

文章编号: 1002-1582(2026)02-0236-09

基于分层融合与推理一致性的RGB-T显著目标检测方法

杨璐霞^{1,2*}, 刘昊杰^{1,2}, 张红瑞^{1,2}, 马永杰³

(1. 太原师范学院 计算机科学与技术学院, 山西 晋中 030619)

(2. 智能优化计算与区块链技术山西省重点实验室, 山西 晋中 030619)

(3. 西北师范大学 物理与电子工程学院, 甘肃 兰州 730070)

摘 要: 针对跨模态显著目标检测中常忽视层级差异与边界恢复精度不足的问题, 提出一种基于分层融合与推理一致性的网络 (HFCR-Net)。该网络以 Swin Transformer 双分支编码器提取多尺度特征; 然后, 设计双门控融合模块 (DGF), 通过像素级空间门控与通道对齐机制实现深层特征的自适应加权融合, 在中低层构建跨模态细化模块 (CMR) 通过交互式引导与可学习的注意力调控机制细化边缘与纹理; 同时设计了推理一致性策略以提升模型的稳定性。实验结果表明, HFCR-Net 在 VT821 数据集上 F_β 、 $S\alpha$ 、 $E\epsilon$ 与 MAE 分别达到了 0.942、0.977、0.985 和 0.010; 在更具挑战性的 VT1000 与 VT5000 数据集上 F_β 分别达到了 0.927 与 0.856, 进一步验证了方法在不同复杂场景下的有效性与鲁棒性。

关 键 词: 显著目标检测; 双门控融合; 跨模态交互

中图分类号: TP394.1; TH691.9

文献标识码: A

Hierarchical fusion with consistent reasoning for RGB-T saliency object detection

YANG Luxia^{1,2*}, LIU Haojie^{1,2}, ZHANG Hongrui^{1,2}, MA Yongjie³

(1. School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China)

(2. Shanxi Key Laboratory of Intelligent Optimization Computing & Blockchain Technology, Jinzhong 030619, China)

(3. College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China)

Abstract: To address the common issues of overlooked hierarchical differences and insufficient boundary restoration accuracy in cross modal salient object detection, a network based on hierarchical fusion and reasoning consistency (HFCR-Net) is proposed. The network employs a Swin Transformer dual branch encoder to extract multi-scale features. Subsequently, a Dual Gated Fusion module (DGF) is designed to achieve adaptive weighted fusion of deep features through pixel-level spatial gating and channel alignment mechanisms. At the middle and low levels, a Cross Modal Refinement module (CMR) refines edges and textures through interactive guidance and a learnable attention regulation mechanism. In addition, a reasoning-consistent inference strategy is introduced to improve model stability. Experimental results demonstrate that HFCR-Net achieves F_β , $S\alpha$, $E\epsilon$ and MAE of 0.942, 0.977, 0.985 and 0.010 on VT821, and attains F_β of 0.927 and 0.856 on the more challenging VT1000 and VT5000, respectively, further validating its effectiveness and robustness in diverse complex scenarios.

Keywords: salient object detection; dual gated fusion; cross modal interaction

收稿日期: 2025-12-25; 收到修改稿日期: 2026-02-10

基金项目: 国家自然科学基金项目 (62066041); 山西省重点研发计划 (202102010101008); 山西省高等学校科技创新项目 (2024L295); 山西省科技战略研究专项重点项目 (202304031401011); 山西省基础研究计划项目 (自由探索类) (202403021222276, 202503021212249); 山西省研究生实践创新项目 (2025SJ412); 山西省研究生教育改革研究课题 (2023JG163)

作者简介: 杨璐霞 (1979—), 女, 教授, 博士, 从事智能信息处理方面的研究。

* 通讯作者: luxia yang@tynu.edu.cn

0 引言

显著目标检测 (salient object detection, SOD) 旨在模拟人类视觉从复杂场景中准确地检测和分割最具吸引力的物体, 在图像检索^[1]、对象识别^[2]、多模态匹配^[3]、视觉跟踪^[4]等计算机视觉任务中发挥着重要作用。

近年来, 显著目标检测已经在RGB领域成果众多, 但是在光照剧烈变化、目标与背景表现相似等复杂场景下, 仅依赖可见光信息往往难以实现鲁棒检测。为此, 融合热红外 (Thermal) 模态的RGB-T显著目标检测应运而生。该方法通过有效互补可见光丰富的纹理细节与热红外对热辐射的敏感性优势, 为提升复杂环境下检测模型的鲁棒性与准确性, 提供了新的解决路径。

在此背景下, 基于Transformer的RGB-T SOD^[5-9]已经取得了显著进展, 超越了传统的卷积神经网络方法^[10-15]。杨超等^[5]引入Transformer架构以增强全局信息建模, 提出了包含降噪、语义指导及跨模态交互的显著性增强框架, 有效缓解递归式聚合带来的低泛化问题。魏明军等^[6]利用Transformer的全局上下文建模能力, 在其提出的CALNet中通过跨层级注意力模块、全局信息建模与多交互解码策略有效利用多模态间的跨层级互补信息, 提升了复杂场景下显著目标的性能。Lv等^[7]基于Swin Transformer网络, 设计了跨模态特征融合模块和交互式特征解码模块, 有效建模RGB与热红外模态间的长距离依赖关系, 显著增强了特征融合效果。Liu^[8]等提出SwinNet模型, 通过空间对齐与通道重校准模块缩小了RGB和热成像模态差异, 并引入边缘引导的解码器补充边界细节, 提高了模型的泛化能力。Tang^[9]提出了HRFormer, 通过设计双流向短连接融合模块优化了模型的输出, 提升了模型在复杂场景下的判别力。

然而, 这些方法在跨模态融合与跨层级协同方面仍存在明显不足, 主要体现在以下两点: 首先, 多数方法在不同语义层级上采用相同的融合模块或统一的融合策略, 未能充分考虑高层语义与中低层细节在融合形式与信息侧重点上的本质差异, 致使跨层级互补不足、模态交互粗糙, 进而影响模型的判别能力。其次, 在模型训练与推理阶段, 现有方法在推理阶段普遍缺乏针对性的优化与校准手段。这导致模型在检测精度和泛化能力上存在明显的局限性。

针对上述问题, 本文提出了一种基于分层融合与推理一致性的RGB-T显著目标检测方法 (Hi-

erarchical Fusion with Consistent Reasoning Network, HFCR-Net), 通过层次化的融合机制, 旨在获得更好的RGB-T SOD性能。首先, 通过双分支的Swin Transformer对RGB图像和热成像图像进行双流特征编码, 并使用统一的卷积层增强感受野。然后, 网络采用自顶向下逐层传播, 其中高层语义阶段采用双门融合模块 (Dual Gate Fusion, DGF) 完成跨模态融合, 并通过解码器将上采样特征与当前阶段对齐; 中低层阶段采用跨模态细化模块 (Cross Modal Refinement, CMR) 进一步细化并增强结构信息。最后, 通过设计包含复合损失函数、TTA与Platt校准以及热红外预训练映射机制的推理一致性策略, 以提高模型跨模态特征提取的稳定性与泛化性。

1 总体网络架构

本文设计的基于分层融合与推理一致性的RGB-T显著目标检测方法 (HFCR-Net) 采用双分支编码-解码架构, 如图1所示。在编码阶段, 采用基于Swin Transformer的双分支编码器分别对可见光与热红外图像进行4个阶段的特征提取, 用于捕获多尺度特征之间的长距离依赖关系, 并使用统一的卷积层扩大感受野。在解码阶段, 对于不同层次的特征采用差异化处理机制, 在高层语义阶段采用双门控融合模块 (DGF) 完成跨模态信息融合, 在中低层阶段采用跨模态细化模块 (CMR) 进一步整合与细化多尺度线索。在优化与推理策略方面, 设计了协同平衡像素级准确性与全局结构一致性的复合损失函数, 采用热红外预训练映射机制稳定热分支训练过程, 并结合测试时增强 (Test Time Augmentation, TTA) 与Platt概率校准策略提升模型的跨域泛化能力与预测置信度, 最终生成显著性预测。

1.1 基于SwinTransformer的编码网络

在机器视觉任务中, 卷积神经网络 (CNN) 因其局部感受野的设计, 在建模长距离依赖关系上存在固有局限。而Transformer凭借自注意力机制, 在自然语言处理中展现出强大的全局上下文建模能力。Swin Transformer作为基于Transformer架构的代表性深度学习模型, 通过引入分层结构和移位窗口机制, 在保持全局建模能力的同时实现高效计算, 为图像识别任务提供了更优的解决方案。

基于上述分析, 本文采用Swin Transformer作为网络骨架, 利用其分层架构所生成的多尺度特征图进行特征交互与融合。具体处理流程如下:

$$A_R = \sigma(\text{Conv}_{3 \times 3}(\text{Max}_c(R))) \quad (3)$$

$$A_T = \sigma(\text{Conv}_{3 \times 3}(\text{Max}_c(T))) \quad (4)$$

$$S = \text{Concat}(A_R, A_T, U(F'_{i+1})) \quad (5)$$

式中, $\text{Max}_c(\cdot)$ 表示通道维度的最大池化操作; $\text{Conv}_{3 \times 3}$ 表示一个 3×3 的标准卷积层; σ 表示 Sigmoid 激活函数。Concat 表示通道维度拼接。

最后, 通过两条并行的卷积路径生成空间门控权重 g_R 和 g_T , 并应用于原始特征 E_i^{rgb}, E_i^t ($i=4,5$), 实现逐像素门控加权, 得到融合后的特征 R' 和 T' 。该加权操作专为深层特征设计, 实现了像素级的自适应融合。该模型依据联合注意力表征 S 所编码的上下文信息动态选择并增强特征响应, 有效抑制了深层特征中的模态噪声, 增强了其语义判别力。最终, R' 和 T' 在通道维度拼接并通过 3×3 卷积降维至原始通道数 C , 得到融合特征 F_n^f , 具体公式如下

$$g_R = \sigma(\text{ReLU}(\text{BN}(\text{Conv}(S)))) \quad (6)$$

$$g_T = \sigma(\text{ReLU}(\text{BN}(\text{Conv}(S)))) \quad (7)$$

$$R' = g_R \odot E_i^{rgb}, T' = g_T \odot E_i^t \quad (8)$$

$$F_n^f = \text{Concat}(\text{Conv}_{3 \times 3}(R'), \text{Conv}_{3 \times 3}(T')) \quad (9)$$

式中, BN 表示批归一化操作, ReLU 表示线性整流激活函数, σ 表示 Sigmoid 激活函数, $\odot$ 表示逐元素相乘, $\text{Conv}_{3 \times 3}$ 表示一个 3×3 的标准卷积层。

1.2.2 跨模态细化模块

中低层特征包含丰富的目标边缘、纹理等空间细节信息, 对恢复显著目标的清晰轮廓至关重要。然而, 在低层特征的融合过程中, 由于不同模态在边界与纹理区域存在空间错位与细节响应冲突, 若直接融合, 易导致细节信息相互抵消或模糊, 进而削弱模型对目标轮廓的精准定位能力。因此, 本文提出一种跨模态细化模块 (CMR), 在实现细节信息的互补与融合, 生成细节丰富的特征表达, 具体如图3所示。

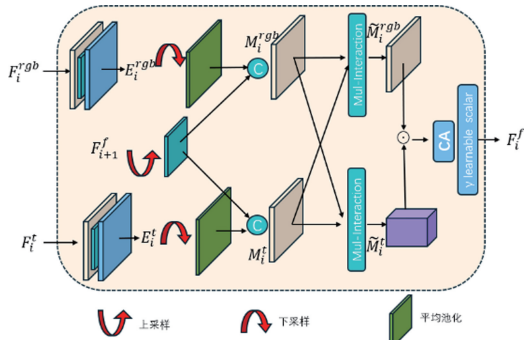


图3 CMR模块架构

首先, 针对富含细节的中低层特征, 对双模态特征 F_i^{rgb} 与 F_i^t 执行与 DGF 模块相同的串联卷积操作、批归一化及 ReLU 激活预处理流程, 生成通道对齐且感受野扩大的特征 E_i^{rgb} 与 E_i^t 。然后, 为消除通道与尺度差异, 双模态特征通过 $\text{Down} \times 2$ (即 $2 \times$ 平均池化) 下采样以匹配上一级融合特征, 实现跨层级上下文对当前层细节的引导与补充。

同时, 上一级富含高级语义的融合特征通过 $\text{Up} \times 2$ (双线性上采样), 使其空间分辨率与当前层级特征对齐, 分别得到初步融合特征表达 M_i^{rgb} 与 M_i^t 。随后, 为了增强中低层特征的细节与结构, M_i^{rgb} 与 M_i^t 输入 CMR 的并行乘性/加性路径以及空间注意力与通道注意力进一步交互与增强, 实现对初步融合特征的精细化增强。

最后, 为了实现对注意力权重的控制, 引入一个可学习缩放因子 γ , 用于调节通道注意力输出的幅度, 并通过残差连接确保中低层原始细节信息得以保留。上述流程可表示为

$$\begin{cases} \tilde{E}_i^{rgb} = \text{Down} \times 2(E_i^{rgb}) \\ \tilde{E}_i^t = \text{Down} \times 2(E_i^t) \\ \tilde{F}_{i+1}^f = \text{UP} \times 2(F_{i+1}^f) \end{cases} \quad (10)$$

$$M_i^{rgb} = \text{Conv}(\text{Concat}(\tilde{E}_i^{rgb}, \tilde{F}_{i+1}^f)) \quad (11)$$

$$M_i^t = \text{Conv}(\text{Concat}(\tilde{E}_i^t, \tilde{F}_{i+1}^f)) \quad (12)$$

$$\tilde{M}_i^{rgb} = \text{Mul}(M_i^{rgb}, M_i^t) \quad (13)$$

$$\tilde{M}_i^t = \text{Mul}(M_i^{rgb}, M_i^t) \quad (14)$$

$$F_i^f = \gamma \cdot \text{CA}(\text{Conv}(\tilde{M}_i^{rgb} \odot \text{SA}(\tilde{M}_i^t))) \quad (15)$$

式中, $\text{Down} \times 2(\cdot)$ 为 $2 \times$ 表示步长为 2 的平均池化, $\text{UP} \times 2$ 表示 2 倍双线性上采样, $\text{Mul}(\cdot)$ 为表示乘法交互操作。

1.3 优化与推理策略

针对当前 RGB-T 显著目标检测在训练过程中存在全局特征与局部细节优化失衡、推理阶段预测置信度偏移导致泛化能力受限, 以及热红外分支因梯度不稳定易引发过拟合等问题, 本文提出了具有推理一致性的优化与推理策略。该策略从梯度优化、概率校准与模态迁移三个层面进行设计, 旨在同步提升模型的检测精度、结构一致性及跨域泛化能力, 具体实现细节如下:

(1) 复合损失函数

为解决单一损失函数难以同时兼顾像素级准确性与全局显著性结构一致性的问题, 本文提出了一种融合二元交叉熵 (BCE) 与 F-measure 的复合显著性损失函数。其中, BCE 损失聚焦像素级

分类准确性, F-measure 损失通过梯度近似实现对全局 F 值的直接优化。该损失函数表达式为

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_F \quad (16)$$

式中, $\mathcal{L}_{\text{total}}$ 为复合显著性损失, 由 $\mathcal{L}_{\text{BCE}}$ 与 $\mathcal{L}_F$ 加权构成, λ_1 与 λ_2 为平衡系数。

本文在 VT5000 的训练集上划分一部分样本作为验证集, 并在 $\lambda_1, \lambda_2 \{0.5, 1, 2\}$ 的组合上进行网格搜索, 并以验证集主要性能指标为准, 同时结合训练过程的收敛稳定性进行选择。最终, 选取 $\lambda_1=1$ 、 $\lambda_2=1$ 作为默认设置。

(2) TTA 与 Platt 校准

为缓解模型在跨数据中出现的置信度偏移问题, 本文提出测试时增强 (Test Time Augmentation, TTA) 与 Platt 概率校准相结合的推理策略。

在 TTA 实施中, 通过对输入执行水平翻转、尺度缩放等多种几何变换, 生成多个测试样本并进行预测, 最后对预测结果进行平均融合, 从而提升预测的稳定性与鲁棒性。Platt 校准则基于训练集验证结果, 对模型输出的 logits 进行 Sigmoid 回归调整, 使预测概率更符合真实标签分布。

(3) 预训练权重的热红外映射机制

由于热红外特征提取初期的梯度不稳定, 直接训练易导致模型过拟合。为充分利用 RGB 预训练模型的迁移能力, 本文对热红外分支的输入通道进行了权重初始化。该机制在加载 SwinTransformer 的 ImageNet 预训练权重时, 针对热成像分支的输入通道自动执行通道映射与权重均衡初始化, 实现过程如下

$$W_T = \frac{1}{3}(W_R + W_G + W_B) \quad (17)$$

式中, W_T 表示热红外输入层权重, W_R, W_G, W_B 分别为 RGB 通道的预训练权重。

2 实验细节

2.1 数据集

为了验证所提方法的有效性, 本文在 VT821、VT1000、VT5000 数据集上进行实验分析。其中, VT821 包含 821 个手动注册的图像对。VT1000 包含 1000 个 RGB-T 图像对, 这些图像由高度对齐的 RGB 和热像仪捕获。VT5000 包含 5000 对 RGB-T 图像对, 并对样本难度属性进行了标注, 覆盖大目标、小目标、多目标、低照度、中心偏置等 11 类典型挑战场景。根据其属性统计结果, 在单一挑战属性样本中, 背景杂乱为 1096 对、热交叉为 858 对、低照度为 535 对、小目标为 340 对、多目标为 491 对, 表明该数据集在上述典型挑战场景上

具有较高覆盖比例, 能够反映复杂条件下的检测难度。

在 3 个通用数据集中, 所采集的图像由不同光照亮度下的各种场景、物体组成, 且包含小目标、动态场景、热交叉、添加噪声等特殊场景。为了公平比较, 使用 VT5000 中 2500 对图像组成的训练数据集进行训练, VT1000、VT821、VT5000 中的其余图像用于测试评估。

2.2 实验环境

本文设计的模型在配备 NVIDIA RTX 4090 (24 GB) GPU 的计算机上实现。训练阶段采用 AdamW 优化器, 初始学习率设为 2×10^{-5} , 训练 epoch 的数量为 50, 批量大小设置为 2。损失函数采用本文提出的复合显著性损失函数 $\mathcal{L}_{\text{total}}$, 其中平衡系数取 $\lambda_1=1$ 、 $\lambda_2=1$ 。推理阶段采用 TTA 与 Platt 校准策略。对于 RGB-T 图像对的输入, 本文将所有的图像像素调整为 384×384 。

2.3 评估指标

本文采用的评估指标分别为加权 F 度量 (F_β)、结构相似性度量 (S_α)、增强对齐度量 ($E\epsilon$) 和平均绝对误差 (MAE)。

S_α 可评估显著性图与真值之间的区域感知和对象感知结构相似性, 公式如下

$$S_\alpha = \alpha \times S_0 + (1 - \alpha) \times S_r \quad (18)$$

式中, S_r 为区域感知的结构相似性, S_0 为对象感知的结构相似性, α 设置为 0.5。

F_β 是精确率 (Precision) 和召回率 (Recall) 的加权调和平均值, 公式表示如下

$$\text{Precision} = \frac{TP}{TP + FP} \quad (19)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (20)$$

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (21)$$

式中: TP 、 FP 和 FN 表示真阳性、假阳性和假阴性, 并且 β^2 被设置为 0.3。

$E\epsilon$ 用于融合全局统计信息和局部像素匹配信息, 公式如下

$$E\epsilon = \frac{1}{N} \sum_{i=1}^N \frac{2 \cdot P_i \cdot G_i}{P_i^2 + G_i^2} \quad (22)$$

式中, N 为像素数量, P_i 为预测显著性值, G_i 为真实值。

MAE 用于测量显著性图 and 真值图之间每像素绝对差的平均值, 公式表示为

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |P_i - G_i| \quad (23)$$

3 实验结果与分析

3.1 定量对比

为全面验证本模型 HFCR-Net 的有效性, 与现有的 RGB-T SOD 算法进行对比分析, 结果如表 1 所示。其中, 表 1 所列性能指标均在启用 TTA

与 Platt 校准策略下测得, 以保证结果的稳定性与可比性。对比方法包括: MIDD^[14]、SwinNet^[8]、HRTransNet^[9]、CSRNet^[16]、MCNet^[17]、OSRNet^[18]、FAWT^[19]、LSNet^[20]、TANet^[21]、ADFNet^[22]、MGAI^[23]、CAVER^[24]、CAFCNet^[25], 最优值以加粗格式标出。

表 1 不同方法在不同数据集上的定量对比

数据集方法	VT821				VT1000				VT5000			
	加权 F 度量 F_{β}	结构相似性度量 S_{α}	增强对齐度量 E_{ϵ}	平均绝对误差 MAE	加权 F 度量 F_{β}	结构相似性度量 S_{α}	增强对齐度量 E_{ϵ}	平均绝对误差 MAE	加权 F 度量 F_{β}	结构相似性度量 S_{α}	增强对齐度量 E_{ϵ}	平均绝对误差 MAE
CSRNet ^[16]	0.918	0.878	0.877	0.024	0.868	0.797	0.811	0.042	0.83	0.821	0.83	0.038
MIDD ^[14]	0.915	0.856	0.822	0.027	0.868	0.763	0.801	0.043	0.76	0.804	0.76	0.045
OSRNet ^[18]	0.926	0.891	0.892	0.022	0.875	0.807	0.823	0.04	0.801	0.813	0.801	0.043
FAWT ^[19]	0.802	0.883	0.907	0.033	0.863	0.933	0.937	0.021	0.807	0.884	0.91	0.036
LSNet ^[20]	0.865	0.912	0.921	0.028	0.775	0.849	0.883	0.049	0.706	0.81	0.835	0.07
TANet ^[21]	0.838	0.902	0.912	0.03	0.754	0.847	0.864	0.047	0.717	0.81	0.852	0.052
ADFNet ^[22]	0.847	0.91	0.921	0.034	0.787	0.864	0.891	0.048	0.717	0.81	0.81	0.077
MGAI ^[23]	0.929	0.893	0.885	0.021	0.883	0.815	0.824	0.034	0.824	0.829	0.824	0.031
SwinNet ^[8]	0.896	0.938	0.947	0.018	0.865	0.912	0.942	0.026	0.847	0.904	0.926	0.03
CAVER ^[24]	0.853	0.898	0.934	0.027	0.904	0.938	0.965	0.017	0.854	0.9	0.94	0.029
HRTransNet ^[9]	0.851	0.906	0.933	0.026	0.898	0.938	0.956	0.017	0.869	0.913	0.948	0.025
CAFCNet ^[25]	0.856	0.891	0.927	0.028	0.913	0.935	0.951	0.017	0.858	0.9	0.939	0.027
MCNet ^[17]	0.876	0.923	0.945	0.021	0.918	0.949	0.971	0.014	0.89	0.925	0.957	0.021
HFCR-Net(Ours)	0.942	0.977	0.985	0.01	0.927	0.968	0.976	0.014	0.856	0.914	0.932	0.033

从表 1 可以看出, HFCR-Net 在多个数据集及评估指标上均达到了较好的效果。

在 VT821 数据集上, 本文模型的 F_{β} 达到了 0.942, 相较于该数据集上 F_{β} 次优的 MGAI 提升了 1.4%。此外, S_{α} 和 E_{ϵ} 也分别达到了 0.977 和 0.985, 与 SwinNet 相比分别提高了 4.16% 与 4.01%。

在 VT1000 数据集上, HFCR-Net 的 F_{β} 值达到 0.927, 优于表现最佳的 MCNet, 性能提升约 0.98%。并且提出的方法在 S_{α} 和 E_{ϵ} 上分别达到了 0.968 与 0.976, 与 MCNet 相比分别提高了 2.00%、0.52%。

在更具挑战性的 VT5000 数据集上, 本文方法的 F_{β} 值为 0.856, MAE 值为 0.033, 较最优方法存在小幅差距。存在差距的原因主要有以下两点: 其一, 在复杂背景干扰下, 模型对于热红外目标与背景之间的可分性降低, 显著区域与非显著区域的响应重叠增多, 从而更容易产生局部误检; 其二, 在多目标与背景纹理丰富的场景中, 中低层跨模态细化模块 CMR 在增强边缘与纹理细节的同时, 对部分高频背景纹理产生响应, 导致精确率略有下降。但是, 本文方法在结构保真度指标 S_{α} 上仍达到了 0.914, 表明所提出的分层融合与细化机制在目标形状与边界恢复方面仍具有优势。

3.2 定性分析

为了直观验证 HFCR-Net 在复杂场景下的泛化

性与鲁棒性, 图 4 给出了与多种现有方法在 VT5000 数据集上的可视化对比结果。

从图 4 可以看出, 在复杂结构目标 (第 1 行) 与细粒度目标检测 (第 2、6 行) 场景中, CSRNet、MGAI 等方法在捕捉镂空结构或细小物体时容易出现结构断裂或漏检。

面对存在动态模糊的图像 (第 2 行) 与多目标邻近干扰 (第 3 行), 部分方法如 OSRNet、ADFNet 等因模态间对齐偏差或感受野有限, 导致目标定位模糊或彼此粘连。本文提出的方法通过 DGF 模块在高层语义层面实现 RGB 与热红外的自适应融合。该方法充分利用热辐射信息的空间稳定性, 有效抑制了模糊干扰。在此基础上, 模型在邻近目标间建立了清晰的边界。

在弱纹理低对比度 (第 4 行) 与单/双模态退化 (第 5~6 行) 等困难场景中, 常见模型如 LSNet、CAFCNet 等易受图像质量下降影响, 出现显著区域提取不全或边界混淆问题。本文方法通过双分支编码-解码架构分别提取并充分利用两种模态的优势特征。当某一模态信息因退化而不可靠时, 模型能够自适应地依赖另一模态的有效信息进行补偿, 从而实现稳定与准确的检测结果。

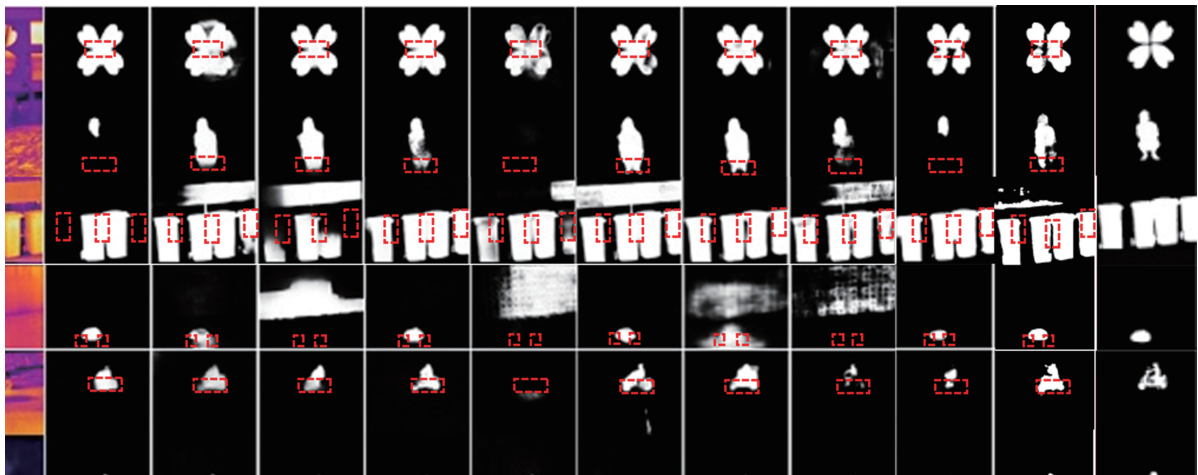


图 4 实验结果定性分析

3.3 消融实验

3.3.1 模块有效性消融

为了系统评估所提出模块的有效性，本文进行了消融实验分析。其中，基线使用 Swin Transformer 作为编码器的基础双流架构，通过逐步移除不同模块的方式验证其贡献。

消融实验结果如表 2 所示，各项设置均在完整模型的基础上单独移除对应模块或训练项进行对比。相较于完整模型，当移除 DGF 模块后， F_{β} 值从 0.851 下降至 0.839，MAE 从 0.053 上升至 0.055，表明 DGF 模块能有效融合高层语义信息。移除 CMR 模块后， E_{ϵ} 指标从 0.913 下降至 0.906，体现出模块在增强结构感知与边界细节方面的作用。移除 F-measure 损失项 $\mathcal{L}_F$ （仅采用 $\mathcal{L}_{BCE}$ 监督），增强对齐度量 S_{α} 从 0.899 下降至 0.894。移除优化与推理策略后，整体性能进一步退化，验证其对模型稳定性与泛化能力的促进作用。

表 2 消融实验定量对比

评估指标名称	加权 F 度量 F_{β}	结构相似性度量 S_{α}	增强对齐度量 E_{ϵ}	平均绝对误差 MAE
W/O DGF	0.839	0.894	0.904	0.055
W/O CMR	0.842	0.896	0.906	0.054
W/O $\mathcal{L}_F$	0.838	0.894	0.902	0.055
W/O optimization strategy	0.833	0.89	0.898	0.056
本文	0.851	0.899	0.913	0.053

为直观验证各核心模块对模型预测效果的贡献，对消融实验结果进行可视化对比，结果如图 5 所示。

从图 5 可以看出，当移除双门控融合模块（W/O DGF）时，高层语义信息的自适应融合能力减弱，导致目标整体轮廓模糊，尤其在热辐射信息主导的区域出现误判。移除跨模态交互模块（W/O CMR）则削弱了细节补偿能力，使得目标边缘与细微结构变得粗糙。移除 F-measure 损失项 $\mathcal{L}_F$ （W/O $\mathcal{L}_F$ ），预测结果的结构

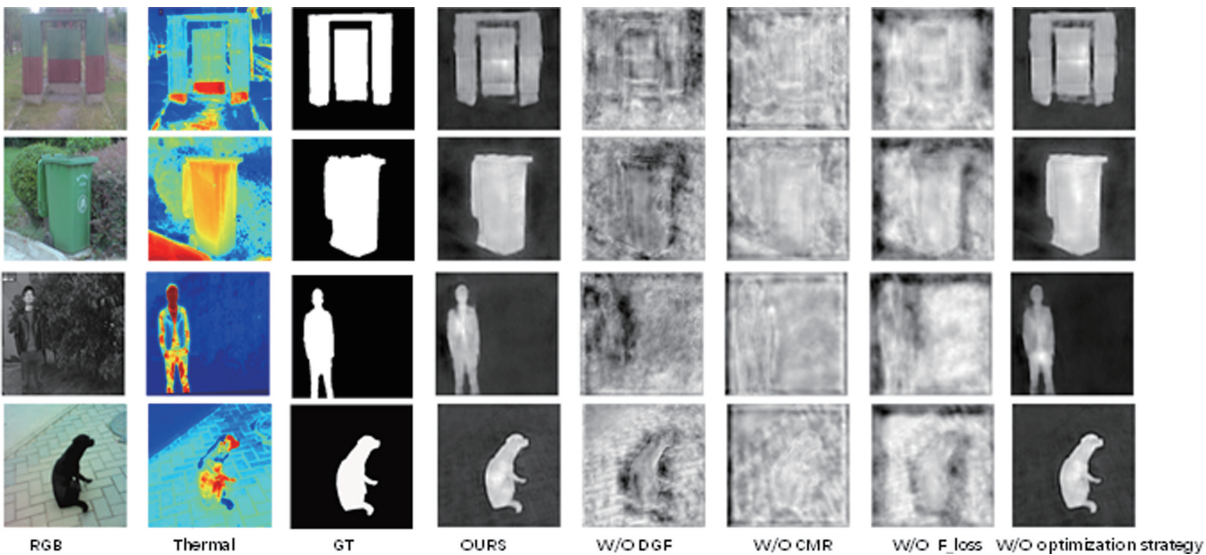


图 5 消融实验定性分析

保真度明显下降,与真实目标的形状一致性变差。当移除优化与推理策略(W/O optimization strategy)时,模型在部分场景中出现显著区域范围的局部外扩,进而导致预测边界模糊或发生轻微位置偏移;同时,预测结果中保留少量非目标区域残留,使得输出轮廓与真实边界的一致性降低。这些可视化对比充分证明了本文所提各模块在提升模型鲁棒性与检测精度方面的必要性。

3.3.2 收敛稳定性分析

为验证优化与推理策略对训练收敛稳定性的促进作用,本文进一步对比了引入该策略前后的复合显著性损失函数 $\mathcal{L}_{total}$ 收敛曲线。

如图 6 所示,曲线 A 为基线模型,曲线 B 为引入推理与优化策略后的模型。两组曲线均随着训练 epoch 逐渐降低,整体呈现正常的收敛趋势,但引入优化与推理策略后, $\mathcal{L}_{total}$ 在前期下降更为迅速,中后期的波动幅度更小,收敛阶段曲线更为平滑。相比之下,基线模型在收敛过程中仍存在更明显的局部起伏,表现出不稳定性。

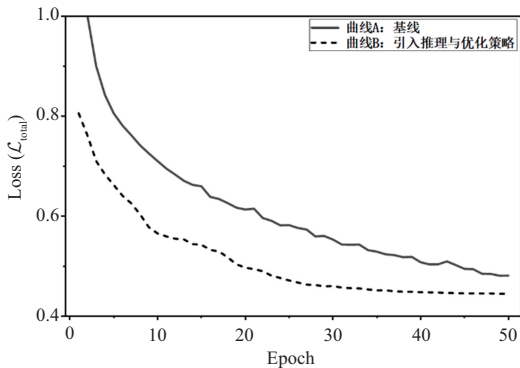


图 6 复合显著性损失函数 $\mathcal{L}_{total}$ 随训练轮次变化曲线对比

上述结果说明,优化与推理策略能够缓解训练过程中的波动,提升模型优化过程的稳定性,从而为后续性能提升提供更可靠的收敛基础。

3.4 复杂度分析

为了客观评估模型的实用性与计算效率,本文采用参数量(Params)、浮点运算数(FLOPs)和每秒帧数(FPS)进行了复杂度分析,结果如表 3 所示。

表 3 模型复杂度与效率分析

模型	参数量 Params /M	浮点运算数 FLOPs/G	帧数/FPS
MGAI	1.86	13.4	46.8
SwinNet	1.92	13.6	47
CAFCNet	1.86	13.1	46.2
HFCR-Net	1.61	12.9	47.8

在参数量与计算复杂度方面,HFCR-Net 仅需 1.61M 参数和 12.9G FLOPs,与参数量均在 1.8 M 左右的 SwinNet、MGAI 及 CAFCNet 等方法相比,

计算开销显著降低。在推理速度方面,47.8 FPS 是在关闭 TTA 条件下测得的单次前向推理速度,用于反映模型的基础实时推理能力。

4 结 论

本文针对 RGB-T 显著目标检测中跨模态融合的层级差异与边界恢复问题,提出了一种分层融合与推理一致性网络(HFCR-Net)。通过设计双门控融合模块(DGF)与跨模态细化模块(CMR),实现了深层语义与浅层细节的差异化融合。此外,结合包含复合损失函数、TTA 与 Platt 校准等推理一致性策略显著提升了模型的收敛稳定性、结构一致性与跨域泛化能力。在三个基准数据集上的实验结果表明,该方法在保持较低计算复杂度的同时,达到了较优的检测性能。但是,本文方法在极端光照等复杂背景的显著目标检测中仍有优化空间,未来工作将着重探索针对此类复杂场景的更强健的特征表示与融合机制。

参考文献:

[1] Li Z X, Xie X M, Ling F, et al. Matching images and texts with multi-head attention network for cross-media hashing retrieval[J]. Engineering Applications of Artificial Intelligence, 2021, 106: 104475.

[2] Zhou W J, Liu J F, Lei J S, et al. GMNet: graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 7790—7802.

[3] Zhuge M, Gao D H, Fan D P, et al. Kaleido-BERT: Vision-language pre-training on fashion domain[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA: IEEE, 2021: 12642—12652.

[4] Borji A, Frintrop S, Sihite D N, et al. Adaptive object tracking by learning background context[C]//Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, RI, USA: IEEE, 2012: 23—30.

[5] 杨超,官铮,王学,等.显著性增强和全局感知的 RGB-T 显著性目标检测[J].计算机辅助设计与图形学学报,2025,37(08): 1359—1370.

Yang Chao, Guan Zheng, Wang Xue, et al. Saliency enhancement and global perception for RGB-T salient object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2025, 37(08): 1359—1370.

[6] 魏明军,魏帅,刘亚志,等.基于跨层级注意力学习的 RGB-T 显著目标检测[J].郑州大学学报(理学版),2025,57(03): 42—48.

Wei Mingjun, Wei Shuai, Liu Yazhi, et al. RGB-T salient object detection based on cross-level attention learning[J]. Journal of Zhengzhou University (Natural Science Edition), 2025, 57(03): 42—48.

[7] Lv C, Zhou X, Wan B, et al. Transformer-based cross-modal

- integration network for RGB-T salient object detection[J]. IEEE Transactions on Consumer Electronics, 2024, 70(2): 4741—4755.
- [8] Liu Z, Tan Y, He Q, et al. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4486—4497.
- [9] Tang B, Liu Z, Tan Y, et al. HRTransNet: HRFormer-driven two-modality salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(2): 728—742.
- [10] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815—828.
- [11] Luo Z, Mishra A K, Achkar A, et al. Non-local deep features for salient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6593—6601.
- [12] Zhang Q, Huang N C, Yao L, et al. RGB-T salient object detection via fusing multi-level CNN features[J]. IEEE Transactions on Image Processing, 2020, 29: 3321—3335.
- [13] Wei J, Wang S, Huang Q. F3Net: Fusion, feedback and focus for salient object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12321—12328.
- [14] Tu Z, Li Z, Li C, et al. Multi-interactive dual-decoder for RGB-thermal salient object detection[J]. IEEE Transactions on Image Processing, 2021, 30: 5678—5691.
- [15] Bi H, Zhang J, Wu R, et al. RGB-T salient object detection via excavating and enhancing CNN features[J]. Applied Intelligence, 2023, 53(21): 25543—25561.
- [16] Huo F, Zhu X, Zhang L, et al. Efficient context-guided stacked refinement network for RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 3111—3124.
- [17] Jiang X, Zhu L, Hou Y, et al. Mirror complementary transformer network for RGB-thermal salient object detection[J]. IET Computer Vision, 2023, 18(1): 15—32.
- [18] Huo F, Zhu X, Zhang Q, et al. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1—12.
- [19] Zhang P, Xu M N, Zhang Z Y, et al. Feature aggregation with transformer for RGB-T salient object detection[J]. Neurocomputing, 2023, 546: 126329.
- [20] Zhou W, Zhu Y, Lei J, et al. LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images[J]. IEEE Transactions on Image Processing, 2023, 32: 1329—1340.
- [21] Liu C, Yang G, Wang Y. TANet: Transformer-based asymmetric network for RGB-D salient object detection[J]. IET Computer Vision, 2023, 17(4): 415—430.
- [22] Tu Z, Ma Y, Li Z, et al. RGBT salient object detection: a large-scale dataset and benchmark[J]. IEEE Transactions on Multimedia, 2022, 25: 4163—4176.
- [23] Song K, Huang L, Gong A, et al. Multiple graph affinity interactive network and a variable illumination dataset for RGBT image salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3104—3118.
- [24] Pang Y, Zhao X, Zhang L, et al. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection[J]. IEEE Transactions on Image Processing, 2023, 32: 892—904.
- [25] Jin D, Shao F, Xie Z, et al. CAFNet: Cross-modality asymmetric feature complement network for RGB-T salient object detection[J]. Expert Systems with Applications, 2024, 247: 123222.