中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-22

论文引用格式:Liu Yu, Feng Yingchao, Zhang Yidan, Li Ning, Diao Wenhui, Hu Yanfeng. DIV: A visible-infrared object detection dataset for weakly aligneddrone scenarios[J/OL]. Journal of Image and Graphics, XXXX:1-22. DOI: 10.11834/jig.260248.(刘煜, 冯瑛超, 张伊丹, 李宁, 刁文辉, 胡岩峰. DIV:面向无人机弱对齐场景的可见光-红外目标检测数据集[J/OL]. 中国图象图形学报,XXXX:1-22. DOI: 10.11834/jig.260248.)[DOI:10.11834/jig.260248]

DIV:面向无人机弱对齐场景的可见光-红外目标检测数据集

刘煜[1,2,3,4],冯瑛超[1,4*],张伊丹[1,4],李宁[1,4],刁文辉[1,4],胡岩峰[1]

1. 中国科学院空天信息创新研究院,北京,100190; 2. 中国科学院大学,北京,100190; 3. 中国科学院大学电子电气与通信工程学院,北京,100190; 4. 目标认知与应用技术国家级重点实验室,北京,100190

摘　要: 无人机凭借全天候、全天时探测优势,在边境监控及灾害救援等领域发挥重要作用。然而,现有的红外-可见光(IR-VIS)多模态目标检测研究过度依赖理想化的"像素级对齐"假设,且数据集普遍存在目标尺度分布极化、类别同质化等问题,导致现有算法在处理具有真实视差及动态尺度演变的无人机航拍数据时,极易出现特征失配、定位漂移及漏检误识。基于以上问题,文中构建了一个面向真实弱配准场景的多模态、多尺度、多类别无人机目标检测基准数据集DIV(drone-based IR-VIS object detection)。该数据集保留了红外与可见光图像之间因无人机传感器安装差异、视角变化及飞行抖动等因素在实际应用中产生的非线性空间偏移。数据集内容涵盖从像素占比极低的微小型目标到显著区域的大型目标,并引入了行人、非机动车、各类车辆等多样化类别。拍摄环境覆盖城市、山区及乡村,并细分为日间、傍晚和弱光三种典型光照场景。同时,通过独立模态的人工高精度标注及多轮交叉验证机制,确保了弱对齐约束下的语义一致性。选取了9种主流多模态目标检测算法在所提数据集上进行基准测试。实验结果表明,在理想对齐数据集中表现优异的方法,在本数据集的弱配准场景下性能出现不同程度的下滑。一些模型方法在特定维度上表现出色,但在应对极端复杂环境时的综合感知能力仍显不足。面对真实环境中的复合挑战,构建一套多维协同优化体系以强化对多模态信息的提取与整合能力,是提升无人机平台感知鲁棒性的关键。提出的数据集有效弥合了多模态探测算法的"理想假设"与无人机数据"现实分布"之间的鸿沟,为构建多维协同优化体系提供了真实的验证平台,在数据层面为弱对齐约束下的跨模态特征融合与鲁棒检测研究提供了关键支撑。DIV数据集发布地址为:https://www.scidb.cn/preview?dataSetId=d1e8909592e04cc2a1095e62f579ead1&version=V1。

关键词: 航空遥感影像;多模态图像融合;跨模态对齐;红外与可见光;目标检测

DIV: A visible-infrared object detection dataset for weakly aligneddrone scenarios

Liu Yu[1,2,3,4], Feng Yingchao[1,4*], Zhang Yidan[1,4], Li Ning[1,4], Diao Wenhui[1,4], Hu Yanfeng[1]

1. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190; 2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190; 3. University of Chinese Academy of Sciences, Beijing 100190; 4. National Key Laboratory of Target Cognition and Application Technology (TCAT), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190

收稿日期:2026-04-30;**修回日期**:2026-06-13

*通信作者:冯瑛超。E-mail: fengyc@aircas.ac.cn

基金项目:国家重点研发计划(2024YFF1401001);国家自然科学基金(62301538);空天院科学与颠覆性技术项目(2025-AIRCAS-SDTP-04)

Supported by: National Key R&D Program of China (2024YFF1401001); National Natural Science Foundation of China (62301538); the Science and Disruptive Technology Program (2025-AIRCAS-SDTP-04)

Abstract: Multimodal detection using unmanned Aerial Vehicles (UAVs) has demonstrated significant practical value in applications such as border surveillance, disaster relief, and urban security. This is primarily due to their advantages of high mobility, rapid deployment, and all-weather observation capabilities. Compared with single-modality perception methods, multimodal object detection can simultaneously utilize the textural information provided by visible (VIS) images and the thermal radiation characteristics captured by infrared (IR) images. This complementary characteristic significantly improves detection performance in complex environments. However, most existing studies are developed under an idealized assumption that IR and VIS images are strictly aligned at the pixel level. Such an assumption rarely holds in real UAV scenarios. In practical applications, factors including sensor installation deviations, viewpoint differences, flight attitude changes, and platform vibrations often introduce nonlinear spatial offsets between modalities. These discrepancies further lead to cross-modal feature misalignment, semantic inconsistency, and severe performance degradation. Existing infrared-visible (IR-VIS) datasets also exhibit several inherent limitations. Most datasets are collected in ground-level or low-altitude scenarios and therefore lack sufficient representation of high-altitude UAV viewpoints. In addition, object-scale distributions are often highly imbalanced, while category settings remain relatively limited. As a result, these datasets cannot adequately reflect the diverse object distributions and complex environmental conditions encountered in real-world UAV tasks. Under such conditions, mainstream multimodal detection methods often suffer from issues such as localization drift, missed detections, and misclassifications when processing real-world UAV data. This is because real-world data typically involves spatial disparities, dynamic scale variations, and complex background interference. These issues severely limit the practical value of existing methods. To address the above limitations, this paper presents DIV, a multimodal UAV object detection dataset specifically designed for real-world weakly aligned scenarios. The dataset was collected using UAV platforms equipped with infrared and visible sensors across diverse real-world environments. Unlike existing strictly aligned datasets, DIV intentionally preserves the spatial inconsistencies caused by systematic errors and dynamic flight factors, thereby providing a more realistic representation of practical deployment conditions. The proposed dataset emphasizes the characteristic of "weak alignment", enabling more effective evaluation of model robustness and generalization under cross-modal offset conditions. In terms of data content, DIV exhibits both diversity and complexity. The dataset contains objects ranging from extremely small objects with very limited pixel coverage to large-scale objects occupying substantial image regions, allowing comprehensive evaluation of scale adaptability. The category settings include person, non-motorized vehicle, and multiple types of motor vehicles, which improves both task complexity and practical relevance. In addition, the dataset covers a variety of environments, including urban, mountainous, and rural areas. To further simulate real-world conditions, the data are subdivided into daytime, nighttime, and low-light scenarios. To evaluate the challenges and effectiveness of DIV, several mainstream multimodal object detection algorithms were selected for experimental analysis. Experimental results demonstrate that methods achieving strong performance on conventional strictly aligned datasets experience noticeable performance degradation on DIV, particularly in small object detection and complex scene perception tasks. These findings indicate that many existing methods heavily depend on accurate cross-modal alignment and lack sufficient capability to model realistic UAV flight conditions. Furthermore, although some approaches perform well under specific metrics or limited scenarios, their overall perception capability remains insufficient when simultaneously confronted with multi-scale targets, multiple object categories, and complex environmental interference. The results further reveal a coupling relationship among weak multimodal alignment, scale variation, and environmental complexity. These factors jointly affect the quality of multimodal feature representation and fusion. Therefore, optimization at a single level alone is insufficient to comprehensively address these challenges. For practical UAV applications, there is an urgent need to develop a multi-dimensional collaborative optimization framework that improves multimodal modeling and fusion capability at the data, feature, and task levels simultaneously. By breaking the conventional ideal-alignment assumption, the proposed DIV dataset effectively bridges the gap between multimodal algorithm research and real-world UAV deployment scenarios. This work provides a foundational benchmark for future research on robust multimodal perception in realistic environments.

Key words: aerial remote sensing images; multimodal image fusion; cross-modal alignment; infrared and visible; object detection

论文引用格式：Liu Y，Feng Y C，Zhang Y D，Li N，Diao W H and Hu Y F. 2026. DIV：A visible-infrared object detection dataset for weakly aligned drone scenarios. Journal of Image and Graphics，xx(xx)：xxxx-xxxx(刘煜，冯瑛超，张伊丹，李宁，刁文辉，胡岩峰. 2026. DIV：面向无人机弱对齐场景的可见光-红外目标检测数据集. 中国图象图形学报，xx(xx)：xxxx-xxxx)[DOI：10.11834/jig.260248]

0 引 言

在无人机(unmanned aerial vehicle，UAV)多模态目标检测任务中，可见光与红外模态的深度融合是实现全天候、全天时探测的核心路径(Ma等，2019)。可见光传感器具有高分辨率空间纹理，能够精细刻画目标(Tang等，2023)；而红外传感器凭借对热辐射的敏感性，在低光、烟雾或伪装等复杂场景下具有独特的显著性表征能力(Guo等，2024)。虽然通过融合不同模态的独特特征可以提供跨模态互补信息，但实现有效的跨模态融合仍具挑战性。有效的信息集成并非特征分布简单地堆叠，无差别地盲目聚合往往会引入无关的、错误的融合表征，引发跨模态“信息污染”。这些噪声会影响融合图像的视觉质量，并且严重削弱融合表征在下游检测任务中的判别力。

如图1所示，在实际应用中，由于双载荷传感器安装位置的差异、无人机平台的震动以及观测距离改变引起的视差，获取的跨模态图像对之间普遍存在空间位置的不对齐。这些时空不一致性加剧了模态间特征匹配失效与模态内可靠性等问题，打破了现有算法对空间像素级对齐的先验假设。因此，探究低信噪比及弱对齐约束下的鲁棒融合和检测方法已成为当前亟待解决的学术难题(Guo等，2024；Li等，2025；Chen等，2024；Yuan等，2024)。

目前，无人机视角下的多模态目标检测研究面临算法模型与真实数据分布严重脱节的困境。现有多数方法过度依赖于理想的对齐假设，忽视了实际场景中会遇到的复杂飞行条件与环境变化。此外，现有方法与数据的脱节不仅限于真实跨模态数据的弱对齐问题。现有的可见光-红外数据集在目标尺度上往往呈现出明显的“极化”现象：大量数据集聚焦于近距离、大尺度的预配准场景，忽视了远距离观测下极小目标的探测难题。在类别维度上，现有的小目标检测(small object detection，SOD)数据集与研究多侧重于行人等单一类别，缺乏对非机动车、多型车辆等复杂目标多样化覆盖。在真实的航拍任务中，目标尺寸会随观测高度与视角的实时变化而波动，需要搜寻与检测的目标种类也不会局限于单一维度，这种尺度敏感性与类别多样性要求检测模型必须具备极高的泛化能力。

这种数据集与真实场景之间的不兼容，进一步诱发了底层特征处理的瓶颈。由于多数融合和检测框架基于“像素级对齐”的理想假设，在弱对齐条

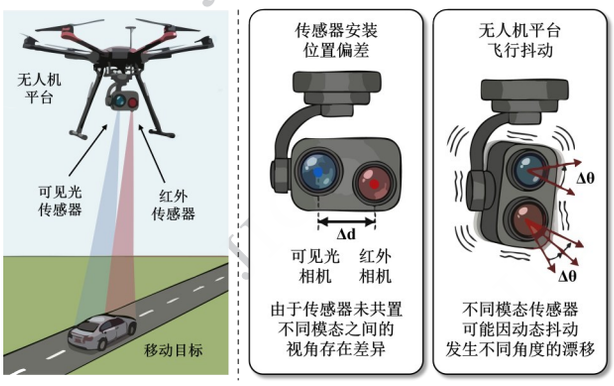


图1 无人机双载荷系统空间对准偏差的成因

Fig. 1 Origins of spatial misalignment in UAV dual-payload systems

件下，强制的像素融合会导致不同模态的语义信息相互干扰乃至抵消，产生视觉伪影和噪声。这不仅掩盖了目标的显著性，更破坏了对目标原有特征的准确提取，导致定位框发生严重漂移。此外，不同尺度的目标对空间偏移也会具有截然不同的敏感性：对于大尺度目标，其对于位移具有一定的容忍度，数十像素的偏差可能仅表现为局部模糊；而对于像素占比极小的目标，微小的像素偏移就足以导致跨模态的特征在空间上完全解耦。由于空间偏差，跨模态互补信息无法聚焦。这种互补信息的空间发散使得模型难以建立稳定的跨模态关联，成为错检与漏检频发的根本原因。

针对上述问题，本文聚焦于真实无人机场景下多模态目标检测这一研究问题，旨在弥合现有算法理想像素级对齐假设与真实数据分布之间的鸿沟。本文的主要贡献如下：

1) 明确了真实无人机场景下多模态融合和检测数据和方法的挑战与问题，并建立了面向弱配准

挑战的系统量化评估。本文指出现有检测研究大多默认红外与可见光图像严格像素级对齐的理想化假设,而真实无人机场景中普遍存在跨模态空间错位问题。围绕这一现实挑战,提出的DIV数据集保留真实空间偏移,为评估多模态算法在非理想条件下的空间鲁棒性提供了标准化的数据支撑和基准,将研究重点从理想条件,扩展至面向真实弱配准条件的鲁棒多模态感知问题;

2)构建了多尺度、多类别与复杂环境特性的无人机多模态IR-VIS目标检测基准数据集。针对无人机检测任务,DIV涵盖了从极低像素占比的微小目标到大尺度目标的宽尺度域,打破了现有数据集在目标尺寸上的"极化"现象。同时,本数据集引入了非机动车、多型车辆等复杂类别,且包含了不同环境以及不同时段,形成了更贴合无人机在真实环境下探测时的数据分布特性;

3)揭示了现有主流多模态检测方法在真实弱配准场景下的性能退化与潜在瓶颈。通过对比实验发现,现有方法在处理跨模态空间错位、极端尺度变化及局部质量退化等复杂条件时,普遍存在跨模态特征关联能力下降与目标定位不稳定等问题。当前大量方法仍高度依赖理想像素级对齐假设,其特征融合机制在真实无人机场景中的鲁棒性与泛化能力仍存在明显不足。相关分析为后续研究从"理想对齐条件下的特征融合"向"面向真实退化环境的容错感知"演进提供了实验依据与研究启示。

1 相关工作

1.1 无人机多模态目标检测方法

随着无人机技术在低空经济、边境巡逻及灾难救援等领域的广泛应用,构建全天候、高鲁棒性的感知系统已成为学术界与工业界的共识。在光照多变、烟雾遮挡等复杂环境下,单一模态传感器(如可见光)往往面临成像失效的风险。红外与可见光图像融合技术通过整合红外模态的目标热辐射特征与可见光模态的高分辨纹理细节,能够有效弥补单模态信息的局限性。因此,如何实现跨模态信息的深度整合及其下游检测任务的高效耦合,已成为提升无人机自主感知能力的核心路径。

基于视觉增强驱动的图像融合方法:该类研究主要聚焦于底层视觉效果的重建,旨在生成符合人类视觉感知的融合图像。早期的主流架构如基于自编码器(AE)的DenseFuse(Li等,2018)、NestFuse(Li等,2020)和RFN-Nest(Li等,2021),通过预训练的编码器解耦并提取模态特有属性。随后,基于卷积神经网络(CNN)的方法(如LRRNet(Li等,2023)、IGNet(Li等,2023)、Dif-Fusion(Yue等,2023))与生成对抗网络(GAN)方法(如FusionGAN(Ma等,2018)、DDcGAN(Ma等,2020)、GANMcC(Ma等,2020))相继涌现,利用对抗博弈或特定的像素级损失函数,试图在热辐射显著性与空间纹理细节之间寻求最优平衡。近年来,以SwinFusion(Ma等,2022)、CDDFuse(Zhao等,2023)和DATfusion(Tang等,2023)为代表的Transformer架构,凭借其强大的全局注意力机制,在长程依赖建模上展现出显著优势。

基于感知任务驱动的协同检测方法:与传统视觉增强方法侧重于"主观观感"不同,该类研究强调融合过程应服务于检测、分割等高层语义理解任务。传统增强范式生成的图像虽然在信噪比和对比度上表现优异,但往往缺乏针对目标检测任务的判别性表征,导致"视觉优美但检测性能平庸"的困境。为了破解这一难题,研究重点已从像素级的低层次增强转向语义级的深度融合。例如,Liu等(2022)通过构建了一个双层优化框架,用于建立图像融合与目标检测之间的内在关系。DetFusion(Sun等,2022)提出了一种检测结果反馈的共享注意力机制,利用这种机制引导融合网络关注对分类和定位更有利的显著区域,该设计有效提升了融合质量与检测精度。文中提出的基于目标感知的损失函数在学习位置信息方面也发挥了关键作用。MoE-Fusion(Cao等,2023)则引入混合专家模型,通过协同局部特征与全局对比分析,实现了异构信息在动态场景下的自适应整合,从而有效提升特征表示能力与检测性能。这类方法通过将检测任务的反馈信号反传至融合网络,推动了图像融合与目标检测从"分段式级联"向"一体化深度耦合"的范式转变。

然而,这些模型大多建立在"空间强对齐"的理想假设之上,现有的模型算法虽然在高度对齐的实验数据集上取得了突破,但在真实场景中仍面临诸多问题。在无人机机动飞行的过程中,不同载荷相机安装位置的差异、机身的高频震动以及观测距离改变引起的视差,这些因素会使得异源图像中出现

目标细节缺失、模态间特征非线性空间错位以及模态内可靠性波动等问题。现有模型在处理此类非理想输入时，往往因特征失配而导致检测性能断崖式下降。

尽管部分研究（如UMFusion（Wang等，2022）、MURF（Xu等，2023）、SuperFusion（Tang等，2022））尝试引入弱配准机制，但其通常仅能应对小尺度的线性偏移，难以表征真实航拍场景中复杂的多自由度几何畸变。这种“模型预设与现实退化”的不兼容现象，根源在于缺乏支持非对齐场景训练的高质量基准数据集。

1.2 IR-VIS目标检测数据集

表1汇总了现有主流IR-VIS目标检测数据集的统计特性。对比分析结果表明，该领域算法泛化性上的瓶颈主要受限于现有数据集在目标尺度分布、对齐精度及环境多样性方面的局限。这些局限性主

表1 现有数据集统计与比较

Table 1 Statistics and comparison of existing datasets

数据集名称	数据模态	图像分辨率	数量规模	类别数量	是否对齐	标注形式	目标相对大小	目标绝对大小	年份
KAIST（Hwang等，2015）	光学/红外	640×512	95328	3	√	HBB	0.091±0.302	52.1±7.2	2015
DLR 3K(Liu等，2015)	光学	5616×3744	20	2	–	OBB	–	–	2015
VEDAI(Razakarivony等，2015)	光学/红外	1024×1024	2420	9	√	OBB	0.037±0.013	18.4±6.4	2015
COWC(Mundhenk等，2016)	光学	2048×2048	32716	1	–	HBB	–	–	2016
CARPK(Hsieh等，2017)	光学	1280×720	1448	1	–	HBB	–	–	2017
UAVDT(Du等，2018)	光学	1080×540	80000	3	–	HBB	–	–	2018
VisDrone(Zhu等，2018)	光学	2000×1500	10209	10	–	HBB	–	–	2018
DOTA(Xia等，2018)	光学	12029×5014	2806	15	–	OBB	–	–	2018
FLIR-aligned(Zhang等，2020)	光学/红外	640×512	18944	3	√	HBB	0.071±0.266	40.4±6.4	2020
DroneVehicle(Sun等，2022)	光学/红外	640×512	56878	5	部分对齐	OBB	0.064±0.253	49.3±7.0	2021
LLVIP(Jia等，2021)	光学/红外	1080×720	30976	1	√	HBB	0.124±0.353	142.3±11.9	2021
M^3FD(Liu等，2022)	光学/红外	1024×768	8400	6	√	HBB	0.071±0.072	61.8±61.8	2022
RGBTDronePerson(Zhang等，2023)	光学/红外	640×512	12250	3	×	HBB	0.020±0.143	11.7±3.4	2023
DVTOD(Song等，2024)	光学/红外	1920×1080/640×512	4358	3	×	HBB	0.106±0.071	60.4±40.4	2024
DIV	光学/红外	1920×1080/640×512	10926	5	×	OBB	0.025±0.014/0.047±0.026	35.8±20.3/26.8±14.8	2025

注：目标相对大小和绝对大小显示的数值中，前者表示均值，后者表示标准差。OBB代表定向边界框，HBB代表水平边界框。

要体现在以下方面：

1） 尺度分布失衡与类别同质化

现有数据集在目标物理属性表征上存在明显的局限性。一方面，经典数据集（如KAIST（Hwang等，2015）、FLIR（Zhang等，2020））多聚焦于地面车载或固定视角场景，目标在图像中占据显著的像素比例。这种“大尺度主导”的数据分布不适用于无人机高空巡检中目标像素极度稀疏的典型特征，导致模型难以学习到微小目标在跨模态下的关键鉴别特征。另一方面，目标类别呈现严重的“单一种类聚集”现象，多局限于行人或特定车辆（如LLVIP（Jia等，2021）、RGBTDronePerson（Zhang等，2023））。在真实巡检任务中，目标类别具有高度的异质性，现有数据集的语义单一性限制了算法在复杂多目标场景下的分类鲁棒性。

2） 跨模态空间特征的“虚假完美”假设

为简化任务难度，LLVIP、M^3FD（Liu等，2022）等多模态数据集通过人工介入实现了可见光和红外图像像素级的完美对齐。这种处理掩盖了从无人机视角捕获的多模态图像之间固有的天然弱对齐关系。人工对齐无法反映真实场景中由传感器安装误差、视场差异及机身震动所导致的像素级空间错位。基于此类“虚假完美”数据训练的模型，缺乏对空间位置扰动的容忍度。一旦部署于存在物理视差的真实机载平台，异源特征在融合层会产生严重的语义干扰，从而导致检测模型在实际应用中面临严重的性能退化。

3） 动态环境下的质量非对称性

现有数据集往往缺乏极端环境的多样化覆盖，尤其是在昼夜交替等非对称质量失衡场景。在真实场景中，可见光模态在低光、傍晚及遮挡环境下失效，红外模态在热背景干扰下对比度会显著降低，这些情况在图像中往往呈现局部性分布，单张图像内部存在区域性质量差异。因此，面对此类动态且不均匀的质量失衡，融合与检测方法需要增强区域级的质量感知与融合能力。部分算法侧重于模态间的全局特征交互，使得模型在面对局部特征失效时难以实现鲁棒的自适应融合。

2 数据集构建与分析

为支持无人机视角下红外-可见光多模态目标检测研究，本文构建了一个无人机多模态目标检测数据集，如图2所示。该数据集通过真实无人机平台采集，涵盖多场景、多光照条件以及多个飞行高度，并在数据设计中保留了跨模态弱配准特性，为研究鲁棒的跨模态融合与目标检测提供了重要实验基础。

2.1 数据集设计

1） 物理属性真实性

在真实无人机平台中，由于传感器安装偏差、视场角差异以及飞行振动等因素，跨模态图像对之间存在不可避免的非线性空间偏差。如果在数据构建阶段强制进行完美配准，虽然能够简化算法设计，但会掩盖真实应用中的关键问题，从而导致模型在实际部署时性能显著下降。

本数据集坚持非强制配准原则，保留了原始成像中的天然弱配准状态。这种设计旨在真实还原无人机载荷在动态飞行中的空间偏移特性，挑战算法在面对模态间语义错位时的特征关联与纠偏能力，从而确保模型在实际部署中具备更高的环境适应性。

2） 多尺度覆盖

针对现有数据集目标尺度“极化”问题，以及航拍目标随无人机飞行高度实时波动的特性，本数据集放弃了单一尺度的采集模式，转而追求全尺度分布的完整性。考虑到安全飞行高度与广域搜索需求，无人机视角中的目标通常呈现极低像素占比与低信噪比特征。

我们在数据构建中特别强化了对微小和小型目标的采样权重：通过多种飞行高度（60m-120m）与不同俯仰角的组合采集，确保数据集中包含足够的像素占比较少的目标。统计显示，数据集中小目标及微小目标的占比超过70%。这一准则不仅还原了航拍任务中目标跨尺度剧烈波动的真实过程，也为突破远距离探测难题提供了高难度的验证基准。

3） 环境多样性

无人机全天时、全场景作业需要模型具备应对模态质量局部退化的能力。不同环境演化会直接改变可见光与红外模态的信噪比分布。

为此，本数据集跨越了城市居民区、复杂高速公路及植被茂密的丛林山区等多型地貌，并精细划分为日间、傍晚以及极低光照的夜间等时段。这种多样化的场景旨在模拟不同光热环境下模态可靠性的

图2　数据集设计与特征属性概述

Fig. 2　Overview of dataset design and feature attributes

动态波动，迫使融合算法从全局简单的特征堆叠转向具有区域感知能力的自适应融合，从而实现在极端情况下的鲁棒感知。

2.2　数据采集与筛选

2.2.1　无人机平台与传感器配置

数据采集采用DJI M300无人机平台，搭载集成式可见光与红外热成像传感器系统。

1）可见光传感器：采用1/2.3英寸CMOS传感器，具有1200万像素分辨率，配置82.9°视场角（field of view，FOV）与24 mm等效焦距，输出分辨率为1920×1080像素。

2）红外传感器：分辨率为640×512像素，配置热成像相机视场角为40.6°，帧率为30 fps，噪声等效温差（NETD）≤50 mK，光圈f/1.0。

2.2.2　采集环境与场景多样性

为确保DIV数据集的泛化性，采集过程涵盖了多种时空维度，具体属性和数据分布比例如表2所示。

2.2.3　数据筛选与质量控制

原始采集数据经过严格的三级筛选流程：

1）物理有效性剔除：移除因无人机剧烈机动导致的运动模糊难辨图像。

表2　DIV数据集的属性与分布统计

Table 2　Attributes and distribution statistics of the DIV dataset

属性	内容与分布
飞行高度	60 m–120 m（动态变化）
观测视角	垂直俯视+多角度倾斜（0°<θ<60°）
光照条件	白天、弱光环境、夜间
场景类型	城市/乡村道路、高速公路、居民住宅区、停车场

2）视觉可见度评价：剔除缺少或遮挡比例超过90%或在双模态中均无法辨识的目标区域。

3）弱对齐特征保留：保留具有典型非线性位移的图像对，以构建真实的“弱对齐”挑战环境。

最终筛选出5463组高质双模态图像对进入标注阶段。

2.3　标注协议与类别定义

2.3.1　标注准则

针对数据集，本研究确立了以下标注准则：

1）跨模态独立标注：不对两个模态进行标签共享，而是根据各个模态进行独立框选。

2）语义一致性约束：由于两种模态的分辨率不同，部分目标可能仅在单一模态中可见，本数据集仅

对两个模态均可观察到的共享目标进行标注。

3）非强制对齐：在目标受弱对齐影响产生位移时，以目标在不同模态图像中的位置为准进行标注，而非强制像素坐标对齐，保留真实的几何偏移。

2.3.2　标注工作流

本研究构建了“双模态标注-交叉核验-多层判别”的标准化工作流。标注人员首先在双模态同步视图下确认目标存在，随后分别在可见光和红外视图中进行独立框选。完成初步标注后，由另外两名标注员进行交叉核验，针对存在争议的微小目标或边缘模糊目标，由高级审核员进行最终判定。

2.3.3　标注格式规范

标注工作使用roLabelImg工具对图像中的目标进行标注，包括五类目标：car、van、truck、person、non-motorized vehicle。采用面向旋转目标的定向边界框（Oriented Bounding Box，OBB）进行数学描述。每个目标实例的标注框表示为一个五元组：$B=(cx,cy,w,h,angle)$。其中，cx,cy 代表标注框几何中心点的像素坐标。w,h 分别代表标注框的宽度与高度。$angle$ 代表标注框的旋转角度，旋转角度为弧度制，定义为水平轴顺时针旋转至矩形框长边所

表3　DIV数据集总览统计表

Table 3　DIV dataset statistics overview

属性	数值
图像总数	10926张（5463对）
实例总数	87140个
目标类别	car (44.5%)　person (24.8%)　van (2.5%) truck (10.8%)　non-motorized vehicle (17.4%)
平均每图实例数	8（单张最多81）
目标尺度	微小（39.4%）　小（31.5%）　中/大（29.0%）
光照分布	白天（60.6%）　弱光（28.7%）　夜间（10.7%）
场景类型	城市/乡村道路（30.9%）　高速公路（47.9%） 城市住宅区（18.8%）　停车场（2.5%）

形成的夹角，水平方向 $angle=0$，得到的角度值是正值，旋转一周为 π，没有负值。

2.3.4　标注校验与小目标处理

针对无人机视角下的小目标（例如，远距离person或non-motorized vehicle），标注团队在标注时给予特别关注和处理。在标注过程中增加放大审查与多次复核机制，以确保小目标被正确且完整标注。经过多轮人工校验，最终获得87140个目标实例。

表4　两个模态中各类别标注目标的数量统计

Table 4　Statistical analysis of annotation bias between different modalities

模态	car	non-motorized vehicle	person	truck	van
可见光	19363	7607	10637	4726	1092
红外	19368	7588	10937	4727	1095

2.4　数据分布与多维特性分析

为了建立DIV数据集的标准化参考基准，本节从数据规模、弱对齐特性、尺度分布、场景光照环境及目标密度五个维度进行系统化分析。

2.4.1　数据集统计概述

表3总结了DIV数据集的整体统计属性，为后续研究提供直观的数据基准。整个数据集涵盖了五个常见的目标类别。针对可见光和红外两种模态，各类别的具体实例数量如表4所示。其中，car、person、non-motorized vehicle占据主要比例，其次是truck和van。

2.4.2　跨模态弱配准与非对称偏移特性

与预对齐数据集不同，本数据集坚持真实性原则，保留了由传感器基线偏差、视场角失配及飞行

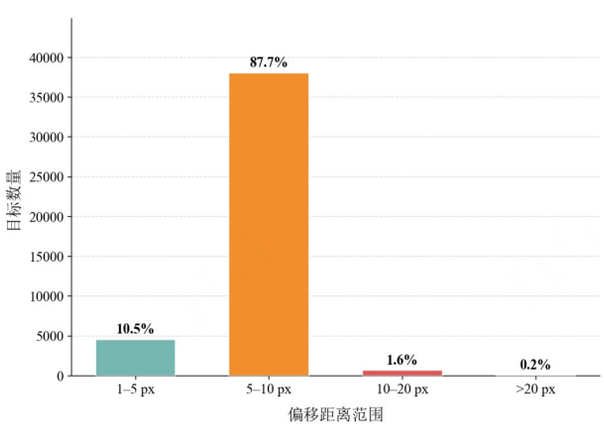


图3　跨模态图像对中目标中心点偏移分布

Fig. 3　Offset distribution statistics of target center points in cross-modal image pairs

平台高频振动引起的自然空间位移。为了进一步定量刻画“弱配准”特征，我们对数据集图像对的目标中心点偏移与交并比（Intersection of Union，

表 5　跨模态图像对 IoU 分布统计

Table 5　Statistical analysis of IoU distribution for multi-modal images

IoU范围	数量	占比
< 0.3	53	1.2%
0.3 – 0.5	2634	6.1%
0.5 – 0.8	11824	27.3%
> 0.8	28303	65.2%

IoU)进行了分布统计(见图3与表5)。此外,表6展示了可见光与红外边界框之间的几何失配情况。具体的空间偏移统计与影响机制分析如下:

1)　整体统计分布分析

图3详细展示了跨模态图像对中目标中心点的偏移距离分布。数据表明,约87.7%的目标中心偏移精确集中在5 – 10像素这一中度干扰区间,10.5%的目标处于1 – 5像素的微弱偏移范围,甚至有1.8%的目标偏移超过10像素。这一分布特征显示DIV中跨模态目标在空间位置上普遍存在偏移,这种弱配准特征更加贴近真实无人机平台下由传感器安装误差、飞行振动及视角差异所导致的自然空间失配。同时,表5统计了双模态标注框IoU分布情况。统计发现,IoU低于0.8的目标占比达到34.8%,其中IoU处于0.3-0.5的严重失配目标占比为6.1%。结合中心点偏移与IoU统计结果可见,DIV数据集包含了由视点差异引起的非线性几何形变。这种弱配准特性为评估多模态融合和检测模型在不同程度空间

表 6　不同模态间标注偏差的统计分析

Table 6　Statistical analysis of annotation bias between different modalities

类别名	Δx 范围	Δy 范围	Δw 范围	Δh 范围	中心点之间最大偏移距离	最小交并比(IOU)
car	[−26.00, 54.05]	[−20.50, 33.00]	[−14.76, 15.83]	[−11.11, 15.00]	54.33	0.105
non−motorized vehicle	[−15.00, 16.00]	[−17.86, 12.00]	[−8.11, 14.10]	[−8.80, 11.44]	22.41	0.113
person	[−12.36, 12.00]	[−13.64, 13.00]	[−10.00, 6.00]	[−13.04, 8.00]	17.09	0.102
truck	[−11.85, 18.00]	[−8.00, 13.00]	[−15.30, 18.09]	[−6.95, 13.03]	18.00	0.361
van	[−13.78, 9.00]	[−12.00, 19.00]	[−11.98, 13.15]	[−8.72, 9.91]	23.75	0.270

偏差下的容错感知能力提供了基础数据支撑。

2)　空间表征的去相关性

数据集中局部区域仍存在显著的几何错位。表5统计结果显示,所有类别的中心点偏移(Δx, Δy)与尺度偏差(Δw, Δh)均呈现出不同程度的随机波动。以car类别为例,其中心点最大偏移高达54.33像素,最低IoU仅为0.105。这表明未经空间重校准的前提下,同一目标的双模态特征在局部可能处于“弱相关”甚至“解离”状态,传统依赖严格像素配准的融合策略在此类数据上易产生定位漂移。

3)　不同目标的尺度敏感性

尽管person类与non-motorized vehicle类的绝对偏移数值较小,但其对异构特征融合的破坏性也不可忽视。对于无人机场景中的微小目标,轻微偏移即可能导致致命的区域错位。例如,针对尺寸仅为10×10像素的行人目标,即使是5像素的偏移,其意味着对于目标的相对偏移率也高达约25%。这种高比例位移易导致双模态特征在融合时产生“虚警偏移”,使得特征响应在空间维度上出现弥散现象。

4)　动态弱对齐的非线性特征

偏移量在横纵坐标上均表现出正负双向分布,且各类别间的偏移规律并非简单的线性平移,而是包含旋转、尺度缩放及动态视差在内的复杂非线性畸变。这种特性要求检测模型必须超越静态对齐假设,构建具备空间重采样能力(如可变形对齐或注意力引导对齐)的鲁棒特征提取网络,以应对真实的感知退化条件。

2.4.3　场景与光照分布

关于场景的覆盖范围,部分图像拍摄于城市街区和居民区,另一部分则拍摄于高速公路或山路,具体情况如图4所示。多样化的拍摄确保了该数据集

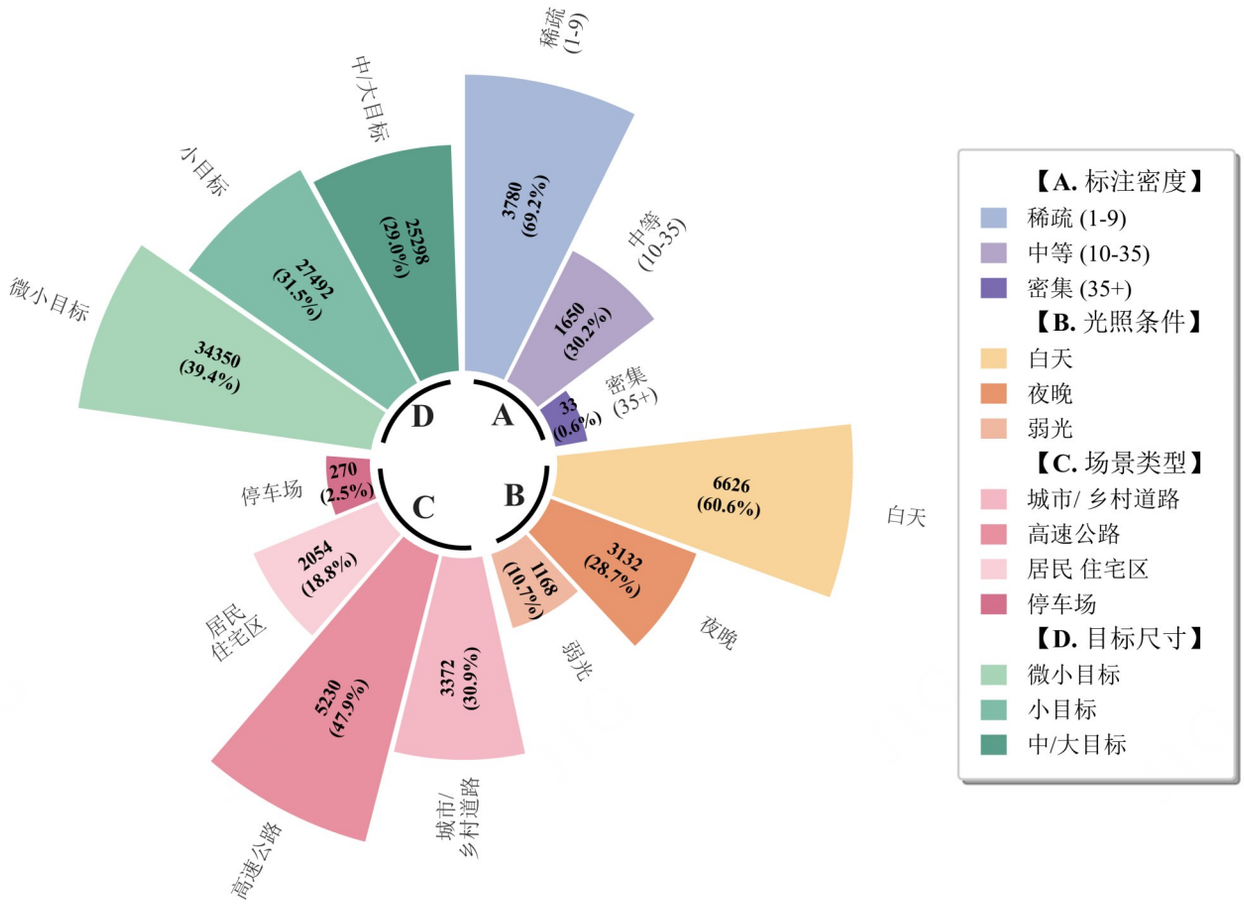


图4　标注密度、光照条件、场景类型及目标尺寸的数据分布

Fig. 4　Data distribution of object size, lighting conditions, scene type and annotation density

既涵盖目标密集的城市区域,同样也包含视野开阔、目标稀疏的山地和乡村地区,从而有效模拟了无人机在各种环境中执行巡检和监视任务的需求。根据采集时间,我们将数据划分为三类光照条件:白天(60.64%)、夜间(28.67%)、弱光环境(10.69%)。夜间与弱光样本合计占数据集总量的近40%,有助于整体评估可见光与红外模态在复杂光照条件下的表征性能差异。如图7所示,在低光条件下,可见光图像中的行人与车辆往往难以区分,而相应的红外模态则保持了强烈的对比度和可识别性。

2.4.4　目标尺度与密度分布

我们参考航拍多模态 TinyPerson 数据集的定义,将有效边长在[2, 20]像素间的目标定义为微小目标,[20, 32]像素间定义为小目标。图5中统计了各目标尺度在不同类别中的分布。结合图4中数据可以观察到,在20像素以下的微小目标在数据集中占比近40%,20-32像素的小目标占比超过30%,这为解决远距离探测下的弱小目标难题提供了充足的样本。与现有数据集不同,本数据集并非局限于小型目标,中、大型目标的数量同样占据数据集实例近30%。

此外,为了客观反映目标在图像中的显著程度,我们分析了目标的相对大小,即目标面积占据图像面积的比例。如图6(b)所示,图中分别展示了各类别分布曲线的覆盖范围,所有目标的均值分布在3‰附近(图中虚线所示)。这种多尺度共存的分布特性,

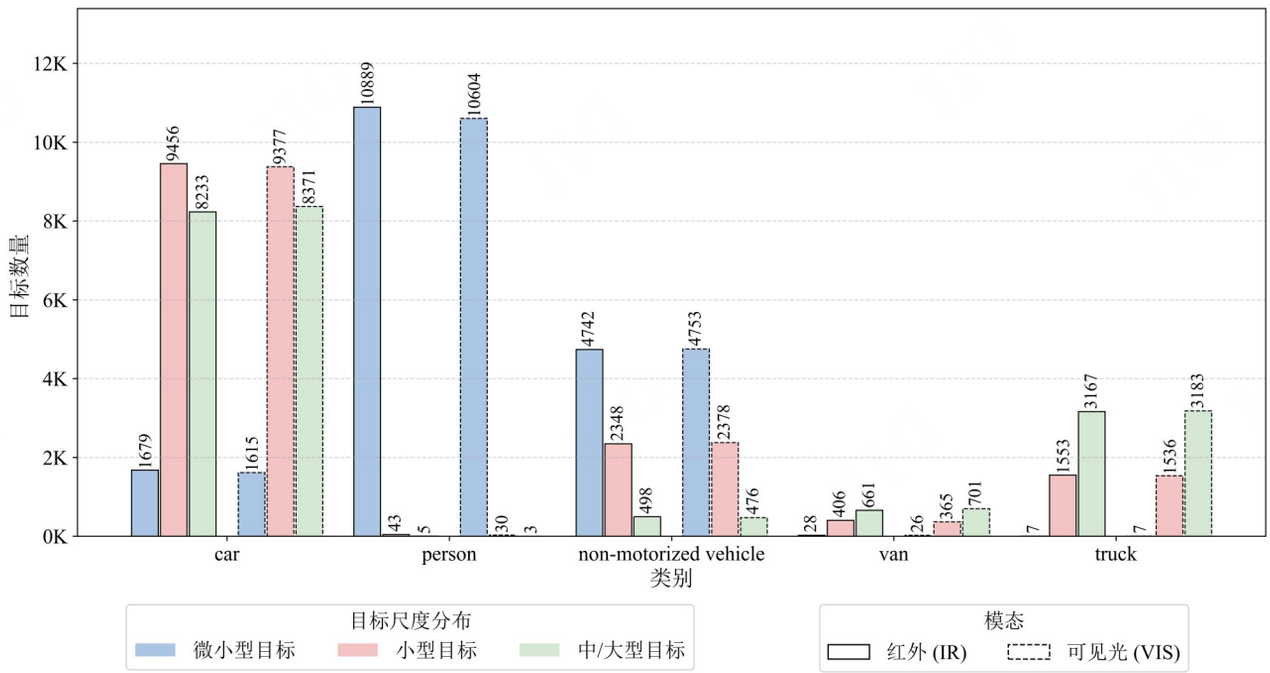
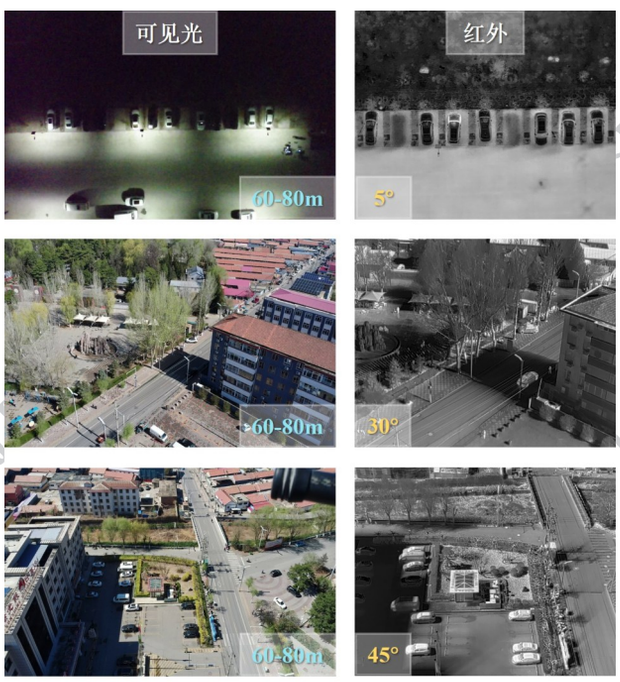
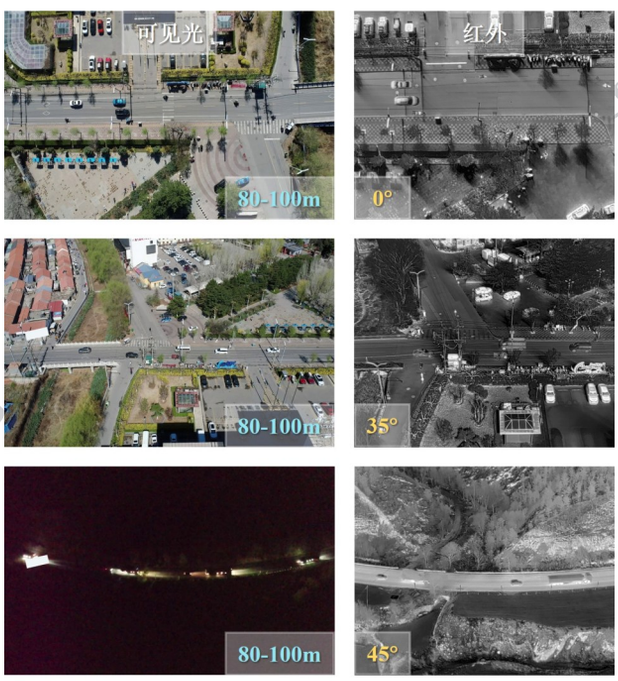


图5　不同模态各目标类别的标注分布。其中，实线表示可见光模态，虚线表示红外。柱内不同颜色则对应不同的目标尺度，柱状图上方数字表示各目标尺度具体数量。

Fig. 5　The distribution of annotations for each object category in different modalities. The solid line represents the visible light mode and the dashed line represents the infrared. Different colors in the column correspond to different object scales, and the numbers above the column indicate the specific number of object scales.



（a）　60-80m飞行高度下不同角度的图像示例

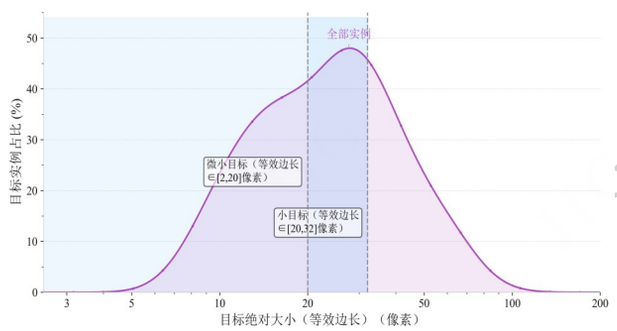


（b）　80-100m飞行高度下不同角度的图像示例

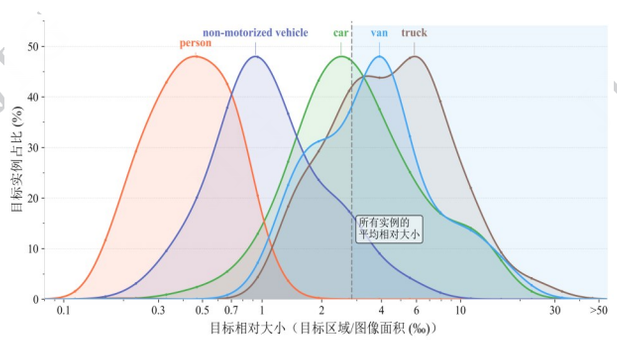
（c） 100-120m飞行高度下不同角度的图像示例

图7 数据集中不同高度与角度的图像示例

Fig. 7 Sample images at different heights and angles in the dataset ((a) sample images taken at different angles from a flight altitude of 60-80 meters; (b) sample images taken at different angles from a flight altitude of 80-100 meters; (c) sample images taken at different angles from a flight altitude of 100-120 meters)



（a） 绝对大小分布



（b） 相对大小分布

((a)distribution of absolute size; (b) distribution of relative size)

图6 目标实例大小的相对与绝对分布

Fig. 6 Relative and absolute distribution of object sizes

要求检测器克服对特定分辨率的依赖，在统一的特征提取框架下，同时具备对极微弱信号的捕获能力，以及对大尺度显著目标的精确回归能力，从而在真实多变的航拍任务中具备更强的空间适应性。

本数据集还涵盖了多种高度和角度下的采集条件，如图7所示。除了垂直俯视视角外，数据还包含在60至120米飞行高度下捕获的多种倾斜视角。这些不同的飞行高度和角度直接影响目标的尺度、形变以及遮挡模式。例如，高空垂直视图虽具有较高的物体密度，但细节信息有限。相反，低空斜视图则容易出现建筑物和物体的遮挡现象。这种多样性增强了数据的空间几何特征，为后续的目标检测及多视图鲁棒性分析研究提供了有力支持。

我们进一步分析了目标的密度分布，各密度范围占比如图4所示。我们将其划分为三个等级：稀疏（1-9个目标），中等（10-35个目标），密集（大于35个目标）。本数据集中每幅图像平均包含约8个目标，单张图像最多可达81个目标。高密度往往伴随着小尺度的聚集与相互遮挡。本数据集涵盖从低密度稀疏场景到高密度重叠场景的多种样本，配合前述的全尺度特征，为研究多模态模型在复杂环境下的检测鲁棒性提供了坚实的数据支撑。

2.5 典型IR-VIS数据集的统计对比

在多模态目标检测领域，已有多个典型的IR-VIS数据集推动了算法的发展。然而，通过对现有主流数据集与本文提出的DIV进行统计对比（见图8、图9和表7），可以发现当前研究仍面临显著的局限性。

2.5.1 指标定义与可视化分析

为了定量表征不同数据集在复杂环境下的任务难度，我们引入了两个关键统计指标：

1）尺度多样性（SD_{norm}）与散点图（图8）含义：

该指标基于目标像素占比的香农熵（Shannon Entropy）。首先计算目标面积与全图面积的比值 $r = \mathrm{Area}_{obj}/\mathrm{Area}_{img}$，随后在对数空间内进行分箱统计并求取信息熵，最后将其归一化[0, 1]区间。SD_{norm} 越接近1，意味着数据集内目标的尺度分布越广泛越

均匀，模型在检测时面临的跨尺度挑战越大。具体如图8所示：(1) 横轴：反映尺度变化的丰富度。数值越大(靠右)，说明目标大小跨度较大，对检测器的“多尺度建模”要求越高；(2) 纵轴：反映场景拥挤度。数值越高(靠上)，说明目标越密集；(3) 气泡大小则体现了数据集的总规模。

2) 对数面积占比分布与直方图(图9)含义：

为了更直观地观察目标在图像中的绝对大小分布，采用的分布直方图展示目标面积占比的对数化特征，具体如图9所示：(1) 横轴：采用对数坐标以涵盖跨越多个数量级的尺度差异。例如，数值为-1代表较大目标(占全图10%)，而数值为-5代表极小目标(占全图0.001%)；(2) 纵轴：落入对应面积区间的目标实例数量。波峰越偏向左侧，说明该数据集越倾向于“小目标检测”任务。

2.5.2 挑战属性维度

除统计分布的差异，表7从平台视角、对齐精细度及标注机制等方面进一步揭示了DIV的特性：

1) 视角与拍摄角度：现有数据集多采用单一平视(如KAIST，LLVIP)或垂直俯视(如VEDAI，

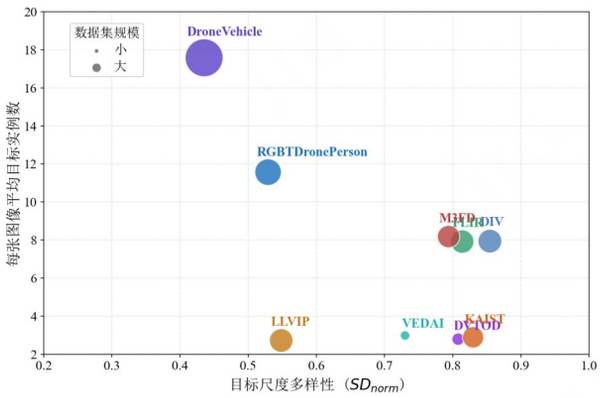


图8 DIV与其他多模态目标检测数据集的目标尺度与密度对比

Fig. 8 Comparison of object scale and density between DIV and other multimodal object detection datasets

DroneVehicle)。DIV突破了单一视角限制，涵盖了垂直俯视与多个倾斜视角。这种多角度组合引入了剧烈的视角诱发畸变，更真实地模拟了无人机在动态巡检中的工作状态。

2) 标注机制：大多数据集(如KAIST，M^{3}FD)采用共享标签，即在对齐图像基础上仅标注单个模态。

表7 主流IR-VIS数据集挑战属性对比

Table 7 Comparisons of challenge attributes across mainstream IR-VIS datasets

数据集	平台视角	目标尺度跨度 (SD_{raw}/SD_{norm})	拍摄高度/角度	对齐精细度	是否共享标签	任务和挑战
KAIST	地面车载	2.757/0.830	平视	精确对齐	是	地面多类行人检测
VEDAI	高空卫星	2.425/0.730	垂直俯视	精确对齐	是	高空多类车辆检测
FLIR-aligned	地面车载	2.702/0.813	平视	精确对齐	是	地面多类目标检测
LLVIP	地面监控	1.820/0.548	平视	精确对齐	是	地面单类行人检测
DroneVehicle	无人机(高空)	1.446/0.435	垂直俯视	部分对齐(65%)	否(双模态独立标注)	高空多类车辆检测
M^{3}FD	地面多场景	2.637/0.794	平视	精确对齐	是	地面多类目标检测
RGBTDronePerson	无人机(高空)	1.758/0.529	垂直俯视	未对齐	是(红外为主)	高空多类行人检测
DVTOD	无人机(低空)	2.682/0.807	平视	未对齐	是(红外为主)	地面多类目标检测
DIV	无人机(高空)	2.838/0.854	垂直俯视、多个倾斜视角	未对齐	否(双模态独立标注)	高空多类目标检测

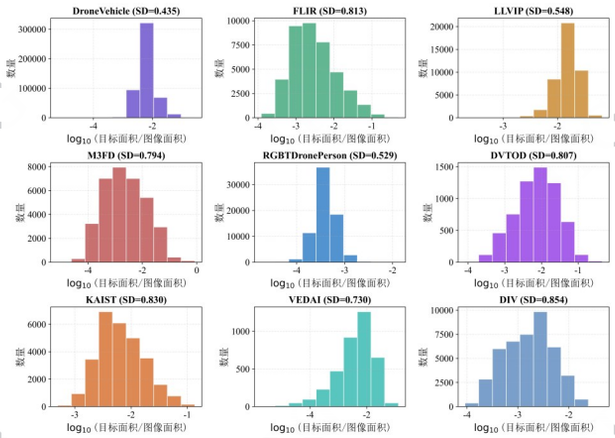


图9　DIV与其他多模态目标检测数据集的目标实例分布对比

Fig. 9　Comparison of object instance distributions between DIV and other multimodal object detection datasets

然而，表7显示DIV采取了双模态独立标注且不共享标签的策略。这种设计保留了可见光与红外模态在真实环境中因传感器视差、物体热特征差异导致的成像不对称性，能有效评估模型在非对齐状态下的特征对齐与融合能力。

3）对齐精细度：相比于追求“精确对齐”的数据集，DIV明确标注为“未对齐”。这要求算法必须具备处理空间错位的能力，更符合真实硬件平台部署时的实际情况。

2.5.3　任务复杂度

结合统计数据，DIV数据集在数据属性及任务难度方面展现出显著优势：

1）尺度覆盖的广度：如图8所示，DIV的 SD_{norm} 高达0.8543，显著高于大规模俯拍数据集DroneVehicle（0.4352）。从图9的分布曲线可见，DIV呈现出一种宽广且平滑的形态，覆盖了从10^{-4}到10^{-1}的全量程范围。这种多尺度覆盖要求模型必须同时具备极小目标捕捉与大尺度目标语义建模的能力，其难度远高于尺度分布单一的地基视角数据集。

2）场景密度的科学平衡：在平均密度方面，DIV（7.96）位于高度密集（DroneVehicle，17.60）与极度稀疏（LLVIP，2.74；VEDAI，3.00）之间。DIV的密度更符合真实巡检中的自然分布，有效评估算法在处理复杂空间关系时的准确性，避免了模型过分拟合于超高密集或稀疏场景而导致的泛化性弱化。

通过上述对比分析，我们识别出当前IR-VIS目标检测研究中存在的关键问题：

1）视角与几何畸变：现有基准多集中于地面视角、固定俯视，无法模拟UAV检测中因高度与偏航角动态变化引起的几何失真。

2）尺度压缩与丢失：典型无人机视角数据集往往只涵盖特定的尺度区间，缺乏对极小目标到大目标平滑过渡的建模，导致模型在应对突发尺度变化时性能下降。

3）对齐程度与质量：在真实无人机作业中，多模态信息常伴随弱对齐和动态质量退化，现有预对齐数据集对此类真实挑战建模不足。

DIV的设计初衷是为了弥补这些空白，旨在解决真实UAV环境下严重的尺度压缩、视点诱发的几何差异以及跨模态弱对齐等核心难题，为更具鲁棒性的多模态目标检测算法提供客观、具有挑战性的评价标准。

2.6　数据集挑战

为解决现有数据的瓶颈问题，我们在设计数据集时特意保留了无人机场景中常见的典型难题，旨在推动跨模态融合和目标检测领域出现新的范式，使其具备更强的鲁棒性和泛化能力。以下总结了其中最突出的三个挑战。

2.6.1　多尺度覆盖下的特征学习与鲁棒性挑战统计

数据统计结果表明，微小目标占本数据集标注目标的40%，且小目标（包含微小目标）占比超过总目标70%。无人机高空视角导致目标尺度极度压缩，可见光图像中的纹理信息严重缺失，而红外图像中的热信号又容易被背景热噪声淹没。这使得目标检测面临低信噪比与低对比度问题。此外，无人机俯视和斜视的视角加剧了物体密集排列和相互遮挡的问题。因此，亟需开发新型的融合策略，以从信噪比极低的红外信号和微弱的可见光轮廓中提取并整合互补特征。这要求检测模型在尺度变化方面具备优越的鲁棒性，并对背景噪声具有强大的抑制能力。

与现有聚焦于单一尺度或预配准场景的数据集不同，本数据集特意保留并覆盖了从极低像素占比的微型目标到近距离显著区域的中、大型目标。这种多尺度目标检测，不仅要求模型具备优越的尺度感知能力，更需在特征融合阶段实现针对不同尺度目标的非对称容错机制。

2.6.2　动态环境中的区域级模态可靠性不均衡

在夜间或弱光环境中，可见光模态可能完全失效，而红外模态仍然有效。同时，在同一图像中还可能出现局部过曝或热干扰现象。因此，多模态检测算法需要具备区域级质量感知与动态融合能力。

该数据集是在多种照明条件下及多个场景中采集的，真实反映了多模态数据在不同环境下动态变化的可靠性。核心挑战在于模态可靠性的空间非均匀性。（例如，夜间红外图像可靠而可见光图像基本无效。）其次，也表现为模态内异质性。（例如，可见光图像在特定区域可能出现局部过曝或极度昏暗。或在红外图像中，沥青道路和建筑墙壁等背景热源会干扰前景物体。）

传统的平均融合或固定权重融合策略因无法感知这些区域级别的质量差异，其性能会下降。为应对这一挑战，提出的新算法应具备自适应的质量感知机制，能够动态评估每个特征块在区域层面的可用性。

2.6.3　传感器固有偏差导致的弱配准与特征错位

在数据采集和预处理阶段，由于本数据集保留了真实的跨模态偏移，这样的跨模态位置偏移问题对基于数据完美配准假设的检测模型构成了直接挑战。尽管全局偏移可能仅涉及几个像素，但对于极小的物体而言，即使是微小的空间错位也会被放大。这导致同一物体在可见光和红外模态下的边界框位置及特征中心存在显著差异。这种特征错位严重削弱了后续特征配准和跨模态特征融合的有效性。因此，未来研究应优先开发对配准误差具有高容忍度的融合策略，并探索在弱对齐条件下的鲁棒目标检测方法。例如，设计容错型特征聚合模块或隐式空间变换网络，能够在特征层面上有效补偿或缓解跨模态数据中的几何错配。

3　基线结果比较与分析

在本节中，使用目标检测中常见的评价指标，我们在所提数据集上对比了9种近年来在多模态可见光-红外图像融合与检测领域的代表性主流基线方法，具体包括：CFT（Fang等，2022）、CSSA（Cao等，2023）、CALNet（He等，2023）、ICAFusion（Shen等，2024）、CMADet（Song等，2024）、C^{2}Former-S^{2}Anet（Yuan等，2024）、MMIDet（Zeng等，2024）、E2E-MFD（Zhang等，2024）、SM3Det（Li等，2026）。文中对上述算法在提出的数据集上进行了重训练，并进行了统一的评估测试，以确保对比实验的公平性与有效性。

3.1　基线方法结果

3.1.1　定量分析

表8展示了9种主流多模态检测方法在DIV数据集上的性能比较结果。实验结果显示，SM3Det在综合评估指标mAP_{50}上取得了50.5%的最佳成绩。这主要归功于其创新的网格级稀疏混合专家（Grid-level Sparse MoE）架构，该架构允许模型在统一的主干网络中，针对不同模态的局部特征动态调用专用专家，从而高效地提取跨模态共享知识并保留模态特有属性。同时，其动态学习率调整策略有效平衡了不同模态与任务间的梯度冲突，验证了其在复杂遥感场景下的整体优越性。

从类别维度深入分析，C^{2}Former-S^{2}ANet和E2E-MFD在person和van类别中表现尤为突出。C^{2}Former架构通过自适应特征采样预测模态间的空间偏移，并利用跨模态交叉注意力在特征层级实现精细的对齐与补全，其特征校准机制在处理弱对齐约束下微小目标时具有鲁棒性。E2E-MFD方法采用了端到端同步融合检测架构，利用其由粗到细的扩散检测头强化了对目标区域的关注。通过梯度矩阵任务对齐确保了融合任务与检测任务的协同优化，使其在处理多尺度目标时具备更强的语义捕获能力。对于non-motorized vehicle和truck等类别，SM3Det展现出显著优势。其MoE结构能够有效整合红外模态的热显著性与可见光模态的几何轮廓，通过参数空间的动态分配，解决了不同目标尺度极化带来的特征建模难题，彰显了其在多任务统一建模下的强大适应性。

通过对表8中基线方法的性能指标进行横向对比可见，所有模型在person类别上的平均精度显著低于car、truck等中、大尺寸目标类别。以整体性能较优的SM3Det和E2E-MFD为例，其car类别AP分别达到83.3%和76.6%，truck类别AP分别达到83.1%和77.4%。而SM3Det方法在person类别上的AP仅为9.27%，性能差距超过70%，E2E-MFD方法在person类别上的精度虽然为所有对比方法的次优结果，但AP结果也仅为21.3%。本研究认为，这种性能塌陷并非偶然，而是由无人机视角和弱对齐

场景下的多种问题共同导致：

1） 极小尺度与弱特征表现的叠加效应

统计分析表明，person类别在DIV数据集中的等效边长像素平均值仅为12.24像素，且微小目标占比高达99%。在无人机高空视角下，person目标在可见光模态中极易受光照不足、遮挡或复杂地物背景的干扰，导致纹理缺失；而在红外模态中，由于人体热辐射在远距离成像时边缘模糊，难以形成清晰的语义中心。这种“双模态弱特征”属性使得模型在特征提取阶段便面临极高的漏检风险。

2） 弱对齐对小目标的“空间位移放大”机理

在弱对齐场景下，红外与可见光图像之间存在的非线性空间偏移对于不同尺度的目标具有不同的影响权重。对于中、大尺寸目标（如truck），5像素的偏移在其整体面积中占比极小，模型仍能通过重叠区域提取到共性特征；但对于等效边长仅为10像素左右的person目标，5像素的偏移意味着跨模态语义中心的相对位移偏差高达50%。这种剧烈的空间错位导致基于理想对齐假设的融合网络无法在像素层级实现有效的特征对齐，进而引发特征空间的严重解耦，在融合过程反而引入了噪声，抑制了检测精度。

3） 复杂环境下的模态依赖

在夜间或强遮挡场景下，person类别的检测高度依赖红外模态提供的显著性热信号。然而，由于DIV数据集保留了真实的弱配准特性，当红外模态识别到行人热红外特征时，其对应的可见光区域往往是一片黑暗或错误的背景。在这种情况下，现有的双模态融合模型往往会因为无法对齐两个模态的检测框，而在预测阶段产生严重的定位漂移或双重重影现象，最终反映在评价指标上即为person类别极低的检测精度。

综上所述，person类别在DIV数据集上的低性能表现，定量地印证了真实无人机航拍场景下目标尺度变换和“弱对齐”问题的严峻性，也进一步凸显了构建DIV数据集以推动抗偏移多模态融合和检测算法研究的必要性。

3.1.2 定性分析

我们对数据集中的代表性场景进行了可视化分析，如图10所示。图中展示了这些模型在涵盖白天、夜间和低光等不同光照条件，以及城市区域、山区高速、乡村道路、停车场等复杂环境下的代表性检测结果。同时所选场景涵盖了多种具有挑战性的条件，例如运动模糊（如高速物体移动导致的模糊效果）以及复杂遮挡（如建筑物和树木的阴影）。此外，无人机以大倾斜角度拍摄导致的几何畸变和尺度缩放问题，也给检测模型带来了更大的分类与定位挑战。

在目标密集的城市场景下，SM3Det表现出了较强的鲁棒性，该模型能够根据局部区域的视觉质量，动态激活不同的专家模块，并且利用动态学习率调整策略有效平衡了不同模态与任务间的冲突。这样的设计增强了模型感知不同目标结构的能力，更有效地弥补了复杂场景中跨模态特征的语义缺陷。C^2Former利用自适应特征采样与跨模态交叉注意力模块在特征层级实现了空间校准，能够自动寻找并对齐异源模态间的关联像素点，确保了在目标位置发生错位时依然能生成精确的边界框。该模型在一定程度上避免了因缺乏对齐机制导致两个模态的检测框错位漂移，减少了定位误差。此外，得益于梯度矩阵任务对齐策略，E2E-MFD的融合图像能为检测任务提供更具判别性的语义特征。不同于传统的回归方式，E2E-MFD基于扩散模型的生成式检测机制能从高斯噪声中通过迭代引导恢复出清晰的目标轮廓，使得模型在背景杂乱的情况下，也能保持稳定预测。

综上所述，SM3Det、C^2Former以及E2E-MFD等方法都在多模态目标检测领域展现了显著的改进效果。SM3Det通过混合专家架构实现了多尺度特征的动态适配；C^2Former通过自适应特征采样与交叉注意力机制缓解了模态间的空间错位；E2E-MFD则利用端到端同步优化策略提升了检测任务对融合过程的引导能力。然而，尽管这些模型在特定维度上表现出色，但在应对极端复杂环境时的综合感知能力仍显不足。在无人机遥感等实际应用中，多模态检测依然面临着多重严峻挑战。在真实飞行环境中，无人机检测面临的挑战往往是复合存在的，如高空

微小目标导致的特征极度压缩、纵横比剧烈变化、复杂背景下的目标相互干扰以及不同模态特征失配等。为应对这些极端情况，仍需要构建一套多维度的协同优化体系，旨在全方位强化算法在复杂挑战场景下对多模态信息的提取与融合能力，并提升无人机平台在动态环境下的检测效能。

表 8　基线方法在 DIV 数据集上的检测结果

Table 8　Detection results of baseline methods on DIV dataset

方法	car	non-motorized vehicle	person	truck	van	mAP_{50}
CFT (Fang 等，2022)	**86.9**	42.3	14.2	76.8	10.9	46.2
CSSA (Cao 等，2023)	84.4	49.3	16.5	76.6	14.6	48.3
CALNet (He 等，2023)	82.3	50.0	14.5	74.1	12.6	46.7
ICAFusion (Shen 等，2024)	85.6	34.0	10.1	69.0	13.6	42.5
CMADet (Song 等，2024)	81.6	40.8	10.2	60.4	9.29	40.5
C^2Former-S^2ANet (Yuan 等，2024)	75.7	27.4	**24.4**	74.4	22.8	44.9
MMIDet (Zeng 等，2024)	85.4	50.5	16.6	77.5	16.6	49.3
E2E-MFD (Zhang 等，2024)	76.6	47.1	21.3	77.4	**25.0**	49.5
SM3Det (Li 等，2026)	83.3	**64.3**	9.27	**83.1**	12.8	**50.5**

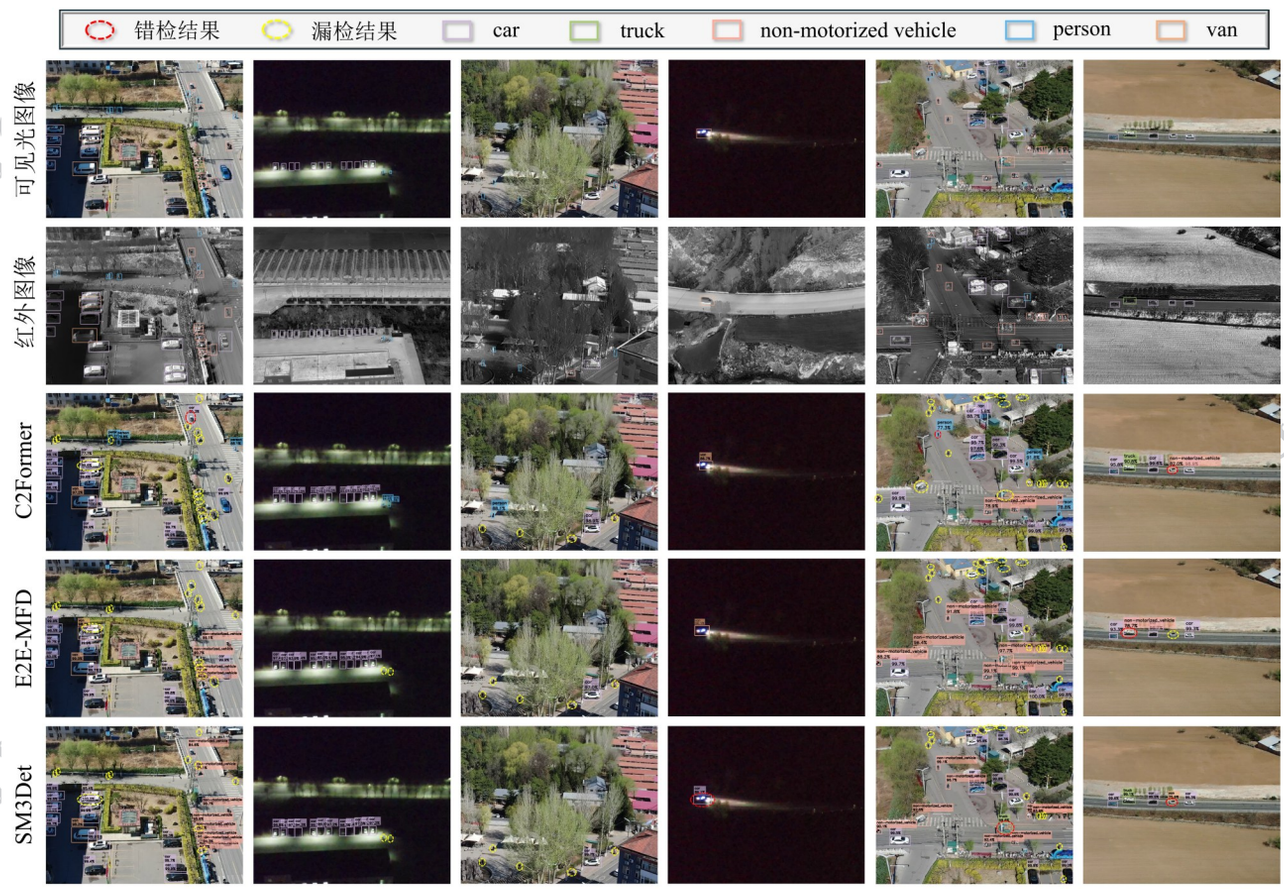


图 10　数据集中不同高度与角度的图像示例

Fig. 10　Sample images at different heights and angles in the dataset

3.1.3　失败案例分析

为了进一步分析现有多模态目标检测方法在真实弱配准场景中的局限性，本文对误检与漏检案例进行了可视化分析，如图 11 所示。结果表明，现有方法在处理跨模态空间错位、微小目标以及复杂背景退化等条件时，仍存在明显性能瓶颈。

1）　弱配准引起的目标定位不稳定

在无人机飞行过程中，由于传感器安装偏差、平台振动及视角变化等因素影响，可见光与红外图像之间往往存在明显空间错位。在这种情况下，传统

基于严格空间对应关系的融合策略容易产生跨模态语义关联失稳问题，导致模型无法准确建立目标区域间的一致性映射关系。如图11中蓝色虚框所示，其结果表现为检测框偏移、目标边界漂移以及局部区域误匹配等现象。

2） 微小目标条件下的跨模态语义解耦

微小目标由于像素占比极低，本身仅包含有限的纹理与结构信息。当跨模态空间偏移进一步存在时，不同模态之间原本有限的语义对应关系会被进一步削弱，导致目标特征难以有效聚合。图11中黄色和红色虚框可以观察到，部分行人及远距离车辆目标出现严重漏检和误检现象。这表明，弱配准与微小目标感知之间存在明显耦合效应，现有方法在低信噪比条件下仍缺乏稳定的跨模态特征建模能力。

3） 复杂背景与模态质量非对称退化

在夜间、弱光、热干扰或复杂背景条件下，不同模态的局部区域往往呈现明显质量差异。例如，可见光模态可能受到低照度或阴影影响，而红外模态则可能受到高温背景或热源干扰。在此条件下，现有全局融合策略容易错误增强低质量区域特征，从而引发背景误激活与类别混淆问题，如图11中失败案例所示，各个场景都存在不同程度的误检漏检现象。这表明，当前多模态检测方法在复杂环境中的区域级模态可靠性感知能力仍存在明显不足。

上述失败案例进一步表明，真实无人机场景中的弱配准、多尺度变化及复杂环境退化并非独立因素，而是具有明显耦合关系。现有大量多模态检测方法仍高度依赖理想像素级对齐假设，其特征融合机制在真实退化条件下的鲁棒性与泛化能力仍有较大提升空间。

3.2 跨数据集泛化性与弱对齐鲁棒性分析

为了深入探究跨模态弱对齐及成像视角对IR-VIS多模态融合和检测性能的具体影响，本文选取了M^{3}FD、DroneVehicle以及提出的DIV数据集进行对比实验，旨在分析不同检测框架在真实弱对齐环境下的的鲁棒性及其对对齐条件的依赖程度。如表1与表7所示，这三个数据集在成像视角、目标尺度分布及对齐特性上构成了从“理想对齐”到

“真实弱对齐”的挑战：M^{3}FD：代表精确对齐的地面多场景视角；DroneVehicle：代表部分对齐的无人机高空俯视场景。DIV：涵盖了高空与低空视角

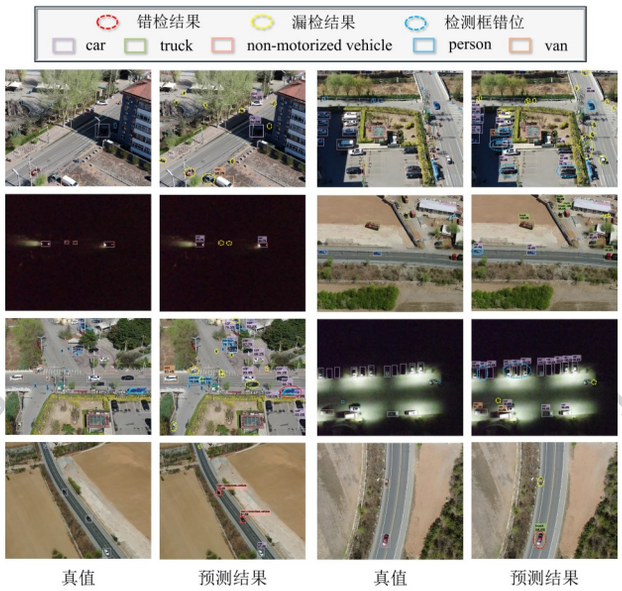


图11 不同场景失败案例

Fig. 11 Failure cases in different scenarios

下，受无人机平台振动、视角不一致及传感器安装偏差影响的真实弱对齐场景。通过对表8、表9及表10中的实验结果进行综合分析，可以发现：

1） 视角、尺度与对齐程度的影响导致性能坍塌

实验结果显示，各基准方法在理想对齐的M^{3}FD上性能优异，最高mAP_{50}达85.1%。当场景切换至DroneVehicle的无人机俯视角时，虽然检测难度增加，但由于目标尺度变化相对较小且仍存在部分对齐支撑，先进方法的检测精度仍能达到约82.3%。然而，在完全处于未对齐状态且目标尺度分布更广泛的DIV数据集上，所有方法的性能均出现断崖式下跌，最高mAP_{50}仅为50.5%。这证明了当前主流检测框架高度依赖“像素级精确对齐”的理想假设，真实场景下的弱对齐与剧烈的尺度变化会导致现有特征融合机制严重失效。

2） 弱对齐与微小目标感知的退化效应

在DIV数据集上，person类别的检测性能下降最为显著，最低仅为9.27%。相比于地面视角M^{3}FD的行人目标和DroneVehicle的车辆目标，DIV中的person目标在无人机视角下像素占比极小。在缺失对齐先验的情况下，跨模态间的空间偏差影响较大偏差范围甚至大于在目标本身的尺寸，导致红外与可见光特征在语义层面上严重失配，产生严重的定

位漂移与漏检。这表明弱对齐与微小目标感知之间存在强烈的耦合效应，是限制真实场景检测精

表 9　基线方法在 M^{3}FD 数据集上的检测结果

Table 9　Detection results of baseline methods on M^{3}FD datasets

方法	People	Car	Bus	Lamp	MotorCycle	Truck	mAP_{50}
CFT	82.2	91.3	91.6	85.1	73.6	86.1	85.0
CSSA	87.2	91.9	89.9	65.3	70.3	85.0	81.6
CALNet	87.4	91.7	88.5	64.1	69.9	83.0	80.8
ICAFusion	83.0	91.0	92.3	85.0	76.5	88.9	85.1
CMADet	88.2	92.2	89.6	70.8	73.2	85.1	83.2
C^{2}Former-S^{2}ANet	70.1	81.0	84.3	68.5	57.5	76.3	73.0
MMIDet	80.6	90.7	89.5	83.3	70.7	85.9	83.5
E2E-MFD	71.4	79.9	86.8	63.1	70.8	83.5	75.9
SM3Det	70.7	82.1	90.6	73.1	70.9	79.6	77.8

表 10　基线方法在 DroneVehicle 数据集上的检测结果

Table 10　Detection results of baseline methods on DroneVehicle datasets

方法	car	freight car	truck	bus	van	mAP_{50}
CFT	81.3	57.9	71.5	86.2	53.8	70.1
CSSA	98.4	67.5	79.2	95.1	61.3	80.3
CALNet	90.2	60.9	73.8	88.7	51.6	73.0
ICAFusion	98.5	74.6	76.4	96.5	65.4	82.3
CMADet	98.2	70.4	78.3	96.8	66.4	82.0
C^{2}Former-S^{2}ANet	90.2	64.4	68.3	89.8	58.5	74.2
MMIDet	98.4	72.8	77.6	96.6	63.1	81.7
E2E-MFD	90.3	64.6	79.3	89.8	63.1	77.4
SM3Det	97.3	66.5	81.4	94.4	60.8	80.1

度的核心瓶颈。

3）高级融合策略的优势随对齐约束减弱而受限

对比发现，在其他两个数据集中表现突出的复杂融合和检测策略，其领先优势在DIV上大幅缩水，不同方法间的性能差距明显减小。这揭示了现有高级算法的技术路线过于依赖对齐先验来建立特征交互。一旦对齐约束退化，模型难以仅通过特征交互补偿空间位移，导致其融合优势难以充分发挥。

4　结论与展望

4.1　研究总结与学术价值

本文构建了一个面向真实无人机场景的弱配准IR-VIS多模态目标检测数据集DIV。与现有大多数基于理想对齐假设的数据集不同，DIV保留了真实飞行环境中由传感器偏差、飞行姿态变化及平台振动所引起的跨模态空间错位，从而更加真实地反映了无人机多模态感知任务中的复杂退化条件。实验结果表明，现有高度依赖严格空间对齐假设的多模态检测方法，其特征融合策略在真实弱配准条件下容易出现跨模态关联能力下降与目标定位不稳定等问题。因此，本文认为，未来多模态感知研究的重点不应仅停留于理想条件下的特征融合与目标检测，而应进一步转向面向真实退化环境的鲁棒感知与语义一致性建模。DIV的提出，为研究弱配准条件下的跨模态几何鲁棒性、动态模态可靠性评估以及任务驱动的一体化感知框架提供了新的数据基础。

4.2　未来研究展望

1）面向弱配准条件的鲁棒语义一致性建模

针对传感器基线偏差与平台震动导致的跨模态位置不可避偏移问题，现有算法过度依赖像素级精确对齐先验，在处理非对称位移数据时存在严重的“信息污染”。未来研究需要进一步探索弱配准条件下的容错型特征关联机制，从“像素级对齐融合”逐步转向“语义级一致性建模”。

2）面向宽尺度与跨模态特征自适应重构

针对航拍微小目标在低信噪比环境下极易被背景噪声淹没以及目标尺度跨度大的问题，现有研究在处理异构特征时仍面临尺度依赖性与模态权重分配僵化的限制。因此，未来需要研究具备尺度鲁棒性的动态特征增强与跨模态语义补全机制，以提升

模型在极端尺度条件下的稳定感知能力。

3） 面向复杂环境退化的区域级模态可靠性感知

在低照度、局部过曝及热干扰等复杂环境中，不同模态的有效信息区域往往呈现明显非对称性，传统的全局加权融合方案在处理局部信息退化场景时已难以奏效。未来研究可进一步探索空间自适应的模态质量评估机制，实现区域级动态融合与按需特征分配，以应对复杂的质量非对称场景。

4） 面向真实任务的一体化协同感知框架

目前的图像融合与目标检测往往处于分段优化状态，这种阶段化脱节往往导致融合结果可能“视觉优美”但“性能平庸”，难以实现任务目标与特征表达之间的统一约束。未来需要进一步探索任务驱动的端到端协同感知框架，使多模态特征生成过程能够直接服务于下游检测任务需求。

参考文献（References）

Ma J Y, Ma Y and Li C. Infrared and visible image fusion methods and applications: A survey. Information fusion, 2019, 45: 153-178. [DOI: 10.1016/j.inffus.2018.02.004]

Tang L F, Xiang X Y, Zhang H, Gong M Q and Ma J Y. Divfusion: Darkness-free infrared and visible image fusion. Information Fusion, 2023, 91: 477-493. [DOI: 10.1016/j.inffus.2022.10.03 4]

Guo L, Rao P, Chen X and Li Y J. Infrared differential detection and band selection for space-based aerial targets under complex backgrounds. Infrared Physics & Technology, 2024, 138: 10517 2. [DOI: 10.1016/j.infrared.2024.105172]

Guo J J, Gao C Q, Liu F C, Meng D Y and Gao X B. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 464-481. [DOI: 10.1007/978-3-031- 73383-3_27]

Li H F, Yang Z Y, Zhang Y F, Jia W, Yu Z T and Liu Y. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47 (5): 3673-3690. [DOI: 10.1109/TPAMI.2025.3535617]

Chen C, Qi J H, Liu X Y, Bin K C, Fu R G and Hu X K. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 26836-26845. [DOI: 10.1109/CVPR52733.2024.02534]

Yuan M X, Shi X R, Wang N, Wang Y Y and Wei X X. Improving RGB-infrared object detection with cascade alignment-guided transformer. Information Fusion, 2024, 105: 102246. [DOI: 10.1016/j.inffus.2024.102246]

Li H and Wu X J. DenseFuse: A fusion approach to infrared and visible images. IEEE transactions on image processing, 2018, 28 (5): 2614-2623. [DOI: 10.1109/TIP.2018.2887342]

Li H, Wu X J and Durrani T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656. [DOI: 10.1109/TIM.2020. 3005230]

Li H, Wu X J and Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. Information Fusion, 2021, 73: 72-86. [DOI: 10.1016/j.inffus.2021.02.023]

Li H, Xu T Y, Wu X J, Lu J W and Kittler J. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(9): 11040-11052. [DOI: 10. 1109/TPAMI.2023. 3268209]

Li J W, Chen J S, Liu J Y and Ma H M. Learning a graph neural network with cross modality interaction for image fusion// Proceedings of the 31st ACM international conference on multimedia. 2023: 4471-4479. [DOI: 10.1145/3581783.361213 5]

Yue J, Fang L Y, Xia S B, Deng Y and Ma J Y. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. IEEE Transactions on Image Processing, 2023, 32: 5705-5720. [DOI: 10.1109/TIP.2023.3322046]

Ma J Y, Yu W, Liang P W, Li C and Jiang J J. FusionGAN: A generative adversarial network for infrared and visible image fusion. Information fusion, 2019, 48: 11-26. [DOI: 10.1016/j. inffus. 2018. 09.004]

Ma J Y, Xu H, Jiang J J, Mei X G and Zhang X P. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing, 2020, 29: 4980-4995. [DOI: 10.1109/TIP.2020. 2977573]

Ma J Y, Zhang H, Shao Z F, Liang P W and Xu H. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-14. [DOI: 10.1109/TIM. 2020.3038013]

Ma J Y, Tang L F, Fan F, Huang J, Mei X G and Ma Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica, 2022, 9 (7): 1200-1217. [DOI: 10.1109/JAS.2022.105686]

Zhao Z X, Bai H W, Zhang J S, Zhang Y L, Xu S and Lin Z D. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5906-5916. [DOI: 10.1109/CVPR52729.2023.00572]

Tang W, He F Z, Liu Y, Duan Y S and Si T Z. DATFuse: Infrared and visible image fusion via dual attention transformer. IEEE Transac-

tions on Circuits and Systems for Video Technology，2023，33(7)：3159-3172.[DOI：10.1109/TCSVT.2023.3234340]

Liu J Y，Fan X，Huang Z B，Wu G Y，Liu R S and Zhong W. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022：5802-5811.[DOI：10.1109/CVPR52688.2022.00571]

Sun Y M，Cao B，Zhu P F and Hu Q H. Detfusion：A detection-driven infrared and visible image fusion network// Proceedings of the 30th ACM international conference on multimedia. 2022：4003-4011.[DOI：10.1145/3503161.3547902]

Cao B，Sun Y M，Zhu P F and Hu Q H. Multi-modal gated mixture of local-to-global experts for dynamic image fusion// Proceedings of the IEEE/CVF international conference on computer vision. 2023：23555-23564.[DOI：10.1109/ICCV 51070.2023.02153]

Wang D，Liu J Y，Fan X and Liu R S. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration//Proceedings of the thirty-first international joint conference on artificial intelligence. 2022：3508-3515.[DOI：10.24963/ijcai.2022/487]

Xu H，Yuan J T，Ma J Y. Murf：Mutually reinforcing multi-modal image registration and fusion. IEEE transactions on pattern analysis and machine intelligence，2023，45(10)：12148-12166.[DOI：10.1109/TPAMI.2023.3283682]

Tang L F，Deng Y X，Ma Y，Huang J and Ma J Y. SuperFusion：A versatile image registration and fusion network with semantic awareness. IEEE/CAA Journal of Automatica Sinica，2022，9(12)：2121-2137.[DOI：10.1109/JAS.2022.106082]

Hwang S，Park J，Kim N，Choi Y and So Kweon，I. Multispectral pedestrian detection：Benchmark dataset and baseline// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015：1037-1045.[DOI：10.1109/CVPR. 2015.7298706]

Zhang H，Fromont E，Lefevre S and Avignon B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks//2020 IEEE International conference on image processing (ICIP). IEEE，2020：276-280.[DOI：10.1109/ICIP40778.2020.9191080]

Jia X Y，Zhu C，Li M Z，Tang W Q and Zhou W L. LLVIP：A visible-infrared paired dataset for low-light vision//Proceedings of the IEEE/CVF international conference on computer vision. 2021：3496-3504.[DOI：10.1109/ICCVW54120.2021.00389]

Zhang Y，Xu C，Yang W，He G J，Yu H，Yu L and Xia G S. Drone-based RGBT tiny person detection. ISPRS Journal of Photogrammetry and Remote Sensing，2023，204：61-76.[DOI：10.1016/j.isprsjprs.2023.08.016]

Liu K and Mattyus G. Fast multiclass vehicle detection on aerial images. IEEE Geoscience and Remote Sensing Letters，2015，12(9)：1938-1942.[DOI：10.1109/LGRS.2015.2439517]

Razakarivony S and Jurie F. Vehicle detection in aerial imagery：A small target detection benchmark. Journal of Visual Communication and Image Representation，2016，34：187-203.[DOI：10.1016/j.jvcir.2015.11.002]

Mundhenk T N，Konjevod G，Sakla W A and Boakye K. A large contextual dataset for classification，detection and counting of cars with deep learning//European conference on computer vision. Cham：Springer International Publishing，2016：785-800.[DOI：10.1007/978-3-319-46487-9_48]

Hsieh M R，Lin Y L and Hsu W H. Drone-based object counting by spatially regularized regional proposal network//Proceedings of the IEEE international conference on computer vision. 2017：4145-4153.[DOI:10.1109/ICCV.2017.446]

Du D W，Qi Y K，Yu H Y，Yang Y F，Duan K W，Li G R，Zhang W G，Huang Q M and Tian Q. The unmanned aerial vehicle benchmark：Object detection and tracking//Proceedings of the European conference on computer vision (ECCV). 2018：370-386.[DOI：10.1007/978-3-030-01249-6_23]

Zhu P F，Wen L Y，Du D W，Bian X，Fan H and Hu Q H. Detection and tracking meet drones challenge. IEEE transactions on pattern analysis and machine intelligence，2021，44(11)：7380-7399.[DOI：10.1109/TPAMI.2021.3119563]

Xia G S，Bai X，Ding J，Zhu Z，Belongie S，Luo J B，Datcu M，Pelillo M and Zhang L P. DOTA：A large-scale dataset for object detection in aerial images//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018：3974-3983.[DOI：10.1109/CVPR.2018.00418]

Sun Y M，Cao B，Zhu P F and Hu Q H. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video Technology，2022，32(10)：6700-6713.[DOI：10.1109/TCSVT. 2022.3168279]

Song K C，Xue X T，Wen H W，Ji Y Y，Yan Y H and Meng Q G. Misaligned visible-thermal object detection：A drone-based benchmark and baseline. IEEE Transactions on Intelligent Vehicles，2024.[DOI：10.1109/TIV.2024.3398429]

Fang Q Y and Wang Z K. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition，2022，130：108786.[DOI：10.1016/j. patcog.2022.108786]

Cao Y，Bin J C，Hamari J，Blasch E and Liu Z. Multimodal object detection by channel switching and spatial attention// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：403-411.[DOI：10.1109/ CVPRW59228.2023.00046]

He X，Tang C，Zou X and Zhang W. Multispectral object detection via cross-modal conflict-aware learning//Proceedings of the 31st ACM International Conference on Multimedia. 2023：1465-1474.[DOI：10.1145/3581783.3612651]

Shen J F，Chen Y F，Liu Y，Zuo X，Fan H and Yang W K. ICAFusion：

Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition, 2024, 145: 109913. [DOI: 10.1016/j.patcog.2023.109913]

Yuan M X and Wei X X. C²former: Calibrated and complementary transformer for rgb-infrared object detection. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12. [DOI: 10.1109/TGRS.2024.3376819]

Zeng Y Q, Liang T F, Jin Y and Li Y D. MMI-Det: Exploring multimodal integration for visible and infrared object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 11198-11213. [DOI: 10.1109/TCSVT.2024.3418965]

Zhang J Q, Cao M X, Xie W Y, Lei J, Li D X, Huang W B, Li Y S and Yang X. E2e-mfd: Towards end-to-end synchronous multimodal fusion detection. Advances in Neural Information Processing Systems, 2024, 37: 52296-52322. [DOI: 10.52202/079017-1658]

Li Y X, Li X, Li Y H, Zhang Y C, Dai Y M, Hou Q B, Cheng M M and Yang J. Sm3det: A unified model for multi-modal remote sensing object detection//Proceedings of the AAAI Conference on Artificial Intelligence. 2026, 40(8): 6717-6725. [DOI: 10.1609/aaai.v40i8.37603]

作者简介

刘煜:女,博士研究生,主要研究方向为计算机视觉与遥感图像智能解译。E-mail: liuyu236@mails.ucas.ac.cn

冯瑛超:男,助理研究员,主要研究方向为计算机视觉与遥感图像智能解译。E-mail: fengyc@aircas.ac.cn

张伊丹:女,助理研究员,主要研究方向为计算机视觉与遥感图像智能解译。E-mail: zhangyidan19@mails.ucas.ac.cn

李宁:男,助理研究员,主要研究方向为计算机视觉与地理应用。E-mail: lining2434@gmail.com

刁文辉:男,副研究员,主要研究方向为遥感图像智能解译。E-mail: diaowh@aircas.ac.cn

胡岩峰:男,研究员,主要研究方向为自然语言处理与遥感图像理解。E-mail: huyanfeng@iphy.ac.cn