E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2026)04-1142-14

论文引用格式:Wang X, Zhang D W, Yang C Z, Wang Z G, Chen H and Zheng Z L. 2026. Three-branch multistage fusion network for RGB-T object tracking. Journal of Image and Graphics, 31(4):1142-1155(王炫, 张大伟, 阳诚砖, 王志刚, 陈灏, 郑忠龙. 2026. 面向RGB-T目标跟踪的三分支多阶段融合网络. 中国图象图形学报, 31(4):1142-1155)[DOI:10.11834/jig.250256]

面向RGB-T目标跟踪的三分支多阶段融合网络

王炫[1],张大伟[1,2*],阳诚砖[1],王志刚[3],陈灏[3],郑忠龙[1]

1. 浙江师范大学计算机科学与技术学院,金华　321004; 2. 浙江大学计算机科学与技术学院,杭州　310027;
3. 中国电信金华分公司大数据AI联合研究院,金华　321004

摘　要:**目的**　目标跟踪是指在一段视频序列中持续跟踪某个目标,单一模态来源限制了跟踪性能的提升,RGB-T(RGB-thermal)目标跟踪通过对可见光与热红外信息的融合来提升精度。但现有的算法往往侧重于模态间融合而忽略了局部与全局特征的融合,导致跟踪性能下降。本文提出一种基于多阶段融合的三分支RGB-T目标跟踪网络,在不同阶段对局部特征与全局特征分别建模,以促进多模态特征的融合与广泛传播。**方法**　设计了一种包含纯卷积分支的三分支多阶段融合网络结构,通过在网络的不同阶段采用不同的融合策略实现特征融合与增强。卷积分支通过卷积融合模块(convolutional fusion module,CFM)来提取和融合局部信息,并利用跨层连接的方式将其传播至深层网络以避免局部特征丢失。同时使用部分参数共享的注意力融合增强模块(attention fusion enhancement module,AFEM)来融合全局特征,从而缓解不同模态数据分布差异造成的影响。通过多阶段的全局和局部特征融合,本文方法能够提取更有效的多模态特征表示,以确保鲁棒的目标跟踪。**结果**　在3个常用跟踪数据集上的对比实验表明,所提方法在LasHeR数据集上的PR、NPR和SR相比于基线算法分别提升了2.4%、1.8%和1.5%,在大部分挑战属性上均有明显的性能提升,同时在RGBT210与RGBT234数据集上依然取得先进的跟踪结果。消融实验表明CFM、AFEM以及三分支结构的有效性。**结论**　本文提出的多阶段融合三分支网络能够对跨模态特征以及局部和全局特征进行更加充分的融合,在单一模态缺陷时能够更好地利用可靠模态的细节信息辅助决策,显著提升了RGB-T目标跟踪性能。

关键词:目标跟踪;注意力;卷积;多模态融合;三分支网络

Three-branch multistage fusion network for RGB-T object tracking

Wang Xuan[1], Zhang Dawei[1,2*], Yang Chengzhuan[1], Wang Zhigang[3], Chen Hao[3], Zheng Zhonglong[1]

1. School of Computer Science and Technology, Zhejiang Normal University, Jinhua　321004, China;
2. College of Computer Science and Technology, Zhejiang University, Hangzhou　310027, China;
3. Big Data and AI Research Institute, China Telecom Jinhua Branch, Jinhua　321004, China

Abstract: **Objective** Object tracking, which aims to track a target continuously in a given video sequence, is one of the important research directions in computer vision. It has been widely used in many fields, such as video surveillance,

收稿日期:2025-06-17;**修回日期**:2025-10-19;**预印本日期**:2025-10-26

*通信作者:张大伟　davidzhang@zjnu.edu.cn

基金项目:国家自然科学基金项目(62272419, 62402449);浙江省自然科学基金项目(LQ23F020010);多模态认知计算安徽省重点实验室(安徽大学)开放基金项目(MMC202409);浙江大学计算机辅助设计与图形系统全国重点实验室开放课题项目(A2421)

Supported by: National Natural Science Foundation of China (62272419, 62402449); Natural Science Foundation of Zhejiang Province, China (LQ23F020010); Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (MMC202409); Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University (A2421)

human-computer interaction, and visual navigation. The fundamental challenge lies in maintaining robust tracking performance across diverse environmental conditions, including extreme lighting variations, occlusions, and dynamic background scenarios. The emergence of deep learning significantly facilitates the development of the field of object tracking. However, coping with the challenges triggered by the attributes of the scene is difficult. RGB tracking uses only visible light information and often struggles with challenging scenarios, including dim illumination and bad weather, particularly in low-light conditions, foggy weather, or nighttime environments where color and texture information becomes unreliable or completely unavailable. In comparison, thermal infrared images use the temperature as a source of information, do not depend on lighting conditions, and can provide considerable details in low-light environments. However, thermal infrared images suffer from low resolution, noise contamination, and a lack of texture details. RGB-thermal (RGB-T) object tracking is a downstream task of visual tracking. It extracts features by fusing visible and thermal infrared images to complement each other, and its core lies in the way of feature fusion. It has emerged as a promising solution to overcome the limitations of single-modality tracking. As the earliest deep learning method, the convolutional neural network (CNN) is widely used in Siamese networks for RGB-T object tracking. After the emergence of the transformer, researchers use a unified network for joint feature learning and relation modeling, thereby facilitating the interaction between the template and search region. However, existing methods tend to single out CNN or a transformer for feature extraction. Moreover, they underutilize complementary local and global feature relationships even though various complex fusion modules are designed to facilitate intermodal information interaction. In particular, CNNs excel at capturing local structural details but lack long-range dependencies, whereas transformers capture global context but may dilute fine-grained details during attention computation. Thus, a three-branch multistage fusion network for RGB-T object tracking, which models shallow local features and deep global features separately and promotes the wide propagation of multistage features, is proposed in this study to address the aforementioned issue. Our approach leverages the complementary strengths of CNN and transformer while mitigating their respective limitations through a carefully designed multistage fusion strategy. **Method** The RGB modality possesses abundant detailed features, whereas the thermal infrared modality, despite lacking detailed information, still has significant global features that cannot be overlooked. We enhance and fuse features by employing different fusion strategies at different layers to achieve a robust representation and prevent interference between local and global features. First, considering that attention may dilute local features, we design a convolutional fusion module (CFM) to extract and integrate local features between adjacent patches. CFM preserves the spatial structure of local information through convolutional operations that maintain the inherent spatial relationships between neighboring patches. This module, which is embedded in an additional branch that directly processes the features after patch embedding, merges this branch into the backbone. The additional branch is designed to operate in parallel with the main transformer backbone, thereby ensuring that local feature extraction does not interfere with global context modeling. The three-branch design improves the retention of local features and facilitates the propagation of shallow local information to deep layers while avoiding attention distractions. In addition, an attention fusion enhancement module (AFEM), a partially weight-shared module, is designed to extract and fuse global features. AFEM employs a dual-path linear projection strategy where one path uses shared parameters to maintain feature consistency across modalities, whereas the other path uses independent parameters to handle modality-specific characteristics adaptively. This design effectively normalizes the feature distributions of RGB and thermal modalities before fusion, thereby addressing the challenge of heterogeneous data distributions between different sensor modalities. AFEM is directly inserted into the backbone, which can integrate the global features of deep networks and avoid affecting local features. Our method facilitates the integration of local and global features and significantly enhances model robustness through the design of additional branches and multistage fusion. The multistage fusion strategy enables progressive refinement of features across different network depths, with shallow layers focusing on local detail preservation and deep layers emphasizing global contextual understanding. **Result** Our method is evaluated on three standard RGB-T object tracking datasets, including LasHeR, RGBT210, and RGBT234. Result shows that our method achieves precision rate (PR)/success rate (SR) of 83.7%/60.6% on RGBT210, and MPR/MSR of 86.4%/83.5% on RGBT234, thereby demonstrating our method's robustness to modality-specific challenges. On LasHeR, our method (72.0%/67.7%/57.0%) achieves 2.4%/1.5% gains in PR/SR compared with the baseline (69.6%/65.9%/55.5%). We evaluate the challenge attributes on

LasHeR, and the results indicate that our method achieves optimal performance under the most challenging attributes. For instance, our method achieves a PR of 59.2% and far surpasses other methods in abrupt illumination variation, thereby demonstrating the effectiveness of our local-global feature fusion strategy. We also perform ablation experiments to demonstrate the effectiveness of the modules. Results show that our designed modules and the three-branch architecture effectively enhance performance. They validate the synergistic enhancement from our novel module ensemble. **Conclusion** In this study, we propose a three-branch multistage fusion network for RGB-T object tracking that obtains rich semantic information by extracting and fusing local and global features. Our method decouples the local and global feature extraction and processes these two features through dedicated pathways before fusion. In this way, our method effectively addresses the fundamental limitation of the existing RGB-T tracking approaches that either overemphasize local details or global context at the expense of the other. With the additional branch, the tracker effectively captures features at different stages and facilitates the propagation of local features. The three-branch architecture enables the network to maintain high-quality local details throughout the entire processing pipeline. This approach is particularly crucial for tracking small objects or objects with fine-grained textures. Experimental results show that our algorithm outperforms state-of-the-art RGB-T object tracking algorithms and effectively improves RGB-T object tracking performance. Our method achieves superior performance across multiple benchmark datasets while maintaining reasonable computational efficiency, with a frames per second (FPS) of 22.7.

Key words: object tracking; attention; convolution; multi-modal fusion; three-branch network

0 引 言

目标跟踪是计算机视觉领域的重要研究内容之一,旨在给定视频序列中连续跟踪目标,已广泛用于智能监控、机器人视觉以及无人驾驶等多个领域(胡世宇 等,2024)。早期的跟踪算法(Bunyak 等,2007;Hare 等,2016;Henriques 等,2015)偏向于根据领域知识和经验使用手工特征跟踪目标,然而这些方法的性能难以在现实场景中达到要求。自深度学习提出后,凭借其对特征强大的建模能力,逐渐取代了传统算法,成为目标跟踪领域的主流。

为了更好地适应现实场景,研究者们开始关注算法在特定挑战场景下的性能,现有数据集覆盖了各种挑战属性,同时也进一步完善了现有评价指标。尽管大部分挑战可以通过算法上的优化来应对,但仍有部分挑战属性如低照度(low illumination,LI)等由场景自身特性引发,同时基于可见光的跟踪器容易受到光照、天气等因素的影响,特别是在夜间难以提供有效信息。为了解决这些挑战,研究人员尝试将其他模态的信息与可见光模态融合,以增强算法性能。红外光具有热效应,能够反映物体的温度,且不受光照影响,能够在夜间识别目标,因此,红外光在目标跟踪领域有着广泛的应用前景。RGB-T(RGB-thermal)目标跟踪综合热红外信息和可见光信息来持续追踪目标,然而热红外图像存在噪声严重、分辨率低、目标不清晰、纹理特征不明显以及难以穿透透明物体等缺陷,相比之下,可见光虽然提供了丰富的纹理信息和色彩信息,但易受光照条件影响,如何合理地融合两个模态是当前RGB-T目标跟踪的研究重点。

早期的RGB-T目标跟踪方法以纯卷积神经网络(convolutional neural network, CNN)为主干,如ConvNet(convolutional neural network)(Li 等,2018)使用双流卷积提取特征,这种方法的本质是将多模态跟踪视为两个独立的跟踪任务,缺乏合适的融合机制导致其性能较低。SiamFT(siamese fusion tracking)(Zhang 等,2019b)虽然引入了融合,但仅限于决策级融合,在特征提取过程中两个模态并未实现充分的信息互补。DAPNet(dense feature aggregation and pruning network)(Zhu 等,2019)尝试通过特征级融合来避免单一模态的限制,充分利用多层次语义信息,但忽略了不同模态的实际贡献,并引入了更多参数。DAFNet(deep adaptive fusion network)(Gao 等,2019)与其相比则更加简洁,它通过加权相加的方式来自适应计算模态权重,这种方法能够避免不可靠模态的干扰。CNN主干并不擅长长距离建模,为了弥补这一缺陷,研究者们开始尝试将Transformer引入RGB-T目标跟踪。CANet(cross-modal attention network)(Yang 等,2021)利用Transformer辅

助CNN提取特征，通过穿插注意力计算的方式增强跨模态特征。这种混合CNN-Transformer的方法在一定程度上提升了跟踪性能，但并未充分发挥Transformer的潜力。

上述方法大多将重点放在不同模态的特征融合上，未能充分考虑模板与搜索区域的交互问题，同时CNN主干在远距离建模上的局限性限制了跟踪器的性能。研究者们开始使用纯Transformer作为主干，以便于更好地捕获图像中的关键部分，建立搜索区域与模板间的复杂特征关联。Wang等人（2023）在孪生网络中引入Transformer，以加强目标模板、搜索区域交互以及模态融合，并设计了模板更新策略，通过融合先前帧跟踪结果来传达目标外观信息，但这种方式容易加剧误差累积并导致漂移现象。TATrack（temporal adaptive RGBT tracking）（Wang等，2024）使用了双流跟踪框架、初始模板与在线模板独立进行跟踪，以避免跟踪失败时在线模板对初始模板造成干扰。ViPT（visual prompt multi-modal tracking）（Zhu等，2023）作为一个通用跟踪框架，采用提示学习的方法，通过引入极少量参数学习不同模态间的互补特征，并将辅助模态作为提示来提升跟踪性能。

基于Transformer的方法尽管在全局特征的利用上相比CNN有所改进，但却忽略了对局部特征的利用。而可见光模态包含大量纹理与细节信息，现有的算法对这些信息的提取仍不够充分。

基于上述情况，本文的目标是在保持原有全局特征的情况下，增强局部特征，并促进局部特征往深层网络的传播。考虑到现有研究对局部信息利用与融合不充分的问题，本文提出一种面向RGB-T目标跟踪的三分支多阶段融合网络（three-branch multi-stage fusion network for RGB-T object tracking，MSFT），其通过在不同的阶段选择性地融合两种特征来跟踪目标。具体而言，MSFT使用三分支结构，通过额外设计的纯CNN分支来提取局部特征，以避免注意力对细节特征的破坏。该分支通过嵌入多层卷积融合模块（convolutional fusion module，CFM）高效提取局部特征，并将其分别叠加至主干网络的不同层中，这种方法能够避免深层网络丢失细节信息。此外，全局特征融合也是必不可少的，因此设计了注意力融合增强模块（attention fusion enhancement module，AFEM），通过交叉注意力来融合全局特征。为了保持两个模态的一致性，使用两个共享参数的模块来建模全局特征，考虑到不同模态可能拥有不同的数据分布，通过额外的线性映射层来对齐特征分布，每层的注意力输出结果与原始特征叠加后进入下一层。通过在3个RGB-T跟踪数据集上的对比实验以及模块消融实验，验证了MSFT的有效性，本文提出的多阶段融合三分支网络能够有效提升跟踪性能。

1 相关工作

1.1 单目标跟踪

随着深度学习的发展，出现了一系列基于CNN的跟踪器。MDNet（multi-domain network）（Nam和Han，2016）是一个基于CNN的多域网络，这是CNN首次应用于目标跟踪领域，其主要思想是在上一帧周围选定若干个候选框，计算每个候选框为目标或背景的概率，选取目标概率最高者作为目标位置。SiamFC（siamese fully-convolutional networks）（Bertinetto等，2016）使用孪生网络来跟踪目标，其原理是使用相同结构的卷积神经网络计算模板和搜索区域的相似度作为响应值，选取响应值最大的区域作为目标区域。考虑到对局部关系的过度依赖以及对全局特征的忽视，STARK（spatio-temporal transformer network for visual tracking）（Yan等，2021）使用了一种以Transformer为关键元素的新架构，深度融合时空长距离信息，该方法通过直接估计目标角点避免了额外的后处理，极大简化了处理流程。尽管STARK仍然采用ResNet（residual network）主干网络，但对Transformer的应用被后续的研究广泛采用。SwinTrack（swin Transformer tracker）（Lin等，2022）是首个使用纯Transformer的跟踪器，它以简单的结构取得了比CNN主干更高的性能，展现了Transformer主干在跟踪任务上的潜力，但SwinTrack仍沿用孪生网络的架构，对模板和搜索区域的特征交互仍不够充分。OSTrack（one-stream tracking）（Ye等，2022）作为一个单流的目标跟踪算法，其核心是将特征学习与关系建模相统一，通过模板—搜索区域双向信息流来动态提取具有判别性的特征。但这种方法并未考虑到对局部特征的利用，模型对细节信息的理解尚不充分。上述方法大多仅关注了局部特征或全局特征，而未能综合考虑两种特征并充分融合与利用。

1.2 RGB-T目标跟踪

首个深度学习RGB-T跟踪器(Li等,2018)采用双流卷积的方式跟踪目标,但其本质是将多模态跟踪视为两个单独的跟踪任务,特征提取与融合缺乏交互。此外热红外模态由于缺乏纹理信息,使用卷积来提取局部特征效果有限。MANet(multi-adapter convolutional network)(Li等,2019b)认为对于多模态跟踪,应采取求同存异的原则,设计了使用大参数量的大核卷积通用适配器来捕获通用的融合特征,使用小参数量的小核卷积模态适配器来自适应细化区分模态独立特征,这种方式有助于避免融合过程中细节信息的丢失。然而盲目地融合特征容易导致噪声影响跟踪结果,有研究(王福田 等,2022)通过门控机制限制不可靠信息的传播,但由于融合特征仅在最终层与主干网络结合,且纯CNN的方式难以对全局特征进行建模,因此该方法对全局特征的利用、模板与搜索区域的交互以及跨模态特征的融合仍然不足。HMFT(hierarchical multi-modal fusion tracker)(Zhang等,2022)通过互补特征分支与判别特征分支,分别提取模态共同特征与模态独立特征并独立作出决策,自适应融合两个响应图。为了弥补CNN主干在远距离建模上的不足,Transformer逐渐应用到RGB-T目标跟踪。TBSI(template-bridged search region interaction)(Hui等,2023)的核心思想是利用模板作为媒介,通过收集和分发目标相关对象来桥接两个模态,这种方法能够增强模板表示,使其包含更多来自对方模态的信息。MCTrack(modalities correlation tracking)(Hu等,2024)在Transformer主干的基础上使用卷积来分别融合模板与搜索区域特征,通过分组卷积在不同模态间建立映射关系,通过参数独立的卷积调整以更好地适应各自模态。尽管加入了对卷积的利用,但深层网络的局部特征较为稀疏,导致卷积作用有限。OneTracker(Hong等,2024)是一个通用的跟踪框架,通过提示学习的方法对模态信息建模,使其能够适用于多种不同模态。BAT(bi-directional adapter for multi-modal tracking)(Cao等,2024)使用提示学习来融合模态特征,通过线性投影层将特征压缩至低维空间,在降低计算量的同时保留关键信息,但压缩特征必然会导致细节信息丢失,不利于局部特征的表达。为了解决此问题,加强对局部特征的利用与融合,本文MSFT以局部特征为出发点,在多模态特征融合的基础上,充分促进局部特征与全局特征的有效融合。

2 多阶段融合三分支网络

MSFT的总体架构如图1所示,这是一种多阶段的特征融合网络,自适应融合局部与全局特征。模型采用一种三分支结构,图中上下侧的编码器序列分别表示可见光和热红外分支,用于提取单模态特征,两者穿插AFEM进行跨模态特征融合。多个CFM单独构成一条分支,用于局部信息的提取,其分支输入为初始模板和搜索区域特征,经过逐层提取后将局部信息补充至可见光和热红外分支。

尽管输入特征已经过patch编码,但每个区块仍保留有原始的局部信息,卷积分支使用CFM来针对性提取局部特征。AFEM使用交叉注意力来提取全局特征,与常规的注意力不同,AFEM使用一个参数独立的和一个参数共享的线性层来共同映射并计算注意力,参数独立线性层能够提高模块的泛化能力,而参数共享的线性层则可以保持特征的一致性。在本研究中,CFM与AFEM的引入使模型兼具局部特征和全局特征的提取能力,有效提高了跟踪性能。

为了使各个模块的作用最大化,不同模块所插入的层数不完全相同。对于同时出现两个模块的双融合层,由于两者的侧重点不同,模块的先后顺序会影响彼此效果,因此选择并行计算的方式,将两者输出相加后再残差连接,这种方式能够避免梯度消失现象。定义$\boldsymbol{X}$表示线性特征序列,$\boldsymbol{X}_{\mathrm{CFM}}$,$\boldsymbol{X}_{\mathrm{AFEM}}$分别表示CFM和AFEM输出结果,当某一层为单融合层时,该层输出为$\boldsymbol{X}+\boldsymbol{X}_{\mathrm{CFM}}$或$\boldsymbol{X}+\boldsymbol{X}_{\mathrm{AFEM}}$;当某一层为双融合层时,输出结果为$\boldsymbol{X}+\boldsymbol{X}_{\mathrm{CFM}}+\boldsymbol{X}_{\mathrm{AFEM}}$。本研究中,在第1层仅使用CFM,在第4、7、10层同时使用CFM和AFEM。

2.1 主干网络

考虑到CNN提取的特征不包含长距离信息,不利于注意力的融合,因此选取OSTrack作为基线。假设搜索区域为$\boldsymbol{I}_x \in \mathbf{R}^{H_x \times W_x \times 3}$,模板$\boldsymbol{I}_z \in \mathbf{R}^{H_z \times W_z \times 3}$,其中,$H$与$W$分别表示图像的高与宽,下标$z$与$x$分别表示模板与搜索区域。首先将其划分为大小为$p \times p$的若干个区块,每个区块表示为$\boldsymbol{P} \in \mathbf{R}^{p \times p \times C}$,其中,$C$为通道数。随后将搜索区域与模板分别展平并拼接到一起,得到一维序列,即目标模板与搜索区域的联

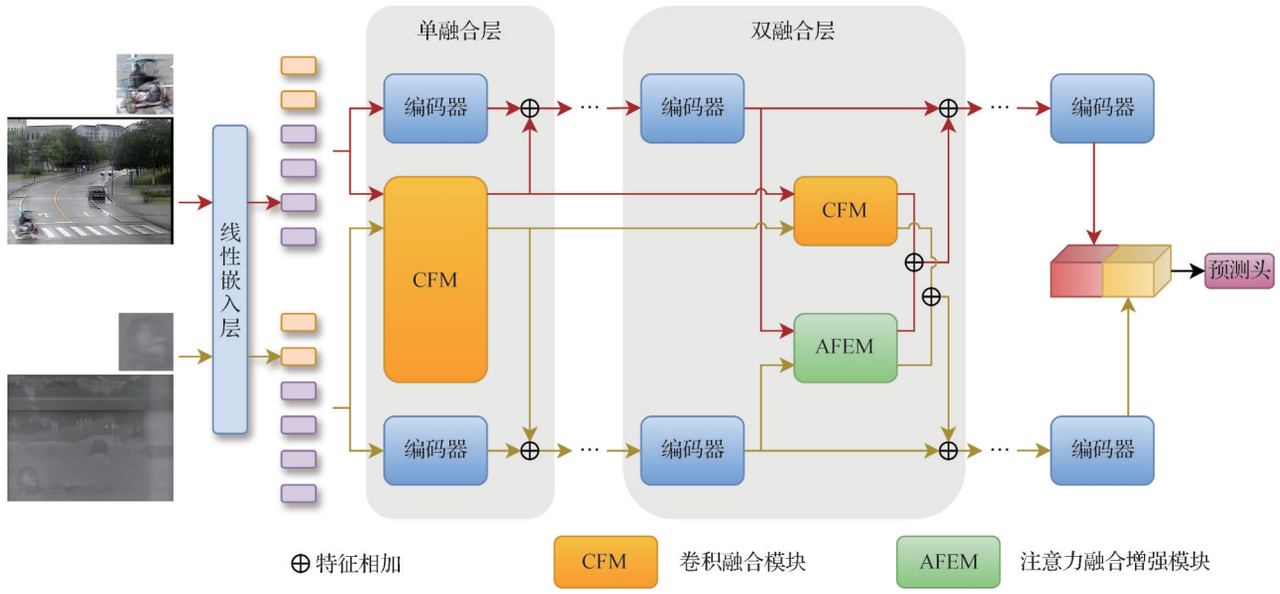

图1　MSFT框架

Fig. 1　Structure of MSFT

合特征，具体为

$$\boldsymbol{P}_z = [\boldsymbol{P}_{z,1}, \boldsymbol{P}_{z,2}, \cdots, \boldsymbol{P}_{z,N_z}] + \boldsymbol{E}$$
$$\boldsymbol{P}_x = [\boldsymbol{P}_{x,1}, \boldsymbol{P}_{x,2}, \cdots, \boldsymbol{P}_{x,N_x}] + \boldsymbol{E}$$
$$\boldsymbol{X} = [\boldsymbol{P}_z, \boldsymbol{P}_x] \tag{1}$$

式中，$N_z = (H_z \times W_z)/p^2$，$N_x = (H_x \times W_x)/p^2$，$N = N_z + N_x$，分别表示模板、搜索区域和总特征数量，$\boldsymbol{E}$表示位置编码。展平后的序列丢失了位置信息，这会增加模型学习难度，因此使用可学习的位置编码$\boldsymbol{E} \in \mathbf{R}^{N \times C}$来补充这部分信息。然后，将$\boldsymbol{X}$输入到12层编码器中提取特征，这种统一的特征提取方式可以极大地促进模板与搜索区域的交互。使用上标来表示层数，如$\boldsymbol{X}^i$表示第i层编码器的输入，$\boldsymbol{X}^{i+1}$表示第i层编码器的输出，*Encoder*表示Transformer编码器，则特征计算方式为

$$\boldsymbol{X}^{i+1} = Encoder_i(\boldsymbol{X}^i), 1 \leqslant i < l \tag{2}$$

MSFT将ViT(vision Transformer)主干应用于可见光和热红外图像，使用两个参数共享的分支分别提取两种特征，分别用$\boldsymbol{X}_r^l$和$\boldsymbol{X}_t^l$表示可见光和热红外分支的输出结果，其中下标r表示可见光，t表示热红外。从最后一层的输出结果$\boldsymbol{X}_r^l$, $\boldsymbol{X}_t^l$中分离出搜索区域$\boldsymbol{P}_{xr}$, $\boldsymbol{P}_{xt} \in \mathbf{R}^{N_x \times C}$，将其二维化后得到搜索区域特征$\boldsymbol{L}_r^l, \boldsymbol{L}_t^l \in \mathbf{R}^{\frac{H_x}{p} \times \frac{W_x}{p} \times C}$，其中下标$r$与$t$分别表示可见光与热红外模态。使用卷积拼接后得到最终特征，并输入预测头得到跟踪预测坐标$\boldsymbol{O}$，具体为

$$\boldsymbol{O} = Head(Conv([\boldsymbol{L}_r^l, \boldsymbol{L}_t^l])) \tag{3}$$

式中，*Head*表示预测头。

2.2　卷积融合模块

常规的卷积融合方式是将特征二维化后使用卷积提取局部特征。但随着编码器层数的加深，特征逐渐由浅层局部特征转变为深层抽象特征。为了促进局部特征的表达，设计了一种将局部特征传递至深层的方式，即将局部特征通过单独的分支传递至深层网络。该分支使用CFM来融合局部特征，其具体结构如图2所示。

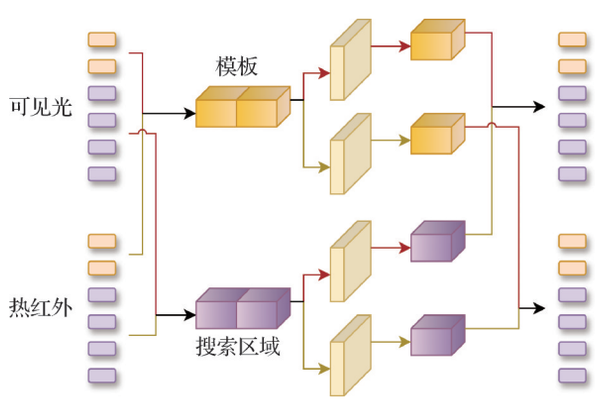

图2　卷积融合模块结构

Fig. 2　Structure of CFM

首先将输入的特征二维化，得到$\boldsymbol{L}_{zr}, \boldsymbol{L}_{zt} \in \mathbf{R}^{\frac{H_z}{p} \times \frac{W_z}{p} \times C}$，$\boldsymbol{L}_{xr}, \boldsymbol{L}_{xt} \in \mathbf{R}^{\frac{H_x}{p} \times \frac{W_x}{p} \times C}$，以便于卷积对特征进行提取，然后将不同模态的模板与搜索区域分别拼接到一起，得到统一特征。随后使用3×3卷

积来压缩通道数，在压缩过程中卷积将能够自适应地选取重要特征，同时较小的感受野使得模型能够更加专注于局部信息。为了使融合特征能够更好地适应各个模态，使用模态独立的卷积将特征映射至两个模态，具体为

$$\begin{cases} \boldsymbol{L}_{zr} = Conv_{zr}(\boldsymbol{L}_z) \\ \boldsymbol{L}_{zt} = Conv_{zt}(\boldsymbol{L}_z) \\ \boldsymbol{L}_{xr} = Conv_{xr}(\boldsymbol{L}_x) \\ \boldsymbol{L}_{xt} = Conv_{xt}(\boldsymbol{L}_x) \end{cases} \tag{4}$$

将来自两个模态的模板与搜索区域特征相互拼接，得到 $\boldsymbol{X}_r = [\boldsymbol{R}_{zr}, \boldsymbol{R}_{xr}]$，$\boldsymbol{X}_t = [\boldsymbol{R}_{zt}, \boldsymbol{R}_{xt}]$即为CFM的输出。通过将 $\boldsymbol{X}_r$，$\boldsymbol{X}_t$加到主干网络上，使得局部特征可以在深层网络中传播，避免细节信息的丢失。

由于编码器将特征抽象化后会削弱细节信息，因此直接使用 $\boldsymbol{X}^1$ 作为第1层CFM的输入，以获得更丰富的局部特征。为了使CFM能够专注于局部特征的提取，将其作为一个单独的纯卷积分支，并将输出结果以跨层叠加的方式添加到主干网络中，以避免注意力对局部特征的干扰。MSFT使用多个CFM来提取特征，每个CFM的输出既被叠加到主干网络中，又作为下一个CFM的输入，这种多层串联的方式有效扩展了卷积的感受野。

2.3 注意力融合增强模块

除局部信息以外，长距离融合在模型中也是必不可少的。注意力融合增强模块的具体结构如图3所示。

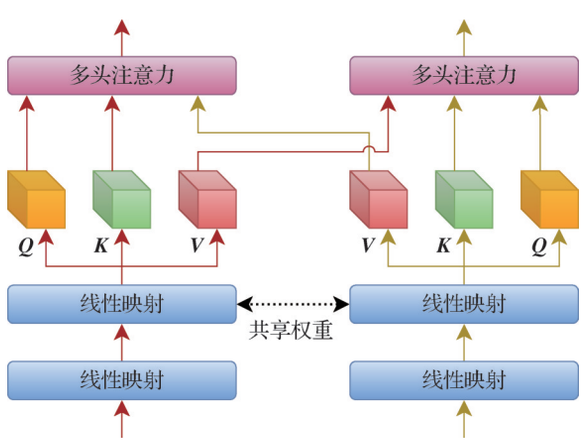

图3 注意力融合增强模块结构图

Fig. 3 Structure of AFEM

与CFM不同，AFEM的作用在于充分发挥注意力的全局建模能力。首先，通过线性层分别计算得到 $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{Q}$，$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{K}$，$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{V}$，式中，$\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$，$\boldsymbol{W}^{V}$为可学习的参数矩阵。随后使用交叉注意力计算式代入计算，具体为

$$\boldsymbol{A} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

式中，d是特征维度，$\boldsymbol{A}$表示注意力计算结果。为了促进模态交互，使用交叉注意力来促进模态融合，即每个模态的 $\boldsymbol{V}$ 均来自于另一模态。通过共享参数使两个分支的侧重点保持一致，以避免两个模态的特征图出现显著偏差。为了使模块更具泛化性，额外使用两个参数独立线性层处理双模态输入，将其映射为特征向量后再进行注意力计算。其目的是将模态独立数据映射为同一分布，以便于共享参数的注意力机制提取特征。

2.4 预测头与损失函数

MSFT采用与OSTrack类似的预测头，首先将搜索区域特征输入到全卷积网络中，得到目标得分图，将得分最高的位置选定为目标响应区域。

本文方法采用的损失函数为以下3种：

1）分类损失。预测每个位置属于前景或背景的概率，使用加权焦点损失计算分类损失以缓解前景和背景样本不平衡问题。

2）广义IoU损失。常规交并比（intersection over union，IoU）损失计算预测框与标注框的覆盖率并作为损失函数，在预测框与标注框完全不重叠时存在梯度消失问题，广义IoU损失可以表示为 $GIoU = 1 - IoU$，使得IoU为0时仍然可以有效指导梯度下降，有助于在预测框与标注框重叠部分较小时更有效地优化边界。

3）L1损失。将预测坐标与标注坐标进行归一化，计算L1距离并作为损失。

为综合以上3种损失函数，采用目前常见的损失函数组合，具体为

$$L = L_{\mathrm{cls}} + \lambda_{\mathrm{iou}} L_{\mathrm{iou}} + \lambda_{\mathrm{L1}} L_{\mathrm{L1}} \tag{6}$$

式中，L_{cls}为分类损失，L_{iou}为广义IoU损失，用于衡量定位精度，L_{L1}为L1损失，λ_{iou}和λ_{L1}为超参数。

3 实 验

3.1 实验细节

模型使用PyTorch实现，其中Python版本为3.10.6，Torch版本为2.1.0。训练在8张Hygon

Z100 DCU上进行，单张显卡的batchsize为16，优化器为AdamW。使用LasHeR作为训练数据集，每轮共包含60 K个视频帧。

MSFT采用OSTrack作为基线算法，其中模板大小设置为128 × 128像素，搜索区域大小为256 × 256像素，patch大小为16 × 16。CFM同时对模板与搜索区域进行融合，其中，卷积核大小为3，padding为1。

为了使模块更好地学习特征建模，使用分段训练的方法。在第1阶段，仅在第4、7、10层添加AFEM来学习全局特征，训练轮次为20轮；在第2阶段，在第1、4、7、10层添加CFM以学习局部特征，训练轮次为15轮。每阶段训练开始时学习率均设置为0.000 1，并在第10轮后降低至原本的10%。

3.2 数据集与评估指标

使用3个权威的数据集来验证MSFT的有效性：RGBT210（Li等，2017）、RGBT234（Li等，2019a）和LasHeR（Li等，2022）。

1）RGBT210数据集使用两个相同参数的相机拍摄，并在拍摄前进行了对齐操作，使得数据集无须人工处理就可以拥有较高的准度。数据集共包含210个视频对，总帧数为104.7 K，其中单视频最大8 K帧。2019年，又在此基础上扩展至234个视频对，得到RGBT234数据集，总帧数116.7 K。

2）RGBT234构建了多个挑战属性，如部分遮挡（partial occlusion，PO）、热交叉（thermal crossover，TC）、相机移动（camera moving，CM）等。

3）LasHeR是一个具有里程碑意义的大规模RGB-T跟踪数据集，包含了总帧数达730 K的1 224个视频对，每一帧都经过空间对齐和手动标注，确保了数据集的精确性。在挑战属性上，LasHeR细化并提出了更多挑战属性，分别为：无遮挡（no occlusion，NO）、部分遮挡、完全遮挡（total occlusion，TO）、透明遮挡（hyaline occlusion，HO）、目标形变（deformation，DEF）、背景杂乱（background clutter，BC）、运动模糊（motion blur，MB）、快速运动（fast motion，FM）、热交叉、低照度、高照度（high illumination，HI）、帧丢失（frame lost，FL）、移出视野（out of view，OV）、低分辨率（low resolution，LR）、光照突变（abrupt illumination variation，AIV）、相机移动、外观相似（similar appearance，SA）、尺度变化（scale variation，SV）和横纵比变化（aspect ratio change，ARC）。其广泛的对象类别、多变的跟踪场景，以及各种跨季节、天气和昼夜的环境因素体现出LasHeR的多样性，使得LasHeR能够更加全面地模拟现实场景。

精确率（precision rate，PR）指算法预测的目标框与标注目标框的中心点之间的欧氏距离低于某个阈值的帧数的百分比，它用于评价定位精度。为了避免图像大小的影响，LasHeR数据集额外增加归一化精确率（normalized precision rate，NPR），将预测框与标注框归一化之后再计算，以更加公平地评价跟踪性能。成功率（success rate，SR）反映了目标框与标注框的重叠率，主要用于评价预测的目标框是否准确地框选出物体。

RGBT234数据集使用最大精确率（maximum precision rate，MPR）与最大成功率（maximum success rate，MSR）作为评估指标，旨在解决模态未对齐导致的结果偏差问题。数据集对两个模态的目标位置均做了标注，在计算MPR与MSR时，分别选取两个模态欧氏距离更小者和重叠率更大者作为跟踪结果，其余计算方式不变。

3.3 实验结果

3.3.1 LasHeR评估结果

为验证MSFT的有效性，在LasHeR数据集上进行了评估，并选取了一些具有代表性的跟踪器进行比较，包括DAPNet（Zhu等，2019）、DAFNet（Gao等，2019）、MANet（Li等，2019b）、CAT（challenge-aware RGBT tracker）（Li等，2020）、HMFT（Zhang等，2022）、APFNet（attribute-based progressive fusion network）（Xiao等，2022）、ProTrack（multi-modal prompt tracker）（Yang等，2022）、CMD（cross-modality distillation）（Zhang等，2023）、ViPT（Zhu等，2023）、TBSI（Hui等，2023）、TBSI-Tiny（Hui等，2023）、SiamTFA（siamese triple-stream feature aggregation network）（Zhang等，2025）、SDSTrack（self-distillation symmetric adapter tracking）（Hou等，2024）、OneTracker（Hong等，2024）、BAT（Cao等，2024）、TATrack（Wang等，2024）、LightFC-T（lightweight full-convolutional siamese tracker）（Li等，2025）和CAFormer（cross-modulated attention transformer）（Xiao等，2025）。表1展示了各项算法在LasHeR测试集上的评估结果，实验结果表明MSFT取得了领先的跟踪性能，其PR、NPR和SR分别为72.0%、67.7%和57.0%，比TATrack分别高出1.8%、1.0%和0.9%，这说明了MSFT具有较大的性能优势。可以注意到，基于

表1　数据集LasHeR上的跟踪结果比较

Table 1　Comparison of tracking results on LasHeR dataset

方法	出处	主干网络	PR/%	NPR/%	SR/%	FPS/(帧/s)
DAPNet(Zhu等,2019)	ACM MM 2019	VGG-M	43.1	38.3	31.4	2
DAFNet(Gao等,2019)	ICCV 2019	VGG-M	44.8	39.0	31.1	23
MANet(Li等,2019b)	ICCV 2019	VGG-M	45.5	38.3	32.6	1.11
CAT(Li等,2020)	ECCV 2020	VGG-M	45.0	39.5	31.4	20
HMFT(Zhang等,2022)	CVPR 2022	ResNet-50	43.6	38.1	31.3	30.2
APFNet(Xiao等,2022)	AAAI 2022	VGG-M	50.0	53.9	36.2	1.9
ProTrack(Yang等,2022)	ACM MM 2022	ResNet-50	53.8	49.8	42.0	–
CMD(Zhang等,2023)	CVPR 2023	ResNet-50	59.0	54.6	46.4	30
ViPT(Zhu等,2023)	CVPR 2023	ViT	65.1	61.7	52.5	–
TBSI(Hui等,2023)	CVPR 2023	ViT	69.2	65.7	55.6	36.2
TBSI-Tiny(Hui等,2023)	CVPR 2023	ViT	61.7	57.8	48.9	40.3
SiamTFA(Zhang等,2025)	T-ITS 2024	SwinT	62.5	–	48.1	37
SDSTrack(Hou等,2024)	CVPR 2024	ViT	66.5	–	53.1	20.86
OneTracker(Hong等,2024)	CVPR 2024	ViT	67.2	–	53.8	–
BAT(Cao等,2024)	AAAI 2024	ViT	70.2	–	56.3	–
TATrack(Wang等,2024)	AAAI 2024	ViT	70.2	66.7	56.1	26.1
LightFC-T(Li等,2025)	arXiv 2025	TinyViT	63.8	59.9	50.1	**81**
CAFormer(Xiao等,2025)	AAAI 2025	ViT	70.0	66.1	55.6	–
本文	中国图象图形学报	ViT	**72.0**	**67.7**	**57.0**	22.7

注:加粗字体表示各列最优结果,“–”表示该指标结果未提供。

CNN主干的网络,如APFNet(Xiao等,2022)、CAT(Li等,2020)、DAFNet(Gao等,2019)等性能差距与ViT主干较为明显,这是CNN主干缺乏模板与搜索区域交互导致的,这也印证了使用Transformer架构的优越性。与另一些基于ViT主干的网络如CAFormer(Xiao等,2025)、TATrack(Wang等,2024)、OneTracker(Hong等,2024)等方法对比,尽管这些方法引入了更加复杂的模态融合或时序特征模块,如CAFormer设计了模态相关的特征交互与搜索模块,TATrack引入了初始分支与在线分支,但这些方法往往局限于对全局特征的提取,并未涉及对局部信息的利用。SDSTrack(Hou等,2024)与BAT(Cao等,2024)则通过微调或提示器的方式补充不可靠模态特征,但并未改变特征提取能力上的不足。而MSFT通过三分支网络,增强了模型的特征提取能力,并使用局部特征补充深层高级语义特征,从而达到更好的性能。

此外,表1中还展示了MSFT与LightFC-T(Li等,2025)、TBSI-Tiny(Hui等,2023)和SiamTFA(Zhang等,2025)等轻量级方法的对比。这些方法通常通过简洁的模块设计来达到更高的效率,但这也导致特征交互不充分。尽管MSFT的效率并未达到最优,但性能上存在较大幅度的领先,达到了领先水平。

图4展示了MSFT和另外4个跟踪器在LasHeR上的各项挑战属性的评估结果。其中各项数字分别代表各个挑战属性的最低PR与最高PR。MSFT在大部分挑战属性下较其他跟踪器均有提升,如相机移动、热交叉等。其中MSFT在光照突变与透明遮挡属性下的提升最为明显,这是因为MSFT更加侧重对局部特征与全局特征的融合,在单一模态缺陷时,该方法能够更好地发掘可靠模态中的细节信息并辅助决策,提升该场景下的跟踪性能。

3.3.2　RGBT210与RGBT234评估结果

表2是MSFT在RGBT210数据集上的跟踪结

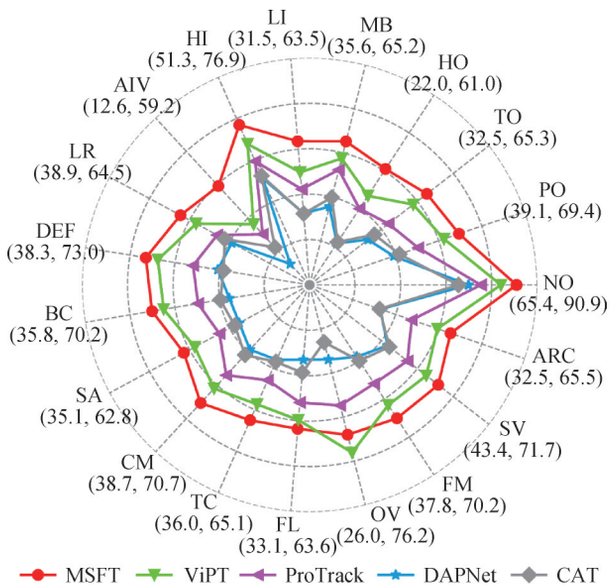

图4　LasHeR上的19个挑战属性的评估结果

Fig. 4　Evaluation results of 19 challenge attributes on LasHeR dataset

果，对比方法包括DAFNet（Gao等，2019）、TFNet（trident fusion network）（Zhu等，2022）、mfDiMP（multi-modal fusion discriminative model prediction）（Zhang等，2019a）、CAT（Li等，2020）、DMCNet（duality-gated mutual condition network）（Lu等，2025）和CAT++（Liu等，2024），表明了该模型在其他数据集上也能保持优越的性能，具有一定的泛化性。

表2　数据集RGBT210上的跟踪结果比较

Table 2　Comparison of tracking results on RGBT210 dataset

方法	PR/%	SR/%
DAFNet（Gao等，2019）	72.6	48.5
TFNet（Zhu等，2022）	77.7	52.9
mfDiMP（Zhang等，2019a）	78.6	55.5
CAT（Li等，2020）	79.2	53.3
DMCNet（Lu等，2025）	79.7	55.5
CAT++（Liu等，2024）	82.2	56.1
本文	**83.7**	**60.6**

注：加粗字体表示各列最优结果。

表3展示了RGBT234数据集下的评估结果，对比方法包括mfDiMP（Zhang等，2019a）、FANet（feature aggregation network）（Zhu等，2021）、JMMAC（jointly modeling motion and appearance cues）（Zhang等，2021）、ProTrack（Yang等，2022）、DAFNet（Gao等，2019）、CAT（Li等，2020）、APFNet（Xiao等，2022）、ViPT（Zhu等，2023）、CAT++（Liu等，2024）、SDSTrack（Hou等，2024）、OneTracker（Hong等，2024）和TATrack（Wang等，2024）。MSFT的MPR与MSR分别为86.4%和63.5%，两者均略微低于利用时序信息的TATrack，但仍优于大部分跟踪器。

表3　数据集RGBT234上的跟踪结果比较

Table 3　Comparison of tracking results on RGBT234 dataset

方法	MPR/%	MSR/%
mfDiMP（Zhang等，2019a）	64.6	42.8
FANet（Zhu等，2021）	78.7	55.3
JMMAC（Zhang等，2021）	79.0	57.3
ProTrack（Yang等，2022）	79.5	59.9
DAFNet（Gao等，2019）	79.6	54.4
CAT（Li等，2020）	80.4	56.1
APFNet（Xiao等，2022）	82.7	57.9
ViPT（Zhu等，2023）	83.5	61.7
CAT++（Liu等，2024）	84.0	59.2
SDSTrack（Hou等，2024）	84.8	62.5
OneTracker（Hong等，2024）	85.7	64.2
TATrack（Wang等，2024）	**87.2**	**64.4**
本文	86.4	63.5

注：加粗字体表示各列最优结果。

3.4　消融实验

为了验证所提模块的有效性，选取LasHeR数据集进行消融实验，具体实验设计如表4所示。为了公平起见，消融实验中将严格按照MSFT的训练方式重新训练。实验结果表明，MSFT相比基线算法提升了2.4%，同时各个模块在单独使用时也能带来性能提升，其中AFEM带来的提升更明显，这是因为AFEM融合了更具判别性的全局特征，更有利于跟踪决策。与基线相比，MSFT的参数量较高，大部分参数主要集中于额外分支。尽管如此，计算量的增长幅度并不大，因此在计算效率上仍然能够保持较好的水平。同时MSFT的性能提升较现有方法更为明显，略微的效率损失在实际应用中是可接受的。

特别地，针对CFM进行了更深入的消融研究，来探讨融合方式造成的影响。通过对CFM中模板

表4 MSFT模块消融

Table 4 Modules ablation of MSFT

方法	PR/%	NPR/%	SR/%	Params/M	FLOPs/G
Baseline(基线)	69.6	65.9	55.5	**85.6**	**54.8**
CFM	70.3	66.1	55.7	255.6	82.0
AFEM	71.2	67.2	56.6	94.5	60.5
CFM+AFEM	**72.0**	**67.7**	**57.0**	264.4	87.6

注:加粗字体表示各列最优结果。

和搜索区域融合方式的修改,设计实验来验证此模块对两者进行融合的必要性。在本消融实验里,只修改了卷积的融合方式,仅融合模板或搜索区域,除此之外与完整模型保持一致。

表5展示了实验结果。可以看出,无论使用哪种方式,仅融合单一部分都会导致性能显著下降,这是因为Transformer将模板与搜索区域视为统一特征来处理,而CFM仅对其中一部分信息进行融合,破坏了数据的原始分布状态。此外,实验还探究了额外分支对性能的影响,从表中可以看出,在没有额外分支的情况下,性能会明显降低,这也说明了注意力会稀释局部特征,不利于CFM对局部特征的融合,因而使用额外的分支来提取并传播局部信息是有必要的。

表5 CFM的融合方式对比

Table 5 Comparison of the fusion approach to CFM

融合方式	三分支	PR/%	NPR/%	SR/%
仅模板	√	70.6	66.6	56.1
仅搜索区域	√	70.5	66.6	56.1
模板+搜索区域	–	70.3	66.6	56.0
模板+搜索区域	√	**72.0**	**67.7**	**57.0**

注:加粗字体表示各列最优结果,“√”和“–”表示启用和未启用三分支结构。

为了探索模块的最佳位置,设计实验以评估模块在不同层插入的性能。表6和表7展示了修改CFM和AFEM层数之后的性能。注意到当CFM的首层非第1层时,跟踪器性能会大幅度降低,仅略高于基线,这是因为局部信息主要存在于神经网络的浅层,因此在浅层融合局部特征对性能的提升最为明显。同时层数过多时也会导致性能下降,这是因为随着层数的加深,局部特征逐渐转变为抽象的全局特征,网络的深层特别是末尾层主要依靠语义特征进行决策,局部特征的作用较小,强行引入局部特征会影响全局特征的表达,从而影响跟踪结果。

由于AFEM的注重点并非局部信息,因此不适合应用于包含大量局部信息的浅层,实验结果表明其更适合应用于中间层。与CFM相同,过多的层数也会导致性能下降,这是因为模型过多地融合多模态特征导致低质模态干扰了另一模态的表达,模态信息仅需要少部分层的融合即可有效传递。为了探究MSFT在面对模态缺陷时的跟踪性能,展开了单一模态缺陷实验,如表8所示。使用LasHeR作为测试集,并将缺陷模态特征置0。实验结果表明,无论是何种模态缺陷,都会导致跟踪性能严重下降,其中可见光模态缺陷时性能下降更明显。其原因是MSFT致力于对细节特征进行融合,而可见光模态包含了更丰富的细节信息。当可见光模态缺陷时,模型难以提取到有效的细节特征,同时缺陷模态在特征融合阶段进一步干扰可靠模态的特征表达,导致跟踪性能受到较大影响。

表6 CFM插入的层数

Table 6 Inserting layers of the CFM

CFM层数	PR/%	NPR/%	SR/%
4, 7, 10	69.8	65.8	55.7
1, 4, 7, 10	**72.0**	**67.7**	**57.0**
1, 4, 7, 10, 12	70.3	66.5	55.9

注:加粗字体表示各列最优结果。

表7 AFEM插入的层数

Table 7 Inserting layers of the AFEM

AFEM层数	PR/%	NPR/%	SR/%
4, 7, 10	**72.0**	**67.7**	**57.0**
1, 4, 7, 10	71.3	67.3	56.7
1, 4, 7, 10, 12	70.4	66.5	56.1

注:加粗字体表示各列最优结果。

表8 模态缺陷实验

Table 8 Modal defects experiments

缺陷模态	PR/%	NPR/%	SR/%
无	**72.0**	**67.7**	**57.0**
RGB	45.9	41.9	37.6
TIR	58.2	54.5	47.0

注:加粗字体表示各列最优结果。

3.5 可视化

图5将MSFT与4个跟踪模型进行了比较，展示了部分场景下的跟踪结果。“rightdarksingleman 877”序列的跟踪难点主要在于小目标，MSFT已针对局部特征进行优化，因此在类似场景下拥有更高的性能，而其余的跟踪器均完全丢失目标。“bikeboyintodark 623”序列中曝光现象较为严重，ViPT与BAT对目标大小的预测存在较大偏差，MSFT能够较准确地预测出目标大小并跟踪目标。“midboyNo_9 1888”序列为多人打篮球视频，存在大量热交叉现象。目标人物的穿着为黑色衣物，而其余跟踪器却将预测框定位至白色衣服人物身上，这是因为模型忽视了可见光模态中的局部信息，转而选取了包含更多显著性特征的热红外模态。而MSFT准确捕获了衣物细节特征，取得较好的跟踪结果。以上跟踪结果验证了本文方法能够充分促进局部信息的提取与融合，能够在挑战场景下有效提升跟踪性能。

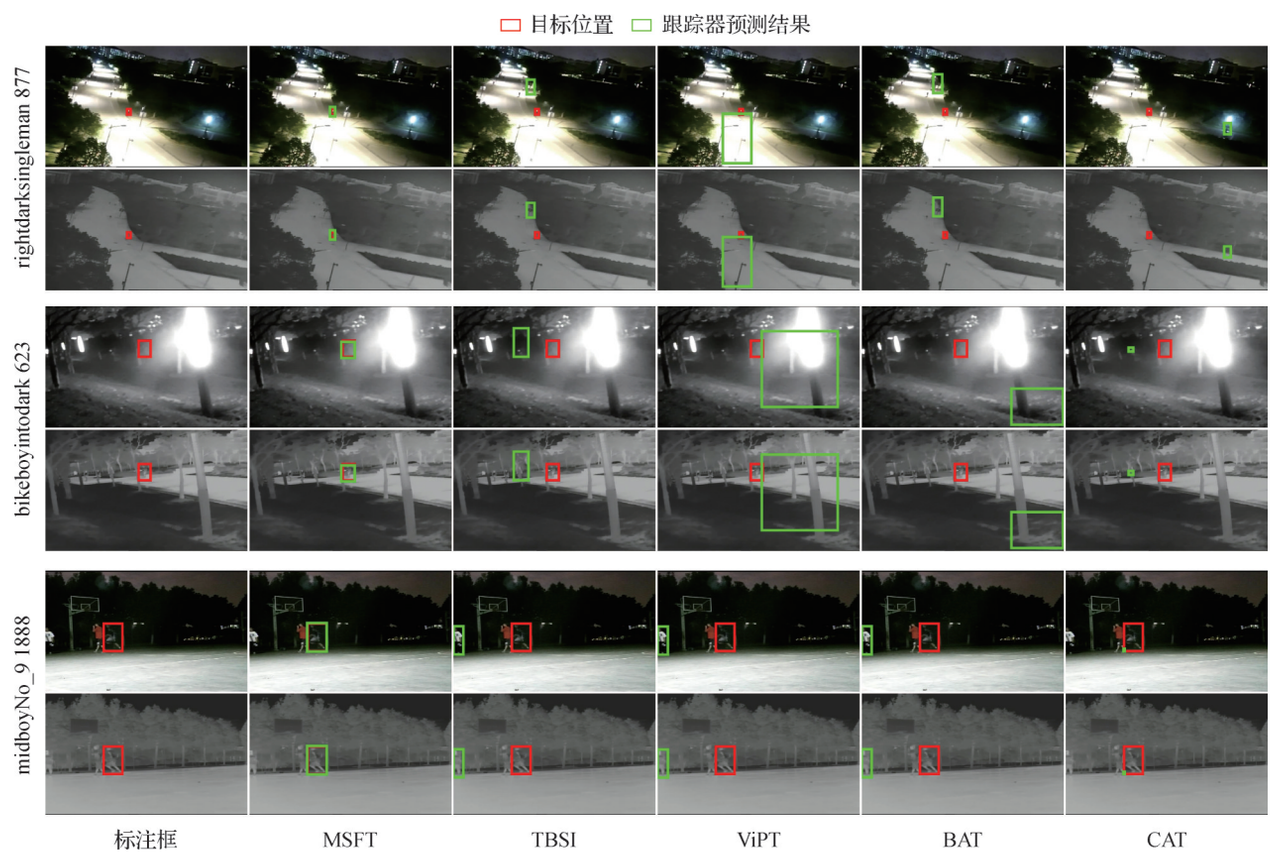

图5 LasHeR数据集上的跟踪结果比较

Fig. 5 Comparison of tracking results on LasHeR dataset

4 结 论

本文提出一种新颖的面向RGB-T目标跟踪的三分支多阶段融合网络，使用单独的卷积分支来集中捕获局部信息，并在不同阶段采取不同的融合策略来增强特征，有效提升了跟踪性能。通过CFM增强与融合局部特征，为深层网络补充局部信息，同时设计了AFEM来融合全局特征，进一步促进了模态的融合。在LasHeR、RGBT210和RGBT234上进行的实验结果表明MSFT已取得较好的跟踪性能。消融实验表明MSFT能够有效避免深层网络丢失局部信息，并广泛促进局部与全局特征融合。

本研究尚存在一些局限性，未来将探索新的训练方式与模块设计，增强泛化性，提高跟踪速度。此外，本研究使用patch编码后的特征作为局部信息的来源，尽管并未进行大范围特征交互，但单个区块内的纹理细节信息仍有所稀释，未来将尝试频域增强与融合的方法保留更多纹理信息。

参考文献(Reference)

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of 2016 European Conference on Computer Vision Workshops. Amsterdam, the Netherlands: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]

Bunyak F, Palaniappan K, Nath S K and Seetharaman G. 2007. Geodesic active contour based fusion of visible and infrared video for persistent object tracking//Proceedings of 2007 IEEE Workshop on Applications of Computer Vision. Austin, USA: IEEE: #35 [DOI: 10.1109/wacv.2007.26]

Cao B, Guo J L, Zhu P F and Hu Q H. 2024. Bi-directional adapter for multimodal tracking//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 927-935 [DOI: 10.1609/aaai.v38i2.27852]

Gao Y, Li C L, Zhu Y B, Tang J, He T and Wang F T. 2019. Deep adaptive fusion network for high performance RGBT tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea (South): IEEE: 91-99 [DOI: 10.1109/iccvw.2019.00017]

Hare S, Golodetz S, Saffari A, Vineet V, Cheng M M, Hicks S L and Torr P H S. 2016. Struck: structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2096-2109 [DOI: 10.1109/TPAMI.2015.2509974]

Henriques J F, Caseiro R, Martins P and Batista J. 2015. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3): 583-596 [DOI: 10.1109/tpami.2014.2345390]

Hong L Y, Yan S L, Zhang R R, Li W Y, Zhou X Y, Guo P X, et al. 2024. OneTracker: unifying visual object tracking with foundation models and efficient tuning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19079-19091 [DOI: 10.1109/cvpr52733.2024.01805]

Hou X J, Xing J Z, Qian Y J, Guo Y W, Xin S, Chen J H, et al. 2024. SDSTrack: self-distillation symmetric adapter learning for multi-modal visual object tracking//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26541-26551 [DOI: 10.1109/cvpr52733.2024.02507]

Hu S Y, Zhao X and Huang K Q. 2024. Visual intelligence evaluation techniques for single object tracking: a survey. Journal of Image and Graphics, 29(8): 2269-2302 (胡世宇, 赵鑫, 黄凯奇. 2024. 单目标跟踪中的视觉智能评估技术综述. 中国图象图形学报, 29(8): 2269-2302) [DOI: 10.11834/jig.230498]

Hu X T, Zhong B N, Liang Q H, Zhang S P, Li N and Li X X. 2024. Toward modalities correlation for RGB-T tracking. IEEE Transactions on Circuits and Systems for Video Technology, 34(10): 9102-9111 [DOI: 10.1109/tcsvt.2024.3396289]

Hui T R, Xun Z Z, Peng F G, Huang J S, Wei X M, Wei X L, et al. 2023. Bridging search region interaction with template for RGB-T tracking//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13630-13639 [DOI: 10.1109/cvpr52729.2023.01310]

Li C L, Liang X Y, Lu Y J, Zhao N and Tang J. 2019a. RGB-T object tracking: benchmark and baseline. Pattern Recognition, 96: #106977 [DOI: 10.1016/j.patcog.2019.106977]

Li C L, Liu L, Lu A D, Ji Q and Tang J. 2020. Challenge-aware RGBT tracking//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 222-237 [DOI: 10.1007/978-3-030-58542-6_14]

Li C L, Lu A D, Zheng A H, Tu Z Z and Tang J. 2019b. Multi-adapter RGBT tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea (South): IEEE: 2262-2270 [DOI: 10.1109/iccvw.2019.00279]

Li C L, Wu X H, Zhao N, Cao X C and Tang J. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. Neurocomputing, 281: 78-85 [DOI: 10.1016/j.neucom.2017.11.068]

Li C L, Xue W L, Jia Y Q, Qu Z C, Luo B, Tang J, et al. 2022. LasHeR: a large-scale high-diversity benchmark for RGBT tracking. IEEE Transactions on Image Processing, 31: 392-404 [DOI: 10.1109/tip.2021.3130533]

Li C L, Zhao N, Lu Y J, Zhu C L and Tang J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 1856-1864 [DOI: 10.1145/3123266.3123289]

Li Y F, Wang B and Li Y. 2025. LightFC-X: lightweight convolutional tracker for RGB-x tracking [EB/OL]. [2025-06-17]. https://arxiv.org/pdf/2502.18143.pdf

Lin L T, Fan H, Zhang Z P, Xu Y and Ling H B. 2022. SwinTrack: a simple and strong baseline for transformer tracking//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1218

Liu L, Li C L, Xiao Y, Ruan R and Fan M H. 2024. RGBT tracking via challenge-based appearance disentanglement and interaction. IEEE Transactions on Image Processing, 33: 1753-1767 [DOI: 10.1109/tip.2024.3371355]

Lu A D, Qian C, Li C L, Tang J and Wang L. 2025. Duality-gated mutual condition network for RGBT tracking. IEEE Transactions on Neural Networks and Learning Systems, 36(3): 4118-4131 [DOI: 10.1109/tnnls.2022.3157594]

Nam H and Han B. 2016. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4293-4302 [DOI: 10.1109/cvpr.2016.465]

Wang F T, Wang W Q, Liu L, Li C L and Tang J. 2023. Siamese trans-

former RGBT tracking. Applied Intelligence, 53(21): 24709-24723[DOI: 10.1007/s10489-023-04741-y]

Wang F T, Zhang S Y, Li C L and Luo B. 2022. RGBT tracking based on dynamic modal interaction and adaptive feature fusion. Journal of Image and Graphics, 27(10): 3010-3021(王福田, 张淑云, 李成龙, 罗斌. 2022. 动态模态交互和特征自适应融合的RGBT跟踪. 中国图象图形学报, 27(10): 3010-3021)[DOI: 10.11834/jig.210287]

Wang H Y, Liu X T, Li Y F, Sun M, Yuan D and Liu J. 2024. Temporal adaptive RGBT tracking with modality prompt//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 5436-5444[DOI: 10.1609/aaai.v38i6.28352]

Xiao Y, Yang M M, Li C L, Liu L and Tang J. 2022. Attribute-based progressive fusion network for RGBT tracking//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 2831-2838[DOI: 10.1609/aaai.v36i3.20187]

Xiao Y, Zhao J C, Lu A D, Li C L, Yin B, Lin Y, et al. 2025. Cross-modulated attention transformer for RGBT tracking//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 8682-8690[DOI: 10.1609/aaai.v39i8.32938]

Yan B, Peng H W, Fu J L, Wang D and Lu H C. 2021. Learning spatio-temporal transformer for visual tracking//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10428-10437[DOI: 10.1109/iccv48922.2021.01028]

Yang J Y, Li Z, Zheng F, Leonardis A and Song J K. 2022. Prompting for multi-modal tracking//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 3492-3500[DOI: 10.1145/3503161.3547851]

Yang Y, Liang H, Yang Y and Feng T. 2021. Cross-modal attention network for RGB-T tracking//Proceedings of the 3rd International Conference on Advances in Computer Technology, Information Science and Communication. Shanghai, China: IEEE: 341-346[DOI: 10.1109/ctisc52352.2021.00068]

Ye B T, Chang H, Ma B P, Shan S G and Chen X L. 2022. Joint feature learning and relation modeling for tracking: a one-stream framework//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 341-357[DOI: 10.1007/978-3-031-20047-2_20]

Zhang J M, Qin Y, Fan S M, Xiao Z and Zhang J. 2025. SiamTFA: siamese triple-stream feature aggregation network for efficient RGBT tracking. IEEE Transactions on Intelligent Transportation Systems, 26(2): 1900-1913[DOI: 10.1109/TITS.2024.3512551]

Zhang L C, Danelljan M, Gonzalez-Garcia A, Van De Weijer J and Shahbaz Khan F. 2019a. Multi-modal fusion for end-to-end RGB-T tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea (South): IEEE: 2252-2261[DOI: 10.1109/iccvw.2019.00278]

Zhang P Y, Zhao J, Bo C J, Wang D, Lu H C and Yang X Y. 2021. Jointly modeling motion and appearance cues for robust RGB-T tracking. IEEE Transactions on Image Processing, 30: 3335-3347[DOI: 10.1109/tip.2021.3060862]

Zhang P Y, Zhao J, Wang D, Lu H C and Ruan X. 2022. Visible-thermal UAV tracking: a large-scale benchmark and new baseline//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8876-8885[DOI: 10.1109/cvpr52688.2022.00868]

Zhang T L, Guo H Y, Jiao Q, Zhang Q and Han J G. 2023. Efficient RGB-T tracking via cross-modality distillation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5404-5413[DOI: 10.1109/cvpr52729.2023.00523]

Zhang X C, Ye P, Peng S Y, Liu J, Gong K and Xiao G. 2019b. SiamFT: an RGB-infrared fusion tracking method via fully convolutional siamese networks. IEEE Access, 7: 122122-122133[DOI: 10.1109/access.2019.2936914]

Zhu J W, Lai S M, Chen X, Wang D and Lu H C. 2023. Visual prompt multi-modal tracking//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9516-9526[DOI: 10.1109/cvpr52729.2023.00918]

Zhu Y B, Li C L, Luo B, Tang J and Wang X. 2019. Dense feature aggregation and pruning for RGBT tracking//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 465-472[DOI: 10.1145/3343031.3350928]

Zhu Y B, Li C L, Tang J and Luo B. 2021. Quality-aware feature aggregation network for robust RGBT tracking. IEEE Transactions on Intelligent Vehicles, 6(1): 121-130[DOI: 10.1109/tiv.2020.2980735]

Zhu Y B, Li C L, Tang J, Luo B and Wang L. 2022. RGBT tracking by trident fusion network. IEEE Transactions on Circuits and Systems for Video Technology, 32(2): 579-592[DOI: 10.1109/tcsvt.2021.3067997]

作者简介

王炫，男，硕士研究生，主要研究方向为计算机视觉和图像处理。E-mail: xuanwang@zjnu.edu.cn

张大伟，通信作者，男，校聘副教授，主要研究方向为多模态学习、图像处理与计算机视觉。

E-mail: davidzhang@zjnu.edu.cn

阳诚砖，男，副教授，主要研究方向为计算机视觉、模式识别与人工智能。E-mail: czyang@zjnu.edu.cn

王志刚，男，工程师，主要研究方向为计算机视觉。

E-mail: 18957990008@189.cn

陈灏，男，工程师，主要研究方向为计算机视觉。

E-mail: 18957990007@189.cn

郑忠龙，男，教授，主要研究方向为机器学习与视觉、模式识别。E-mail: zhonglong@zjnu.edu.cn