arXiv:2601.00598v1 [cs.CV] 2 Jan 2026

面向具身 RGB–红外感知的模态主导性感知优化

刘先辉[1]，蒋思琪[2],谢毅[3],林雨晴[1],刘晓[4,5]1 同济大学电子与信息工程学院[2] 同济大学计算机科学与技术学院[3] 亚利桑那大学电气与计算机工程系[4] 苏州大学未来科学与工程学院[5] 苏州大学省属高校通用人工智能与大模型重点实验室{xianhui l@163.com, 2432013@tongji.edu.cn, saliu@suda.edu.cn}

摘要—可见光–红外（RGB–IR）多模态感知是具身多媒体系统在复杂物理环境中运行的基础。尽管近期的跨模态融合方法推动了 RGB–红外检测的发展，但由模态非对称特性造成的优化动态仍未得到充分研究。在实际应用中，信息密度和特征质量的差异会引发持续的优化偏差，导致训练过度侧重于主导模态，并阻碍有效融合。为量化这一现象，我们提出模态主导指数（MDI），通过联合建模特征熵和梯度贡献来衡量模态主导性。基于 MDI，我们构建了模态主导性感知跨模态学习（MDACL）框架，用于调节跨模态优化。MDACL 采用分层跨模态引导（HCG）增强特征对齐，并通过对抗均衡正则化（AER）平衡融合过程中的优化动态。在三个可见光–红外基准数据集上进行的大量实验表明，MDACL 能有效缓解优化偏差，并取得最先进性能。

索引词—**多模态感知、可见光-红外、模态不平衡**

I. 引言

融合可见光（RGB）与红外（IR）输入的多模态感知，对于在复杂物理环境中运行的具身智能系统至关重要。对于自主机器人等真实世界智能体而言，必须在低照度和雾霾等恶劣条件下保持稳健的目标检测能力，因为此时RGB感知会显著退化。红外成像通过提供稳定的热信息来补充RGB，因此RGB–IR融合对于在动态环境中实现可靠感知尤为有效。因此，

可见光–红外检测已成为具身多媒体系统的关键组成部分。尽管其潜力巨大，但要训练出能够有效利用两种模态的统一检测器仍然颇具挑战，主要原因在于两种模态固有的光谱域差异和数据异质性显著。

针对跨模态融合，已有一系列研究被提出并取得了显著进展，这些研究大致可分为两类。一类着重于建模模态一致性和互补性 [1]–[4]。例如，CMRFusion [1]和 ICAFusion [2] 设计了专门的模块，用于解耦不同模态间的共享特征和特定特征。另一类旨在缓解 RGB

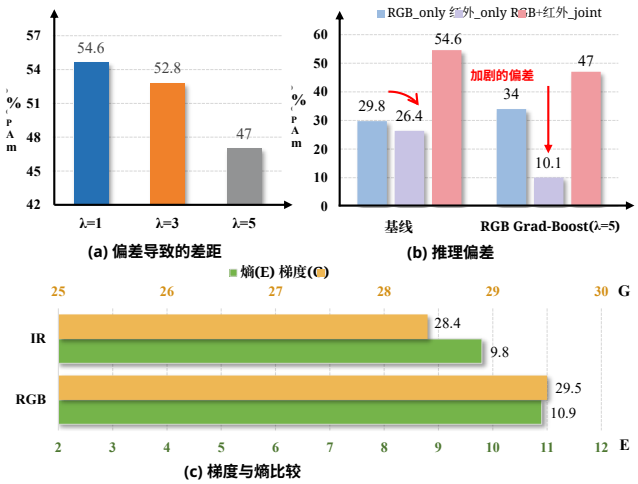

图 1. 可见光–红外检测中的优化偏差现象示意图。(a) M3FD 数据集在不同优化设置下的性能比较，包括 RGB+IR 联合训练基线（$\lambda = 1$）以及两种手动放大 RGB 梯度的变体（$\lambda = 3$ 和 $\lambda = 5$）。(b) 基线 RGB+IR 模型和 RGB Grad-Boost（$\lambda = 5$）变体在 M3FD 上使用仅 RGB、仅红外和 RGB+IR 输入时的推理性能。(c) 在 M3FD 上进行联合训练期间，RGB 和红外模态的平均梯度贡献以及目标区域特征熵。

与红外特征在空间或语义表征上的差异 [5]–[7]。OAFA [5] 和 DAMSDet [6]等代表性方法旨在通过隐式对齐 RGB 和红外模态来解决错位问题，从而提升检测性能。

尽管已有显著进展，大多数现有方法仍隐含地假设各模态贡献均衡，因此由非对称 RGB–红外模态引发的优化动态在很大程度上尚未得到探索。然而，信息密度和特征质量的差异往往会导致持续的优化偏差，使学习过程偏向某一模态。如图 1(a) 所示，在 M3FD 上进一步放大主导 RGB 模态的梯度会持续降低性能，且偏差越大，性能下降越明显。图 1(b) 进一步表明，在这种有偏的优化下，联合训练的检测器使用仅 RGB 输入时的表现明显优于使用仅红外输入时的表现，这说明训练过度依赖 RGB 模态。为进行定量分析，图 1(c) 显示，熵更高且梯度贡献更强的模态会获得

更高的优化偏好，这表明这些因素与优化偏差之间存在正相关关系。

受这些发现启发，我们首先设计了一种简单而有效的度量指标——**模态主**导指数（**MDI**），通过联合建模特征熵与梯度能量，明确衡量优化偏差。此外，我们提出了一种新颖的**模态**主导**感**知**跨**模态**学**习（**MDACL**）框架，以解决可见光-红外检测中的训练优化偏差问题，并提升其泛化性能。具体而言，MDACL 包含两个关键组件：分层跨模态引导（HCG）和对抗均衡正则化（AER）策略。为增强两种模态之间的一致性，HCG 采用双阶段交互设计，在低层特征上引导面向结构的对齐，并在高层特征上引导语义驱动的跨模态一致性，从而有效缓解特征级错位。此外，鉴于先前研究 [8], [9] 已表明，模态间的优化失衡会导致次优收敛行为，我们引入 AER 并设计了一种简单而高效的逆权重解，以调整优化动态，促进更加均衡的学习过程。为验证 MDACL 的有效性，我们在三个可见光-红外检测基准数据集上进行了广泛实验。总而言之，我们的贡献主要包括以下三个方面：

- 据我们所知，我们率先通过全面的定量分析，发现并缓解了可见光-红外检测中的优化偏差问题。
- 我们提出了一种新颖的**模**态**主**导**感**知**跨**模态**学**习（**MDACL**）框架，由两个核心组件构成：分层跨模态引导（HCG），旨在增强跨模态一致性并缓解错位；以及对抗均衡正则化（AER）策略，旨在主动调节优化动态，从而实现更稳定的跨模态学习过程。
- 大量实验表明，MDACL在三个基准数据集上均实现了最先进性能，相较于现有方法取得了显著提升。

II. 相关工作

可见光-红外检测的跨模态融合。作为可见光–红外检测中的一项关键技术，跨模态融合已成为重要的研究方向。现有方法主要应对两项挑战：（1）在保留模态特定特征的同时，利用模态互补性；（2）缓解跨模态失配。相应地，已有工作大致可分为两条路线。第一条路线侧重于解耦共享表征与模态特定表征。CMRFusion [1] 显式建模每种模态的共有分支与独有分支，而 CDDFuse [4] 通过相关性驱动损失约束特征分解。此外，基于注意力的机制也被广泛采用，以捕获全局跨模态交互和互补信息 [2], [10]。第二条路线旨在通过跨模态特征对齐来缓解空间或语义差异。OAFA [5]显式建模跨模态空间偏移以实现对齐，而 DAMS-Det [6]采用可变形交叉注意力，以适应复杂场景中的模态错位。尽管取得了显著进展，但大多数现有的 RGB–IR融合策略都忽视了训练过程中的非对称优化动态。

跨模态学习中的优化动态。在跨模态学习中，一个关键挑战是不同模态的收敛不同步，这可能导致性能欠佳 [8] 和训练过程不稳定。一系列研究 [9], [11], [12] 指出，性能更优的模态往往会主导梯度更新，并无意中抑制较弱模态的学习。为缓解此类梯度冲突，近期方法引入了一些正则化策略，例如帕累托整合 [13]和自适应梯度调制 [14]。然而，在可见光-红外目标检测领域，优化动态在解决模态不平衡问题中的作用仍未得到充分探索。

III. 问题表述

给定一对RGB-IR图像$\mathbf{I} = \{\mathbf{I}^{\text{rgb}}, \mathbf{I}^{\text{ir}}\}$，RGB-IR目标检测的目标是预测目标对象 $\mathcal{O} = \{(\mathbf{b}_i, c_i)\}_{i=1}^{N}$的位置，其中$\mathbf{b}_i \in \mathbb{R}^4$ 是一个边界框，$c_i \in \{1, \ldots, K\}$ 是其类别标签。我们旨在从数据集 $\mathcal{D} = \{(\mathbf{I}_j, \mathcal{O}_j)\}$中学习一个由 θ 参数化的模型 $\mathcal{F}_\theta$。形式上，标准目标可表述如下：

$$\theta^* = \arg\min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{I},\mathcal{O})\in\mathcal{D}} \mathcal{L}_{\text{det}} \left(\mathcal{F}_\Theta(\mathbf{I}^{\text{rgb}}, \mathbf{I}^{\text{ir}}), \mathcal{O}\right). \quad (1)$$

对于大多数RGB-IR检测框架，$\mathcal{F}_\theta$ 包含三个组成部分：模态特定编码器 $\mathcal{G}_*$、跨模态融合模块 $\mathcal{H}$ 和标准检测头 $\mathcal{D}et$：

$$\mathcal{F}_\theta(I^{RGB}, I^{IR}) = \mathcal{D}et(\mathcal{H}\phi(\mathcal{G}\psi(I^{RGB}), \mathcal{G}\chi(I^{IR}))), \quad (2)$$

其中，$\mathcal{G}\psi$ 和 $\mathcal{G}\chi$ 分别能够提取RGB特征和红外特征。为了充分利用多模态信息，现有方法 [2], [3], [15]大多侧重于融合模块 $\mathcal{H}_\phi$的设计，却忽视了内在的优化动态问题。这一局限促使我们显式地建模并平衡模态特定学习信号。

IV. 方法

A. 概述

本文从优化角度探讨可见光–红外检测，旨在缓解由模态非对称特性引起的训练偏差。我们不再假设模态贡献均衡，而是显式建模模态主导性，并据此调节跨模态优化。如图2所示，我们提出了一种统一的主导性感知框架，用于估计模态主导性、引导跨模态对齐，并在严重模态差异下稳定特征融合。下文将在第 IV-B 小节介绍模态主导指数（MDI），在第 IV-C 小节介绍分层跨模态引导，并在第 IV-D 小节介绍对抗均衡正则化。

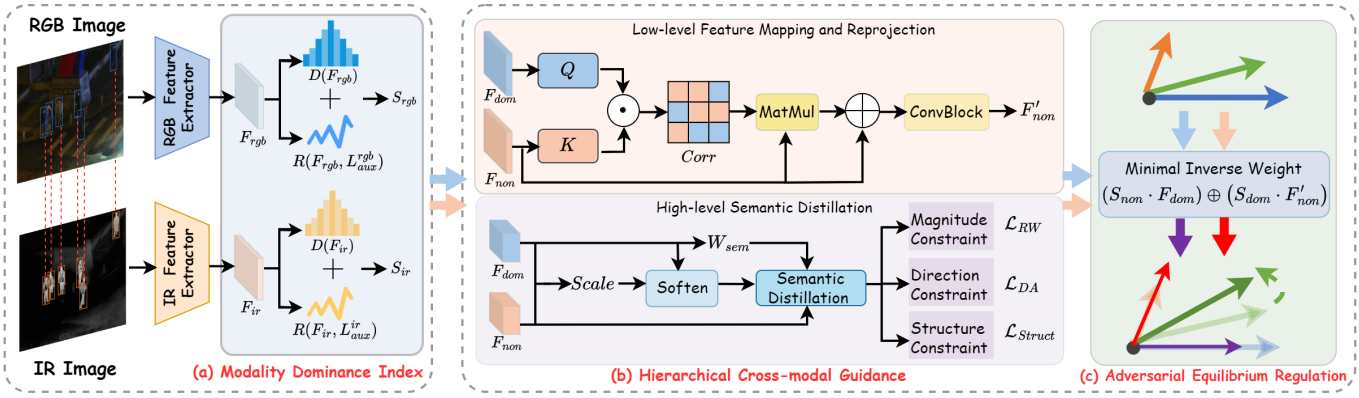

图2. 所提出的**MDACL** 框架概览。RGB和红外图像由双流骨干网络处理，随后通过（**a**）**模态主导指数**（**MDI**）估计模态主导性。主导性分数用于指导（**b**）**分层跨模态引导**（**HCG**）进行跨模态特征对齐，并通过（**c**）**对抗均衡正则化**（**AER**）实现平衡特征融合和稳定优化。

B. 模态主导性指数

在 RGB-红外检测任务中，模态之间固有的差异通常会导致不平衡的优化动态，最终阻碍有效的跨模态特征协作。基于这一认识，我们引入模态主导指数（MDI），以动态量化训练过程中各模态的贡献。令 F_i 表示模态 $i \in \{RGB, IR\}$ 的特征图。所提出的 MDI 捕捉模态质量的两个互补方面：**(1) 表征多样性**。为评估各模态固有的信息丰富度，我们定义多样性函数 $D(F_i)$，用于度量其特征激活的统计离散度。具有更丰富且分布更均匀的特征激活的模态将获得更高的多样性得分，这表明其具有更强的潜在贡献。

（2）任务响应敏感性。表征丰富性不足以反映模态的重要性。因此，我们定义任务响应函数 $R(F_i,\ L^i_{aux})$，用于评估检测任务对各模态的敏感程度。响应分数越高，表示轻微扰动会导致检测损失发生更大变化，这意味着该模态与任务的相关性更高。

模态主导指数 S 通过对多样性项和响应项进行归一化并线性组合得到：

$$S_i = \delta \cdot D(F_i) + (1-\delta) \cdot R(F_i, L^i_{aux}), \tag{3}$$

其中，δ 用于平衡表征多样性与任务响应敏感性。MDI 的计算过程见算法1。MDI值越高，表明该模态在当前训练情境中越占主导地位。

C. 分层跨模态引导

1) 低层特征映射与重投影：由于RGB与红外成像机制之间存在光谱差异，编码纹理模式和空间结构的低层特征往往会出现跨模态错位。为增强结构层面一致性，我们将非主导模态特征 F_{non} 动态投影至由主导模态 F_{dom}定义的结构空间中，其中模态主导性由MDI决定。

算法1模态主导指数（MDI）

要求：模态特征 $\{F_i\}_{i\in\{rgb,ir\}}$，**辅助检测器**$g(\cdot)$，**真实标注** M_{GT}，**平衡因子** δ**确保：模态主导性得分** S_{rgb}, S_{ir} **1: 对于** $i \in \{rgb, ir\}$ **执行 2:** $D(F_i) \leftarrow$ **熵(Softmax(展平**(F_i)**))) 3:** $L^i_{aux} \leftarrow \|g(F_i) - M_{GT}\|^2$**4:** $R(F_i, L^i_{aux})$ $\leftarrow \left\|\partial L^i_{aux}/\partial F_i\right\|_2$**5: 结束循环 6: 归一化** $\{D(F_i)\}$ **和** $\{R(F_i, L^i_{aux})\}$**7:** $S_i \leftarrow \delta \cdot D(F_i) + (1-\delta) \cdot R(F_i, L^i_{aux})$ **8: 返回** $\{S_{rgb}, S_{ir}\}$

具体而言，我们通过独立的卷积将 F_{non} 和 F_{dom} 转换为低维查询（Q）和键（K）表征。随后，通过对 Q 和 K进行缩放点积计算跨模态空间相关矩阵 $Corr \in \mathbb{R}^{HW\times HW}$，并进行Softmax归一化。在相关矩阵 $Corr$的引导下，我们可以将 F_{non} 重新投影到由 F_{dom} 定义的结构流形上，从而得到对齐后的表征 F_{reproj}:

$$F_{reproj} = \text{MatMul}(Corr, F_{non}). \tag{4}$$

为在对齐结构的同时保留模态特定特征，我们通过残差相加将重新投影后的特征 F_{reproj} 与原始非主导特征 F_{non} 进行融合，并使用轻量级卷积块对结果进行细化。

$$F'_{non} = \text{ConvBlock}(F_{non} + F_{reproj}). \tag{5}$$

2) 高层语义蒸馏：不同于以结构为重点的低层阶段，高层阶段旨在引导 F_{non} 获得 F_{dom}的语义丰富性。为此，我们设计了跨模态语义蒸馏损失 $\mathcal{L}_{Distill}$，其中 F_{dom} 作为教师 F_T，而 F_{non} 作为学生 F_S。所提出的 $\mathcal{L}_{Distill}$ 被构建为一种多目标复合损失函数，同时针对语义对齐和结构鲁棒性保持。

为缓解模态差异显著时可能产生的有害知识迁移，我们首先计算初始特

征方差 Δ，即 F_T 与 F_S 之间的方差，以动态生成缩放因子 Scale $\propto e^{-\Delta}$，从而减弱教师信号。

$$\Delta = \frac{1}{CHW} \parallel F_T - F_S \parallel_2^2, \tag{6}$$

其中， C， H和 W 分别表示特征的通道数、高度和宽度。

语义对齐监督。为确保 F_S 能够准确逼近语义空间中的 F_T，我们引入两个互补的损失项， $\mathcal{L}_{RW}$ 和 $\mathcal{L}_{DA}$，共同约束特征幅值和方向的一致性。

区域加权 L_2 损失 $\mathcal{L}_{RW}$ 通过最小化通道归一化表示之间的平方差来约束特征的幅值。我们进一步根据 F_T的激活强度生成加权图 W_{sem} ，对判别区域赋予更强的监督，从而实现任务感知语义对齐：

$$\mathcal{L}_{RW} = \|W_{sem} \odot (F_S - F_T)\|_2^2 . \tag{7}$$

W_{sem} 通过将教师特征$\|F_T\|_{c,\ 2}$ 的通道维度 L_2范数除以其空间均值 $\mathbb{E}(\cdot)$进行归一化得到：

$$W_{sem} = \frac{\|F_T\|_{c,2}}{\mathbb{E}\left(\|F_T\|_{c,2}\right)}. \tag{8}$$

为确保语义维度上的方向一致性，余弦相似度损失 $\mathcal{L}_{DA}$ 通过最小化 F_S 和 F_T的特征向量之间的夹角，确保学生学习教师的语义流形：

$$\mathcal{L}_{DA} = 1 - \frac{1}{N} \sum_i \frac{F_{Si} \cdot F_{Ti}}{||F_{Si}||_2 ||F_{Ti}||_2}, \tag{9}$$

其中， N 为空间位置的总数。

结构保持约束。为避免在语义蒸馏过程中对 F_S 过度平滑，我们引入基于梯度的结构保持项，使 F_S 与 F_T之间的空间变化率保持一致，从而隐式地维持特征图的结构一致性：

$$\mathcal{L}_{Struct} = |\text{Grad}(F_S) - \text{Grad}(F_T)| , \tag{10}$$

其中， $Grad(F)$表示特征 F的平均绝对空间梯度，通过有限差分近似计算：

$$Grad(F) = \mathbb{E}\left[|F_x - F_{x-1}|\right] + \mathbb{E}\left[|F_y - F_{y-1}|\right]. \tag{11}$$

语义蒸馏损失 $\mathcal{L}_{Distill}$ 是上述各项的加权总和：

$$\mathcal{L}_{Distill} = \alpha \mathcal{L}_{RW} + \beta \mathcal{L}_{DA} + \gamma \mathcal{L}_{Struct}, \tag{12}$$

其中， α， β， γ 为用于加权各项约束的超参数。

D. 对抗均衡调节

在可见光-红外融合中，传统方法往往会为主导模态分配更大的融合权重。然而，这种“优势放大”会扭曲优化动态，导致网络过度依赖单一模态，同时削弱另一模态的贡献。为缓解这种不平衡，我们从博弈论中汲取灵感，并提出对抗均衡调节（AER）策略。

从博弈论的角度来看，两种模态可以被视为合作—竞争智能体。过于主导的模态会使系统偏离最优联合解，而维持相互调节、互补的交互，则能使模型趋近帕累托最优状态。这一认识凸显了多模态融合的一项关键原则：我们应适当抑制主导模态，并增强较弱模态的作用，从而实现更加均衡、高效的合作均衡。

基于这一认识，我们设计了一种简单而有效的实现方案——最小逆权重（MIW）方案。在特征融合过程中，MIW 利用模态主导指数 S 作为调节信号，为非主导模态 F_{non}分配较高的融合权重，同时降低主导模态 F_{dom}的贡献：

$$F_{fused} = (S_{non} \cdot F_{dom}) \oplus (S_{dom} \cdot F'_{non}) . \tag{13}$$

最小逆权重形式使网络能够在反向传播过程中维持“对抗均衡”，且计算开销可以忽略不计。

尽管 MIW 是所提出的 AER 策略的一种直接实现，但它有效验证了核心思想。未来，我们将探索更先进的调节范式，以进一步增强可见光-红外多模态学习中的动态优化均衡。

V. 实验与分析

A. 实验设置

数据集与评估指标。我们在三个广泛使用的可见光-红外检测基准数据集上进行实验：LLVIP [16], M3FD [17], 和 FLIR [18]。为评估整体检测性能，我们采用两种常用指标：平均精度均值（mAP）和 mAP50。其中，mAP 表示在交并比阈值从 0.50 到 0.95 范围内计算的平均精度（AP）的均值，mAP50 则表示交并比阈值为 0.50 时的 AP。

实现细节。所有实验均在 NVIDIA RTX 4090 GPU 上进行。默认采用 SGD 优化器，动量为 0.937，权重衰减为 0.0005。初始学习率设为 0.01，并使用余弦退火调度器逐步衰减。数据预处理时，所有输入图像均调整为 640×640。为确保公平比较，所有结果均报告五次实验的结果。

表 I M3FD 上与其他方法的比较：最佳结果以粗体标出，次佳结果加下划线。

模型	数据类型	骨干网络	平均精度均值50 ↑	平均精度均值 ↑
TarDAL [17]	红外+RGB	CSPDarknet53	80.7	54.1
CDDFuse [4]	红外+RGB	CSPDarknet53	81.2	53.6
KCDNet [19]	红外+RGB	CSPDarknet53	83.2	56.3
DAMSDet [6]	红外+RGB	ResNet50	80.2	52.9
EMMA [7]	红外+RGB	CSPDarknet53	82.9	55.4
CRSIOD [20]	红外+RGB	CSPDarknet53	<u>84.0</u>	<u>57.2</u>
YOLOv8l-红外 [21]	IR	CSPDarknet53	79.5	53.1
YOLOv8l-RGB [21]	RGB	CSPDarknet53	80.9	52.5
Ours	红外+RGB	CSPDarknet53	**86.8**	**60.5**

表 II LLVIP 上与其他方法的比较：最佳结果以粗体标出，次佳结果加下划线。

模型	数据类型	骨干网络	平均精度均值50 ↑	平均精度均值 ↑
ICAFusion [2]	红外+RGB	CSPDarknet53	95.2	60.1
LUT-Fuse [22]	红外+RGB	CSPDarknet53	94.1	61.4
Fusion-Mamba [23]	红外+RGB	CSPDarknet53	97.0	64.3
UniRGB-IR [15]	红外+RGB	Transformer	96.1	63.2
CSAA [10]	红外+RGB	ResNet50	94.3	54.2
Text-IF [24]	红外+RGB	Transformer	94.1	60.2
FFM [3]	红外+RGB	CSPDarknet53	<u>97.6</u>	<u>64.8</u>
YOLOv8l-红外 [21]	IR	CSPDarknet53	94.6	61.7
YOLOv8l-RGB [21]	RGB	CSPDarknet53	91.8	53.6
Ours	红外+RGB	CSPDarknet53	**97.9**	**66.5**

表 III FLIR 上与其他方法的比较：最佳结果以粗体标示，次佳结果加下划线。

模型	数据类型	骨干网络	平均精度均值50 ↑	平均精度均值 ↑
ICAFusion [2]	红外+RGB	CSPDarknet53	79.2	41.4
CSAA [10]	红外+RGB	ResNet50	79.2	41.3
CrossFormer [25]	红外+RGB	CSPDarknet53	79.3	42.1
UniRGB-IR [15]	红外+RGB	Transformer	<u>81.4</u>	<u>44.1</u>
FFM [3]	红外+RGB	CSPDarknet53	<u>81.4</u>	42.3
YOLOv8l-IR [21]	IR	CSPDarknet53	72.9	38.3
YOLOv8l-RGB [21]	RGB	CSPDarknet53	66.3	28.2
Ours	红外+RGB	CSPDarknet53	**83.2**	**44.6**

B. 实验结果

M3FD 上的结果。M3FD 是一个广泛使用的可见光–红外基准数据集，涵盖多样的场景和恶劣天气条件。如表 I 所示，我们的方法始终优于所有竞争方法，取得了 86.8% 的 mAP50 和 60.5% 的平均精度均值，分别领先此前最先进方法 CRSIOD 2.8% 和 3.3%。这些结果凸显了在应对复杂环境变化时，明确缓解跨模态结构错位和优化失衡的重要性，也验证了我们的方法在 M3FD 的严苛条件下的有效性。**LLVIP 上的结果**。LLVIP 是一个在低照度条件下采集的大规模数据集，在这类条件下，模态不平衡十分常见。如表 II 所示，我们的方法在 LLVIP 上取得了新的最先进结果，mAP50 达到 97.9%，平均精度均值达到 66.5%，并始终优于 Fusion-Mamba 和 FFM 等近期强劲的竞争方法。这一持续提升得益于我们提出的模态主导性感知学习策略，该策略能够明确调节跨模态优化动态，并在严重模态差异下促进稳定、有效的特征融合。

FLIR上的结果。FLIR是一个真实世界的RGB–红外基准数据集，其特点是背景杂乱以及模态之间存在显著的外观差异。如表 III 所示，我们的方法在FLIR上取得了更高的检测精度，达到 83.2% 的 mAP50 和 44.6% 的平均精度均值，优于所有现有方法。与 ICAFusion 在特征交互时隐含假设模态贡献均衡不同，我们的方法明确识别并调节 RGB 与红外模态之间的非对称优化行为，从而实现更有效的特征对齐，提升检测性能。

C. 定性分析

样本可视化。我们将定性检测结果可视化，并与具有代表性的方法进行比较 [2], [6],

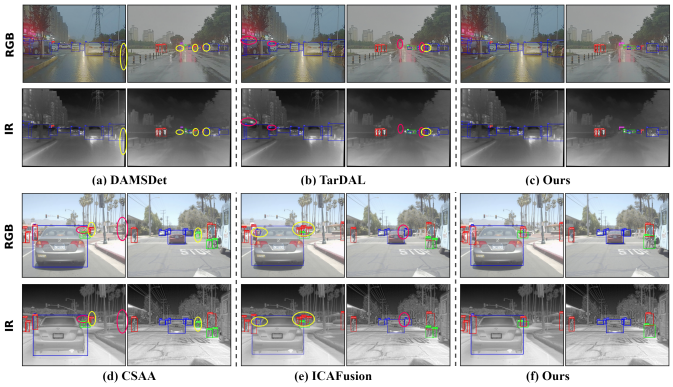

图 3. M3FD 和 FLIR 数据集上部分可见光-红外检测方法的可视化结果。(a)–(c) 展示 M3FD 数据集的结果，(d)–(f) 展示 FLIR 数据集的结果。黄色椭圆圈出的目标为假阳性，红色椭圆圈出的目标为漏检。

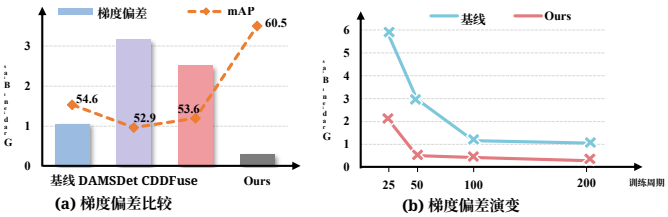

图 4. 梯度偏差比较及其对模型性能的影响。(a) 四种 RGB–IR 检测器的性能与平均梯度偏差之间的关系。(b) 不同训练轮次下平均梯度偏差的演变。

[10], [17]。如图 3 所示，在低照度和严重遮挡等具有挑战性的场景下，我们的方法始终能够减少漏检和假阳性，展现出稳健且准确的检测性能。

梯度比较。为评估我们的方法在缓解优化偏差方面的有效性，我们分析了训练期间的梯度偏差，并将其与具有代表性的基线方法进行比较。梯度偏差定义为 RGB 和 IR 分支的梯度贡献之间绝对差值的平均值：每 100 个训练步计算一次，并在整个训练过程中取平均。如图 4(a) 所示，检测性能与梯度偏差呈现出明显的反向关系。具体而言，我们的方法在取得最高性能的同时实现了最低的梯度偏差，表明其具有更加均衡的优化行为。图 4(b) 进一步展示了训练阶段中的梯度偏差演变。与基线相比，我们的方法在整个训练过程中始终保持较低的梯度偏差，体现了其在调节跨模态优化动态和促进有效 RGB–IR 融合方面的稳健性。

表 IV LLVIP HCG 上不同配置的消融研究

MDI	Low-Map	High-Distill	MIW	平均精度均值50 ↑	平均精度均值 ↑
				95.5	63.4
✓	✓			96。3 +0.8	64.6 +1.2
✓		✓		96。5 +1.0	64.8 +1.4
✓	✓	✓		97.0 +1.5	65.2 +1.8
✓			✓	96.4 +0.9	64.5 +1.1
✓	✓	✓	✓	**97.9** +2.4	**66.5** +3.1

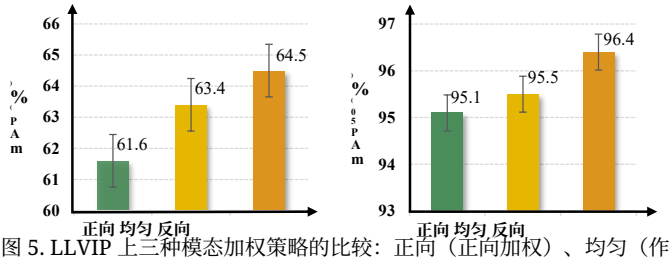

图 5. LLVIP 上三种模态加权策略的比较：正向（正向加权）、均匀（作为基线的均匀加权）和反向（本文提出的反向加权）。

D. 消融研究

核心组件。表 IV 总结了 LLVIP 上的消融结果。由于 MDI 用于量化模态主导性，因此设计为与 HCG 或 MIW 联合使用。两种 HCG 引导机制分别增强结构层面和语义层面对齐，单独使用时各自带来约 +1% 的 mAP50 提升，同时启用时则能进一步提升性能。与此同时，MIW 在优化过程中重新平衡模态贡献，带来额外且稳定的提升。整合 MDI、HCG 和 MIW 后，整体性能达到最佳，显著优于基线，凸显了所有组件的有效性及协同作用。**权重分配策略**。 图 5 展示了三种权重分配策略的比较。采用均匀加权的基线模型取得了 95.5% 的 mAP50 和 63.4% 的 mAP。采用正向分配策略进一步强调主导模态后，性能下降至 95.1% 的 mAP50 和 61.6% 的 mAP，这表明偏重主导模态可能会妨碍有效学习。相比之下，反向分配策略显著提升了性能，使 mAP50 达到 96.4%，mAP 达到 64.5%。这些结果证实，通过强化较弱模态来重新平衡优化，能够有效缓解训练导致的模态不平衡，并获得更稳健的检测性能。

VI. 结论

在本研究中，我们从以优化为核心的视角重新审视了可见光-红外检测，并着重阐明模态非对称特性如何影响多模态训练动态。实证结果表明，主导模态往往会吸引不成比例的优化关注，从而阻碍有效的跨模态融合。为量化这一现象，我们提出了模态主导性指数，这是一种简洁且易于解释的模态级优化不平衡度量指标。基于这一洞见，我们提出了MDACL框架，将分层跨模态引导与平衡感知正则化相结合，以共同对齐表征并平衡优化。在多个基准数据集上进行的大量实验，证明了我们方法的有效性和鲁棒性。我们认为，这项工作凸显了优化感知建模对于可见光-红外检测之外的具身多模态感知的重要性。

参考文献

[1] Chao Yang、Chao Tian、Guoqing Zhu、Qiang Wang 和 Zhenyu He，“Cmrfusion：通过跨模态掩码重建”，载于*2025*年*IEEE*国际多媒体与博览会会议 *(ICME)*，2025年，第1–6页。[2] Jifeng Shen、Yifei Chen、Yue Liu、Xin Zuo和Heng Fan，“Icafusion：用于多光谱目标检测的迭代式交叉注意力引导特征融合”，*Pattern Recognition*，第145卷，第109913页，2024年。[3] Zeyu Wang、Huiying Xu、Yun Liu、Chen Li、Xinzhong Zhu、Xiaolei Zhang和Hongbo Li，“从频域视角重新思考跨模态融合Mamba”，载于*2025*年*IEEE*国际多媒体与博览会会议 *(ICME)*。IEEE，2025年，第1–6页。[4] Zixiang Zhao、Haowen Bai、Jiangshe Zhang、Yulun Zhang、Shuang Xu、Radu Timofte和Luc Van Gool，“CDDFuse：面向多模态图像融合的相关性驱动双分支特征分解”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年。[5] Chen Chen、Jiahao Qi、Xingyue Liu、Kangcheng Bin、Ruigang Fu、Xikun Hu和Ping Zhong，“面向无人机的多模态目标检测的弱错位自适应特征对齐”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第26836–26845页。[6] Junjie Guo、Chenqiang Gao、Fangcen Liu、Deyu Meng和Xinbo Gao，“DAMSDet：采用竞争式查询选择和自适应特征融合的动态自适应多光谱检测变换器”，载于欧洲计算机视觉会议。Springer，2024年，第464–481页。[7] Zixiang Zhao、Haowen Bai、Jiangshe Zhang、Yulun Zhang、Kai Zhang、Dongdong Chen、Radu Timofte和Luc Van Gool，“等变多模态图像融合”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第25912–25921页。[8] Weiyao Wang、Du Tran和Matt Feiszli，“是什么让多模态分类网络的训练变得困难？”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第12695–12705页。[9] Yake Wei、Ruoxuan Feng、Zihe Wang和Di Hu，“通过样本级模态评估增强多模态协作”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第27338–27347页。[10] Yue Cao、Junchi Bin、Jozsef Hamari、Erik Blasch和Zheng Liu，“通过通道切换与空间注意力实现多模态目标检测”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第403–411页。[11] Xiaohui Zhang、Jaehong Yoon、Mohit Bansal和Huaxiu Yao，“通过交替进行单模态适应实现多模态表征学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第27456–27466页。[12] Ruize Xu、Ruoxuan Feng、Shi-Xiong Zhang和Di Hu，“Mmcosine：面向均衡音视频细粒度学习的多模态余弦损失”，载于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2023年，第1–5页。[13] Yake Wei和Di Hu，“Mmpareto：借助无害的单模态辅助提升多模态学习”，*arXiv*预印本 *arXiv:2405.17730*，2024年。[14] Hong Li、Xingyu Li、Pengbo Hu、Yinuo Lei、Chunxiao Li和Yi Zhou，“通过自适应梯度调制提升多模态模型性能”，载于*IEEE/CVF*国际计算机视觉会议论文集，2023年，第22214–22224页。[15] Maoxun Yuan、Bo Cui、Tianyi Zhao、Jiayi Wang、Shan Fu、Xue Yang和Xingxing Wei，“UniRGB-IR：通过适配器微调实现可见光-红外语义任务的统一框架”，载于第*33*届*ACM*国际多媒体会议论文集，2025年，第2409–2418页。[16] Xinyu Jia、Chuang Zhu、Minzhen Li、Wenqi Tang和Wenli Zhou，“LLVIP：面向低照度视觉的可见光-红外配对数据集”，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第3496–3504页。[17] Jinyuan Liu、Xin Fan、Zhanbo Huang、Guanyao Wu、Risheng Liu、Wei Zhong和Zhongxuan Luo，“面向目标检测的红外与可见光融合：目标感知双重对抗学习及多场景多模态基准”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第5802–5811页。[18] Heng Zhang、Elisa Fromont、S´ebastien Lefevre和Bruno Avignon，“采用循环融合与细化模块的目标检测多光谱融合”，载于*2020*年*IEEE*国际图像处理会议 *(ICIP)*。IEEE，2020年，第276–280页。[19] Haoyu Wang、Shiyuan Qu、Zhenzhuang Qiao和Xiaomin Liu，“KCDNet：模态信息不平衡场景中的多模态目标检测”，*IEEE*仪器与测量汇刊，2024年。[20] Huiying Wang、Chunping Wang、Qiang Fu、Dongdong Zhang、Renke Kou、Ying Yu和Jian Song，“基于图像特征的无人机航空图像跨模态定向目标检测”，*IEEE*地球科学与遥感汇刊，第62卷，第1–21页，2024年。

[21] Rejin Varghese和M Sambath，“Yolov8：一种性能和鲁棒性均得到增强的新型目标检测算法”，载于*2024*年数据工程与智能计算系统进展国际会议 *(ADICS)*。IEEE，2024年，第1–6页。[22] Xunpeng Yi、Yibing Zhang、Xinyu Xiang、Qinglong Yan、Han Xu和Jiayi Ma，“LUT-Fuse：通过蒸馏至可学习查找表实现超快速红外与可见光图像融合”，载于*IEEE/CVF*国际计算机视觉会议论文集，2025年，第14559–14568页。[23] Wenhao Dong、Haodong Zhu、Shaohui Lin、Xiaoyan Luo、Yunhang Shen、Guodong Guo和Baochang Zhang，“用于跨模态目标检测的Fusion-Mamba”，*IEEE*多媒体汇刊，2025年。[24] Xunpeng Yi、Han Xu、Hao Zhang、Linfeng Tang和Jiayi Ma，“Text-IF：利用语义文本引导实现退化感知交互式图像融合”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第27026–27035页。[25] Seungik Lee、Jaehyeong Park和Jinsun Park，“CrossFormer：用于多模态目标检测的交叉引导注意力”，《模式识别快报》，第179卷，第144–150页，2024年。