的难题。具体的方法包括：数据增强，多尺度融合，超分辨率，上下文建模，注意力机制，聚焦检测。

1) 基于图像增强的小目标检测方法

在深度学习领域，训练样本的不足会直接导致模型性能的显著下降。因此，利用大量数据进行训练是确保模型获得强大泛化能力的关键。然而，鉴于小目标标注成本高昂，现有数据集的标注样本数量远不能满足需求。数据增强技术是丰富数据集多样性、提升模型鲁棒性与泛化能力的常用策略，同时也有助于缓解因数据集中不同尺度目标分布不均衡而导致的检测精度下降问题。早期的数据增强主要依赖于基础的几何变换（如旋转、缩放、裁剪和平移）和颜色变换，这些方法通过对图像像素进行重分布来增加数据多样性。尽管已有很多经典的数据增强方法被提出，如 Mosaic^[1]、MixUp^[2] 和 CutMix^[3]，但这些通用方法对中大型目标的性能提升通常优于小目标，在小目标检测场景下效果有限。因此，越来越多的学者开始专注于研究针对小目标检测的专用数据增强技术。Kisantal 等人^[4]提出了两种创新策略：一是对包含小目标的图像进行过采样，二是对小目标进行多次复制-粘贴。他们通过系统实验比较了不同复制策略，发现复制全部小目标的效果最优，而仅复制单个小目标虽然能提升小尺度目标检测，但会对大尺度目标产生负面影响。为进一步优化上述方法，Chen 等人^[5]提出了自适应数据增强技术，通过引入反馈机制解决随机复制粘贴可能导致的背景失配与目标尺寸失配问题。Xiao 等人^[6]则提出了 Copy-Reduce-Paste 方法，该方法通过将所有大尺寸目标缩放至小目标范畴，有效平衡了训练过程中的尺度分布。

在数据处理层面，Ünel 等人^[7]提出了一种创新的分块方法，通过裁剪图像增大小目标的相对面积，同时保留完整图像输入以确保对大目标的检测能力，最终通过融合多尺度检测结果实现性能提升。针对预训练与微调阶段的数据尺度不匹配问题，Yu 等人^[8]提出了尺度匹配策略，通过智能裁剪减小尺度差距，这一方法显著提升了 FPN 检测器 5% 的检测性能。Lin 等人^[9]提出一种新的尺度感知模块（scale-aware network for semantic segmentation of high-resolution aerial images, SAN）通过自适应重采样策略解决了遥感影像中的尺度不连续性问题，而 Zoph 等人^[10]提出的自搜索学习框架则采用强化学习方法自动探索最优数据增强策略组合，为小目标检测的性能优化开辟了新途径。

尽管数据增强方法旨在通过增加小目标数量来缓解正样本稀缺这一核心问题，但其自身存在明显的局限性。这类方法通常表现出性能提升的不稳定与较差的迁移性，即某种增强策略在特定数据集或模型上有效，但难以泛化至其他场景。其根本原因在于，许多数据增强方法（如简单的复制-粘贴）可能引入不真实的上下文或背景失配问题，未能从根本上优化模型对于小目标本质特征的学习能力。

2) 基于多尺度融合的小目标检测方法

在基于深度学习的目标检测中，小目标的弱特征表示是导致其检测性能不佳的主要原因。经过卷积和池化的多次下采样操作后，最终特征图中包含的小目标特征信息极少。此外，神经网络会生成不同分辨率的特征图，深层特征图具有更大的感受野、更强的语义信息，但其分辨率较低，丢失了大量细节信息。相反，浅层特征图分辨率较高，

保留了丰富的细节和位置信息，但缺乏足够的语义信息。为了同时利用深层特征的强语义性和浅层特征的精细空间细节，研究者们主要沿着两条技术路径进行探索：一是构建专用尺度检测器，通过多分支架构或定制化训练方案使不同层级负责不同尺度的目标；二是进行分层特征融合，整合不同深度的特征以构建对小目标更强大的表征能力。这两种路径的核心都在于最大限度地减少特征提取过程中的信息损失。

(1) 专用尺度检测器

该类方法的本质是令网络中不同深度的特征图专注于检测相应尺度的目标，从而实现更优的感受野匹配。早期工作如 Yang 等人^[11]提出的尺度依赖池化 (scale-dependent pooling, SDP)，通过为小目标选择合适的特征层进行后续操作。MS-CNN^[12]则在不同中间层生成候选区域，使每一层专注于特定尺度范围的目标。单阶段检测器如 YOLO 系列^[13]，通过添加并行分支，利用高分辨率特征负责小目标预测。具有里程碑意义的工作是 Lin 等人^[14]提出的特征金字塔网络 (Feature Pyramid Networks, FPN)。FPN 通过自上而下的路径和横向连接，构建了具有强语义信息的多尺度特征金字塔，并依据目标尺寸将其分配至不同金字塔层级进行检测。这种简单高效的设计已成为现代检测器的标准组件，并催生了一系列卓越的变体，如 NAS-FPN^[15]和 Recursive-FPN^[16]。此外，研究者还尝试组合多个尺度专用检测器，Li 等人^[17]构建了并行子网络，其中小尺度子网络专门用于检测小尺寸行人。TridentNet^[18]构建了具有不同膨胀卷积的并行多分支架构，使每个分支对特定尺度目标具有最优感受野。QueryDet^[19]则设计了级联查询策略，有效避免了在低层特征上的冗余计算，使得能够高效地利用高分辨率特征图检测小目标。在训练策略上，研究者们也提出了针对性的数据准备方法。Singh 等人^[20]提出了尺度归一化图像金字塔，该范式只对落在预定尺度范围内的目标实例进行训练，其余则被忽略，从而确保小目标能在最合理的尺度上被处理而不影响中大目标的性能。后续的 Sniper^[21]从多尺度图像金字塔中采样芯片以进行高效训练。Najibi 等人^[22]提出了一种由粗到细的检测流程。Chen 等人^[23]则设计了一种反馈驱动的训练范式，动态指导数据准备并平衡小目标的训练损失。

(2) 分层特征融合

深度神经网络天然生成具有不同空间分辨率的层次化特征图，低层特征蕴含丰富的细节和定位信息，而高层特征则捕获更强的语义信息。对于小目标检测任务，深层特征可能因小目标响应消失而失效，而浅层特征又易受光照、形变等因素干扰，分类辨识度低。为克服此困境，特征融合方法通过整合不同深度特征，以获得对小目标更鲁棒的表征。受 FPN 启发，PANet^[24]通过增加自底向上的路径增强，利用精准的定位信号丰富了深层特征。Tan 等人^[25]提出了双向特征金字塔网络，以更高效、直观的方式进行多尺度特征融合，旨在为小目标提供更合适的表征，并实现更好的精度效率权衡。Zhang 等人^[26]将多个深度上的 RoI 池化特征与全局特征拼接，为小目标获得更具判别力的表示。Woo 等人^[27]提出的 StairNet 利用反卷积放大特征图，这种基于学习的

上采样函数能产生比传统基于核的方法更精细的特征，并促进不同金字塔层级间信息的有效传播。M2Det^[28]构建并行分支以级联方式描述从浅到深的特征，并利用 U 形模块捕获小目标的更多细节。Liu 等人^[29]提出的 IPG-Net，将图像金字塔得到的不同分辨率图像输入特定变换模块，以补充空间信息和细节。Gong 等人^[30]设计了一种基于统计的融合因子，用以控制相邻层级间的信息流。针对 FPN 类方法中存在的梯度不一致问题，SSPNet^[31]通过强调不同层级的特定尺度特征，并利用 FPN 中相邻层级的关系来实现更合理的特征共享。

专用尺度检测器致力于在最合理的尺度上处理小目标，而基于融合的方法旨在弥合低金字塔层级与高层级之间的空间和语义差异。然而，前者通常直接将不同尺寸目标映射到对应层级，这可能因单层信息不足而误导检测器，后者则面临网络内信息流并不总是对小目标表征有利的挑战。研究者的目标是既要赋予低层特征更多语义，又要防止小目标的原始响应被深层信号所淹没。如何在增强语义与保留细节之间取得最佳平衡，仍是当前多尺度融合技术面临的核心困境。

3) 基于超分辨率的小目标检测方法

小目标检测的本质困境在于其有限的像素数量所导致的特征信息匮乏，尽管多尺度融合等方法试图从空间维度上弥补信息损失，但直接从图像层面提升小目标的分辨率与质量是另一条更直接的解决思路。超分辨率技术旨在从低分辨率图像中恢复出对应的高分辨率图像。高分辨率图像能够提供更多原始场景的细节，这为小目标检测提供了有力的支撑。近年来，基于生成对抗网络 (GAN)^[32]的算法在图像超分领域取得了显著进展，并被成功应用于小目标检测任务中。典型的基于 GAN 的超分辨率框架包含两个核心子网络：一个生成器网络和一个判别器网络。生成器负责以低分辨率的小目标特征或图像块作为输入，生成高分辨率输出，试图“欺骗”判别器。判别器则作为一个对手，努力区分输入是来自真实的高分辨率图像，还是生成器产生的“伪造”超分图像。通过这种对抗性训练，生成器被驱动学习从低分辨率小目标到高分辨率表示的复杂映射，从而恢复出对小目标分类与定位至关重要的细节信息。

Perceptual GAN^[33]是首次将 GAN 应用于小目标检测任务的代表性工作。该模型引入了一个新颖的条件生成器，它以小目标的低层特征作为输入，生成具有更多细节的超分特征表示。生成器内部包含多个残差块，用于学习小目标与相似大目标之间的特征差异。判别器包含两个分支：对抗分支负责区分生成的小目标超分区域与真实的大目标区域，感知分支则直接在生成的超分表示上执行常规的目标检测任务。在训练中，两个分支都致力于最小化各自的损失，生成器被训练成尽可能使判别器做出错误判断，从而形成有效的对抗学习。SOD-MTGAN^[34]为了解决生成图像不够清晰的问题，在标准 GAN 框架中引入了细化模块。判别器被设计为一个多任务判别器网络，同时执行三个任务：区分图像的真伪、预测目标类别以及优化边界框回归。分类损失和回归损失会反向传播到生成器，从而直接指导生成器产生更易于被准确分类和定位的超分图像。生成器的总损失函数由对抗损失、像素级均方误差损失、分类损失和边界框回归损失

共同构成，这种多目标优化强制要求重建的图像不仅在视觉上逼真，而且必须包含对检测任务有益的高频细节。JCS-Net^[35]专注于小尺度行人检测，它将分类子网络和超分辨率子网络集成在一个统一的架构中。超分辨率子网络采用了类似于 VDSR^[36]的残差架构，通过探索大尺度行人与小尺度行人之间的关系，来恢复小尺度行人的细节信息。因此，重建后的小尺度行人特征既包含了原始信息，也融入了超分子网络输出的增强信息。在训练阶段，该模型通过结合分类损失和超分辨率损失进行端到端优化，并使用多层通道特征^[37]与多尺度表示来进一步增强检测性能。

使用超分辨率的方法能够有效增强图像的细节信息，但是 GAN 本身训练难度大，生成器与判别器之间的动态平衡不易达成，训练过程的不稳定会直接影响最终模型的性能，在训练过程中，如果生成器只能产生有限多样性的样本，学习过程可能会过早停滞，导致生成质量下降，进而增加最终检测的误差。

4) 基于上下文建模的小目标检测方法

小目标由于自身像素有限、特征表达能力弱，仅凭其自身信息难以实现精准的识别与定位。在此背景下，利用其周围的上下文信息成为提升小目标检测性能的关键途径之一。上下文信息提供了目标所处环境或其与周围元素的关联性线索，能够为判别小目标提供辅助依据。根据信息提取范围的不同，上下文建模可分为全局语义建模和局部上下文建模，全局语义建模关注整幅图像的统计特性与场景语义，为理解目标出现的宏观环境提供指导，局部上下文建模聚焦于目标邻近区域的细节特征，通过利用目标与周边像素、纹理或相邻物体之间的空间与语义关系，来补偿小目标自身的信息缺失，从而增强其表征的区分度与鲁棒性。

局部上下文主要指目标紧邻区域的像素级信息，如边缘、颜色和纹理等。一种直接且有效的利用方式是通过扩大检测窗口来包含更多的周围环境。Cai 等人^[12]提出将目标的边界框扩大至原区域的 1.5 倍，以此引入更丰富的周边信息。这些额外的上下文信息随后与目标自身的特征相结合，共同输入检测层进行预测。Fu 等人^[38]通过采用带有“跳跃连接”的反卷积层来增强特征的语义信息。与简单地在卷积层上堆叠反卷积层不同，该方法设计的反卷积层更浅，并使用了逐元素乘积操作进行特征融合，从而在小目标检测上取得了更好的效果。CAB-Net^[39]通过设计独特的上下文感知模块，采用金字塔扩张卷积有效融合了多层次上下文信息，同时保持了特征图的原始分辨率，显著提升了小目标检测精度。SEPN^[40]构建了专门的上下文增强模块来提取多尺度特征，该模块核心由并行双分支构成：ASPP 分支用于扩大感受野以捕捉更广泛的上下文信息，而金字塔卷积分支则有效补充了 ASPP 卷积过程中可能丢失的细节特征，二者协同工作共同增强了模型对小目标的表征能力。Corseil^[41]等人提出了一种基于 YOLOv5 的时空算法，通过利用视频序列中可用的时序上下文信息，在人员检测和空中监控领域内提高小尺寸运动物体的识别率。Zhang 等人^[42]则采用了自上而下的路径，将高层特征图的丰富语义信息传递并融合到细节更丰富的低层特征中。这种信息流动方式有效地增强了用于检测小目标的低层特征的语义表示能力。

尽管上下文信息能有效补偿小目标自身特征的不足，但其引入并非总是有益的。冗余或不相关的上下文信息会引入显著的噪声，干扰模型对目标本身的判断，甚至导致性能下降。

5) 基于注意力机制的小目标检测方法

人类视觉系统能够快速聚焦于场景中的关键部分并忽略无关信息，这种高效的认知机制被称为视觉注意力机制。受此启发，注意力机制在计算机视觉领域得到了广泛研究与应用。其核心在于通过为特征图的不同部分分配差异化权重，从而突出有价值的信息区域，同时抑制无关或冗余的部分。在小目标检测任务中，目标像素稀少，其特征极易被复杂的背景和噪声模式所淹没。因此，可以利用注意力机制这一优势，引导模型聚焦于那些可能包含小目标的微小区域，从而在特征层面最大限度地减少背景污染，增强小目标的表征响应。此方法设计灵活，即插即用，能够被嵌入到几乎所有主流的目标检测架构中，因此被广泛使用。

研究者们提出了多种创新的注意力模型来应对小目标检测的挑战，受人类认知过程启发的 KB-RANN 模型^[43]，将长短期注意力神经网络引入检测框架，通过模拟人类视觉系统的持续性关注机制，使模型能够聚焦于图像中的关键区域。在此基础上，SCRDet^[44]提出了一个创新性的旋转目标检测架构，其中集成了经过监督训练的像素注意力和通道注意力模块，实现了从小目标区域中有效分离噪声的突破。随着 anchor-free 检测范式的兴起，FBR-Net^[45]通过引入层级注意力机制，在 FCOS^[46]检测器的基础上实现了特征金字塔各层级间的自适应特征均衡，显著提升了复杂场景下的小目标检测能力。与此同时，研究者们从不同角度完善了注意力机制的应用：Lu 等人^[47]设计的双路径模块通过并行处理实现了关键特征的增强与非目标信息的抑制，MSCCA^[48]通过用增强通道注意力块替代传统卷积组件，构建了参数效率更高的轻量级检测器，而 Li 等人^[49]提出的跨层注意力模块则通过加强不同层级特征间的交互，获得了对小目标更强烈的特征响应。这些方法虽采用不同的技术路径，但其核心都是通过注意力权重的重新分配来增强小目标特征表示的有效性。

尽管注意力机制为小目标检测带来了显著的性能提升，但其应用仍面临两大挑战。首先，性能的提升往往以高昂的计算开销为代价。注意力机制中的相关运算（如大规模矩阵乘法）会大幅增加模型的计算复杂度和推理时间。其次，当前大多数注意力范式缺乏明确的、直接的监督信号，其参数优化过程依赖于最终检测任务的损失回传，有可能限制了注意力模块学习到最精准的聚焦区域。

6) 基于聚焦检测的小目标检测方法

在高分辨率图像中，小目标的分布呈现显著的非均匀性，传统的分块检测策略会导致大量计算资源浪费在无目标的空背景区域。为了解决这一问题，研究者提出聚焦检测的方法：通过先定位潜在目标区域再进行精细检测的两阶段流程，有效提升检测

精度。这类方法的核心思想是打破传统处理高分辨率图像的固定流程，首先提取出可能包含目标的候选区域，随后在这些区域上执行后续检测任务。这种机制确保了小目标能够在更高分辨率下被处理，从而缓解了因下采样导致的信息丢失问题，显著提升了小目标的特征表示质量。

早期代表性工作 ClusDet^[50]通过充分挖掘目标间的语义和空间信息生成聚类区域，进而实施检测。随后的研究沿袭了这一思路并从不同角度进行深化：Duan 等人^[51]与 Li 等人^[52]分别利用像素级监督进行密度估计，生成能精确表征目标分布的密度图，CRENet^[53]设计了自适应聚类算法来搜索目标聚集区域。另一方面，针对固定尺寸输入导致的漏检问题，有研究采用实时切片方法检测高分辨率航拍图像中的行人与车辆。与之理念相似，Deng 等人^[54]和 Xu 等人^[55]分别通过设计超分辨率网络和强化学习框架，实现了对局部区域的分辨率提升和自适应缩放聚焦。Leng 等人^[56]在传统区域挖掘基础上引入了区域特异性上下文学习模块，有效增强了对挑战性区域内小尺寸目标的感知能力。F&D 框架^[57]通过聚焦网络检测候选区域并将其裁剪缩放到更高分辨率，为实现小目标的精准检测提供了系统化解决方案。

与传统的滑动窗口机制相比，聚焦检测方法通过自适应裁剪和灵活缩放操作实现了计算资源的优化配置，较小目标可在更高分辨率下处理以保留细节，而较大目标则在相对较低分辨率下检测，这显著减少了推理过程中的内存占用并降低了背景干扰。然而，该类方法必须解决“聚焦何处”这一关键问题。现有方案主要依赖于人工附加标注或辅助架构（如分割网络和高斯混合模型），前者需要繁重的标注工作，后者则使端到端优化过程复杂化。

1.2.3 机载光电系统目标跟踪研究现状

单目标跟踪是机载光电系统实现持续监视、精确定位与智能决策的核心，在军事侦察、安防巡逻等任务中具有不可替代的关键作用。该功能旨在视频序列中持续定位特定目标，其工作模式是在首帧获取目标初始位置信息后，在后续帧中实现自主定位与持续跟踪。单目标跟踪算法的发展历程经历了从基于相关滤波的传统判别式方法，到引入深度卷积网络的判别式跟踪器，再到基于孪生网络的模板匹配方法。近年来，Transformer 架构的引入推动了跟踪技术的发展，自注意力机制能够有效建模复杂时空依赖关系，显著提升算法对复杂背景下目标形变及快速运动等挑战的适应能力。下面将分别介绍判别式模型的方法，基于孪生网络的方法和基于 Transformer 的方法。

1) 基于判别式模型的单目标跟踪方法

判别式跟踪器将单目标跟踪问题定义为一个二分类任务，其核心是训练一个外观模型，通过最小化判别性损失函数，来区分包含目标的正面样本与来自背景区域的负面样本。此类方法的一个关键特性在于其在线学习和模板更新机制，这使得跟踪器能够在跟踪过程中实时适应目标的外观变化、及环境改变。早期的判别式跟踪器主要依

赖于手工特征（如 HOG）和简单分类器（如支持向量机或岭回归），后续研究则转向使用深度特征和基于优化的预测模型。其中，基于相关滤波的跟踪器在判别式跟踪器发展中发挥了关键作用。

相关滤波类方法的核心是在线学习一个线性滤波器，通过求解一个岭回归问题，使得该滤波器能够从目标周围的背景区域中有效区分出目标图像块。其最主要的创新在于利用快速傅里叶变换将计算转换到频域，并利用循环互相关的性质，实现了极其快速的滤波器训练和更新。在跟踪过程中，相关滤波器被应用于一个以目标上一帧位置为中心的搜索窗口上，滤波器输出的最大响应值位置即被判定为跟踪的当前位置。在每一帧处理完毕后，跟踪器会在线更新滤波器权重，使模型能够动态适应目标外观变化，部分方法还可通过选择最高相关输出对应的尺度实现目标尺度估计。**MOSSE**^[58]是早期相关滤波跟踪器的代表之一。它提出了一种简单且实时的跟踪方法，对尺度、光照、姿态和非刚性形变具有鲁棒性。相较于之前需要大量训练样本的相关滤波方法，**MOSSE**仅使用单帧图像训练滤波器，显著降低了数据需求。**KCF**^[59]在**MOSSE**的基础上引入了核化技巧与多通道特征（如 HOG）支持，进一步提升了判别能力和特征表示。**KCF**充分利用了图像块平移后的循环结构，通过应用离散傅里叶变换，降低了存储和计算复杂度，即使使用更丰富的特征表示也能实现实时处理。**SRDCF**^[60]针对常规互相关滤波跟踪器因循环卷积假设而产生的边界效应问题，引入了空间正则化项，根据滤波器系数的空间位置对其进行惩罚。这使得模型能够从包含更丰富负样本的更大图像区域中学习，同时将注意力集中于目标本身。该方法通过在频域中利用正则项的稀疏性，并采用高斯-塞德尔求解器进行在线优化，保持了计算效率。

随着深度学习的兴起，相关滤波开始与神经网络架构深度融合。**CFNet**^[61]将在线相关滤波器作为一个可微分层集成到一个浅层孪生网络中，实现了跟踪模型与特征表示的端到端学习，其关键创新在于将相关滤波器视为一个封闭的优化模块，并通过反向传播将其嵌入网络。后续的高级判别式跟踪框架不再以相关滤波为核心，**DiMP**^[62]通过改进目标模型的学习过程，增强了跟踪器从背景干扰中区分目标的能力。它将目标模型学习视为一个判别性损失函数优化问题，并利用元学习优化器在线更新模型。此外，**DiMP**集成了一个并行的 IoU 预测分支用于精确的边界框估计。**PrDiMP**^[63]在**DiMP**的基础上，通过将目标中心定位和边界框回归重新表述为概率回归任务，进一步提升了跟踪器的鲁棒性。它直接通过网络架构对目标状态的条件概率密度进行建模，而不假设预定义的分布，使跟踪器能够表示标注本身及目标状态的不确定性。

2) 基于孪生网络的单目标跟踪方法

孪生网络跟踪器是单目标跟踪领域的一个重要范式，其核心思想是将跟踪任务定义为目标模板与搜索区域之间的相似性匹配问题。典型的孪生网络由两个权值共享的分支构成：模板分支处理第一帧中的目标图像块，搜索分支处理当前帧中的候选区域。这两个分支通过一个共享的主干网络将输入映射到一个公共的特征空间，随后通过计算两者之间的相似度来定位目标。此类方法通常在大规模数据集上进行离线训练，以

学习一个通用的匹配函数，从而在在线推理时无需复杂的模型更新即可实现快速高效的跟踪。孪生网络跟踪器通过引入新的回归头、更新机制、更深的主干网络以及注意力模块等创新，在鲁棒性和准确性上不断进步。

SiamFC^[64]开创性地将用于相似性学习的全卷积孪生网络引入跟踪领域。该网络通过两个相同的分支提取模板和搜索区域的特征，再经过一个互相关层，生成一个密集的响应图来指示目标的位置。这种架构无需在线更新模型，并通过使用图像金字塔来应对尺度变化。SA-Siam^[65]提出了一个双重的孪生网络结构，通过融合互补的表观特征和语义特征来提升模型的泛化能力。该网络包含两个独立训练的分支：表观分支保持SiamFC的结构进行相似性学习，语义分支从一个预训练的分类网络中提取高层语义特征。两个分支仅在推理时进行融合，并引入通道注意力机制来增强语义分支中的目标特异性表示，从而提升了对外观变化的鲁棒性。SiamRPN^[66]将区域提议网络引入孪生框架，是推动该领域发展的关键一步。RPN的加入使得网络能够同时进行前景背景分类和边界框回归，从而实现了更精确的尺度与长宽比估计，摒弃了SiamFC中使用的多尺度搜索策略。该模型将跟踪建模为一个局部一次性检测任务，其中模板分支充当元学习器，为搜索分支生成检测核。DaSiamRPN^[67]针对训练数据中语义背景与非语义背景不均衡的问题，提出了干扰物感知的训练策略。在离线训练中，它通过引入来自同类和不同类别的语义负样本对，使网络学习更具判别力的表征。在推理时，它采用由局部到全局的搜索策略，并结合困难负样本挖掘，有效增强了跟踪器的短期精度与长期鲁棒性。SiamRPN++^[68]解决了早期孪生网络无法使用ResNet^[69]等深层主干网络的限制，实现了更深网络的端到端训练。此外，该方法通过多层特征聚合以及轻量的深度互相关模块，在提升精度的同时减少了参数量，稳定了训练过程。

SiamFC++^[70]通过解耦分类与回归分支，将粗略的目标定位与精确的边界框预测分离开。该模型采用无锚框的逐像素预测策略，避免了对目标先验尺度的依赖，并引入一个质量评估分支来估计边界框预测的可靠性，以解决高分类置信度与差定位结果不匹配的问题。SiamAttn^[71]针对固定模板在应对目标外观变化时的局限性，在孪生架构中引入了可变形孪生注意力模块。该模块结合了可变形自注意力和交叉注意力，分别用于建模帧内上下文以及模板与搜索区域之间的相互依赖关系，从而实现了对目标模板的自适应更新，提升了在背景杂乱情况下的鲁棒性。SiamDMU^[72]明确地提出了一个双掩码模板更新方法。该方法在SiamRPN++的框架上，增加了一个由掩码增强块和模板更新块构成的模板更新模块。该模块利用语义分割和长期运动信息，在图像级别而非特征级别更新模板，保留了高分辨率的空间细节，从而在严重外观变化下实现鲁棒跟踪。Siam R-CNN^[73]为长期跟踪设计了一个两阶段的孪生重检测框架。它摒弃了围绕先前预测进行局部搜索的策略，转而执行全图范围的全局重检测。其核心创新是轨迹动态规划算法，该算法联合考虑来自第一帧模板和前一帧的重检测结果，形成时空轨迹，从而实现鲁棒的目标关联与干扰物抑制。

3) 基于 Transformer 的单目标跟踪方法

Transformer 架构因其在自然语言处理任务中的成功而被引入计算机视觉领域，并在语义分割、目标检测、图像分类等任务中展现出卓越性能。与依赖局部感受野的卷积神经网络不同，Transformer 采用全局注意力能更好地捕获长距离上下文信息。基于 Transformer 的跟踪方法利用编码器-解码器架构、自注意力等关键组件来增强特征表示和目标定位能力。根据架构设计，现有 Transformer 跟踪器可分为两大类：混合 Transformer 跟踪器和完全 Transformer 跟踪器。

混合 Transformer 跟踪器将 Transformer 模块嵌入到已有的孪生或判别式跟踪框架中，以增强其性能，解决 CNN 架构中感受野有限、全局上下文建模不足等问题。TransT^[74]是早期将注意力架构引入跟踪的代表性工作。它完全取代了孪生框架中传统的基于互相关的特征融合方式，设计了一个由基于多头自注意力的 (Ego-Eontext Augment, ECA) 模块和基于多头交叉注意力的 (Cross-Feature Augment, CFA) 模块构成的融合网络。这种设计能更好地捕获全局上下文，在整合模板与搜索区域特征时保留语义信息，从而在外貌变化和相似物体干扰下实现鲁棒跟踪。TrDiMP 与 TrSiam^[75]通过为判别式和孪生跟踪器引入 Transformer 架构来建模视频帧间的时序依赖。它们设计了一个并行的编码器解码器框架，编码器用于增强多帧间的模板特征，解码器则将历史模板的空间掩码和特征传播到当前搜索区域，共享注意力权重和轻量级设计保证了计算效率。

完全 Transformer 跟踪器不依赖原有的孪生或判别式跟踪框架，从头构建独立架构。为了克服卷积网络难以捕捉长距离依赖关系的问题，STARK^[76]引入了具有编码器解码器架构的模型。其编码器通过联合处理初始模板、动态更新模板和当前搜索区域的特征，捕获全局上下文关系，利用多头自注意力强化时空编码。一个轻量级解码器学习单个查询向量来预测目标位置，并提出了一个基于角点的全卷积预测头，简化了跟踪流程。CSWinTT^[77]引入了多尺度循环移位窗口注意力机制，将注意力计算从像素级提升到窗口级，以保持物体结构并实现更局部化的匹配。其循环移位策略结合空间正则化注意力掩码，在生成多样窗口样本的同时抑制了边界效应。MixFormer^[78]提出了一种紧凑的端到端架构，其核心是混合注意力模块，该模块能够同时执行自注意力和交叉注意力操作，统一了特征提取和信息集成阶段。它采用 CvT^[79]作为主干网络，并引入非对称注意力方案以提升效率。SwinTrack^[80]是一个基于 Swin Transformer^[81]的单目标跟踪框架。它将模板和搜索区域特征拼接后通过共享主干进行联合建模，并引入一个运动令牌来捕获目标的历史轨迹，从而在解码器中增强时序感知能力。两阶段跟踪器分别从模板区域和搜索区域独立提取特征，并在后期进行特征融合以建立关联模型，其对目标的感知能力较弱，目标与背景的区分度有限。为解决这一问题，OSTrack^[82]提出了一种单流、单阶段的 Transformer 框架，该框架在初始阶段实现模板与搜索区域之间的双向信息流动，将特征提取与关系建模统一起来。该方法通过直接拼接两个输入，利用自注意力机制学习具有目标感知能力的特征，无需再设计独立的交叉注意力模块。

SeqTrack^[83]引入了新颖的 Seq2Sqe 学习框架，将目标跟踪视为自回归序列生成任务。该方法将边界框坐标离散化为令牌序列，并使用普通编码器解码器 Transformer 学习生成它们，编码器联合提取模板和搜索图像的特征，而因果解码器自回归地预测边界框。

传统的单目标跟踪方法依赖于针对特定模态的设计，这些方法采用定制化架构，参数冗余且性能有限。OneTracker^[84]通过引入一种统一高效的双模态（RGB 与 RGB+X）跟踪框架来上述问题。该框架采用模块化的两阶段设计：其核心是基础跟踪器，一个在大规模 RGB 跟踪数据集上进行预训练的 Transformer 的模型。为将模型扩展至其他模态，OneTracker 集成了一个提示跟踪器模块（Prompt Tracker module），将额外输入（如深度、热成像、分割掩码或语言描述）视作任务提示。这一功能通过引入跨模态跟踪提示器（Cross-Modality Tracking Prompters）与跟踪任务感知层实现，仅更新轻量级适配器参数而保持主干基础模型冻结，实现了高效的参数微调。该设计支持基于提示的多模态融合，能够在无需修改核心模型结构的前提下具备任务自适应性，使 OneTracker 成为适用于多输入模态、多样化跟踪场景的高效可扩展解决方案。

1.3 论文的主要贡献

1. 提出面向可见光航拍图像的高性能小目标检测网络

无人机航拍图像中存在极端尺度变化、密集小目标分布及复杂背景干扰等挑战，通用检测器在此类场景下存在显著性能差距，提高输入图像分辨率能提升检测精度但是会大幅增加计算量。现有小目标检测方法在保留对小目标检测至关重要的细粒度特征方面存在结构缺陷，难以在精度与速度之间取得理想平衡。针对上述问题，本文提出针对航拍可见光图像的小目标检测网络 BAP-DETR，通过设计双重注意力模块，采用通道分离策略实现卷积与注意力机制的并行优化，突破了简单的“拆分-处理-合并”范式，实现了更具交互性的特征优化与注意力建模机制，增强了模型从复杂航拍图像中提取差异化特征的能力。创新设计的双融合编码器配备频域感知融合模块，在有效融合高级语义信息的同时保留关键的低层特征，有效增强了对小目标特征的保留能力与多尺度特征的整合效果。此外，结合倒数归一化 Wasserstein 距离与 CIoU 优化损失函数，在不增加推理阶段计算成本的前提下，提升模型对于小目标的检测精度。我们在 VisDrone、UAVDT 和 AI-TOD 三个主流航拍图像数据集上对模型进行了全面的评估，BAP-DETR 在保持较高推理速度的同时，平均检测精度较基线方法提升 6.9%，计算负载降低 17.5%，为无人机可见光图像的小目标检测提供了有效的解决方案。

2. 设计面向无人机红外图像的轻量化小目标检测网络

红外图像具有空间分辨率较低，且缺乏丰富的纹理细节和颜色信息以及噪声显著等特点，在无人机航拍场景下，由于观测距离远，大多数目标在图像中仅呈现为像素占比极小的点状或微小区域，显著的尺度变化进一步加大了检测难度，现有网络在复杂背景下难以对这些小目标进行精准定位和分类。此外，现有神经网络及计算密集型的 Transformer 架构在算力受限的边缘设备上难以实现实时处理，其高计算复杂度成为实

际部署的主要瓶颈。针对上述挑战，本文提出了一种面向无人机红外图像的小目标检测网络 MFF-DCNet。该网络的核心创新包括：基于深度分离卷积的跨阶段 Transformer (Depth-wise Cross-stage Transformer, DCFormer) 和多尺度特征聚焦 (Multi-Feature Focus, MFF) 颈部结构。DCFormer 模块通过深度可分离卷积与跨阶段特征融合的结合，在显著降低计算开销的同时，有效增强了主干网络对多尺度上下文信息的建模能力。优化特征提取过程，提升了模型在复杂场景下对小目标的特征判别能力，并且为在边缘设备上的实时部署提供了可能。重新设计的 MFF 颈部结构通过构建新颖的特征聚合机制，增强了跨尺度特征的整合能力，使模型能够更好地融合不同层级的语义信息和空间细节。这种设计显著提升了模型对多尺度目标的检测性能，特别是在复杂背景下对微小红外目标的精准识别能力。在 HIT-UAV 数据集上，MFF-DCNet 与专用无人机目标检测网络 SuperYOLO 相比提升了 5.8% AP，FPS 提升了 10%，与基线网络 YOLO 系列相比，AP 提升最多 12.8%，与最新的 DETR 及其衍生方法相比，AP 提升 8.2%，FPS 是 RT-DETR 的 3 倍。在 DrongVehicle 数据集上，与基线网络相比，AP 提升 5.3%，与 RT-DETR 相比，AP 提升 11.9%。此外，我们的方法在 NVIDIA Jetson Orin NX 边缘设备上实现了 39.6 FPS 的实时性能，展现出其在资源受限物联网环境中的实际部署能力。

3. 开发复杂场景下的抗遮挡长时目标跟踪算法

对于遮挡目标的长时跟踪是无人机智能光电系统中的一项关键技术挑战。随着无人机在复杂环境中的应用不断增加，目标遮挡情况变得日益普遍，给目标跟踪带来了显著的困难。特别是在城市环境中，由输电线（塔）、交通灯、高架桥等复杂地物所造成的目标遮挡问题尤为突出。这些障碍物不仅会遮挡视线，影响目标的可见性，还可能导致目标在跟踪过程中出现频繁的丢失和再识别的困难。现有目标跟踪数据集并不能全面覆盖城市环境中复杂遮挡场景的多样性和复杂性，同时现有的目标跟踪算法普遍缺乏可靠的目标判丢和重捕获机制，在遇到目标被完全遮挡的情况时，极易出现跟踪失败，在实际工程应用中难以满足长时稳定跟踪的需求。针对上述问题，本文提出了一种具备目标丢失判断与重捕获能力的抗遮挡长时跟踪框架。该框架采用模块化设计，可复用于任何基础跟踪器，通过引入平均峰值相关能量和结构相似性指数构建了可靠的目标跟踪置信度评估体系。框架包含专门设计的遮挡判定模块和目标重捕获机制，有效解决了目标消失后的重新识别问题，同时，我们还提出了一种动态模板更新机制，根据跟踪置信度调整模板更新策略，增强了模型对目标外观变化的适应能力。同时我们还构建了一套城市环境下无人机视角目标遮挡测试数据集 MMUOT-1050 (Multi-Modal UAV Occlusion Tracking Benchmark)，一共包含 353 段可见光视频序列和 697 段红外视频序列，每一段视频序列都包含目标被完全遮挡的场景，能有效检测跟踪算法对遮挡目标的跟踪性能，在 MMUOT-1050 上，可见光视频测试中，AS-Tracker 跟踪成功率达到 86.38%，与基线算法相比提升 11.01%，在红外视频测试中，AS-Tracker 跟踪成功率达到 91.93%，与基线算法相比提升 11.73%，实验结果表明，该框架在可见光与红外双模态数据下均能显著提升跟踪成功率，为解决复杂遮挡场景下的长时跟踪难题提供了有效方案。

4. 构建完整的智能光电系统软硬件集成方案

在算法创新的基础上，本文进一步完成了从理论方法到工程实践的跨越，设计并实现了一套完整的机载智能光电系统解决方案。该系统采用模块化设计理念，集成了多光谱传感器数据采集、嵌入式边缘计算和上位机软件三大核心模块。在硬件实现层面，通过优化传感器选型，实现了可见光与红外多光谱数据的同步采集与预处理，选用瑞芯微 RK3588 作为核心计算模组，深入利用瑞芯微官方提供的 RKNN 工具链，对训练好的深度学习模型进行量化、转换与加速处理，显著提升了神经网络在嵌入式平台上的推理效率，有效满足了端侧实时计算的严苛要求。

在软件层面，本文设计并实现了一套完整的端侧运行程序框架，不同于部署孤立的算法模块。该框架采用基于生产者-消费者模式的流水线架构，通过观察者模式实现模块间的松耦合通信，运用工厂模式统一管理各类处理单元的实例化。核心架构包含数据接入层、处理引擎层和服务总线的三层设计：数据接入层采用多路复用 I/O 模型实现多传感器数据的并行采集，处理引擎层通过策略模式封装各类智能算法，支持运行时动态切换，服务总线基于事件驱动架构，采用消息队列中间件实现模块间的异步通信。通过引入无锁环形缓冲区优化数据流传输，结合内存池技术预分配关键数据结构，并采用优先级抢占式任务调度策略，提升系统实时性能。框架通过设计统一的通信接口层，与上位机系统建立了基于多种可选协议（TCP/IP, UDP, HTTP, 串口）的双向数据通道，采用协议缓冲器进行高效数据序列化，确保检测结果与系统状态的实时上传，同时支持上位机控制指令的可靠接收与解析。该架构通过前后端分离的设计理念，实现了数据处理与用户交互的逻辑解耦，为使用者提供了流畅的实时监控体验。

这一系列高度模块化的设计不仅确保了系统的可靠性和可维护性，更为算法更新和功能扩展提供了极大的灵活性，建立了完整的嵌入式 AI 软件栈解决方案。

1.4 论文的组织结构

本文包含六个章节，每个章节的主要内容如下所述：

第 1 章为绪论。首先阐述了低空经济发展对无人机智能化的迫切需求，进而剖析了作为无人机核心感知部件的智能光电系统所面临的关键技术挑战。在此基础上，系统综述了机载光电系统及其核心算法（目标检测与跟踪）的国内外研究现状与发展趋势。最后，明确了本文的研究目标与主要贡献。

第 2 章介绍了智能光电系统及相关算法的基础理论。本章系统阐述了机载多光谱光电系统的基本原理、核心组成与发展历程，并深入剖析了深度学习框架下小目标检测与鲁棒目标跟踪的基础模型与关键技术，为后续章节的算法创新提供理论基础。

第 3 章提出了一种基于双通道处理网络的可见光航拍图像小目标检测算法（BAP-DETR）。本章首先分析了无人机可见光图像中因极端尺度变化、目标密集及复杂背景导致检测性能下降的核心问题。针对现有 Transformer 检测架构在特征细化与多尺度融合方面的不足，提出了基于通道分离的双重注意力处理模块与频率感知融合编码器。通

过公开数据集上的对比实验与消融研究，验证了该算法在精度与效率上的综合优势。

第 4 章提出了一种面向无人机红外图像的高效小目标检测网络（MFF-DCNet）。本章聚焦于红外图像低分辨率、低对比度及高噪声特性带来的独特挑战。通过设计深度交叉阶段 Transformer 骨干网络与多特征聚焦颈部结构，强化了对弱小目标的特征表征能力。实验部分不仅在公开数据集上证明了其领先的检测精度，更在嵌入式平台上验证了其满足实时性要求的部署可行性。

第 5 章致力于智能光电系统的工程实现与复杂场景下的跟踪算法集成。本章首先从系统工程角度，详细阐述了基于软硬件协同设计的智能光电系统整体架构，包括多光谱数据同步采集、基于 RK3588 平台与 RKNN 工具链的算法部署优化，以及采用先进设计模式的端侧全功能软件框架。继而，针对动态场景中的遮挡难题，提出了一种融合重识别机制与自适应模板更新的抗遮挡长时目标跟踪算法，并通过实际场景测试验证了其鲁棒性。

第 6 章为总结与展望。系统梳理并总结了本文在算法与系统层面的主要研究工作与贡献，客观分析了当前研究的局限性，并对多模态感知前沿技术、算法与硬件的深度协同优化以及群体智能等未来发展方向进行了展望。

2 智能光电处理算法与系统设计研究基础

本章主要介绍智能光电系统的基本组成以及其中涉及的核心功能，目标检测和目标跟踪的相关算法研究基础。

2.1 智能光电系统基本组成

智能光电系统，作为无人机实现环境感知与智能决策的核心，是一个深度融合了光学、机械、电子、计算机与人工智能的复杂综合性系统。其基本工作原理是建立在光电转换与信息处理的基础之上，系统通过各类光电传感器，捕获目标反射或辐射的电磁波信号，经由一系列软硬件模块的处理，完成光信号到电信号的初级转换。现代智能光电系统的关键差异在于其信息处理架构的改变：它将传统上由地面站或后端服务器承担的高级处理功能（如图像增强、特征提取、目标识别与跟踪）前移至机载端侧。这一转变的核心在于系统内部集成了高性能的嵌入式计算平台（如 Nvidia Jetson 系列，华为 Atlas 系列，瑞芯微 RK 系列），能够在数据产生的源头对原始图像进行实时、智能化的处理与分析，直接形成可供决策的结构化信息。这种一体化设计，不仅极大降低了对数据链带宽的依赖，减少了信息传输的延迟，更使无人机平台获得了在复杂、动态环境中进行即时态势理解和自主响应的能力，从而真正实现了从“被动成像设备”到“主动感知节点”的智能化跃迁。

智能光电系统的硬件架构是实现其物理功能的基础，通常遵循模块化、集成化设计原则，以适应无人机平台对载荷尺寸、重量和功耗的严苛限制。其核心硬件模块主要包括光学传感单元、惯性稳定平台以及信息处理单元。

光学传感单元负责从环境中捕获原始光学信息。为突破单一波段感知的局限，满足全天候、全时段及复杂场景下的探测需求，现代光电系统的演进方向已从早期的单一传感器，发展为多光谱、多传感器深度融合的架构。可见光高清成像与长波热红外成像系统已逐渐成为现代机载光电系统的标配，可见光成像系统通常基于高分辨率的 CCD 或 CMOS 图像传感器构建，配备电动连续变焦镜头，以实现从广域搜索到细节识别的无缝切换。其作用是在昼间良好光照条件下提供富含纹理和色彩信息的高清图像。红外热成像系统是系统实现夜间和恶劣气象条件下工作的关键，它们通过探测目标与背景的热辐射差异来生成热图像。在此基础上，为进一步提升系统在复杂环境下的感知能力，现代光电系统还逐步集成中波红外、短波红外、微光、多光谱成像仪激光测距仪等特种传感器。中波红外传感器对高温目标（如发动机喷口、导弹尾焰）极为敏感，是实现早期预警与精准识别的关键，短波红外传感器具有穿透薄雾、烟尘和水汽的能力，提升系统在恶劣天气下的感知能力，微光传感器能在极低照度下将微弱光子信号大幅增益，生成可供人眼判读的夜景图像。多光谱成像仪能在多个狭窄、连续的波段内同时对目标成像，从而提升对伪装目标、特定物质的识别与分类能力，激光测距仪

通过测量激光脉冲的往返时间精确计算目标距离,为定位提供关键参数。机载光电系统主要用于对地面对目标进行识别跟踪、激光测距和照射等,主要指标包括以下部分或全部:红外作用距离、红外探测视场、红外探测波长、可见光作用距离、可见光探测视场、可见光探测波长、稳定精度、跟踪精度、随动精度、瞄准线指示精度、搜索范围、搜索角速度、搜索角加速度、跟踪速度、跟踪加速度、视场切换时间、激光工作波长、激光测距范围、激光照射距离、激光照射频率、激光照射精度、光轴平行度、工作准备时间、供电电源、系统重量和外形尺寸等。其中与光学传感单元相关的最重要的系统指标是可见光与红外作用距离,下面将具体分析这两个指标

1) 红外作用距离

人眼通过显示器看到红外成像系统获得的目标图像的基本条件包括:1. 目标具有一定空间频率。2. 目标与背景的温差经过大气衰减,到达红外系统探测器上时仍大于或等于系统对该频率的最小可分辨温差(MRTD),即

$$\begin{cases} \Delta T = \Delta T_e \cdot \tau_a \geq MRTD(f) \\ \frac{H}{2n_e \cdot R} \geq \theta \end{cases} \quad (2-1)$$

其中, ΔT 为经过大气衰减后, 目标与背景的温差, ΔT_e 为目标与背景的实际温差, τ_a 为 R 距离上的大气平均透过率, $MRTD(f)$ 为系统对空间频率 f 的最小可分辨温差, H 为目标尺寸, n_e 为不同观察等级要求时的目标等效线对数, R 为目标与系统的距离, θ 为瞬时视场。

对于目标的发现距离和识别距离,目前光电系统都是根据 JOHNSON 法则,即发现目标和识别目标 50% 概率所需线对数 n_e ,

表 2-1 JOHNSON 法则规定的发现和识别判据

鉴别等级	50% 概率所需线对数
发现	1.0 ± 0.25
识别	4.0 ± 0.8

红外系统的瞬时视场由探测器面元数量和总视场决定,表达式为

$$\theta = \frac{\alpha}{n} \quad (2-2)$$

其中, α 为系统偏航或俯仰方向的总视场, n 为探测器偏航或俯仰方向的面元数量。

由上可知,红外系统在设计中既要考虑能量方面的要求,又要考虑空间分辨率的要求,MRTD 是一个综合考核的指标,可以综合描述系统的空间分辨率和温度灵敏度特

性，与目标的空间频率、系统传递函数和等效噪声温差等相关。一般情况下，对于近距离探测目标，红外系统空间分辨率是主要矛盾，对于远距离探测目标，温度灵敏度即能量因素是主要矛盾。

2) 可见光作用距离

可见光探测利用的是目标对阳光的反射光，与红外探测相似，获得可观察的目标图像的基本条件是：1. 具有一定空间频率的目标，2. 经过大气衰减后到达 CCD 的照度满足最低照度要求，目标与背景的对比度经大气衰减后满足成像要求。即空间分辨能力和能量分辨能力，具体理论计算公式为

$$\left\{ \begin{array}{l} E = \frac{1}{4}E_0\rho_t\tau_a\tau_0\left(\frac{D}{f}\right)^2 \geq E_M \\ C = C_0 \cdot \tau_a \cdot \tau_0 \geq C_M \\ \frac{H}{2n_e \cdot R} \geq \theta \end{array} \right. \quad (2-3)$$

其中， E 为到达 CCD 像面的照度， E_0 为环境照度， ρ_t 为目标反射率， τ_a 为大气透过率， τ_0 为系统光学透过率， $\frac{D}{f}$ 为可见光系统相对孔径， E_M 为 CCD 工作允许的最低照度， C 为经过大气传输后目标与背景的对比度， C_0 为目标与背景的对比度， C_M 为人眼能够分辨的最低对比度， H 为目标尺寸， n_e 为不同观察等级要求时的目标等效线对数， R 为目标与系统的距离， θ 为瞬时视场。

惯性稳定平台是机载光电系统硬件架构中实现高精度指向与稳定成像的核心模块，其核心功能在于隔离无人机飞行中的高频振动、姿态变化等运动干扰，为光学传感器提供一个相对稳定的基准，从而确保获取清晰、稳定的图像与视频数据。为实现这一目标，平台在机械上通常采用精密的框架式结构，例如常见的两轴（方位、俯仰）四环架设计：内万向架直接承载并稳定光电传感器，尽可能隔离载体振动，外万向架则将内环与外部环境扰动进一步隔离。该系统主要由框架结构、敏感测量元件、伺服驱动机构及控制单元组成。敏感元件（如陀螺仪）实时测量载体或框架自身的角运动，这些运动信息被送至控制单元（如基于单片机的数字控制器）。控制单元计算补偿指令，并驱动安装在框架轴上的力矩电机产生反向力矩，从而主动抵消扰动，使光电传感器的视轴在惯性空间中保持稳定锁定或实现平滑跟踪。随着技术发展，为在性能与体积重量间取得更好平衡，出现了“半捷联稳定”等先进形式，其利用载机惯性导航系统的信息辅助稳定，减少了平台上的陀螺数量，有助于实现系统的小型化和轻量化。惯性稳定平台的设计与性能，直接决定了整个机载光电系统在动态环境下成像的清晰度、目标跟踪的准确性。

智能机载光电系统的信息处理单元已超越简单的数据采集与转发，成为一个集采集、处理、智能分析于一体的综合模块。它不仅负责采集来自可见光、红外、激光等多种传感器的海量数据并进行去噪、增强、稳像等预处理，还需在尺寸、重量和功耗严格受限的机载环境下，完成目标检测、跟踪及多传感器融合等复杂的计算任务。现代信息

处理单元普遍采用异构计算架构，主流设计通常集成通用 CPU 用于复杂逻辑与控制、高性能 GPU 用于并行图像处理与 AI 推理、专用神经网络处理器 NPU 用于高效执行深度学习算子。在工程实现上，常使用 FPGA 作为高速数据采集、预处理和逻辑控制的协处理器，与 AI 主控芯片构成异构系统，以实现性能与效率的最佳平衡。为适配各类传感器，处理单元还提供丰富的工业级接口，如多路 MIPI CSI（相机串行接口）、GMSL（千兆多媒体串行链路）用于摄像头接入，以及 CAN 总线、ETH、RS-232/422/485 等用于与其他设备通信。在具体产品形态上，信息处理单元常以核心计算模组（System on Module, SOM）或整机系统的形式存在。下面列举几种面向高性能边缘计算场景的代表性模组：

表 2-2 边缘计算模组

模组名称	核心架构与算力	关键特性
Nvidia Jetson	CPU + CUDA 核心 + Tensor 核心，算力覆盖数 TOPS 至数百 TOPS	拥有最成熟的 CUDA 开发生态，提供从开发、部署、调试的全栈工具，便于复杂神经网络的部署。平台丰富（如 Nano, Xavier, NX, Orin, Thor）
华为 Ascend	昇腾 CPU + 达芬奇架构 NPU，Atlas 200I A2 模块提供 8-20 TOPS INT8 算力	高集成度与能效比，单模块集成 CPU、AI 计算、编解码等功能。全国产化，适用于有自主可控要求的项目
瑞芯微 RK	ARM 架构 CPU + 自研 NPU，提供 1-6 TOPS INT8 算力，内置高性能 ISP	强大的多媒体处理与接口能力，支持多路摄像头输入和 8K 编解码。高性价比与低功耗，内置 ISP 简化了成像系统设计。支持多种操作系统和开发框架，适用于轻量级 AI 应用
寒武纪 MLU	基于寒武纪 MLUv02 架构的专用 AI 推理模组，提供 16 TOPS INT8 峰值算力	专为边缘 AI 设计，支持主流深度学习框架，低功耗、紧凑型设计，是 AI 核心硬件国产化替代的选项之一

2.2 机载光电系统目标检测算法研究基础

目标识别是指在图像或视频序列中检测并定位特定类别的目标物体。对于机载光电系统而言，目标识别是实现智能感知、态势理解与自主决策的基石，其性能直接决定了无人机能否在复杂的低空环境中，从海量图像数据中精准地“看见”并“理解”特定目标。本节将系统性地介绍目标识别算法的研究基础，围绕目标检测基本方法、常用数据集与核心评价指标进行详细介绍。

2.2.1 基本方法

深度学习的崛起为目标检测领域带来了根本性变革。其强大的自动特征学习能力能够从海量数据中直接提取具有高度判别性的多层次特征，克服了手工设计特征的局限性，端到端的模式将特征提取、候选区域生成、分类与回归整合进一个统一的网络中进行训练，显著提升了系统性能与优化效率，此外，深度学习模型展现出卓越的可扩展性与大数据适应能力，性能随着数据规模和模型复杂度的增加而持续提升。这些优势共同驱动了目标检测技术完成了从依赖人工特征设计到大规模数据驱动学习的范式转变。早期深度学习目标检测网络普遍以卷积神经网络（CNN）为核心架构，利用其局部连接、权值共享的归纳偏置，高效地处理网格化图像数据，发展出以 Faster R-CNN 为代表的两阶段检测器和以 YOLO 系列为代表的单阶段检测器，在 ImageNet、MS COCO 等大规模数据集上取得了突破性进展。2017 年 Google 提出的 Transformer 架构，凭借其注意力机制的核心设计，在自然语言处理领域取得了革命性成功，2020 年，Dosovitskiy 等人首次将 Transformer 架构应用于图像分类任务，提出了 Vision Transformer 模型，开始挑战卷积的统治地位。相比于卷积操作的局部性和静态权重，Transformer 能够更灵活地捕捉图像中长距离的依赖关系，并更好地理解复杂场景中目标的上下文信息，发展出了更简洁的端到端框架，最具代表性的便是 Detection Transformer (DETR)，DETR 摒弃了传统检测器中手工设计的锚框（Anchor）和非极大值抑制（NMS）等复杂后处理步骤，使用一个 Transformer 编码器-解码器架构直接将图像特征映射为目标集合，实现了真正的端到端目标检测。

与此同时，目标检测的任务边界本身也在不断变化。传统的检测任务通常被定义为“闭集检测”，即模型只能识别和定位在训练集中预先定义好的有限类别。这严重限制了其在真实开放世界中的应用，因为系统总会遇到训练集之外的类别。2021 年 Joseph 等人首次提出了“开放世界目标检测”这一新范式，要求模型能够识别已知类别，同时主动发现并标注未知类别，并能在获得增量信息后逐步学习这些新类别，代表性的工作有 OW-DETR，YOLO-World 等。此外，为了更自然、灵活地理解和定位物体，目标检测正与自然语言等多模态信息深度融合，通过将图像区域与自然语言描述在语义空间进行对齐，使模型能够根据任意文本描述来检测目标，从而天然地具备了识别无限类别的潜力，催生了开放词汇检测和指代检测，极大地提升了人机交互的直观性和系统在开放场景中的通用性，代表性工作有 OVR-CNN、RegionCLIP 等。

1) 基于卷积网络的目标检测

基于卷积神经网络的目标检测方法，在深度学习时代早期主导了该领域的发展。其核心思想是利用卷积核自动学习从原始像素到高级语义特征的层次化表示，并在此基础上完成目标定位与分类。这类方法的性能优势、工程成熟度及在边缘设备上的优化便利性，使其至今仍在众多实际工程系统中广泛应用。

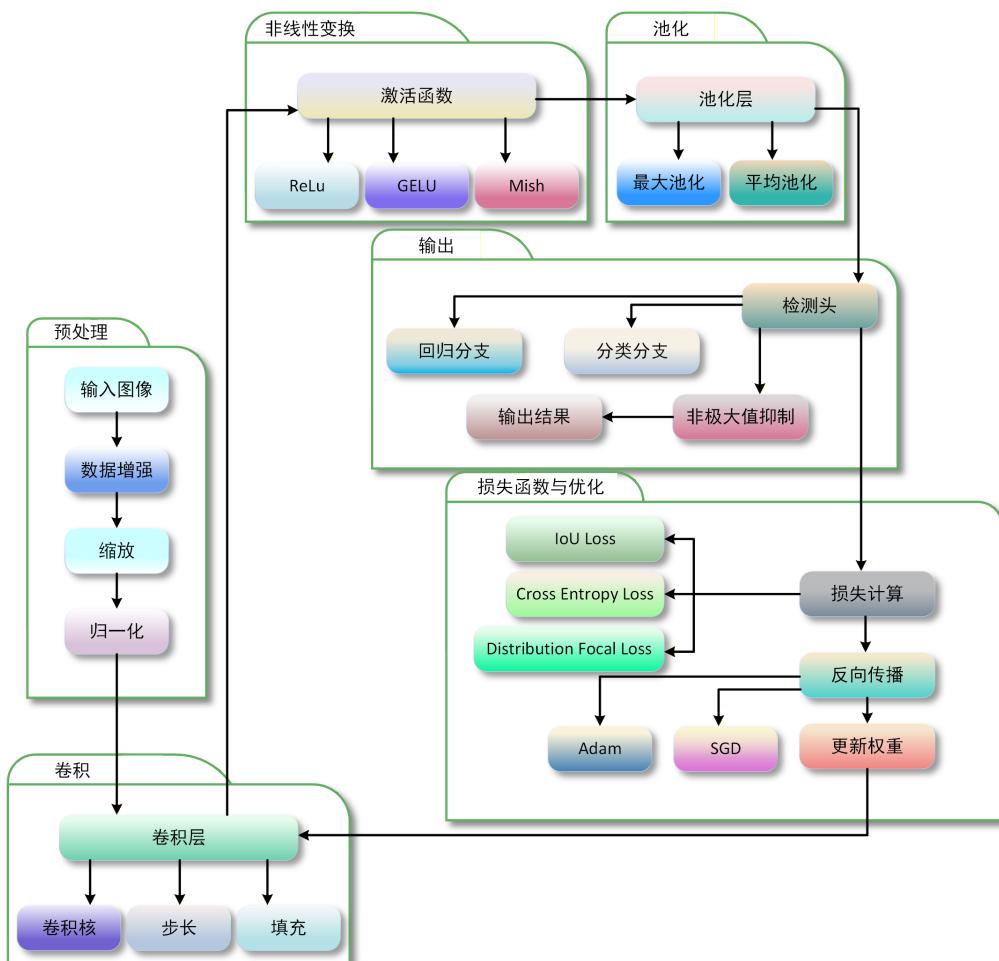


图 2-1 卷积神经网络架构示意图

如图2-1所示，卷积神经网络通过一种层次化的前向传播实现从原始图像输入到目标预测的端到端映射，由低层到高层、由具体到抽象渐进式提取图像特征。整个处理流程始于输入图像的预处理：原始图像首先被转换为网络可处理的多维张量，这一过程通常包含图像尺寸标准化、像素值归一化至特定范围，以及可选的标准化操作（如减去均值、除以标准差）。为提高模型泛化能力并防止过拟合，训练阶段常在线施加一系列数据增强操作，如随机裁剪、水平翻转、色彩抖动等，以此在不增加原始数据规模的前提下有效扩充训练样本的视觉多样性。预处理后的张量随后进入多层特征提取阶段。该阶段由交替堆叠的卷积层、池化层及非线性激活函数构成。卷积层通过具有局部连接与权重共享特性的滤波器在输入特征图上滑动计算，实现初级图像特征（如边缘、纹理）的检测，对于一个输入特征图 $X \in \mathbb{R}^{H \times W \times C_{in}}$ 和一个卷积核 $K \in \mathbb{R}^{k \times k \times C_{in} \times C_{out}}$ ，其输出特征图 Y 在位置 (i,j) 处的计算可表示为：

$$Y(i, j, c_{out}) = \sum_{c_{in}=1}^{C_{in}} \sum_{m=1}^k \sum_{n=1}^k X(i+m, j+n, c_{in}) K(m, n, c_{in}, c_{out}) \quad (2-4)$$

其输出的特征图经过非线性激活函数（如 ReLU 及其变体）的变换，引入表达复杂映射所需的非线性能力。池化层（如最大池化，平均池化）则对特征图进行空间下采样，在保留显著特征的同时逐步扩大后续层的感受野，并赋予特征一定程度的平移不变性。通过这种“卷积-激活-池化”模块的重复堆叠，网络能够自动构建起一个从局部细节到全局语义的、层次深化的多级特征表示。在早期的检测架构中，由主干网络提取的深层高级语义特征被直接送入检测头以完成最终的分类与定位。随着对多尺度目标，特别是小目标检测需求的日益迫切，研究者们在主干网络与检测头之间引入了 Neck（颈部）结构层（如特征金字塔网络 FPN 及其变体 PANet、BiFPN）。Neck 的核心功能是进行多尺度特征融合与增强，它通过自上而下、自下而上或双向融合路径，将主干网络中不同深度的特征图进行有效聚合，使输出特征同时具备丰富的空间细节和高层语义信息。经 Neck 优化处理后的特征图再送至检测头，显著提升了模型处理尺度变化的能力。

在两阶段检测器（如 Faster R-CNN）中，检测头由区域提议网络（Region Proposal Network, RPN）和全连接分类回归网络构成。RPN 在特征图上采用滑动窗口机制，为每个位置预设一组不同尺度和长宽比的锚框作为先验，执行前景/背景的二分类判断并进行初步的边界框回归，从而生成高质量的候选区域。这些候选区域通过感兴趣区域池化层被映射并裁剪为固定尺寸的特征块，最后由全连接网络完成精细的多类别分类和边界框坐标回归。而在单阶段检测器（如 YOLO、SSD）中，检测头被设计为直接在特征图的每个空间位置进行密集预测。该结构通常在多个尺度的特征图上预设密集的锚点，并利用卷积层一次性并行输出每个位置的类别概率分布和边界框偏移量，从而实现极高的推理速度。尽管省去了显式的区域提议步骤，但这类方法依赖精心设计的特征金字塔和后处理策略来保证在多尺度目标上的检测精度。

目标检测网络的训练以前向传播计算损失、反向传播更新模型参数的方式进行。为同时优化目标的识别与定位精度，损失函数通常由分类损失与回归损失两部分加权构成。其中，分类损失（如交叉熵损失）负责监督模型对目标类别的判别能力，回归损失则确保边界框坐标的预测准确性，早期多采用如 Smooth L1 Loss，而目前更广泛使用与评价指标直接相关的 IoU 损失（如 GIoU、CIoU）。为进一步提升边界框定位精度，研究者提出了分布焦点损失函数（Distribution Focal Loss），核心思想是将边界框位置的回归建模为一个离散概率分布的学习问题，而非直接回归一个确定的连续值，DFL 常与 CIoU Loss 结合使用，共同构成回归损失。

训练完成的模型在推理时，网络会对输入图像输出大量密集的初始预测框。为得到清晰唯一的检测结果，必须经过非极大值抑制（Non-Maximum Suppression, NMS）。NMS 的核心是基于置信度排序与 IoU 筛选：首先按分类置信度对所有预测框降序排列，选取最高置信度的框作为保留结果，随后剔除所有与其 IoU 超过设定阈值的其他框（视为对同一目标的冗余预测），此过程迭代进行，直至处理完所有框，最终输出一组互不重叠的高质量检测结果。

2) 基于 Transformer 的目标检测

Transformer 架构在自然语言处理领域取得了革命性成功，随后研究者们开始将其引入计算机视觉领域。通过其强大的特征建模能力和全局信息交互机制，克服了传统卷积神经网络在长距离依赖建模和局部感受野上的局限性。Vision Transformer 通过将图像重塑为图像块（Patch）序列，使其能够直接处理二维视觉数据，不依赖于预设的锚框或滑动窗口，而是通过注意力机制让模型自主“关注”与目标相关的所有图像区域，实现了从“局部感知”到“全局推理”的转变，此外，使用基于匈牙利算法的二分图匹配损失进行训练，使模型直接输出一组无冗余的预测结果，从而避免了非极大值抑制等复杂后处理流程，实现真正的端到端目标检测。

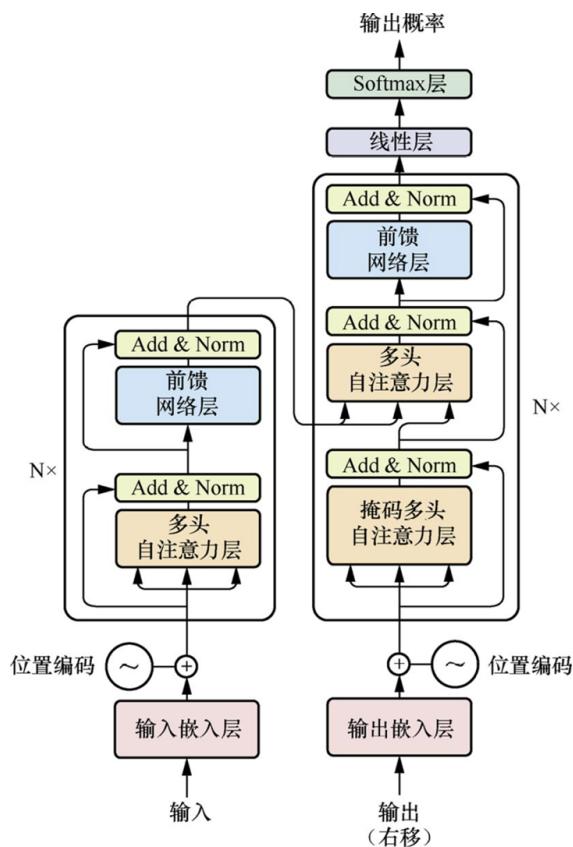


图 2-2 Transformer 架构示意图

Transformer 最初为自然语言处理中 Seq2Seq 任务设计，其核心在于多头自注意力机制，该机制允许模型动态地权衡序列中所有元素（在视觉任务中即为图像块或特征向量）之间的关系，从而捕捉全局上下文，整体架构如图所示2-2。对于输入序列 $X \in \mathbb{R}^{n \times d}$ ，自注意力通过三个可学习的线性变换矩阵 W^Q , W^K , W^V ，将其分别投影为查询 (Query) 向量 Q 、键 (Key) 向量 K 和值 (Value) 向量 V :

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \quad (2-5)$$

注意力权重通过查询与键的点积计算，并经 Softmax 归一化，最终输出为值向量的加权和：

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2-6)$$

其中， $\sqrt{d_k}$ 为键值的维度开方，作为缩放因子，用于防止点积过大导致梯度消失。多头自注意力将上述过程并行执行 h 次（每个头使用不同的投影矩阵），并将结果拼接后再投影，使模型能同时关注来自不同表示子空间的信息：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2-7)$$

其中， $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ ， W^O 为输出投影矩阵。

为了处理二维图像，需要将图像 $I \in \mathbb{R}^{H \times W \times C}$ 转换为序列形式。主流方法有两种：一种是将图像划分为固定大小的图像块（Patch），如 ViT 中将图像划分为 $P \times P$ 的块，并将每个块展平为一维向量，形成长度为 $N = \frac{HW}{P^2}$ 的序列。另一种是直接使用卷积神经网络提取的特征图作为输入序列，如 DETR 模型中使用 ResNet 作为主干网络，提取的二维特征图被展平为一维序列输入 Transformer 编码器。

由于自注意力机制本身具有置换不变性，必须注入位置信息以保留图像的空间结构。通常将可学习或预设的位置编码 $E_{pos} \in \mathbb{R}^{N \times d}$ 与输入序列相加，位置编码可以是正弦余弦函数形式（如原始 Transformer 中使用的）或可训练的嵌入向量。

$$Z_0 = X + E_{pos} \quad (2-8)$$

编码器由多个相同层堆叠而成，每层包含两个核心子层，分别是多头自注意力层和前馈网络，用于对输入序列进行深度全局建模，每个自注意力层后接一个前馈网络，通常由两个线性变换与一个非线性激活函数（如 GELU）组成，每个子层均采用残差连接和层归一化以稳定训练和加速收敛，编码器最终输出深度编码后的特征序列 $Z_{enc} \in \mathbb{R}^{N \times d}$ 。

解码器接收编码器输出的特征序列和一组可学习的目标查询（Query）向量 $Q_{obj} \in \mathbb{R}^{M \times d}$ ，其中 M 为预设的最大检测目标数。目标查询向量与传统检测器中预定义在图像网格和尺度上的锚框不同，目标查询向量是与位置解耦的、内容驱动的抽象实体。它们通过学习，掌握的是“目标应该是什么样”的语义概念，而非“目标可能在哪里”的空间先验。

解码器层的核心是掩码多头自注意力和编码器-解码器交叉注意力。在自然语言处理中，掩码用于防止解码器在生成下一个词时“看到”未来的信息，而在视觉应用中，掩码用于自监督预训练的图像建模或用于调控注意力范围的注意力。在交叉注意力中，目标查询作为 Q_{obj} ，编码器输出分别通过线性变换生成 K 和 V ，实现目标查询与全局图像特征的交互，从而定位和识别目标。

$$\text{CrossAttention}(Q_{obj}, Z_{enc}) = \text{Softmax} \left(\frac{Q_{obj}(Z_{enc}W^K)^T}{\sqrt{d_k}} \right) Z_{enc}W^V \quad (2-9)$$

生成的 $M \times N$ 注意力权重矩阵表示每个目标查询对编码器输出中每个位置特征的关注程度，通过训练，每个查询学会将高权重分配给与其所代表目标相关的图像区域，这使得每个目标查询能够有选择地从编码器输出的全局特征中收集与特定目标最相关的信息。

预测头由两个独立的全连接层（或小型 FFN）并行构成，分别作用于每个解码后的查询向量，一个线性层将查询向量投影到 $C + 1$ 维，经过 Softmax 激活函数后得到每个类别的概率分布（ C 为目标类别数，+1 表示背景类）。另一个线性层将查询向量投影到 4 维，通常通过一个 Sigmoid 激活函数将输出边界框坐标归一化到 [0,1] 范围内，代表归一化后的边界框中心坐标和宽高。最终模型直接输出 M 个无序的预测集合，通过二分图匹配损失（如匈牙利匹配算法结合交叉熵和 L1 损失）进行端到端训练，无需非极大值抑制等后处理。

2.2.2 常用数据集

目标检测算法的快速发展，与大规模、高质量数据集的建立和迭代密不可分。以 PASCAL VOC、ImageNet、MS COCO 为代表的通用场景数据集，在类别多样性、场景复杂性和标注规模上不断发展，为算法研究提供了坚实的基础。然而，当聚焦于无人机特有的航拍视角时，可供研究使用的专用数据集不论在目标种类还是数据规模上，均与通用数据集存在显著差距，此外，航拍红外小目标数据集更为稀缺，已成为制约该细分领域发展的关键瓶颈之一。本节将系统梳理现有的、适用于无人机平台的红外与可见光航拍图像目标检测数据集。

1) VisDrone

VisDrone 数据集是由天津大学 AISKEYEYE 团队创建的一个大规模、专注于无人机视觉的基准数据集。该数据集旨在为无人机平台上的视觉算法研发与评估提供支持，其数据通过多种无人机平台（包括 DJI Mavic, Phantom 系列等）在中国 14 个不同城市的各种城市与郊区环境中采集，涵盖了多样的天气、光照条件以及稀疏到拥挤的不同场景密度。它是该领域迄今为止最广泛的数据集，共包含 263 个视频剪辑（总计 179,264 帧）和 10,209 张静态图像。训练集包含 6471 张图像，验证集包含 548 张图像，此外测试集还进一步划分为“test-challenge”（1,580 张）和“test-dev”（1,610 张）。标注对象详细区分为行人、人、汽车、厢式货车、公交车、卡车、摩托车、自行车、带篷三轮车和三轮车等 10 个类别，数据集中大部分标注集中在车和人，因此存在显著的类别不平衡问题，例如，带篷三轮车的实例数量还不到汽车实例数量的四十分之一。所有标注均附加了遮挡、截断等丰富属性。自发布以来，该数据集已成为无人机视觉领域最具影响力的数据集之一，极大推动了该方向的研究与发展。



图 2-3 Visdrone 数据集示例

2) UAVDT

UAVDT 数据集是一个专为无人机巡检领域目标检测任务设计的数据集。该数据集由中国科学院大学、哈尔滨工业大学和美国德克萨斯大学圣安东尼奥分校的研究团队于 2018 年联合发布，训练集包含 23258 张图像，测试集包含 15069 张图图像，涵盖汽车、卡车、公交车和其他车辆共 4 个类别。其核心特点在于提供了超越常规类别标签的、极为丰富的多维度标注信息，每张图像和标注对象都附带序列标签和目标 ID。同时，数据可从摄像机视角（前视、侧视、鸟瞰）、目标截断程度、飞行高度、遮挡程度和天气条件（日光、夜间、雾天）等多个维度进行划分与评估。这种精细的标注体系使得 UAVDT 能够支撑分析和改进算法在复杂真实场景下的性能，使其成为车辆检测方向的关键数据集之一。



图 2-4 UAVDT 数据集示例

3) AI-TOD

AI-TOD 是一个专为航空图像中极小目标检测任务设计的基准数据集。该数据集包含 28036 张航拍图像，标注了 8 个类别的超过 70 万个目标实例，类别包含飞机、桥梁、储罐、船舶、游泳池、车辆、行人、风车。其最显著的特点是目标尺寸极小，平均大小仅约 12.8 像素，远小于其他航空影像数据集，其中高达 86% 的对象小于 16 像素，对现有检测算法构成了独特挑战。AI-TOD 主要被用于开发和评估专门针对微小目标的检测算法。

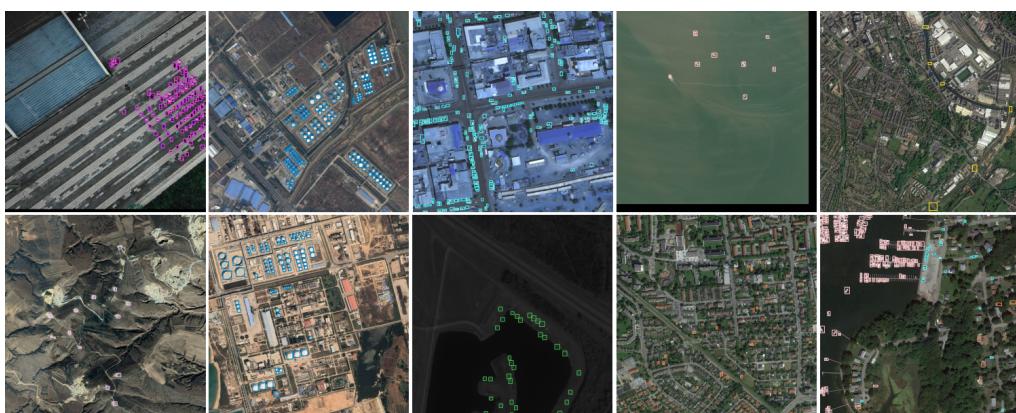


图 2-5 AI-TOD 数据集示例

4) HIT-UAV

HIT-UAV 是首个公开可用的、专用于检测人员和车辆的高空无人机红外热成像数据集。该数据集包含 2898 张图像（训练集 2029 张，测试集 579 张，验证集 290 张），标注了人员、汽车、自行车、其他车辆和难以确定的物体共 5 个类别的 24899 个对象。所有数据均使用搭载 DJI Zenmuse XT2 相机的无人机在 60 至 130 米高空采集，包含了白天、夜晚、多种地点与摄像机视角的场景，专注于红外热成像模态，这使得其在低光照、夜间或恶劣天气条件下具有独特优势。为提升实用性，数据集创新性地为每个对象同时提供了旋转边界框（解决空中目标重叠问题）和标准边界框两种标注。该数据集有效推动了无人机在复杂环境下，特别是全天候条件下的目标检测应用研究。

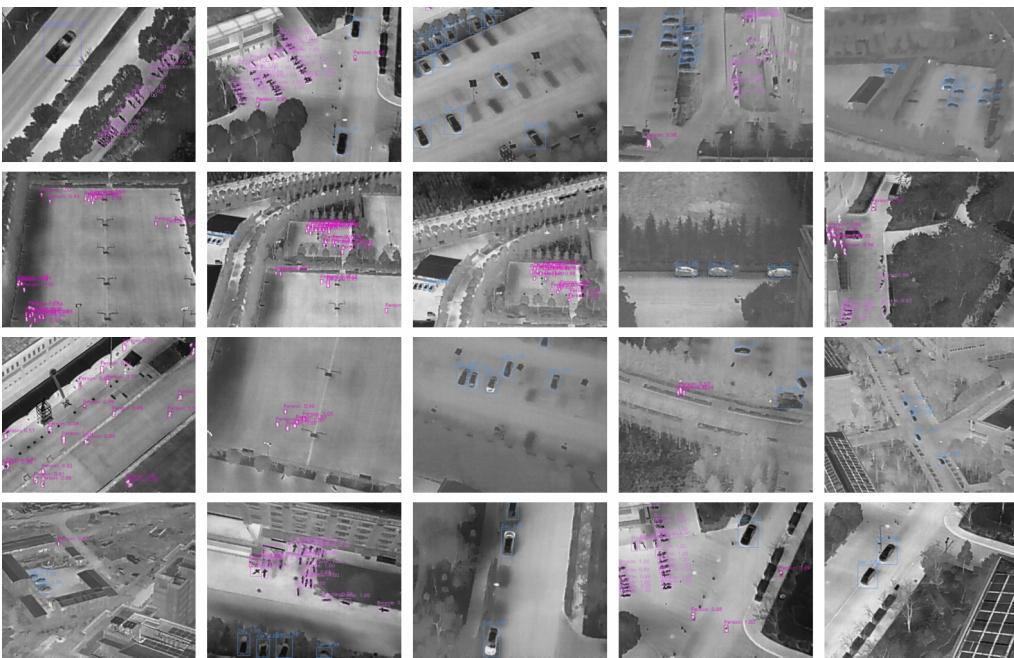


图 2-6 HIT-UAV 数据集示例

5) DroneVehicle

DroneVehicle 是一个专为无人机视角下的跨模态车辆检测研究而设计的大规模数据集。该数据集同时提供了可见光和红外两种模态的图像，各占一半，共计 56,878 张图像，专门用于研究不同光照与天气条件下的车辆检测。数据集中标注了汽车（car）、卡车（truck）、公交车（bus）、厢式货车（van）和货运车（freight car）这五类车辆，标注实例总数超过 90 万个，并采用了更贴合空中视角物体形状的旋转边界框（oriented bounding boxes）。**DroneVehicle** 主要服务于推动基于无人机的 RGB-红外感知算法的研究。



图 2-7 DroneVehicle 数据集示例

2.2.3 评价指标

评价指标为理解模型的能力边界与优化方向提供了标准化的度量框架，经过多年发展，研究者们定义了一系列标准化的评价指标，本节将系统梳理目标检测领域的核心评价指标，阐述其计算原理及其在算法评估中的具体意义。

一个预测框是否正确通过计算预测框与真值框的交并比（Intersection over Union, IoU）来确定，其定义为预测框与真值框的交集面积与并集面积之比：

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{B_{pred} \cap B_{gt}}{B_{pred} \cup B_{gt}} \quad (2-10)$$

早期标准中认为预测框与真值框 IoU 大于 0.5 即为正确。然而，随着检测算法性能的提升，单一的 IoU 阈值已无法全面反映模型在不同定位精度下的表现。为此，后续评测标准引入了多阈值评估机制，如 COCO 数据集采用从 0.5 到 0.95（步长为 0.05）的 10 个 IoU 阈值，计算平均精度，以更细粒度地衡量模型在不同定位要求下的性能。

在构建评估指标之前，通常需要根据预测结果与真实标签的对应关系，统计四种预测结果：真正例（True Positive, TP）、假正例（False Positive, FP）、真负例（True Negative, TN）和假负例（False Negative, FN）。如图2-8所示，通过这四类预测情况可以构成混淆矩阵（Confusion Matrix）来可视化模型的分类性能。

		Predict 模型预测	
		0	1
Real 实际情况	0	Ture Negative 真阴性	False Positive 伪阳性
	1	False Negative 伪阴性	True Positive 真阳性

图 2-8 预测结果混淆矩阵

基于以上四种情况，可以定义一系列核心评价指标：

1) 精确率 (Precision)

精确率表示模型预测为正的样本中实际为正类的比例，定义为：

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2-11)$$

2) 虚警率 (False Positive Rate, FPR)

虚警率表示模型预测为正的样本中实际为负类的比例，定义为：

$$\text{FPR} = \frac{FP}{TP + FP} \quad (2-12)$$

3) 召回率 (Recall)

召回率表示模型正确识别出的正样本占所有实际正样本的比例，定义为：

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2-13)$$

4) 平均精度 (Average Precision, AP)

平均精度是通过计算不同召回率下的精确率曲线下面积来衡量模型整体性能的指标。具体计算方法为：

$$\text{AP} = \int_0^1 \text{Precision}(r) dr \quad (2-14)$$

其中， $\text{Precision}(r)$ 表示在召回率 r 下的精确率。AP 值越高，表示模型在不同召回率下的精确率表现越好，这里的 AP 是对单个类别计算，衡量模型对某一个类别的检测能力。

5) 平均精度均值 (Mean Average Precision, mAP)

平均精度均值是对所有类别的平均精度进行平均得到的指标，定义为：

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i \quad (2-15)$$

其中， N 为类别总数， AP_i 为第 i 类的平均精度。mAP 值越高，表示模型在所有类别上的整体性能越好。

在 PASCAL VOC 评价体系中，AP 与 mAP 是两个独立的指标，mAP 是所有类别 AP 的平均值，在 COCO 数据集评价体系中，只保留了 AP 的概念，但是这里的 AP 的计算方式实际是 PASCAL VOC 中的 mAP，并且更加精细化，即对所有类别和多个 IoU 阈值下的 AP 进行平均。

2.3 机载光电系统目标跟踪算法研究基础

目标跟踪的任务是在连续的图像序列中，对特定目标进行持续锁定与状态估计，是智能光电系统实现持续态势感知的关键功能，其性能决定了系统能否在动态复杂的真实环境中保持稳定、可靠的观测能力。机载光电系统所处的动态空基环境对目标跟踪算法提出了严苛的要求。平台的高速运动、高度变化带来的尺度与视角剧烈变动，复杂的地面背景与遮挡的干扰，以及机载计算资源固有的严格限制，对跟踪算法的设计提出了更高要求。本节将系统介绍单目标跟踪算法的研究基础，涵盖基本方法、常用数据集

与核心评价指标。

2.3.1 基本方法

主流跟踪算法经历了从依赖手工特征与轻量级互相关运算的判别式相关滤波，到利用大规模数据驱动的孪生网络，再到通过全局注意力实现上下文建模的 Transformer 架构的转变。判别式相关滤波方法因其卓越的计算效率成为机载平台实时跟踪的首选。该类方法通过在线学习一个滤波模板，在频域利用快速傅里叶变换实现与候选区域的密集相关运算，从而快速定位目标。其优势在于计算速度极快，能满足高频帧率下的实时处理需求。孪生网络将目标跟踪构建为一次性的模板匹配学习问题，通过一个共享权重的卷积神经网络，分别提取模板帧中目标区域和搜索帧中候选区域的特征，然后进行互相关操作，生成响应图以定位目标。

然而，孪生网络的深度模型结构带来了显著的计算负载，为实现其在高频帧率下的实时运行，必须依赖特定边缘计算平台的专用工具链进行加速优化。这一过程伴随着模型精度的损失，优化后的网络性能虽可勉强满足实时性要求，但对系统资源占用巨大。Transformer 架构的模型即使经过上述加速优化，在当前的机载边缘计算设备上仍难以达到实时处理要求。在性能表现上，孪生网络在标准评测数据集上的精度指标相比相关滤波方法通常有大幅提升。但深入工程实践发现，在大部分常规跟踪场景中，目标特征明显、运动相对平缓且无严重遮挡，经过调优的相关滤波方法与孪生网络在实际跟踪效果上差异并不显著，二者均能可靠完成任务。在极端挑战场景下，例如目标被完全、长时间遮挡后重现，两类方法均面临跟踪失败的风险，其根本原因在于它们都难以解决判定目标丢失与重识别问题。仅在少数情况下，例如目标发生突然形变，孪生网络能表现出比相关滤波方法肉眼可见的性能优势。在这一节中，我们将系统介绍单目标跟踪的两大主流方法：基于相关滤波的目标跟踪和基于孪生网络的目标跟踪。

1) 基于相关滤波的目标跟踪

基于相关滤波跟踪算法的核心，相关滤波器 (Correlation Filter, CF)，最早应用于信号处理领域，用来描述两个信号之间的相关性。在视觉跟踪中，该框架将跟踪问题转化为一个模板匹配的滤波学习问题，目标在线学习一个线性滤波器，使得当该滤波器与目标图像块进行相关运算时，产生一个理想的响应图，在目标中心产生峰值，而在背景区域响应平缓，通过在图像中寻找响应最大值的位置，即可实现目标的快速定位。相关滤波跟踪器的关键优势在于利用循环卷积和快速傅里叶变换，将密集采样和复杂的时域运算转化为高效的频域元素级操作，实现极快的跟踪速度，这使其在计算资源受限的机载平台上广泛应用。

其理论基础可表述如下：给定一组训练样本 $\{x_i, y_i\}_{i=1}^N$ ，其中 x_i 为尺寸 $N_1 \times N_2$ 的特征图， y_i 为对应的期望响应图，通常定义为以目标位置为中心的高斯函数。目标是学习一个尺寸同样为 $N_1 \times N_2$ 滤波器 w ，使得对于每个样本 x_i ，其与滤波器的相关运算结

果接近期望响应 y_i 。相关运算定义为：

$$x * w(n_1, n_2) = \sum_{l_1=0}^{N_1-1} \sum_{l_2=0}^{N_2-1} x((n_1 - l_1)_{N_1}, (n_2 - l_2)_{N_2}) w(l_1, l_2) \quad (2-16)$$

滤波器的学习通过最小化以下损失函数实现：

$$L(w) = \sum_{i=1}^N \|x_i * w - y_i\|^2 + \lambda \|w\|^2 \quad (2-17)$$

其中， λ 为正则化参数，用于防止过拟合。通过求解该优化问题，可以得到滤波器 w 的闭式解：

$$\mathcal{F}(w) = \frac{\sum_{i=1}^N \mathcal{F}(x_i) \odot \overline{\mathcal{F}(y_i)}}{\sum_{i=1}^N \mathcal{F}(x_i) \odot \overline{\mathcal{F}(x_i)} + \lambda} \quad (2-18)$$

其中， \mathcal{F} 表示傅里叶变换， \odot 表示元素级乘法， $\overline{\mathcal{F}(x_i)}$ 表示 $\mathcal{F}(x_i)$ 的复共轭，该解完全由元素级乘除法和 FFT/IFFT 操作构成。由于 FFT 的复杂度为 $O(N \log N)$ ，相比时域的 $O(N^2)$ 运算，极大提升了计算效率。跟踪时，在图像中提取图像块 z ，其响应图计算如下：

$$R = \mathcal{F}^{-1}(\mathcal{F}(w) \odot \mathcal{F}(z)) \quad (2-19)$$

其中， \mathcal{F}^{-1} 表示傅里叶逆变换，响应图 R 中最大值的位置即为目标位置。

为了实现在线学习并适应目标外观变化，滤波器 w 在每一帧通过加权移动平均的方式进行更新：

$$w_t = (1 - \gamma)w_{t-1} + \gamma w_{new} \quad (2-20)$$

其中， w_{new} 为当前帧计算得到的新滤波器， γ 为学习率，控制模型更新速度，较大的学习率使模型更快地适应新外观，但对噪声更敏感，较小的学习率使模型更稳定，但可能无法跟上剧烈的外观变化。

在此基础上，后续多个经典方法从不同层面对相关滤波跟踪器进行了改进。在特征层面，研究从最初的灰度特征发展至 HOG、颜色等手工特征，并最终成功融合了深度卷积特征。在模型层面，为缓解边界效应，研究者们提出了空间正则化、背景感知等策略，约束滤波器在目标中心区域学习，有效抑制了由循环移位引入的虚假背景干扰，为适应目标尺度变化，引入了多尺度搜索或独立的尺度滤波器。在效率层面，通过引入核函数、采用因式分解的卷积算子以及设计稀疏更新策略，在提升跟踪精度的同时，显著降低了计算与存储开销，增强了算法的实用性。

2) 基于孪生网络的目标跟踪

孪生网络是近年来视觉目标跟踪领域另一主流框架，其核心思想是将跟踪视为一个相似性学习问题。与在线更新模型的判别式相关滤波不同，孪生网络跟踪器采用离

线训练、在线匹配的策略，预先在大量数据上学习一个通用的、具有判别力的特征嵌入空间和匹配函数，在跟踪时，通过计算第一帧目标模板与后续帧搜索区域在该空间中的相似性，直接定位目标。

其基本架构由两个权重共享的子网络分支构成，分别用于提取目标模板和搜索区域的特征，给定一个模板图像 x 和一个搜索图像区域 z ，孪生网络的目标是学习一个匹配函数 $g_\rho(\cdot)$ ，输出一个响应图 $g_\rho(x, z)$ ，其峰值位置对应于搜索区域中目标的位置，在 SiamFC 中，该函数定义为：

$$g_\rho(x, z) = f_\rho(x) * f_\rho(z) + b \quad (2-21)$$

其中， $f_\rho(\cdot)$ 为共享参数的卷积神经网络， b 为可学习的偏置项， $*$ 表示互相关操作，响应图 $g_\rho(x, z)$ 的每个位置值表示搜索区域中对应位置与模板的相似度。

离线训练的目标是优化网络参数 ρ ，使得正样本对的响应值高，负样本对的响应值低，通常采用逐元素的逻辑损失：

$$L(c, v) = \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-c \cdot v)) \quad (2-22)$$

其中 v 是响应图中某位置的相似度得分， c 为该位置的标签，取值为 $+1$ 表示正样本， -1 表示负样本。整个训练集由大量样本对 $\{(x_i, z_i, y_i)\}_{i=1}^N$ 组成，最终的训练目标为最小化以下总损失：

$$\min_{\rho} \frac{1}{N} \sum_{i=1}^N L(c, g_\rho(x_i, z_i)) \quad (2-23)$$

在跟踪过程中，第一帧中的目标区域被提取为模板 x ，后续每一帧从图像中提取搜索区域 z ，通过计算响应图 $g_\rho(x, z)$ ，找到峰值位置作为目标位置。

后续工作中研究者对于此基础框架的优化主要集中在以下几个方面：在特征提取层面，通过引入零填充消除和分层聚合技术（如 SiamRPN++），成功将 ResNet 等深层主干网络融入孪生框架，实现了高分辨率细节与高层语义信息的有效融合，突破了浅层网络（如 AlexNet）的表征瓶颈。在目标状态估计层面，其方法从计算密集且粗糙的多尺度搜索，发展为引入区域提议网络（RPN）的锚框回归（如 SiamRPN），实现了位置与尺度的联合精确预测，并进一步发展为更灵活简洁的无锚框回归（如 SiamBAN、Ocean），通过直接预测边界框参数，减少了超参数并提升了泛化能力。在模型自适应层面，为克服初始模板僵化的问题，研究从简单的线性加权平均更新，发展到利用外部记忆模块动态存储与读取历史模板（如 MemTrack），乃至训练专门的元更新网络（如 UpdateNet）来学习通用的模板更新策略，显著增强了对目标外观剧烈变化的鲁棒性。这些改进显著提升了跟踪精度和鲁棒性，然而，其计算复杂度和资源需求也随之增加，在机载平台的实时应用中面临挑战。

2.3.2 常用数据集

本节将介绍单目标跟踪领域中常用的评测数据集，这些数据集涵盖了多种应用场景和挑战因素，为算法的开发与评估提供了标准化的平台。

1) OTB

OTB (Object Tracking Benchmark) 数据集是单目标跟踪领域一个具有里程碑意义的基准数据集，它首次为跟踪算法提供了统一的量化标准，极大地推动了该领域的发展。该数据集最初于 2013 年以 OTB-50 (包含 50 个视频序列) 发布，后于 2015 年扩展为 OTB-100 (包含 100 个序列)。其视频涵盖了光照变化、尺度变化、遮挡、形变、快速运动等 11 类常见的跟踪挑战属性。OTB 的一个显著特点是其数据混合了约 25% 的灰度序列和 75% 的彩色序列，以测试算法在不同输入下的性能。它首创并广泛使用了一次通过评估 (One-Pass Evaluation, OPE)、以及考虑时间与空间扰动的鲁棒性评估方法，其核心评估指标是精确度图 (Precision Plot) 和成功率图 (Success Plot)，后者曲线下面积 (AUC) 常作为算法排名的关键依据。作为早期权威基准，OTB 至今仍被广泛用于算法验证。



图 2-9 OTB 数据集示例

2) VOT

VOT (Visual Object Tracking Challenge) 数据集是单目标跟踪领域的权威基准之一，它并非一个固定不变的数据集，而是一个以年度竞赛形式更新、分支多样的评测体系。自创办以来，VOT 每年会发布新的挑战数据集（如 VOT2022、VOT2025 等），每个版本包含数十个精心挑选的、包含各种挑战因素的视频序列。VOT 不仅提供了经典的精确度图 (Precision Plot) 和成功率图 (Success Plot)，还引入了期望平均重叠率 (EAO) 这样的综合指标来衡量跟踪器的整体表现。除了主流的 RGB 的短时、长时跟踪任务，VOT 也拓展到了 RGB-D (深度) 等多模态领域，以促进算法在更复杂场景下的鲁棒性研究。

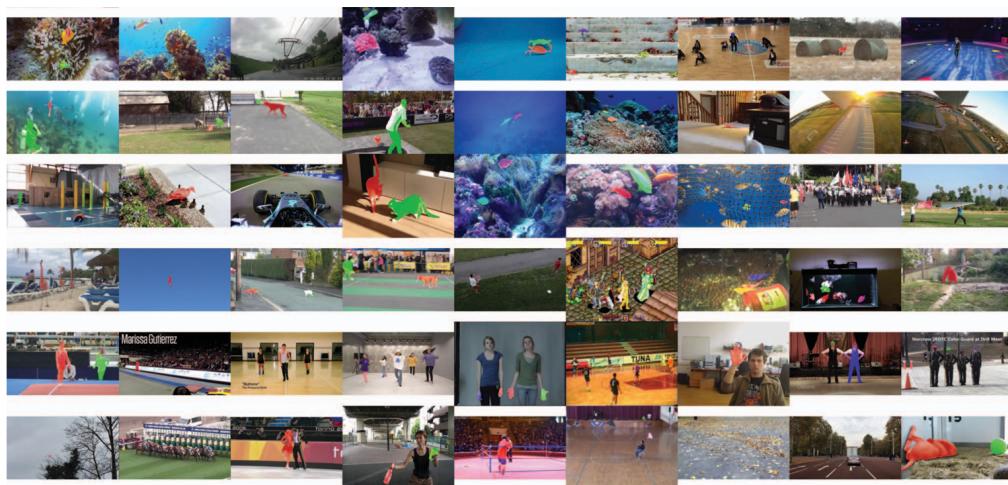


图 2-10 VOT 数据集示例

3) LaSOT

LaSOT (Large-scale Single Object Tracking) 是单目标跟踪领域一个高质量、大规模的数据集，最初于 CVPR 2019 发布。为克服早期数据集规模小、视频短、类别不均衡等局限，其设计核心在于大规模、长时跟踪和类别平衡，数据集包含 1550 个视频与超过 387 万帧图像，平均每个视频长达约 2500 帧，支持对跟踪器长时稳健性的评估，它涵盖了 85 个物体类别，且每个类别严格包含 20 个视频，有效抑制了类别偏差。LaSOT 的每一帧都提供了高质量、手工标注的密集边界框，并且创新性地为每个视频配备了自然语言描述，以促进结合语言特征的跟踪研究。该数据集已成为训练深度网络跟踪模型主流基准之一。



图 2-11 LaSOT 数据集示例

4) GOT-10k

GOT-10k 是一个由中国科学院自动化研究所发布的大规模、多样性的通用目标跟踪数据集，旨在为类无关的通用短时跟踪器提供一个统一的训练和评估平台。其核心特点在于“规模大”与“多样性高”，数据集提供了超过 10,000 段视频片段和超过 150 万个手工标注的边界框，并创新性地依据 WordNet 语义层次结构构建，覆盖了超过 560 个移动物体类别和 87 种运动模式，以确保对现实世界物体的广泛且相对无偏见的覆盖。GOT-10k 最具影响力的贡献是引入了严格的“单次评估”(one-shot protocol)，即训练集和测试集的物体类别完全无重叠，这迫使算法必须学习通用的跟踪能力，而非记忆特定物体，从而能更真实地评估跟踪器的泛化性能。

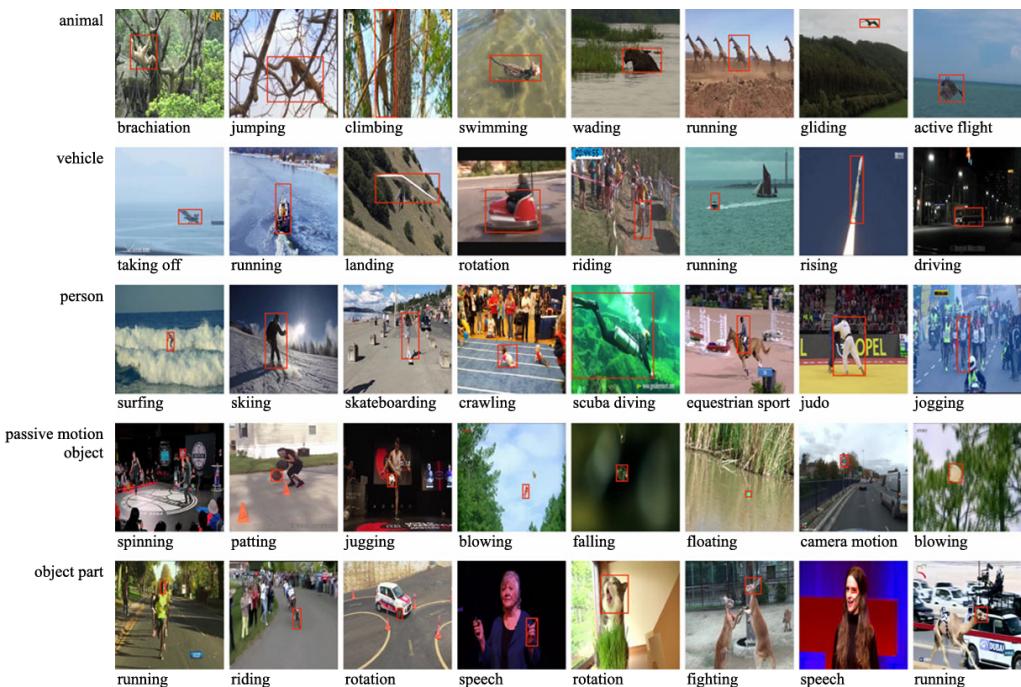


图 2-12 GOT-10k 数据集示例

5) UAV123

UAV123 数据集是专为评估无人机 (UAV) 视角下目标跟踪算法性能而设计的基准数据集，共包含 123 个由低空无人机拍摄的视频序列，总计超过 11 万帧图像。该数据集的核心价值在于其独特的无人机俯拍视角，这与 OTB、VOT 等基于地面视角的传统数据集有本质区别，能更好地模拟和测试算法在真实空中监视场景下面临的挑战，如目标尺度剧烈变化、复杂背景和运动模糊等。所有序列均提供了精确的手工标注边界框，并且数据集还专门包含了一个用于长时跟踪评估的子集 UAV20L，从而使其成为推动空中目标跟踪技术发展的重要资源。

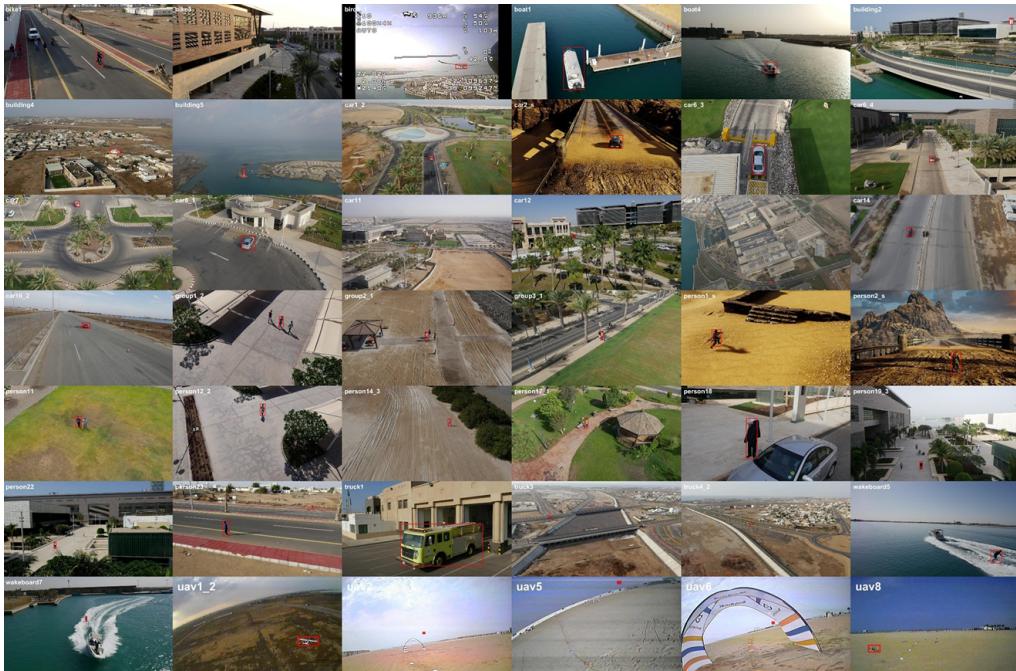


图 2-13 UAV123 数据集示例

2.3.3 评价指标

单目标跟踪算法评估主要包括两个方面，分别是跟踪精度和跟踪成功率，跟踪精度反应了跟踪器定位目标的准确程度，跟踪成功率则反映了跟踪器在整个序列中成功跟踪目标的能力。常用的评价指标包括以下几种：

1) 中心位置误差 (Center Location Error, CLE)

中心位置误差是指跟踪器预测的目标中心位置与真实目标中心位置之间的欧氏距离，计算公式为：

$$\text{CLE} = \sqrt{(x_{pred} - x_{gt})^2 + (y_{pred} - y_{gt})^2} \quad (2-24)$$

其中， (x_{pred}, y_{pred}) 为预测目标中心位置， (x_{gt}, y_{gt}) 为真实目标中心位置。CLE 值越小，表示跟踪器定位越准确。

2) 重叠率 (Overlap Rate, OR)

重叠率是指跟踪器预测的目标边界框与真实目标边界框之间的交并比 (IoU)，计算公式为：

$$\text{OR} = \frac{B_{pred} \cap B_{gt}}{B_{pred} \cup B_{gt}} \quad (2-25)$$

其中， B_{pred} 为预测边界框， B_{gt} 为真实边界框。OR 值越大，表示跟踪器预测的边界框与真实边界框越接近。

3) 精确度 (Precision)

精确度是指在所有测试帧中，中心位置误差小于某一阈值（通常为 20 像素）的帧数占总帧数的比例，计算公式为：

$$\text{Precision} = \frac{\mathcal{N}_{CLE < \text{threshold}}}{\mathcal{N}_{total}} \quad (2-26)$$

精确度值越高，表示跟踪器定位目标的能力越强。

4) 成功率 (Success Rate)

成功率是指在所有测试帧中，重叠率大于某一阈值（通常为 0.5）的帧数占总帧数的比例，计算公式为：

$$\text{Success Rate} = \frac{\mathcal{N}_{OR > \text{threshold}}}{\mathcal{N}_{total}} \quad (2-27)$$

成功率值越高，表示跟踪器在整个序列中成功跟踪目标的能力越强。

2.4 本章小结

本章构建了智能光电处理算法与系统设计的理论框架与研究基础，分析了智能光电系统的基本组成与工作原理。在核心算法层面，系统梳理了从卷积神经网络到 Transformer 架构的发展历程，以及每种架构的基本原理。本章还梳理了单目标跟踪算法的研究基础，围绕相关滤波和孪生网络两大框架，分析了其各自的原理、优势以及在机载平台上面临的工程挑战。同时，本章介绍了支撑算法研发与评估的常用数据集与评价指标，特别是航拍视角下的目标检测与跟踪数据集。

3 基于双重注意力处理的高效可见光航拍图像小目标检测算法

无人机航拍图像中的目标检测面临着一系列严峻挑战，包括极端尺度变化、密集分布的小目标以及复杂的背景，这导致通用目标检测器在此类场景下性能显著下降。最直接的解决方案是提升输入图像的分辨率，但这会急剧增加计算负担。现有方法由于在网络架构上存在缺陷，难以在保持对小目标检测至关重要的细粒度特征的同时，实现精度与速度的平衡。为此，本章节设计了一种基于 RT-DETR 框架的优化模型架构，提出了基于通道分离的双重注意力处理模块，通过通道分离策略，实现了卷积操作与注意力机制并行处理，从而显著增强了模型从复杂航拍图像中提取判别性特征的能力，同时引入频率感知融合模块，能够有效保留关键的低层细节特征，并将其与高层语义信息深度融合。

3.1 引言

作为无人机的核心感知单元，机载智能光电系统在对地观测、目标识别与态势感知等关键任务中发挥重要作用。目标检测算法使无人机能够识别并定位图像中的物体，从而增强其自主环境感知能力。随着卷积神经网络与视觉 Transformer 的快速发展，通用目标检测器在 MS COCO^[85]等通用图像数据集上取得了显著进步。然而，面对航拍图像时，通用检测器的性能出现显著下降。例如，当前最受欢迎的基于 CNN 的检测器 YOLOv11-M^[86]，在 MS COCO 数据集上的 AP_{50} 为 51.5%，但在 VisDrone^[87]数据集上仅为 43.1%，基于 Transformer 的端到端检测器 RT-DETRv2-S^[88]在 MS COCO 上 AP_{50} 为 63.8%，在 VisDrone 上则为 45.3%。这一明显的性能差距表明，无人机视角下的目标检测技术仍需改进。

与常规图像相比，由于无人机拍摄高度和角度的变化，航拍图像中的目标通常表现出以下三个显著特征^[89]：1) 小目标占比极高；2) 小目标往往在特定区域密集聚集；3) 同类目标的尺度变化极为剧烈。如图3-1所示，VisDrone 数据集中大多数目标的尺寸小于 20 像素。密集区域中的目标可能相互重叠与遮挡，导致漏检与误检。这些因素严重影响了检测性能，并制约了无人机自主感知系统的可靠性。

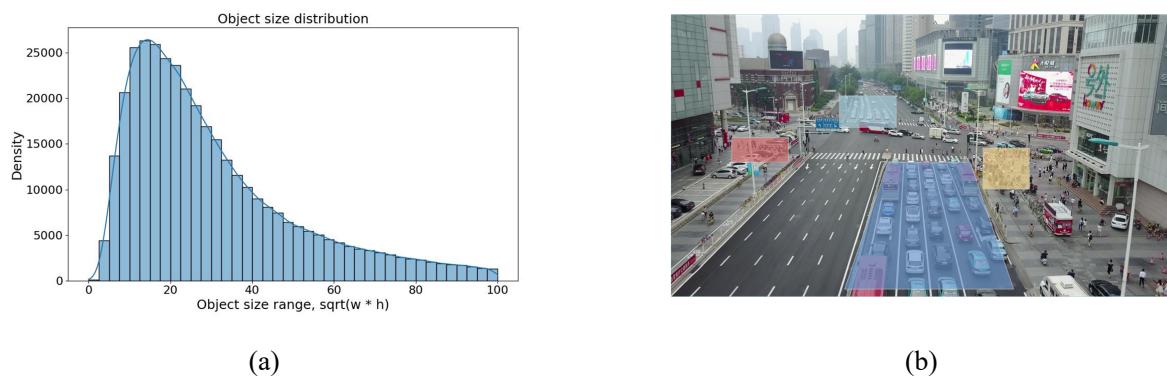


图 3-1 (a)VisDrone 数据集中目标的平均尺寸分布，其中，横轴的目标尺寸由边界框宽度与高度乘积的平方根计算得出。如图所示，绝大多数目标的尺寸集中在 20 像素以内。
 (b) 小目标在特定区域呈现密集分布

检测密集分布的小目标是一项极具挑战的任务，这主要源于其有限的有效像素难以提供具有差异化的视觉特征，且在高密度聚集时易产生相互遮挡与背景混淆，导致传统检测方法难以实现稳定、可靠的识别与定位。常见解决方案包括多尺度特征融合^[20-21]、数据增强^[4]、背景建模^[90]或者聚焦检测^[50]。然而，这些方法通常难以解决密集小目标差异化特征表征不足的问题，而直接提升图像分辨率或进行裁剪的方法则会带来难以承受的计算开销，且计算常浪费在背景区域。此外，基于 CNN 的检测器存在超参数敏感性问题，在目标重叠区域容易因非极大值抑制而产生误差^[91]。对于航拍图像，极端尺度变化和复杂背景加剧了这些问题。尽管如 DETR^[92]等端到端检测器通过集合预测摒弃了 NMS 和锚框，但其对航拍场景下小目标的适应能力仍然不足。DQ-DETR^[93]作为首个专为小目标检测设计的类 DETR 模型，虽提升了精度，却以收敛速度慢和高计算量为代价。RT-DETR^[88]通过混合编码加速推理，但其架构面对密集小目标检测仍存在局限。通过分析其具体架构，我们发现其主干网络缺乏针对小目标细粒度特征的专门优化，且混合编码器仅利用高层特征，此时小目标已历经多次下采样，导致对检测至关重要的空间与纹理信息严重流失。因此，在无人机目标检测任务中同时实现实时速度与高精度仍是一个挑战。

基于上述分析，本章提出一种基于实时端到端 RT-DETR 框架的优化架构。我们的创新设计旨在针对性解决核心挑战：1) 增强差异化特征学习：利用新设计的双重注意力处理块重构主干网络，以生成更丰富、更具判别力的特征表示；2) 优化多尺度融合与小目标特征保留：提出了双重融合特征编码器，显式结合高分辨率特征以保留对小目标至关重要的细节。此外，采用频率感知融合模块来更智能地整合跨尺度特征，增强上下文建模与特征兼容性；3) 优化小目标相似性度量：将倒数归一化 Wasserstein 距离与 CIoU 相结合构建损失函数，提供了与尺度无关的边界框相似性度量，提升了密集小目标的定位精度。大量实验证实，我们的方法为航拍图像检测建立了新的效率-精度均衡标杆。

本章的主要贡献总结如下：

- 在 ResNet 主干网络中引入了双重注意力处理块。该模块将注意力机制与并行分支处理相结合，超越了简单的“分割-处理-合并”模式，确保了更具交互性的特征提取与注意力建模。
- 设计了双重融合特征编码器，以有效保留小目标的特征并改善多尺度特征融合。关键创新包括引入低层特征图、重构融合路径以及设计频率感知融合模块。
- 提出了 RNWD-CIoU 损失，该混合损失利用经过倒数归一化的 Wasserstein 距离，为边界框相似性提供了与尺度无关的度量方式，在不引入任何额外推理计算开销的情况下，提升了小目标及密集目标的检测性能。
- 在三个公开航拍图像数据集（Visdrone^[87], UAVDT^[94], AI-TOD^[95]）上的大量实验表明，BAP-DETR 以最小的计算负载和参数量，实现了最优性能，特别地，在 VisDrone 数据集上，BAP-DETR 的 *AP* 比基线模型提升了 6.9%，同时减少了 17.5% 的计算量。

3.2 相关工作

3.2.1 基于 CNN 的小目标检测方法

与通用数据集中的图像相比，航拍图像中的目标检测通常更具挑战性。这主要因为图像中存在数量更多的小目标，其有限的像素使得特征提取极为困难，且在卷积过程中，小目标的特征容易被背景或其他目标干扰。此外，这些小目标在图像中分布不均，多数呈密集聚集状态，这不仅使精确定位更为困难，也导致了更高的漏检率。

为应对小目标检测难题，研究者们提出了多种解决方案，主要包括样本导向^[4]、基于注意力^[47]、多尺度融合^[20]及聚焦检测^[50,96]等方法。这些方法的核心大多围绕如何解决因像素有限导致的低质量特征表示问题。然而，在存在极端尺度变化和密集目标的航拍场景中，这些通用方法的性能仍显不足。样本导向方法常面临性能提升不稳定和迁移性差的问题，基于注意力的方法虽凭借其灵活的嵌入设计而备受推崇，可便捷地集成到各类架构中，但其性能提升往往以复杂的关联运算所带来的沉重计算开销为代价，多尺度融合架构旨在以合适的尺度处理小目标，但不同尺度特征融合时易引入噪声或冲突，可能导致小目标的特征在融合过程中被淹没或扭曲。

针对航拍图像中密集小目标的特定挑战，一些研究提出了专门设计。例如，Query-Det^[19]设计了级联查询策略，以避免在低层特征上的冗余计算，从而能高效地在高分辨率特征图上检测小目标，但其精度依赖于初始预测的准确性。DMNet^[97]和 Dynamic Anchor^[96]等方法利用密度图来检测目标并学习尺度信息。CEASC^[98]将全局上下文集成到稀疏卷积网络中，以增强无人机图像的目标检测。SAHI^[99]采用了裁剪策略，虽提升了精度，但也增加了计算复杂度和处理时间，其均匀裁剪方法未能考虑目标分布的非均匀性，导致检测所有图像块耗时巨大，效率降低。ClusDet^[50]利用特定模块搜索可能包含目标的聚类区域，但这类方法通常训练成本高、推理速度慢，且仍需依赖非极大值抑制进行后处理，这进一步降低了推理速度，并引入了可能影响速度和精度稳定性的

超参数，尤其在处理密集小目标时更为明显。

3.2.2 基于 DETR 的小目标检测方法

DETR^[92]是首个基于 Transformer 架构的端到端目标检测模型。它通过二分图匹配机制，避免了手工锚框设计与复杂的后处理步骤，在性能上达到了与当时主流 CNN 检测器相当的水平，但因其训练收敛速度缓慢而备受制约。为此，一系列类 DETR 模型被提出以改进此问题。为加速收敛，Deformable-DETR^[100]引入了可变形注意力机制，使每个查询仅关注特征图上一小组关键采样点，从而显著提高了注意力计算效率与模型训练速度。DN-DETR^[101]采用了去噪训练，通过在输入中加入带有噪声的标注并让模型学习恢复，有效降低了二分图匹配的难度，加快了收敛进程。另一方面，为了降低模型的计算负担，Efficient DETR^[102]和 Sparse DETR^[103]等研究专注于减少编码器-解码器层数或优化查询数量。DQ-DETR^[93]是首个专门为小目标检测设计的类 DETR 模型，它通过动态调整查询数量并增强查询的位置感知能力来精准定位小目标。然而，这些方法在追求性能的同时，通常计算开销仍然较大，难以满足实时处理的需求。

RT-DETR^[104]是首个在速度和精度上均超越同期传统 CNN 检测器的实时端到端目标检测器。其核心创新在于设计了一个高效的混合编码器，该编码器融合了基于注意力的同尺度特征交互模块与基于 CNN 的跨尺度特征融合模块。这一设计显著降低了计算成本，成功将 DETR 框架拓展至实时检测场景。尽管 RT-DETR 在通用目标检测上表现出色，但由于其对小目标及复杂航拍场景的适应性有限，在处理无人机图像特有的挑战时仍显不足。针对无人机场景的后续改进工作试图弥补这一差距，但各有折衷。RT-DETR-UAVs^[91]优先保障实时性能，却在检测精度上有所牺牲，导致模型在复杂背景下难以准确识别小目标。UAV-DETR^[105]引入了结合频率增强与语义对齐的多尺度特征融合模块，以应对小目标检测挑战，但频域信息的利用可能导致不同特征图间的语义与空间信息错位，进而引发误检。HCTD^[106]提出了一种专为无人机检测设计的混合 CNN-Transformer 架构，但由于其主干网络未针对小目标优化，且对低层特征增强关注不足，在保留密集小目标所需的高分辨率空间特征方面仍存在局限。

3.3 基于双重注意力处理的目标检测网络

3.3.1 整体架构

本章提出的网络整体结构如图3-2所示。该网络以实时端到端检测器 RT-DETR 为基础，其架构主要由三部分组成：一个经 BAPB 增强的 ResNet 主干网络、一个新式的双重融合特征编码器以及一个带有辅助预测头的 Transformer 解码器。主干网络基于 ResNet 构建。在每个 ResNet 阶段的核心 ResBlock 中，集成了双重注意力处理块与 SE 注意力层。BAPB 模块旨在增强 ResNet 主干网络的特征提取能力，其设计思想是：将输入特征图沿通道维度分割为两个独立的处理流，每个流分别进行独立的卷积处理和

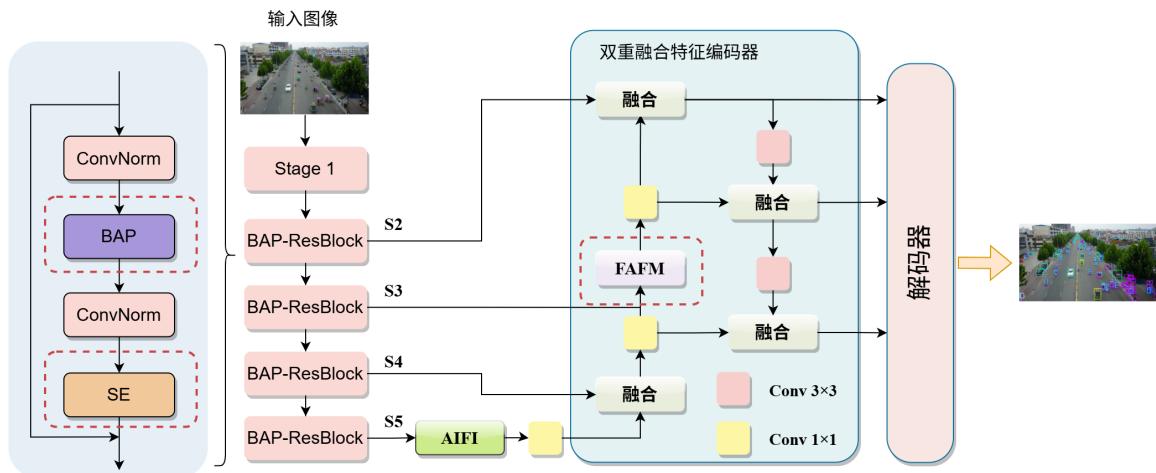


图 3-2 BAP-DETR 网络架构示意图

基于注意力的特征提取，最后将两个流的输出融合，从而为后续网络层生成判别性更强、上下文信息更丰富的输出特征图。

具体而言，主干网络最终会生成四个层级的特征图： S_2 、 S_3 、 S_4 和 S_5 ，记为 $S_i = R^{C \times H \times W}$ 。其中， S_2 、 S_3 、 S_4 和 S_5 的空间尺寸分别对应输入图像的 $1/4$ 、 $1/8$ 、 $1/16$ 和 $1/32$ 。这些多尺度特征是后续处理的基础。双重融合特征编码器负责集成主干网络输出的多尺度特征。在原始 RT-DETR 中，其高效混合编码器对高层特征图 S_5 （蕴含最丰富的语义信息）使用了基于注意力的同尺度特征交互模块，这极大地减少了 Transformer 编码器层带来的计算开销，是本模型实时性的关键之一。然而，随着网络加深，小目标的特征响应会逐渐减弱。相反，低层特征图（如 S_2 ）虽然语义抽象程度低，但保留了更丰富的纹理、边缘等高频细节信息，这对小目标的精确定位至关重要。因此，我们的编码器同时将高层特征 S_5 和低层特征 S_2 作为输入。双融合特征编码器包含一个独特的特征融合模块，频率感知融合模块（Frequency-Aware Fusion module, FAFM）。在 RT-DETR 的原始设计中，其高效混合编码器在自顶向下和自底向上的融合路径中使用了相同的融合块。在自底向上的上采样路径中，简单的上采样操作容易导致特征边界模糊和空间信息丢失。为解决这一问题，我们利用 FAFM 模块构建了全新的双融合特征编码器，能够更智能、更有效地融合不同尺度的特征，特别是增强对高频细节的保留与利用，以提升对航拍图像中小目标的检测能力。

3.3.2 双重注意力处理块

无人机航拍图像中目标尺度变化、小目标密集分布和背景复杂等挑战，对 ResNet 中标准的残差块构成了显著考验，导致其在此类场景下性能下降。为解决这一问题，我们提出了双重注意力处理块（Bipartite Attentive Processing Block, BAPB），该模块是一种全新的架构，通过将相互依赖的注意力机制与并行分支处理相结合，来优化特征提取过程。BAPB 的结构如图3-2所示。

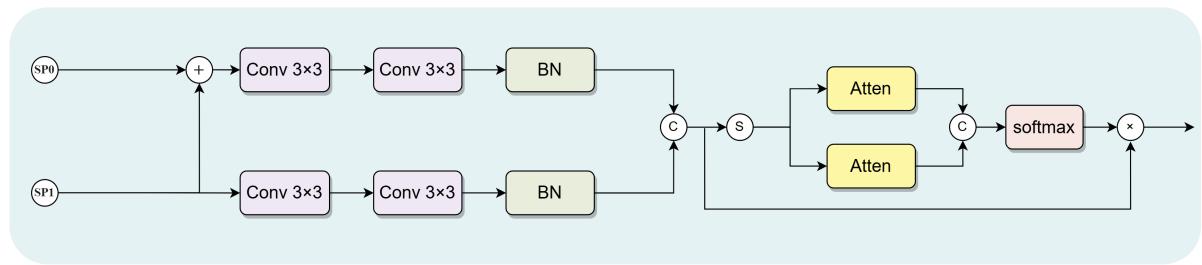


图 3-3 双重注意力模块示意图

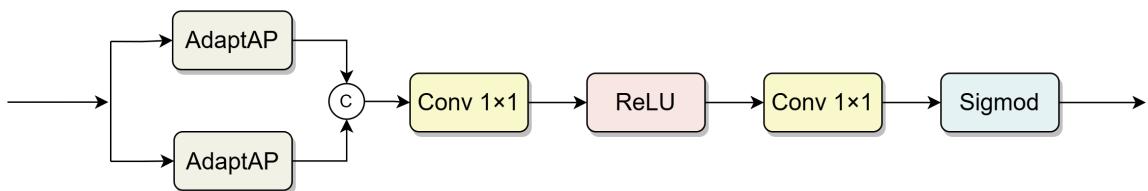


图 3-4 注意力分支示意图

输入特征图首先沿通道维度被均匀分割为两部分，分别记为 SP_0 和 SP_1 ，并进入两个并行处理分支。在 SP_0 分支中， SP_0 首先与 SP_1 进行逐元素相加，其和作为该分支的后续输入。这一操作在初始阶段即引入了跨通道的交互。两个分支随后分别独立地通过两个连续的 3×3 卷积层和一个批归一化层进行处理。通道分离策略将特征图解耦为不同的子成分，使每个分支能够专注于其输入特征的不同特性。最后，两个分支的处理输出沿通道维度进行拼接，生成优化后的融合特征 F_{fused} 。

为自适应地增强 F_{fused} 中的信息，我们对其进行进一步的处理。首先，将 F_{fused} 再次沿通道维度分割为 F_0 和 F_1 ，并分别输入两个专用的注意力分支。每个分支内部是一个注意力块，其结构如图3-4所示，具体而言， F_0 和 F_1 同时输入两个并行的自适应平均池化层，分别产生尺寸为 1 和 2 的池化输出，池化结果经拼接后，通过一个由 1×1 卷积层、ReLU 激活函数、第二个 1×1 卷积层和 Sigmoid 激活函数组成的序列，生成中间注意力权重，该过程可表述为：

$$\text{Atten}(F_i) = \sigma(W_2 \delta(W_1 \cdot \text{Concat}[\text{Pool}_1(F_i), \text{Pool}_2(F_i)])) \quad (3-1)$$

其中， W_1 和 W_2 表示 1×1 卷积， δ 为 ReLU 函数， σ 为 Sigmoid 函数。随后，两个分支产生的中间注意力权重被拼接，并通过 SoftMax 函数进行归一化，以计算最终的注意力权重：

$$W_{atten} = \text{SoftMax}(\text{Concat}[\text{Atten}(F_0), \text{Atten}(F_1)]) \quad (3-2)$$

最终，权重 W_{atten} 与特征图 F_{fused} 相乘，通过学习到的权重对不同通道的重要性进行调整。BAP 模块通过并行的卷积处理与注意力机制，有效地融合了空间与上下文信

息，使模型能够聚焦于特征图中信息最丰富的区域。与标准残差块相比，该设计能进行更有效的特征学习，最终提升了模型在复杂航拍场景下的差异化特征表征能力。

3.3.3 双融合特征编码器

原始 RT-DETR 中的高效混合编码器通过仅使用单层 Transformer 编码器处理 S5 特征图，优化了基于注意力的特征融合，从而兼顾了计算效率与精度，而其基于 CNN 的跨尺度特征融合模块则遵循了传统的多尺度特征融合模式，通过卷积层整合特征，其中粗粒度特征仅通过最近邻上采样后便与高分辨率特征简单拼接。这种方法存在两个显著影响预测精度的问题：类内不一致性与边界偏移，此外，简单的插值操作也常常导致特征被过度平滑。

双融合特征编码器的结构如图3-2所示。自底向上和自顶向下路径中的融合块分布记为 F_{ij}^{bu} 和 F_{ij}^{td} ，其中 i 和 j 代表来自第 i 和第 j 个特征图的输入。在自底向上路径中，我们引入了一个额外的融合块 $F_{2,3}^{bu}$ ，将低层特征 S2 与其前一层的特征进行融合。低层特征 S2 包含更丰富的空间信息和小目标特征，这些细节能够显著提升复杂场景中小目标的定位与识别精度。将 S2 集成到特征融合过程中，可以极大改善检测性能，尤其是在小目标密集聚集的场景中。为进一步提升模型的小目标检测能力，我们直接将 $F_{2,3}^{bu}$ 的输出作为解码器的输入之一。同时， $F_{2,3}^{bu}$ 的结果也被送入自顶向下路径进行下采样，并与自底向上路径各层的输出进行融合。在自顶向下路径中，我们移除了 $F_{4,5}^{td}$ 模块。这是因为经过深层卷积处理后，小目标特征极易被背景及其他目标影响，导致网络难以捕获差异化信息。最终，编码器的输出由原来的 $F_{3,4}^{bu}, F_{3,4}^{td}, F_{4,5}^{td}$ 变为 $F_{2,3}^{bu}, F_{2,3}^{td}, F_{3,4}^{td}$ 。此项调整在增加模型计算负载的同时也提升了精度，但实验表明这一权衡是值得的：具体而言，该修改增加了 9.5 GFLOPs 的计算量，带来了 AP2.6% 的提升。相比之下，若直接将主干网络替换为 ResNet-50 而保持网络其余部分不变，计算量将增加 76.0 GFLOPs，但 AP 仅能提升 1.7%。

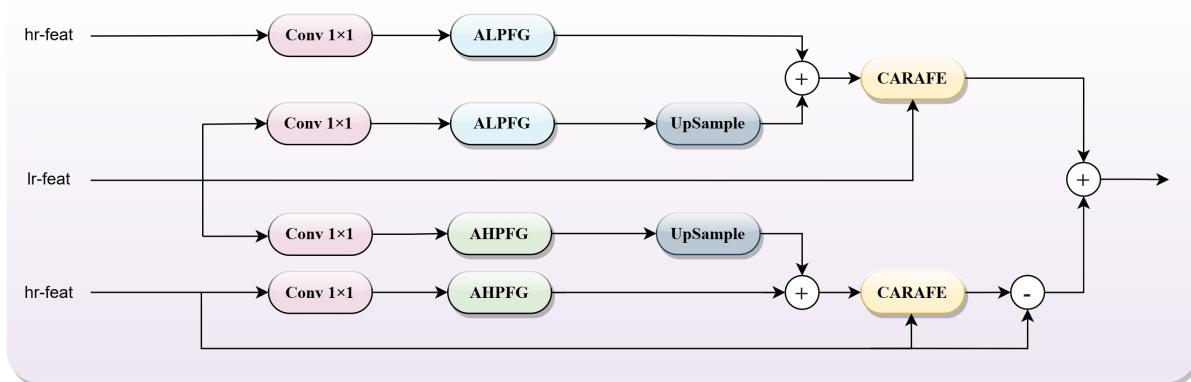


图 3-5 频率感知融合模块示意图

此外，我们将自底向上路径中的融合块 $F_{3,4}^{bu}$ 替换为频率感知融合模块。此修改增强了网络不同层级特征的整合能力，从而全面提升模型性能。FAFM 的结构如图3-5所

示，其设计灵感来源于 Freqfusion^[107]，采用了一个自适应低通滤波器生成器（Adaptive Low-Pass Filter Generator, ALPFG）和一个自适应高通滤波器生成器（Adaptive High-Pass Filter Generator, AHPFG）构成。ALPFG 允许特征的低频成分通过并抑制高频成分，而 AHPFG 则用于从特征图中提取高频成分，从而克服标准卷积层的局限性。FAFM 的输入来自两个不同尺寸的特征图，输出为一个与高分辨率输入特征图尺寸相同的融合特征。在融合前，两个输入特征图均经过一个 1x1 卷积层进行通道压缩，这有助于减少特征图的通道数并促进更有效的特征集成。我们采用 CARAFE^[108]作为低通滤波器，它能有效适应输入特征图的空间特性，在滤除噪声的同时增强模型保持重要空间信息的能力。高频成分则通过从压缩后的特征图中减去低频部分获得。FAFM 的引入显著增强了网络内多层级特征的融合质量。然而，增加更多模块并不必然带来性能提升。通过实验，我们发现将 $F_{3,4}^{bu}$ 替换为 FAFM 是最优选择，具体结果将在实验部分讨论。

3.3.4 损失函数

传统的 IoU 损失在某些情况下无法为网络优化提供有效梯度，例如当预测框与真实框不重叠，或一个框完全包含另一个框时。这两种情况在小目标检测中尤为常见。虽然 CIoU 和 DIoU 可以部分缓解此问题，但它们本质上仍基于 IoU 度量，对小目标的位置偏差极为敏感。具体而言，对于小目标，微小的位置偏移就会导致 IoU 值显著下降，而对于正常尺寸的目标，相同的偏移所引起的 IoU 变化则相对平缓。为克服这一问题，我们采用 RNWD-CIoU 来计算边界框损失，该损失由倒数归一化 Wasserstein 距离^[109]与标准的 CIoU 共同构成。我们将边界框建模为二维高斯分布 $N(\mu, \sigma)$ ：

$$\mu = \begin{bmatrix} x \\ y \end{bmatrix}, \sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \quad (3-3)$$

其中， (x, y) 为边界框中心坐标， w 和 h 分别为边界框的宽度和高度。两个边界框之间的相似性可以转化为对应高斯分布之间的距离。两个高斯分布之间的 Wasserstein 距离平方计算如下：

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left[x_a, y_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[x_b, y_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right\|_2^2 \quad (3-4)$$

随后，我们利用倒数形式的归一化将 Wasserstein 距离映射到 0 至 1 的区间，作为边界框的相似性度量。与指数归一化形式相比，该归一化方法在计算效率上更具优势，且衰减更慢，有助于缓解梯度消失问题。

$$\text{RNWD}(\mathcal{N}_a, \mathcal{N}_b) = \frac{1}{1 + \sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}} \quad (3-5)$$

相比于 IoU, RNWD 在检测小目标时具有多项优势：尺度不变性、对位置偏移的平滑性，以及能够有效度量非重叠或相互包含的边界框之间的相似性。最终的边界框损失定义为：

$$\text{Loss}_{\text{bbox}} = \lambda * \text{Loss}_{\text{ciou}} + (1 - \lambda) * \text{Loss}_{\text{RNWD}} \quad (3-6)$$

其中， λ 为权重系数，用于控制两部分损失的贡献比例。在处理密集小目标的场景时，适当降低 λ 值可以增强 $\text{Loss}_{\text{RNWD}}$ 对整体损失的贡献，从而提升模型对微小目标的定位鲁棒性。

3.4 实验结果与分析

3.4.1 数据集与评价指标

我们在两个无人机航拍图像基准数据集 VisDrone^[87]与 UAVDT^[94]，以及一个面向微小目标检测的遥感数据集 AI-TOD^[95]上进行了充分的实验。VisDrone-2019-DET 的训练集包含 6471 张图像，验证集包含 548 张图像，所有图像均在不同高度和地点由无人机拍摄。每张图像使用边界框标注了十个预定类别：行人、人、汽车、厢式货车、公交车、卡车、摩托车、自行车、带篷三轮车以及三轮车。UAVDT 训练集包含 23258 张图像，测试集包含 15069 张图像，该数据集涵盖了不同天气条件、飞行高度、拍摄角度和遮挡场景下的图像，包含汽车、公交车和卡车三个类别。AI-TOD 数据集由 28036 张尺寸为 800×800 的图像构成，划分为训练集（11214 张）、验证集（2804 张）和测试集（14018 张）。它包含八个目标类别：飞机、桥梁、储罐、船舶、游泳池、车辆、人和风车。

VisDrone 数据集涵盖了复杂城市场景下各种无人机拍摄的情况，而 UAVDT 数据集则主要关注交通和人群场景。与 VisDrone 和 UAVDT 中无人机捕获的航拍图像不同，AI-TOD 主要由光学遥感图像组成，这带来了拍摄角度、背景环境和目标尺度的显著差异，其目标平均尺寸仅为 12.8 像素，且 86% 的目标小于 16 像素，这些特性使得 AI-TOD 比典型的无人机数据集更具挑战性。这些数据集对于无人机图像分析与目标检测领域的研究与开发具有重要价值，尤其是在涉及小目标和复杂背景的场景中。

我们采用广泛使用的 COCO 风格目标检测评估指标 AP 和 AP_{50} 来衡量精度， AP 是通过对 10 个 IoU 阈值（从 0.5 到 0.95，步长为 0.05）上的精度取平均值计算得出。 AP_{50} 则是在 IoU 阈值为 0.5 时计算的平均精度。此外，为全面评估模型，我们还采用 GFLOPs、参数量和 FPS 等指标来衡量模型的复杂度。GFLOPs 是基于 640×640 的输入分辨率计算的。

3.4.2 实现细节

实验在如下环境中进行：CUDA 11.7，NVIDIA GeForce RTX 4090 显卡，Python 3.9 以及 PyTorch 2.7.0。与传统的 CNN 模型相比，类 DETR 模型通常需要更长的训练时间，且收敛速度较慢。在单张 RTX 4090 GPU 上以批次大小为 16 进行训练时，可能会出现

表 3-1 Visdrone 数据集结果对比

模型	出处	GFLOPs	Params	<i>AP</i>	<i>AP</i> ₅₀
基于 CNN 的通用检测器					
YOLOv8-M	—	78.9	25.9	24.6	40.7
YOLOv9-M ^[110]	2024 ECCV	76.8	20.1	25.2	42.0
YOLOv10-M ^[111]	2024 NeurIPS	59.1	15.4	24.5	40.5
YOLOv11-M ^[86]	—	67.7	20.1	25.9	43.1
RetinaNet ^[112]	2017 ICCV	88.5	35.6	21.8	39.3
FSAF ^[113]	2019 CVPR	246.7	—	26.3	50.3
基于 Transformer 的通用检测器					
DETR ^[92]	2020 ECCV	187.0	60.0	24.1	40.1
Deformable DETR ^[100]	2020 ICLR	173.0	40.0	27.1	42.2
Sparse DETR ^[103]	2022 ICLR	121.0	40.9	27.3	42.5
RT-DETR ^[104]	2024 CVPR	136.0	42.0	28.4	47.0
UAV 图像专用检测器					
HRDNet ^[114]	2021 ICME	—	62.4	28.1	49.2
QueryDet ^[19]	2022 CVPR	212.0	36.2	28.3	48.1
CEASC ^[98]	2023 CVPR	150.1	—	28.7	50.7
ClusDet ^[50]	2019 ICCV	207.0	30.2	26.7	50.6
NWD-RKA ^[115]	2022 ISPRS	246.0	—	27.4	46.2
UAV-DETR ^[105]	2024 arXiv	170.0	42.0	31.5	51.1
DMNet ^[97]	2020 CVPRW	224.4	—	28.2	47.6
BAP-DETR-S	—	140.2	25.0	33.2	51.6
BAP-DETR-M	—	179.7	34.8	35.3	53.5

内存溢出错误，导致训练意外终止。若从断点恢复训练，最终模型的 *AP* 性能会出现显著下降。为确保训练过程的完整性并避免因内存溢出而中断，可采用的方案包括：使用多 GPU 并行训练，或在单 GPU 上减小批次大小。实验发现，使用双 GPU 训练所得的 *AP* 通常比单 GPU 训练低约 0.2，这一下降可能归因于数据分布不均、梯度更新同步问题以及批量归一化效果减弱等因素。因此，我们最终选择了单 GPU 训练策略，设定批次大小为 8，训练轮数为 150。模型优化采用 AdamW 优化器，动量设置为 0.9，权重衰减为 0.0001。初始学习率和最终学习率分别为 0.0001 和 0.001。在训练过程中，采用了包括 HSV 色彩调整、平移及尺度变换在内的数据增强技术。

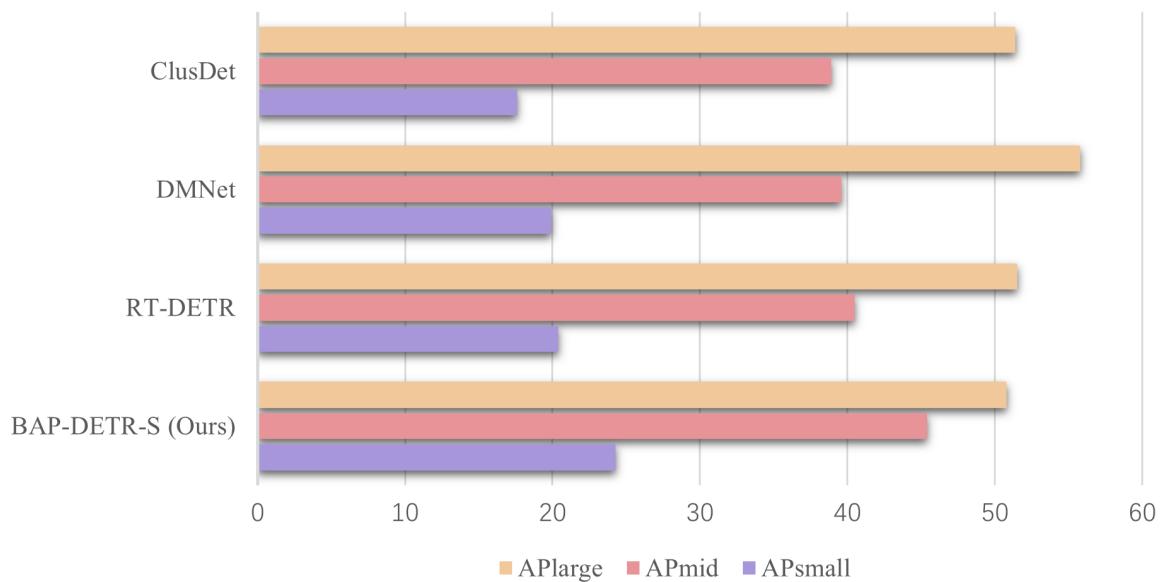


图 3-6 各尺度精度对比

3.4.3 与 SOTA 的对比实验

在本节中,我们在三个公开数据集(VisDrone、UAVDT 和 AI-TOD)上,将 BAP-DETR 与当前最先进的检测器进行对比分析,这些对比方法包括基于 CNN 和基于 Transformer 的通用检测器,以及专门为无人机图像设计的检测器。为展示模型的可扩展性以及在效率与精度间的权衡,我们提出了两个变体: BAP-DETR-S (使用带有 BAPB 的 ResNet-18 主干网络) 和 BAP-DETR-M (使用带有 BAPB 的 ResNet-34 主干网络)。所有对比方法的性能指标均源自其原始出版物。

如表3-1所示, BAP-DETR-S 在所有对比的无人机图像检测器中实现了最低的计算复杂度,同时相比其基线模型,精度仍有 4.8% AP 的显著提升。如图3-6所示,与其他最先进的检测器相比, BAP-DETR-S 在小目标和中等目标检测上展现了绝对优势。重要的是,我们计算量为 179.7 GFLOPs 的 BAP-DETR-M 变体,其计算负载与 CEASC、QueryDet 等近期先进方法相当甚至略低,但却提供了明显更优的检测性能。此外,与采用计算密集的聚焦检测(如 ClusDet)的检测器相比, BAP-DETR-M 在显著降低计算需求的同时,实现了更高的 AP 指标(例如 8.6% AP 的提升)。尽管我们的双融合编码器引入了高分辨率特征图,但由此带来的计算负载增加,相对于复杂度相似或更高的同类方法,换取了可观的精度增益。

为进一步证明 BAP-DETR 的泛化能力,我们还在用于微小目标检测的遥感数据集 AI-TOD 上进行测试,结果如表3-3所示。BAP-DETR-S 在 AI-TOD 上的 AP_{50} 达到 52.7%,优于基线 RT-DETR。而 BAP-DETR-M 则以 55.8% 的 AP_{50} 在所有对比方法中取得最高值。值得注意的是,尽管 NWD-RKA 在 AI-TOD 上表现出色,但其在 VisDrone 上的 AP_{50}

表 3-2 UAVDT 数据集结果对比

模型	出处	GFLOPs	Params	AP	AP_{50}	FPS
RT-DETR ^[104]	2024 CVPR	136.0	42.0	16.3	29.1	85.4
ClusDet ^[50]	2019 ICCV	207.0	30.2	13.7	26.5	16.5
DMNet ^[97]	2020 CVPRW	224.4	—	14.7	24.6	15.4
CEASC ^[98]	2023 CVPR	150.1	—	17.1	30.9	64.6
BAP-DETR-S	—	140.2	25.0	19.9	31.4	69.4
BAP-DETR-M	—	179.7	34.8	22.1	33.1	45.9

表 3-3 AI-TOD 数据集结果对比

模型	出处	GFLOPs	Params	AP	AP_{50}	FPS
RT-DETR ^[104]	2024 CVPR	136.0	42.0	18.8	47.4	85.4
HANet ^[116]	2023 TCSVT	—	26.4	22.1	53.7	—
QueryDet ^[19]	2022 CVPR	212.0	36.2	12.2	29.3	23.6
NWD-RKA ^[115]	2022 ISPRS	246.0	—	23.4	53.5	20.1
DETR ^[92]	2020 ECCV	187.0	60.0	18.4	41.4	27.8
BAP-DETR-S	—	140.2	25.0	26.7	52.7	69.4
BAP-DETR-M	—	179.7	34.8	27.5	55.8	45.9

仅为 46.2%，远低于 BAP-DETR-S。这些结果表明了我们的方法在不同航空成像场景下的泛化能力。

表3-2和表3-3还对比了 BAP-DETR 与其他方法的推理速度。BAP-DETR-S 达到了 69.4 FPS，比基线 RT-DETR 略慢，BAP-DETR-M 以一定的速度换取了更高的精度，但仍保持 45.9 FPS 的实时性能，两个模型在速度上均优于其他专用检测器。总体而言，BAP-DETR 在精度和速度之间取得了良好的平衡，使其适用于实时应用场景。

3.4.4 消融实验

双重注意力处理块(BAPB)、双融合编码器(Dual-fusion encoder)以及 RNWD-CIoU 损失函数是 BAP-DETR 的三个核心组件。为验证各模块的有效性，我们在 VisDrone 数据集上进行了系统的消融实验。由于 BAP-DETR-S 采用 ResNet-18 作为主干网络，为公平对比，我们将基线模型 RT-DETR 的主干网络也替换为 ResNet-18。

1) 核心组件消融实验

表3-4详细展示了各个组件对最终性能的贡献。双重注意力处理块和双融合编码器均能显著提升模型的 AP 值，同时也带来了计算负载的增加。这两者共同贡献了 6.3%

的 AP 提升。相比之下，RNWD-CIoU 损失函数在不增加模型参数量和计算复杂度的情况下，也为性能带来了有限的改进。与基线模型相比，BAP-DETR-S 模型实现了 6.5% 的 AP 提升，这证明了每个组件都对整体性能产生了积极贡献。

BAP-DETR-S 确实比 ResNet-18 基线模型增加了计算负载，但从量化收益看，这些增加的计算量能带来可观的精度收益：BAP-DETR-S 以增加 80.2 GFLOPs 的代价，换来了 6.5% 的 AP 增益。若采用其他性能提升策略作为对比，例如将主干网络直接从 ResNet-18 替换为 ResNet-50 而保持其他部分不变，计算负载将增加 76.0 GFLOPs，但 AP 仅能提升 1.7%。这充分证明，我们的结构修改在单位计算成本上带来了更优越的性能回报。

表 3-4 核心组件消融实验结果

BAPB	Dual-fusion encoder	RNWD	AP_{50}	AP	GFLOPs	Params
			44.6	26.7	60.0	20.0
✓			49.1	30.4	130.7	21.1
✓	✓		51.1	33.0	140.2	25.0
✓	✓	✓	51.6	33.2	140.2	25.0

2) 频率感知融合模块消融实验

为验证频率感知融合模块的有效性并确定其在编码器中的最佳配置，我们进行了一系列消融实验。原始 RT-DETR 编码器包含四个融合块（自底向上与自顶向下路径各两个）。在引入低层特征 S2 后，自底向上路径新增了一个融合块 $F_{2,3}^{bu}$ ，使得可替换的融合块变为三个： $F_{2,3}^{bu}$, $F_{3,4}^{bu}$ 和 $F_{4,5}^{bu}$ 。我们设计了四种实验配置：分别用 FAFM 单独替换上述三个融合块，以及同时替换所有三个融合块。实验结果如表3-5所示。

表 3-5 频率感知融合模块消融实验结果

$\mathcal{F}_{2,3}^{bu}$	$\mathcal{F}_{3,4}^{bu}$	$\mathcal{F}_{4,5}^{bu}$	AP_{50}	AP	GFLOPs	Params
			44.6	26.7	60.0	20.0
✓			46.8	29.2	60.9	20.3
	✓		47.6	30.4	61.0	20.2
		✓	46.6	29.1	60.8	20.2
✓	✓	✓	47.2	30.1	65.8	23.7

实验结果表明，在各种配置中引入 FAFM 均能提升 AP 与 AP_{50} ，其中，在中层特征 (S3-S4) 对应的 $F_{3,4}^{bu}$ 处应用 FAFM 效果最为显著，取得了 30.4% 的 AP 和 47.6% 的 AP_{50} 。单独替换三个融合块所产生的计算负载相对一致。然而，同时替换所有三个融合块虽进一步提高了计算复杂度，但性能增益并未成比例增加。这说明，虽然 FAFM 能

增强特征融合，但累积添加多个模块可能导致计算资源的无谓消耗，而无法带来相应的精度提升。

这一反直觉结果的原因可能在于，添加多个 FAFM 模块可能会破坏网络固有的特征层次。FAFM 的设计原理是：在高层特征图中衰减目标内部的高频分量以减少类内不一致性，同时在低层特征图中增强高频分量以维持清晰的边界。在 S2 这类低层特征中应用 FAFM，不仅会放大有用的高频细节，也会放大噪声和背景干扰。当与其他 FAFM 模块的放大效应叠加时，会导致噪声累积，从而对模型产生干扰。而在高层特征中，连续应用 FAFM 会导致高频成分被过度衰减，产生过度平滑的特征，从而丢失精确定位所需的结构细节。中层特征 S3 和 S4 恰好包含了空间细节与语义信息的平衡，因此对频率感知增强的响应最为积极和有效。本实验证实了我们的设计选择：仅在关键的 $F_{3,4}^{bu}$ 位置引入单一 FAFM 模块，是实现性能与效率最优平衡的最佳策略。

3) RNWD 消融实验

将 RNWD 整合到现有的 CIoU 损失函数中，是网络的另一项改进。消融实验结果 3-6 展示了 RNWD 权重 λ 对模型性能的影响。结果表明，引入 RNWD 能带来性能指标的轻微提升：当 λ 值为 0.3 和 0.7 时， AP_{50} 达到 44.9%，当 $\lambda = 0.7$ 时， AP 达到 26.8%。与 BAPB 和 FAFM 带来的显著改进相比，RNWD 的提升相对有限，但其优势在于不引入任何额外的模型参数或计算负载。

表 3-6 RNWD 权重系数消融实验结果

Loss function	AP_{50}	AP
CIOU	44.6	26.7
CIOU+RNWD($\lambda=0.3$)	44.9	26.6
CIOU+RNWD($\lambda=0.5$)	44.8	26.7
CIOU+RNWD($\lambda=0.7$)	44.9	26.8
CIOU+RNWD($\lambda=0.9$)	44.8	26.7

表 3-7 RNWD 消融实验结果

Loss function	AP_{50}	AP
CIOU	44.6	26.7
+GWD	43.8	26.2
+GWD exp norm	44.9	26.6
+KL Divergence	44.7	26.8
+RNWD	44.9	26.8

为更进一步验证 RNWD 的有效性，我们将其与 IoU 系列之外的其他先进损失函数进行了比较，包括 GWD、指数归一化 GWD 以及 KL 散度损失。如表 3-7 所示，GWD

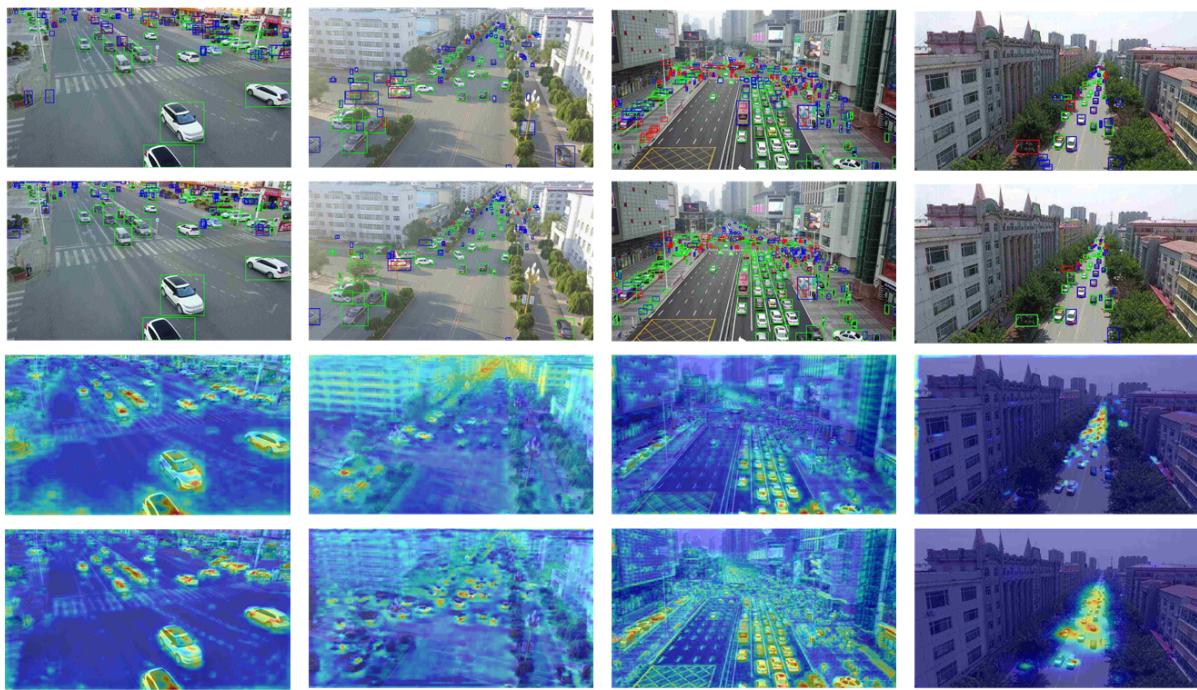


图 3-7 第一行呈现了基线方法的检测结果，第二行则展示了 BAP-DETR 的检测结果。其中，绿色、蓝色和红色边界框分别表示正确检测、误检与漏检。第三行和第四行分别为基线方法与 BAP-DETR 的热力图可视化结果。

因其快速增长的梯度趋势而对大误差过于敏感，导致其性能不佳。指数归一化 GWD 在 AP_{50} 指标上表现出竞争力，但在 AP 上略低于 RNWD，与此同时，KL 散度损失取得了与 RNWD 相近的 AP ，却在 AP_{50} 上表现不足。RNWD 在两个指标上均提供了最均衡、稳定的性能，验证了其设计的有效性。

3.4.5 可视化结果

为直观展示 BAP-DETR 的有效性，图3-7中呈现了若干具有代表性的检测样例。图中第一、二行分别展示了基线模型与 BAP-DETR 模型的检测结果，第三、四行则分别展示了两者对应的、基于边界框预测进行反向传播所生成的热力图。从图中可以直观观察到，基线模型中存在的多个误检与漏检，均被 BAP-DETR 有效消除。与基线模型相比，模型在热力图上表现出对密集小目标及其周围环境更强的聚焦能力。这表明，BAP-DETR 不仅提升了分类性能，还增强了对上下文线索的敏感度，使得模型对关键区域的关注更加集中和精准，有效抑制了背景噪声的干扰。通过可视化对比，BAP-DETR 在复杂航拍场景下对小目标，尤其是密集分布的小目标，展现出了优于基线的、更鲁棒和精确的感知能力。

3.5 本章小结

本章针对机载光电系统目标检测中的核心挑战，提出了一种基于 RT-DETR 的创新型架构 BAP-DETR，旨在优化特征提取与融合过程。通过引入双重注意力处理块，BAP-DETR 能够同时捕获对小目标检测至关重要的细粒度细节与全局上下文信息。结合频率感知融合模块的双融合编码器，在有效整合多尺度特征的同时，保留了高分辨率的空间信息。所提出的 RNWD-CIoU 损失函数解决了传统 IoU 度量在密集小目标场景下的固有局限。BAP-DETR 通过两个变体（侧重精度的 BAP-DETR-M 与侧重效率的 BAP-DETR-S）实现了灵活的速度精度调节。两个变体均在精度上超越了现有方法，其中侧重效率的变体相比专用的无人机图像检测器，计算负载显著降低。在三个公开航拍数据集上的广泛评估证实，BAP-DETR 实现了业界领先的效率精度平衡。

4 基于多特征聚焦与跨阶段 Transformer 的红外小目标检测网络

在机载光电系统中，红外成像传感器是实现全天时、全天候环境感知的核心组件。机载平台下的红外小目标检测面临三重严峻挑战：首先，红外图像本身存在分辨率低、缺乏纹理细节、对比度差的固有限制，其次，从空中俯视的复杂动态背景对小目标检测造成干扰，导致算法虚警率高，最后，无人机载荷对功耗、重量和计算资源的严格限制，要求算法必须在极高的实时性与足够的检测精度之间取得平衡。当前多数基于像素级分割的检测网络，虽在静态基准测试中表现良好，但其庞大的计算量和对高分辨率特征图的需求，使其难以直接部署于机载嵌入式平台进行实时处理。为应对这些挑战，本章提出一种专为无人机平台设计的高效红外小目标检测网络 MFF-DCNet。该网络通过协同优化特征提取与融合机制，实现精度与速度的兼顾，核心创新在于：基于深度分离卷积的跨阶段 Transformer（Depth-wise Cross-stage Transformer，DCFormer）和多特征聚焦（Multi-Feature Focus，MFF）颈部结构。DCFormer 模块通过深度可分离卷积与跨阶段特征融合的结合，在显著降低计算开销的同时，有效增强了主干网络对多尺度上下文信息的建模能力。优化特征提取过程，提升了模型在复杂场景下对小目标的特征判别能力，并且为在边缘设备上的实时部署提供了可能。重新设计的 MFF 颈部结构通过构建新颖的特征聚合机制，增强了跨尺度特征的整合能力，使模型能够更好地融合不同层级的语义信息和空间细节。这种设计显著提升了模型对多尺度目标的检测性能，特别是在复杂背景下对微小红外目标的精准识别能力。

4.1 引言

红外探测系统凭借其被动成像，抗干扰能力强及全天时工作的独特优势，显著提升了无人机的自主感知能力。然而，红外图像普遍存在空间分辨率低、缺乏色彩与丰富纹理细节的局限^[117]，导致为高分辨率可见光图像设计的通用深度网络难以直接迁移并提取有效的判别性特征，在典型的无人机应用场景中，地面目标在红外图像中仅占据极少的像素，表现为信噪比极低的弱小目标。同时，复杂的地物背景会引入大量与目标热辐射特征相似的杂波，使得在低信噪比条件下实现精准的“目标-背景”分类变得困难。传统的神经网络，尤其是计算密集的 Transformer 架构，其高复杂度在算力受限的边缘设备上难以实现实时性能，构成了实际部署的显著瓶颈。因此，在嵌入式设备上实现高精度、高实时性的红外小目标检测是一个亟待解决且具有重要实际价值的研究课题。

针对小目标检测问题，研究者们提出了多种解决方案，如数据增强、背景建模以及聚焦检测等 Kou2023survey, Liu2024YOLC, Yang2019ClusteredOD.^[118]。然而，当这些方法直接应用于红外图像时，其性能往往会因红外成像的物理特性而显著衰减。数据增强策略在可见光图像中能有效增加小目标样本，但在红外图像中可能破坏目标与背景之间的热对比

度关系，导致性能提升不稳定且泛化能力有限。背景建模利用目标周围区域提供辅助信息，然而在杂波丰富的红外背景中，过度依赖上下文极易引入干扰噪声，反而模糊了目标本身的特征。聚焦检测类的方法计算成本高，且将大量计算浪费在对广阔背景区域的处理上。

现有的红外小目标检测方法可分为两大技术路线：基于分割的方法与基于检测的方法。基于分割的方法在天空、海面等纯净背景下表现良好^[119]。然而，当应用于具有复杂背景的无人机航拍图像时，这类方法的误报率显著升高。红外传感器的远距离成像导致小目标信噪比较低，且目标越小，其像素表现越模糊、不确定性越高，在复杂环境中获取精确的像素级标注也变得更加困难。此外，分割网络通常采用的编码器解码器结构会引入巨大的计算开销，无法在资源受限的嵌入式平台上达到实时处理的要求。基于检测的方法专注于直接识别与定位小目标。其中两阶段模型虽能达到较高精度，但常受计算复杂度高的困扰^[120]，而单阶段模型则因其高效性在嵌入式系统中日益流行。为提升上下文建模能力，Transformer 架构^[81,121]被引入小目标检测网络，以解决捕获长程依赖关系的挑战。其自注意力机制通过计算特征的相关性，使网络能够聚焦于目标区域并捕获更广泛的上下文信息。后续工作中对编码器做出了进一步优化，例如在编码器中引入局部感知块的 Local Perception Swin Transformer (LPPSW)^[122]，以及融合了全局-局部特征交错模块的双网络结构 (Dual network structure with Interweaved Global-Local, DIAG)^[123]。这些方法均致力于优化特征提取，以应对复杂航拍图像中的小目标检测难题。虽然这些改进提升了检测精度，但通过复杂自注意力机制带来的性能提升，往往以计算开销的大幅增加为代价。因此，基于 Transformer 的方法在部署于红外小目标检测系统时，特别是在计算资源有限的边缘计算场景中，仍面临显著局限。

针对上述问题，本章提出 MFF-DCNet，一种基于 YOLOv11 的高效红外小目标检测网络。通过两项关键创新 MFF-DCNet 实现了业界领先的效率精度平衡：深度可分离跨阶段 Transformer 模块 (Depth-wise Cross-stage Former, DCFormer) 与多特征聚焦颈部结构 (Multi-Feature Focus, MFF)。DCFormer 通过深度可分离卷积对标准 Transformer 编码器进行改进，在降低计算复杂度的同时提升了特征提取能力。MFF 颈部结构重新定义了原有框架的颈部结构，特征聚合机制跨尺度选择并融合差异化特征，有效抑制了冗余信息。

本文的主要贡献总结如下：

- 重新设计了整个颈部结构，提出了多特征聚焦模块 MFF。这是一种新颖的特征聚合结构，能有效增强不同尺度间特征信息的融合，从而提升模型对多尺度目标（尤其是复杂环境中的红外小目标）的检测能力。
- 通过引入 DCFormer 模块增强了主干网络的特征提取能力。该先进的增强模块集成了深度可分离的空间特征提取与跨阶段特征融合，在降低计算成本的同时优化了多尺度上下文建模，提升了复杂场景下的红外小目标检测精度。
- 在 HIT-UAV^[124]和 DroneVehicle^[125]数据集上进行了充分的实验。MFF-DCNet 在

检测精度 (AP_{5095} 达到 57.4%，较同类先进方法提升 5.8%) 与处理效率 (帧率提升 10%) 上均取得显著进步。同时，该网络在 NVIDIA Jetson Orin NX 边缘计算模块上达到了 39.6 FPS 的稳定实时处理能力，验证了其满足实际机载任务严苛的实时性、可靠性与低功耗要求，具备直接的工程应用价值。

4.2 相关工作

4.2.1 面向嵌入式平台的红外小目标检测

传统基于建模的方法，如稀疏分解与背景建模，均建立在一个先验假设之上，即小目标可以从结构化背景中被有效分离。与基于深度学习的方法相比，这类方法在面对典型的城市无人机复杂运行环境时，性能严重下降。基于深度学习的红外小目标检测方法主要分为分割方法和检测方法两条技术路线。分割方法^[119]将该任务建模为一个正负样本极不平衡的二值语义分割问题，代表性方法可归类为超分辨率、多尺度表征、上下文信息和尺度感知训练等。近期工作 LRRNet^[126]试图将深度学习与传统的稀疏分解和背景建模相结合。然而，这些分割网络的推理速度通常很慢^[127]，即便在标准的消费级 GPU 上平均帧率也低于 10FPS，嵌入式平台的算力通常不足普通桌面级 GPU 的十分之一，难以支持高复杂度的分割网络实时运行。

嵌入式系统在计算能力、内存和能耗方面面临严格限制。尽管存在众多适用于消费级 GPU 的高性能算法，但它们往往难以直接部署于无人机等边缘设备上。在边缘设备上部署高性能红外检测算法的主要挑战，在于高计算需求与硬件约束之间的冲突^[128]，这推动着研究向轻量化和专用化模型发展。单阶段检测模型因其高效性在嵌入式系统中日益流行，例如 MobileNet^[129]、ShuffleNet^[130]和 GhostNet^[131]。MobileNet 通过使用深度可分离卷积减少了参数量，ShuffleNet 使用分组卷积将输入通道划分为更小的组，降低了计算复杂度和参数量，GhostNet 引入 Ghost 模块，通过先使用较少卷积核生成主要特征图，再生成额外的特征图，实现了参数量和计算量的大幅降低。为提高小目标检测精度，这些轻量级主干网络常与多尺度特征学习^[14,132]或超分辨率技术结合使用。例如，SuperYOLO^[133]采用对称紧凑的多模态融合技术整合多种数据模态（RGB 与红外），并融入超分辨率学习以获取高分辨率特征表示，YOLC^[134]利用针对聚类区域的局部尺度模块，但在处理稀疏分布的目标时效果欠佳。

YOLO 系列模型在速度与准确性之间取得了出色平衡，是嵌入式平台的理想选择 Xiong2024AdaptiveFeatureFusion。近期，基于 DETR^[92]框架的实时检测器（如 RT-DETR^[104]）被提出。然而，在没有对应领域成熟预训练模型的情况下，DETR 极难应用于新领域，因此目前应用最广泛的实时目标检测器仍是 YOLO 系列。本章提出的 MFF-DCNet 基于 YOLOv11 构建，以架构效率为核心设计原则，DCFormer 模块通过深度可分离设计减少参数，MFF 模块在不依赖昂贵计算成本的编解码器结构的前提下增强了特征判别力。与基于注意力的 Transformer 方法相比，MFF-DCNet 实现了更少的参数量和计算负载，使其更适用于资源受限的嵌入式设备。

4.2.2 多尺度特征学习方法

深度卷积神经网络会生成具有不同空间分辨率的层级化特征图。其中，低层特征蕴含更丰富的细节和定位信息，而高层特征包含更强的语义信息。对于红外小目标检测而言，随着网络深度的增加，小目标的特征表示在最终的特征图中会逐渐减弱。由于红外图像固有的特性（如分辨率低、纹理信息弱），这一问题在红外小目标检测中被进一步放大。为此，一种有效的解决方案是多尺度特征学习，通过整合不同深度的特征来增强对小目标的表征能力。

特征金字塔网络（Feature pyramid Network, FPN）^[14]通过构建自顶向下的路径与横向连接，在不同尺寸的特征图上进行预测，并根据目标尺寸将不同尺度的目标分配到对应的金字塔层级。这一多尺度预测的方式被广泛集成于各类目标检测网络中。受到 FPN 的启发，PANet^[24]通过引入双向路径来丰富特征层次，利用精确的定位信号强化深层特征，以更直观方式实现多尺度特征融合的优化。AugFPN^[135]则提出残差特征增强与一致性监督，以缩小不同尺度特征间的语义差距。双向特征金字塔网络（Bi-directional Feature Pyramid Network, BiFPN）引入了双向连接，允许信息在网络中同时进行自顶向下和自底向上的流动，确保了对小目标差异化特征的提取，并提升了网络效率。EfficientDet^[25]提出了加权的双向 FPN，通过引入可学习的权重来评估不同输入特征的重要性，从而实现融合过程中多尺度特征图均衡贡献。空间金字塔池化网络（Spatial Pyramid Pooling, SPP）^[132]通过引入空间金字塔池化层，实现从任意尺寸的图像中生成固定长度的特征表示，提升了图像分类与目标检测任务的精度与效率。

传统 CNN 在处理多尺度目标时面临挑战，视觉 Transformer 利用层级化的自注意力机制来构建跨尺度的特征表示，而不过度依赖空间降采样，这为红外小目标检测提供了另一种思路。多尺度 ViT（Multiscale ViT^[136]）将 CNN 结构中的多尺度特征提取思想与 Transformer 结合，实现多尺度特征提取。金字塔 ViT（Pyramid ViT^[137]）使用渐进缩小的金字塔结构来减少大尺寸特征图的计算量，可作为 CNN 主干网络的替代方案，在目标检测中展现出优越性能。CrossViT^[138]则采用双分支 Transformer 处理不同尺寸的图像块，生成多尺度图像特征，并通过交叉注意力机制进行特征交互学习。

尽管上述多尺度方法整合了不同层级的特征，但它们通常不加区分地融合所有层次的特征，这会带来计算成本的增加，并可能放大背景噪声，从而导致性能下降。这些局限性凸显了当前方法需要一种更精细、更具选择性的融合策略，以专注于跨尺度中最具信息量的特征。

4.3 基于多特征聚焦与跨阶段 Transformer 的检测网络

4.3.1 模型架构

本节将详细介绍基于多特征聚焦与跨阶段 Transformer 的目标检测网络 MFF-DCNet，其整体框架如图4-1所示。该网络专为红外航拍图像中的小目标检测任务设计，核心包

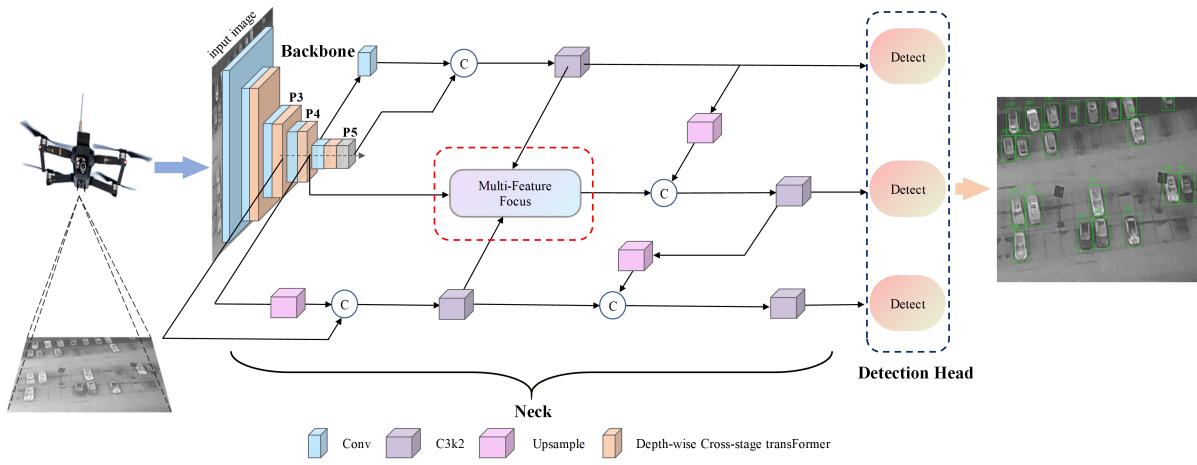


图 4-1 MFF-DCNet 网络架构示意图

含一个集成了新型 DCFormer 模块的增强型主干网络和一个设计有 MFF 模块的创新颈部结构。针对 YOLOv11 主干网络中 C3k2 模块存在的计算效率低下以及跨阶段特征融合能力不足的问题，本章设计了 DCFormer 模块作为其替代。该模块融合了 Transformer 的设计思想，利用深度可分离卷积作为轻量化的特征混合器，并结合跨阶段残差连接。这一设计通过空间解耦的操作显著降低计算成本，同时通过跨阶段特征重组与高效的长程依赖捕获，增强了模型的上下文建模能力。与此同时，为从根本上改善小目标检测性能，我们设计了一个包含多特征聚焦（Multi-Feature Focus, MFF）模块的全新颈部结构。当前应对多尺度挑战的主流方法是构建特征金字塔来整合不同尺度的特征。然而，对于小目标而言，其特征信息在经过连续的卷积层后会逐渐衰减，导致在最终特征图中保留的有效像素极少。因此，在检测头之前有策略地保留并增强高分辨率特征信息，无疑是提升小目标检测能力的关键。MFF 模块正是通过优化特征聚合过程来实现这一目标，它显著增强了网络对于微小目标的检测能力。

MFF 模块通过一个结构化的融合过程来聚合多尺度特征。具体而言，在主干网络完成特征提取后，我们获得特征图 P3、P4 和 P5。这些特征图的分辨率分别为输入图像尺寸的 1/8、1/16 和 1/32。其中，P3 层捕获空间细节，P4 层在空间细节与语义丰富性之间取得平衡，而 P5 层则专注于高层语义信息。P4 层因其居中的位置，成为双向特征传播的核心枢纽。我们为 P4 层创建了两个处理分支。一个分支对 P4 特征图进行 3×3 卷积以实现下采样，将其与经过 SPPF 模块和 C2PSA 模块处理的 P5 层特征融合，随后通过 DCFormer 模块处理，输出一个尺寸为 $20 \times 20 \times 512$ 的特征图。SPPF 模块是传统空间金字塔池化（SPP）的优化版本，在保留 SPP 核心功能的同时提升了计算效率，它通过级联结构串行执行多个相同核尺寸的最大池化操作，C2PSA 模块则整合了跨阶段局部连接（Cross Stage Partial）结构与注意力机制，以增强特征表征能力。P4 层的另一个分支通过最近邻插值进行上采样，与 P3 层特征合并，并经由 DCFormer 模块处理，生成一个 $80 \times 80 \times 128$ 的特征图。上述操作得到的特征，连同原始的 P4 层特征，一并送入 MFF 模块进行进一步的处理。

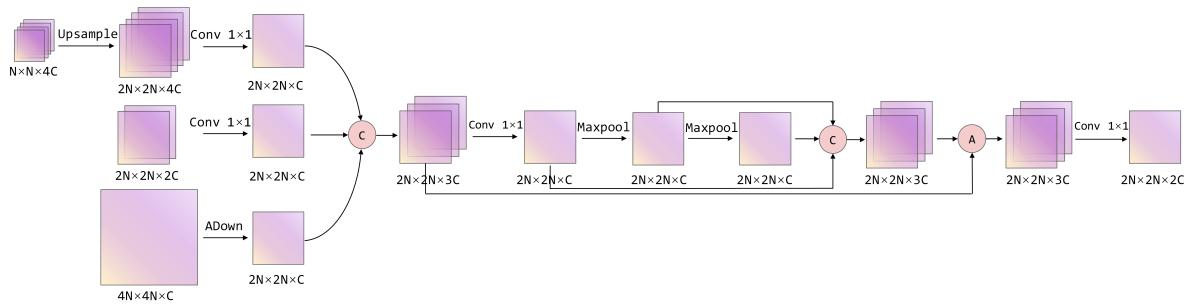


图 4-2 多特征聚焦模块示意图

4.3.2 多特征聚焦模块

多特征聚焦模块是特征聚焦颈部结构中的核心组件，负责对来自三个不同尺度的特征进行集成与融合。其详细结构如图4-2所示，完整的算法流程总结于算法1。具体而言，对于尺寸为 $20 \times 20 \times 512$ 的P5层输出特征，首先通过最近邻上采样将其空间尺度扩大一倍至 $40 \times 40 \times 512$ ，随后经过一个 1×1 卷积进行通道调整，得到尺寸为 $40 \times 40 \times C$ 的特征。对于尺寸为 $40 \times 40 \times 256$ 的P4层输出特征，同样使用 1×1 卷积调整通道数，得到尺寸为 $40 \times 40 \times C$ 的特征。对于尺寸为 $80 \times 80 \times 128$ 的P3层输出特征，则先通过一个Adown模块进行下采样，再进行通道调整，最终得到尺寸为 $40 \times 40 \times C$ 的特征，其中， C 被设定为P4层特征通道数的一半，在本文中取值为128。此步骤将所有输入特征在空间维度和通道维度上对齐，为后续的特征融合做好准备。

经过上述处理，所有三个尺度的特征均被对齐到统一的尺度 $40 \times 40 \times C$ 。随后，这三个尺度相同的特征图被拼接起来，形成初始的融合特征图，其尺寸为 $40 \times 40 \times 384$ 。该拼接后的特征接着通过一个 1×1 卷积进行处理，以将通道数调整至128，从而得到尺寸为 $40 \times 40 \times 128$ 的特征。此后，使用一个 5×5 的池化窗口、步长为1并进行边缘填充（以保持输出特征图尺寸）执行两次最大池化操作。将前述三个操作（包括一次 1×1 卷积和两次最大池化）产生的特征进行拼接。具体而言，这次拼接融合了：经过 1×1 卷积后的特征（ $40 \times 40 \times 128$ ）、第一次最大池化后的特征以及第二次最大池化后的特征，最终得到一个尺寸为 $40 \times 40 \times 384$ 的特征。优化后的特征首先与初始拼接阶段产生的融合特征进行相加，随后通过 1×1 卷积对合并后的特征进行进一步优化，同时保持特征尺度不变，输出最终的融合特征图，其尺寸为 $40 \times 40 \times 256$ 。

4.3.3 基于深度分离卷积的跨阶段 Transformer

YOLOv11的主干网络主要由连续的标准卷积和C3k2模块构成。C3k2模块通过跨阶段局部连接和可变核卷积进行特征整合，其瓶颈结构是负责通道变换和局部上下文聚合的核心。然而，传统的C3k2模块在红外小目标检测中存在关键局限：其串行的卷积结构不可避免地会削弱小目标的判别性特征表示，这一问题在目标缺乏显著纹理、且与复杂背景对比度低的红外图像中尤为突出。

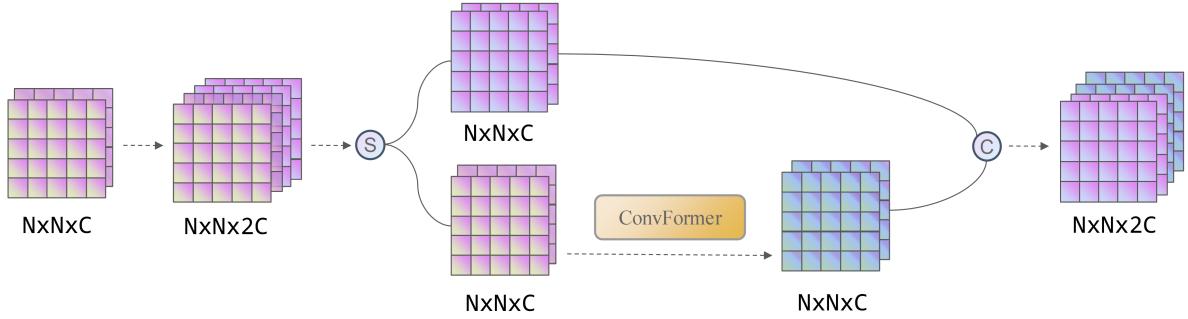


图 4-3 DCFormer 架构示意图

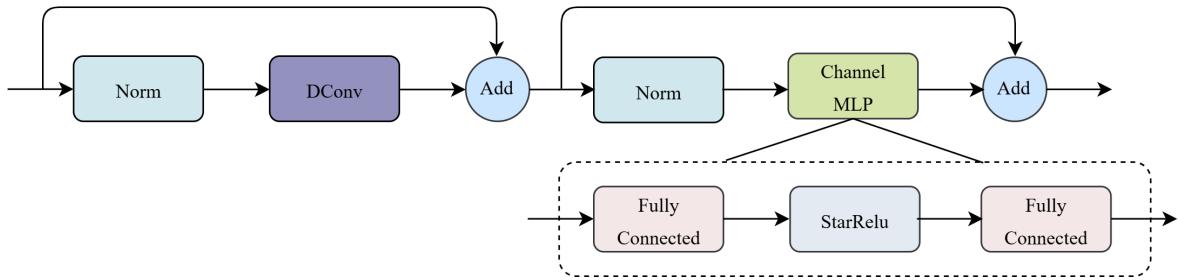


图 4-4 ConvFormer 架构示意图

DCFormer 模块通过重新设计特征传播路径，利用深度可分离卷积和受 Transformer 启发的跨阶段特征提取机制，有效应对了上述局限。DCFormer 在降低计算成本的同时，增强了对上下文信息的捕获和跨阶段特征提取能力。如图4-3所示，其处理流程始于一个卷积-批归一化-SiLU 激活模块（CBS），所得特征图沿通道维度被均匀分割为两部分，其中一部分通过 ConvFormer 块（其详细结构见图4-4），ConvFormer 采用深度可分离卷积作为高效的特征混合器，以替代标准的自注意力机制。另一部分则与 ConvFormer 的输出沿通道维度进行拼接，从而实现有效的跨阶段信息传播。最后，拼接后的特征再经由另一个 CBS 模块处理。DCFormer 的架构结合了卷积与 Transformer 思想，优化了特征提取过程，特别是增强对红外小目标的检测能力。值得注意的是，该设计相比原始的 C3k2 模块，所需的计算成本更低。

DCFormer 架构的一个核心组件是 ConvFormer，其设计灵感来源于 MetaFormer^[139]框架，作为主要的特征提取单元。与依赖计算密集型自注意力机制的传统 Transformer 不同，ConvFormer 通过深度可分离卷积^[140]实现了一种高效的特征混合策略。该架构用具有线性复杂度的深度可分离卷积替代了具有二次复杂度的注意力操作，使其能够在保持实时处理能力的同时，捕获细微的红外目标特征。ConvFormer 的详细结构如图4-4所示，它保持了标准 Transformer 编码器的结构，包含两个主要部分：一是用于空间信息提取的特征混合器（token mixer），二是带有残差连接的多层感知机（Multi-Layer Perceptron，MLP）。在 ConvFormer 中，特征混合器由深度可分离卷积实现，可以表示为：

$$X = DConv(Norm(X)) + X. \quad (4-1)$$

其中， X 表示输入特征图， $Norm(\cdot)$ 表示归一化操作， $DConv(\cdot)$ 表示深度可分离卷积。通过这种设计，ConvFormer 能够高效地捕获空间上下文信息，同时保持较低的计算复杂度。随后，通过一个带有 StarReLU 激活函数的双层 MLP 进行特征变换：

$$X = \sigma(Norm(X)W_1)W_2 + X. \quad (4-2)$$

其中， W_1 和 W_2 分别表示 MLP 的权重矩阵， $\sigma(\cdot)$ 表示 StarReLU 激活函数。

在 ConvFormer 中采用深度可分离卷积作为特征混合器，主要基于其在计算效率方面的显著优势，这对于实时红外检测系统至关重要。与标准卷积对所有输入通道执行卷积运算不同，深度可分离卷积将此过程分解为两步：深度卷积和逐点卷积。深度卷积独立处理每个输入通道，而逐点卷积则负责跨通道集成输出。这种分解在保持表征能力的同时，显著减少了参数量和计算成本。假设输入特征图尺寸为 $W_i \times H_i \times C_i$ ，卷积核尺寸为 $K_w \times K_h \times C_i$ ，输出特征图尺寸为 $W_o \times H_o \times C_o$ ，对于标准卷积，单个卷积核包含 $K_w \times K_h \times C_i$ 个参数和一个偏置项，共有 C_o 个卷积核，其参数量和计算量如下：

$$Params_{std_conv} = (K_w \times K_h \times C_i + 1) \times C_o. \quad (4-3)$$

$$FLOPs_{std_conv} = K_w \times K_h \times C_i \times W_o \times H_o \times C_o. \quad (4-4)$$

对于深度卷积，单个卷积核的维度为 $K_w \times K_h \times 1$ ，并带有一个偏置项。使用 C_i 个卷积核，输出特征图维度为 $W_o \times H_o \times C_i$ ，参数量和计算量如下：

$$Params_{depth_conv} = (K_w \times K_h \times 1 + 1) \times C_i. \quad (4-5)$$

$$FLOPs_{depth_conv} = K_w \times K_h \times W_o \times H_o \times C_i. \quad (4-6)$$

对于逐点卷积，单个卷积核的维度为 $1 \times 1 \times C_i$ ，并带有一个偏置项，使用 C_o 个卷积核，参数量和计算量如下：

$$Params_{point_conv} = (1 \times 1 \times C_i + 1) \times C_o. \quad (4-7)$$

$$FLOPs_{point_conv} = C_i \times W_o \times H_o \times C_o. \quad (4-8)$$

对于深度可分离卷积，总参数量和总计算量为：

$$Params_{Dconv} = (K_w \times K_h + 1) \times C_i + (C_i + 1) \times C_o. \quad (4-9)$$

$$FLOPs_{Dconv} = W_o \times H_o \times C_i \times (K_w \times K_h + C_o). \quad (4-10)$$

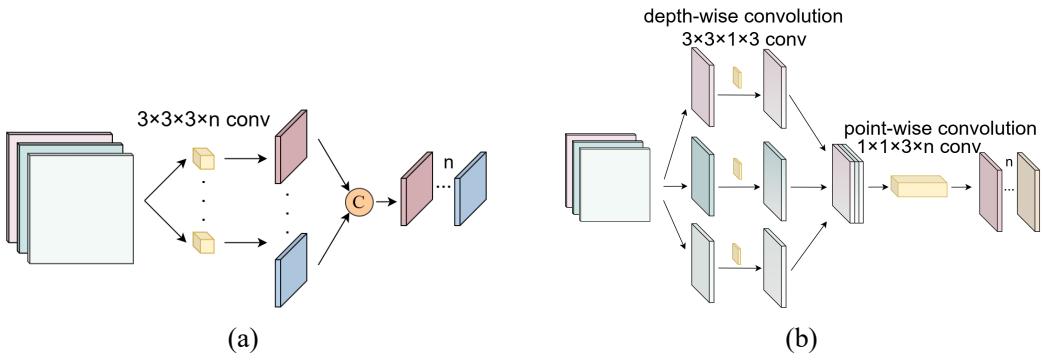


图 4-5 (a) 标准卷积 (b) 深度可分离卷积

假设 $K_w = K_h = K$, 通过将传统卷积替换为深度可分离卷积, 理论计算量可降低至标准卷积的 $\frac{1}{C_o} + \frac{1}{K^2}$, 图4-5直观对比了标准卷积与深度卷积的处理过程。

4.4 实验结果与分析

4.4.1 数据集与评价指标

我们采用 HIT-UAV 和 DroneVehicle 这两个数据集对所提出的网络进行评估。HIT-UAV 数据集包含 2898 张红外图像, 标注类别分为五类: 行人、汽车、自行车、其他车辆以及“其他”类别(指标注者无法准确归类的物体)。DroneVehicle 数据集则包含 28439 张图像, 提供 44163 个标注框, 类别包括汽车、公交车、卡车。该数据集专为无人机监控场景设计, 图像从不同高度和角度拍摄, 涵盖多种复杂背景。遵循 COCO 数据集的目标尺寸定义, 我们将尺寸小于 32×32 像素的物体归类为小目标。在此基础上, 我们进一步将小目标细分为两个子类: 尺寸在 16×16 像素以下的物体被定义为微小目标, 尺寸在 16×16 到 32×32 像素之间的物体被定义为小目标。这一细化的分类方案有助于更精确地评估模型, 特别是针对小目标的检测性能。模型训练使用一块 2080Ti GPU, 超参数设置如下: 初始学习率为 0.01, 最终学习率为 0.0001, 动量与权重衰减分别设为 0.9 和 0.0005。我们采用平均精度(Average Precision, AP)作为核心评估标准。在计算目标检测的 AP 时, 首先根据置信度对预测边界框进行排序, 随后计算精确率与召回率, 绘制 P-R 曲线, 最后计算每个类别曲线下的面积。 AP_{50-95} 通过在 0.5 至 0.95 区间内(步长 0.05)的多个 IoU 阈值上计算平均精确率, 提供了对模型检测性能更综合的评估。 AP_{50} 计算模型在 IoU 阈值为 0.5 时的性能。二者的计算方法如下:

$$AP_{50} = \int_0^1 Precision(r) dr \quad (4-11)$$

$$AP_{50-95} = \frac{1}{10} \sum_{i=0}^9 AP_{IOU=0.5+0.05 \cdot i} \quad (4-12)$$

此外，我们将 AP 进一步细分为 AP_{Tiny} 、 AP_{Small} 、 AP_{Medium} 和 AP_{Large} 。这些指标均在 IoU 阈值为 0.5 的条件下计算，从而能够更细致地评估模型检测不同尺寸目标的能力。

表 4-1 HIT-UAV 数据集结果对比

模型	出处	AP_{50} (%)	AP_{50-95} (%)	参数量	计算量	FPS
SuperYOLO ^[133]	TGRS 2023	83.7	51.6	7.7	20.89	41.7
QueryDet ^[19]	CVPR 2022	72.1	45.6	36.2	212.0	2.7
YOLC ^[134]	TITS 2024	74.0	46.8	67.8	—	1.8
CFPT ^[141]	TGRS 2025	82.4	52.5	37.17	83.13	13.7
CFINet ^[142]	ICCV 2023	69.4	43.3	43.96	111.59	15.7
YOLOv7	CVPR 2023	71.4	44.6	36.9	104.5	25.8
YOLOv8s	—	82.7	55.1	11.1	28.4	38.7
YOLOv11s	—	82.8	55.3	9.5	21.7	43.6
YOLOv12s	—	84.0	55.8	9.2	21.2	33.1
RT-DETR ^[104]	CVPR 2024	79.1	49.2	28.5	100.6	18.1
DEIM-s ^[143]	CVPR 2025	83.6	56.3	10.2	24.8	23.2
DEIMv2-s ^[144]	CVPR 2025	82.7	55.4	9.7	25.4	19.6
D-FINE-s ^[145]	ICLR 2025	84.1	57.2	10.2	24.8	25.2
MFF-DCNet	—	85.7	57.4	8.8	21.9	45.9

4.4.2 与 SOTA 的对比实验

1) HIT-UAV 实验结果

在 HIT-UAV 数据集上的实验结果如表4-1所示。MFF-DCNet 的 AP_{50} 达到 85.7%，帧率 (FPS) 为 45.9，优于所有对比方法。与基线模型 YOLOv11s 相比，我们在 AP_{50-95} 、参数量、GFLOPs 及 FPS 等各项指标上均表现更优。与 SuperYOLO 等专为无人机图像设计的检测器相比，MFF-DCNet 在仅增加 1.01GFLOPs 的情况下，实现了 AP_{50-95} 5.8% 的提升。至关重要的是，这一精度提升并未牺牲推理速度，MFF-DCNet 比 SuperYOLO 的帧率提升了 10%。相较于基于 Transformer 的方法 (RT-DETR、DEIM-s、DEIMv2-s 和 D-FINE-s)，MFF-DCNet 在推理速度上展现出明显优势，相比 RT-DETR，我们的 AP_{50} 提升了 8.2%，同时推理速度提高了三倍。这些 Transformer 方法通常结构复杂，导致了较高的计算开销。图4-6以 AP_{50} 为横轴、FPS 为纵轴直观呈现了上述结果，我们的方法在精度和速度上展现了全面领先，验证了所提架构在同步提升检测性能、模型效率和实时能力方面的有效性。

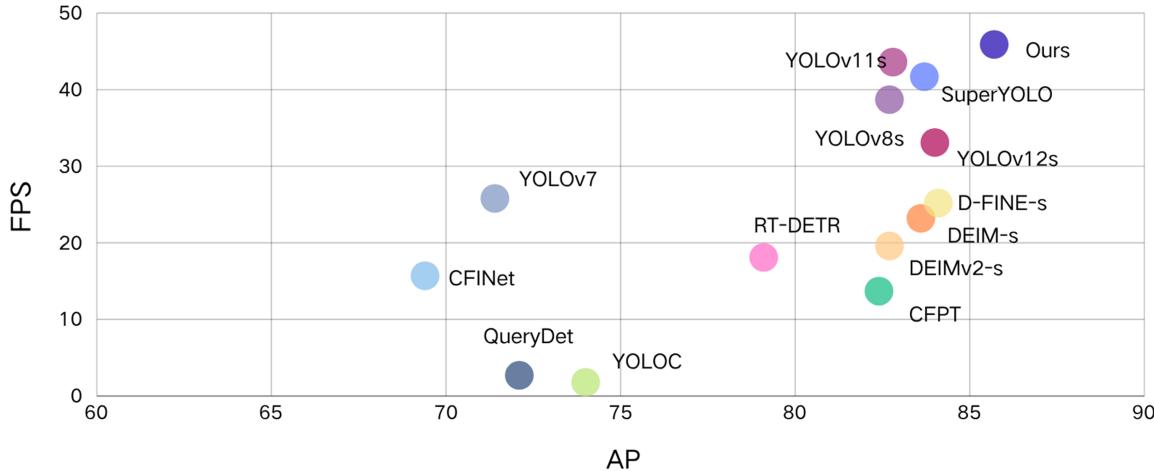


图 4-6 HIT-UAV 可视化结果对比

2) DroneVehicle 实验结果

在 DroneVehicle 数据集上的实验结果如表4-2所示。MFF-DCNet 取得了最高的检测精度，其 AP_{50} 达到 82.1%， AP_{50-95} 达到 59.6%，显著优于所有对比的先进方法。此外，我们的方法保持了最低的参数量，并实现了最快的推理速度，使其成为对比中最具计算效率的模型，这对于嵌入式部署而言是一个关键优势。

表 4-2 DroneVehicle 数据集结果对比

模型	出处	AP_{50} (%)	AP_{50-95} (%)	参数量	计算量	FPS
YOLOv11s	-	79.0	54.3	9.5	21.7	43.6
YOLOv12s	-	79.1	55.0	9.2	21.2	33.1
RT-DETR ^[104]	CVPR 2024	71.3	47.7	28.5	100.6	18.1
DEIM-s ^[143]	CVPR 2025	79.3	54.5	10.2	24.8	23.2
DEIMv2-s ^[144]	CVPR 2025	77.8	53.5	9.7	25.4	19.6
D-FINE-s ^[145]	ICLR 2025	78.4	53.8	10.2	24.8	25.2
MFF-DCNet	-	82.1	59.6	8.8	21.9	45.9

3) 嵌入式平台帧率测试

为评估 MFF-DCNet 的实际部署能力，我们在具有代表性的嵌入式平台 NVIDIA Jetson Orin NX 上进行了帧率测试。所有模型均在输入尺寸为 640×640 的条件下，使用 TensorRT 加速进行测试。结果如表4-3所示，MFF-DCNet 实现了最高的推理速度，达 39.6 FPS，优于所有对比的先进方法。RT-DETR、DEIM 和 D-FINE 等基于 Transformer 的方法，其参数量和计算成本均高于 MFF-DCNet。其二次计算复杂度的自注意力操作对于资源受限的嵌入式平台并非理想选择。相比之下，我们的 MFF-DCNet 在精度与速

度之间保持了良好平衡，使其更适用于计算能力有限的无人机实际应用场景。

表 4-3 在 NVIDIA Jetson Orin NX 的帧率测试结果

YOLOv11	YOLOv12	RT-DETR	DEIMv2	D-FINE-s	Ours
32.9	27.8	18.7	13.0	38.0	39.6

4.4.3 消融实验

1) 核心组件消融实验

本节我们通过实验验证 MFF 和 DCFormer 对整体性能的贡献。所有实验均在 HIT-UAV 数据集上、采用相同设置进行。基线模型为 YOLOv11s。为定量评估模型检测小目标的能力，我们在表中展示了 AP , AP_{Tiny} 和 AP_{Small} 指标。

实验结果如表4-4和表4-5所示，每个组件都贡献显著。与基线模型相比，单独引入 MFF 可使 AP_{Small} 指标获得 8.7% 的显著提升，凸显了其在增强小目标特征表征方面的能力。同时，单独引入 DCFormer 可使 AP_{Small} 提升 6.8%，证明了其能有效增强主干网络的特征提取能力。同时集成 MFF 和 DCFormer 模块在所有指标上均取得了最佳性能， AP_{Small} 从 64.2% 提高至 75.4%。这些结果共同验证了我们的方法成功应对了红外小目标检测中的关键挑战。

表 4-4 HIT-UAV 数据集的核心组件消融实验结果

baseline	MFF	DCFormer	AP _{50-95 (%)}	AP50 _{Tiny (%)}	AP50 _{Small (%)}
✓			55.3	57.6	64.2
✓	✓		55.8	59.1	72.9
✓		✓	57.0	58.8	71.0
✓	✓	✓	57.4	60.0	75.4

表 4-5 DroneVehicle 数据集的核心组件消融实验结果

baseline	MFF	DCFormer	AP _{50-95 (%)}	AP50 _{Tiny (%)}	AP50 _{Small (%)}
✓			54.3	12.3	45.8
✓	✓		59.0	12.7	43.0
✓		✓	58.1	12.4	41.6
✓	✓	✓	59.6	13.1	47.3

2) 多特征聚焦模块消融实验

本章提出的 MFF 模块，通过高效聚合来自不同分辨率的特征图以形成统一、全面的集成表征，成为网络颈部结构的核心组件。该设计旨在增强网络捕获对目标检测至关重要的复杂细节和空间信息的能力。在本节，我们在原始 YOLOv11s 网络架构基础上，集成了 MFF 模块以及其他先进的特征聚合模块进行对比，包括双向特征金字塔网络 (Bidirectional Feature Pyramid Network, BiFPN)^[25]、注意力尺度序列融合 (Attentional Scale Sequence Fusion, ASF)^[146] 和渐进式特征金字塔网络 (Asymptotic Feature Pyramid Network, AFPN)^[147]。

表 4-6 多特征聚焦模块消融实验结果

模型	AP50 _{Tiny}	AP50 _{Small}	AP50 _{Medium}	AP50 _{Large}	参数量	AP ₅₀₋₉₅
YOLOv11s	57.6	64.2	88.2	71.7	9.5	55.3
YOLOv11s-BiFPN	58.0	65.1	89.1	69.2	7.1	55.3
YOLOv11s-ASF	60.3	68.1	82.8	72.0	9.8	55.9
YOLOv11s-AFPN-P345	51.9	64.8	85.4	64.5	9.5	53.4
YOLOv11s-MFF	59.1	72.9	87.0	73.0	9.3	55.8

实验结果如表4-6所示。我们以 YOLOv11s 为基线，评估了不同特征聚合模块对目标检测性能的影响。在对微小目标 AP_{Tiny} 的提升上，MFF 的确稍差于 ASF 模块，两者相差 1.2%。然而，MFF 对小目标 AP_{Small} 的改进非常显著，将其从基线的 64.2% 提升至 72.9%，这比 ASF 高出 4.8%。对于中等目标 AP_{Medium} ，MFF 相比 ASF 也有 4.2% 的更大提升。尽管 BiFPN 在 AP_{Medium} 上取得了最大的改进，但其对小目标和整体 AP 的贡献非常有限，这对于机载场景下的检测任务价值不高。总体而言，MFF 带来的性能提升最符合我们对模型的预期。

为进一步说明 MFF 模块在特征聚合中的作用，我们将其移植到早期版本的 YOLO 架构中，并在相同的数据集上进行了测试。具体而言，我们将 MFF 模块集成到 YOLOv5s 和 YOLOv9s 网络的颈部结构，得到 YOLOv5s-MFF 和 YOLOv9s-MFF，并对比了其与原始版本在检测性能上的差异。

表 4-7 MFF 对于 YOLOv5s 和 YOLOv9s 的效果验证

模型	AP50 _{Tiny}	AP50 _{Small}	AP50 _{Medium}	AP50 _{Large}	AP ₅₀₋₉₅
YOLOv5s	52.8	60.2	85.3	65.0	51.7
YOLOv5s-MFF	56.5	70.8	83.1	66.5	54.4
YOLOv9s	59.2	66.0	87.6	70.3	55.7
YOLOv9s-MFF	57.6	74.3	88.5	65.3	57.1

结果如表4-7所示，YOLOv5s-MFF 与 YOLOv9s-MFF 的检测精度均得到显著提升，尤其在具有挑战性的小目标检测场景中。具体而言：在 YOLOv5s 上，MFF 模块使 AP_{Small}

表 4-8 DCFormer 消融实验结果

模型	AP50 _{Tiny}	AP50 _{Small}	AP50 _{Medium}	AP50 _{Large}	参数量	AP ₅₀₋₉₅
YOLOv11s	57.6	64.2	88.2	71.7	9.5	55.3
YOLOv11s-MFF-DLKA	56.5	70.1	88.9	58.7	10.6	56.7
YOLOv11s-MFF-EMSC	57.0	69.5	88.3	65.9	9.1	55.3
YOLOv11s-MFF-FADC	55.7	70.0	86.6	65.9	9.3	55.7
YOLOv11s-MFF-DCFormer	60.0	75.4	91.2	68.0	8.8	57.4

提升了 10.6%，同时改善了 AP_{Tiny} 与 AP_{Large} 指标。在 YOLOv9s 上，MFF 模块将 AP_{Small} 从 66.0% 提升至 74.3%， AP_{Medium} 从 87.6% 提升至 88.5%。虽然在某些尺度下的精度会下降，但整体性能特别是对小目标的精度提升显著，验证了 MFF 模块在不同 YOLO 架构中的通用性和有效性。

3) DCFormer 消融实验

本节将评估 DCFormer 的有效性，我们将其与当前其他先进的主干网络改进模块（包括 DLKA^[148]、EMSC^[149]、FADC^[150]）进行对比。结果如表4-8所示。我们的 YOLOv11s-MFF-DCFormer 模型实现了 57.4% 的 AP_{50-95} ，相比基线模型有显著提升。该模型的综合性能通过雷达图4-7进行了可视化展示，该图对比了不同目标尺度下的检测精度以及整体的 AP_{50-95} 指标，我们的模型在所有维度上均展现出性能优势。此外，我们的模型参数量最少，仅为 8.8M，使其在资源受限的边缘平台上部署时具备显著优势。

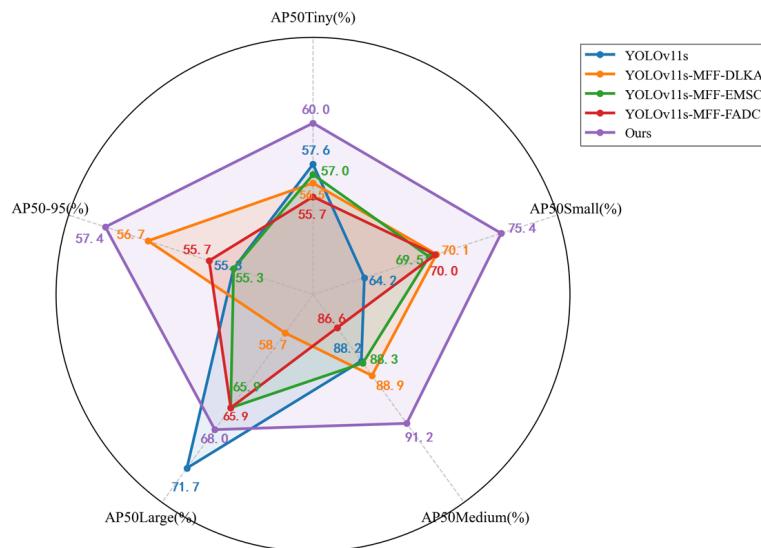


图 4-7 不同尺度目标检测性能对比雷达图。为增强视觉区分度，各坐标轴数值已进行归一化处理。

4.4.4 可视化结果

本节通过定性分析展示我们的方法在红外小目标检测中的优势。

图4-8展示了MFF-DCNet与多种先进检测器的结果对比，包括基于CNN的YOLOv11、YOLOv12，以及基于Transformer的RT-DETR、D-FINE。与基线YOLOv11相比，MFF-DCNet表现出显著改善。可以观察到，在第一行和第四行样例中，仅有MFF-DCNet未产生误检，而在第四行和第六行中，仅有MFF-DCNet未出现漏检。图4-9进一步给出了基线模型与MFF-DCNet的检测结果及对应热力图的定性对比。热力图显示，我们的模型对目标区域展现出更优的聚焦能力，同时对背景噪声和杂乱信息的抑制也更为有效。

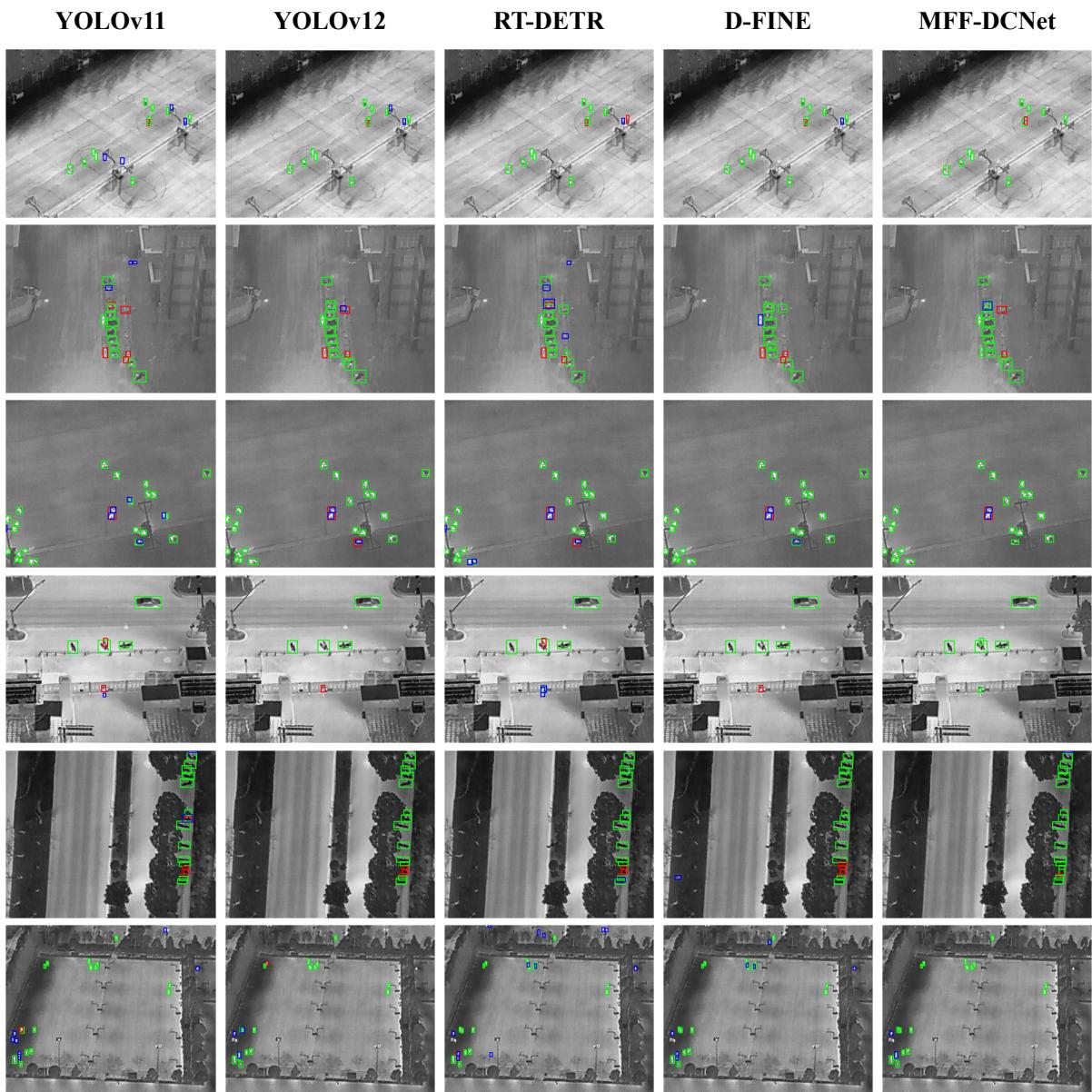


图 4-8 与先进检测器的检测结果可视化对比。其中，绿色、蓝色与红色边界框分别表示正确检测、误检以及漏检。

我们的方法相较于 YOLO 系列等基于 CNN 的检测器，在保留红外小目标的差异性特征方面具有优势。此类检测器虽然在通用目标检测上表现出色，但其串行的卷积结构容易在网络层级中逐渐削弱红外小目标本就有限的特征。MFF 模块通过对多尺度特征的聚合与增强，有效弥补了这种信息衰减，使网络能够保留对小目标检测至关重要的精细特征。相较于 RT-DETR、D-FINE 等基于 Transformer 的检测器，我们的方法在全局上下文建模与计算效率之间取得了更优的平衡。自注意力操作的二次复杂度会带来过高的计算与内存开销，使其难以满足嵌入式平台的实时性与功耗约束。DCFormer 在提供有效上下文建模能力的同时，保持了线性复杂度，使其更适合无人机的实时应用场景。

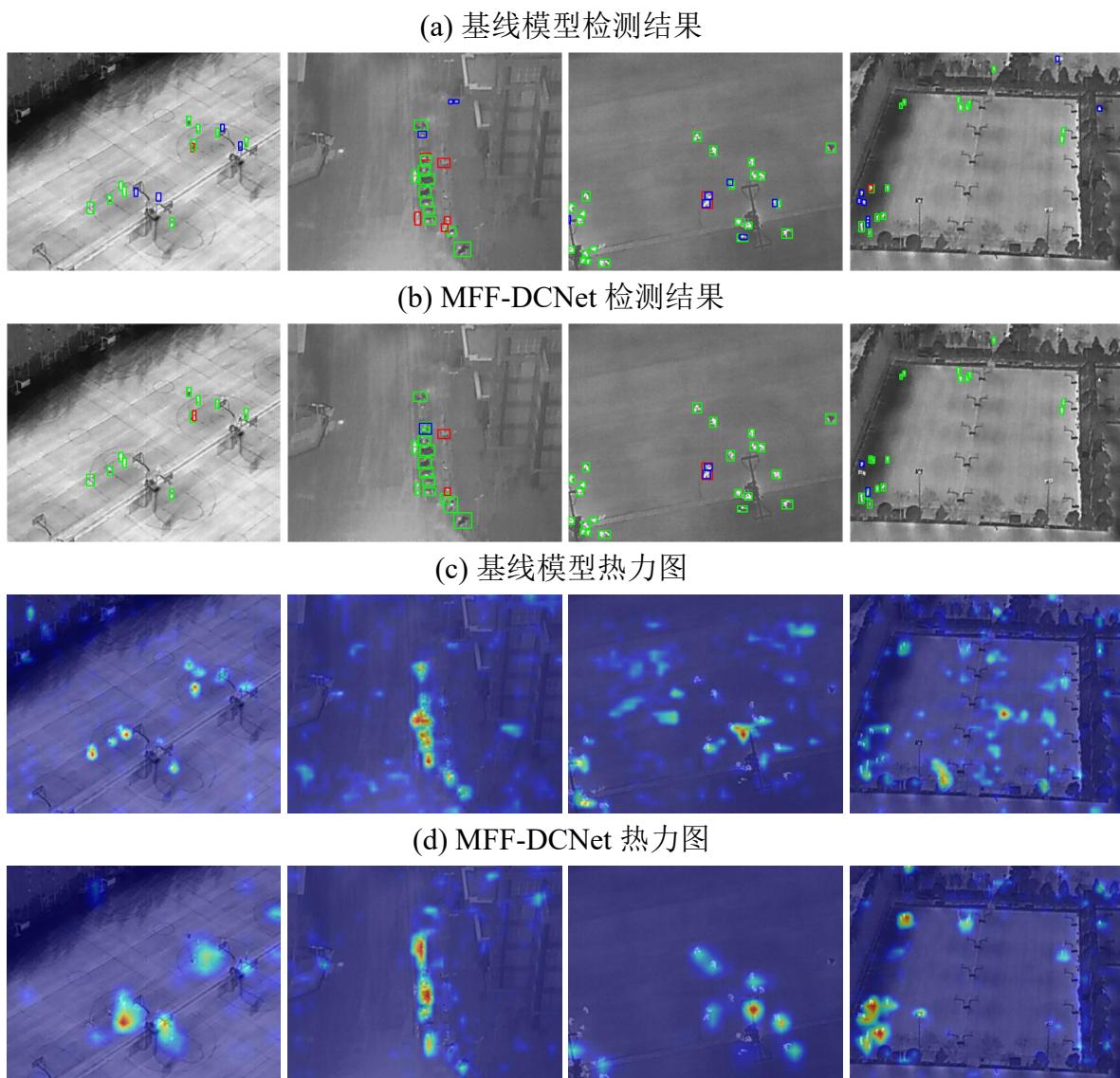


图 4-9 基线模型与 MFF-DCNet 检测结果及热力图结果对比，(a)、(b) 行展示目标检测结果，其中，绿色边界框代表正确检测的目标，红色边界框表示漏检，蓝色边界框则为误检。(c)、(d) 行展示对应热力图，热力图显示了网络在进行预测时的关注区域，MFF-DCNet 通过更精准的特征聚焦机制，在提升目标检测准确性的同时，有效降低了漏检与误检。

4.5 本章小结

本章提出了 MFF-DCNet，一种专为无人机红外小目标检测设计的神经网络。提出的多特征聚焦模块实现了多尺度特征信息的有效融合，提升了模型在复杂环境中对不同尺寸目标，尤其是小目标的检测能力。深度可分离跨阶段 Transformer 模块的引入，进一步增强了模型对空间关系与上下文信息的提取能力，从而提升了对红外图像中细微特征变化的感知灵敏度。在两个公开数据集上的实验表明，我们的网络取得了领先的性能，并在处理速度上实现了显著提升。

致 谢

漫漫求学路，最让人回味的，莫属于读博这几年。回首初入交大时情不自禁的喜悦，经历了硕博八年洗礼后，依旧幸福感满存。谨以此文聊表感激之心。

衷心感谢我的导师孙宏滨教授在博士期间对我的悉心指导与关怀。在刚进组时，孙老师就将新发现号的搭建工作交与我全权负责，帮助我从系统的角度对无人驾驶整体研究有了深刻认识；在博二时，孙老师就让我担任了发现号车队队长并认真细致地指导我们准备每年的未来挑战赛，极大地提高了我的组织管理能力；在科研工作中，孙老师指点迷津，引领我做好科研探索。孙老师严谨务实的科研态度，一丝不苟的治学精神，高屋建瓴的学术见地，勤奋谦虚的个人品质都深深感染着我，激励着我，使我受益终生。

感谢我们敬爱的郑南宁院士。郑老师对于无人驾驶车队的关心和指导使我们整个车队的技术水平得到不断提高。感谢我博士前两年的合作导师辛景民教授的关怀，感谢魏平教授在智能车未来挑战赛备赛和比赛过程中的悉心教导，感谢王乐教授在轨迹预测方面的支持，感谢薛建儒教授、兰旭光教授、任鹏举教授、杜少毅教授、徐林海高级工程师、陈仕韬助理教授、王芳芳工程师以及其他所有人工智能学院老师在我读博期间给予的帮助和支持。

感谢课题组张旭翀师兄和汪航师兄对我科研工作一直以来的帮助，两位师兄扎实的理论功底和极强的解决问题能力都给我留下了很深印象。感谢沈源、张婧、刘丹为我们的学习生活提供的便利。

感谢王潇、史菊旺、李庚欣、陶中幸、张璞等师兄师姐在科研上的关照。感谢冯洋、杨帅、吴金强、冯超、向钊宏、陈达、张志浩、王玉学、韩伟光、权柄章、钱成龙、葛冲、陈科、李诚、罗鑫凯、陈煜炜、王申奥、李天航等师弟在发现号无人驾驶平台开发和无人车比赛中的付出。感谢戴赫、孙长峰、郑方、段景海、石刘帅等师弟在小论文上的帮助。感谢同届张剑、杨少飞、李宝婷的帮助。感谢唐浩雯师妹在科研生活中的交流与帮助。感谢好友冯立琛、丁兆伦、雷洁、马晨、荣韧闲暇时度过的快乐时光。感谢和我一起在创新港并肩战斗的赵博然，在科研和为人处世方面都对我产生了很大影响。

最后，感谢我的父母和家人多年来对我学习和生活上的关心和支持，是你们的坚强后盾让我能够全身心地投入到科研探索中。感恩一路有你们相伴，你们永远是我内心最温暖的港湾。

参考文献

- [1] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J/OL]. ArXiv, 2020, abs/2004.10934. <https://api.semanticscholar.org/CorpusID:216080778>.
- [2] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond Empirical Risk Minimization[C/OL]. 2018. <https://openreview.net/forum?id=r1Ddp1-Rb>.
- [3] Yun S, Han D, Oh S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6023-6032.
- [4] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.
- [5] Chen C, Zhang Y, Lv Q, et al. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2019: 100-108. DOI: 10.1109/ICCVW.2019.00018.
- [6] Xiao J, Guo H, Zhou J, et al. Tiny object detection with context enhancement and feature purification [J]. Expert Systems with Applications, 2023, 211: 118665.
- [7] Ünel F Ö, Özkalayci B O, Çığla C. The Power of Tiling for Small Object Detection[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019: 582-591. DOI: 10.1109/CVPRW.2019.00084.
- [8] Yu X, Gong Y, Jiang N, et al. Scale Match for Tiny Person Detection[C/OL]. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020: 1246-1254. DOI: 10.1109/WACV45572.2020.9093394.
- [9] Lin J, Jing W, Song H. SAN: Scale-aware network for semantic segmentation of high-resolution aerial images[J]. arXiv preprint arXiv:1907.03089, 2019.
- [10] Zoph B, Cubuk E D, Ghiasi G, et al. Learning data augmentation strategies for object detection[C]. European conference on computer vision. 2020: 566-583.
- [11] Yang F, Choi W, Lin Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2129-2137. DOI: 10.1109/CVPR.2016.234.
- [12] Cai Z, Fan Q, Feris R S, et al. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection[C]. Leibe B, Matas J, Sebe N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 354-370.
- [13] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[EB/OL]. 2018. <https://arxiv.org/abs/1804.02767>. arXiv: 1804.02767 [cs.CV].
- [14] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 936-944. DOI: 10.1109/CVPR.2017.106.
- [15] Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 7029-7038. DOI: 10.1109/CVPR.2019.00720.

- [16] 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). 2021: 10208-10219. DOI: 10.1109/CVPR46437.2021.01008.
- [17] Li J, Liang X, Shen S, et al. Scale-Aware Fast R-CNN for Pedestrian Detection[J/OL]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996. DOI: 10.1109/TMM.2017.2759508.
- [18] Li Y, Chen Y, Wang N, et al. Scale-Aware Trident Networks for Object Detection[J]. ICCV 2019, 2019.
- [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). 2022: 13658-13667. DOI: 10.1109/CVPR52688.2022.01330.
- [20] Singh B, Davis L S. An Analysis of Scale Invariance in Object Detection - SNIP[J/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 3578-3587. <https://api.semanticscholar.org/CorpusID:4615054>.
- [21] Singh B, Najibi M, Davis L S. SNIPER: Efficient Multi-Scale Training[J]. NeurIPS, 2018.
- [22] Najibi M, Singh B, Davis L S. AutoFocus: Efficient Multi-Scale Inference[J]. ICCV, 2019.
- [23] Chen Y, Zhang P, Li Z, et al. Dynamic Scale Training for Object Detection[EB/OL]. 2021. <https://arxiv.org/abs/2004.12432>. arXiv: 2004.12432 [cs.CV].
- [24] Liu S, Qi L, Qin H, et al. Path Aggregation Network for Instance Segmentation[C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768. DOI: 10.1109/CVPR.2018.00913.
- [25] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 10778-10787. DOI: 10.1109/CVPR42600.2020.01079.
- [26] Zhang H, Wang K, Tian Y, et al. MFR-CNN: Incorporating Multi-Scale Features and Global Information for Traffic Object Detection[J/OL]. IEEE Transactions on Vehicular Technology, 2018, 67(9): 8019-8030. DOI: 10.1109/TVT.2018.2843394.
- [27] Woo S, Hwang S, Kweon I S. StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection[J/OL]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017: 1093-1102. <https://api.semanticscholar.org/CorpusID:13681687>.
- [28] Zhao Q, Sheng T, Wang Y, et al. M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network[C]. The Thirty-Third AAAI Conference on Artificial Intelligence,AAAI. 2019.
- [29] Liu Z, Gao G, Sun L, et al. IPG-Net: Image Pyramid Guidance Network for Small Object Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020: 4422-4430. DOI: 10.1109/CVPRW50498.2020.00521.
- [30] Gong Y, Yu X, Ding Y, et al. Effective Fusion Factor in FPN for Tiny Object Detection[C/OL]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021: 1159-1167. DOI: 10.1109/WACV48630.2021.00120.
- [31] Hong M, Li S, Yang Y, et al. SSPNet: Scale Selection Pyramid Network for Tiny Person Detection From UAV Images[J/OL]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5. DOI: 10.1109/LGRS.2021.3103069.
- [32] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. NIPS'14: Pro-

- ceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. Montreal, Canada: MIT Press, 2014: 2672-2680.
- [33] Li J, Liang X, Wei Y, et al. Perceptual Generative Adversarial Networks for Small Object Detection [J/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1951-1959. <https://api.semanticscholar.org/CorpusID:6704804>.
- [34] Bai Y, Zhang Y, Ding M, et al. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network[C]. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 210-226.
- [35] Pang Y, Cao J, Wang J, et al. JCS-Net: Joint Classification and Super-Resolution Network for Small-Scale Pedestrian Detection in Surveillance Images[J/OL]. IEEE Transactions on Information Forensics and Security, 2019, 14(12): 3322-3331. DOI: 10.1109/TIFS.2019.2916592.
- [36] Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1646-1654. DOI: 10.1109/CVPR.2016.182.
- [37] Cao J, Pang Y, Li X. Learning Multilayer Channel Features for Pedestrian Detection[J/OL]. IEEE Transactions on Image Processing, 2017, 26(7): 3210-3220. DOI: 10.1109/TIP.2017.2694224.
- [38] Fu C Y, Liu W, Ranga A, et al. DSSD : Deconvolutional Single Shot Detector[EB/OL]. 2017. <https://arxiv.org/abs/1701.06659>. arXiv: 1701.06659 [cs.CV].
- [39] 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. DOI: 10.1109/TCYB.2020.3004636.
- [40] Sun J, Gao H, Wang X, et al. Scale Enhancement Pyramid Network for Small Object Detection from UAV Images[J/OL]. Entropy, 2022, 24(11). <https://www.mdpi.com/1099-4300/24/11/1699>. DOI: 10.3390/e24111699.
- [41] Corsel C W, van Lier M, Kampmeijer L, et al. Exploiting Temporal Context for Tiny Object Detection[C/OL]. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). 2023: 1-11. DOI: 10.1109/WACVW58289.2023.00013.
- [42] Zhang S, Wen L, Bian X, et al. Single-Shot Refinement Neural Network for Object Detection [C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4203-4212. DOI: 10.1109/CVPR.2018.00442.
- [43] Yi K, Jian Z, Chen S, et al. Feature Selective Small Object Detection via Knowledge-based Recurrent Attentive Neural Network[EB/OL]. 2019. <https://arxiv.org/abs/1803.05263>. arXiv: 1803.05263 [cs.CV].
- [44] Yang X, Yang J, Yan J, et al. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 8231-8240. DOI: 10.1109/ICCV.2019.00832.
- [45] Fu J, Sun X, Wang Z, et al. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(2): 1331-1344. DOI: 10.1109/TGRS.2020.3005151.
- [46] Tian Z, Shen C, Chen H, et al. FCOS: A Simple and Strong Anchor-Free Object Detector[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933. DOI: 10.1109/TPAMI.2020.3032166.
- [47] Lu X, Ji J, Xing Z, et al. Attention and Feature Fusion SSD for Remote Sensing Object Detection

- [J/OL]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-9. DOI: 10.1109/TI.M.2021.3052575.
- [48] Ran Q, Wang Q, Zhao B, et al. Lightweight Oriented Object Detection Using Multiscale Context and Enhanced Channel Attention in Remote Sensing Images[J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 5786-5795. DOI: 10.1109/JSTARS.2021.3079968.
- [49] Li Y, Huang Q, Pei X, et al. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery[J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 2148-2161. DOI: 10.1109/JSTARS.2020.3046482.
- [50] 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). 2019: 8310-8319. DOI: 10.1109/ICCV.2019.900840.
- [51] Duan C, Wei Z, Zhang C, et al. Coarse-grained Density Map Guided Object Detection in Aerial Images[C/OL]. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021: 2789-2798. DOI: 10.1109/ICCVW54120.2021.00313.
- [52] Li C, Yang T, Zhu S, et al. Density Map Guided Object Detection in Aerial Images[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020: 737-746. DOI: 10.1109/CVPRW50498.2020.00103.
- [53] Wang Y, Yang Y, Zhao X. Object Detection Using Clustering Algorithm Adaptive Searching Regions in Aerial Images[C]. Computer Vision – ECCV 2020 Workshops. Springer International Publishing, 2020: 651-664.
- [54] Deng S, Li S, Xie K, et al. A Global-Local Self-Adaptive Network for Drone-View Object Detection [J/OL]. IEEE Transactions on Image Processing, 2021, 30: 1556-1569. DOI: 10.1109/TIP.2020.3045636.
- [55] Xu J, Li Y, Wang S. AdaZoom: Adaptive Zoom Network for Multi-Scale Object Detection in Large Scenes[EB/OL]. 2021. <https://arxiv.org/abs/2106.10409>. arXiv: 2106.10409 [cs.CV].
- [56] Leng J, Mo M, Zhou Y, et al. Pareto Refocusing for Drone-View Object Detection[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1320-1334. DOI: 10.1109/TCSVT.2022.3210207.
- [57] Koyun O C, Keser R K, Akkaya İ B, et al. Focus-and-Detect: A small object detection framework for aerial images[J/OL]. Signal Processing: Image Communication, 2022, 104: 116675. DOI: <https://doi.org/10.1016/j.image.2022.116675>.
- [58] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters [C/OL]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 2544-2550. DOI: 10.1109/CVPR.2010.5539960.
- [59] Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596. DOI: 10.1109/TPAMI.2014.2345390.
- [60] Danelljan M, Häger G, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C/OL]. 2015 IEEE International Conference on Computer Vision (ICCV). 2015: 4310-4318. DOI: 10.1109/ICCV.2015.490.
- [61] Valmadre J, Bertinetto L, Henriques J, et al. End-to-End Representation Learning for Correlation

- Filter Based Tracking[C/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5000-5008. DOI: 10.1109/CVPR.2017.531.
- [62] Bhat G, Danelljan M, Van Gool L, et al. Learning Discriminative Model Prediction for Tracking [C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 6181-6190. DOI: 10.1109/ICCV.2019.00628.
- [63] Danelljan M, Van Gool L, Timofte R. Probabilistic Regression for Visual Tracking[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 7181-7190. DOI: 10.1109/CVPR42600.2020.00721.
- [64] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]. Computer Vision – ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 850-865.
- [65] He A, Luo C, Tian X, et al. A Twofold Siamese Network for Real-Time Object Tracking[C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4834-4843. DOI: 10.1109/CVPR.2018.00508.
- [66] Li B, Yan J, Wu W, et al. High Performance Visual Tracking with Siamese Region Proposal Network [J/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980. <https://api.semanticscholar.org/CorpusID:52255840>.
- [67] Zhu Z, Wang Q, Bo L, et al. Distractor-aware Siamese Networks for Visual Object Tracking[C]. European Conference on Computer Vision. 2018.
- [68] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4277-4286. DOI: 10.1109/CVPR.2019.00441.
- [69] 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). 2016: 770-778. DOI: 10.1109/CVPR.2016.90.
- [70] Xu Y, Wang Z, Li Z, et al. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12549-12556. <https://ojs.aaai.org/index.php/AAAI/article/view/6944>. DOI: 10.1609/aaai.v34i07.6944.
- [71] Yu Y, Xiong Y, Huang W, et al. Deformable Siamese Attention Networks for Visual Object Tracking [C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 6727-6736. DOI: 10.1109/CVPR42600.2020.00676.
- [72] Liu J, Wang H, Ma C, et al. SiamDMU: Siamese Dual Mask Update Network for Visual Object Tracking[J/OL]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(2): 1656-1669. DOI: 10.1109/TETCI.2024.3353674.
- [73] Voigtlaender P, Luiten J, Torr P H, et al. Siam R-CNN: Visual Tracking by Re-Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 6577-6587. DOI: 10.1109/CVPR42600.2020.00661.
- [74] Chen X, Yan B, Zhu J, et al. Transformer Tracking[C]. CVPR. 2021.
- [75] Wang N, Zhou W, Wang J, et al. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking[C/OL]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 1571-1580. DOI: 10.1109/CVPR46437.2021.00162.

- [76] Yan B, Peng H, Fu J, et al. Learning Spatio-Temporal Transformer for Visual Tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 10448-10457.
- [77] Song Z, Yu J, Chen Y P P, et al. Transformer Tracking with Cyclic Shifting Window Attention [C/OL]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 8781-8790. DOI: 10.1109/CVPR52688.2022.00859.
- [78] Cui Y, Jiang C, Wang L, et al. MixFormer: End-to-End Tracking with Iterative Mixed Attention [C/OL]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 13598-13608. DOI: 10.1109/CVPR52688.2022.01324.
- [79] Wu H, Xiao B, Codella N, et al. CvT: Introducing Convolutions to Vision Transformers[C/OL]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 22-31. DOI: 10.1109/ICCV48922.2021.00009.
- [80] Lin L, Fan H, Zhang Z, et al. SwinTrack: A Simple and Strong Baseline for Transformer Tracking [C/OL]. Advances in Neural Information Processing Systems: vol. 35. 2022: 16743-16754. https://proceedings.neurips.cc/paper_files/paper/2022/file/6a5c23219f401f3efd322579002dbb80-Paper-Conference.pdf.
- [81] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J/OL]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9992-10002. <https://api.semanticscholar.org/CorpusID:232352874>.
- [82] Ye B, Chang H, Ma B, et al. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework[C]. ECCV. 2022.
- [83] Chen X, Peng H, Wang D, et al. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking[C/OL]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 14572-14581. DOI: 10.1109/CVPR52729.2023.01400.
- [84] Hong L, Yan S, Zhang R, et al. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning[C/OL]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 19079-19091. DOI: 10.1109/CVPR52733.2024.01805.
- [85] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. 2014: 740-755.
- [86] Jocher G, Qiu J, Chaurasia A. Ultralytics YOLO[CP/OL]. 8.0.0. 2023. <https://github.com/ultralytics/ultralytics>.
- [87] Cao Y, He Z, Wang L, et al. VisDrone-DET2021: The vision meets drone object detection challenge results[C]. Proceedings of the IEEE/CVF International conference on computer vision. 2021: 2847-2854.
- [88] Lv W, Zhao Y, Chang Q, et al. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer[J]. arXiv preprint arXiv:2407.17140, 2024.
- [89] Leng J, Ye Y, Mo M, et al. Recent Advances for Aerial Object Detection: A Survey[J]. ACM Computing Surveys, 2024, 56(12): 1-36.
- [90] Li Y, Wu P, Zhang M. Rethinking the sparse mask learning mechanism in sparse convolution for object detection on drone images[J]. Computer Vision and Image Understanding, 2025: 104432.
- [91] Tan L, Liu Z, Liu H, et al. A Real-Time Unmanned Aerial Vehicle (UAV) Aerial Image Object Detection Model[C]. 2024 International Joint Conference on Neural Networks (IJCNN). 2024: 1-7.

- [92] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. European conference on computer vision. 2020: 213-229.
- [93] Huang Y X, Liu H I, Shuai H H, et al. Dq-detr: Detr with dynamic query for tiny object detection [C]. European Conference on Computer Vision. 2024: 290-305.
- [94] Du D, Qi Y, Yu H, et al. The unmanned aerial vehicle benchmark: Object detection and tracking [C]. Proceedings of the European conference on computer vision (ECCV). 2018: 370-386.
- [95] Wang J, Yang W, Guo H, et al. Tiny Object Detection in Aerial Images[C/OL]. 2020 25th International Conference on Pattern Recognition (ICPR). 2021: 3791-3798. DOI: 10.1109/ICPR48806.2021.9413340.
- [96] Xu X, Mao Z, Wang X, et al. Dynamic Anchor: Density Map Guided Small Object Detector for Tiny Persons[J]. Computer Vision and Image Understanding, 2025, 255: 104325.
- [97] Li C, Yang T, Zhu S, et al. Density map guided object detection in aerial images[C]. proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 190-191.
- [98] Du B, Huang Y, Chen J, et al. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 13435-13444.
- [99] Akyon F C, Altinuc S O, Temizel A. Slicing aided hyper inference and fine-tuning for small object detection[C]. 2022 IEEE international conference on image processing (ICIP). 2022: 966-970.
- [100] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection [J]. arXiv preprint arXiv:2010.04159, 2020.
- [101] Li F, Zhang H, Liu S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13619-13627.
- [102] Yao Z, Ai J, Li B, et al. Efficient detr: improving end-to-end object detector with dense prior[J]. arXiv preprint arXiv:2104.01318, 2021.
- [103] Roh B, Shin J, Shin W, et al. Sparse detr: Efficient end-to-end object detection with learnable sparsity[J]. arXiv preprint arXiv:2111.14330, 2021.
- [104] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.
- [105] Zhang H, Liu K, Gan Z, et al. UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery[J]. arXiv preprint arXiv:2501.01855, 2025.
- [106] Xue H, Tang Z, Xia Y, et al. HCTD: A CNN-transformer hybrid for precise object detection in UAV aerial imagery[J]. Computer Vision and Image Understanding, 2025: 104409.
- [107] Chen L, Fu Y, Gu L, et al. Frequency-aware feature fusion for dense image prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
- [108] Wang J, Chen K, Xu R, et al. CARAFE: Content-Aware ReAssembly of FEatures[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 3007-3016. DOI: 10.1109/ICCV.2019.00310.
- [109] Wang J, Xu C, Yang W, et al. A normalized Gaussian Wasserstein distance for tiny object detection [J]. arXiv preprint arXiv:2110.13389, 2021.
- [110] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]. European conference on computer vision. 2024: 1-21.

- [111] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.
- [112] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [113] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 840-849.
- [114] Liu Z, Gao G, Sun L, et al. HRDNet: High-resolution detection network for small objects[C]. 2021 IEEE international conference on multimedia and expo (ICME). 2021: 1-6.
- [115] Xu C, Wang J, Yang W, et al. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark[J/OL]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 79-93. DOI: <https://doi.org/10.1016/j.isprsjprs.2022.06.002>.
- [116] Guo G, Chen P, Yu X, et al. Save the Tiny, Save the All: Hierarchical Activation Network for Tiny Object Detection[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34: 221-234. DOI: 10.1109/TCSVT.2023.3284161.
- [117] Zhao M, Li W, Li L, et al. Single-frame infrared small-target detection: A survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 87-119.
- [118] Tong K, Wu Y. Deep learning-based detection from the perspective of small or tiny objects: A survey[J]. Image and Vision Computing, 2022, 123: 104471.
- [119] Kou R, Wang C, Peng Z, et al. Infrared small target segmentation networks: A survey[J]. Pattern Recognition, 2023, 143: 109788.
- [120] Tong K, Wu Y. Deep learning-based detection from the perspective of small or tiny objects: A survey[J]. Image and Vision Computing, 2022, 123: 104471.
- [121] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [122] Xu X, Feng Z, Cao C, et al. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation[J/OL]. Remote Sensing, 2021, 13(23). DOI: 10.3390/rs13234779.
- [123] Xue J, He D, Liu M, et al. Dual Network Structure With Interweaved Global-Local Feature Hierarchy for Transformer-Based Object Detection in Remote Sensing Image[J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 6856-6866. DOI: 10.1109/JSTARS.2022.3198577.
- [124] Suo J, Wang T, Zhang X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data, 2023, 10(1): 227.
- [125] Sun Y, Cao B, Zhu P, et al. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713. DOI: 10.1109/TCSVT.2022.3168279.
- [126] Zhang G, Xu G, Chen S, et al. It's Not the Target, It's the Background: Rethinking Infrared Small-Target Detection via Deep Patch-Free Low-Rank Representations[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-13. DOI: 10.1109/TGRS.2025.3608239.
- [127] Dai Y, Wu Y, Zhou F, et al. Attentional local contrast networks for infrared small target detection [J]. IEEE transactions on geoscience and remote sensing, 2021, 59(11): 9813-9824.

- [128] Min X, Zhou W, Hu R, et al. LWUAVDet: A Lightweight UAV Object Detection Network on Edge Devices[J]. IEEE Internet of Things Journal, 2024, 11(13): 24013-24023.
- [129] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. ArXiv, 2017, abs/1704.04861.
- [130] Zhang X, Zhou X, Lin M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856. DOI: 10.1109/CVPR.2018.00716.
- [131] Han K, Wang Y, Tian Q, et al. GhostNet: More Features From Cheap Operations[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1577-1586. DOI: 10.1109/CVPR42600.2020.00165.
- [132] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
- [133] Zhang J, Lei J, Xie W, et al. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.
- [134] Liu C, Gao G, Huang Z, et al. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images[J]. IEEE Transactions on Intelligent Transportation Systems, 2024.
- [135] Guo C, Fan B, Zhang Q, et al. Augfpn: Improving multi-scale feature learning for object detection [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12595-12604.
- [136] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6824-6835.
- [137] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.
- [138] Chen C F R, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 357-366.
- [139] Yu W, Si C, Zhou P, et al. Metaformer baselines for vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(2): 896-912.
- [140] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
- [141] Du Z, Hu Z, Zhao G, et al. Cross-Layer Feature Pyramid Transformer for Small Object Detection in Aerial Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-14.
- [142] Yuan X, Cheng G, Yan K, et al. Small object detection via coarse-to-fine proposal generation and imitation learning[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6317-6327.
- [143] Huang S, Lu Z, Cun X, et al. DEIM: DETR with Improved Matching for Fast Convergence[J/OL]. 2025: 15162-15171. DOI: 10.1109/CVPR52734.2025.01412.
- [144] Huang S, Hou Y, Liu L, et al. Real-Time Object Detection Meets DINOV3[J]. arXiv, 2025.
- [145] Peng Y, Li H, Wu P, et al. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement[C]. The Thirteenth International Conference on Learning Representations. 2025.

- [146] Kang M, Ting C M, Ting F F, et al. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. *Image and Vision Computing*, 2024, 147: 105057.
- [147] Yang G, Lei J, Tian H, et al. Asymptotic Feature Pyramid Network for Labeling Pixels and Regions [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(9): 7820-7829.
- [148] Azad R, Niggemeier L, Hüttemann M, et al. Beyond self-attention: Deformable large kernel attention for medical image segmentation[C]. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2024: 1287-1297.
- [149] Rahman M M, Munir M, Marculescu R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 11769-11779.
- [150] Chen L, Gu L, Zheng D, et al. Frequency-adaptive dilated convolution for semantic segmentation [C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 3414-3425.
- [151] Xiong X, He M, Li T, et al. Adaptive Feature Fusion and Improved Attention Mechanism-Based Small Object Detection for UAV Target Tracking[J]. *IEEE Internet of Things Journal*, 2024, 11(12): 21239-21249.

攻读学位期间取得的研究成果

I. 学术论文

- [1] **Weihuang Chen**, Zhigang Yang, Lingyang Xue, Jinghai Duan, Hongbin Sun, Nanning Zheng. Multimodal pedestrian trajectory prediction using probabilistic proposal network[J]. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2022. (SCI 1 区, IF: 5.859, DOI: 10.1109/TCSVT.2022.3229694)
- [2] **Weihuang Chen**, Fang Zheng, Liushuai Shi, Yongdong Zhu, Hongbin Sun, Nanning Zheng. Multiple goals network for pedestrian trajectory prediction in autonomous driving[C]. IEEE International Conference on Intelligent Transportation Systems (ITSC), 2022:717–722.
- [3] **Weihuang Chen**, Fangfang Wang, Hongbin Sun. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving[C]. Asian Conference on Machine Learning (ACML), 2021:454–469. (引用: 10)
- [4] **Weihuang Chen**, Yuwei Chen, Shen'ao Wang, Tianhang Li, Xuchong Zhang, Hongbin Sun. Motion planning using trajectory tree network for autonomous driving[J]. IEEE Transactions on Vehicular Technology (Tvt), 2023, Under review. (投稿号: VT-2023-00733)
- [5] Cheng Li, **Weihuang Chen**, Xinkai Luo, Fangfang Wang, Jingmin Zhang, Yanlong Yang, Hongbin Sun. Optimal preview distance control using model prediction for autonomous vehicle[C]. CAA International Conference on Vehicular Control and Intelligence (CVCI). 2021:1–8.

II. 专利

- [6] 孙宏滨、**陈炜煌**、王玉学、章浩飞、李煊、吴彝丹, 一种面向多场景的自动驾驶规划方法及系统 [P], 专利授权号: ZL202110276175.5

III. 科研获奖

- [7] 第一届全国研究生智能挑战赛, 三等奖, 2019 年。(队长)
- [8] 第六届中国研究生智慧城市技术与创意设计大赛, 二等奖, 2019 年。
- [9] 第十二届中国智能车未来挑战赛, 全国第 5 名, 发现号自动驾驶平台, 2020 年。(队长)

IV. 参与项目

- [10] 国家重点研发计划项目 (2018.05-2023.04): “下一代深度学习理论、方法与关键技术”(项目编号: 2017YFA0700800)
- [11] 国家自然科学基金重大项目 (2018.01-2022.12) : “极限工况下的人机协同机理及切换控制”(项目编号: 61790563)
- [12] 横向项目 (2021.03-2021.09): “基于深度学习的传感器数据融合”(项目编号: 202103136)

答辩委员会会议决议

轨迹预测与规划是自动驾驶领域的重要研究问题。论文开展了基于深度神经网络的轨迹预测和运动规划方法研究，选题具有重要的研究与应用价值。主要创新点如下：

1. 提出了一种基于时空 Transformer 网络的单模态轨迹预测模型，提升了密集交通环境下不同类别交通参与者的轨迹预测能力。
2. 提出了一种基于概率性候选轨迹网络的多模态轨迹预测模型，提高了交通参与者的多模态轨迹预测速度和精度。
3. 提出了一种基于安全轨迹树网络的运动规划模型，提高了自动驾驶车辆的运动规划性能。

论文写作认真，结构清晰，论述清楚，工作量饱满，表明作者已掌握本学科宽广坚实的基础理论和系统深入的专业知识，独立从事科研工作的能力强，是一篇高质量的博士学位论文。

答辩中讲述清晰，回答问题正确，经答辩委员会讨论和无记名投票表决，一致同意通过学位论文答辩，并一致建议授予陈炜煌同学工学博士学位。

常规评阅人名单

本学位论文共接受 3 位专家评阅，其中常规评阅人 2 名，名单如下：

魏平 教授 西安交通大学

邓成 教授 西安电子科技大学

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 交回学校授予的学位证书；
2. 学校可在相关媒体上对作者本人的行为进行通报；
3. 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。
4. 本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 学校可在相关媒体上对本人的失察行为进行通报；
2. 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。
3. 本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

（本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用）