文章编号:1671-4598(2024)09-0001-08 DOI:10.16526/j.cnki.11-4762/tp.2024.09.001 中图分类号:TP391 文献标识码:A

RGBT 多模态视觉跟踪方法综述

杨晓丽,张馨月,于 涛,高 鹏,王茂励
(济宁师范大学 网络空间安全学院,山东 济宁 273165)

摘要:RGBT 视觉跟踪是指融合可见光和热红外多模态图像信息进行视觉跟踪的新兴热点研究课题,合理融合可见光和热红外图像的互补信息可以提高跟踪器的性能和鲁棒性;人工智能技术的发展推动了 RGBT 多模态视觉跟踪的发展,深度学习技术逐渐代替传统目标跟踪方法,在精确度与速度方面更具有优势;对近年来 RGBT 多模态视觉跟踪进行了全面综述,整理了 RGBT 多模态视觉跟踪的发展历程,归纳和讨论了相关算法,具体包括基于相关滤波的方法和基于深度学习的方法;回顾了 RGBT 多模态视觉跟踪数据集的发展历史,介绍了算法性能评估指标,分析了不同方法在评估数据集上的性能,展望了 RGBT 多模态视觉跟踪的未来研究趋势;旨在为相关研究者提供全面的概览和参考,以促进 RGBT 多模态视觉跟踪领域的研究和发展。

关键词:计算机视觉;RGBT 视觉跟踪;信息融合;相关滤波;深度学习

Survey of RGBT Multimodal Visual Tracking Methods

YANG Xiaoli, ZHANG Xinyue, YU Tao, GAO Peng, WANG Maoli
(School of Cyber Science and Engineering, Jining Normal University, Jining 273165, China)

Abstract: RGB thermal (RGBT) visual tracking is an emerging hot research topic on visual tracking, it fuses visible and thermal infrared multimodal image information, and the reasonable fusion of complementary information of visible and thermal infrared images can improve the performance and robustness of trackers. Artificial intelligence technology has promoted the development of RGBT multimodal visual tracking, and deep learning technology gradually replaces traditional target tracking methods, with more advantages of accuracy and speed. Comprehensively overview the development of RGBT multimodal visual tracking, summarize and discuss related algorithms, specifically including correlation filtering-based methods and deep learning-based methods, review the development history of RGBT multimodal visual tracking datasets, introduce algorithm performance evaluation indexes, analyze the performance of different algorithms on evaluation datasets, and look forward to the future research trends of RGBT multimodal visual tracking methods. This paper aims to provide a comprehensive overview and reference for related researchers to promote research and development in RGBT multimodal vision tracking fields.

Keywords: computer vision; RGBT visual tracking; information fusion; correlation filter; deep learning

0 引言

随着人工智能和深度学习技术的发展,视觉跟踪已成为计算机视觉领域中的重要研究方向之一,广泛应用于智能安防[1]、人脸识别[2]、人机交互[3]和自动驾驶[4]等领域。目前,基于可见光(Visible RGB)单模态视觉跟踪方法的研究成果较为丰富,但相关方法容易受到光照变化、背景杂乱等因素的影响,跟踪效果较差。相比之下,热红外(TIR,thermal infrared)图像对光照变化不敏感,尤其具有对雾和霾等恶劣天气的穿透能力。图 1 比较了同一场景下的可见光图像和热红外图像,鉴于 RGB 图像和 TIR 图像存在互补的优势,融合 RGB 与 TIR 的可见光-热红外(RGB-Thermal,RGBT)视觉跟踪方法成为近年来备受关注的研究方向。RGBT 多模态视觉跟踪方法的发展历程如图 2 所示。早在 2004 年就出现了基于 RGBT 信息融合的视觉跟踪算法。然而,早期基于传统机器学习的视觉跟踪算法存在精确度和实时性等方面的缺陷。随着人工智能的不断发展,深度学习相关的研究成果和应用不断涌现,为 RGBT 多模态视觉跟踪方法的设计提供了新的思路和借鉴,越来越多的研究成果发表于国内外高水平期刊和会议。本

(a) 可见光图像 (b) 热红外图像

图 1 RGBT 图像示例

收稿日期:2023-08-29; **修回日期**:2023-10-12。
基金项目:中国博士后科学基金面上资助项目(2023M732022);山东省自然科学基金青年基金项目(ZR2021QF061);广东省自然科学基金面上项目(2020A1515010706);曲阜师范大学科研基金项目(167-602801)。
作者简介:杨晓丽(1998-),女,硕士。
通讯作者:高 鹏(1991-),男,博士,副教授。
引用格式:杨晓丽,张馨月,于 涛,等.RGBT 多模态视觉跟踪方法综述[J].计算机测量与控制,2024,32(9):1-8,35.

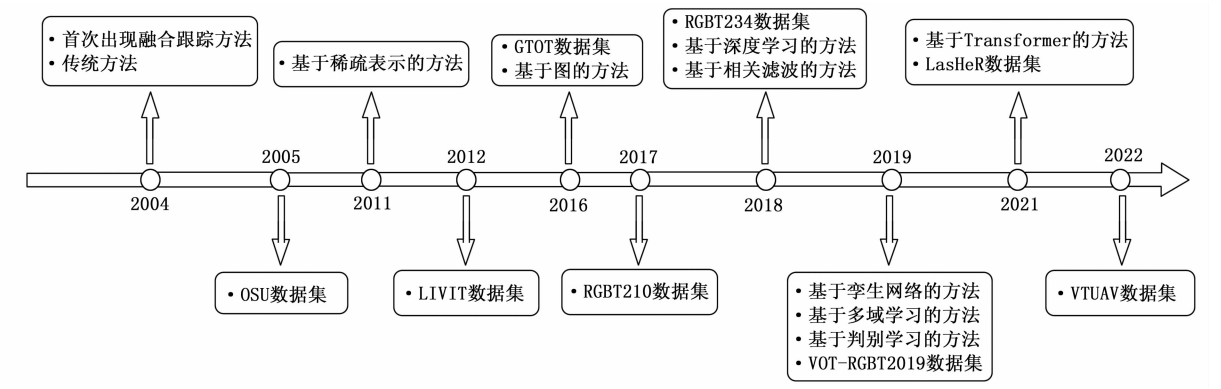

图 2 RGBT 多模态视觉跟踪的发展历程

文总结和分析了 RGBT 多模态视觉跟踪方法研究现状，并展望了未来的研究趋势。

1 RGBT 多模态视觉跟踪的国内外发展现状

1.1 基于相关滤波的方法

文献［5］提出的 MOSSE 模型首次将相关滤波应用于通用视觉跟踪领域。该方法利用输出结果的最小均方误差在初始帧训练滤波模版，后续帧与训练好的滤波器做互相关运算，根据响应值判断相关性，进而确定目标。响应值最大的位置即为目标在当前帧的位置。文献［6］研究了基于相关滤波的 RGBT 视觉跟踪，提出了一种使用 RGB 和 TIR 多模态信息进行视觉跟踪的新型软一致性相关滤波器，并使用快速傅里叶变换（FFT，fast fourier transform）降低了计算开销，构造了一种新的加权机制来融合两种模态信息的响应图，从而得到当前帧的目标位置。该方法的速度为 50 帧/秒（FPS，frames per second），满足实时性要求。文献［7］提出了一种基于交叉模态相关滤波器的 RGBT 视觉跟踪方法，为每种模态设计专门的相关滤波器，并进一步考虑了 RGB 和 TIR 两种模态之间的相互依赖性，引入低秩约束来协同训练滤波模版，设计了一种高效的 ADMM（交替方向乘数法）算法对模型进行优化。然而该方法没有考虑模态权重，容易丢失目标。

除了上述直接使用相关滤波技术进行 RGBT 多模态视觉跟踪的方法外，也有一些研究工作将相关滤波与其他技术相结合进行跟踪。文献［8］等人提出了一种鲁棒融合跟踪方法，由基于相关滤波器的跟踪模块和基于直方图的跟踪模块组成，通过自适应加权机制融合两个模块的响应图得到最终跟踪结果。该方法在确定两个模块的权重时考虑了时间信息，利用 KL 散度来衡量当前帧的响应图和上一帧响应图之间的相关性。为进一步提高相关滤波跟踪器的性能，分别从特征改进[9]、多核算法[10]、分块算法[11]、尺度估计[12]、样本标签[13]和边界效应[14]等方面入手对基于相关滤波的 RGBT 多模态视觉跟踪方法进行改进。

1.2 基于深度学习的方法

近年来，随着深度学习技术在各个领域取得成功，越来越多研究者开始将其应用于视觉跟踪领域。多域学习网络（MDNet，multi-domain learning network）[15]、全卷积孪生网络（Siamese Network）[16]、判别预测框架（DiMP，discriminative model prediction）[17]和 Transformer 框架[18]已成为当前视觉跟踪领域的主导方法。

1.2.1 基于多域学习网络的方法

文献［15］将 MDNet 的概念引入到视觉跟踪中，即将每个序列视为单个域，并将域信息融入到学习过程中，采用在线更新技术获取目标的特定信息，取得了鲁棒跟踪性能。图 3 给出了基于 MDNet 的 RGBT 视觉跟踪方法的总体框架，该方法中的特征提取器和融合模块通过离线训练得到，用于最终分类的全连接层则进行在线训练。

各类基于 MDNet 的 RGBT 多模态视觉跟踪方法的不同主要体现在融合模块。文献［19］提出了一种质量感知的特征聚合网络，采用最大池化操作将每种模态中的分层深度特征转换到相同分辨率，并根据不同模态的可靠性集成不同模态的特征。文献［20］为了利用不同卷积特征的互补性，提出了一种递归融合模块来密集聚合所有模态的特征。但是该方法没有考虑不同卷积层和不同模态特征对整体表征能力的贡献，直接聚合所有模态特征会引入冗余信息和噪声。文献［21］进一步设计了一个轻量级自适应融合模块，融合来自不同模态和卷积层的特征，一定程度上抑制了冗余信息和噪声。文献［22］提出了一种多适配器卷积网络，设计了通用适配器、模式适配器和响应适配器，实现了信息共享和实例感知的特征学习。文献［23］提出了一种跨模态模式传播的跟踪方法，实现了空间跨模态信息和时间上下文传播。文献［24］通过模态感知的注意力网络实现了跨模态信息交互，实现了多模态信息的协同互补学习。

与以上 RGBT 多模态视觉跟踪方法不同，最新的方法可以应对 RGBT 跟踪的模态共享问题（例如，快速移动、尺度变化和遮挡）和模态特定问题（例如，光照变化和热交叉），使得网络的构建更加具体。文献［25］提出了一个引导模块，将判别特征从一种模态转移到另一种模态，提升弱模态的判别能力，所有模态的特征以自适应方式聚合在一起。文献［26］为每个异构属性设计了一个属性驱动的残差分支，从而为不同模态下的特征构建残差表示，并提出了属性集成网络自适应聚合这些表示。文献［27］提

出了一种基于属性的渐进融合网络，根据各个属性分支（例如，热交叉、光照变化、尺度变化、遮挡、快速移动）的自适应融合特性，设计了基于 Transformer 的编解码器来增强聚合特征和特定模态特征。

1.2.2　基于孪生网络的方法

基于孪生网络的视觉跟踪方法是通过相似度度量策略进行视觉跟踪。文献［16］所提出的 SiamFC 模型是一种全卷积孪生网络。该模型通过计算模板帧和搜索帧之间的区域相似度来预测目标的位置。研究人员将该孪生网络从单模态视觉跟踪扩展到 RGBT 视觉跟踪。首先利用特征提取器提取 RGB 和 TIR 模态的特征，然后采用多模态融合模块实现特征聚合，最后采用相似度度量策略进行分类和回归。

实现多模态任务的关键是 RGB 和 TIR 特征的有效结合，如何处理多模态信息的融合是从将视觉跟踪方法从单模态扩展到多模态的关键。文献［28］以 SiamFC 为基础，设计了两个孪生网络分别处理 RGB 和 TIR 图像，然后将两个网络提取的特征进行级联，自适应融合 RGB 和 TIR 图像的特征。文献［29］提出了具有多层融合的动态孪生网络，由动态连体网络处理两个模态的图像，自适应融合多层语义特征。然而，直接将单模态的孪生网络应用于多模态视觉跟踪任务，会出现跟踪性能和鲁棒性差的问题。为此，文献［31］在文献［30］的基础上构建了信息互补和干扰感知的 RGBT 多模态视觉跟踪方法。该方法设计了互补感知多模态特征融合模块来捕获 RGB 和 TIR 特征之间的跨模态信息，同时使用干扰感知区域建议选择模块进一步增强了跟踪鲁棒性。

1.2.3　基于判别预测的方法

基于孪生网络的 RGBT 多模态视觉跟踪方法仅仅使用了目标模板特征，忽略了背景信息的有效使用，因此预测模型存在判别性不足的问题。以文献［17］为代表的基于判别预测的方法改善了这一问题，其设计了一种具有判别能力的损失函数，通过端到端的训练来学习损失函数中的参数，并设计了专门的优化过程，能够在较少的迭代中学习到一个强判别能力的模型。文献［32］将单模态判别预测扩展到 RGBT 多模态视觉跟踪，使用判别损失以端到端的方式进行训练，使得目标预测网络能够学习如何融合来自两种模态的信息，融合机制包括像素级、特征级和决策级。然而，该方法没有使用专门的 RGBT 数据集进行训练，而是通过图像翻译合成的方法将 GOT-10k 数据集合成成对的 TIR 图像。文献［33］提出了一种新的 RGBT 分层多模态融合跟踪方法，将多种模态融合策略统一到一个层次融合框架中。

1.2.4　基于 Transformer 的方法

2017 年，文献［18］完全抛弃了 CNN 和 RNN 等神经网络结构，仅采用注意力机制进行机器翻译任务，并取得了良好的效果。Google 大脑团队提出的 ViT 模型[34]证明了 Transformer 框架不需要依赖 CNN 结构，就可以在图像分类任务上获得很好的效果。这项工作将图片分割成多个块，使用线性嵌入序列作为 Transformer 的输入，效果可以与基于卷积神经网络的先进方法持平。文献［35］将 Transformer 引入目标检测领域，实现了 ViT 中没有完成的多分类任务。具体来说，该方法将卷积神经网络与 Transformer 结合起来直接预测最终的检测结果。通过二分匹配向预测结果分配唯一的真实边界框，没有匹配的预测生成一个无目标的分类预测结果。文献［36］首次将 Transformer 框架引入到 RGBT 多模态视觉跟踪，提出了一种跨模态协作上下文表示的双流混合结构。通过编码器融合不同分辨率下的局部特征和全局表示，并利用空间和通道自注意协作机制获取上下文信息，实现稳定的多模态信息融合。为进一步充分利用 RGB 和 TIR 图像的互补特性，文献［37］提出了跨模态注意网络，为两种模态的特征信息设计了一种基于注意力机制校正的融合模块，不仅能获得丰富的多模态特征信息，而且有效减少了计算冗余。

目前，将 Transformer 框架应用于 RGBT 多模态视觉跟踪的研究较少。研究者通常使用基于 Transformer 的视觉跟踪为基础，将其从单模态扩展到 RGBT 多模态，其中需要解决的关键问题是如何利用注意力机制对两个模态之间的互补信息进行有效的挖掘，以及如何设计融合策略实现两个模态信息的有效结合。最近，文献［38］构建了两个相同的跟踪模型分别处理 RGB 和 TIR 图像，最后对两个模态的信息进行融合。其使用经过预训练的 CVT[39]作为主干网络，对两种模态下的模板图像、在线模板图像和搜索区域图像提取特征，并逐元素取最大值进行融合。当前的研究热点主要集中于借鉴文献［38］的设计思想，构建 RGB 和 TIR 模态下多任务跟踪模型，或者借鉴其他性能出色的 Transformer 跟踪器思想，将其从单模态扩展到多模态跟踪任务，这些都为实现高性能的 RGBT 多模态信息融合的视觉跟踪提供了宝贵的思路和参考。

2　RGBT 多模态视觉跟踪的数据集发展现状

RGBT 多模态视觉跟踪是一项数据驱动的任务，大规模数据集不仅为模型方法的训练提供充足的样本，而且还能对跟踪方法的性能进行测试和评估。2016 年之前，评估跟踪性能的实验部分仅使用几个甚至单个 RGBT 视频，无法覆盖足够具有挑战的复杂场景，如光照变化、相似物体的干扰、物体变形、遮挡等。同时，所使用视频没有统一的标注，数据的准确性和可靠性不高。用于客观公正评估跟踪性能的大规模数据集应该具有以下属性：

1）包含大量对齐的 RGB 和 TIR 视频图像；

2）应给出每一帧视频图像所对应目标的精确标注，即要有显示目标位置和大小的边界框；

3）应涵盖各种具有挑战的复杂场景，如光照不足、快速运动和遮挡等。

本文汇总了 6 个公开的 RGBT 视觉跟踪数据集，分别是 GTOT[40]、RGBT210[41]、RGBT234[42]、VOT-RGBTIR 2019/2020[43]、LasHeR[44]和 VTUAV[33]。每个数据集的详细信息如表 1 所列。

表 1 RGBT 多模态视觉跟踪领域的数据集

数据集	序列数	平均帧数	最小帧数	最大帧数	总帧数/K	分辨率	训练集	长时序列	分割掩码	发布年份
GTOT[40]	50	157	40	376	7.8	384×384	×	×	×	2016
RGBT210[41]	210	498	40	4 140	104.7	630×460	×	×	×	2017
RGBT234[42]	234	498	40	4 140	116.7	630×460	×	×	×	2019
VOT-RGBT2020[43]	60	334	40	1 335	40.2	630×460	×	×	×	2019
LasHeR[44]	1 224	600	57	12 806	734.8	630×480	×	×	√	2021
VTUAV[33]	500	3 329	196	27 213	1 700	1 920×1 080	√	√	√	2021

2016 年，文献 [40] 创建了第一个用于评估 RGBT 跟踪性能的专用数据集 GTOT。该数据集收集了 50 个不同场景的 RGBT 视频序列，如办公区域、公共道路、游泳池等，并且精确标注了目标边界框。由于 GTOT 数据集的视频采集设备较为落后，成像质量较差。文献 [41] 在 2017 年建立了更加全面的 RGBT210 数据集。该数据集包含 210 个 RGBT 视频序列，总帧数将近 21 万，每个视频的不超过 8000 帧。相比于 GTOT 数据集，RGBT210 数据集的视频序列对齐精度高，且对一些遮挡情况进行更为精确的标注。此外，RGBT 还将各种具有挑战性的跟踪场景分为 12 类，如表 2 所列。2018 年，文献 [42] 在 RGBT210 数据集的基础上增加了 24 个新的视频序列，扩展为 RGBT234 数据集。

2019 年，著名的 VOT 视觉跟踪挑战赛引入一个新的赛道，即 VOT-RGBT[43]。该赛道组委会在 RGBT234 数据集中选取 60 个 RGBT 视频序列作为比赛数据集。与 RGBT234 不同的是，VOT-RGBT 使用 TIR 图像作为主要模态，RGB 图像作为辅助模态，评估指标是基于 TIR 图像的真实目标边界框计算得到。

上述数据集虽然推动了 RGBT 视觉跟踪的研究和发展。然而，这些数据集也存在一些缺点。首先，数据集的规模较小，一般用于跟踪性能的评估，不适用于训练；第二，RGBT210 和 RGBT234 中 TIR 图像的分辨率和成像特性相同，这导致在实际应用中难以全面评估跟踪过程中可能遇到的各种分辨率。

2021 年，文献 [44] 提出了一个大规模的 RGBT 视觉跟踪数据集 LasHeR。该数据集包含 1 224 个 RGBT 视频序列，总帧数超过 73 万。视频序列从广泛的对象类别、场景复杂性和跨季节、天气、昼夜的环境因素中进行高度多样化的捕捉，目标边界框采用人工密集标注，显著提升了 RGBT 视觉跟踪方法的训练和评估质量。同年，文献 [33] 构建了一个更大规模且属性类别更多的 RGBT 跟踪数据集 VTUAV。该数据集有训练集和测试集组成，训练集包含 207 个短时视频序列和 43 个长时视频序列，测试集包含 76 个短时视频序列和 74 个长时视频序列，可以应用于短时跟踪、长时跟踪和语义分割预测，以实现更客观公平的性能评估。此外，该数据集还在视频帧图像和序列层次上进行了全面的属性分类，能够满足训练更为鲁棒的跟踪方法的要求。

RGBT 视觉跟踪数据集的日趋完善，必将推动 RGBT 多模态视觉跟踪方法迅速发展。

表 2 RGBT 数据集的各类属性汇总

模态	属性名称	具体含义
RGB	遮挡(OCC,occlusion)	目标未被遮挡(NO,no occlusion)
		目标被部分遮挡(PO,partial occlusion)
		目标被遮挡比例超过 80%(HO,heavy occlusion)
		目标被完全遮挡(FO,full occlusion)
	尺度变化(SV,scale variation)	相邻帧目标真实边界框的缩放比例为[0.5, 1]
	快速移动(FM,fast motion)	相邻帧目标真实边界框的中心距离超过 20 个像素
	低光照(LI,low illumination)	目标区域的光照条件弱
	低分辨率(LR,low resolution)	目标真实边界框内的像素少于 400
	变形(DEF,deformation)	目标发生非刚体变形
	目标模糊(TB,target blur)	运动使得目标模糊(MB,motion blur)
		低光照使得目标模糊(IB,illumination blur)
	摄像机移动(CM,camera moving)	摄像机移动使得成像效果较差
	背景干扰(BC,background clustering)	视频背景中有与目标纹理或颜色相似的干扰物体
	极端照明(EI,extreme illumination)	光照极亮或极暗
	超出视野(OV,out of view)	目标在视频中消失
TIR	热交叉(TC,thermal crossover)	目标与背景中的物体有相似的热辐射
	模态分离(TVS,thermal visible separation)	可见光图像和热红外图像的边界框不重合

3　RGBT 多模态视觉跟踪的关键技术

3.1　图像信息融合

图像信息融合旨在将来自多张图像的信息组合成一张图像，为应用程序提供信息更加丰富的数据源。目前的图像信息融合技术一般可分为像素级、特征级和决策级。图像信息融合技术可应用于多个领域，如医学图像融合[45]、多聚焦图像融合[46]、遥感图像融合[47]、多曝光图像融合[48]、RGBT 图像融合[49]。

3.1.1　像素级融合

像素级融合是将 RGB 和 TIR 两种模态下的图像像素逐一计算融合，得到新的像素值。像素级融合的操作较为简单，但会引入噪声和无用的细节信息，计算量较大，同时两种模态的图像必须保证完全对齐。

3.1.2　特征级融合

特征级融合是通过特征提取网络提取 RGB 和 TIR 图像的特征，然后使用某种融合策略对这些特征进行融合，最后将融合后的特征用于后续视觉跟踪。特征级融合可以减少噪声，且对图像的对齐要求不高。

3.1.3　决策级融合

决策级融合通常由两个特征提取网络或者两个不同的跟踪器分别处理 RGB 和 TIR 图像，得到两种模态的初步跟踪结果，然后按照某种设计好的策略来将这两个初步结果融合，获得最优的结果。决策级融合可以综合利用 RGB 和 TIR 两个模态的图像信息，提高跟踪效果。

3.2　RGBT 多模态信息融合

RGB 图像在恶劣天气或者低光照情况下的成像效果较差，这会导致跟踪效果不佳。TIR 图像是根据物体辐射的红外光获取的图像，对光照变化不敏感，能够弥补 RGB 图像成像效果不佳的劣势。然而，TIR 图像所包含的边缘信息较模糊，与 RGB 图像相比，不能很好地保留物体的纹理和色彩等细节。因此，基于两种信息的互补特性，在 RGB 视频图像进行视觉跟踪的基础上，加入 TIR 视频图像，并通过图像融合的方法将来自两种模态下的图像信息合并到一起，从而为跟踪方法提供更全面、更可靠的表征信息。这种方法被称为 RGBT 多模态视觉跟踪。

RGBT 多模态信息融合亟需解决的关键问题主要有：

1）热交叉。热交叉是指目标与背景的温度或形态接近时，TIR 图像内目标与背景难以区分，当目标与背景轨迹交叉时，无法准确定位目标位置。

2）低光照和极端光照。低光照和极端光照现象都是 RGB 图像受光照条件的影响，无法在夜晚或强光环境下捕获有效的目标信息，造成成像质量差或目标不可见。

3）空间不对齐。由于 RGBT 数据通常由两个不同的成像平台采集，因此成像范围及角度有所差异，已有 RGBT 目标跟踪数据集预处理的第一步就是对多模态图像进行空间配准，但多模态图像的空间不对齐问题在已有的公开数据集上广泛存在，容易影响多模态特征之间的有效交互，并干扰目标定位。

在实际的 RGBT 视觉跟踪场景中，固有挑战和特有挑战通常同时出现。RGB 图像和 TIR 图像对这些通用挑战的影响取决于跟踪场景的特有挑战属性。为了实现准确鲁棒的 RGBT 目标跟踪，算法在设计过程中不仅需要考虑如何应对目标跟踪任务中的通用挑战，还需要考虑如何利用 RGBT 数据的互补信息以应对 RGBT 目标跟踪任务中的特有挑战。

相比传统的 RGBT 视觉跟踪算法，基于深度学习的 RGBT 视觉跟踪算法获益于 CNN 的特征提取和表示能力，获得比传统算法更优的跟踪结果，吸引计算机视觉领域研究人员的广泛关注。多模态应用主要增长原因包括：低成本高质量图像技术的发展，用模态融合可以克服一些传感器图像质量不高的缺陷；信号处理与分析算法的发展，例如稀疏表示，多尺度分解等对融合算法的提升较大；不同应用中获得的互补图像的数量和多样性不断增加，例如遥感领域卫星图像可以获得不同空间位置、时间、频谱的各种图像。

RGBT 多模态视觉跟踪中的关键技术主要有：

1）RGB 和 TIR 图像信息的精准匹配。当前的 RGBT 视觉跟踪方法都要求对 RGB 和 TIR 图像进行精准匹配。

2）提升每种模态信息的可靠性。为了利用 RGB 和 TIR 图像的互补信息，应该确保每种模态下信息的可靠性。

3）有效利用 RGB 和 TIR 互补信息。有效利用互补信息意味着要利用具有强大表征能力的算法提取 RGB 和 TIR 特征，设计高效的融合策略将提取的多模态信息融合在一起，可以在不同级别的策略上进行尝试，选取性能较高的融合方式。

4）多模态信息融合的速度。多模态互补信息的融合需要占用一定的硬件资源，耗费一定的时间，可能会影响 RGBT 视觉跟踪的速度。因此，如何高效融合特征并进行快速融合跟踪，是研究 RGBT 多模态视觉跟踪需要关注的一个重要问题。

4　RGBT 多模态视觉跟踪的评估指标

在 RGB 视觉跟踪领域常用的评估指标包括精确率（PR，precision rate）、成功率（SR，success rate）、鲁棒性（Robustness）和期望平均重叠（EAO，expect average overlap），这些评估指标同样适用于评估 RGBT 视觉跟踪方法。

PR 表示目标中心位置误差（CLE，center location error）在给定的阈值距离内的图像帧数占总帧数的百分比。CLE 表示目标预测中心位置与真实中心位置之间的欧氏距离。不同的阈值导致不一样的百分比，一般阈值设定为 20 个像素。

SR 表示目标的预测边界框与真实边界框之间的重叠区域大于某一阈值的帧数占比。重叠区域的计算公式为：

$$O(B_p, B_g) = \frac{|Area(B_p \cap B_g)|}{|Area(B_p \cup B_{\text{}}g)|} \tag{1}$$

式中，B_p 和 B_g 分别表示预测边界框和真实边界框的像素，

$O(B_p, B_g)$ 取值范围为 [0, 1]，一般阈值设定为 0.5。

5 RGBT 多模态视觉跟踪的方法对比与分析

由于 RGBT234 数据集相比于 RGBT210 数据集扩展了 24 个新的视频序列，因此大部分的 RGBT 多模态视觉跟踪方法使用 RGBT234 进行性能评估，导致 RGBT210 的公开结果较少。数据集 VTUAV 于 2022 年正式公开，因此该数据集的结果也相对较少。

表 3 列出了 16 个跟踪器在 RGBT234 数据集上的 PR/SR 评分结果，包括总体性能和 12 个基于属性的性能，数据结果均来自已有的 RGB-T 融合跟踪新方法文献。从表 3 可以看出，深度学习技术的发展推动着跟踪器的不断进步，这主要得益于深度学习卓越的特征表示能力。

表 4 显示了 19 个具有代表性的深度学习的 RGBT 多模态视觉跟踪方法在数据集 GTOT、RGBT234 和 LasHeR 上的性能评估结果。

根据评估结果，基于 MDNet 的方法普遍获得较好的性能，这主要得益于多域学习网络框架的每个域单独迭代训练学习与领域无关的表示，共享层在每次迭代中更新以自适应的学习特定领域的信息，表现出具有较强的鲁棒性。基于 DiMP 的方法通过为预测器设计判别损失函数，不断优化该函数得到模型优化器，增强模型的判别能力，能更好的将目标与背景区分开来，因此该类方法也取得了较好的跟踪性能。

6 RGBT 多模态视觉跟踪的发展展望

6.1 RGBT 多模态视觉跟踪的数据集发展

6.1.1 更大规模的数据集

现有的 RGBT 视觉跟踪数据集中仅有 VTUAV 数据集包含少量的长时跟踪序列，因此采集更加多样化且同时包含大量短时序列和长时序列的数据集对 RGBT 视觉跟踪领域的发展是十分重要的。

表 3 RGBT 多模态视觉跟踪方法在 RGBT234 数据集上性能评估结果(PR/SR)

跟踪方法	NO	PO	HO	LI	LR	TC	DEF	FM	SV	MB	CM	BC	ALL
文献[28]	84.8/62.0	72.7/50.6	57.6/40.7	68.8/47.4	69.6/46.5	70.7/50.3	69.7/51.8	62.0/42.2	71.1/50.5	59.0/43.3	63.2/45.7	60.5/40.3	68.8/48.6
文献[29]	83.8/64.2	73.5/54.3	59.4/43.4	59.3/42.4	66.4/46.5	70.6/53.0	69.5/53.2	65.3/46.9	72.7/55.5	64.5/48.7	66.4/49.9	57.8/39.3	69.7/51.7
文献[30]	88.4/66.4	84.2/63.9	66.2/48.7	81.8/61.2	70.9/49.9	67.4/47.7	77.9/59.2	61.4/45.3	77.7/59.3	63.6/47.9	73.3/54.7	74.0/52.9	76.0/56.9
文献[15]	86.2/61.1	76.1/51.8	61.9/42.1	67.0/45.5	67.0/45.5	75.6/51.7	66.9/47.3	58.6/36.3	73.5/50.5	65.4/46.3	64.0/45.4	64.4/43.2	72.2/49.5
文献[22]	88.7/64.6	81.6/56.6	68.9/46.5	76.9/51.3	70.8/48.7	75.4/54.3	72.0/52.4	69.4/44.9	77.7/54.2	72.6/51.6	71.9/50.8	73.9/48.6	77.7/53.9
文献[19]	88.2/65.7	86.6/60.2	66.5/45.8	80.3/54.8	75.0/51.0	76.6/54.9	72.2/52.6	68.1/43.6	78.5/56.3	70.0/50.3	72.4/52.3	75.7/50.2	78.7/55.3
文献[20]	90.0/64.4	82.1/57.4	66.0/45.7	77.5/53.0	75.9/51.5	76.8/54.3	71.7/57.8	67.0/44.3	78.0/54.2	65.3/46.7	66.8/47.4	71.7/48.4	76.6/53.7
文献[21]	90.0/63.6	85.9/58.8	68.6/45.9	81.2/54.2	81.8/53.8	81.1/58.3	74.1/51.5	74.0/46.5	79.1/54.4	70.8/50.0	72.3/50.6	79.1/49.3	79.6/54.4
文献[24]	92.7/66.5	81.1/57.2	70.9/48.8	77.7/52.7	78.3/52.3	77.0/56.3	73.1/51.4	72.8/47.1	78.7/56.1	71.6/52.5	71.7/51.7	77.8/50.1	79.0/55.4
文献[26]	91.7/65.8	86.3/61.2	70.8/49.1	80.2/55.1	83.1/55.6	78.9/58.9	74.3/52.9	77.6/50.3	79.0/56.2	72.7/53.0	75.7/53.5	78.9/52.7	80.9/57.1
文献[23]	95.6/67.8	85.5/60.1	73.2/50.3	86.2/58.4	86.5/57.1	83.5/58.3	75.0/54.1	78.6/50.8	81.5/57.2	75.4/54.1	75.6/54.1	83.2/53.8	82.3/57.5
文献[26]	93.2/66.8	85.1/59.3	70.0/48.0	81.0/54.7	82.0/53.9	82.0/53.9	76.2/54.1	73.1/47.0	79.7/56.6	68.3/49.0	75.2/52.7	81.1/51.9	80.4/56.1
文献[52]	92.3/67.1	89.5/63.1	74.5/52.1	85.3/58.7	85.4/57.9	87.2/61.2	77.9/56.5	80.0/52.4	84.6/59.8	77.3/55.9	80.1/57.6	83.8/55.9	83.9/59.3
文献[53]	93.2/69.4	84.1/61.1	67.7/48.3	84.0/58.8	77.1/51.7	74.9/52.6	70.6/52.9	61.0/41.7	83.7/61.6	75.1/54.9	76.2/55.6	68.7/48.5	79.0/57.3
文献[32]	73.6/50.5	71.9/48.3	53.0/33.5	64.7/43.8	62.8/37.5	69.5/44.9	60.7/42.7	55.7/35.4	68.3/47.0	61.9/42.0	61.4/41.9	54.7/33.4	64.6/42.8
文献[36]	90.1/65.8	77.9/54.6	64.5/43.1	76.9/46.8	75.4/49.7	89.4/64.3	67.5/47.0	63.5/43.9	80.5/57.5	68.2/53.5	66.2/48.6	67.4/42.2	73.2/52.6

表 4　RGBT 多模态视觉跟踪方法在不同数据集上的性能评估结果

跟踪方法	GTOT		RGBT234		LasHeR	
	PR	SR	PR	SR	PR	SR
文献[19]	0.891	0.728	0.787	0.553	0.442	0.309
文献[20]	0.882	0.707	0.766	0.537	0.431	0.214
文献[21]	0.891	0.712	0.796	0.544	0.449	0.311
文献[22]	0.894	0.724	0.777	0.539	0.457	0.330
文献[24]	0.880	0.714	0.790	0.554	0.483	0.352
文献[52]	0.909	0.733	0.839	0.593	0.491	0.357
文献[23]	0.926	0.738	0.823	0.575	—	—
文献[25]	0,889	0.717	0.804	0.561	0.451	0.317
文献[55]	0.901	0.723	0.800	0.554	0.467	0.317
文献[26]	0.904	0.739	0.809	0.571	—	—
文献[27]	0.905	0.739	0.827	0.579	0.500	0.362
文献[28]	0.826	0.700	0.688	0.486	—	—
文献[30]	0.877	0.732	0.760	0.569	—	—
文献[56]	0.890	0.695	0.719	0.525	—	—
文献[57]	0.766	0.628	0.567	0.384	—	—
文献[32]	0.594	0.490	0.785	0.559	0.447	0.344
文献[33]	0.912	0.749	0.788	0.568	—	—
文献[37]	0.858	0.702	0.768	0.553	—	—
文献[36]	—	—	0.732	0.526	—	—

6.1.2　更可靠的数据集

大多数的 RGBT 多模态视觉跟踪方法对 RGB 和 TIR 图像的对齐程度有很高的要求。精确的边界框一般需要人工标注，构建数据对齐的大规模数据集需要耗费巨大的代价，因此可以通过设计高效的跟踪方法获得高质量的边界框标注。

6.2　RGBT 多模态视觉跟踪的方法发展

6.2.1　结合深度学习与相关滤波的方法

基于深度学习的 RGBT 多模态视觉跟踪方法在该领域表现出了卓越的性能。然而，该类方法在实时性上仍有改进空间。基于相关滤波的 RGBT 多模态视觉跟踪方法通过傅里叶变换实现了快速的相关操作，具有较高的实时性。在未来，可以将深度学习与相关滤波结合，充分利用两者的优势获得更高性能的 RGBT 多模态视觉跟踪方法。

6.2.2　基于迁移学习的方法

如今大多数高性能的视觉跟踪方法在 RGB 单模态下训练得到。因此，可以在 RGB 单模态视觉跟踪方法的基础上，利用 TIR 图像结合迁移学习进行微调，使其能够跟踪 TIR 视频中的目标。

6.2.3　基于无监督或弱监督的方法

基于深度学习的 RGBT 多模态视觉跟踪方法大多都是使用监督方法进行训练，因为训练需要精确标注的真实边界框。而标注边界框是一项繁琐且耗时的工作，需要耗费大量的人力成本。因此，可以考虑研究基于无监督或者弱监督的 RGBT 多模态视觉跟踪方法。

6.2.4　基于 Transformer 的方法

最近，基于 Transformer 的 RGB 单模态视觉跟踪方法取得了出色的跟踪性能。Transformer 借助于注意力机制，能够有效的建立全局依赖关系，更好的整合空间和时序信息，生成具有强判别性的时空特征。可以将 Transformer 应用到 RGBT 多模态视觉跟踪中，以期取得更好的跟踪性能。

7　结束语

近年来，RGBT 多模态视觉跟踪受到了广泛的关注并取得了显著的进展。本文对当前深度学习的 RGBT 多模态视觉跟踪进行了综述，介绍了 RGBT 视觉跟踪数据集的发展以及常用的评估指标，并对比和分析了不同跟踪方法在几种公开的大规模数据集上的评估结果。根据分析结果，本文最后对 RGBT 多模态视觉跟踪的未来研究方向提出了展望和预期，为后续开展相关研究提供了思路和建议。

参考文献：

[1] ALI A，JALIL A，NIU J，et al. Visual object tracking-classical and contemporary approaches [J]. Frontiers of Computer Science，2016，10：167－188.

[2] WRIGHT J，YANG A Y，GANESH A，et al. Robust face recognition via sparse representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2008，31 (2)：210－227.

[3] MUELLER M，SMITH N，GHANEM B. A benchmark and simulator for uav tracking [C] // Computer Vision-ECCV 2016：14th European Conference，Amsterdam，The Netherlands，October 11－14，2016，Proceedings，Part I 14. Springer International Publishing，2016：445－461.

[4] LAURENSE A，GOH Y，GERDES C. Path-tracking for autonomous vehicles at the limit of friction [C] //2017 American Control Conference (ACC)，IEEE，2017：5586－5591.

[5] BOLME D S，BEVERIDGE J R，DRAPER B A，et al. Visual object tracking using adaptive correlation filters [C] //2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition，IEEE，2010：2544－2550.

[6] WANG Y，LI C，TANG J. Learning soft-consistent correlation filters for RGB-T object tracking [C] //Pattern Recognition and Computer Vision：First Chinese Conference，PRCV 2018，Guangzhou，China，November 23－26，2018，Proceedings，Part IV 1，Springer International Publishing，2018：295－306.

[7] ZHAI S，SHAO P，LIANG X，et al. Fast RGB-T tracking via cross-modal correlation filters [J]. Neurocomputing，2019，334：172－181.

[8] LUO C，SUN B，YANG K，et al. Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme [J]. Infrared Physics & Technology，2019，99：265－276.

[9] DANELLJAN M，SHAHBAZ KHAN F，FELSBERG M，et al. Adaptive color attributes for real-time visual tracking [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2014：1090－1097.

[10] TANG M, FENG J. Multi-kernel correlation filter for visual tracking [C] //Proceedings of the IEEE International Conference on Computer Vision, 2015: 3038 - 3046.

[11] LI Y, ZHU J, HOI H. Reliable patch trackers: Robust visual tracking by exploiting reliable patches [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 353 - 361.

[12] LI Y, ZHU J. A scale adaptive kernel correlation filter tracker with feature integration [C] //ECCV Workshops (2), 2014, 8926: 254 - 265.

[13] BIBI A, MUELLER M, GHANEM B. Target response adaptation for correlation filter tracking [C] //Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part VI 14, Springer International Publishing, 2016: 419 - 433.

[14] KIANI GALOOGAHI H, SIM T, LUCEY S. Correlation filters with limited boundaries [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4630 - 4638.

[15] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4293 - 4302.

[16] BERTINETTO L, VALMADRE J, HENRIQUES F, et al. Fully-convolutional siamese networks for object tracking [C] //Computer Vision-ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14, Springer International Publishing, 2016: 850 - 865.

[17] BHAT G, DANELLJAN M, GOOL L V, et al. Learning discriminative model prediction for tracking [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6182 - 6191.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30.

[19] ZHU Y, LI C, TANG J, et al. Quality-aware feature aggregation network for robust RGBT tracking [J]. IEEE Transactions on Intelligent Vehicles, 2020, 6 (1): 121 - 130.

[20] ZHU Y, LI C, LUO B, et al. Dense feature aggregation and pruning for RGBT tracking [C] //Proceedings of the 27th ACM International Conference on Multimedia, 2019: 465 - 472.

[21] GAO Y, LI C, ZHU Y, et al. Deep adaptive fusion network for high performance RGBT tracking [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[22] LONG C, LU A, HUA A, et al. Multi-adapter RGBT tracking [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[23] WANG C, XU C, CUI Z, et al. Cross-modal pattern-propagation for RGB-T tracking [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 7064 - 7073.

[24] ZHANG H, ZHANG L, ZHUO L, et al. Object tracking in RGB-T videos using modal-aware attention network and competitive learning [J]. Sensors, 2020, 20 (2): 393.

[25] LI C, LIU L, LU A, et al. Challenge-aware RGBT tracking [C] //Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXII 16, Springer International Publishing, 2020: 222 - 237.

[26] ZHANG P, WANG D, LU H, et al. Learning adaptive attribute-driven representation for real-time RGB-T tracking [J]. International Journal of Computer Vision, 2021, 129: 2714 - 2729.

[27] XIAO Y, YANG M, LI C, et al. Attribute-based progressive fusion network for rgbt tracking [C] //Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36 (3): 2831 - 2838.

[28] ZHANG X, YE P, PENG S, et al. SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks [J]. IEEE Access, 2019, 7: 122122 - 122133.

[29] ZHANG X, YE P, PENG S, et al. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion [J]. Signal Processing: Image Communication, 2020, 84: 115756.

[30] ZHANG T, LIU X, ZHANG Q, et al. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32 (3): 1403 - 1417.

[31] LI B, WU W, WANG Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4282 - 4291.

[32] ZHANG L, DANELLJAN M, GONZALEZ-GARCIA A, et al. Multi-modal fusion for end-to-end rgb-t tracking [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[33] ZHANG P, ZHAO J, WANG D, et al. Visible-thermal UAV tracking: A large-scale benchmark and new baseline [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8886 - 8895.

[34] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. ArXiv Preprint ArXiv: 2010.11929, 2020.

[35] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C] //Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I 16, Springer International Publishing, 2020: 213 - 229.

[36] LIU X, LUO Y, YAN K, et al. CMC2R: Cross-modal collaborative contextual representation for RGBT tracking [J]. IET Image Processing, 2022, 16 (5): 1500 - 1510.

(下转第 35 页)