无人机视角多源目标检测数据集 UAV-RGBT及算法基准

汪进中[1],戴 顺[1],张秀伟[1*],田雪涛[1,2],邢颖慧[1],汪 芳[1],尹翰林[1,3],张艳宁[1]
(1. 西北工业大学计算机学院,陕西西安 710072;2. 西安爱生技术集团有限公司,陕西西安 710065;
3. 西北工业大学深圳研究院,广东深圳 518063)

摘 要: 基于无人机(Unmanned Aerial Vehicle,UAV)平台的可见光(Red Green Blue,RGB)和热红外(Thermal infrared,T)多源目标检测,可实现全天时、全天候的目标侦察,在军用和民用领域有着重要的应用价值. 受限于数据拍摄获取和处理的复杂性,当前少有公开的UAV视角RGB-T多源目标检测数据集,一定程度上限制了UAV视角RGB-T多源目标检测算法的研究和应用. 与此同时,UAV应用场景复杂多变,其飞行高度、速度、焦距和背景等快速变化,所拍摄目标在图像上呈现出尺度多样、稠密/稀疏分布不均衡、类别不平衡等特点,具有一定的挑战性. 此外,在诸如目标侦察、交通监控等高时效性应用场景中,算法需在保证高精度的同时实现实时目标检测,因此,算法的设计必须充分考虑精度与速度之间的平衡. 针对上述问题,本文构建了一个跨季节、跨昼夜、多类别、多尺度的大规模UAV视角RGB-T多源图像数据集UAV-RGBT,包含20个类别、5 117对RGB-T图像和超11万个标注,有助于推进UAV视角多源目标检测算法的研究. 同时,基于YOLOv8n模型,本文提出了一种UAV视角多源目标检测(UAV-based Dual-branch Multispectral object Detection,UAV-DMDet)模型,其通过多源交叉注意力融合和多源特征分解组合方法有效促进了多源特征的深度融合,较好地实现了模型参数量、检测速度和检测精度的均衡. 实验结果表明:在UAV-RGBT数据集上,UAV-DMDet模型较单源YOLOv8n模型,在RGB和T模态方面,mAP@0.5分别提高了3.61%、11.03%,mAP@0.5:0.95分别提高了0.84%、6.76%;在DroneVehicle数据集上,mAP@0.5和mAP@0.5:0.95较主流算法I^2MDet提高了2.66%和12.36%;在检测速度方面,以640×640分辨率图像为例,UAV-DMDet模型在单张GeForce RTX 3090显卡上FP32精度推理速度可达31帧/s,在华为昇腾710处理器上FP16精度推理速度可达58帧/s,可有效应用于UAV视角RGB-T多源实时目标检测任务.

关键词: 无人机(UAV);可见光-热红外(RGB-T)多源目标检测;数据集;多源特征融合;YOLOv8
基金项目: 国家自然科学基金(No.61971356);陕西省自然科学基础研究计划(No.2024JC-DXWT-07,No.2024JC-YBQN-0719);陕西省重点研发计划(No.2023-YBGY-012);广东省基础与应用基础研究基金(No.2024A1515030186)
中图分类号: TP389.1;TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0686-19
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240602

UAV-RGBT Multispectral Object Detection Dataset and Algorithm Benchmark

WANG Jin-zhong[1], DAI Shun[1], ZHANG Xiu-wei[1*], TIAN Xue-tao[1,2], XING Yin-hui[1],
WANG Fang[1], YIN Han-lin[1,3], ZHANG Yan-ning[1]
(1. *School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an*, *Shaanxi* 710072, *China*;
2. *Xi'an ASN Technology Group Co., Ltd.*, *Xi'an*, *Shaanxi* 710065, *China*;
3. *Shenzhen Research Institute*, *Northwestern Polytechnical University*, *Shenzhen*, *Guangdong* 518063, *China*)

Abstract: Unmanned aerial vehicle (UAV)-based multispectral object detection utilizing both visible (RGB) and thermal infrared (T) images, makes all-weather and all-day target monitoring possible, serving critical roles in military and civilian applications. However, due to the complexity of data acquisition and processing, there is currently a lack of publicly available UAV-based RGB-T multispectral object detection datasets, which to some extent limits its research and application. Meanwhile, UAV operational scenarios are characterized by complex and variable conditions, including rapid changes

收稿日期:2024-06-26;修回日期:2024-09-29;责任编辑:朱梅玉
*通讯作者:张秀伟

in flight altitude, speed, focal length, and background. So, the captured targets exhibit diverse scales, uneven (dense/sparse) distributions, and category imbalances in images, which presents significant challenges for accurate detection. Furthermore, real-time requirement should be guaranteed in applications such as reconnaissance and traffic monitoring. Therefore, it is the key to keep a trade-off between accuracy and speed in the algorithmic design of UAV RGB-T object detector. To address these issues, this paper introduces a large-scale UAV-based RGB-T multispectral dataset named UAV-RGBT, which spans across seasons and day-night cycles, and includes multiple categories and scales. Specifically, UAV-RGBT comprises 20 categories with 5 117 pairs of RGB-T images and over 110 000 annotations, which is conducive to advancing research in UAV-based multispectral object detection algorithms. Moreover, based on the YOLOv8n model, the UAV-based dual-branch multispectral object detection (UAV-DMDet) model is proposed to promote deep fusion of multispectral features through a multi-modal cross-attention fusion module and a multi-modal feature decomposition combination module. This approach achieves a batter trade-off among model parameter size, detection speed, and accuracy. Experimental results demonstrate that the UAV-DMDet model improves the mAP@0.5 on the UAV-RGBT dataset by 3.61% and 11.03% in the visible and thermal modalities, respectively, and enhances the mAP@0.5:0.95 by 0.84% and 6.76%, respectively. On the DroneVehicle dataset, the UAV-DMDet model outperforms the mainstream algorithm I^2MDet, with mAP@0.5 and mAP@0.5:0.95 improvements of 2.66% and 12.36%, respectively. Furthermore, with 640 × 640 resolution images as input, the UAV-DMDet model achieve FP32 precision inference speed of 31 frames per second on a GeForce RTX 3090 GPU, and FP16 precision inference speed of 58 frames per second on a Huawei Ascend 710 processor, making it effectively applicable for real-time UAV-based RGB-T multispectral object detection tasks.

Key words: unmanned aerial vehicle (UAV); visible and thermal infrared multispectral object detection; dataset; multi-modal feature fusion; YOLOv8

Foundation Item(s): National Natural Science Foundation of China (No.61971356); Natural Science Basic Research Program of Shaanxi Province (No.2024JC-DXWT-07, No.2024JC-YBQN-0719); Key Research and Development Program of Shaanxi Province (No.2023-YBGY-012); Basic and Applied Basic Research Foundation of Guangdong Province (No.2024A1515030186)

1 引言

随着无人机(Unmanned Aerial Vehicle,UAV)和机载传感器技术的发展,使用UAV执行紧急搜救、线路巡检、环境侦察、交通监控等任务逐渐在军用和民用领域得到普及. 为了适应雨雾天气和夜间低照度等复杂气候环境,使用可见光(Red Green Blue,RGB)和热红外(Thermal infrared,T)传感器进行多源协同探测已成为一种典型应用范式. 与此同时,随着人工智能的发展,一大批基于深度学习的目标检测技术在工业检测、视频监控、自动驾驶等领域已经广泛应用,由于深度学习是一种数据驱动的技术,对于特定任务通常需要大量的训练数据,因此构建基于UAV视角的RGB-T多源目标检测数据集显得尤为重要.

在数据集方面,现有公开的RGB-T多源图像数据集大多针对自动驾驶[1]、行人检测[2]等任务而制定,由车载传感器或监控设备拍摄,与UAV视角差异较大. 而公开的UAV视角目标检测数据集大多只包含RGB或T单个模态,已知的UAV视角RGB-T多源数据集仅有文献[3]构建的DroneVehicle数据集,但其拍摄场景相对单一、拍摄高度和拍摄视角相对固定、仅标注了用于车辆检测的5个类别,应用场景有限. 对此,为探索复杂背景、多类目标的现实世界应用场景,迫切需要构建一个涵盖丰富场景、跨昼夜、多尺度的UAV视角RGB-T多源图像数据集,用于UAV视角多源协同目标检测方法研究.

在目标检测算法方面,基于深度学习的RGB-T目标检测算法大多从RGB单源目标检测算法演变而来,大体可分为基于卷积神经网络(Convolutional Neural Networks,CNN)的算法和基于Transformer的算法. 其中,基于CNN的算法包括基于回归的一阶段算法和基于区域候选的两阶段算法[4]. 比较有代表性的一阶段算法有YOLO(You Only Look Once)系列[5,6]和SSD(Solid State Drive)系列[7],检测速度较快. 两阶段的主流算法有更快的基于区域的CNN(Faster Region-based Convolutional Neural Networks,Faster R-CNN)[8]和级联R-CNN(Cascade Region-based Convolutional Neural Networks,Cascade R-CNN)[9],检测精度较高. 基于Transformer的代表算法有视觉变换器(Vision Transformer,ViT)[10]和检测变换器(DEtection TRansformer,DETR)[11],其不需要锚框等先验设计和非极大值抑制等后处理操作,实现了端到端的目标检测.

由于UAV应用场景的复杂性,其飞行过程中平台快速移动,光照条件、拍摄背景、拍摄视角等快速变化,所拍摄获取的目标大小不一、分布不均匀,对目标检测算法提出了较高的性能要求[4]. 与此同时,为实现跨昼夜、雨雾天气等恶劣条件下的目标检测,提取RGB和T图像特征并进行互补融合也一直是RGB-T多源目标检测算法的研究重点. 针对RGB和T模态在不同光照条件下的成像差

异，文献[3]提出了一种不确定感知跨模态检测器(Uncertainty-Aware Cross-Modality Detector，UA-CMDet)，通过计算RGB-T模态标注交并比(Intersection Over Union，IOU)来量化各模态的不确定性权重，同时设计了一种光照感知的跨模态非极大值抑制算法，以解决不同场景下不同模态预测的不确定性，降低模态差异场景下的误检率. 针对RGB-T特征跨模态校正和融合难的问题，文献[12]基于Transformer强大的相关性建模优势，提出了C^2Former网络，通过自适应特征采样策略和跨模态交叉注意力机制进行模态互补融合；文献[13]探索了卷积核尺寸、膨胀系数与特征提取频率之间的关系，设计了万花筒模块用于提取图像中不同频率特征信息并进行模态融合. 针对UAV视角RGB-T航拍图像存在的弱配准问题，文献[14]从辐射畸变和拍摄时差两个方面分析了问题原因，并设计了双流特征对齐检测器(Two-Stream Feature Alignment Detector，TSFADet)，通过平移、缩放和旋转操作来对齐和校准RGB-T特征图. 针对UAV航拍图像检测速度要求高的特点，文献[15]提出了一种多模态知识蒸馏(Multi-modal Knowledge Distillation，MKD)方法，在多模态特征互补融合的基础上，实现了轻量级模型的有效学习.

上述方法在模型精度、推理速度等方面有效地提升了UAV视角目标检测算法的性能，但由于UAV应用场景的复杂性和执行任务的时效性，较难实现检测性能和推理速度的均衡，仍然有进一步提高的空间. 针对上述问题，本文进行了如下创新.

(1)构建了一个大规模的UAV视角RGB-T多源图像数据集UAV-RGBT，该数据集具有跨季节、跨昼夜、多类别、多尺度、适用范围广等特点，包含5 117对RGB-T图像，并对人、各型车辆、路灯、电线杆、红绿灯、高压输电塔、桥梁、操场等20个类别进行了超11万个实例标注，可广泛应用于诸如智慧交通、环境侦察、无人驾驶航空器等UAV视角目标检测算法评估和研究工作.

(2)针对多源目标检测任务中多源特征互补融合难的问题，本文从高层语义交叉融合和低层特征分解组合两个方面设计了多源交叉注意力融合模块和多源特征分解组合模块. 其中，多源交叉注意力融合模块通过计算RGB和T特征之间的交叉注意力，引导高层多源语义特征进行交互融合和特征增强. 多源特征分解组合模块将低层RGB和T特征分别分解为基础特征和细节特征，通过不同形式的组合，在实现基础特征互补增强的同时，避免了不同模态细节特征表征差异所带来的噪声影响，较好地实现了RGB和T双分支的特征深度融合.

(3)立足RGB-T多源特征的差异性，基于轻量化的YOLOv8n[5]模型，本文设计了一个双分支UAV视角多源目标检测网络(UAV Dual-branch Multispectral object Detection，UAV-DMDet)，通过双分支多源特征提取和高低层多源特征融合，较好地实现了模型精度和检测速度的均衡. 实验表明：UAV-DMDet在UAV-RGBT数据集和DroneVehicle[3]数据集上均取得了最佳的检测性能，在640 × 640分辨率图像情况下，UAV-DMDet在GeForce RTX 3090单卡上达到了31帧/s的检测速度，在华为昇腾710处理器上达到了58帧/s的检测速度，且相关算法及模块可迁移应用于诸如YOLOv10[16]等其他算法框架，具有较好的通用性，可有效应用于UAV航拍图像的实时、高精度目标检测.

2 UAV视角RGB-T目标检测数据集构建

2.1 基于UAV视角的RGB-T目标检测数据集现状

截至目前，公开的UAV视角目标检测数据集大多只包含RGB或T单个模态，且重点针对交通监控[17,18]、车辆检测[19,20]等任务而设计，很少有包含RGB和T模态对齐的多源数据集，相关统计如表1所示.

表1中，UAV视角RGB目标检测数据集主要有7类，分别是文献[19]构建的停车场车辆计数(CAR ParKing counting，CARPK)数据集、文献[21]构建的无人机视觉目标检测(Visual detection of Drones-DETection，VisDrone-DET)数据集、文献[17]构建的UAV检测与跟踪(UAV for Detection and Tracking，UAVDT)数据集、文献[20]构建的异方差信号检测(Unequal Variance Signal Detection，UVSD)数据集、文献[18]构建的用于人工智能与机器人的无人机空中(Aerial UAV dataset for Artificial Intelligence and Robotics，AU-AIR)数据集、文献[22]构建的高分辨率UAV图像中的多尺度目标检测(Multi-scale Object detection in High Resolution UAV images，MOHR)数据集和文献[23]构建的用于车辆监控的航空影像(Vehicle Surveillance via Aerial Imagery，VSAI)数据集，大体情况如下.

(1)CARPK、UAVDT、UVSD和VSAI数据集仅标注了如小汽车、卡车和公共汽车等车辆目标，主要用于车辆计数和车辆检测相关研究. VisDrone-DET和AU-AIR数据集对自行车、摩托车、三轮车等车辆目标进行了细化，并标注了行人目标. MOHR数据集聚焦于自然场景，对建筑物、坍塌和洪涝灾害等场景进行了标注，可用于灾害预警、石油运输管线巡检等相关领域.

(2)由于RGB传感器对光照较为敏感，上述大部分数据集拍摄于白天或夜晚弱光照环境，缺少暗夜环境，多场景应用受限.

(3)VisDrone-DET、MOHR和VSAI数据集拍摄自多个城市，场景较为丰富，较好地表征了真实世界复杂场景的现状.

T数据集方面，仅有文献[24]构建的ASL-TID数据集和文献[25]构建的HIT-UAV数据集. 其中，ASL-TID数据集由T相机固定在高架平台，仿照UAV 10~30 m飞行高度

表1　基于UAV视角的目标检测数据集现状

数据集	模态	图像数量	类别/类	标注数量/×10³	分辨率	拍摄场景	拍摄时段	拍摄高度/m	拍摄视角	平均标注数量	最大标注数量	年份
CARPK[19]	RGB	1 448张	1(小汽车)	89.8	1 280×720	停车场	白天	40	—	62.00	188	2017
VisDrone-DET[21]	RGB	10 209张	10(行人和各型车辆)	471.2	2 000×1 500	城市、乡村	白天、夜晚	—	—	46.16	914	2018
UAVDT[17]	RGB	约80 000张	3(小汽车、卡车、公共汽车)	280.5	1 080×540	广场、主干道、收费站、高速公路等城市场景	白天、夜晚	—	—	10.52	—	2018
UVSD[20]	RGB	5 874张	1(车辆)	98.6	960×540至5 280×2 970	城市道路、住宅、停车场、高速公路和校园	白天	10~150	任意视角	16.78	—	2020
AU-AIR[18]	RGB	32 823张	8(行人及各型车辆)	132.0	1 920×1 080	十字路口等交通场景	白天	5~30	−90°~−45°(以水平方向为0°,垂直地面为−90°)	4.02	—	2020
MOHR[22]	RGB	10 631张	5(车辆、建筑和自然灾害等)	90.0	5 472×3 078 7 360×4 192 8 688×5 792	郊区、山区、雪地、沙漠	白天	200~400	—	8.47	—	2021
VSAI[22]	RGB	444张	2(小型车辆和大型车辆)	49.7	4 000×3 000 5 472×3 648 4 056×3 040	沙漠、城市、山地、郊区、河边和海边	白天、夜晚	55~500	−90°~0°	111.96	—	2022
ASL-TID[23]	T	4 381张	3(人、猫和马)	7.8	324×256	户外场景	白天、夜晚	10~30	—	1.78	—	2014
HIT-UAV[24]	T	2 898张	5(行人和各型车辆)	24.9	640×512	学校、停车场、道路、操场	白天、夜晚	60~130	−90°~−30°	8.59	—	2023
DroneVehicle[3]	RGB+T	28 439对	5(小汽车、面包车、公共汽车、卡车、厢式货车)	452.5(RGB) 500.5(T)	640×512	城市道路、住宅区、停车场、高速公路等	白天、夜晚、暗夜	80~120	−75°、−60°和−45°	17.59	206	2021
UAV-RGBT(本文)	RGB+T	5 117对	20(人、各型车辆、路灯、高压输电塔、桥梁、操场等)	57.9(RGB) 54.0(T)	1 920×1 080	校园、城市道路、高速公路、居民区、停车场、公园、机场、乡镇、农田、河流等10余种	白天、夜晚、暗夜	50~500	−90°~10°	11.53	136	2024

制作而成,仅对人类、猫和马3类目标进行了标注,属于UAV视角T目标检测的早期探索;HIT-UAV数据集由UAV在低空(60~130 m)拍摄,对人、车辆等5类目标进行了标注,可用于夜晚场景UAV搜救任务相关研究.

已知的UAV视角RGB-T多源目标检测数据集仅有2021年文献[3]构建的DroneVehicle数据集,该数据集使用大疆M200 UAV搭载禅思XT2摄像机拍摄,共包含28 439对RGB-T图像,场景涵盖城市道路、住宅小区、停车场和高速公路,标注包含小汽车、面包车、公共汽车、卡车和厢式货车共5类超95万个目标,拍摄时段包含白天、夜晚和暗夜无光照环境,拍摄高度介于80~120 m,拍摄视角固定为−75°、−60°和−45°三个视角,促进了复杂场景UAV

视角RGB-T车辆检测相关算法研究.

综合来看,上述UAV视角数据集还存在以下不足.

(1)模态单一. 现有的UAV视角数据集大多只包含RGB或T单一模态,由于RGB和T相机的成像机理差异,RGB相机在夜晚等弱光照条件下几乎失效;而T相机通过传感器来探测目标物体发射或反射的T辐射,在夜间成像方面具有明显的优势,但其成像缺乏色彩、纹理等细节信息,分辨率较低. 因此,通过构建RGB-T多源图像数据集,可充分利用RGB和T图像优势,有效提升复杂条件下的目标检测能力.

(2)目标/场景/采集方式相对单一. 大部分UAV视角数据集针对特定目标类别而采集,拍摄场景相对单一. 同时,采用相对固定的飞行高度和俯仰角进行拍摄,所获取的目标尺度相对固定. 这种聚焦于单一场景或目标类别的数据集,限制了其在真实世界复杂场景下的适用性.

针对上述不足,本文使用UAV采集获取了多个城市、不同背景、多种飞行高度、不同拍摄视角、不同目标尺度且跨昼夜的RGB-T多源目标检测数据集,称之为UAV-RGBT,其具体构建过程包括数据采集、数据预处理和数据标注等阶段.

2.2 数据采集

UAV视角目标检测通常面临飞行高度动态变化、拍摄焦距按需调整、恶劣气候及跨昼夜运行等应用场景,对数据采集提出了较高的要求. 对此,选用大疆M300RTK高性能四旋翼UAV,搭载禅思H20N RGB-T外相机进行数据拍摄,两种型号的设备如图1所示.

立足真实场景应用需求,数据采集时,UAV飞行高度介于50~500 m,相机俯仰角介于−90°~10°,相机焦距

(a) 大疆M300RTK UAV (b) 禅思H20N相机

图1 数据采集设备示意

在1倍、2倍和4倍之间切换,确保了所获取目标的尺度多样性. 拍摄场景包括校园、城市道路、高速公路、居民区、停车场、公园、机场、乡镇、农田、河流等. 拍摄时段包含白天、夜晚和暗夜无光照环境,同时横跨春、夏、秋、冬4个季节,基本覆盖了UAV多用途应用环境,相关示例如图2所示. 与HIT-UAV等数据采集方式不同,本文未采用拍摄视频后抽取关键帧的形式,而是由飞手针对场景特点,逐张拍摄感兴趣目标,确保了拍摄质量,同时也更加契合UAV实际应用特点. 最终,拍摄获取了5 117对基于UAV视角的RGB-T高分辨率图像.

2.3 数据预处理

通过禅思H20N相机拍摄获取的RGB图像分辨率为2 688 × 1 512像素,T图像分辨率为640 × 512像素. 由于RGB和T相机成像差异,如图3所示,拍摄获取的RGB-T图像视场范围未能严格对齐;加之UAV飞行处于运动状态,导致获取的RGB-T图像存在位置和角度偏差[14]. 针对上述问题,本文使用文献[26]提出的RGB-T图像配准方法,对RGB-T图像进行了配准处理,并将RGB-T图像共视区域统一缩放至1 920 × 1 080像素,确保了图像的高分辨率.

图2 UAV-RGBT数据集RGB-T多场景示意(图中矩形框为目标标注框)

图3　UAV-RGBT数据集RGB-T图像视场差异及预处理流程示意

2.4　数据标注

由于RGB和T成像机理的差异性,RGB图像原始分辨率较高,对小目标区分度较好;而T图像分辨率较差,但对夜晚场景表示能力较强. 因此,使用唯一标注来同时表征RGB和T图像的目标分布是不合理的. 对此,本文使用LabelImg工具对RGB图像和T图像分别进行了标注,使用水平框(左上角和右下角x,y坐标)表示目标位置信息;对于遮挡或截断(物体位于图像边界,部分处于图像外)情况,遮挡或截断比例超过物体本身2/3的目标则不进行标注. 最终,立足现实世界目标检测需要,对人、小汽车、公共汽车、面包车、卡车、半挂车、厢式货车、工程车、自行车、三轮车、骑车的人、路灯、电线杆、红绿灯、高压输电塔、桥梁、立交桥、篮球场、羽毛球场和操场共20个常见类别进行了标注,共得到RGB标注57 879个、T标注54 009个,各类别示意如图4所示,标注数量统计如表2所示.

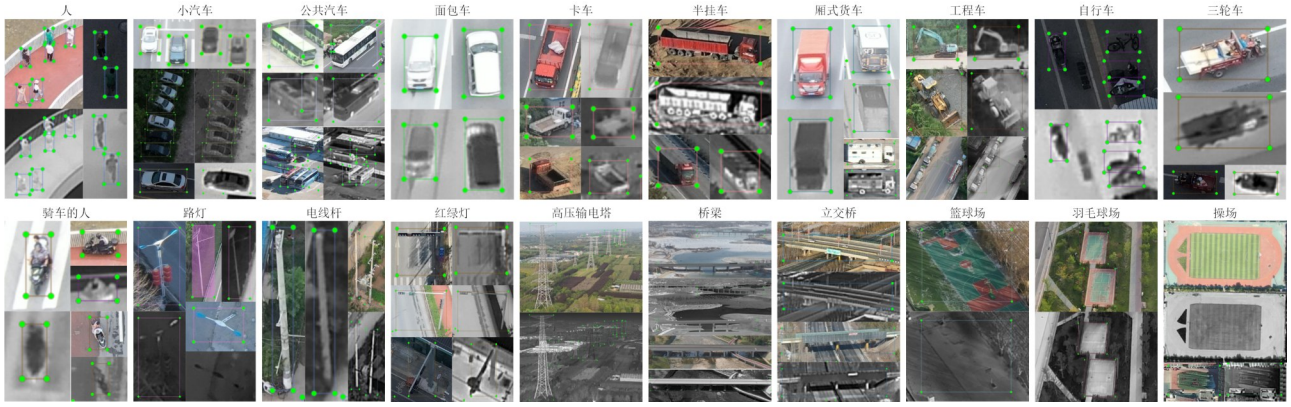


图4　UAV-RGBT数据集20个类别RGB-T图像示意

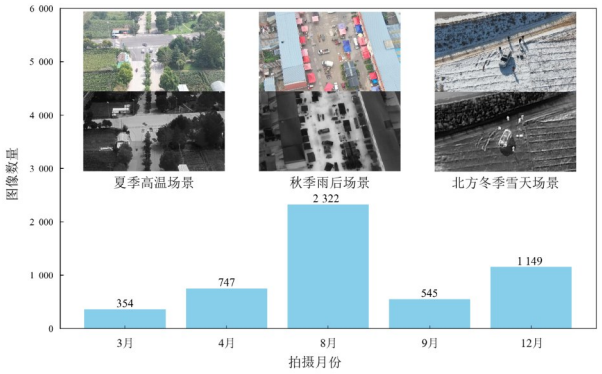
表2　UAV-RGBT数据集标注类别数量统计

模态	人	小汽车	公共汽车	面包车	卡车	半挂车	厢式货车	工程车	红绿灯	高压输电塔
RGB	4 477	29 397	1 042	2 173	2 506	1 138	478	552	473	815
T	3 723	29 178	1 036	2 021	2 456	1 092	480	514	461	861
模态	自行车	三轮车	骑车的人	路灯	电线杆	桥梁	立交桥	羽毛球场	篮球场	操场
RGB	2 159	850	1 986	6 006	3 105	142	65	244	237	34
T	1 705	754	1 810	4 384	2 803	154	63	244	232	38

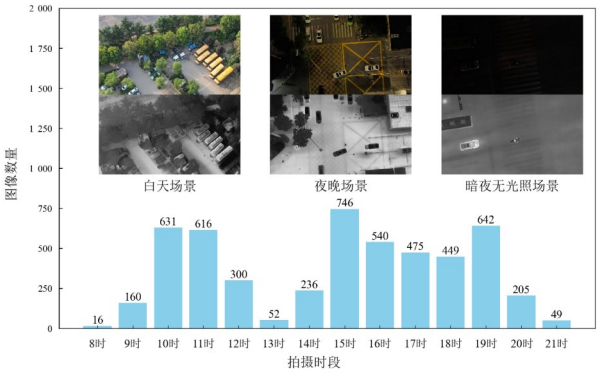
2.5　统计分析

UAV-RGBT数据集具有跨季节、跨昼夜、多场景和多尺度等特点:(1)在季节和气候方面,如图5(a)所示,所拍摄数据横跨春、夏、秋、冬4个季节,包含晴天、阴天、雨天、雪天等多种气候条件,基本涵盖了西北地区全年常见气候条件;(2)在数据获取时间方面,如图5(b)所示,拍摄时段覆盖8时~21时,包含白天、夜晚和暗夜无光照等多种光照条件,凸显了不同时段RGB和T图像的差异性,为

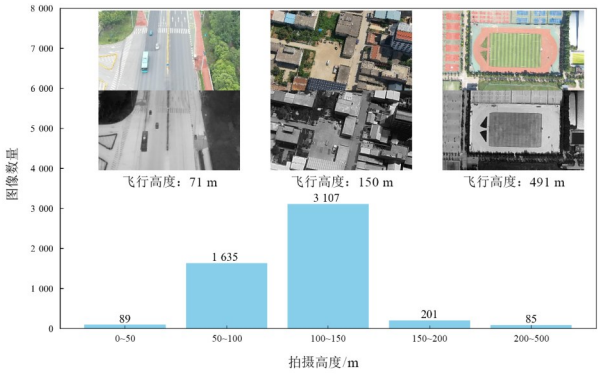
RGB-T多源融合检测提供了较好的数据基础;(3)在场景分布方面,如图2所示,包含校园、城市道路、高速公路、居民区、停车场、公园、机场、乡镇、农田、河流等10余种拍摄场景,极大地丰富了数据的多样性;(4)在目标尺度分布方面,如图5(*c*)和图5(*d*)所示,拍摄高度覆盖50~500 m不同高度,相机俯仰角介于-90°~10°,从不同高度和不同视角采集了目标多尺度图像. 表3统计了各类别最小尺寸(宽高)、最大尺寸(宽高)以及最大最小面积比等尺度信息,除自行车、篮球场和操场外,其他类别最大最小面积比均超过100倍. 同时,图6绘制了各类别目标尺寸(面积)分布的箱型图,可以看出各类目标尺寸分布差异较大,呈现多尺度特性.



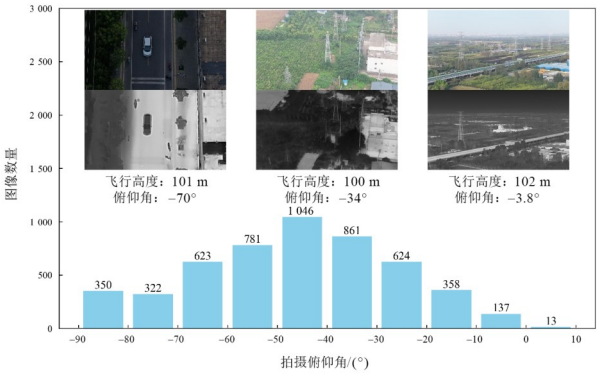
(*a*) 数据拍摄月份分布



(*b*) 数据拍摄时段分布



(*c*) 数据拍摄高度分布



(*d*) 数据拍摄俯仰角分布

图5 UAV-RGBT数据集数据分布情况

3 UAV-DMDet多源目标检测算法

3.1 UAV-DMDet算法总体框架

以往的相关研究[12]通常使用双分支骨干网络和单一检测头对RGB-T图像进行检测,这种使用单一检测结果表征RGB-T两个模态目标分布的方法存在一定争议,例如在暗夜无光照条件下,RGB模态失效,图像呈现纯黑样式,用T的检测结果来指示目标在RGB图像中的位置是不合理的,对实际应用存在一定误导. 此外考虑到UAV目标检测高时效性的应用需求,本文基于轻量化的单源YOLOv8n[5]模型,构建了双分支跨模态目标检测算法UAV-DMDet. 如图7所示,UAV-DMDet由骨干网络(backbone)、颈部网络(neck)和检测头(head)三部分组成,使用RGB和T并行的双分支结构,有效避免了使用单一检测结果表征两个模态属性所导致的二义性.

在骨干网络中,ConvModule和C2f等卷积层用于提取目标多尺度特征,其中P1~P5分别为输入图像的2倍、4倍、8倍、16倍和32倍下采样特征图. 对于P5尺度的特征,本文设计了多源交叉注意力融合(Multi-modal Cross Attention Fusion,MCAF)模块,通过计算RGB-T特征相似度,实现对RGB-T特征的跨模态融合. 颈部网络使用了路径聚合网络(Path Aggregation Network,PANet[27]),对来自骨干网络的P3~P5这3个尺度特征进行自顶向下和自底向上的多路径聚合,以实现不同尺度的特征融合. 同时,针对双分支结构需要,本文设计了多源特征分解组合(Multi-modal Feature Decomposition Combination,MFDC)模块,将RGB和T图像分别拆解为基础特征和细节特征,之后将4类特征进行不同形式的组合,并分别送入RGB和T分支颈部进行特征融合和增

表3　UAV-RGBT数据集各类别尺寸信息统计(宽、高像素)

类别	最小尺寸	最大尺寸	面积比	类别	最小尺寸	最大尺寸	面积比
人	5×8	98×156	382.20	自行车	12×21	79×207	64.89
小汽车	9×7	283×391	1 756.40	三轮车	13×27	273×206	160.22
公共汽车	18×25	689×589	901.82	骑车的人	5×11	141×108	276.87
面包车	19×13	269×255	277.71	路灯	8×32	378×641	946.48
卡车	36×10	396×638	701.80	电线杆	8×34	174×906	579.57
半挂车	24×38	915×529	530.74	桥梁	72×25	1 673×1 074	998.22
厢式货车	18×20	1 263×429	1 505.08	立交桥	117×25	1 755×600	360.00
工程车	19×22	702×291	488.71	羽毛球场	189×86	1 794×1 075	118.65
红绿灯	27×29	743×371	352.05	篮球场	85×36	253×305	25.22
高压输电塔	12×35	308×1 017	745.80	操场	220×84	1 466×1 020	80.92

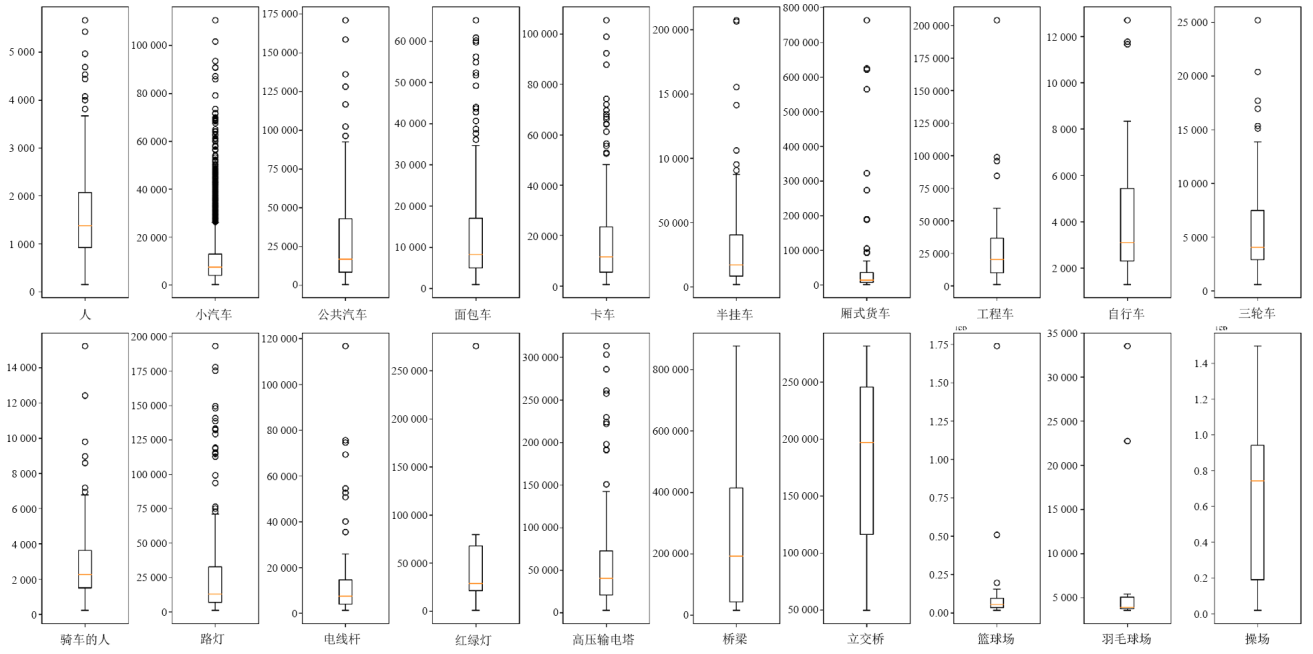


图6　UAV-RGBT数据集20类目标尺寸分布箱型图

强. 检测头使用解耦的方法,利用3组不同大小的检测器,分别对P3~P5这3个尺度的特征图进行目标位置预测和目标类别预测. 此外,在训练阶段,本文分别使用RGB和T标注对双分支网络进行共同监督.

3.2　基础模块

UAV-DMDet算法基础结构和其中的ConvModule和C2f等卷积层来自YOLOv8算法. YOLOv8是Ultralytics公司于2023年1月发布并维护的算法模型,其整合了前期YOLO系列算法的优势,同时支持图像分类、目标检测、实例分割、目标跟踪和关键点检测等计算机视觉任务,算法精度在多个任务中处于领先地位. 根据模型深度和宽度差异,按照参数量从小到大,YOLOv8算法包含n、s、m、l和x共5个版本,模型越大,精度越高,其计算时效性也越差.

ConvModule、C2f、SPPF和检测头等模块网络结构如图8所示. 其中,ConvModule模块采用Conv2d卷积、BatchNorm2d和SiLU激活函数堆叠的形式,可对输入特征进行降采样,在减小特征图尺寸的同时增加通道数,并增强模型的非线性表示能力. C2f模块借鉴了CSP和ELAN结构,使用跳层连接和Split操作,并根据模型大小对DarknetBottleneck模块进行多次堆叠,在减少计算量和内存消耗的同时具有更优秀的特征提取能力;空间金字塔池化融合(Spatial Pyramid Pooling Fusion,SPPF)模块使用多个MaxPool2d串行结构,可以在不同尺度的特征图上进行特征提取,提高了网络的感受野和特征表达能力. 检测头采用了无锚的解耦头结构,使用回归分支和预测分支分别预测目标位置和目标类别,来自颈部的特征分别通过3×3卷积和1×1卷积提取信息,预测并计算目标位置损失(BboxLoss)和类别损失(ClsLoss).

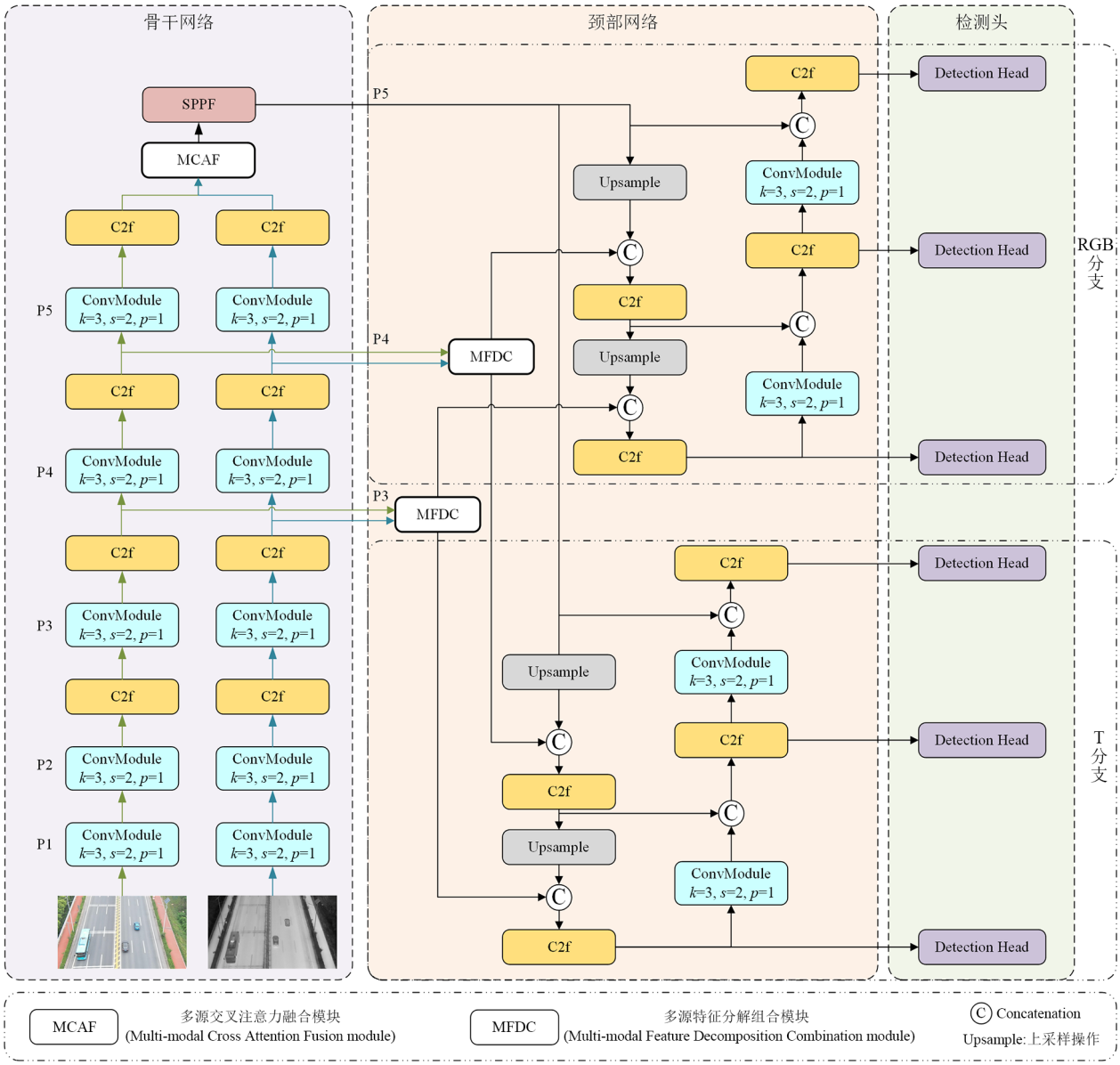


图 7 UAV-DMDet 算法网络结构图

3.3 多源交叉注意力融合模块

在双分支骨干网络中，为实现 RGB 和 T 特征的融合，以往的相关研究通常使用通道拼接(Concatenation)[3]或逐元素相加(Addition)[12,14]的方式，上述方法仅实现了 RGB 和 T 特征的组合，未能充分考虑到模态之间的差异性. 受交叉注意力机制的启发，本文提出了一种 MCAF，通过跨模态特征交互和多源特征增强两个阶段，计算 RGB 和 T 模态间的相关性，并实现多源特征的充分融合.

如图 9 所示，跨模态特征交互部分由 RGB(左侧)和 T(右侧)两个分支构成，各包含一个多头注意力机制模块(Multi-Head Attention，MHA)[28]. 在 RGB 分支中，T 特征 $\boldsymbol{f}_{\mathrm{ir}}$ 作为查询向量 $\boldsymbol{Q}$，RGB 特征 $\boldsymbol{f}_{\mathrm{vis}}$ 作为键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 计算模态间交叉注意力，目的是使用 T 特征引导 RGB 特征来捕获模态间的相关性，之后对 RGB 特征 $\boldsymbol{f}_{\mathrm{vis}}$ 进行残差连接得到 $\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{att}}$；与 RGB 分支类似，在 T 分支中，RGB 特征 $\boldsymbol{f}_{\mathrm{vis}}$ 作为查询向量 $\boldsymbol{Q}$，T 特征 $\boldsymbol{f}_{\mathrm{ir}}$ 作为键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 进行交叉注意力融合，并对 T 特征 $\boldsymbol{f}_{\mathrm{ir}}$ 进行残差连接得到 $\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{att}}$，两个分支计算过程如式(1)和式(2)所示. 在多源特征增强阶段，如式(3)和式(4)所示，跨模态特征 $\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{att}}$ 和 $\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{att}}$ 首先在通道维度拼接并经过层归一化(Layer Normalization，LN)[29]，之后融合特征 $\boldsymbol{f}_{\mathrm{att}}$ 经过多头注意力机制和前馈网络(Feed-forward Network，FN)[28]进行特征增强并建立残差连接，最终得到跨模态增强的融合特征 $\boldsymbol{f}'_{\mathrm{att}}$.

$$\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{att}}=\boldsymbol{f}_{\mathrm{vis}}+\mathrm{MHA}(\boldsymbol{f}_{\mathrm{ir}},\boldsymbol{f}_{\mathrm{vis}},\boldsymbol{f}_{\mathrm{vis}}) \tag{1}$$

$$\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{att}}=\boldsymbol{f}_{\mathrm{ir}}+\mathrm{MHA}(\boldsymbol{f}_{\mathrm{vis}},\boldsymbol{f}_{\mathrm{ir}},\boldsymbol{f}_{\mathrm{ir}}) \tag{2}$$

$$\boldsymbol{f}_{\mathrm{att}}=\mathrm{LN}\left(\mathrm{Concat}(\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{att}},\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{att}})\right) \tag{3}$$

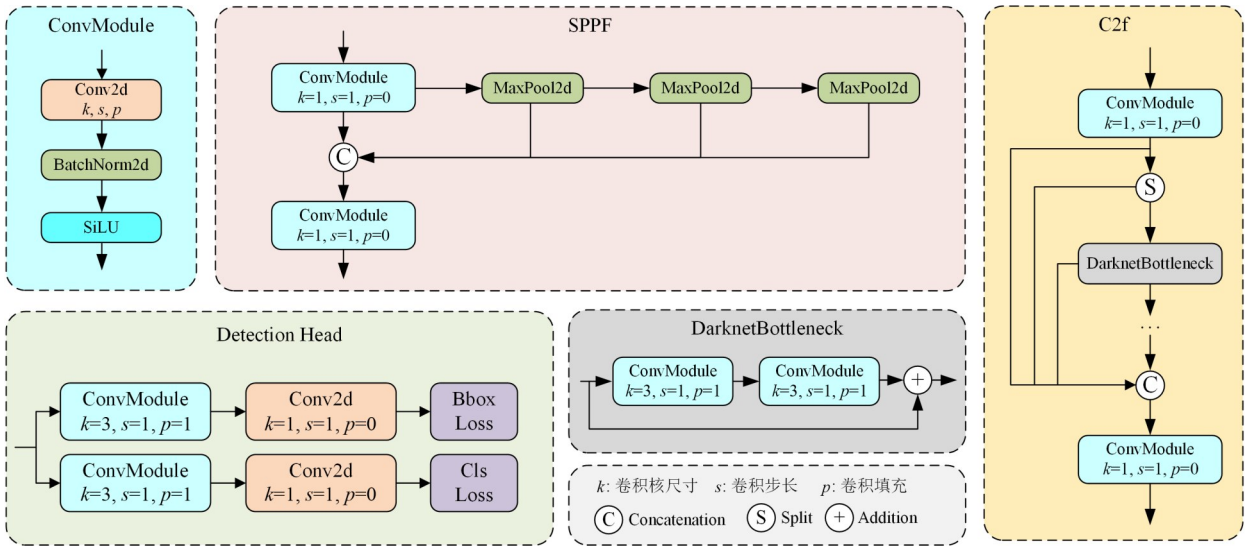


图8　UAV-DMDet算法基础模块结构图

$$\boldsymbol{f}'_{\mathrm{att}}=\boldsymbol{f}_{\mathrm{att}}+\mathrm{FN}\left(\boldsymbol{f}_{\mathrm{att}}+\mathrm{MHA}(\boldsymbol{f}_{\mathrm{att}})\right) \tag{4}$$

MHA计算过程如图9所示,首先利用3个线性变换矩阵将输入特征转换为查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$,之后通过点积对$\boldsymbol{Q}$和$\boldsymbol{K}$中的每个元素进行相似性匹配,通过Softmax归一化后与值向量$\boldsymbol{V}$加权得到注意力输出. 多头注意力机制是指使用多个不同的可学习线性映射函数将$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$线性映射到不同的嵌入空间,然后对映射到不同空间的向量分别应用注意力机制,最后进行特征拼接和线性投影. FN由两个线性层和高斯误差线性单元(Gaussian Error Linear Unit,GELU)激活函数组成,目的是进一步提取特征,并提高模型的泛化能力.

3.4　多源特征分解组合模块

图像的频率特征包括低频特征和高频特征两类. 其中,低频特征表示图像中亮度或灰度变化缓慢的区域,通常包含图像的全局结构和整体形状等基础特征;而高频特征表示图像中相邻像素之间灰度差异大,即频率变化快的区域,通常包含图像的细节成分及噪声等细节特征. RGB-T多源图像拍摄自同一场景,但由于传感器成像机理差异,RGB-T图像在低频信息上表现出一定的相关性,均包含了如背景、大尺度目标以及布局等共性特征;而在高频信息上相互独立,代表着各自模态的细节特征,如RGB图像中的色彩、纹理和细节等信息,以及T图像中的热辐射信息[30].

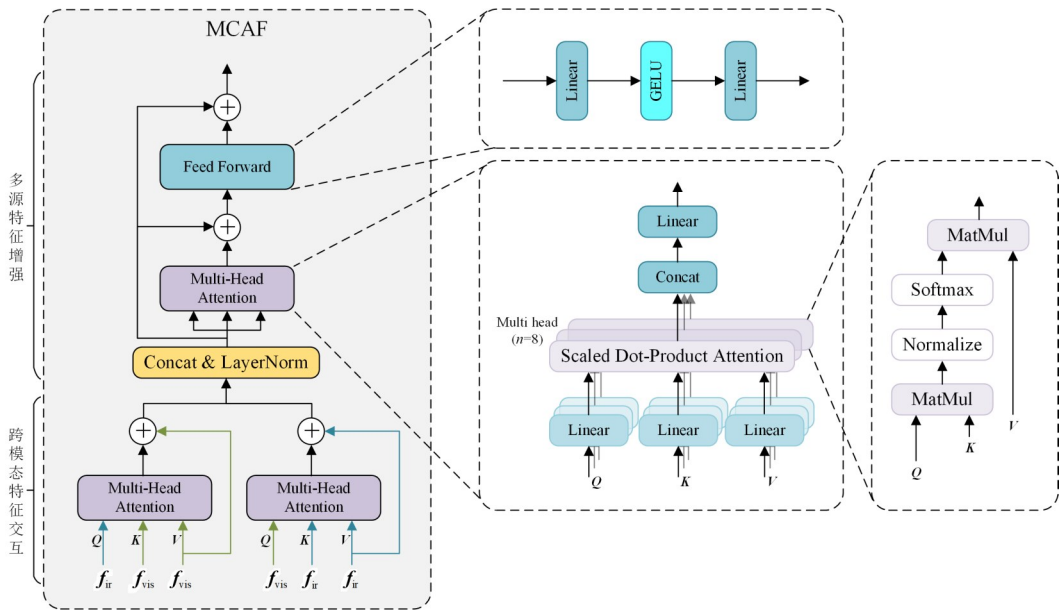


图9　MCAF网络结构图

若能将RGB-T多源图像分解为低频基础特征和高频细节特征，并将某一模态的低频基础特征引入另一模态，则可提高另一模态特征的丰富度和准确度. 受文献[30]的启发，本文提出了一种MFDC模块，通过将RGB-T图像特征分解为基础特征和细节特征，并在RGB-T双分支颈部进行不同形式的互补组合，来促进两个分支的跨模态特征融合.

多源特征分解组合模块网络结构如图10所示，主要由基础特征提取器(Base Feature Extractor，BFE)和细节特征提取器(Detail Feature Extractor，DFE)两部分组成. RGB特征$\boldsymbol{f}_{\text{vis}}$和T特征$\boldsymbol{f}_{\text{ir}}$首先分别经过基础特征提取器和细节特征提取器，得到RGB基础特征$\boldsymbol{f}_{\text{vis}}^{B}$、RGB细节特征$\boldsymbol{f}_{\text{vis}}^{D}$、T基础特征$\boldsymbol{f}_{\text{ir}}^{B}$和T细节特征$\boldsymbol{f}_{\text{ir}}^{D}$；之后将上述4类特征进行组合，得到融合特征$\boldsymbol{f}_{\text{vis}}^{f}$和$\boldsymbol{f}_{\text{ir}}^{f}$，并分别送入RGB分支颈部和T分支颈部进行特征融合和增强.

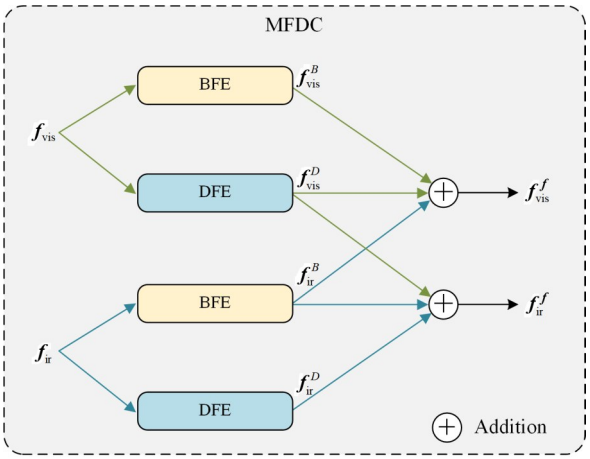


图10 MFDC模块网络结构图

如式(5)和式(6)所示，RGB分支融合特征$\boldsymbol{f}_{\text{vis}}^{f}$由RGB基础特征$\boldsymbol{f}_{\text{vis}}^{B}$、T基础特征$\boldsymbol{f}_{\text{ir}}^{B}$和RGB细节特征$\boldsymbol{f}_{\text{vis}}^{D}$相加得到，T分支融合特征$\boldsymbol{f}_{\text{ir}}^{f}$由RGB基础特征$\boldsymbol{f}_{\text{vis}}^{B}$、T基础特征$\boldsymbol{f}_{\text{ir}}^{B}$和T细节特征$\boldsymbol{f}_{\text{ir}}^{D}$相加得到. 上述方法在当前模态中引入另一模态的基础特征，丰富了全局、结构等低频基础特征，实现了跨模态的特征融合；同时考虑到两个模态细节特征之间的差异性以及高频细节特征可能包含噪声等无效信息，避免了引入另一模态细节特征可能导致的特征冲突和噪声干扰等问题.

$$\boldsymbol{f}_{\text{vis}}^{f}=\boldsymbol{f}_{\text{vis}}^{B}+\boldsymbol{f}_{\text{ir}}^{B}+\boldsymbol{f}_{\text{vis}}^{D} \tag{5}$$

$$\boldsymbol{f}_{\text{ir}}^{f}=\boldsymbol{f}_{\text{vis}}^{B}+\boldsymbol{f}_{\text{ir}}^{B}+\boldsymbol{f}_{\text{ir}}^{D} \tag{6}$$

3.4.1 基础特征提取器(BFE)

考虑到Transformer注意力机制相较CNN拥有全局感受野的优势，且能够建模特征的长距离依赖关系，如图11所示，本文使用较为高效的Restormer[31]模块作为BFE. BFE主要包含深度卷积多头转置注意力模块(Multi-Dconv head Transposed Attention，MDTA)[31]和深度卷积门控FN(Gated-Dconv Feed-forward Network，GDFN)[31]两个子模块. 如式(7)所示，输入特征$\boldsymbol{f}_{t-1}\in\mathbb{R}^{H\times W\times C}$首先经过层归一化，然后经过MDTA模块进行局部和全局的跨通道特征增强，通过残差连接后得到$\boldsymbol{f}_{t}$；如式(8)所示，$\boldsymbol{f}_{t}$首先经过层归一化，然后通过GDFN进行门控特征增强，残差连接后得到基础特征$\boldsymbol{f}_{\text{BFE}}$.

$$\boldsymbol{f}_{t}=\text{MDTA}\left(\text{LN}(\boldsymbol{f}_{t-1})\right)+\boldsymbol{f}_{t-1} \tag{7}$$

$$\boldsymbol{f}_{\text{BFE}}=\text{GDFN}\left(\text{LN}(\boldsymbol{f}_{t})\right)+\boldsymbol{f}_{t} \tag{8}$$

MDTA通过将深度可分离卷积(Dconv)和自注意力机制[28]相结合，在利用深度可分离卷积计算局部注意力的同时，使用自注意力机制计算跨通道的协方差以生成全局注意力. 以输入特征$\boldsymbol{f}\in\mathbb{R}^{H\times W\times C}$为例，首先通过$1\times1$卷积将特征通道升维到$3C$，然后使用深度可分离卷积在丰富局部特征的同时生成$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$向量；$\boldsymbol{Q}$和$\boldsymbol{K}$重塑尺寸后通过点积形式计算通道相似性，生成跨通道的转置注意力特征图$\boldsymbol{A}\in\mathbb{R}^{C\times C}$，并将$\boldsymbol{A}$作用于$\boldsymbol{V}$后使用$1\times1$卷积映射输出.

与常规FN不同，GDFN通过引入深度可分离卷积和门控机制提高了模型表征学习能力. 以输入特征$\boldsymbol{f}\in\mathbb{R}^{H\times W\times C}$为例，首先通过$1\times1$卷积拓展特征通道为$2C'$，然后使用深度可分离卷积丰富局部特征并拆分为上下两组平行路径的特征，其中一组特征使用GELU非线性激活后与另一组特征进行逐元素乘积形成门控机制，最后使用1×1卷积对通道维度进行复原.

3.4.2 细节特征提取器(DFE)

针对特征提取过程中细节特征可能丢失的问题，受文献[30]的启发，本文采用可逆神经网络[32,33]作为细节特征提取器. 可逆神经网络具有双射构造和高效可逆的特点，可以被视为一个无损的特征提取模块. 如图12所示，本文使用的仿射耦合可逆神经网络包含局部分支(上分支)和长程分支(下分支)两个部分，输入特征$\boldsymbol{f}_{i,0}\in\mathbb{R}^{H\times W\times C}$在通道维度首先被拆分为局部特征$\boldsymbol{g}_{i,0}$和长程特征$\boldsymbol{l}_{i,0}$以作为两个可逆结点，$\boldsymbol{g}_{i,0},\boldsymbol{l}_{i,0}\in\mathbb{R}^{H\times W\times C/2}$，其中，$i$代表RGB或T模态$i\in(\text{vis},\text{ir})$. 如式(9)所示，对于长程分支，本文使用加法变换得到结点特征$\boldsymbol{l}_{i,1}$；如式(10)所示，对于局部分支，使用增强仿射变换得到$\boldsymbol{g}_{i,1}$，最后将局部特征$\boldsymbol{g}_{i,1}$和长程特征$\boldsymbol{l}_{i,1}$在通道维度拼接得到细节特征$\boldsymbol{f}_{i,1}$.

$$\boldsymbol{l}_{i,1}=\boldsymbol{l}_{i,0}+F_1(\boldsymbol{g}_{i,0}) \tag{9}$$

$$\boldsymbol{g}_{i,1}=\boldsymbol{g}_{i,0}\odot\exp\left(F_2(\boldsymbol{l}_{i,1})\right)+F_3(\boldsymbol{l}_{i,1}) \tag{10}$$

$$\boldsymbol{f}_{i,1}=\text{Concat}\left(\boldsymbol{l}_{i,1},\boldsymbol{g}_{i,1}\right) \tag{11}$$

其中，exp(·)为指数函数；$\odot$为哈达玛德积(逐元素乘积)；F_1、F_2和F_3为任意映射函数，为均衡特征提取能力和计算时效性，本文采用了类似MobileNetV3[34]的网络模块. 如

图12所示,本文去除了MobileNetV3模块中的归一化层,由1×1卷积、ReLU激活函数、3×3深度可分离卷积以及压缩-激发注意力(SEmodule,SE)[35]等模块进行特征提取.以输入特征$\boldsymbol{f}\in\mathbb{R}^{H\times W\times C}$为例,映射函数$F$处理过程如式(12)所示;以输入特征$\boldsymbol{f}_{\mathrm{SE}}^{\mathrm{in}}\in\mathbb{R}^{H\times W\times C}$为例,压缩-激发注意力模块处理过程SE(·)如式(13)所示.其中,$W_p^{(\cdot)}$为1×1卷积,$W_d^{(\cdot)}$为深度可分离卷积,σ为ReLU激活函数,Avg为平均池化,ρ为Hardsigmoid激活函数.

$$\boldsymbol{f}_F=\sigma\left(W_p^1\left(\mathrm{SE}\left(\sigma\left(W_d^0\left(\sigma\left(W_p^0(\boldsymbol{f})\right)\right)\right)\right)\right)\right) \tag{12}$$

$$\boldsymbol{f}_{\mathrm{SE}}^{\mathrm{out}}=\boldsymbol{f}_{\mathrm{SE}}^{\mathrm{in}}\odot\left(\rho\left(W_p^3\left(\sigma\left(W_p^2\left(\mathrm{Avg}(\boldsymbol{f}_{\mathrm{SE}}^{\mathrm{in}})\right)\right)\right)\right)\right) \tag{13}$$

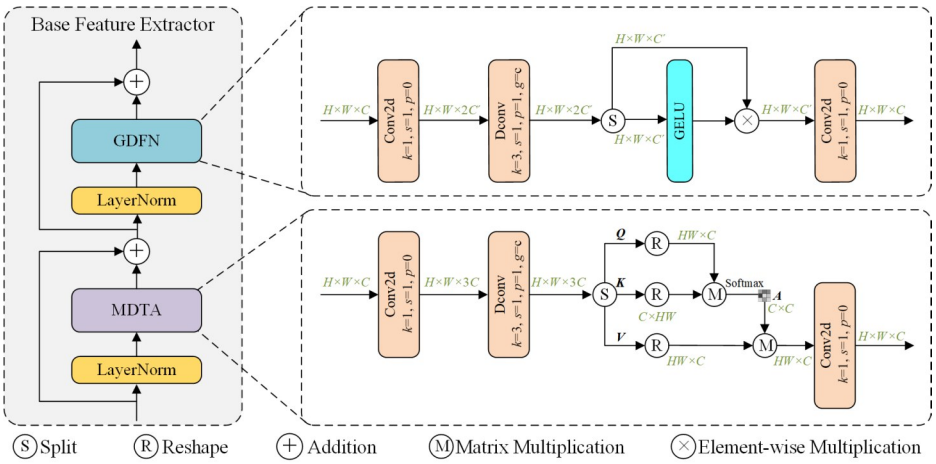


图11　BFE网络结构图

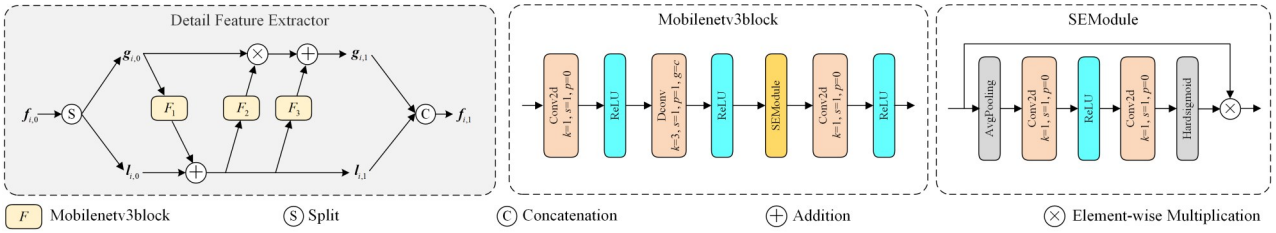


图12　DFE网络结构图

3.5　损失函数

如式(14)所示,UAV-DMDet损失函数由分类损失$\mathcal{L}_{\mathrm{cls}}$、回归损失$\mathcal{L}_{\mathrm{box}}$、分布聚焦损失$\mathcal{L}_{\mathrm{dfl}}$和多源特征分解损失$\mathcal{L}_{\mathrm{decomp}}$构成,$i\in(\mathrm{vis},\mathrm{ir})$代表RGB和T模态.

$$\begin{aligned}\mathcal{L}_{\mathrm{total}}=&\lambda_{\mathrm{cls}}*\sum_i\mathcal{L}_{\mathrm{cls}}^i+\lambda_{\mathrm{box}}*\sum_i\mathcal{L}_{\mathrm{box}}^i\\&+\lambda_{\mathrm{dfl}}*\sum_i\mathcal{L}_{\mathrm{dfl}}^i+\lambda_{\mathrm{decomp}}*\mathcal{L}_{\mathrm{decomp}}\end{aligned} \tag{14}$$

其中,分类损失$\mathcal{L}_{\mathrm{cls}}$为二元交叉熵损失.回归损失$\mathcal{L}_{\mathrm{box}}$为完整交并比(Complete Intersection over Union,CIoU)损失,其计算过程如式(15)~式(18)所示,IoU为预测框B^{pred}和真实框B^{gt}的交并比,$\rho^2(B^{\mathrm{pred}},B^{\mathrm{gt}})$为预测框和真实框中心坐标的欧氏距离,$c$为同时包含预测框和真实框的最小闭包区域对角线距离,v用于度量预测框和真实框长宽比的相似性,α为平衡参数.

$$\mathrm{CIoU}=\mathrm{IoU}-\left(\frac{\rho^2\left(B^{\mathrm{pred}},B^{\mathrm{gt}}\right)}{c^2}+\alpha v\right) \tag{15}$$

$$v=\frac{\pi}{4}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w^{\mathrm{pred}}}{h^{\mathrm{pred}}}\right) \tag{16}$$

$$\alpha=\frac{v}{(1-\mathrm{IoU})+v} \tag{17}$$

$$\mathrm{IoU}=\frac{\left|B^{\mathrm{pred}}\cap B^{\mathrm{gt}}\right|}{\left|B^{\mathrm{pred}}\cup B^{\mathrm{gt}}\right|} \tag{18}$$

如式(19)所示,分布聚焦损失(Distribution Focal Loss,DFL)将预测框的坐标位置建模为概率分布,以交叉熵的形式对距离真实框坐标y最近的两个位置y_i和y_{i+1}的概率S_i和S_{i+1}进行优化,目的是让网络快速聚焦于目标位置的邻近区域分布,概率密度尽可能靠近目标位置.

$$\begin{aligned}\mathrm{DFL}(S_i,S_{i+1})=&-(y_{i+1}-y)\ln(S_i)\\&+(y-y_i)\ln(S_{i+1})\end{aligned} \tag{19}$$

在多源特征分解组合模块中,为了实现基础特征和细节特征的分解,本文引入了多源特征分解损失$\mathcal{L}_{\mathrm{decomp}}$[30],目的是增加多源基础特征之间的相关性,减少多源细节特征之间的相关性,促进多源特征的分解.

如式(20)所示，$\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{D}}$和$\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{D}}$为RGB和T细节特征，$\boldsymbol{f}_{\mathrm{vis}}^{\mathrm{B}}$和$\boldsymbol{f}_{\mathrm{ir}}^{\mathrm{B}}$为RGB和T基础特征，$\mathcal{CC}(\cdot)$为皮尔逊相关系数运算符，介于-1~1(越靠近1，相关性越高)，$\epsilon=1.01$以确保计算结果为正. 以输入特征$\boldsymbol{f}_1,\boldsymbol{f}_2\in\mathbb{R}^{H\times W\times C}$为例，$\mathcal{CC}(\cdot)$相关性计算过程如式(21)所示，其中$\mu_{f_1}$和$\mu_{f_2}$为均值.

$$\mathcal{L}_{\mathrm{decomp}}=\frac{\left(\mathcal{CC}(\boldsymbol{f}_{\mathrm{vis}}^{D},\boldsymbol{f}_{\mathrm{ir}}^{D})\right)^2}{\mathcal{CC}(\boldsymbol{f}_{\mathrm{vis}}^{B},\boldsymbol{f}_{\mathrm{ir}}^{B})+\epsilon} \tag{20}$$

$$\mathcal{CC}(\boldsymbol{f}_1,\boldsymbol{f}_2)=\frac{\sum(\boldsymbol{f}_1-\mu_{f_1})(\boldsymbol{f}_2-\mu_{f_2})}{\sqrt{\sum(\boldsymbol{f}_1-\mu_{f_1})^2\sum(\boldsymbol{f}_2-\mu_{f_2})^2}} \tag{21}$$

4 实验设计与结果分析

4.1 数据集

实验使用基于UAV视角的UAV-RGBT数据集和DroneVehicle数据集. 对于UAV-RGBT数据集，本文按照8∶2比例将其划分为训练集和测试集，训练集包含4 094对RGB-T图像和90 226个实例标注，测试集包含1 023对RGB-T图像和21 662个实例标注，均包含人、车辆等20类航拍检测目标. DroneVehicle数据集包含小汽车、面包车、公共汽车、卡车和厢式货车共5类航拍检测目标，其训练集包含17 990对RGB-T图像，测试集包含8 980对RGB-T图像.

4.2 实验设置

(1)实验环境

处理器型号为Intel(R) Xeon(R) Silver 4210 CPU@2.20 GHz，计算机内存256 GB，显卡型号NVIDIA GeForce RTX 3090，操作系统Ubuntu 18.04，Python3.9.18，Pytorch1.11.0深度学习框架，CUDA版本11.3.58，YOLOv8算法版本为8.0.170.

(2)参数设置

训练阶段，对RGB和T图像均使用mosaic数据增强(最后50个训练周期停用)、随机扰动和随机翻转等数据增强方法，图像尺寸缩放为640×640像素；对于UAV-RGBT数据集，使用随机梯度下降(Stochastic Gradient Descent，SGD)优化器，初始学习率为0.1，权重衰减设置为0.000 5，训练周期为300，批处理大小为32，损失函数式(14)中的超参数λ_{cls}、λ_{box}、λ_{dfl}和$\lambda_{\mathrm{decomp}}$分别设置为1.0、7.5、1.5和20；对于DroneVehicle[3]数据集，使用SGD优化器，初始学习率为0.01，权重衰减设置为0.000 5，训练周期为200，批处理大小为32；损失函数式(14)中的超参数λ_{cls}、λ_{box}、λ_{dfl}和$\lambda_{\mathrm{decomp}}$分别设置为0.5、7.5、1.5和20.

(3)实验设计思路

为验证UAV-DMDet算法的优越性，共设计2组对比实验，从算法精度、速度、模型参数量等方面评估其在UAV-RGBT数据集和DroneVehicle数据集上的效果. 为验证MCAF和MFDC模块的有效性，设计2组消融实验，分别在UAV-RGBT数据集和DroneVehicle数据集上进行充分验证. 为验证UAV-DMDet在其他算法框架上的通用性，设计1组扩展实验，将单源YOLOv10n[16]模型改进为多源算法，验证UAV-DMDet在诸如YOLOv10等其他算法框架上的有效性和通用性. 为验证UAV-DMDet在边缘设备上的运行效率，设计1组扩展实验，在华为昇腾710处理器上进行部署和验证.

4.3 评估指标

为评估算法在UAV对地目标检测任务上的先进性和适用性，本文使用平均精度均值(mean Average Precision，mAP)评估算法的检测精度，使用帧每秒(Frame Per Second，FPS)评估算法的运行时间，并使用Parms和每秒十亿次浮点运算(Giga Floating point Operations Per second，GFLOPs)评估模型参数量和复杂度. mAP值越高，说明算法检测性能越好；FPS值越高，说明算法对于UAV视角实时目标检测的适用性越强；Parms越小，表示模型越轻量；GFLOPs越低，表示模型所需计算资源越少，更适合UAV平台的部署和应用.

mAP指所有类别平均精度(Average Precision，AP)的均值，而每个类别的AP与其查准率(Precision)和查全率(Recall)相关. 查准率和查全率计算式如式(22)和式(23)所示，其中，TP(True Positive)为正确检测出的目标数，FP(False Positive)为检测错误的目标数，FN(False Negative)为未能检测出的正确目标数，判定目标检测正确与否的标准是预测的边界框和真实边界框的IoU大于指定阈值. 实验中，使用mAP@0.5和mAP@0.5∶0.95两个指标共同评价检测精度. 其中，mAP@0.5为IoU阈值为0.5时检测精度；mAP@0.5∶0.95是以0.05为步长，计算IoU阈值在0.5~0.95的平均检测精度.

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{22}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{23}$$

4.4 对比实验

4.4.1 UAV-RGBT数据集对比实验

为综合验证UAV-DMDet算法性能，共选取两阶段代表算法Faster-RCNN，一阶段代表算法YOLOv5n和YOLOv8n，Transformer代表算法RT-DETR[36]，以及RGB-T多源目标检测代表算法ICAFusion[37]和SAMS-YOLO[38]进行比较，所有实验均使用NVIDIA GeForce RTX3090显卡，并分别在RGB标注和T标注进行了评估，实验结果如表4所示.

相比于单源目标检测算法，UAV-DMDet通过互补融合RGB和T多源特征，在两个模态上均实现了检测精度的提升. 与YOLOv8n相比，UAV-DMDet在RGB模态上mAP@0.5和mAP@0.5∶0.95分别提高了3.61个百分点和0.84个百分点，在T模态上分别提高了11.03个百分点和

表4　UAV-RGBT数据集对比实验结果

模型		RGB模态 mAP@0.5/%	T模态 mAP@0.5:0.95/%	RGB模态 mAP@0.5/%	T模态 mAP@0.5:0.95/%	FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
Faster-RCNN[8]	单源(两阶段)	60.22	30.90	48.57	24.00	6.20	948.35	28.47
RT-DETR[36]	单源(Transformer)	66.80	47.17	54.83	35.28	19.66	186.30	60.94
YOLOv5n[6]	单源(一阶段)	63.80	40.17	55.08	32.03	91.74	4.20	1.79
YOLOv8n[5]	单源(一阶段)	67.72	45.71	59.70	36.66	90.53	8.10	3.01
UAV-DMDet (本文)	多源(一阶段)	71.33(+3.61)	46.55(+0.84)	70.73(+11.03)	43.42(+6.76)	31.31	18.90	8.40
ICAFusion[37]	多源(一阶段)	74.73	48.70	74.09	44.81	27.36	—	120.31
SAMS-YOLO[38]	多源(一阶段)	72.67	48.76	71.96	44.92	23.96	283.30	97.76
UAV-DMDet-L (本文)	多源(一阶段)	75.13	52.78	73.27	47.96	23.38	364.60	99.19

6.76个百分点. 其中,T模态提升较大的主要原因为T图像分辨率较低,缺乏色彩、纹理等信息,通过融合RGB特征极大地丰富了特征信息.

在推理效率方面,UAV-DMDet在单张NVIDIA GeForce RTX3090显卡上检测速度达到了31.31帧/s,实现了UAV航拍RGB-T图像的实时检测. 此外,与当前主流多源目标检测算法相比,基于YOLOv8l模型构建的UAV-DMDet-L算法在检测精度和模型参数量等方面有着一定的优势. 相比于同等参数量的SAMS-YOLO算法,UAV-DMDet-L在RGB模态上mAP@0.5和mAP@0.5:0.95分别提高了2.46个百分点和4.02个百分点,在T模态上分别提高了1.31个百分点和3.04个百分点,且保持了与SAMS-YOLO相当的推理效率;与ICAFusion相比,UAV-DMDet-L在RGB模态上mAP@0.5和mAP@0.5:0.95分别提高了0.4个百分点和4.08个百分点,在T模态上mAP@0.5:0.95提高了3.15个百分点,且参数量减少了21.19 MB.

4.4.2　DroneVehicle数据集对比实验

UAV-DMDet算法和相关主流算法在DroneVehicle数据集上的实验结果如表5所示(均为T模态评估结果). 与I^{2}MDet方法相比,UAV-DMDet在mAP@0.5和mAP@0.5:0.95分别提高了2.66个百分点和12.36个百分点,模型参数量仅不到I^{2}MDet的1/5,在单张NVIDIA GeForce RTX3090显卡上检测速度达到了31.04帧/s,可实现实时车辆检测.

表5　DroneVehicle数据集对比实验结果

模型	小汽车	卡车	面包车	公共汽车	厢式货车	mAP@0.5/%	mAP@0.5:0.95/%	FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
UA-CMDet[3]	87.51	60.70	37.95	87.08	46.80	64.01	—	9.12	—	—
Oriented R-CNN[39]	89.90	56.60	46.90	89.60	54.40	67.52	42.60	—	41.13	107.26
RoI Transformer[40]	90.10	60.40	52.20	89.70	58.90	70.29	43.57	—	—	—
CIAN(OBB)[41]	89.98	62.47	49.59	88.90	60.22	70.23	—	21.70	—	—
AR-CNN(OBB)[42]	90.08	64.82	51.51	89.38	62.12	71.58	—	18.20	—	—
TSFADet[14]	89.88	67.87	53.99	89.81	63.74	73.06	—	18.60	109.80	104.70
ViT-B+RVSA[43]	89.70	52.30	44.40	88.00	51.00	65.07	42.63	—	60.46	134.69
C^{2}Former-S^{2}ANet[12]	90.20	68.30	58.50	89.80	64.40	74.20	—	—	100.90	132.50
AFFCM[44]	90.16	73.40	64.92	89.86	64.86	76.64	—	—	—	—
I^{2}MDet[13]	96.30	73.40	58.60	93.20	65.00	77.30	46.20	—	31.87	48.90
UAV-DMDet(本文)	98.26	78.47	61.54	95.38	66.17	79.96	58.56	31.04	18.90	8.39

4.5　消融实验

4.5.1　基准模型

为验证MCAF和MFDC的有效性,首先去除UAV-DMDet网络中的MCAF模块和MFDC模块以建立基线模型. 具体而言,将MCAF模块替换为骨干网络P5尺度多源特征通道拼接操作;去除MFDC模块,将来自骨干网络的P3和P4尺度RGB特征送入RGB分支颈部,将T特征送入T分支颈部.

4.5.2　消融实验结果

表6和表7展示了不同模块在UAV-RGBT数据集和DroneVehicle数据集上对模型精度、参数量和检测速度等方面的影响.

表6 UAV-RGBT数据集消融实验结果

模型	RGB模态		T模态		FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
	mAP@0.5/%	mAP@0.5:0.95/%	mAP@0.5/%	mAP@0.5:0.95/%			
YOLOv8n[5]	67.72	45.71	59.70	36.66	90.53	8.10	3.01
基线模型	68.42(+0.70)	45.03(−0.68)	66.58(+6.88)	40.93(+4.27)	46.84	16.30	6.08
+MCAF	69.42(+1.00)	45.23(+0.20)	67.42(+0.84)	41.22(+0.29)	46.81	16.70	8.19
+MFDC	70.60(+2.18)	46.58(+1.55)	69.83(+3.25)	44.00(+3.07)	36.87	18.50	6.29
UAV-DMDet	71.33(+2.91)	46.55(+1.52)	70.73(+4.15)	43.42(+2.49)	31.31	18.90	8.40

表7 DroneVehicle数据集消融实验结果

模型	RGB模态		T模态		FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
	mAP@0.5/%	mAP@0.5:0.95/%	mAP@0.5/%	mAP@0.5:0.95/%			
YOLOv8n[5]	75.19	51.87	78.69	57.06	88.40	8.10	3.01
基线模型	78.56(+3.37)	56.40(+4.53)	79.18(+0.49)	57.33(+0.27)	46.91	16.30	6.08
+MCAF	79.23(+0.67)	57.05(+0.65)	79.40(+0.22)	57.70(+0.37)	46.85	16.70	8.18
+MFDC	79.02(+0.46)	57.33(+0.93)	79.74(+0.56)	58.24(+0.91)	33.38	18.40	6.29
UAV-DMDet	79.81(+1.25)	58.00(+1.60)	79.96(+0.78)	58.56(+1.23)	31.04	18.90	8.39

(1)相较于YOLOv8n单源模型,双分支基线模型在RGB和T模态mAP@0.5和mAP@0.5:0.95指标方面提高了0.27~6.88个百分点,充分证明了双分支网络通过跨模态特征融合可以极大地提高RGB和T模态的检测精度.

(2)添加MCAF模块后,在两类数据集上RGB和T模态mAP@0.5和mAP@0.5:0.95提高了0.2~1.0个百分点,证明了交叉注意力融合模块对骨干网络多源特征融合的有效性.

(3)添加MFDC模块之后,在两类数据集上RGB和T模态mAP@0.5和mAP@0.5:0.95提高了0.46~3.25个百分点,精度提升较大,证明在颈部将特征分解为基础特征和细节特征,并在当前模态中引入另一模态的基础特征可以实现跨模态的特征融合,并可避免因另一模态细节特征差异性所带来的噪声问题. 由于其中细节特征提取模块使用了带仿射耦合层的可逆神经网络,计算效率相对较低,对模型检测速度影响较大.

(4)同时添加MCAF和MFDC模块后,在UAV-RGBT数据集上RGB和T模态mAP@0.5和mAP@0.5:0.95分别提高2.91、1.52、4.15和2.49个百分点,在DroneVehicle数据集上相关指标分别提高1.25、1.60、0.78和1.23个百分点,证明了两个模块组合使用的有效性.

4.5.3 可视化结果

为验证MCAF和MFDC模块的有效性,本文基于UAV-RGBT数据集进行了可视化结果分析. 图13为两个场景下基线模型和添加子模块后各模型在RGB和T图像上的检测结果. 如图13中白色椭圆区域所示,相较于基线模型,添加MCAF和MFDC模块后模型漏检和错检现象得到进一步改善.

4.6 扩展实验

4.6.1 YOLOv10框架应用验证

为验证双分支算法框架以及MCAF和MFDC模块的通用性,本文基于主流的单源YOLOv10n模型进行了多源算法实现,称之为UAV-DMDet-v10,并分别在UAV-RGBT数据集和DroneVehicle数据集上进行了实验验证,实验结果如表8和表9所示.

由实验结果可知,相比于单源YOLOv10n模型,双分支结构的基线模型在RGB和T模态mAP指标上均有不同程度的提高,再次验证了双分支结构多源模型相较于单源模型在多场景适用性和模型精度方面的优势. 在基线模型基础上添加MCAF和MFDC模块后,模型精度得到进一步提高,验证了所提出模块在YOLOv10等其他目标检测框架上的有效性.

基于YOLOv8n的UAV-DMDet模型与基于YOLOv10n的UAV-DMDet-v10模型在精度、速度及参数量方面的对比如表10所示. 相比于UAV-DMDet,UAV-DMDet-v10在DroneVehicle数据集mAP@0.5精度指标上有一定提升,但在其他方面精度不及UAV-DMDet,同时UAV-DMDet-v10推理速度不及UAV-DMDet. 综合来看,基于YOLOv8n的UAV-DMDet模型在精度和速度方面实现了较好的均衡.

4.6.2 边缘设备效率验证

基于Pytorch框架实现的UAV-DMDet模型参数量大小为8.40 MB、计算参数量为18.90 GFLOPs,使用FP32计算精度、640×640分辨率图像尺寸,在单张GeForce RTX 3090显卡上其推理速度可达31帧/s,满足地面设备实时处理要求. 为进一步验证UAV-DMDet算法在UAV平台等边缘设备的适用性,本文基于华为昇腾710处理器,对UAV-DMDet模型进行了部署实验,实验结果表明:在输入图像尺寸640×640分辨率、FP16计算精度情况下,UAV-DMDet模型推理速度可达58.19帧/s,实现了边缘设备RGB-T多源目标实时检测.

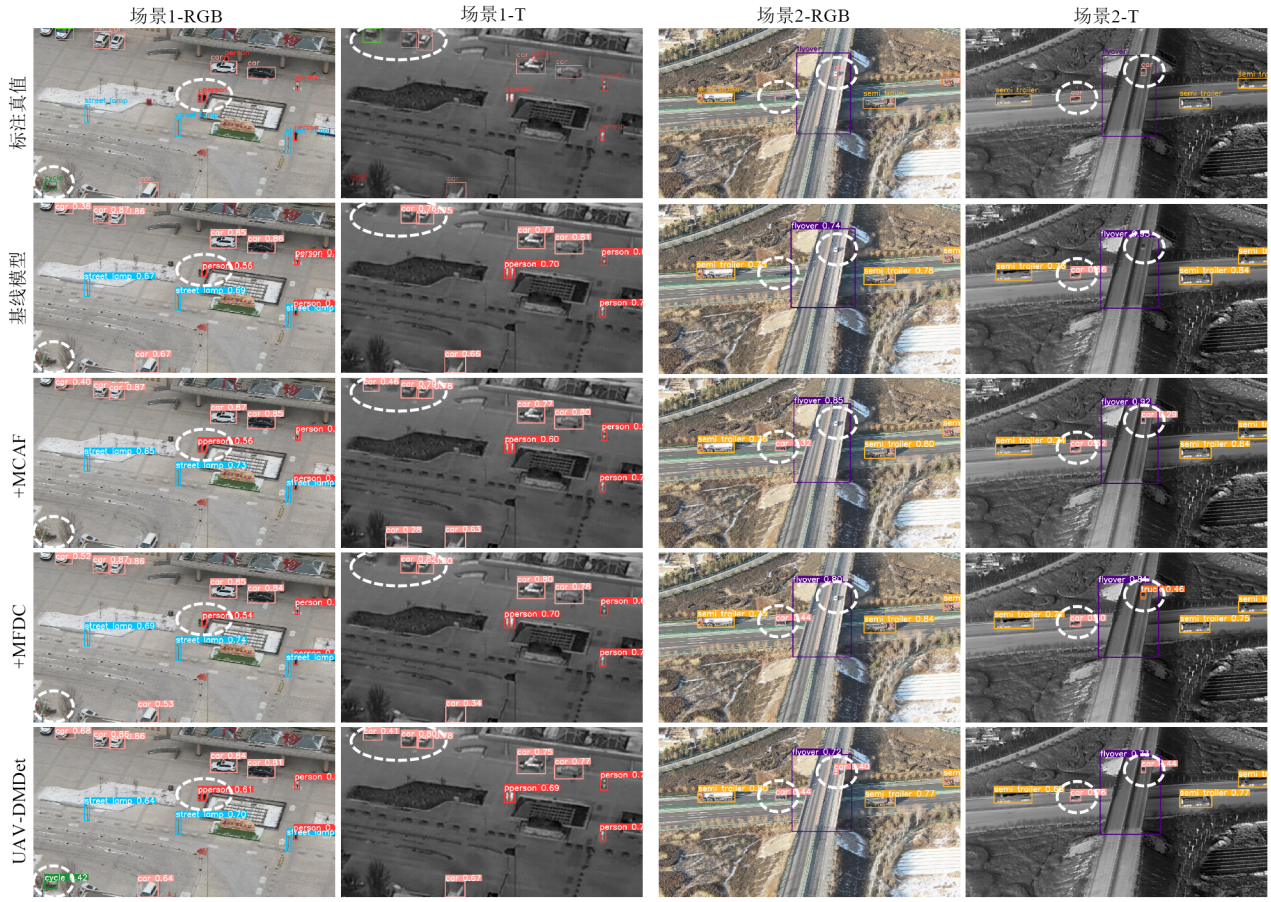


图13　UAV-RGBT数据集可视化效果图

表8　UAV-RGBT数据集通用性实验验证结果

模型	RGB模态		T模态		FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
	mAP@0.5/%	mAP@0.5:0.95/%	mAP@0.5/%	mAP@0.5:0.95/%			
YOLOv10n[44]	64.96	43.95	58.97	36.60	95.95	8.30	2.70
基线模型	66.84(+1.88)	43.68(−0.27)	65.96(+6.99)	40.45(+3.85)	53.71	16.50	5.22
+MCAF	68.60(+1.76)	44.52(+0.84)	66.63(+0.67)	40.80(+0.35)	39.16	16.90	7.33
+MFDC	68.8(+1.96)	44.21(+0.53)	68.26(+2.3)	42.12(+1.67)	25.30	20.70	5.44
UAV-DMDet-v10	69.98(+3.14)	45.46(+1.78)	68.62(+2.66)	43.02(+2.57)	23.62	21.20	7.54

表9　DroneVehicle数据集通用性实验验证结果

模型	RGB模态		T模态		FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
	mAP@0.5/%	mAP@0.5:0.95/%	mAP@0.5/%	mAP@0.5:0.95/%			
YOLOv10n[44]	75.79	52.17	79.58	57.60	98.75	8.20	2.70
基线模型	78.68(+2.89)	56.29(+4.12)	79.72(+0.14)	57.36(−0.24)	49.62	16.40	5.21
+MCAF	78.49(−0.19)	56.39(+0.10)	79.90(+0.18)	57.49(+0.13)	46.11	16.80	7.31
+MFDC	80.16(+1.48)	57.46(+1.17)	80.34(+0.62)	57.94(+0.58)	25.98	20.70	5.43
UAV-DMDet-v10	80.68(+2.00)	57.57(+1.28)	80.83(+1.11)	58.10(+0.74)	21.74	21.10	7.53

5　结论

针对当前UAV视角RGB-T多源图像数据集场景单一、标注类别少等问题,本文采集构建了一个跨季节、跨昼夜、多尺度的大规模UAV视角RGB-T多源图像数据集UAV-RGBT,其飞行场景涵盖了校园、城市道路、高速公路、居民区、停车场、公园、机场、乡镇、农田、河流等10余种常见场景,飞行高度介于50~500 m,并对车辆、路灯、电线杆、红绿灯、高压输电塔、桥梁、操场等20个类别进行了超11万个标注,可广泛应用于智慧交通、无人驾驶航空器,以及其他低空经济所需的UAV视

表 10 UAV-DMDet与UAV-DMDet-v10对比分析

模型	数据集	RGB模态		T模态		FPS/(Frame·s^{-1})	GFLOPs/G	Parms/MB
		mAP@0.5/%	mAP@0.5:0.95/%	mAP@0.5/%	mAP@0.5:0.95/%			
UAV-DMDet	UAV-RGBT	71.33	46.55	70.73	43.42	31.31	18.90	8.40
UAV-DMDet-v10		69.98(−1.35)	45.46(−1.09)	68.62(−2.11)	43.02(−0.4)	23.62	21.20	7.54
UAV-DMDet	DroneVehicle	79.81	58.00	79.96	58.56	31.04	18.90	8.39
UAV-DMDet-v10		80.68(+0.87)	57.57(−0.43)	80.83(+0.87)	58.10(−0.46)	21.74	21.10	7.53

角目标检测算法研究工作. 与此同时,本文提出了一种双分支RGB-T多源目标检测算法UAV-DMDet,其通过双分支多源特征提取和高低层多源特征融合,实现了RGB和T图像的有效检测. 其中,针对RGB-T特征跨模态融合难的问题,本文设计了一种MCAF模块,通过计算模态间的交叉注意力来引导高层语义特征实现跨模态的特征融合;并设计了一种MFDC模块,将多源特征分解为基础特征和细节特征,通过不同形式的组合,以促进低层RGB和T分支的跨模态融合检测. 实验表明:所设计的UAV-DMDet算法,在UAV-RGBT数据集和DroneVehicle数据集上均取得了最佳的检测性能,并在地面设备GeForce RTX 3090显卡和边缘设备昇腾710处理器上实现了实时目标检测,可有效应用于RGB-T多源目标检测任务. 事实上,本文所构建的MCAF和MFDC模块属于即插即用的通识模块,可用于如YOLOv10等其他目标检测框架,也可用于如RGB-T图像融合、语义分割等其他多源视觉任务. 此外,本文提出的UAV-DMDet算法在地面设备及边缘设备进行实时目标检测时图像分辨率大小为640×640像素,对于更高分辨率的数据其计算复杂度还有优化的空间,进一步提高模型效率是后续的研究方向之一.

参考文献

[1] GONZÁLEZ A, FANG Z J, SOCARRAS Y, et al. Pedestrian detection at day/night time with visible and FIR cameras: A comparison[J]. Sensors, 2016, 16(6): 820.

[2] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1037-1045.

[3] SUN Y M, CAO B, ZHU P F, et al. Drone-based RGB-infrared cross-modality vehicle detection *via* uncertainty-aware learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713.

[4] HAN Y Q, LIU H P, WANG Y F, et al. A comprehensive review for typical applications based upon unmanned aerial vehicle platform[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 9654-9666.

[5] JOCHER G, CHAURASIA A, QIU J, et al. YOLO by ultralytics[EB/OL]. (2023-01-23)[2024-06-26]. https://github.com/ultralytics/ultralytics.

[6] JOCHER G. YOLOv5 by Ultralytics[EB/OL]. (2020-01-01)[2024-06-26]. https://github.com/ultralytics/yolov5.

[7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[M]//Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[8] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[9] CAI Z W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6154-6162.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2024-06-26]. https://arxiv.org/abs/2010.11929.

[11] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers[M]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[12] YUAN M X, WEI X X. C^{2}Former: Calibrated and complementary transformer for RGB-infrared object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5403712.

[13] ZHANG N, LIU Y M, LIU H, et al. Oriented infrared vehicle detection in aerial images *via* mining frequency and semantic information[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5002315.

[14] YUAN M X, WANG Y Y, WEI X X. Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 509-525.

[15] HUANG Z C, LI W, TAO R. Multimodal knowledge distillation for arbitrary-oriented object detection in aerial images[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Piscataway: IEEE, 2023: 1-5.

[16] WANG A, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[EB/OL]. (2024-05-30)[2024-06-26]. https://arxiv.org/abs/2405.14458v2.

[17] DU D W, QI Y K, YU H Y, et al. The unmanned aerial vehicle benchmark: Object detection and tracking[M]// Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 375-391.

[18] BOZCAN I, KAYACAN E. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2020: 8504-8510.

[19] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4165-4173.

[20] ZHANG W, LIU C S, CHANG F L, et al. Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images[J]. Remote Sensing, 2020, 12(11): 1760.

[21] ZHU P F, WEN L Y, BIAN X, et al. Vision meets drones: A challenge[EB/OL]. (2018-04-23)[2024-06-26]. https://arxiv.org/abs/1804.07437v2.

[22] ZHANG H J, SUN M S, LI Q, et al. An empirical study of multi-scale object detection in high resolution UAV images[J]. Neurocomputing, 2021, 421: 173-182.

[23] WANG J H, TENG X C, LI Z, et al. VSAI: A multi-view dataset for vehicle detection in complex scenarios using aerial images[J]. Drones, 2022, 6(7): 161.

[24] PORTMANN J, LYNEN S, CHLI M, et al. People detection and tracking from aerial thermal views[C]//2014 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2014: 1794-1800.

[25] SUO J S, WANG T Y, ZHANG X Z, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data, 2023, 10(1): 227.

[26] ZHANG X W, LI Y P, QI Z S, et al. Learning multi-domain feature relation for visible and Long-wave Infrared image patch matching[EB/OL]. (2023-08-09)[2024-06-26]. https://arxiv.org/abs/2308.04880v1.

[27] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8759-8768.

[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[29] BA J L, KIROS J R, HINTON G E. Layer normalization[EB/OL]. (2016-07-21)[2024-06-26]. https://arxiv.org/abs/1607.06450v1.

[30] ZHAO Z X, BAI H W, ZHANG J S, et al. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 5906-5916.

[31] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 5718-5729.

[32] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real NVP[EB/OL]. (2017-02-27)[2024-06-26]. https://arxiv.org/abs/1605.08803v3.

[33] ZHOU M, HUANG J, FANG Y C, et al. Pan-sharpening with customized transformer and invertible neural network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3553-3561.

[34] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1314-1324.

[35] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[36] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[EB/OL]. (2024-04-03)[2024-06-26]. https://arxiv.org/abs/2304.08069v3.

[37] SHEN J F, CHEN Y F, LIU Y, et al. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection[J]. Pattern Recognition, 2024, 145: 109913.

[38] WANG J Z, TIAN X T, DAI S, et al. RGB-T object detection via group shuffled multi-receptive attention and multi-modal supervision[EB/OL]. (2024-05-29)[2024-06-26]. https://arxiv.org/abs/2405.18955v1.

[39] XIE X X, CHENG G, WANG J B, et al. Oriented R-CNN for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3500-3509.

[40] DING J, XUE N, LONG Y, et al. Learning RoI transformer for oriented object detection in aerial images[C]//2019 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 2844-2853.
[41] ZHANG L, LIU Z Y, ZHANG S F, et al. Cross-modality interactive attention network for multispectral pedestrian detection[J]. Information Fusion, 2019, 50: 20-29.
[42] ZHANG L, ZHU X Y, CHEN X Y, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 5127-5137.
[43] WANG D, ZHANG Q M, XU Y F, et al. Advancing plain vision transformer toward remote sensing foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 61: 5607315.
[44] WU Y F, GUAN X R, ZHAO B Y, et al. Vehicle detection based on adaptive multimodal feature fusion and cross-modal vehicle index using RGB-T images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 8166-8177.

作者简介

汪进中 男，1995 年生，甘肃民勤人. 现为西北工业大学计算机学院硕士研究生. 主要研究方向为图像处理、多源目标检测、深度学习等.
E-mail: wangjinzhong@mail.nwpu.edu.cn

张秀伟 女，1981 年生，新疆塔城人. 现为西北工业大学计算机学院教授. 主要研究方向为计算机视觉、多源信息协同处理、深度学习等.
E-mail: xwzhang@nwpu.edu.cn