控制与决策
Control and Decision

双模板跨模态交互与前景选择的高效RGB-T目标跟踪

柳长源, 范培栋, 兰朝凤

引用本文:
柳长源, 范培栋, 兰朝凤. 双模板跨模态交互与前景选择的高效RGB-T目标跟踪[J]. 控制与决策, 2025, 40(12): 3725-3733.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0495

双模板跨模态交互与前景选择的高效 RGB-T 目标跟踪

柳长源†，范培栋，兰朝凤

(哈尔滨理工大学 测控技术与通信工程学院，哈尔滨 150080)

摘　要: 利用可见光 (RGB) 和热红外 (TIR) 双模态信息间的互补性可以克服单模态跟踪在恶劣环境下的局限性. 目前基于 RGB-T 的目标跟踪方法不能充分利用模态间信息, 而且额外模态的引入会导致计算量增大. 为此, 提出双模板跨模态交互与前景选择的高效 RGB-T 目标跟踪网络, 对两个模态的模板图像进行融合构建融合模板图像分支, 利用融合模板图像特征和两个模态模板图像特征作为模态交互的纽带, 克服不同模态图像中心存在的偏差导致两种模态信息利用不充分问题; 利用极性感知线性注意力构建 Transformer 编码器, 减少 ViT(Vision Transformer) 中的多头注意力机制带来的复杂计算量, 提高模型的效率; 通过极性感知线性注意力返回的注意力构建前景选择模块, 去除无关背景特征, 提高跟踪精度的同时减少背景特征带来的计算量. 实验结果表明, 所提出网络在 LasHeR 数据集上跟踪成功率达到 57.1%, 精确率达到 71.2%, 相较于模板连接搜索区域交互算法 (TBSI) 分别提升 1.1% 和 1.5%, 跟踪速度相较于 TBSI 提升 3.5%, 在 RGB-T 目标跟踪任务中取得了较好效果.

关键词: 目标跟踪；RGB-T；双模板跨模态交互；极性感知线性注意力；前景选择；跟踪速度

中图分类号: TP391.4　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0495

引用格式: 柳长源, 范培栋, 兰朝凤. 双模板跨模态交互与前景选择的高效 RGB-T 目标跟踪 [J]. 控制与决策, 2025, 40(12): 3725-3733.

Efficient RGB-T object tracking network with dual-template cross-modality interaction and foreground selection

LIU Chang-yuan†，*FAN Pei-dong*，*LAN Chao-feng*

(College of Measurement and Control Technology and Communication Engineering，Harbin University of Science and Technology，Harbin 150080，China)

Abstract: To overcome single-modal tracking limitations in adverse environments, we exploit complementary information from visible (RGB) and thermal infrared (TIR) modalities. However, existing visible and thermal infrared (RGB-T) tracking frameworks often inadequately leverage inter-modal correlations or efficiently mitigate computational overhead from dual-modal fusion. We propose an efficient RGB-T tracker with dual-template cross-modality interaction and foreground selection. The template images from the two modalities are fused to construct a merged template branch, and both the fused template features and the individual modal template features are used as a bridge for cross-modal interaction, thereby addressing the center misalignment between modalities and fully leveraging information from both sources. To reduce computational burden from Vision Transformer(ViT)'s multi-head attention, we construct a polarity-aware linear attention Transformer encoder. Additionally, a foreground selection module processes PolaFormer's attention maps to eliminate background features, enhancing precision while lowering computational load. On the LasHeR dataset, the proposed method achieves success rate and precision of 57.1% and 71.2%, respectively. This represents an improvement of 1.1% and 1.5% over TBSI, with a tracking speed 3.5% higher, surpassing state-of-the-art RGB-T tracking approaches.

Keywords: object tracking； RGB-T； dual-template cross-modality interaction； polarity-aware linear attention； foreground selection；tracking speed

0　引　言

单目标跟踪是指在给定目标的情况下, 在后续的连续视频帧中持续准确地跟踪目标对象. 近年来, 随着目标跟踪技术的快速发展, 其在无人驾驶[1]、社

收稿日期: 2025-05-13；**录用日期:** 2025-08-19.

基金项目: 黑龙江省交通运输厅科技项目 (HJK2024B002)；黑龙江省“优秀青年教师基础研究支持计划”重点项目 (YQJH2024064).

†通信作者. E-mail: liuchangyuan@hrbust.edu.cn.

区安防[2]等民用领域以及船舶综合舰桥系统[3]等军事领域发挥了十分重要的作用. 目前基于 RGB 的目标跟踪在光照变化、背景杂乱等复杂环境中难以准确地跟踪目标[4-5]. 随着热红外相机的出现, 研究者发现 TIR 图像与 RGB 图像间存在互补性[6], 同时利用两种模态图像能够提升单模态目标跟踪算法的跟踪准确性与鲁棒性.

近年来, 随着卷积神经网络 (CNN) 不断发展, 基于 MDNet(multi-domain network)[7-9] 的 RGB-T 目标跟踪算法和基于 Siamese 网络[10-12] 的 RGB-T 目标跟踪算法成为两大主流跟踪方法. 基于 MDNet 的跟踪算法跟踪准确率较高, 但是该方法采用的在线学习机制导致其运行速度较慢. 基于 Siamese 网络的跟踪算法跟踪速度较快, 但是该方法采用简单的互相关运算导致跟踪准确率较低. 得益于 Transformer 架构近年来在计算机视觉领域 (CV) 的应用, 其中的注意力机制可以有效地将目标跟踪中模板图像和搜索图像进行全局关系建模, 从而减少关键信息丢失, 因此 Transformer 目前已经成为跟踪领域主流的框架.

Cao 等[13] 提出一个通用的双向适配器, 在冻结的 Transformer 块交叉提示多个通道交互, 动态提取两个模态的互补信息. Zhu 等[14] 提出一种基于视觉提示的多模态跟踪算法 (VIPT), 侧重于学习与模态相关的线索, 使冻结的预训练基础模型能够适应下游多模型跟踪任务. Hui 等[15] 提出一种模板连接搜索区域交互的算法 (TBSI), 利用融合模板作为 RGB 和 TIR 图像搜索区域间跨模态交互的媒介, 用来引导 RGB 搜索区域与 TIR 搜索区域进行跨模态交互. 尽管上述算法利用互补模态的特征一定程度上提高了跟踪的准确性, 但是忽略了不同模态图像间存在的偏移, 导致模态间交互不充分. 同时也没有考虑搜索序列块送入跟踪头时无关背景的干扰, 从而导致背景序列块引入冗余并干扰跟踪的准确性. 针对单模态场景, Ye 等[16] 提出基于 Transformer 注意力图的候选消除模块 (CE), 用以抑制背景干扰, 但是该模块插入到每个单独的编码器中, 缺少全局注意力图的整合, 可能会导致关键信息丢失. CE 模块适用于单一模态, 目前尚缺乏在双模态跟踪中实现背景消除的方法.

针对上述问题, 本文在 TBSI 的基础上提出双模板跨模态交互与前景选择的高效 RGB-T 目标跟踪网络 (DIFTrack), 目标是利用融合模板图像与每个模态模板图像交互后的双模板作为纽带, 减小模态间图像的偏移差距, 对两个模态信息进行充分利用. 利用极性感知线性注意力提高模板-搜索图像匹配和特征提取的效率, 并通过该过程返回的注意力图筛选前景特征, 减小背景对模型的干扰.

1 双模板跨模态交互与前景选择网络

基线模型 TBSI 整体结构由双分支 ViT 特征提取网络、模态间特征交互 TBSI 模块以及跟踪头 3 部分组成.

本文提出的 DIFTrack 网络整体结构如图 1 所示, 相较于 TBSI 主要包括 3 个改进部分:

1) 融合两个模态模板图像, 生成融合模板图像, 提出双模板跨模态交互模块 (DTIM), 充分利用融合模板图像和两种模态模板图像信息作为后续模态间交互的纽带, 有效改善不同模态相机图片中心存在偏移导致融合不充分的问题.

2) ViT 中的多头注意力运算量大, 计算成本高昂, 本文通过使用极性感知线性注意力 (PolaFormer),

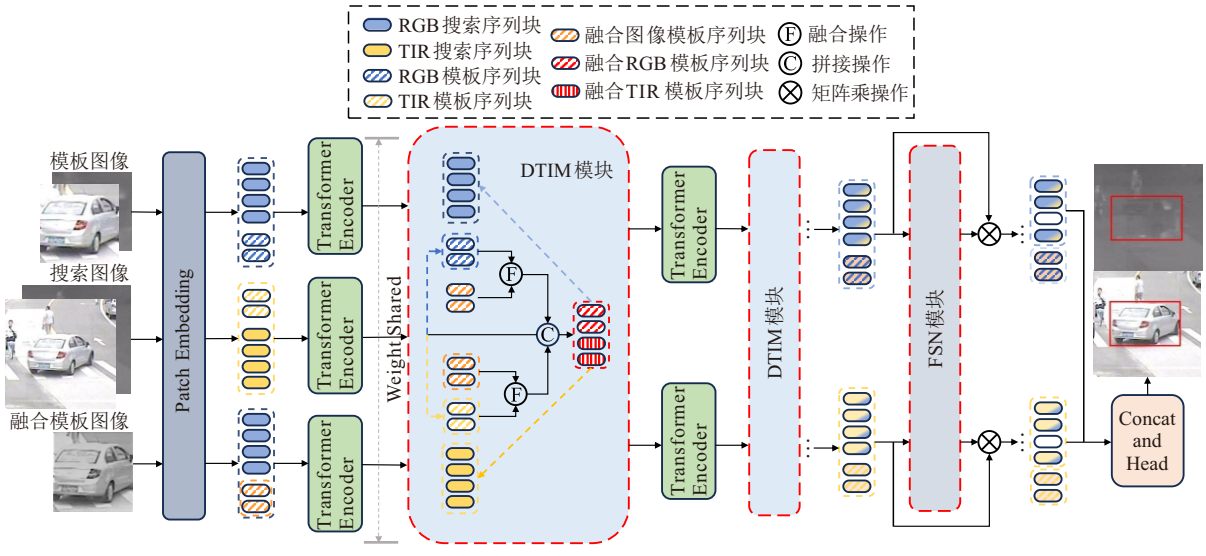


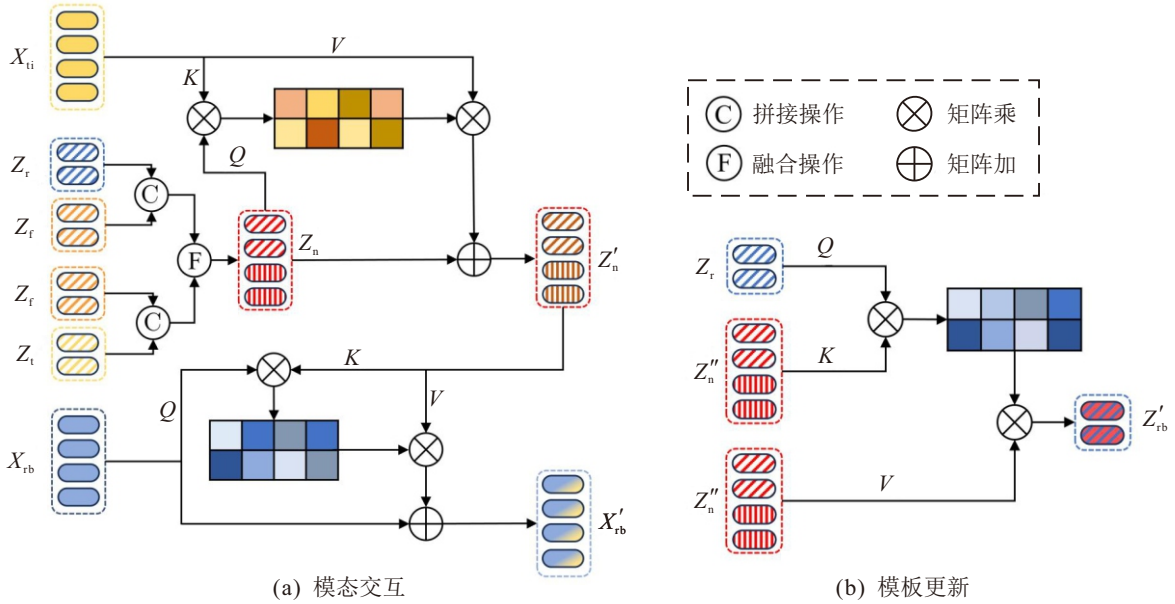
图1 DIFTrack 整体结构

在提高模型表达能力的同时提高模型的效率.

3) 当前多模态融合跟踪均是将前景和背景一起进行融合, 为了解决背景存在的干扰, 提出前景选择网络 (FSN) 去除冗余, 有效减少无关背景对跟踪对象的干扰, 解决跟踪过程计算量大的同时提高了跟踪的精度.

1.1　双模板跨模态交互模块

DTIM 通过引入融合模板作为中介, 构建了 RGB 与融合间、融合与 TIR 间的双向交互路径, 使两种模态能够在共享语义空间中进行特征对齐, 从而提升跨模态匹配的准确率. 首先, 利用文献 [17] 所述融合策略, 即通过使用一个双分支的 Transformer-CNN 框架提取不同模态的全局低频特征和局部高频特征, 对两种模态的模板图像进行融合; 然后, 将融合模板图像与两种模态的输入图像一起进行相同的切块及其他处理, 并送入 Transformer 块进行特征提取; 最后, 将两个模态与融合图像的特征一并送入 DTIM 模块. DTIM 模块的具体结构如图 2 所示.



(a) 模态交互　　(b) 模板更新

图2　DTIM 模块

1.1.1　双模板融合

为了后续两个模态间高效的特征交互, 利用模态间互补性提高跟踪准确率. 首先, 将 RGB 模板图像 Z_r、TIR 模板图像 Z_t 分别与融合模板图像 Z_f 进行融合, 得到 $Z_{rf} \in \mathbb{R}^{N_z \times C}$ 与 $Z_{tf} \in \mathbb{R}^{N_z \times C}$, 聚合单一模态与融合模态的特征, 丰富目标的初始位置信息, 作为纽带有效提升跟踪精度; 然后, 将 Z_{rf} 与 Z_{tf} 在序列维度上进行拼接, 准备进行模态间特征融合. 该过程可表示为

$$\begin{cases} Z_{rf} = [Z_r; Z_f]W_m; \\ Z_{tf} = [Z_t; Z_f]W_m; \\ Z_n = [Z_{rf}; Z_{tf}]. \end{cases} \tag{1}$$

其中: Z_f 表示融合模板图像; $W_m \in \mathbb{R}^{2C \times C}$ 表示使用的线性层参数, 用于将融合模板图像与单一模态图像进行简单融合.

1.1.2　双模态交互

为了充分利用 RGB 模态和 TIR 模态图像的互补性, 从而提高模型对目标的识别准确性, 模态交互模块设计至关重要. 本文模态交互模块在模板图像与两个模态的搜索区域图像间采用交叉注意力的方法, 插入到连续的 Transformer 块中, 聚合两个模态的有效特征, 便于跟踪头输出目标位置信息. 从 TIR 到 RGB 模态交互模块如图 2 (a) 所示.

将 TIR 图像的搜索区域特征 X_{ti} 与双模板融合后的特征 Z_n 进行多头交叉注意力计算, 对 Z_n 进行更新得到 Z'_n, Z'_n 含有 TIR 图像丰富的目标特征, 该过程可表示为

$$\begin{cases} Z'_{tr} = \\ \mathrm{LN}\left(Z_n + \mathrm{Softmax}\left(\dfrac{(Z_n W_q)(X_{ti} W_k)^{\mathrm{T}}}{\sqrt{C}}\right)(X_{ti} W_v)\right); \\ Z'_n = \mathrm{LN}(Z'_{tr} + \mathrm{MLP}(Z'_{tr})). \end{cases} \tag{2}$$

其中: W_q、W_k、W_v 分别表示查询向量、键向量、数值向量投影层的可学习参数, LN 表示 LayerNorm 操作, MLP 表示多层感知机. 用 $\mathrm{MHSA}(X, Y)$ 表示多头交叉注意力, 其中 X 表示查询向量, Y 表示键向量与数值向量.

在获得包含 TIR 目标特征的丰富信息后, 将其与 RGB 特征 X_{rb} 交互便完成了从 TIR 到 RGB 的搜

索图像双模态交互, 该过程可表示为

$$\begin{cases} Z_{\rm tr}'' = {\rm LN}(X_{\rm rb} + {\rm MHSA}(X_{\rm rb}, Z_{\rm n}')), \\ X_{\rm rb}' = {\rm LN}(Z_{\rm tr}'' + {\rm MLP}(Z_{\rm tr}'')). \end{cases} \tag{3}$$

从 RGB 的搜索区域特征到 TIR 的搜索区域特征完成交互的过程与上述从 TIR 到 RGB 的过程类似, 使用从 TIR 到 RGB 过程得到的$Z_{\rm n}'$作为新的模板输入, 得到与两个模态的搜索图像均交互的融合模板$Z_{\rm n}''$和从 RGB 到 TIR 搜索图像交互后的特征.

1.1.3 模板更新

RGB 和 TIR 图像的搜索区域特征已经完成模态间的交互, 但是原始两个模态的模板图像特征只含单一模态, 无法提供准确目标特征信息, 因此将融合模板$Z_{\rm n}''$与两个模态的模板特征进行交互更新对于提高模型的鲁棒性是必要的. 更新 RGB 图像的模板特征得到$Z_{\rm rb}'$的过程如图 2 (b) 所示, 可表示为

$$\begin{cases} Z_{\text{rgb-up}} = {\rm LN}(Z_{\rm r} + {\rm MHSA}(Z_{\rm r}, Z_{\rm n}'')), \\ Z_{\rm rb}' = {\rm LN}(Z_{\text{rgb-up}} + {\rm MLP}(Z_{\text{rgb-up}})). \end{cases} \tag{4}$$

更新 TIR 图像的模板过程与上述更新 RGB 的过程类似, 更新后得到 TIR 图像模板特征$Z_{\rm ti}'$, 将$Z_{\rm rb}'$与$X_{\rm rb}'$连接, $Z_{\rm ti}'$与$X_{\rm ti}'$连接作为 DTIM 模块的输出.

1.2 极性感知线性注意力模块

传统 Softmax 注意力机制会平滑所有输入值, 导致强信号被稀释、弱信号被放大, 影响注意力集中效果. Meng 等[18] 提出了极性感知线性注意力, 通过将查询-键向量分为正负两部分进行独立建模, 强化了对关键区域的响应, 提高了注意力图的质量, 同时采用线性操作进一步降低了计算复杂度. 极性感知线性注意力模块的具体结构如图 3 所示.

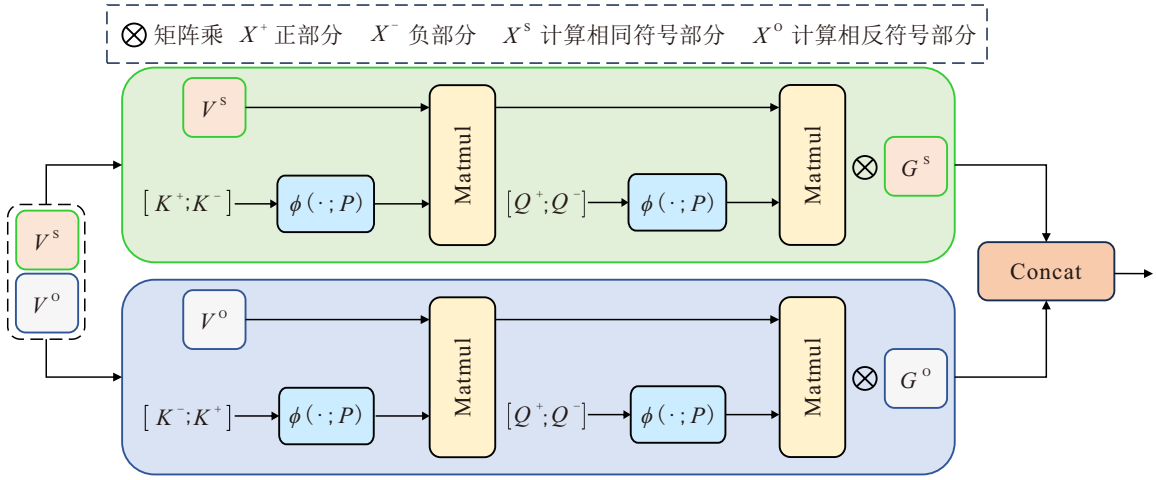


图3 极性感知线性注意力模块

ViT 中基于 Softmax 的多头注意力机制, 图像$X \in \mathbb{R}^{H\times W\times C}$经过 PE(patch embadding) 等操作得到$X \in \mathbb{R}^{N\times D}$. 其中: N表示序列长度, D表示维度. 每个头的计算量为$O(N^2d)$, 而极性感知线性注意力将查询-键向量分为正负两部分, 可表示为

$$\begin{cases} q = q^+ - q^-,\ q_i^+ = \max(q_i, 0); \\ k = k^+ - k^-,\ k_i^+ = \max(k_i, 0). \end{cases} \tag{5}$$

其中q与k分别表示查询向量Q与键向量K中的某一序列.

通过对正负两部分向量的计算可以有效减少信息丢失, 随后对查询-键的每个正负元素通过核函数进行特征映射, 达到基于 Softmax 注意力中指数函数的效果, 解决线性注意力熵较高的问题. 将数值向量在通道维度上分为两部分$V^{\rm s}$与$V^{\rm o}$, 分别用于处理相同和相反符号的计算, 随后合并正负分量得到输出的注意力为

$$o_t = \left[\frac{\phi([q_t^+;q_t^-])\sum_{i=1}^{N}\phi([k_i^+;k_i^-])^{\rm T}v_i^{\rm s}}{\phi([q_t^+;q_t^-])\sum_{j=1}^{N}\phi([k_j^+;k_j^-])^{\rm T}} \otimes G^{\rm s}; \right.$$
$$\left. \frac{\phi([q_t^+;q_t^-])\sum_{i=1}^{N}\phi([k_i^-;k_i^+])^{\rm T}v_i^{\rm o}}{\phi([q_t^+;q_t^-])\sum_{j=1}^{N}\phi([k_j^-;k_j^+])^{\rm T}} \otimes G^{\rm o} \right]. \tag{6}$$

其中: [;] 表示将相同和相反符号的计算进行连接, $\otimes$表示矩阵乘法, ϕ表示线性注意力中核函数, $G^{\rm s} \in \mathbb{R}^{N\times d/2}$、$G^{\rm o} \in \mathbb{R}^{N\times d/2}$表示两个可学习的极性感知系数矩阵.

极性感知线性注意力通过核函数将查询向量和键向量进行特征映射, 使得式 (6) 分子分母的乘法均为线性操作, 计算量减少至$O(Nd^2)$. 本文将 ViT 中多头注意力替换为 PolaFormer 并且返回每个头得到

的注意力矩阵供后续前景选择使用, 在不降低模型表达能力的同时减少了计算量, 提高了模型跟踪的效率.

1.3　前景选择模块

由于 Transformer 强大的全局建模能力可以通过注意力机制逐步关注搜索图像中的目标, 但是也会捕捉目标与背景不相关区域间的关系, 计算得到的注意力权重非但不会对目标定位提供有用的贡献, 反而会引入冗余增加模型的复杂度和计算量. 针对此问题, 本文提出前景选择模块 FSN 以过滤掉部分无关背景, 进而提高模型的跟踪效率. 图 4 为前景选择器模块具体结构, 此处仅以单个注意力头示意前景选择模块的总体结构.

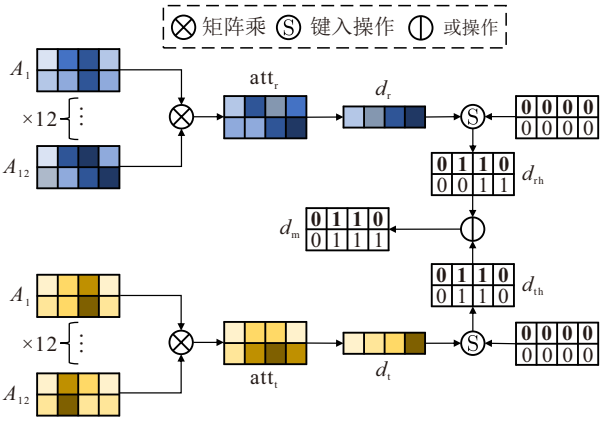


图4　前景选择模块

本文设计的前景选择模块 FSN 根据主干网返回的注意力图进行前景选择, 模块插入到跟踪头之前, 连续进行两次的前景选择, 每次特征选择后经过一个聚合模块捕捉丰富的语义信息, 提高模型在实际应用中的鲁棒性和泛化能力. 特征选择模块首先需要将得到的不同层注意力图列表进行连乘, 找到连乘后得到的注意力图中具有最高注意力值的前$N\times$ ratio个特征, 得到特征的索引为

$$\begin{cases}\text{att} = A_1 \times A_2 \times \cdots \times A_i, \\ d = \text{topk}(\text{att}, N \times \text{ratio}).\end{cases} \tag{7}$$

其中: A_i为第i层的注意力图, att为连乘后的注意力图, topk为排序操作, N 为注意力图列表中元素的长度, ratio为一个可以自定义的比例系数, $N \times \text{ratio}$ 为选择的特征数量. 得到前景特征的索引d后, 创建一个布尔型的掩码, 通过索引d对掩码对应位置置 1, 其余位置置 0, 上述操作对每个头的注意力图遍历执行, 最后得到所有通道单独选择的特征掩码组合d_l为

$$d_l = \bigvee_{i}^{h} \text{mask}_i. \tag{8}$$

其中: mask_i表示第i个头选定的特征掩码, $\bigvee$表示或操作. 将 RGB 模态的特征与 TIR 模态的特征分别经过上述操作得到每个模态的特征掩码d_{rh}与d_{th}, d_{rh}与d_{th}相或得到最终的索引d_{m}, 通过这个索引进行前景的选择, 该过程可表示为

$$\begin{cases} d_{\text{m}} = d_{\text{rh}} \bigvee d_{\text{th}}, \\ N_{\text{r}} = N_{\text{r}} \times d_{\text{m}}, \\ N_{\text{t}} = N_{\text{t}} \times d_{\text{m}}, \end{cases} \tag{9}$$

其中N_{r}与N_{t}表示选定的前景特征. 将选定的前景特征送入特征聚合模块, 分别对 RGB 和 TIR 两个模态的前景特征进行独立处理, 再进行交互聚合, 最后输出X_{output}同样需要应用掩码, 整个模块需要连续执行两次上述操作, 第 2 次与上述第 1 次流程基本一致. 通过连续两次前景选择, 可以有效地去除无关噪声, 提高模型的跟踪效率.

2　实验结果及分析

2.1　实验数据集

实验采用 RGB-T 目标跟踪数据集 LasHeR[19] 作为实验数据集, 该数据集由 1 224 个 RGB 和 TIR 视频序列组成, 含有 32 个目标类别, 包含遮挡、形变和热交叉等 19 个属性挑战, 超过 730 K 帧对. LasHeR 将 1 224 个序列分为 979 个序列的训练集和 245 个序列的测试集, 相较之前的 RGB-T 数据集, LasHeR 数据集规模更大, 包含的挑战更多.

本文采用 RGB-T 目标跟踪数据集 RGBT234 进一步验证所提出模型的有效性, 该数据集由 234 个 RGB 和 TIR 视频序列对组成, 总帧数为 234 K, 每个序列的最大帧为 8 K, 包含遮挡、低照度、热交叉、快速移动等 12 个挑战属性.

2.2　实验环境参数设置及评价指标

本文实验基于 Linux 系统, 实验硬件环境采用 IntelCore i9-13900K CPU, 单张 GeForce NVIDIA RTX 4080 显卡, 显存 16 G. Python 版本为 3.8, 使用 pytorch 1.7.1 深度学习框架. DTIM 模块插入到 ViT 主干的第 4、第 7、第 10 层, 输入搜索图像尺寸为 256×256 像素, 模板图像尺寸为 128×128 像素, 采用 AdamW 优化器进行模型训练, ViT 主干学习率设置为 1e-5, 其他参数的学习率设置为 1e-4, 10 个轮次后衰减 10 倍. 训练轮数设置为 15, 批处理大小设置为 8.

实验指标主要包括成功率 (SR)、精确率 (PR) 和标准化精确率 (NPR) 3 个指标. 成功率指预测边界框与真实边界框的重叠部分大于某个阈值的数量与所有预测帧数量的百分比; 精确率指预测边界框的中心点位置与真实中心点位置的偏差小于某个阈值

的百分比; 标准化精确率指在精确率的基础上, 对所得误差进行归一化处理, 消除不同目标大小或分辨率带来的影响. 3 个指标值越高代表跟踪效果越好.

2.3 消融实验及分析

为了验证本文设计模块的有效性, 在 LasHeR 数据集上对所提出的模块进行消融实验, 实验结果如表 1 所示, 其中 TBSI 表示基线模型用于两个模态间交互的模块.

表1 LasHeR 数据集上各模块消融实验

实验	TBSI	DTIM	FSN	PolaFormer	SR / %	PR / %	NPR / %	FPS / %
基线模型	√				56.0	69.7	65.8	34.1
实验1		√			56.7	70.7	66.7	31.3
实验2	√		√		56.3	70.2	66.2	36.3
实验3	√		√	√	56.5	70.3	66.5	39.2
实验4		√		√	56.8	70.9	66.8	35.1
实验5		√	√		57.0	71.0	67.2	33.9
实验6		√	√	√	57.1	71.2	67.4	37.6

由表 1 可见: 将基线模型用于两个模态间交互的 TBSI 模块替换为本文提出的 DTIM 模块, SR 提高了 0.7%, PR 提高了 1%; 在原基线模型的基础上添加设计的 FSN 模块, 精确率与平均精确率分别提升 0.5% 与 0.3%, 并且跟踪速度也得到提升; 由实验 2 和实验 3 可知, 在替换极性感知线性注意力后, 在 3 个评价指标上均有提升, 跟踪速度也再次得到提升; 由实验 5 可知, 同时采用本文设计的 DTIM 模块和 FSN 模块后, 相较于基线模型成功率提升了 1%, 精确率提升了 1.3%; 所设计的所有模块在原网络改进后, 精确率提升至 71.2%, 成功率提升至 57.1%, 标准化精确率提升至 67.4%, 并且 FSN 模块和极性感知线性注意力模块加快了模型的跟踪速度, 降低了模板图像融合带来的额外计算量.

2.4 对比实验及分析

为了对本文所提出网络的整体性能进行验证, 在 LasHeR 测试集上主要使用 SR、PR 和 NPR 与近年主流 RGB-T 目标跟踪方法进行比较, 实验结果如表 2 所示.

由表 2 可见: 本文网络相较于基线模型 TBSI, SR、PR 和 NPR 分别提高 1.1%、1.5% 和 1.6%; 与其他基于 ViT 主干的方法相比, 在 3 个评价指标上均取得最优结果, 其中相较于 BAT 与 TATrack, PR 高出 1%, SR 分别高出 0.8% 和 1%; 与 VIPT 相比, 本文网络 SR 高出 4.6%, PR 高出 6.1%; 在 3 个评价指标上, 本文方法均优于基于 ResNet 与 VGG 主干的方法, 体现了本文网络在跟踪任务上的优越性.

表2 与主流算法在 LasHeR 数据集上对比实验

方法	主干	SR / %	PR / %	NPR / %
DMCNet[20]	VGG-M	35.5	49.0	43.1
APFNet[9]	VGG-M	36.2	50.0	43.9
CAT++ [21]	VGG-M	35.6	50.9	44.4
MFNet[22]	ResNet-50	46.7	59.7	55.4
CMD [23]	ResNet-18	46.4	59.0	54.6
mfDIMP [24]	ResNet-50	46.7	59.9	—
VIPT[14]	ViT-Base	52.5	65.1	61.7
TBSI[15]	ViT-Base	56.0	69.7	65.8
BAT [13]	ViT-Base	56.3	70.2	66.4
TATrack [25]	ViT-Base	56.1	70.2	66.7
本文网络	ViT-Base	57.1	71.2	67.4

为了进一步验证模型的有效性, 将本文网络和其他算法在 RGBT234 数据集上的跟踪结果进行对比, 实验结果如表 3 所示.

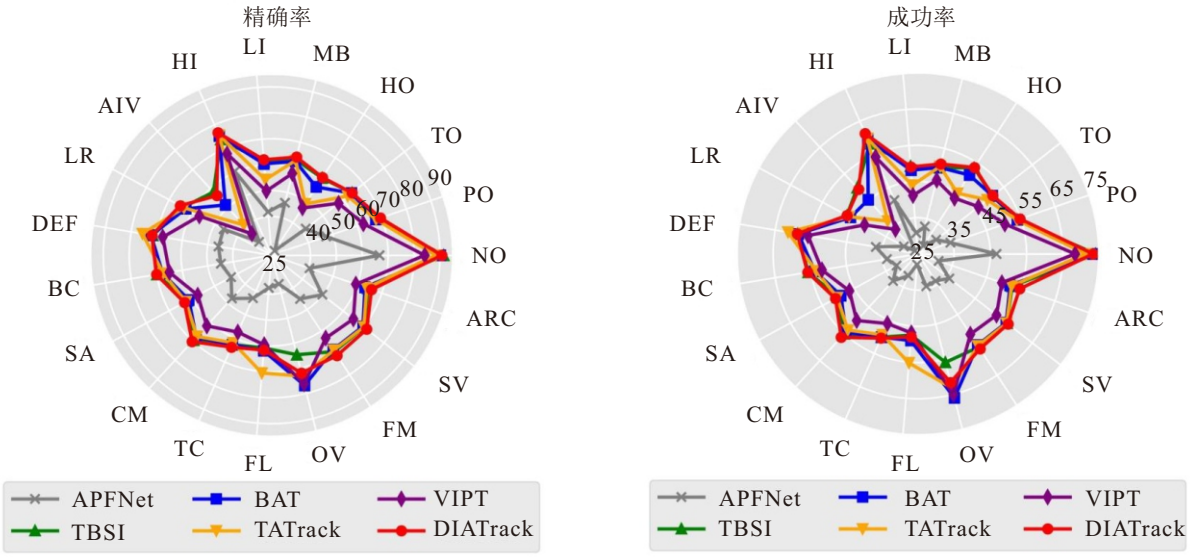
表3 与主流算法在 RGBT234 数据集上对比实验

方法	主干	SR / %	PR / %
DMCNet[20]	VGG-M	59.3	83.9
APFNet[9]	VGG-M	57.9	82.7
CAT++ [21]	VGG-M	59.2	84.0
SiamMT [12]	ResNet-50	54.2	79.5
MFNet[22]	ResNet-50	60.1	84.4
CMD [23]	ResNet-18	58.4	82.4
mfDIMP [24]	ResNet-50	59.1	84.2
VIPT[14]	ViT-Base	61.7	83.5
TBSI[15]	ViT-Base	63.7	87.1
BAT [13]	ViT-Base	64.1	86.8
TATrack [25]	ViT-Base	64.4	87.2
本文网络	ViT-Base	64.9	88.0

由表 3 可见: 本文网络在 RGBT234 数据集上的 SR 达到 64.9%, PR 达到 88.0%. 相较于基线模型 TBSI, SR 与 PR 分别提高 1.2% 与 0.9%; 与 TATrack 相比, SR 高出 0.5%, PR 高出 0.8%; 与 VIPT 相比, SR 高出 3.2%, PR 高出 4.5%. 综上所述, 本文方法不仅在 SR 和 PR 两个核心指标上超越了当前主流方法, 而且在相同主干网络的对比中展现出明显优势, 充分验证了本文网络在 RGB-T 多模态目标跟踪任务中的有效性和优越性.

LasHeR 数据集包含 19 个挑战属性, 分别是无遮挡 (NO)、部分遮挡 (PO)、完全遮挡 (TO)、透明遮挡 (HO)、运动模糊 (MB)、低照度 (LI)、高照度 (HI)、突然光照变化 (AIV)、低分辨率 (LR)、变形 (DEF)、背景杂乱 (BC)、相似外观 (SA)、相机移动 (CM)、热交叉 (TC)、帧丢失 (FL)、视野外 (OV)、快速移动 (FM)、比例变化 (SV) 和纵横比变化 (ARC). 为了评估本文所提出算法在 LasHeR 数据集各种挑战属性下的性

能, 将本文算法与基线模型 TBSI 以及当前 RGB-T 跟踪领域具有代表性的其余 4 种算法进行对比实验, 实验结果如图 5 所示. 由图 5 可见, 本文所提出的网络 DIFTrack 在大多数属性下精确率和成功率均领先其余 5 种算法. 其中在 MB、BC、FM、SV 等属性下具有明显优势, 说明通过本文所设计的 DTIM 模块, 能够充分提取模态间的互补信息, 增加模型的抗干扰能力.



(a) 精确率属性　　(b) 成功率属性

图5　在 LasHeR 数据集上不同算法属性对比

2.5　可视化结果及分析

2.5.1　跟踪结果可视化

为验证本文算法跟踪的具体效果, 将本文算法与 APFNet、BAT、VIPT、TBSI 在 LasHeR 数据集上进行定性分析, 在 4 个典型跟踪序列上的跟踪结果如图 6 所示. ab_blkskirtgirl 序列需要跟踪被树木遮挡的女孩, 如图 6 (a) 所示. 在初始帧被完全遮挡和最后帧被部分遮挡的条件下, 本文所提出的网络与真值框最接近, 优于其余 4 种算法. bikeboyintodark 和 rightdarksingleman 序列是经典的在黑暗和光亮环境下进行切换的序列, 包括光照变化、高照度和形变等挑战属性. 如图 6 (b) 所示, bikeboyintodark 序列的第 557 帧, 在遮挡和光照的影响下, 其余 4 种算法均未能正确框选目标, 而本文网络通过对两种模态信息的充分利用, 可有效提升极端条件下的跟踪性能. 图 6 (d) 所示为在街道场景下跟踪快速移动的男孩, BAT、TBSI 和 VIPT 均存在错误跟踪相似目标的情况, 而本文网络依然可以稳定跟踪正确目标.

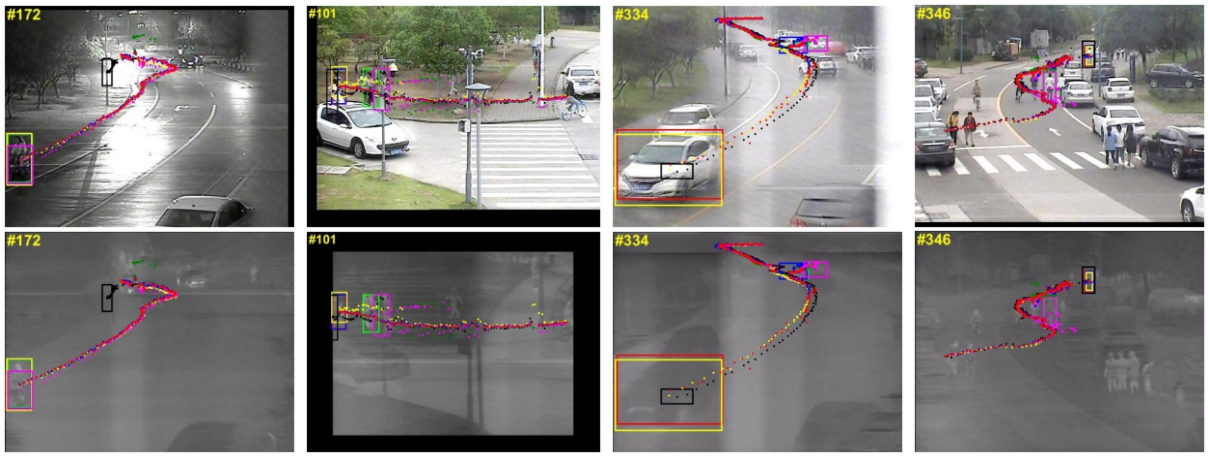
(a) ab_blkskirtgirl　(b) bikeboyintodark　(c) rihgtdarksingleman　(d) whiterunningboy

真值　APFNet　BAT　VIPT　TBSI　本文网络

图6　4 个序列上的可视化跟踪结果

2.5.2　跟踪轨迹可视化

为了直观展示算法跟踪性能, 将本文算法与其余 4 种算法在 LasHeR 数据集上进行跟踪轨迹对比分析, 在 4 个典型跟踪序列上的跟踪结果如图 7 所示. 由图 7 可见, 本文网络在目标发生快速运动、背景干扰严重、光照变化剧烈和目标部分遮挡等复杂场景下, 依然能够保持稳定且准确的跟踪轨迹, 而其他对比算法则出现不同程度的漂移或丢失现象. 在



(a) bikefromlight (b) couple (c) raincarturn (d) whiterunningboy

真值 APFNet BAT VIPT TBSI 本文网络

图7 4 个序列上的可视化跟踪轨迹结果

图 7 (c) 中, 基线模型 TBSI、BAT 和 VIPT 在车辆交汇产生遮挡时, 均丢失原有目标, 本文网络依然能准确跟踪目标, 这表明本文所提出算法在鲁棒性和精度方面优于现有方法, 能够更有效地处理实际应用中常见的复杂跟踪场景.

3 结 论

本文提出了一种双模板跨模态交互与前景选择的高效 RGB-T 目标跟踪网络, 用于解决目前 RGB-T 目标跟踪领域存在的两种模态信息交互不充分、计算量大的问题. 在 LasHeR 数据集上进行相关实验, 实验结果表明, 本文网络相较于基线模型, 跟踪速度提升了 3.5%, 成功率与精确率提升了 1.1% 与 1.5%, 表明了本文网络的有效性和在复杂场景下的鲁棒性. 与现阶段主流算法的实验结果进行对比, 验证了本文网络在目标跟踪任务上的优越性. 可视化实验进一步验证了 DIFTrack 在具有挑战的场景下也能准确跟踪目标, 具有较好的跟踪性能. 从算法性能提升的内在机制看, DIFTrack 的成功可归因于 3 大核心设计所形成的协同作用: 1) 模态对齐机制. 融合模板作为桥梁, 建立了 RGB 与 TIR 之间的双向信息流, 有效缓解了模态不对齐问题. 2) 注意力优化机制. 极性感知线性注意力通过符号敏感建模提升了注意力表达能力, 并降低了计算开销. 3) 特征聚焦机制. 前景选择模块通过注意力图筛选关键特征, 提升了模型对目标的关注度, 同时减少了背景干扰带来的冗余计算. 这 3 个设计共同作用, 使模型在复杂场景下具备更强的鲁棒性和更高的跟踪效率.

参考文献 (References)

[1] 张平, 迟志诚, 陈一凡, 等. 用于自动驾驶车辆的融合注意力机制多目标跟踪算法[J]. 汽车安全与节能学报, 2021, 12(4): 516-521.
(Zhang P, Chi Z C, Chen Y F, et al. Multiple object tracking algorithm integrated with attention mechanism for autonomous vehicles[J]. Journal of Automotive Safety and Energy, 2021, 12(4): 516-521.)

[2] Tang Q, Liang J. Maneuvering multitargets tracking system using surveillance multisensors[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-12.

[3] 张天宇. 浅谈目标跟踪技术及应用前景[J]. 中国新通信, 2021, 23(9): 31-33.
(Zhang T Y. A brief discussion on object tracking technology and its application prospects[J]. China New Telecommunications, 2021, 23(9): 31-33.)

[4] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]. Computer Vision – ECCV 2016 Workshops. Cham: Springer, 2016: 850-865.

[5] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8971-8980.

[6] Alldieck T, Bahnsen C, Moeslund T. Context-aware fusion of RGB and thermal imagery for traffic monitoring[J]. Sensors, 2016, 16(11): 1947.

[7] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 4293-4302.

[8] Zhu Y B, Li C L, Tang J, et al. Quality-aware feature aggregation network for robust RGBT tracking[J]. IEEE Transactions on Intelligent Vehicles, 2021, 6(1): 121-130.

[9] Xiao Y, Yang M M, Li C L, et al. Attribute-based progressive fusion network for RGBT tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2831-2838.

[10] Zhang X C, Ye P, Peng S Y, et al. SiamFT: An RGB-infrared fusion tracking method via fully convolutional

Siamese networks[J]. IEEE Access, 2019, 7: 122122-122133.

[11] Feng L L, Song K C, Wang J Y, et al. Exploring the potential of Siamese network for RGBT object tracking[J]. Journal of Visual Communication and Image Representation, 2023, 95: 103882.

[12] 齐咏生, 姜政廷, 刘利强, 等. SiamMT: 基于自适应特征融合机制的可修正 RGBT 目标跟踪算法[J]. 控制与决策, 2025, 40(4): 1312-1320.
(Qi Y H, Jiang Z T, Liu L Q, et al. SiamMT: A modifiable RGBT target tracking algorithm based on adaptive feature fusion mechanism[J]. Control and Decision, 2025, 40(4): 1312-1320.)

[13] Cao B, Guo J L, Zhu P F, et al. Bi-directional adapter for multimodal tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 927-935.

[14] Zhu J W, Lai S M, Chen X, et al. Visual prompt multi-modal tracking[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 9516-9526.

[15] Hui T R, Xun Z Z, Peng F G, et al. Bridging search region interaction with template for RGB-T tracking[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 13630-13639.

[16] Ye B T, Chang H, Ma B P, et al. Joint feature learning and relation modeling for tracking: A one-stream framework[C]. Computer Vision – ECCV 2022. Cham: Springer, 2022: 341-357.

[17] Zhao Z X, Bai H W, Zhang J S, et al. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 5906-5916.

[18] Meng W K, Luo Y D, Li X, et al. PolaFormer: Polarity-aware linear attention for vision transformers[J/OL]. 2025, arXiv: 2501.15061.

[19] Li C L, Xue W L, Jia Y Q, et al. LasHeR: A large-scale high-diversity benchmark for RGBT tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 392-404.

[20] Lu A D, Qian C, Li C L, et al. Duality-gated mutual condition network for RGBT tracking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4118-4131.

[21] Liu L, Li C L, Xiao Y, et al. RGBT tracking via challenge-based appearance disentanglement and interaction[J]. IEEE Transactions on Image Processing, 2024, 33: 1753-1767.

[22] Zhang Q, Liu X R, Zhang T L. RGB-T tracking by modality difference reduction and feature re-selection[J]. Image and Vision Computing, 2022, 127: 104547.

[23] Zhang T L, Guo H Y, Jiao Q, et al. Efficient RGB-T tracking via cross-modality distillation[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 5404-5413.

[24] Zhang L C, Danelljan M, Gonzalez-Garcia A, et al. Multi-modal fusion for end-to-end RGB-T tracking[C]. 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, 2019: 2252-2261.

[25] Wang H Y, Liu X T, Li Y F, et al. Temporal adaptive RGBT tracking with modality prompt[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5436-5444.

作者简介

柳长源 (1970−), 男, 副教授, 博士, 硕士生导师, 主要研究方向为图像处理与模式识别, E-mail: liuchangyuan@hrbust.edu.cn;

范培栋 (2001−), 男, 硕士生, 主要研究方向为图像处理与目标跟踪, E-mail: 3186682515@qq.com;

兰朝凤 (1981−), 女, 教授, 博士, 主要研究方向为机器视觉与信号处理, E-mail: lanchaofeng@126.com.