

文章编号 1004-924X(2026)19-3104-15

基于特征补偿与增强融合的非对齐 可见光-红外目标检测

周永根^{1,2}, 林坚普^{1,2}, 吕珊红^{1,2}, 林志贤^{1,2,3}, 林珊玲^{1,2*}, 郭太良^{2,3}

(1. 福州大学 先进制造学院, 福建 泉州 362251;

2. 中国福建光电信息科学与技术创新实验室, 福建 福州 350116;

3. 福州大学 物理与信息工程学院, 福建 福州 350116)

摘要: 可见光-红外双模态目标检测在低照度、复杂恶劣场景中具有显著优势, 但现有方法普遍依赖图像严格配准, 难以处理实际应用中广泛存在的跨模态空间错位问题。为此, 提出一种非对齐场景下的轻量化检测框架 CEFNet, 以突破传统像素级强制对齐的局限。该框架首先通过频域解耦特征补偿模块, 利用离散小波变换实现特征频域解耦, 再将跨模态高频差异作为状态空间模型的软门控信号, 阻断局部错位噪声在全局序列中的扩散, 在无需像素级强制对齐的前提下完成高效抗噪特征补偿。其次, 通过多分支非对称卷积增强模块扩大感受野, 强化不同模态的独有特征表达。最后, 构建差模引导融合模块, 自适应抑制共模冗余背景并放大高判别性互补特征。实验结果表明, CEFNet 在非对齐数据集 DV-TOD 上达到 87.1% mAP50 和 53.1% mAP50-95; 相较于当前最优 CDC-YOLOFusion, 参数量降低 90%, 计算量减少 79%, 精度提升 0.6%~1%, 并在 M³FD, LLVIP 等数据集上表现出优异的稳定性与泛化能力。

关键词: 可见光-红外目标检测; 空间错位; 特征补偿; 特征增强; 特征融合

中图分类号: TP391.4 **文献标识码:** A

doi: 10.3724/OPE.20263419.3104

CSTR: 32169.14.OPE.20263419.3104

Misaligned visible-infrared object detection based on feature compensation and enhancement fusion

ZHOU Yonggen^{1,2}, LIN Jianpu^{1,2}, LÜ Shanhong^{1,2}, LIN Zhixian^{1,2,3},

LIN Shanling^{1,2*}, GUO Tailiang^{2,3}

(1. School of Advanced Manufacturing, Fuzhou University, Quanzhou 362251, China;

2. Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China,
Fuzhou 350116, China;

3. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China)

* Corresponding author, E-mail: sllin@fzu.edu.cn

Abstract: Visible-infrared object detection offers substantial advantages in low-light conditions and challenging environments. However, most existing methods rely on precise image registration, limiting their applicability to the cross-modal spatial misalignment commonly encountered in practice. To overcome this

收稿日期: 2026-06-10; 修订日期: 2026-08-11.

基金项目: 国家重点研发计划资助项目(No. 2022YFB3603705); 国家自然科学基金青年基金资助项目(No. 62101132)

limitation, CEFNet, a lightweight detection framework for misaligned visible - infrared images, is proposed without requiring explicit pixel-level alignment. First, a frequency-domain decoupled feature compensation module is introduced, in which features are decomposed using the discrete wavelet transform. Cross-modal high-frequency discrepancies are then employed as soft gating signals for a state-space model to suppress the propagation of local misalignment noise during global sequential modeling, enabling efficient, noise-robust feature compensation. Second, a multibranch asymmetric convolution enhancement module is designed to expand the receptive field and strengthen modality-specific feature representations. Finally, a difference-guided fusion module is developed to adaptively suppress shared, redundant background information and emphasize discriminative complementary features. Experimental results demonstrate that CEFNet achieves an mAP50 of 87.1% and an mAP50-95 of 53.1% on the misaligned DV-TOD dataset. Compared with the state-of-the-art CDC-YOLOFusion framework, the parameter count and computational complexity are reduced by 90% and 79%, respectively, while detection accuracy is improved by 0.6%-1.0%. Further evaluations on M³FD and LLVIP demonstrate enhanced robustness and generalization, supporting the applicability of CEFNet to efficient multimodal object detection under cross-modal spatial misalignment.

Key words: visible-infrared object detection; spatial misalignment; feature compensation; feature enhancement; feature fusion

1 引 言

目标检测作为计算机视觉领域的基础核心任务,在自动驾驶^[1]、智能安防监控^[2]、无人机遥感等场景中具备关键应用价值。传统检测方法高度依赖可见光(Red-Green-Blue, RGB)图像,其丰富的色彩、纹理与细节信息可为检测提供充足特征支撑。然而,在低照度、强烟雾及暴雨等恶劣成像条件下,可见光图像极易出现特征模糊,信噪比下降等问题,导致检测精度显著降低。红外(Infrared, IR)成像依靠目标与环境间的热辐射差异实现感知,对光照变化、遮挡与复杂天气具备天然稳定性。由此可见,可见光与红外模态存在高度互补性,二者融合可弥补单一感知缺陷。因此,可见光-红外多模态融合检测已成为研究热点。Fang等^[3]率先提出跨模态融合 Transformer结构,同步实现模态内与模态间特征融合。Shen等^[4]在其基础上进一步构建跨模态双重交叉注意力,增强全局特征建模能力并强化跨模态互补信息挖掘。Wu等^[5]利用双向模态引导的注意力机制促进跨模态信息交互,并结合浅层空间与边缘特征增强,提升了复杂场景下分类与定位精度。然而,现有基于Transformer的多模态融合方法虽然具备较强的全局建模能力,但其自注意力

机制计算复杂度较高,在高分辨率与多尺度特征融合场景下易带来较大的计算开销。近年来,基于状态空间模型(State Space Model, SSM)的Mamba^[6]架构凭借线性复杂度、高效长程依赖建模以及全局信息感知能力,逐渐引入视觉任务。相比传统卷积与Transformer结构,Mamba能够在较低计算成本下实现更高效的跨模态特征交互,为可见光-红外目标检测提供了新的融合思路。Dong等^[7]首次探索了状态空间模型在跨模态融合中的应用,提出在隐状态空间中进行特征交互,以克服模态差异并显著提升目标检测性能。Liu等^[8]针对离散特征空间交互局限,将隐式神经表征引入跨模态融合,学习连续映射关系与跨尺度无损对齐策略,充分利用多语义层级互补信息。上述方法推动了多模态检测技术发展,但大多建立在可见光与红外图像严格配准的前提。

在实际采集场景中,受传感器时间不同步,空间分辨率差异,成像视角偏差与平台运动等因素影响,可见光与红外图像间普遍存在像素错位、目标偏移、尺度失配等未对齐问题,进而引发融合特征语义混淆、噪声干扰等问题,严重降低检测性能,甚至出现融合效果不及单模态的情况。为此,研究人员针对跨模态错位问题开展大量对齐与融合研究。Yuan等^[9]提出校准互补

Transformer(C²Former),利用跨模态交叉注意力学习全局特征相关性,以获得对齐且互补的融合特征。Liu等^[10]结合交叉Mamba设计偏移引导融合机制,利用多尺度特征实现对轻微模态偏移的自适应缓解,同时也减少了计算开销并提高了效率。Qian等^[11]提出双分支不对称交叉注意力融合方法DACFusion,基于空间位置对应关系实现特征交互与非对称融合,并结合大核卷积扩大感受野。Song等^[12]提出CMADet对齐检测框架,通过形变场估计与双向特征校正融合模块实现双模态特征对齐。尽管上述方法在一定程度上缓解了错位影响,但仍存在显著局限:基于交叉注意力的方法易将局部错位噪声扩散至全局,导致深层特征污染;基于形变场估计的方法高度依赖局部空间梯度,在低照度、低对比度区域易因监督信号缺失出现偏移发散。

除模态错位外,可见光-红外检测面临另一核心挑战:双模态固有特征差异显著,且场景适应性不同。现有方法虽尝试动态调节融合过程,但大多难以实现互补信息的最大化利用。Zeng等^[13]提出MMI-Det检测算法,通过轮廓增强与聚焦融合模块提升特征提取能力。Wang等^[14]设计跨尺度动态卷积融合模块CDCF,自适应提取双模态特征并引入交互损失尝试强化独有特征学习。刘远航等^[15]提出双模态优势特征融合模块,利用可学习的动态权重机制对双模态特征进行校准,从而实现了特征层面的深度融合。代小波等^[16]提出一种基于频率感知的特征融合模块,克服了模态特征差异并实现多尺度特征融合。然而,这些方法在表征与融合机制上仍存在局限。一方面,缺乏针对模态特定优势(如可见光的多向精细纹理与红外的全局热辐射轮廓)的定向结构化提取,导致模态潜力挖掘不充分;另一方面,在融合阶段缺乏对共模冗余与差模互补特征的解耦处理,这导致网络在复杂多变的环境下极易产生特征干扰,难以保持稳定的检测性能。

综上所述,现有可见光-红外融合检测方法主要存在三点不足:一是显式对齐方法易引入噪声扩散与结构畸变,难以应对复杂未对齐场景;二是网络感受野与建模能力有限,跨模态多方向特征提取不充分;三是融合策略对互补特征挖掘不足,易受背景与冗余信息干扰。为此,本文提

出一种适用于非对齐场景的渐进式可见光-红外目标检测框架CEFNet,该方法突破传统先对齐、后融合的固定范式,构建频域补偿、特征增强与差分聚合一体化的融合机制,可自适应不同程度的模态错位场景。

本文主要贡献为提出特征补偿模块(Feature Compensation Module,FCM)。与现有在空间域执行像素级强制对齐的思路不同,从频域解耦视角重构跨模态对齐与融合范式,借助小波变换将粗粒度稳定结构与含错位噪声的细粒度突变特征分离,并将跨模态高频差异作为Mamba状态空间模型的软门控信号,通过选择性门控机制阻断错位噪声扩散,在保留原始结构信息的同时实现高效抗噪融合;提出特征增强模块(Feature Enhancement Module,FEM),利用多分支非对称卷积强化模态特有优势,并借鉴差分放大器机制,设计差分引导特征融合模块(Feature Fusion Module,FFM),自适应抑制共模冗余背景,精准放大高判别性差模互补特征;在DVTOD,M³FD,LLVIP等包含不同错位程度与复杂环境的主流数据集上进行大量实验,结果表明:CEFNet检测精度显著优于当前主流算法,同时在参数量、计算量与推理速度间实现良好平衡,具备更强的稳定性与实用性。

2 原 理

如图1所示,以YOLOv11^[17]架构为基础,将其拓展为双分支并行结构,以实现可见光与红外图像的联合处理。模型采用五阶段对称骨干网络(Stage1~Stage5)完成双模态特征提取,逐级提取低层边缘细节与高层语义信息的多尺度特征。随后,为抑制空间错位带来的特征混淆问题,在Stage2~Stage4的中间层特征处引入特征补偿模块FCM。该模块摒弃传统像素级强制对齐思路,采用离散小波变换实现特征频域解耦,并结合跨模态Mamba交互机制,将高频差异信息作为软门控信号,主动抑制局部错位噪声在全局序列中的传播,从源头缓解由像素偏移导致的特征混叠问题。在深层特征Stage3,4,5尺度上依次嵌入特征增强模块FEM与特征融合模块FFM。FEM采用多分支非对称大核卷积,在轻量化前提下增强不同模态下目标的专属形态特征;在此基础上,

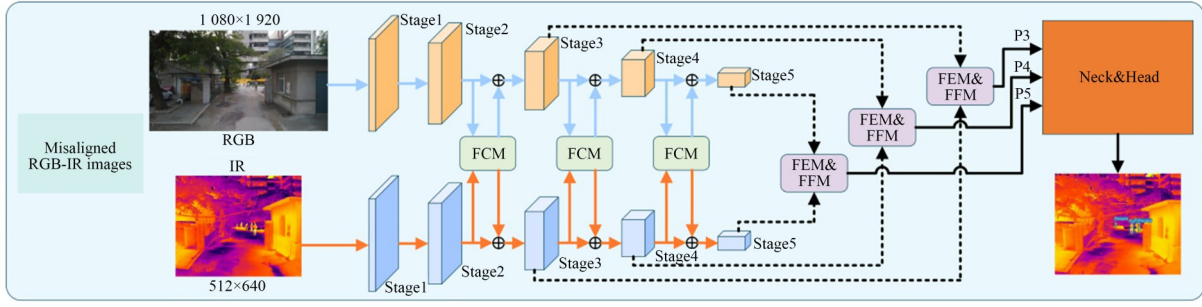


图1 CEFNet的网络结构

Fig. 1 Network architecture of CEFNet

FFM借鉴差分放大器工作机理,自适应抑制模态间共模冗余背景,精准放大高判别性差模互补特征,实现跨模态信息的深度高效融合。最后,将P3,P4,P5三路多尺度融合特征经路径聚合特征金字塔网络(Path Aggregation Feature Pyramid Network, PAFPN)完成特征聚合,沿用YOLOv11的解耦检测头输出目标分类与定位结果。

2.1 特征补偿模块 FCM

受成像机理差异影响,可见光和红外图像在空间结构、纹理丰富度和热辐射分布上差异显著,易引发模态语义失配和空间错位,从而阻碍了直接特征融合的有效性。复杂场景中空间错位与语义偏移共存,会进一步弱化模态特征关联,模糊特征对应关系,盲目学习模态相关性易产生模态冲突,引入额外噪声并加剧特征混淆。为解决这些问题,提出跨模态特征补偿模块FCM,放弃了传统的显式对齐范式,转而在特征融合中期自适应学习模态特异性变化。该模块将错位感知能力直接融入融合过程,聚集强化弱目标特征表达,无需前置对齐操作。FCM的结构如图2所示。

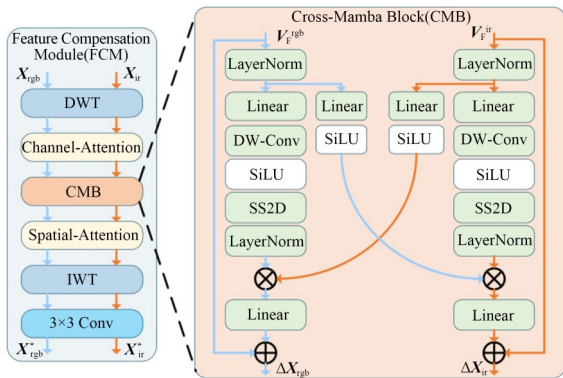


图2 FCM模块结构

Fig. 2 Structure of FCM module

2.1.1 基于小波变换的频域结构化解耦

在直接进行特征交互前,仅在空间域对未对齐的多模态特征进行处理,往往难以有效区分目标的固有结构与传感器视差引发的位置噪声。为突破空间域的这一局限性,本文引入基于多分辨率分析(Multiresolution Analysis, MRA)理论的离散小波变换(Discrete Wavelet Transform, DWT)。作为一种强大的时频域分析工具,小波变换可通过一组正交高通与低通滤波器组,在无信息损失的前提下将图像特征表征精准解耦为两类子带。低频子带表征图像全局平滑结构与语义近似的信息,高频子带则捕获边缘纹理与局部突变信息(如由空间错位产生的伪边缘和散粒噪声),同时弥补了传统傅里叶变换空间局部定位能力不足的缺陷。

在FCM模块中,首先利用DWT对输入的可见光特征 $X_{rgb} \in \mathbb{R}^{B \times C \times H \times W}$ 和红外特征 $X_{ir} \in \mathbb{R}^{B \times C \times H \times W}$ 进行频域分解,经单层DWT分解后,特征图被结构化解耦为包含全局粗粒度结构近似的低频子带(X_{LL})以及包含细粒度局部错位细节的高频子带(X_{LH}, X_{HL}, X_{HH}),4个子带的大小均为 $\mathbb{R}^{B \times C \times H/2 \times W/2}$ 。其中, B, C, H 和 W 分别表示批量大小、通道数、特征图高度和宽度。该过程可表示为:

$$\{X_{LL}^{rgb}, X_{LH}^{rgb}, X_{HL}^{rgb}, X_{HH}^{rgb}\} = \text{DWT}(X_{rgb}), \quad (1)$$

$$\{X_{LL}^{ir}, X_{LH}^{ir}, X_{HL}^{ir}, X_{HH}^{ir}\} = \text{DWT}(X_{ir}). \quad (2)$$

为精准识别模态间的错位噪声与特异性特征,将各高频子带采用相加方式以获取频域聚合特征 $X_F^{rgb} \in \mathbb{R}^{B \times C \times H/2 \times W/2}$ 与 $X_F^{ir} \in \mathbb{R}^{B \times C \times H/2 \times W/2}$,并在不增加通道维度的前提下实现水平、垂直和对角方向的高频信息的有效聚合。随后,通过计算模态间的差异变化量,并结合通道注意机制与非

线性激活函数 (Rectified linear Unit, ReLU), 提取出最具信息量的模态特异性变化特征 $V_F^{rgb} \in \mathbf{R}^{B \times 2C \times H/2 \times W/2}$ 与 $V_F^{ir} \in \mathbf{R}^{B \times 2C \times H/2 \times W/2}$:

$$V_F^{rgb} = \text{Concat}(X_F^{ir} - \text{ReLU}(X_F^{rgb}), X_F^{rgb}), \quad (3)$$

$$V_F^{ir} = \text{Concat}(X_F^{rgb} - \text{ReLU}(X_F^{ir}), X_F^{ir}), \quad (4)$$

其中: $\text{Concat}(\cdot)$ 表示沿通道维度的拼接, $\text{ReLU}(\cdot)$ 为非线性激活函数。这一操作隔离了共模的平滑背景, 突出了跨模态的高频冲突与特异性细节。

2.1.2 跨模态 Mamba 调制

为增强跨模态交互、抑制模态未对齐带来的干扰特征, 将提取出的差异特征 V_F^{rgb} 和 V_F^{ir} 输入由 VSS 模块^[18]构建的交叉 Mamba 调制块 (Cross Mamba Block, CMB) 中。在连续状态空间方程中, 隐状态的更新易受到局部突变噪声的污染。为了打破这一限制, 与依赖显式对应关系的传统交叉注意力不同, CMB 采用频率感知的交叉调制策略, 如图 2 所示。CMB 将来自另一模态的高频差异信息作为软门控信号注入到 Mamba 的选择性扫描机制中。输入特征 V_F^{rgb} 和 V_F^{ir} 经过线性投影后分为两个并行的分支: 表征分支用于模态特征学习, 依次经过 3×3 深度卷积 (DW-Conv)、SiLU 激活函数、核心的 SS2D 模块以及层归一化 (LayerNorm) 处理, 保留模态特有的信息; 门控分支仅通过一个 SiLU 激活函数, 用来调节互补模态信息, 动态地抑制与任务无关的激活。CMB 采用交叉门控方式, 将红外门控分支输出与可见光表征分支经 SS2D 处理后的特征进行逐元素相乘, 反之亦然。该过程不直接修改 Mamba 状态空间模型中的参数矩阵, 而是经 SS2D 完成全局序列建模后, 对其输出特征进行跨模态自适应筛选。该调制同时作用于空间位置和通道响应, 当跨模态高频差异较大时, 门控机制削弱由错位引起的伪边缘、目标偏移及局部噪声响应, 而在语义一致区域保留有效的互补信息。最后, 两个经过跨模态门控的表征在通道维度上进行合并, 并通过共享线性层实现模态协同, 再经特征拆分与残差连接, 输出增强了弱目标且抑制了噪声的补偿特征 ΔX_{rgb} 和 ΔX_{ir} 。

这一协同机制确保了 Mamba 在严重错位的情况下依然能进行稳定的特征交互, 不仅保留了长程依赖建模的优势, 更从源头上消除了未对齐

带来的干扰表示。在此基础上, 将 CMB 学习到的变化量 ΔX_{rgb} 和 ΔX_{ir} 分别与对应的低频 (LL) 特征图进行逐元素相加, 以修复其结构语义, 并重新与高频特征拼接, 得到状态重构特征 X_{rgb}^{sta} 和 X_{ir}^{sta} , 其大小均为 $\mathbf{R}^{B \times 4C \times H/2 \times W/2}$:

$$X_{rgb}^{sta} = \text{Concat}(X_{LL}^{rgb} + \Delta X_{rgb}, X_{HL}^{rgb}, X_{HL}^{rgb}, X_{HH}^{rgb}), \quad (5)$$

$$X_{ir}^{sta} = \text{Concat}(X_{LL}^{ir} + \Delta X_{ir}, X_{HL}^{ir}, X_{HL}^{ir}, X_{HH}^{ir}). \quad (6)$$

为进一步保障跨模态的语义一致性, 本文计算两个模态在频域内的余弦相似度矩阵 S_{cos} :

$$S_{cos} = \frac{X_{rgb}^{sta} \cdot X_{ir}^{sta}}{\|X_{rgb}^{sta}\| \|X_{ir}^{sta}\|}. \quad (7)$$

这种基于余弦相似度的特征增强机制, 能够选择性地放大学习到的模态特异性变化, 从而凸显在语义或结构上高度一致的空间区域。增强后的特征表示为:

$$X_{rgb}^{enh} = X_{rgb}^{sta} + X_{ir}^{sta} \odot \text{ReLU}(S_{cos}), \quad (8)$$

$$X_{ir}^{enh} = X_{ir}^{sta} + X_{rgb}^{sta} \odot \text{ReLU}(S_{cos}). \quad (9)$$

最后, 对增强后的各频率分量量子带进行小波逆变换 (Inverse Wavelet Transform, IWT), 并通过 3×3 卷积映射回空间域, 输出最终的补偿特征 X_{rgb}^* 和 X_{ir}^* , 其大小均为 $\mathbf{R}^{B \times C \times H \times W}$ 。

2.2 特征增强模块 FEM

在复杂的现实场景中, 由于模态异构特性, 目标尺度跨度大且方向多样, 而传统卷积对不同方向目标的适应性不足。此外, 同一目标在不同模态图像中的激活程度也存在差异, 这些因素共同导致了模态特征不平衡问题。为了解决这些挑战, 本文设计了一个特征增强模块 FEM, 如图 3 所示。首先, 采用带有批归一化 (Batch Normalization, BN) 和 SiLU (Sigmoid Linear Unit) 激活函数的 1×1 标准卷积层, 旨在不改变特征图尺寸的情况下增强特征的非线性表达能力。进一步, 分别利用核大小为 1×63 的横向卷积层, 核大小为 63×1 的纵向卷积层, 膨胀率为 3 的 3×3 空洞卷积层以及 1×1 深度点卷积层来捕捉不同区域的特征信息。其中, 1×63 与 63×1 非对称大核卷积通过大核分解方式扩大感受野, 相比直接采用 63×63 卷积可显著降低参数量和计算量, 并分别增强水平与垂直方向的结构特征建模能力。为避免大核卷积过度平滑小目标边缘, FEM 同时引入空洞卷积和深度点卷积以补充局部细节, 从而在全局上下文建模和局部细节保持之间取得

平衡。随后,应用拼接操作对提取的信息进行融合,并通过通道混洗操作(Channel Shuffle, CS)对融合后的特征进行重排,以提升模型的表达能力,同时使用 1×1 卷积层进行特征降维。最后,应用残差结构,将初始信息添加到增强后的特征中,进一步扩充特征表达。

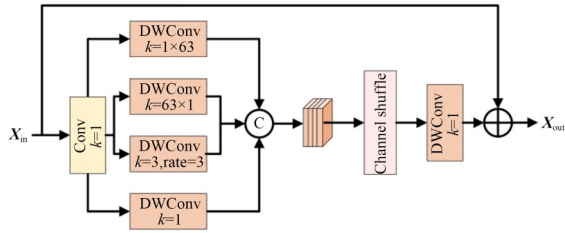


图3 特征增强模块的详细架构
Fig. 3 Structure of FEM module

2.3 特征融合模块 FFM

虽然FEM通过多分支非对称大核卷积强化了模态特定优势,但在双模态互补特征挖掘与跨模态信息交互仍有提升空间。受差分放大器原理的启发,提出差分引导的特征融合模块FFM,结构如图4所示。传统融合方法(如直接求和或拼接)易引入差异信息和冗余信息,而FFM聚焦于模态差异化特征的深度挖掘与建模,以此强化双模态互补关联和信息交互,同时可在空间与通道双维度动态分配注意力权重,精准聚焦显著目标特征,提升特征融合的针对性与有效性。

FFM的核心为差分特征增强部分。根据差分放大器抑共模、扩差模原理,两模态共有的环境信息被视为共模分量,而目标在不同光谱下的表征差异(如红外热辐射与可见光纹理细节)则被视为差模信号,该结构能够提取有效差异特征。这一特性已在各类多模态感知研究中得到有效验证。基于此,将可见光特征图 X_{rgb} 与红外特征图 X_{ir} 作为FFM的双路输入,通过对两个模态特征图进行相减操作以获取差分模态特征 X_{dif} 。然后,通过平均池化和最大池化操作对差分增强特征的空间信息进行聚合,以获得全局感受野。聚合后的空间信息随即通过一个共享多层感知机进行处理,生成通道注意力权重图。通过将生成的通道注意力权重图分别与可见光特征 X_{rgb} 与红外特征 X_{ir} 执行逐元素相乘,完成差分特征的自适应精炼。最后,引入残差求和融合机制,将保留模态固有属性的基础特征与挖掘得到

的模态差异互补特征进行深度融合,以确保融合后的特征 X_{fuse} 能同时涵盖模态间的共性关联与特异性差异,从而保留各模态中原始有效特征信息的完整性,实现高质量双模态特征融合。

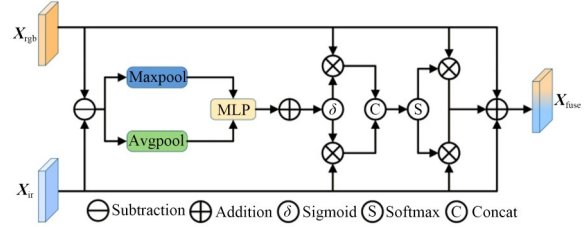


图4 特征融合模块的详细架构

Fig. 4 Structure of FFM module

2.4 损失函数

本文模型的损失函数与YOLOv11^[17]保持一致,主要由分布焦点损失 L_{dif} 、目标分类损失 L_{cls} 以及目标定位损失 L_{loc} 三个部分组成。其损失函数公式如下:

$$L_{all} = \lambda_{dif} L_{dif} + \lambda_{cls} L_{cls} + \lambda_{loc} L_{loc}, \quad (10)$$

其中: L_{all} 包含3个部分, λ 为表征各部分权重的超参数。这些权重可根据实际情况在训练前进行调整。在本文中,这3个部分的权重分别被设定为1.0、0.5和0.05。

3 实验结果与分析

3.1 数据集

本文采用了一个非对齐数据集DVTOD^[12]验证模型在非对齐任务下的有效性和准确性,并采用两个近似对齐数据集M³FD^[19]和LLVIP^[20]来验证模型在对齐任务下的泛化能力。

DVTOD数据集是首个由无人机采集的非对齐RGB-IR目标检测数据集。该数据集共包含2 179对图像,涵盖了不同时间段及季节条件下汽车、行人、自行车等目标类别。该数据集被划分为1 606对训练集和573对测试集。

M³FD数据集是一个存在轻微未对齐现象的基准数据集,在复杂交通和遮挡环境等场景下采集,包含4 200对分辨率为 $1\,024 \times 768$ 的可见光-红外图像对。该数据集分别有3 360对和840对图像被用于模型的训练和评估。

LLVIP数据集是一个专注于在弱光条件下进行多光谱行人检测的数据集。它包含15 488

个对齐的 RGB-IR 图像对,其中 12 025 个图像用于训练,3 463 个图像用于测试。

3.2 实验设置

所有实验均在 NVIDIA 4090D 图形处理器上进行,所提出的模型基于 PyTorch 框架实现。在训练过程中,网络中的输入图像分辨率均被调整为 640×640 像素,对于 DVTOD、M³FD 和 LL-VIP 数据集,分别将训练轮次设置为 300,300 和 100, Batch size 统一设置为 16。遵循 YOLOv11^[17] 的默认训练配置,模型采用随机梯度下降 (Stochastic Gradient Descent, SGD) 算法进行优化,初始学习率设为 0.01,动量为 0.937,权重衰减为 0.000 5。

3.3 评价指标

为了准确评估网络的检测性能,本文采用平均精度 (Average Precision, AP) 和平均精度均值 (Mean Average Precision, mAP) 作为评价指标。AP 是基于精确率 (P) 和召回率 (R) 计算的复合指标,通过计算精确率-召回率曲线 (PR 曲线) 下的面积得出。mAP 是所有类别 AP 值的算术平

均数,反映了模型在多类别任务中的综合性能。计算方法如式 (11)~式 (13) 所示:

$$P = \frac{N_{ii}}{N_i}, R = \frac{N_{ii}}{N_i}, \quad (11)$$

$$AP = \int_0^1 P(R) dR, \quad (12)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N \int_0^1 P_i(R) dR, \quad (13)$$

其中: N_{ii} 表示被正确检测到的样本数量, N_i 表示所有被检测为类别 i 的样本数量, N_i 表示实际属于类别 i 的所有样本数量, P 代表精确率, R 代表召回率, N 表示类别总数。

为了准确评估模型的推理速度与计算成本,采用帧率 (FPS)、网络参数量 (Params) 以及浮点运算次数 (FLOPs) 作为评价指标。

3.4 对比结果与分析

3.4.1 非对齐任务 DVTOD 数据集上的比较结果

为了验证所提出的算法在非对齐任务下的有效性,首先比较了不同模型在 DVTOD 上的性能,如表 1 所示。实验结果表明,受限于 DVTOD 数

表 1 DVTOD 数据集上本文方法与最新方法的比较

Tab. 1 Comparison of proposed method with state-of-the-art methods on DVTOD dataset (%)

Methods	Modality	Backbone	Person	Car	Bicycle	mAP50	mAP50-95
YOLOv5-S	RGB	YOLOv5	20.8	44.1	15.3	26.7	11.7
YOLOv8-S		YOLOv8	21.5	50.8	21.4	31.2	13.8
YOLOv11-S		YOLOv11	28.2	51.1	28.1	35.8	15.6
YOLOv12-S		YOLOv12	29.2	53.3	20.4	34.3	15.1
YOLOv13-S		YOLOv13	23.1	46.0	20.2	29.8	14.6
YOLOv5-S	IR	YOLOv5	84.2	70.7	85.6	80.2	47.2
YOLOv8-S		YOLOv8	84.6	71.9	87.1	81.2	48.3
YOLOv11-S		YOLOv11	84.8	71.3	86.5	80.9	47.9
YOLOv12-S		YOLOv12	85.1	70.3	82.9	79.5	46.5
YOLOv13-S		YOLOv13	83.4	69.4	87.2	80.0	46.6
CFT ^[3]	RGB+IR	YOLOv5	90.9	74.8	82.1	82.6	47.6
ICAFusion ^[4]		YOLOv5	89.8	81.9	82.9	84.9	45.2
COMO ^[10]		YOLOv8	90.5	79.8	83.7	84.7	52.7
CMA-Det ^[12]		YOLOv5	90.4	81.6	83.0	85.0	46.7
CDC-YOLOFusion ^[14]		YOLOv5	91.6	83.6	84.2	86.5	52.1
SuperYOLO ^[21]		YOLOv5	78.8	56.2	76.8	70.6	38.3
DEYOLO ^[22]		YOLOv8	88.8	79.8	84.8	84.5	52.7
JTMDet ^[23]		YOLOv5	90.2	77.2	77.4	81.6	46.0
MSTF ^[24]		YOLOv5	88.6	73.1	78.3	80.0	43.5
CEFNet(Ours)		YOLOv11	91.4	84.5	85.4	87.1	53.1

据集在弱光环境下的成像质量,可见光模态的检测性能明显低于红外模态。由于红外成像能够提供更显著的目标热对比度,而可见光图像富含纹理细节,结合两者的多模态检测算法取得了最优结果,有效证实了异源信息在目标表征上的互补性。在与现有主流先进检测模型 CFT^[3],ICAFusion^[4]及 CMA-Det^[12]等的横向对比中,尽管本文方法并非在所有单一类别上均达到最优,但在 mAP50 指标以及在更为严格的 mAP50-95 指标上分别达到了 87.1% 和 53.1%,远高于以往优秀模型。这一突出表现的根本原因在于,本文方法从特征解耦与抗噪补偿的底层逻辑上,克服了现有主流算法在错位场景下的表征瓶颈。具体而言,CFT^[3]和 ICAFusion^[4]都是利用注意力机制来捕获全局上下文依赖,在错位场景下进行特征交

互极易将局部错位噪声扩散至全局,导致性能下降;而 CMA-Det^[12]虽借助形变场估计获得了一定的对齐增益,但其在低照度、低对比度等恶劣环境下极易因缺乏有效监督信号导致形变场发散,进而引发对齐机制失效。定性可视化结果如图 5 所示(彩图见期刊电子版),对于行人(People)类别,未对齐问题以及现有对齐融合策略的缺陷通常会导致相邻目标发生合并,其他方法经常将站距较近的个体检测为单个目标。相比之下,本文的小波分离结构保留了清晰的边缘和边界,使该类别的 mAP50 达到了 91.4%。此外,差分融合策略抑制了冗余的跨模态特征,降低了表征的模糊性。同时对于受空间错位影响最严重的汽车(Car)类别,本文方法取得了 84.5% 的 mAP50,证明了其对严重空间未对齐现象也具有较稳定的稳定性。

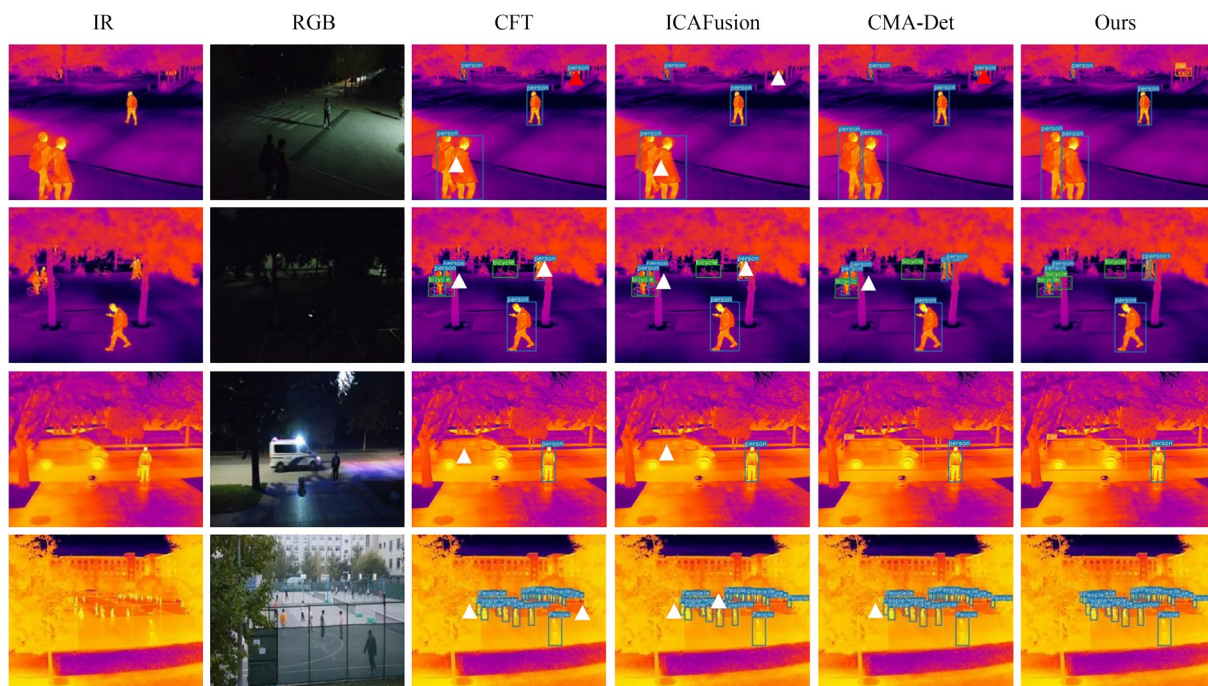


图 5 本文方法与 CFT,ICAFusion 和 CMA-Det 在 DVTOD 数据集上的可视化对比(白色三角形标记漏检目标,红色三角形标记误检目标)

Fig. 5 Visual comparison of our proposed method with CFT, ICAFusion, and CMA-Det on the DVTOD dataset(White triangles mark missed targets, and red triangles mark false positives)

3. 4. 2 M3FD 数据集上的比较结果

为进一步验证本文算法在可见光-红外融合目标检测上的有效性,将本文提出的方法与其他优秀的模型在对齐任务 M³FD 数据集上进行了对比,结果如表 2 和图 6 所示。表中数据表明,本

文方法在 mAP50 和 mAP50-95 指标上均优于其他竞争方法,这意味着该方法在对齐任务下也具有良好的迁移性与泛化能力,为可见光与红外图像融合检测提供了一种更为有效的解决方案。为了直观地反映不同方法的检测性能,对预测结

表 2 M³FD数据集上本文算法与最新算法的比较结果

Tab. 2 Comparison of proposed method with state-of-the-art methods on M³FD dataset (%)

Methods	Modality	mAP50	mAP50-95
CFT ^[3]	RGB+IR	68.2	42.5
ICAFusion ^[4]		67.8	41.9
COMO ^[10]		85.7	59.7
SuperYOLO ^[21]		85.7	55.2
MSTF ^[24]		52.9	30.1
WtCAFNet ^[25]		85.2	58.2
SLBAF ^[26]		78.9	44.3
EAEF ^[27]		80.1	—
RGB-X ^[28]		81.5	—
DAMSDet ^[29]		80.2	52.9
CEFNet(ours)		86.1	60.3

果进行了可视化,如图 6 所示。可以观察到,在浓烟或密集遮挡等极端环境下,前沿的 ICAFu-sion^[4]方法因跨模态特征感知不足漏检了行人(People)、汽车(Car)类别,而本文提出的方法则能够成功识别这些目标。在面对包含大量干扰源的复杂交通或弱光场景时,本文方法能够生成精确的预测边界框,而 CFT^[3]和 ICAFusion^[4]均出现了不同程度的漏检和误检现象。WtCAF-Net^[25]虽然降低了漏检率,却在背景复杂的区域产生了明显的误检,这表明其在融合过程中缺乏对冗余信息与背景噪声的抑制。相比之下,本文提出的方法在各类场景中均表现出极高的稳定性,充分证明了其在复杂开放环境下的卓越检测能力。

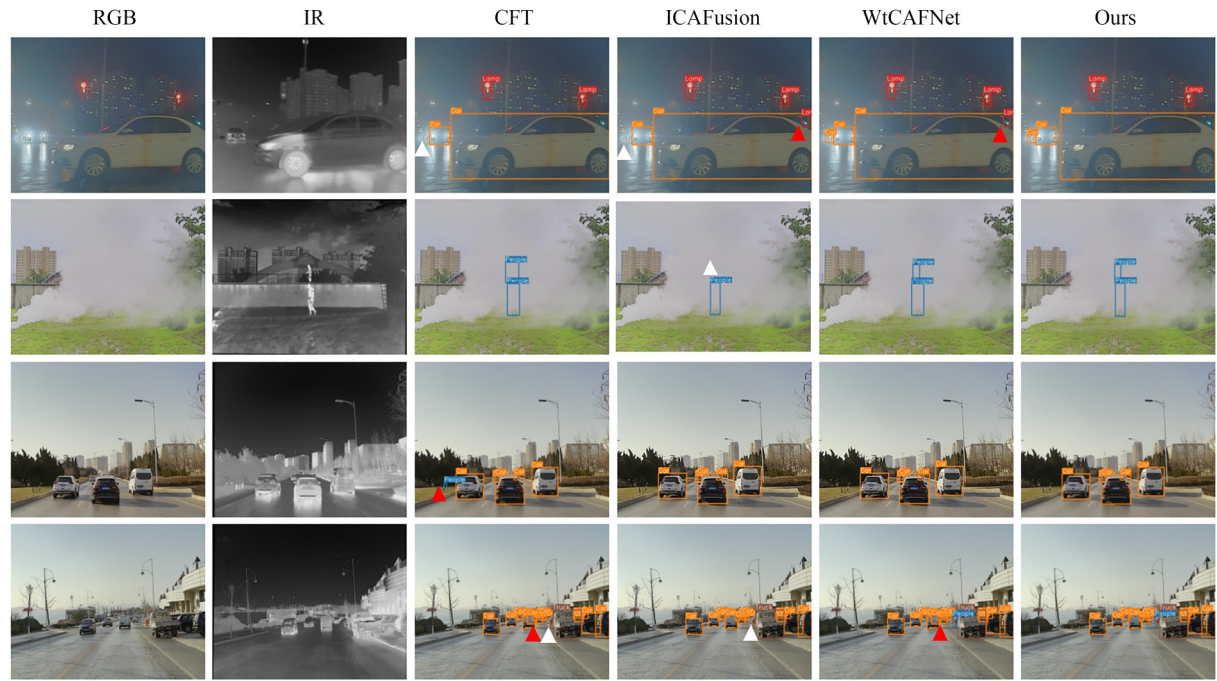


图 6 本文方法与 CFT, ICAFusion 和 WtCAFNet 在 M³FD 数据集上的可视化对比(白色三角形标记漏检目标,红色三角形标记误检目标)

Fig. 6 Visual comparison of our proposed method with CFT, ICAFusion, and WtCAFNet on M³FD dataset(White triangles mark missed targets, and red triangles mark false positives)

3. 4. 3 LLVIP数据集上的比较结果

为了进一步验证本文方法的自适应能力,本文在模态对齐良好的 LLVIP 数据集上进行了对比实验,实验结果如表 3 所示。由实验结果可得,红外单模态检测算法性能显著优于可见光单模态检测算法。这主要归因于 LLVIP 数据集包含

大量低光照复杂场景,此类场景下可见光图像易出现特征退化、细节丢失等问题,难以提取有效目标特征。而红外图像不受光照条件干扰,可稳定提供清晰的目标轮廓与特征信息。与单模态检测算法相比,各类红外与可见光融合的多模态目标检测方法均表现出更优异的检测性能,充分

表 3 LLVIP数据集上本文算法与最新算法的比较结果

Tab. 3 Comparison of proposed method with State-of-the-art methods on LLVIP dataset (%)

Methods	Modality	mAP50	mAP50-95
YOLOv5	RGB	90.9	52.7
YOLOv8		90.4	54.3
DDQ-DETR ^[30]		86.1	46.7
YOLOv5		96.6	63.4
YOLOv8		96.0	65.8
DDQ-DETR ^[30]		93.9	58.6
CFT ^[3]	IR	97.5	63.6
ICAFusion ^[4]		96.3	62.3
TarDAL ^[19]		94.9	61.2
YOLOFusion ^[31]		95.4	62.8
TINet ^[32]		94.3	60.6
RSDet ^[33]		95.8	61.3
CSAA ^[34]	RGB+IR	94.3	59.2
CEFNet(Ours)		96.8	64.7

证明了多模态融合策略能够有效利用异源图像间的互补信息,从而大幅提升低照度环境下的目标检测精度。相较于现有的融合算法,本文方法在对齐良好的数据集上依然展现出了优越的性能。这主要得益于本文所设计FFM模块,该模块能够有效缓解跨模态特征图之间的表征差异,完成高质量的特征融合,充分挖掘并利用两个模态的有效信息,最终实现更优的目标检测效果。

3.5 消融结果与分析

3.5.1 模型各组件的消融实验

为了深入探究本文所设计各个模块(FCM, FEM, FFM)对网络整体性能的具体贡献及模块间的协同机制,在空间错位现象最为严重的DV-TOD数据集上进行了消融实验,结果如表4所示。结合表1中YOLOv11在红外图像上的检测结果可知,当红外图像和可见光图像直接输入到基于YOLOv11简单特征拼接双流基线网络时,模型检测性能相较于YOLOv11单模态红外检测模型虽有一定提升,但提升幅度有限。这是因为简单的拼接操作没有充分考虑到跨模态图像之间固有的空间错位和语义差异,融合特征易受未对齐噪声干扰,没有充分发挥多模态互补的潜力。在基线网络中加入FEM后,模型检测精度提升至84.7%,这得益于该模块可通过多尺度、多方向卷积操作充分挖掘图像局部细节和全局结构特征,显著增强了模型的特征表征能力。在此基础上加入FCM模块后,模型性能得到了更大幅度的提升,原因在于FCM借助离散小波变换对特征进行解耦,使模型在特征融合中期阶段学习不同模态的特征差异化信息,显著增强了弱目标的特征表示。最后,集成FFM模块使模型整体检测性能达到最优状态,充分验证了所提出的FFM模块在聚焦跨模态差异特征并抑制冗余信息进行深度融合的有效性,同时也说明FCM的特征补偿为FFM的差分融合提供了更干净的模态表征,避免了未对齐噪声被差模运算进一步放大,验证了先补偿后融合这一渐进式设计范式的合理性。

表 4 在 DVTOD 数据集上的消融结果

Tab. 4 Ablation results on DVTOD dataset

FCM	FEM	FFM	mAP50/%	mAP50-95/%	Params/M	FLOPs/G	FPS
×	×	×	83.4	49.4	14.22	33.33	227
✓	×	×	85.2	50.6	14.37	33.72	192
×	✓	×	84.7	49.6	15.20	38.65	116
×	×	✓	85.6	51.6	14.27	33.38	215
✓	✓	×	86.2	51.2	15.35	39.04	106
✓	×	✓	86.6	52.3	14.42	33.77	185
×	✓	✓	85.9	52.6	15.25	38.70	108
✓	✓	✓	87.1	53.1	15.40	39.09	100

3.5.2 FCM 模块的消融实验

为了深入探究所提出的 FCM 模块的架构设计合理性和稳定性,本文对其中两个关键超参数进行了广泛的敏感性分析:小波分解层数及 Mamba 模块堆叠深度。实验结果如图 7 所示,以此验证本文超参数选取的有效性,并展示了模型在不同参数范围内的性能。

3.5.2.1 小波分解层数

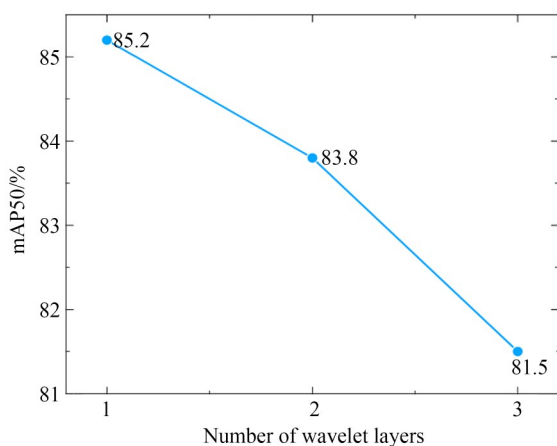
图 7(a)展示了在 Mamba 深度保持恒定为 1 的情况下小波层数对模型性能的影响。由实验结果可知,随着小波层数从 1 增加到 3,模型 mAP 持续下降。这表明单层小波分解足以提取任务所需的高频空间特征,更深的分解层数可能会导致特征表征过于平滑或过滤掉对目标精准定位至关重要的细粒度细节,从而导致性能下降。因此,本文将 FCM 中小波分解层数设定为 1,以实现最佳检测精度。

3.5.2.2 Mamba 模块深度

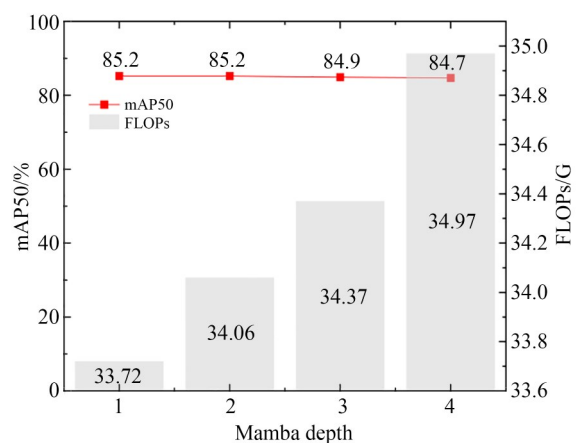
图 7(b)呈现了 Mamba 模块的堆叠数量从 1 调整到 4 时,模型检测精度(mAP50)和浮点运算次数(FLOPs)的变化。当仅使用单个 Mamba 模块时,模型在保持极低计算开销(33.72 GFLOPs)的同时,取得了 85.2% 的最优 mAP50 性能。这说明,在小波高频差异特征的引导下,浅层 Mam-

ba 序列扫描结构即可高效捕获跨模态的长距离依赖关系,并成功切断局部错位噪声的扩散路径。同时进一步增加模块的堆叠深度并没有带来任何额外的收益:当堆叠深度增加到 2 时,模型的 mAP50 停滞在 85.2%,但计算开销却上升至 34.06 GFLOPs;当深度增加到 3 和 4 时,mAP50 分别降至 84.9% 和 84.7%,出现持续性小幅性能下降,同时计算开销则持续增加至 34.97 GFLOPs。这一现象表明,在处理包含高频细节的视觉任务时,盲目堆叠状态空间模型 Mamba 模块会导致严重的特征过度平滑效应,使得原本锐利的跨模态高频边缘特征在深层扫描中被过度稀释;同时,冗余的序列扫描使得网络容易对错位噪声产生过拟合,并带来了额外的计算开销。因此,本文将 FCM 中 Mamba 模块的堆叠深度设定为 1,以实现检测精度与推理效率的最佳平衡。

为进一步直观验证 FCM 模块的特征优化与对齐效果,开展了特征可视化对比实验,如图 8 所示。从图中红框区域可见,FCM 有效地减少了模态之间的信息不平衡(彩图见期刊电子版)。该结果充分证明了 FCM 对跨模态特征进行频域无损解耦后,能够通过跨模态 Mamba 交互机制实现特征层面的语义对齐,从而消除由视差带来的位置噪声与冲突。



(a) 小波分解层数
(a) Number of wavelet decomposition levels



(b) Mamba 模块堆叠深度
(b) Mamba module stacking depth

图 7 FCM 模块中关键超参数对检测性能与计算开销的影响

Fig. 7 Impact of key hyperparameters in FCM module on detection performance and computational overhead

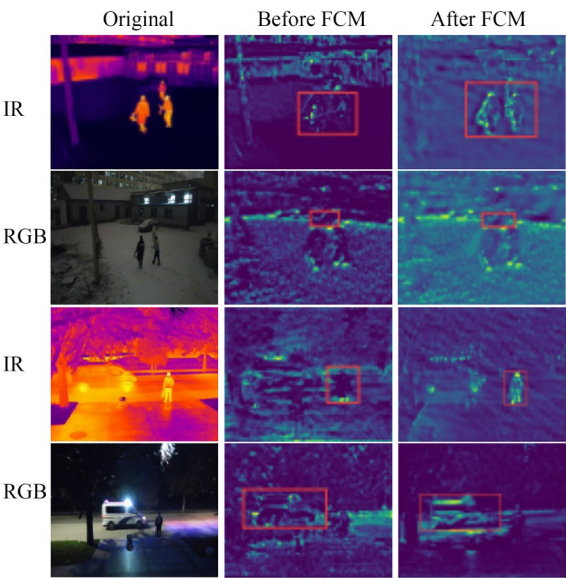


图 8 FCM 模块前后的特征图可视化(特征图的尺寸为 80×80)

Fig. 8 Feature map visualization before and after FCM module(Where the size of the feature maps is 80×80)

3. 5. 3 FEM 模块的消融实验

为了进一步探究FEM模块中大核卷积的计算复杂度及其对底层硬件资源消耗的影响,本文对比了不同卷积核配置下的模型参数量、显存占用与检测性能,实验结果如表 5 所示。实验结果表明,直接将卷积核从 3×3 扩大到 63×63 虽然能将 mAP50 提升 0.3%,但会导致 FLOPs 从 37.59 G 增至 73.27 G,严重降低了模型推理效率。相比之下,本文采用的非对称大核分解策略(1×63&63×1)不仅获得了最高的检测精度 84.7%,且其计算量(38.65 G)和 GPU 显存占用(7.61 G)均与基础模型相当。这证明了非对称大核卷积在有效捕捉全局特征的同时,保持了较好的效率平衡。

表 5 不同卷积核尺寸对模型性能与计算开销的影响

Tab. 5 Impact of different kernel sizes on model performance and computational cost

Kernel size	mAP50/%	Params/	FLOPs/	GPU
		M	G	Memory
3×3	83.8	14.05	37.59	7.78
63×63	84.1	18.13	73.27	7.88
1×63&63×1	84.7	15.20	38.65	7.61

3. 5. 4 FFM 模块的消融实验

为深入探究特征融合模块 FFM 在复杂场景下噪声抑制能力、自适应特征筛选效果及其内在作用机理,通过注意力热力图可视化方式对模块的工作机制进行直观验证。热力图以红色强度表示区域对预测结果的贡献度,即模型关注度。如图 9 所示,图(a)为白天复杂背景,图(b)为夜间弱光场景。可以发现,基准模型由于缺乏自适应的特征筛选机制,其注意力极易被背景中具有强反射或高热辐射的无关物体(如夜间场景中的树冠区域)所分散,导致不可忽视的特征冗余。相比之下,本文提出的 CEFNet 展现出了强大的目标聚焦与背景噪声抑制能力,有效地将双模态特征解耦为共模背景冗余分量与差模互补优势分量,在充分挖掘双模态互补信息的同时,削弱前景目标与复杂背景间的特征混淆度,实现精准的前景聚焦与背景去噪。

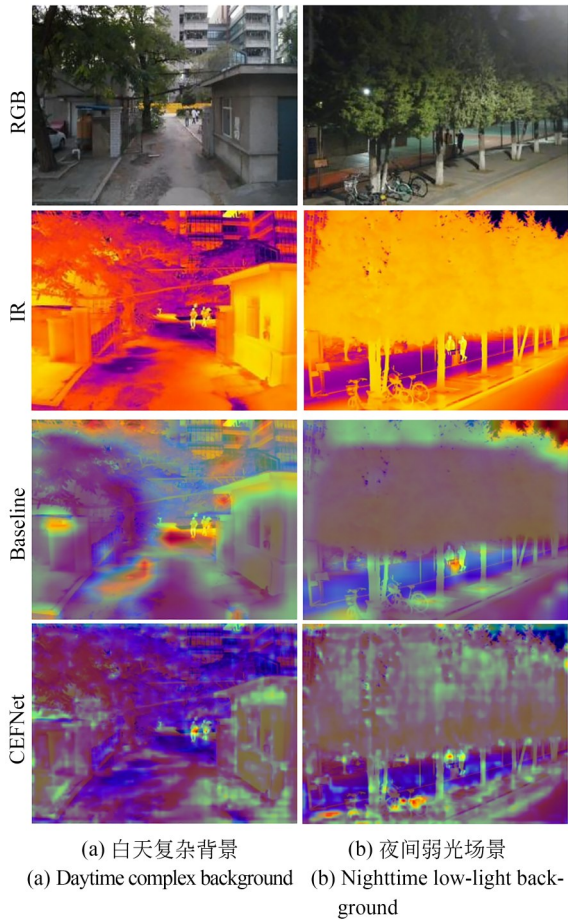


图 9 热力图对比

Fig. 9 Heatmap comparison

3.6 模型复杂度比较

为了全面评估本文所提出的模型,对模型参数量(Parameters)和浮点运算次数(FLOPs)在DVTOD数据集上进行了横向对比实验,结果如图10所示。实验选取CFT^[3],ICAFusion^[4],COMO^[10],CMA-Det^[12],CDC-YOLOFusion^[14],SuperYOLO^[21],DEYOLO^[22],JTM-Det^[23]以及MSTF^[24]等多种主流前沿融合检测算法作为对比基线。由实验结果可知,本文模型参数量低于CFT,ICAFusion,CMA-Det,DEYOLO,CDC-YOLOFusion,JTMDet,MSTF和COMO,仅略高于轻量化模型SuperYOLO。在计算复杂度方面,本文模型的FLOPs略高于CMA-Det模型,但检测精度相较于所有对比算法均实现显著提升。综合分析可得,本文模型在参数量轻量化、计算开销与检测精度之间实现了优质平衡。

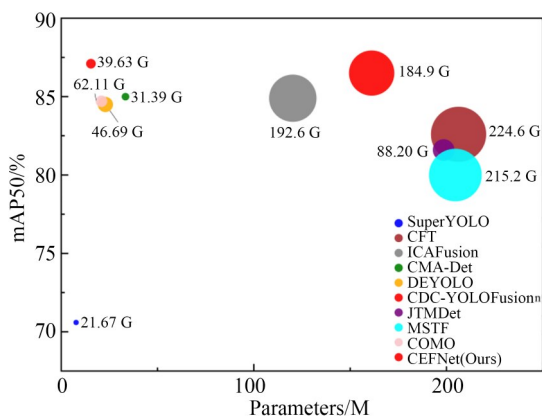


图10 DVTOD数据集上本文方法与其他模型在检测精度、模型参数大小和FLOPs方面的比较(圆圈越大,表示模型FLOPs越大)

Fig. 10 Comparison of our proposed method with other models on the DVTOD dataset in terms of detection accuracy, model parameter size, and FLOPs (A larger circle indicates more model FLOPs)

4 结 论

针对复杂现实场景下可见光-红外多模态目

标检测中普遍存在的空间错位、特征异构与背景干扰等严峻挑战,本文提出了一种免对齐渐进式多模态目标检测方法CEFNet。在特征补偿阶段,本文设计的特征补偿模块FCM利用离散小波变换对特征进行频域无损解耦,并将跨模态的高频差异转化为Mamba的选择性门控信号,主动阻断了局部错位噪声向全局扩散的路径,实现了高效的抗噪补偿。在特征增强与深度融合阶段,构建了特征增强模块FEM与特征融合模块FFM形成渐进式融合优化体系:FEM利用多分支非对称大核卷积,在极低参数量下获得了更大的感受野,深度强化了目标的形态特征;在此基础上,FFM模块通过差分放大自适应抑制了双模态间的共模冗余背景,并精准放大了极具判别力的差模互补特征,实现了跨模态语义特征的高效聚合。实验结果表明,本文方法在非对齐任务DVTOD数据集上取得了87.1%的mAP50和53.1%的mAP50-95,显著优于现有主流前沿检测算法。同时,在涵盖轻微偏移的M³FD数据集与对齐良好的LLVIP数据集的泛化性测试中,进一步验证了所提方法在检测稳定性与计算效率方面具备的综合优势。

尽管本文方法在现有多模态基准数据集上展现出了优异的性能,但在多模态泛化与极限抗干扰能力方面仍有广阔的探索空间。未来的研究工作将主要围绕以下两个方向展开:一是探索本文提出的方法在其他异构感知组合(如RGB-Depth、RGB-毫米波雷达、RGB-高光谱等)中的适用性;二是深化极限环境下的稳定性评估,如人工添加不同严重等级的噪声、运动模糊、图像压缩伪影以及极端尺度变化等,进一步探索并提升模型在复杂开放世界下的容错下限。

作者贡献声明:

周永根:算法模型提出,论文构思与撰写;

林坚普、吕珊红:实验指导,论文审阅;

林志贤:项目管理;

林珊玲:实验程序编写并修改,论文审核及修改;

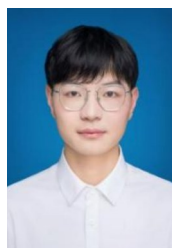
郭太良:实验与论文指导。

参考文献:

- [1] SWAMY S K, SCHWARZ K, HARTMANN M, *et al.* RGB and IR imagery fusion for autonomous driving [C]. *Multimodal Image Exploitation and Learning* 2023. April 30-May 5, 2023. Orlando, USA. SPIE, 2023: 30.
- [2] WEI X B, SONG K C, YANG W K, *et al.* A visible-infrared clothes-changing dataset for person re-identification in natural scene[J]. *Neurocomputing*, 2024, 569: 127110.
- [3] FANG Q Y, HAN D P, WANG Z K. Cross-modality fusion transformer for multispectral object detection [EB/OL]. 2021: *arXiv*: 2111.00273. <https://arxiv.org/abs/2111.00273>.
- [4] SHEN J F, CHEN Y F, LIU Y, *et al.* ICAFu-sion: Iterative cross-attention guided feature fusion for multispectral object detection[J]. *Pattern Recognition*, 2024, 145: 109913.
- [5] WU W C, ZHANG H X, ZHANG X W, *et al.* Lightweight modal-guided cross-attention fusion network for visible-infrared object detection [J]. *Pattern Recognition*, 2026, 177: 113350.
- [6] GU A, DAO T. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. 2023: *arXiv*: 2312.00752. <https://arxiv.org/abs/2312.00752>.
- [7] DONG W H, ZHU H D, LIN S H, *et al.* Fusion-mamba for cross-modality object detection [J]. *IEEE Transactions on Multimedia*, 2025, 27: 7392-7406.
- [8] LIU Z N, ZHU C Y, LI Y C, *et al.* A novel implicit cross-attention framework for RGB-T object detection [J]. *Expert Systems with Applications*, 2026, 314: 131570.
- [9] YUAN M X, WEI X X. C²Former: calibrated and complementary transformer for RGB-infrared object detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5403712.
- [10] LIU C, MA X, YANG X C, *et al.* COMO: Cross-mamba interaction and offset-guided fusion for multimodal object detection [J]. *Information Fusion*, 2026, 125: 103414.
- [11] QIAN J C, QIAO B Y, ZHANG Y K, *et al.* DACFusion: Dual Asymmetric Cross-Attention guided feature fusion for multispectral object detection[J]. *Neurocomputing*, 2025, 635: 129913.
- [12] SONG K C, XUE X T, WEN H W, *et al.* Mis-aligned visible-thermal object detection: a drone-based benchmark and baseline[J]. *IEEE Transactions on Intelligent Vehicles*, 2024, 9(11): 7449-7460.
- [13] ZENG Y Q, LIANG T F, JIN Y, *et al.* MMI-det: exploring multi-modal integration for visible and infrared object detection [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(11): 11198-11213.
- [14] WANG Z A, LIAO X H, YUAN J, *et al.* CDC-YOLO Fusion: leveraging cross-scale dynamic convolution fusion for visible-infrared object detection [J]. *IEEE Transactions on Intelligent Vehicles*, 2025, 10(3): 2080-2093.
- [15] 刘远航, 侯志强, 李新月, 等. 基于YOLOv8s的可见光和红外双模态融合目标检测算法[J]. *光子学报*, 2026, 55(4): 0410005.
LIU Y H, HOU ZH Q, LI X Y, *et al.* Visible and infrared dual-modal fusion object detection algorithm based on YOLOv8s [J]. *Acta Photonica Sinica*, 2026, 55(4): 0410005. (in Chinese)
- [16] 代小波, 任思羽, 何小海, 等. 基于特征增强和对齐融合的可见光-红外图像融合目标检测[J]. *工程科学学报*, 2025, 47(12): 2554-2565.
DAI X B, REN S Y, HE X H, *et al.* Visible-infrared fusion object detection based on feature enhancement and alignment fusion[J]. *Chinese Journal of Engineering*, 2025, 47(12): 2554-2565. (in Chinese)
- [17] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements [EB/OL]. 2024: *arXiv*: 2410.17725. <https://arxiv.org/abs/2410.17725>.
- [18] LIU Y, TIAN Y J, ZHAO Y Z, *et al.* VMamba: visual state space model[C]. *Advances in Neural Information Processing Systems* 37. December 10-15, 2024. Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 103031-103063.
- [19] LIU J Y, FAN X, HUANG Z B, *et al.* Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022: 5792-5801.

- [20] JIA X Y, ZHU C, LI M Z, *et al.* LLVIP: a visible-infrared paired dataset for low-light vision[C]. 2021 *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. October 11-17, 2021. Montreal, BC, Canada. IEEE, 2021: 3489-3497.
- [21] ZHANG J Q, LEI J, XIE W Y, *et al.* SuperY-OLO: super resolution assisted object detection in multimodal remote sensing imagery [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5605415.
- [22] CHEN Y S, WANG B R, GUO X Y, *et al.* DEYOLO: dual-feature-enhancement YOLO for Cross-modality object detection[C]. *Pattern Recognition*. Cham: Springer, 2025: 236-252.
- [23] LI C, PENG X M. Joint Transformer and Mamba fusion for multispectral object detection[J]. *Image and Vision Computing*, 2025, 156: 105468.
- [24] HOU S, YANG M, ZHENG W S, *et al.* Multi-Spectral Transformer Fusion *via* exploiting similarity and complementarity for robust pedestrian detection[J]. *Pattern Recognition*, 2025, 162: 111383.
- [25] HU J, NI L, PENG B, *et al.* WtCAFNet: a wavelet transform and cross-attention modality-adaptive fusion network for multispectral object detection [J]. *Signal Processing*, 2026, 242: 110446.
- [26] 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.
- [27] LIANG M J, HU J J, BAO C Y, *et al.* Explicit attention-enhanced fusion for RGB-thermal perception tasks[J]. *IEEE Robotics and Automation Letters*, 2023, 8(7): 4060-4067.
- [28] DEEVI S A, LEE C, GAN L, *et al.* RGB-X object detection *via* scene-specific fusion modules [C]. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. January 3-8, 2024, Waikoloa, HI, USA. IEEE, 2024: 7351-7360.
- [29] GUO J J, GAO C Q, LIU F C, *et al.* DAMS-Det: dynamic adaptive multispectral detection transformer with Competitive query selection and Adaptive feature fusion[C]. *Computer Vision-EC-CV 2024*. Cham: Springer, 2025: 464-481.
- [30] ZHANG S L, WANG X J, WANG J Q, *et al.* Dense distinct query for end-to-end object detection [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 17-24, 2023, Vancouver, BC, Canada. IEEE, 2023: 7329-7338.
- [31] FANG Q Y, WANG Z K. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery[J]. *Pattern Recognition*, 2022, 130: 108786.
- [32] ZHANG Y, YU H, HE Y J, *et al.* Illumination-guided RGBT object detection with inter- and intra-modality fusion[J]. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 2508013.
- [33] ZHAO T Y, YUAN M X, WEI X X. Removal and selection: Improving RGB-infrared object detection *via* coarse-to-fine fusion [EB/OL]. 2024: *arXiv*: 2401.10731. <https://arxiv.org/abs/2401.10731>.
- [34] CAO Y, BIN J C, HAMARI J, *et al.* Multimodal object detection by channel switching and spatial attention [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 17-24, 2023, Vancouver, BC, Canada. IEEE, 2023: 403-411.

作者简介:



周永根(2001—),男,湖南湘潭人,硕士研究生,主要从事计算机视觉及多模态图像目标检测方面的研究。E-mail:zyg185551373@163.com

通讯作者:



林珊玲(1991—),女,福建泉州人,博士,副教授,硕士生导师,2014年、2020年于福州大学分别获得学士和博士学位,主要从事显示驱动和图像处理等方面的研究。E-mail:sllin@fzu.edu.cn