

DOI: 10.19650/j.cnki.cjsi.J2514453

基于模态引导与自适应对比学习的 无人机目标检测方法^{*}

孙明晓, 杨子祯, 李传龙, 栾添添, 梁洪杰
(哈尔滨理工大学自动化学院 哈尔滨 150080)

摘 要:针对现有多模态目标检测方法在模态选择、空间建模及跨模态一致性约束上的不足,尤其是在低光照条件、目标完全或部分遮挡、复杂背景干扰等具有挑战性的实际应用场景中出现检测精度显著下降的问题,创新性地提出了一种基于模态引导与自适应对比学习的无人机目标检测方法。首先,设计了模态引导选择模块,该模块通过构建全局语义感知的通道注意力机制动态评估不同模态的贡献,实现动态权重分配与特征融合,有效解决了传统固定权重融合策略在环境变化时出现的模态贡献度失衡问题。其次,提出了模态增强模块,该模块基于单分支的空间自注意力增强机制,在标准多头自注意力结构中引入局部-全局协同的相对位置偏置并结合归一化残差结构,增强对融合特征的空间结构感知能力,从而提升复杂背景和遮挡场景下的目标判别性能。最后,提出了检测感知自适应跨模态对比学习策略,该策略以检测框为监督单元,结合检测感知权重与模态自适应温度调节,对跨模态特征进行显式约束,强化语义一致性并提升模型鲁棒性。实验结果表明,该方法在 Drone Vehicle 与 LLVIP 数据集上实现目标检测精度 mAP_{50} 分别为 78.6% 和 98.3%, 相比现有方法均有明显提升。在搭建的无人机平台上实现 12.61 fps 的实时推理性能,验证了方法的有效性和实用价值。

关键词: 多模态目标检测; 模态引导选择; 自适应对比学习; 红外图像; 无人机

中图分类号: TH74 **文献标识码:** A **国家标准学科分类代码:** 460.4020

UAV object detection method based on modality-guided selection and adaptive contrastive learning

Sun Mingxiao, Yang Zizhen, Li Chuanlong, Luan Tiantian, Liang Hongjie
(School of Automation, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: Existing multimodal object detection methods often suffer from limitations in modality selection, spatial modeling, and cross-modal consistency, particularly under challenging conditions such as low illumination, target occlusion, and complex backgrounds. To address these issues, this paper proposes a UAV object detection method based on modality-guided selection and adaptive contrastive learning. First, a modality-guided selection module is designed, which employs global semantic-aware channel attention to dynamically evaluate modal contributions and enable adaptive feature fusion, thereby effectively resolving the modality imbalance issue inherent in conventional fixed-weight fusion strategies. Second, a Modality Enhancement Module is introduced, incorporating locally-globally coordinated relative positional biases and normalized residual connections into a single-branch self-attention structure to enhance spatial perception in complex and occluded scenes. Finally, a detection-aware adaptive cross-modal contrastive learning strategy is proposed, utilizing detection boxes with modality-adaptive temperature scaling to explicitly align multimodal features and improve semantic consistency. Experimental results demonstrate that the proposed method achieves mAP_{50} scores of 78.6% on the Drone Vehicle dataset and 98.3% on the LLVIP dataset, outperforming existing approaches. Deployment on a real UAV platform achieves 12.61 fps, validating both the accuracy and practical utility of the framework.

Keywords: multimodal object detection; modality-guided selection; adaptive contrastive learning; infrared; UAV

0 引言

随着无人机(unmanned aerial vehicle, UAV)平台在智能感知、安防监控等领域的广泛应用,基于无人机的目标检测技术逐渐成为智能视觉系统研究的热点方向^[1-3]。无人机凭借其视野灵活、部署便捷和覆盖范围广的优势,能够在复杂环境下实现地面目标的高效感知^[4]。然而,与固定摄像头或车载平台相比,无人机在飞行过程中面临光照变化剧烈、背景复杂以及目标尺度差异显著等问题,导致目标检测任务的准确性与鲁棒性受到显著影响^[5-6]。尤其在夜间、逆光、雾霾及遮挡等复杂条件下,可见光(red green blue, RGB)图像质量急剧下降,传统单模态检测方法易出现感知能力退化与检测性能不稳定等问题^[7]。

为应对上述挑战,多模态感知与深度学习技术的结合为无人机目标检测提供了新的解决思路。红外(infrared, IR)图像能够通过热辐射信息在低光照或夜间环境下稳定成像,与RGB图像在光谱特性上具有天然互补性^[8]。两种模态的融合有望在复杂环境下提升无人机视觉系统的目标检测精度与鲁棒性。因此,融合可见光与红外信息的多模态目标检测方法正逐渐成为无人机智能感知领域的重要研究方向^[9-10]。

近年来,研究者围绕UAV在复杂光照环境下的图像退化与感知不稳定问题,开展了大量RGB-IR多模态融合检测的研究。典型的多模态检测方法主要采用双分支结构,分别提取RGB与IR图像的特征,然后在中间层或决策层进行融合。Fang等^[11]提出一种多模态融合检测方法,通过引入自注意力机制,实现了模态内外特征的有效融合,提升了检测性能。Chen等^[12]提出双重增强模型,通过语义和空间权重分配模块聚合跨模态信息,增强了特征表征能力。郝博等^[13]提出一种基于红外与可见光图像的目标检测算法,采用双分支结构并在主干网络中加入注意力机制,提升了对小目标的检测能力。Cheng等^[14]提出一种超轻量级双模态网络,通过双模态融合模块自适应融合可见光与红外特征,显著提升无人机在复杂光照及天气条件下的检测鲁棒性。常凯旋等^[15]提出一种双模态融合无人机目标检测算法,通过双分支特征融合模块融合红外与可见光图像特征,并在颈部网络层引入注意力机制,有效提升了小目标检测精度并降低了误检率。孙晓永等^[16]提出一种面向多无人机协同的多模态目标检测方法,通过任务驱动的卷积融合网络实现双光图像融合,消除了不良光照条件对图像质量的影响。郭润泽等^[17]通过耦合光照与对比度的多尺度差分注意力机制,显著提升了弱光环境下的检测精度。然而,现有融合方法仍存在明显局限。首先,多数方法在

特征融合时未能充分考虑模态重要性,通常采用固定权重或简单拼接策略,难以适应不同场景下模态质量变化。其次,在融合特征处理阶段,传统卷积结构难以有效捕捉远距离依赖与高层次语义关联,限制了模型对复杂场景的解析能力。此外,部分方法还存在结构复杂、参数量大的问题。

针对现有多模态目标检测在模态选择、空间建模及跨模态一致性上的不足,提出一种基于模态引导与自适应对比学习的无人机目标检测方法。该方法基于YOLOv8框架,首先设计模态引导选择模块(modality-guided selection module, MGSM),动态评估不同模态的贡献,实现模态层面的自适应加权融合;随后,提出模态增强模块(modality-enhanced transformer, MET),利用局部-全局协同相对位置偏置与归一化残差机制,增强对融合特征的空间结构感知能力;最后,引入检测感知自适应跨模态对比学习策略(detection-aware adaptive contrastive learning, DA-CL),通过检测框级语义监督、检测感知权重及模态自适应温度调节,增强跨模态语义特征的一致性。

1 多模态融合目标检测网络结构设计

1.1 网络总体结构

多模态目标检测系统网络总体结构如图1所示,首先,采用双分支主干网络分别提取RGB和红外图像的初始特征,得到两个分支的中间特征图。

其次,在获取两个模态的初始特征后,MGSM模块利用全局语义感知的通道注意力动态加权不同模态,实现自适应模态融合;为增强对融合特征的空间结构感知能力,模态增强模块MET基于改进多头自注意力机制对融合特征进行空间建模与语义结构增强,结合局部-全局协同的相对位置偏置及归一化残差结构,有效提升了复杂背景和遮挡目标下的识别性能。为缓解融合后可能存在的模态特征漂移问题,训练阶段引入检测感知自适应跨模态对比学习策略,以检测框为监督单元,通过检测感知权重和模态自适应温度调节,对RGB与红外同一区域特征进行显式约束,强化跨模态语义一致性,提升模型泛化与鲁棒性。

最后,增强后的融合特征被送入检测头进行多尺度预测,输出目标的类别、边界框及置信度信息。该结构实现了模态融合、模态增强、模态一致性的层层递进建模思路,在保证检测精度的同时兼顾了实时性。

1.2 模态引导选择模块

模态引导选择模块旨在解决多模态融合中模态贡献度失衡问题。在弱光条件下,RGB图像易出现质量

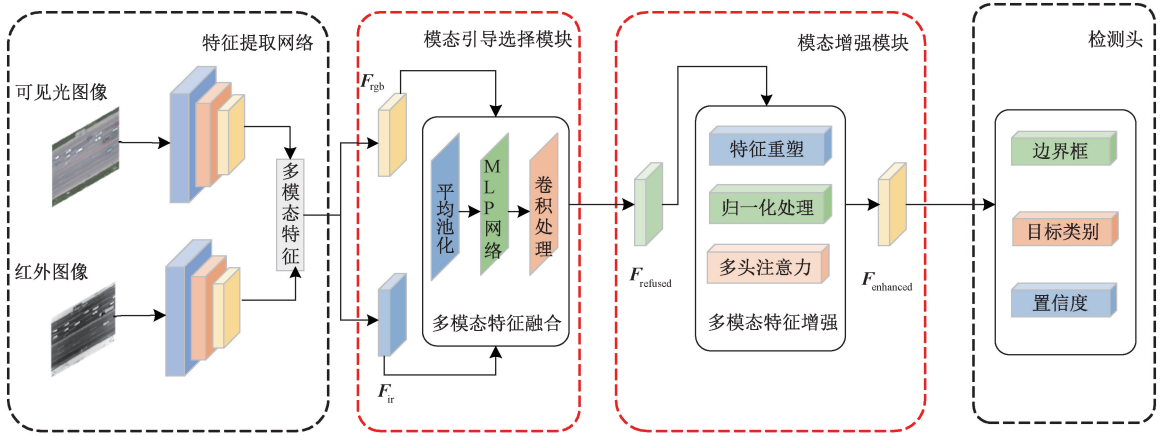


图1 多模态目标检测整体网络结构

Fig. 1 The overall architecture of multimodal detection

退化,而红外模态能提供稳定的热辐射信息;在正常光照下,RGB模态则包含更丰富的纹理与颜色特征。现有通道注意力机制多基于单模态统计信息,缺乏跨模态语义感知,易引入冗余特征导致模态失准。为此,提出MGSM模块,通过全局语义感知的通道注意力机制,实现跨模态权重的自适应调控。模态引导选择模块网络结构如图2所示。

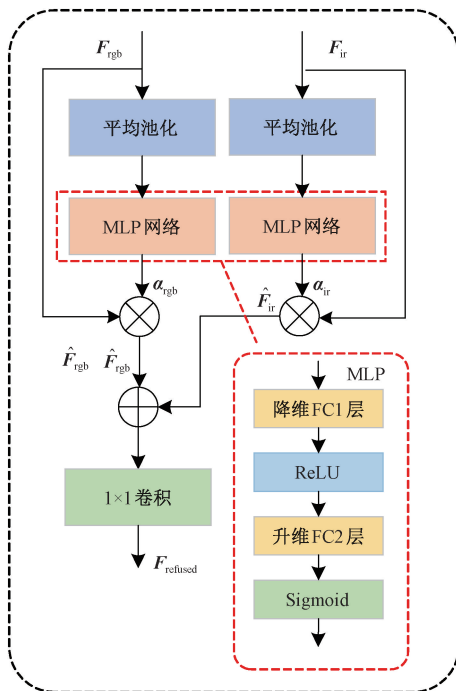


图2 模态引导选择模块网络结构

Fig. 2 Modal guidance selection module

对于两种模态的特征图 $F_{rgb}, F_{ir} \in \mathbf{R}^{B \times C \times H \times W}$, B 是批次大小, C 是通道数, H 和 W 是特征图的大小, 首先分别进

行全局平均池化, 获取每个通道的全局描述向量 $z_{rgb}, z_{ir} \in \mathbf{R}^{B \times C}$, 这个过程如式(1)和(2)所示。

$$z_{rgb} = \text{GAP}(F_{rgb}) \quad (1)$$

$$z_{ir} = \text{GAP}(F_{ir}) \quad (2)$$

随后, 分别将该向量送入由两层全连接层 (fully connected, FC) 和激活函数组成的 MLP 网络中, 得到自适应通道权重向量 $\alpha_{rgb}, \alpha_{ir} \in [0, 1]^{B \times C}$ 。该 MLP 网络首先通过一个降维 FC1 层和 ReLU 激活函数提取中间特征, 再通过升维 FC2 层和 Sigmoid 激活函数生成通道权重向量, 该过程如式(3)和(4)所示。

$$\alpha_{rgb} = \sigma(W_2 \times \text{ReLU}(W_1 \times z_{rgb})) \quad (3)$$

$$\alpha_{ir} = \sigma(W_2 \times \text{ReLU}(W_1 \times z_{ir})) \quad (4)$$

式中: σ 是 Sigmoid 激活函数; W_1 和 W_2 是两个全连接层的权重矩阵。与固定加权不同, 这里的通道权重 α_{rgb} 和 α_{ir} 完全由输入特征动态生成, 因此能够根据场景条件自适应地调节模态贡献度。接下来通道权重 $\alpha_{rgb}, \alpha_{ir}$ 分别与原始模态特征相乘, 得到加权后的特征图 $\hat{F}_{rgb}$ 和 $\hat{F}_{ir}$, 如式(5)和(6)所示。

$$\hat{F}_{rgb} = \alpha_{rgb} \times F_{rgb} \quad (5)$$

$$\hat{F}_{ir} = \alpha_{ir} \times F_{ir} \quad (6)$$

最后两个加权特征图在通道维度进行拼接, 并通过一个 1×1 的卷积层进行特征压缩和融合, 最终得到融合特征图 $F_{fused} \in \mathbf{R}^{B \times C \times H \times W}$ 如式(7)所示。

$$F_{fused} = \text{Conv}_{1 \times 1}([\hat{F}_{rgb}; \hat{F}_{ir}]) \quad (7)$$

通过上述设计, MGSM 模块在融合前即可实现基于全局语义的模态选择, 有效抑制了冗余模态干扰, 并保证在光照退化、目标遮挡和背景复杂的情况下, 网络始终能够选择最优模态特征参与融合。MGSM 的核心优势在于其跨模态的动态权重分配策略, 使融合特征在不同场景中具有更强的判别性和鲁棒性, 为后续的特征增强提供更加可靠的输入。

1.3 模态增强模块

MGSM 模块实现了 RGB 与 IR 特征的动态加权融合,但其通道层面的线性组合难以有效建模空间维度中的目标关系与上下文信息。为此,提出模态增强模块。该模块摒弃传统 Transformer 的直接套用,采用单分支空间自注意力增强机制,在多头自注意力中引入局部-全局协同相对位置偏置,并融合归一化残差结构,在保持长程依赖建模优势的同时,显著提升了复杂场景下的空间结构感知与目标判别能力。

MET 模块的结构如图 3 所示,该模块的输入为 MGSM 输出的融合特征图 $F_{\text{fused}} \in \mathbf{R}^{B \times C \times H \times W}$ 。

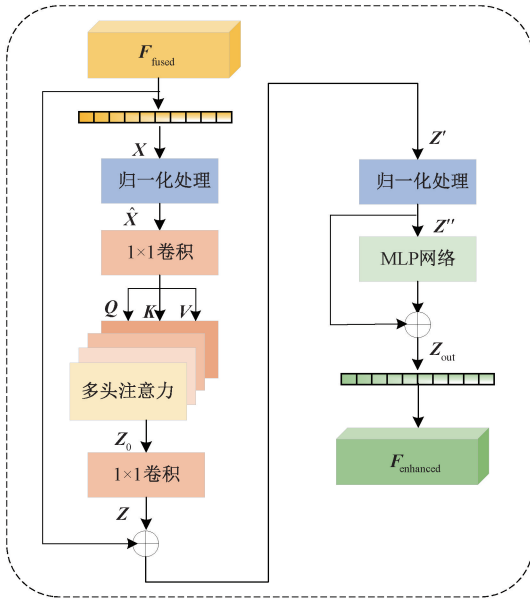


图3 模态增强模块网络结构

Fig. 3 Modal enhancement module structure diagram

首先将输入特征在空间维度上展平为序列形式如式(8)所示。

$$X = \text{Flatten}(F_{\text{fused}}) \in \mathbf{R}^{B \times C \times H \times W} \quad (8)$$

随后,对输入序列进行层归一化处理,并通过 1×1 卷积操作生成自注意力机制所需的查询 Q 、键 K 和值 V 张量,该过程如式(9)和(10)所示。

$$\hat{X} = \text{LayerNorm}(X) \quad (9)$$

$$Q, K, V = \text{Conv}_{1 \times 1}(\hat{X}) \in \mathbf{R}^{B \times C \times N} \quad (10)$$

将 Q, K, V 张量划分为 h 个注意力头,每个子空间的维度为 $d = C/h$ 。第 i 个头的注意力计算公式如式(11)所示。

$$\text{head}_i = \text{SoftMax}\left(\frac{Q_i K_i^T + A}{\sqrt{d}}\right) V_i \quad (11)$$

其中, $Q_i, K_i, V_i \in \mathbf{R}^{B \times d \times N}$, $A \in \mathbf{R}^{N \times N}$ 表示相对位置偏置矩阵,每个元素 $b(p, q)$ 对应序列中位置 p 与 q 的相对

位移。若 p 与 q 相邻,则 $b(p, q)$ 值较大,模型强化局部依赖;若距离较远,依赖于全局注意力建模。所有注意力头的输出在通道维度拼接,并通过卷积整合为最终的多头注意力,如式(12)和(13)所示。

$$Z_0 = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \in \mathbf{R}^{B \times C \times N} \quad (12)$$

$$Z = \text{Conv}_{1 \times 1}(Z_0) \quad (13)$$

为了增强语义表达并提升训练稳定性, MET 模块进一步引入了残差连接与前馈网络。首先将注意力输出与原始输入进行残差连接,如式(14)所示。

$$Z' = Z + X \quad (14)$$

然后对残差结果进行再次归一化处理,并输入至包含两层全连接层的 MLP 网络,如式(15)和(16)所示。

$$Z'' = \text{LayerNorm}(Z') \quad (15)$$

$$Z_{\text{out}} = \text{MLP}(Z'') + Z'' \quad (16)$$

最终,将序列形式的输出特征恢复为空间结构形式,得到增强后的特征图如式(17)所示。

$$F_{\text{enhanced}} = \text{Reshape}(Z_{\text{out}}) \in \mathbf{R}^{B \times C \times H \times W} \quad (17)$$

1.4 检测感知自适应跨模态对比学习策略

尽管 MGSM 模块可实现模态自适应融合, MET 模块增强了空间结构感知,但二者主要在特征生成与建模层面工作,仍缺乏针对跨模态语义一致性的显式约束。DA-CL 则在训练阶段引入检测任务先验,以检测框为监督单元,在语义空间中直接优化 RGB 与 IR 特征对齐,从而进一步增强跨模态表征的鲁棒性。

具体而言,对于每个检测框,分别从 RGB 与 IR 特征图中提取同一区域的特征表示,构造正样本对 $(f_i^{\text{rgb}}, f_i^{\text{ir}})$,并从背景区域采样负样本 f_j 。基于余弦相似度定义的改进对比损失函数如式(18)所示。

$$L_{\text{DA-CL}} = -\frac{1}{B} \sum_{i=1}^B \omega_i \log \frac{\exp(\text{sim}(f_i^{\text{rgb}}, f_i^{\text{ir}})/\tau_i)}{\sum_{j=1}^N \exp(\text{sim}(f_i^{\text{rgb}}, f_j)/\tau_i)} \quad (18)$$

式中: B 表示批次大小; N 为候选样本集合的大小; $\text{sim}(\cdot)$ 为特征间的余弦相似度函数; τ_i 为自适应温度系数; ω_i 为检测框级权重。

为突出语义可靠的目标样本,引入检测感知权重 ω_i ,定义如式(19)所示。

$$\omega_i = \alpha \cdot s_i + (1 - \alpha) \cdot \text{IoU}_i \quad (19)$$

式中: s_i 表示第 i 个检测框的分类置信度; IoU_i 表示预测框与真实框的重叠率; $\alpha \in [0, 1]$ 为平衡系数,这样设计能够使得高置信度且定位精度较高的目标在对比学习过程中获得更大权重,从而引导模型聚焦于语义可靠的目标区域。此外,为避免固定温度带来的训练不稳定性,提出模态自适应温度调节如式(20)所示。

$$\tau_i = \tau_0 \cdot (1 - \lambda \cdot \text{sim}(f_i^{\text{rgb}}, f_i^{\text{ir}})) \quad (20)$$

式中: τ_0 为基础温度超参数; λ 为调节系数。当模态间相似度较低时, τ_i 值增大,使得分布更加平滑以保持梯度稳

定;当模态间相似度较高时, τ_i 值减小,使分布更加尖锐,从而强化对齐约束。在训练过程中,对比损失与检测损失 L_{det} 共同优化,整体目标函数如式(21) 所示。

$$L_{\text{total}} = L_{\text{det}} + \beta \cdot L_{\text{DA-CL}}$$

(21)

式中: β 为调节系数,用于控制对比学习损失对整体优化目标的影响。

2 实验与结果

2.1 数据集

为全面评估多模态目标检测算法在不同场景下的适应性与鲁棒性,选取 Drone Vehicle^[18] 和 LLVIP (low-light visible-infrared paired)^[19] 两个具有代表性的 RGB-IR 成对图像数据集进行验证。采用 Drone Vehicle 数据集对算法进行主体性能测试与模块有效性分析。该数据集包含约 28 000 对 RGB 与红外图像,覆盖城市道路、十字路口等多种真实环境,标注 5 类车辆目标,具有目标尺度变化大、遮挡频繁和背景复杂等特点。为验证算法在不同目标类型及光照条件下的泛化能力,进一步选用 LLVIP 数据集进行补充实验。该数据集包含 15 488 对 RGB 与红外图像,采集自夜间街道、校园等低照度场景,标注目标均为行人。其在模态表现与检测任务方面与 Drone Vehicle 差异显著,能够有效检验模型在跨类别、弱光等挑战条件下的适应性与鲁棒性。

2.2 评价指标与实验环境

为全面衡量提出的多模态目标检测算法在不同数据集和任务场景中的性能表现,实验采用平均精度 (average precision, AP) 和所有类别平均精度的均值 (mean average precision, mAP) 作为评价指标,以全面衡量检测算法在复杂场景下的综合性能。

在模型训练阶段使用的硬件平台配置如下:CPU 采用 Intel Core i7-14700HX,GPU 使用 NVIDIA RTX 4070;软件环境为 Ubuntu 20.04,基于 PyTorch 1.13.1 框架开展实验。在训练过程中输入图像尺寸均为 512×512,采用 SGD 优化器,初始学习率设置为 0.001,Batch Size 设为 8,训练轮数为 100 Epochs。

2.3 消融实验

为验证所提出的各个模块在多模态目标检测中的有效性,在 Drone Vehicle 数据集上进行了消融实验,实验过程中,以 YOLOv8n 作为基线模型,逐步引入 MGSM 模块、MET 模块以及 DA-CL 模块,并对不同模块组合下的检测性能进行对比评估。实验结果如表 1 所示。

由表 1 可见,单模态下的检测性能有限,说明在复杂场景下单一模态难以提供稳定语义信息。引入 MGSM 模块后,mAP₅₀ 与 mAP_{50~95} 均显著提升,验证了其动态

表 1 不同方法的消融实验结果
Table 1 Ablation experimental results of different methods (%)

模态	模型配置	mAP ₅₀	mAP _{50~95}
RGB	YOLOv8n	54.5	33.2
IR	YOLOv8n	58.1	37.5
RGB+IR	+MGSM	72.1	50.3
RGB+IR	+MGSM+MET	76.8	53.9
RGB+IR	+MGSM+MET+(DA-CL)	78.6	56.4

权重分配与特征融合的有效性。进一步加入 MET 模块带来更明显增益,尤其 mAP_{50~95} 提升显著,体现其在空间结构建模方面的优势。最后引入 DA-CL 模块,模型性能达到最优,mAP₅₀ 与 mAP_{50~95} 分别提升至 78.6% 和 56.4%,充分验证了检测感知自适应跨模态对比学习在增强模态一致性方面的作用。为了进一步验证所提方法在特征建模方面的有效性,对最佳模型配置下的中间特征图进行了可视化分析,结果如图 4 所示。

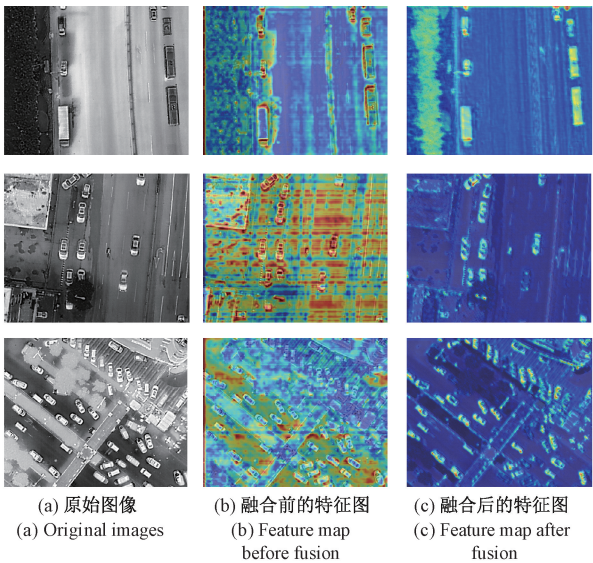


图 4 特征图可视化对比
Fig. 4 Comparison of feature map visualization

图 4(a) 为输入图像,图 4(b) 展示了未引入 3 个模块时的特征图响应,可见模型在低照度和复杂背景下响应不稳定,目标边界模糊且易受到背景干扰。相比之下,图 4(c) 引入 3 个模块后,模型能更集中地关注于目标区域,背景干扰减弱,目标语义信息更突出、边界更清晰。这表明 MGSM 模块增强了模态互补性,MET 模块提升了模型对融合特征的空间结构感知能力,而 DA-CL 模块进一步加强跨模态特征的一致性。

2.4 与现有方法比较

为验证所提方法在多模态目标检测任务中的有效性和鲁棒性,分别在 Drone Vehicle 和 LLVIP 两个数据集上进行了系统性对比实验,对比方法涵盖两类:一类为单模态检测方法,包括 Fast R-CNN 和 YOLOv8n,分别在 RGB 与 IR 图像上测试建立基线性能;另一类为多模态检测方法,包括近年来提出的 CFT、DEYOLO、CAMF^[20] 和 Infusion-Net^[21]。这些方法虽同样采用 RGB-IR 作为输入,但在跨模态融合机制上存在差异,具有较强的可比性。

表 2 为各算法在 Drone Vehicle 数据集上的检测结果。单模态方法性能普遍受限, Fast R-CNN 在 RGB 与 IR 模式下的 mAP_{50} 分别为 55.1% 与 59.5%, YOLOv8n 的表现为 54.5% 与 58.1%, 表明单一模态难以应对复杂场景。多模态方法性能显著提升, DEYOLO 算法与 Infusion-Net 算法的 mAP_{50} 分别达到 72.8% 与 71.3%, mAP_{50-95} 分别为 50.3% 与 49.4%。本文方法表现最佳, mAP_{50} 达到 78.6%, mAP_{50-95} 达到 56.4%, 相较于 DEYOLO 算法分别提升 5.8% 与 6.1%, 这一结果充分证明了所提方法的有效性与优势。

表 2 不同算法在 Drone Vehicle 数据集的对比实验结果
Table 2 Experiments on the Drone Vehicle dataset (%)

算法	模态	mAP_{50}	mAP_{50-95}
Fast R-CNN	RGB	55.1	34.9
Fast R-CNN	IR	59.5	39.1
YOLOv8n	RGB	54.5	33.2
YOLOv8n	IR	58.1	37.5
DEYOLO	RGB+IR	72.8	50.3
Infusion-Net	RGB+IR	71.3	49.4
CAMF	RGB+IR	69.2	47.1
CFT	RGB+IR	68.7	46.8
本文方法	RGB+IR	78.6	56.4

表 3 展示了 LLVIP 数据集上的实验结果,单模态方法已取得较高精度,如 Fast R-CNN 在 IR 模式下 mAP_{50} 达 95.2%。多模态方法在光照变化与遮挡场景中表现更稳健, DEYOLO 和 Infusion-Net 的 mAP_{50} 分别达到 97.5% 和 97.3%。本文方法取得最优结果, mAP_{50} 达 98.3%, mAP_{50-95} 达 58.4%, 进一步验证了其在低照度复杂场景下的精度与鲁棒性优势。

本文方法在 Drone Vehicle 与 LLVIP 两个具有不同场景的数据集上均取得了优于现有方法的性能,验证了其有效性。具体而言, MGSM 模块通过动态加权融合增强了模态互补性; MET 模块借助局部-全局位置偏置与

表 3 不同算法在 LLVIP 数据集的对比实验结果
Table 3 Experiments on LLVIP dataset (%)

算法	模态	mAP_{50}	mAP_{50-95}
Fast R-CNN	RGB	91.7	49.9
Fast R-CNN	IR	95.2	54.6
YOLOv8n	RGB	90.5	48.2
YOLOv8n	IR	94.1	53.5
DEYOLO	RGB+IR	97.5	57.6
Infusion-Net	RGB+IR	97.3	57.2
CAMF	RGB+IR	96.7	56.8
CFT	RGB+IR	96.4	56.3
本文方法	RGB+IR	98.3	58.4

归一化残差提升了空间结构感知能力; DA-CL 策略则通过特征对齐约束缓解了模态偏差。三模块协同工作,共同提升了复杂环境下的检测精度与鲁棒性。

2.5 检测结果可视化

为了直观比较算法性能,选取表现最优的 DEYOLO 算法和本文方法进行可视化对比,图 5 展示了 Drone Vehicle 数据集上的检测结果。

图 5(a) 是 RGB 图像,图 5(b) 是对应的红外图像,图 5(c) 是 DEYOLO 算法的检测结果,图 5(d) 是本文方法的检测结果。由检测结果可知, DEYOLO 算法在部分场景中表现不稳定,第 1 行出现目标被重复检测,第 2 行将建筑物误判为车辆,第 3 行部分遮挡目标存在误检和漏检,说明其在复杂背景和遮挡情况下判别能力有限。相比之下,本文方法检测结果更为准确可靠,尤其在遮挡场景下仍能完整识别目标,体现了 MGSM 模块的动态模态选择、MET 模块的空间结构增强以及 DA-CL 策略的特征一致性约束在复杂环境中的协同优势。

可视化实验表明,本文方法不仅在量化指标上优于现有方法,在实际检测中也展现出更优的目标完整性、鲁棒性和场景适应能力。

2.6 无人机平台算法部署与验证

为验证本文算法在真实环境中的性能,基于四旋翼无人机平台构建了完整的实时目标检测系统,如图 6 所示。

该系统搭载 Intel RealSense D435i 相机同步采集 RGB 与红外图像,通过 ROS 话题机制传输图像流。边缘计算单元 Jetson Orin NX 负责模型推理,检测结果经图传链路实时回传至地面站。飞行控制系统采用 Pixhawk 飞控保障飞行稳定性,并配备 GPS 模块提供空间定位信息,确保在真实场景中实现可靠的多模态目标检测。

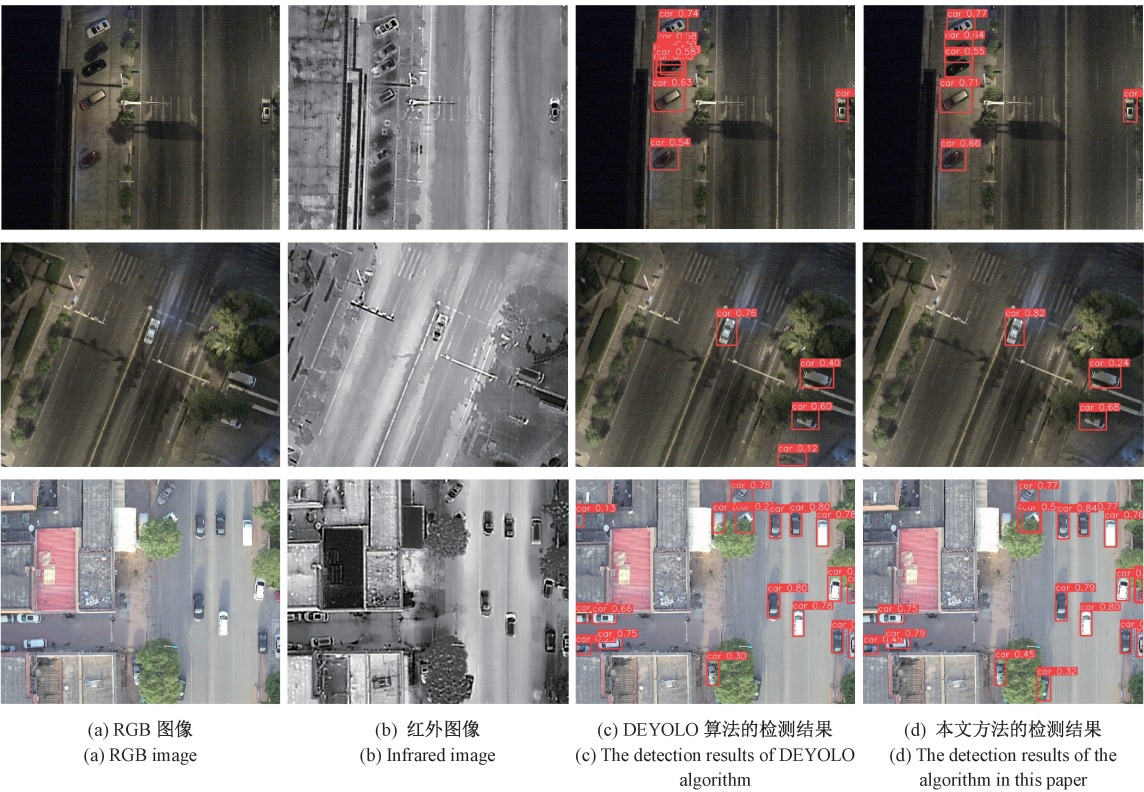


图 5 在 Drone Vehicle 数据集上检测结果
Fig. 5 Detection results on the Drone Vehicle dataset

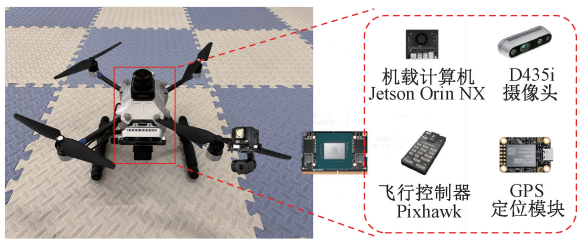


图 6 实验验证平台
Fig. 6 Experimental verification platform

图 7 展示了无人机在实际飞行的过程中采集的多模态目标检测结果,图 7(a)为输入的 RGB 图像,图 7(b)为对应的红外图像,图 7(c)为多模态目标检测结果。

由图 7 可知,在低光照、复杂背景以及无人机飞行过程中机体抖动导致图像模糊等典型无人机场景下,本文方法仍能准确识别目标,边界框定位精确且结果稳定,充分体现了所提算法在实际飞行环境中的检测精度与鲁棒性,能够满足无人机端多模态视觉感知任务的应用需求。

为评估实时推理性能,在相同无人机平台上对比了各算法的推理速度。实验结果如表 4 所示,本文方法推理速度略低于单模态 YOLOv8n 算法,这主要归因于多模

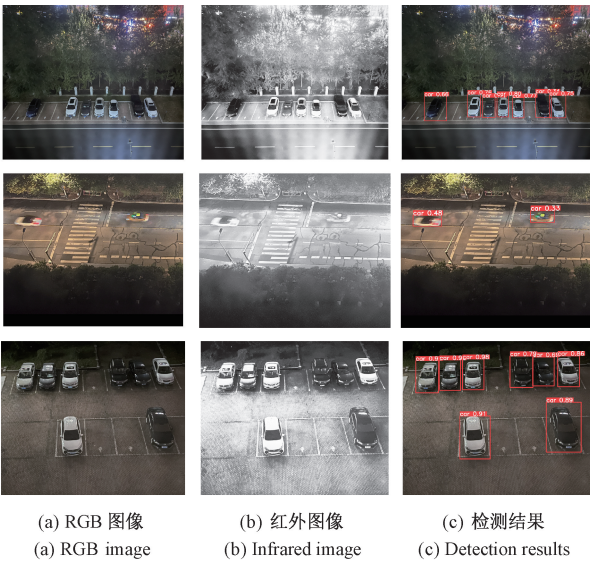


图 7 无人机多模态目标检测结果
Fig. 7 Results of the UAV multimodal target detection

态检测网络参数量通常显著大于单模态目标检测网络。然而和同类型的多模态目标检测算法 DEYOLO 比较,本文方法在推理帧率和延迟时间方面均表现出更优的性能。

表 4 不同算法推理性能对比

Table 4 Comparison of Inference of Different Algorithms

模型	模态	推理速度/fps	延迟时间/ms
YOLOv8n	单模态	22. 26	44. 92
DEYOLO	多模态	10. 33	96. 81
本文方法	多模态	12. 61	79. 30

3 结 论

针对现有多模态目标检测在模态选择、空间建模及跨模态一致性约束上的不足,本研究提出一种基于模态引导与自适应对比学习的无人机目标检测方法。主要贡献为:

1) 设计 MGSM 模块,通过全局语义感知的通道注意力机制动态评估模态可靠性,实现动态加权融合。

2) 提出 MET 模块,基于单分支空间自注意力增强机制,在标准多头自注意力中引入局部-全局位置偏置,并结合归一化残差结构,增强模型对融合特征的空间结构感知能力。

3) 构建 DA-CL 策略,以检测框为监督单元,引入检测感知权重与模态自适应温度调节,对跨模态特征进行显式约束,强化 RGB 与 IR 特征的语义一致性。

实验结果表明,在 Drone Vehicle 数据集上,本文方法在 mAP₅₀ 和 mAP₅₀₋₉₅ 分别达到 78. 6% 和 56. 4%,相比于 DEYOLO 算法分别提升 5. 8% 和 6. 1%,验证了所提方法的有效性和鲁棒性。未来研究可进一步探索模型轻量化、多模态时序信息融合以及跨域泛化能力,以推动多模态检测方法在更复杂动态场景中的应用。

参考文献

[1] 刘克顺,左晓桐,张玉华,等. 基于深度学习的无人机可见光目标检测研究综述[J]. 航空兵器, 2024, 31(6): 23-35.

LIU K SH, ZUO X T, ZHANG Y H, et al. A survey on object detection from UAV visible light images based on deep learning[J]. Aero Weaponry, 2024, 31(6): 23-35.

[2] 李琼,考月英,张莹,等. 面向无人机航拍图像的目标检测研究综述[J]. 图学学报, 2024, 45(6): 1145-1164.

LI Q, KAO Y Y, ZHANG Y, et al. A survey on object detection in UAV aerial images[J]. Journal of Graphics, 2024, 45(6): 1145-1164.

[3] 栾添添,吕奉坤,班喜程,等. 高动态环境下的傅里叶梅林变换视觉 SLAM 算法[J]. 仪器仪表学报, 2023, 44(7): 242-251.

LUAN T T, LYU F K, BAN X CH, et al. Fourier-Mellin transform visual SLAM algorithm in highly dynamic environments[J]. Chinese Journal of Scientific Instrument, 2023, 44(7): 242-251.

[4] 邬开俊,万智博,杜娟娟,等. 基于多尺度融合与通道压缩的无人机红外目标检测算法[J]. 红外与激光工程, 2025, 54(7): 243-255.

WU K J, WAN ZH B, DU J J, et al. UAV infrared object detection algorithm based on multi-scale fusion and channel compression[J]. Infrared and Laser Engineering, 2025, 54(7): 243-255.

[5] 崔阳洁,屠展,杨彬淇,等. 无人机遮挡目标检测与协同跟踪方法[J]. 机器人, 2025, 47(3): 427-437.

CUI Y J, TU ZH, YANG B Q, et al. UAV occluded object detection and cooperative tracking method[J]. Robot, 2025, 47(3): 427-437.

[6] 陈海永,刘博洋,晏行伟. 基于 YOLO-SCAT 的可见光图像无人机小目标检测算法[J]. 应用光学, 2025, 46(3): 505-514.

CHEN H Y, LIU B Y, YAN X W. YOLO-SCAT-based small object detection algorithm for UAV visible light images[J]. Journal of Applied Optics, 2025, 46(3): 505-514.

[7] 王迎龙,孙备,丁冰,等. BG-YOLO:复杂大视场下低慢小无人机目标检测方法[J]. 仪器仪表学报, 2025, 46(2): 255-266.

WANG Y L, SUN B, DING B, et al. BG-YOLO: A low-altitude, slow-moving, and small UAV object detection method in complex wide-field scenarios[J]. Chinese Journal of Scientific Instrument, 2025, 46(2): 255-266.

[8] 李阳,张杰,李强强. 基于改进 YOLOv5 的红外图像目标检测算法[J]. 红外技术, 2025, 47(8): 1009-1017.

LI Y, ZHANG J, LI Q Q. Infrared image object detection algorithm based on improved YOLOv5[J]. Infrared Technology, 2025, 47(8): 1009-1017.

[9] 王晶,黄智鑫,徐京邦,等. 面向地下环境机器人的多模态目标检测方法[J]. 机器人, 2025, 47(4): 537-547.

WANG J, HUANG ZH X, XU J B, et al. Multimodal object detection method for underground environment robots[J]. Robot, 2025, 47(4): 537-547.

[10] 孙明晓,王鑫源,栾添添,等. 动态环境下基于深度图的无人车激光 SLAM 算法[J]. 仪器仪表学报, 2025, 46(3): 101-109.

SUN M X, WANG X Y, LUAN T T, et al. Depth map-based LiDAR SLAM algorithm for unmanned vehicles in

- dynamic environments[J]. Chinese Journal of Scientific Instrument, 2025, 46(3): 101-109.
- [11] FANG Q Y, HAN D P, WANG ZH K. Cross-modality fusion transformer for multispectral object detection[J]. ArXiv preprint arXiv: 2111.00273, 2021.
- [12] CHEN Y SH, WANG B R, GUO X Y, et al. DEYOLO: Dual-feature-enhancement YOLO for cross-modality object detection[C]. Pattern Recognition, Part XVII, 2024: 236-252.
- [13] 郝博, 谷继明, 刘力维. 基于BF-YOLOv5的红外及可见光图像融合的目标检测[J]. 光电与控制, 2024, 31(5): 72-76.
- HAO B, GU J M, LIU L W. Object detection based on BF-YOLOv5 for infrared and visible image fusion[J]. Electronics Optics & Control, 2024, 31(5): 72-76.
- [14] CHENG X L, GENG K K, WANG Z W, et al. SLBAF-Net: Super-Lightweight bimodal adaptive fusion network for UAV detection in low recognition environment[J]. Multimedia Tools and Applications, 2023, 82(30): 47773-47792.
- [15] 常凯旋, 黄建华, 孙希延, 等. 基于双模态图像融合的无人机光学小目标检测算法[J]. 激光与光电子学进展, 2025, 62(4): 279-293.
- CHANG K X, HUANG J H, SUN X Y, et al. UAV small optical target detection algorithm based on dual-modal image fusion[J]. Laser & Optoelectronics Progress, 2025, 62(4): 279-293.
- [16] 孙晓永, 孙备, 郭润泽, 等. 面向多无人机协同的多模态目标检测方法[J]. 仪器仪表学报, 2025, 46(2): 209-220.
- SUN X Y, SUN B, GUO R Z, et al. Multimodal object detection method for multi-UAV collaboration[J]. Chinese Journal of Scientific Instrument, 2025, 46(2): 209-220.
- [17] 郭润泽, 孙备, 孙晓永, 等. 无人机弱光条件下多模态融合目标检测方法[J]. 仪器仪表学报, 2025, 46(1): 338-350.
- GUO R Z, SUN B, SUN X Y, et al. Multimodal fusion object detection method for UAVs under low-light conditions[J]. Chinese Journal of Scientific Instrument, 2025, 46(1): 338-350.
- [18] 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.
- [19] JIA X Y, ZHU CH, LI M ZH, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 3489-3497.
- [20] TANG L F, CHEN Z ANG, HUANG J, et al. CAMF: An interpretable infrared and visible image fusion network based on class activation mapping[J]. IEEE Transactions on Multimedia, 2023, 26: 4776-4791.
- [21] YUN J S, PARK S H, YOO S B. Infusion-net: Inter-and intra-weighted cross-fusion network for multispectral object detection[J]. Mathematics, 2022, 10(21): 3966.

作者简介



孙明晓, 2010年、2018年于哈尔滨工程大学分别获得工学学士学位、博士学位, 现为哈尔滨理工大学副教授, 主要研究方向为机器视觉、无人机目标检测。

E-mail: sunmingxiao@hrbust.edu.cn

Sun Mingxiao received his B. Eng. and Ph. D. degrees both from Harbin Engineering University in 2010 and 2018, respectively. He is currently an associate professor at Harbin University of Science and Technology. His main research interests include machine vision, and object detection in UAVs applications.



杨子祯, 2022年于扬州大学获得学士学位, 现于哈尔滨理工大学攻读硕士研究生, 主要研究方向为机器视觉、无人机目标检测。

E-mail: zizhenyang00@163.com

Yang Zizhen received his B.Sc. degree from Yangzhou University in 2022. He is currently pursuing his M.Sc. degree at Harbin University of Science and Technology. His main research interests include machine vision, and object detection in UAVs applications.



栾添添(通信作者), 2011年、2018年于哈尔滨工程大学分别获得工学学士学位、博士学位, 现为哈尔滨理工大学副教授, 主要研究方向为机器视觉、目标检测。

E-mail: luantiantian@hrbust.edu.cn

Luan Tiantian (Corresponding author) received her B. Eng. and Ph. D. degrees both from Harbin Engineering University in 2011 and 2018, respectively. She is currently an associate professor at Harbin University of Science and Technology. Her main research interests include machine vision, and object detection.